Novel Analysis Approaches to Context-Dependent Molecular Networks

Dissertation

zur Erlangung des Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.) der
Naturwissenschaftlich-Technischen Fakultäten der
Universität des Saarlandes

eingereicht von
Dorothea Emig

Saarbrücken, Februar 2011

Tag des Kolloquiums:	12. April 2011
Dekan:	Prof. Dr. Holger Hermanns
Vorsitzender des Prüfungsausschusses:	Prof. Dr.-Ing. Thorsten Herfet
Gutachter:	Dr. Mario Albrecht
	Prof. Dr. Dr. Thomas Lengauer
Beisitzer:	Dr. Jan Baumbach

Research is to see what everybody else has seen,
and to think what nobody else has thought.

Albert von Szent-Györgyi

Abstract

Proteins are key players in all kinds of biological processes and accurate knowledge of their presence and their interactions is fundamental for understanding the functioning of the cells. Over the last years, many large-scale studies have been performed in order to unravel the complete human interactome. However, the results of these studies usually depend on the cellular conditions, in which the protein interactions were detected. Furthermore, additional biological mechanisms or temporal and spatial constraints contribute to the context-dependent formation of protein interactions.

In this thesis, we focus on different biological aspects that are important for the formation of protein-protein interactions. We first analyze protein interactions in a structural context and demonstrate that interacting proteins may collide in three-dimensional space, rendering the interaction impossible. Second, we investigate the tissue-specific formation of protein interactions. We analyze the ability of different technologies such as microarray platforms and next-generation RNA-sequencing to reliably detect tissue-specific gene expression. We further use gene expression data to identify tissue-specific protein interactions and their functional implications. Finally, we concentrate on protein variants that arise by alternative splicing events. We describe our software DomainGraph that allows for visually exploring protein variants and their interactions in different biological conditions.

Kurzfassung

Proteine übernehmen viele wichtige Funktionen in biologischen Prozessen. Daher ist genaues Wissen über ihre Interaktionen essentiell, um die Funktionsweise von Zellen zu verstehen. In den letzten Jahren wurden viele Experimente durchgeführt, um die Gesamtheit des menschlichen Interaktoms zu ermitteln. Die Ergebnisse solcher Studien sind jedoch abhängig von der biologischen Umgebung, in der die Proteininteraktionen nachgewiesen wurden. Außerdem werden viele Proteininteraktionen aufgrund zeitlicher und räumlicher Einschränkungen nur in einem bestimmten biologischen Kontext gebildet.

In dieser Arbeit betrachten wir verschiedene biologische Aspekte, die eine wichtige Rolle für die Interaktionen zwischen Proteinen spielen können. Zuerst analysieren wir Proteininteraktionen in einem strukturellen Kontext. Wir zeigen auf, dass interagierende Proteine im dreidimensionalen Raum kollidieren können und dadurch Interaktionen verhindert werden können. Des Weiteren untersuchen wir die gewebespezifische Ausbildung von Proteininteraktionen. In diesem Zusammenhang vergleichen wir zunächst Möglichkeiten, Genexpression mit Hilfe verschiedener Technologien wie Microarrayanalyse und Hochdurchsatz-Sequenzierung zu detektieren. Die Ergebnisse dieser Studie benutzen wir, um gewebespezifische Proteininteraktionen zu identifizieren und diese funktionell zu charakterisieren. Im letzten Teil der Arbeit konzentrieren wir uns auf Proteinvarianten, die sich durch alternatives Spleißen ergeben. Wir beschreiben unsere Software DomainGraph, die die visuelle Analyse von Proteinvarianten und deren Interaktionen unter verschiedenen biologischen Bedingungen ermöglicht.

Acknowledgments

First of all I would like to thank my supervisor Mario Albrecht for his continuous support and advice during my PhD thesis. Furthermore, I would like to thank Thomas Lengauer for giving me the opportunity to do my PhD at MPII. Working in such a great environment was an incredible experience.

Furthermore, I would like to thank all my collaborators for so many great discussions: First of all, Melissa Cline who supported my work from the beginning of my PhD studies. A very warm thank-you goes to Nathan Salomonis for many great and fruitful discussions. Furthermore, I would like to thank Jan Baumbach who gave me the opportunity for a wonderful research visit at UC Berkeley, and Oliver Sander, Gabriele Mayr and Karsten Klein for the joint work and for their support. Finally, I would like to say thank you to my students Anne Kunert and Tim Kacprowski for their help.

Special thanks go to my current and previous officemates, Andreas Schlicker, Fidel Ramírez and Gabriele Mayr for many useful discussions and also for all the fun times we had together. Furthermore, I want to thank the other team members, Hagen Blankenburg, Sven-Eric Schelhorn and Nadezhda Doncheva for many great discussions and the good times. Additionally, I am thankful to all members of D3 and of the IRG1 for making this such a great place to work.

Last but not least, I am deeply grateful to my family, especially my Mum, for their continuous and loving support and for always believing in me and encouraging me. I would also like to thank my friends, most of all Laura Dietz, Jan Baumbach and Andreas Schlicker, for making the time in Saarbrücken unforgettable.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The completion of the sequencing of the human genome in 2003 was a milestone towards understanding the principles of the human organism. Almost ten years later, about 30,000 genes have been identified in the human genome, of which about 22,000 genes encode for proteins and about 8,000 for RNAs (Bonetta (2010)). However, we are still far from understanding the functioning of the cells. This is mainly due to the fact that, while the genome is a mostly static entity shared by all cells of an organism, the expression of the gene products, proteins for example, depends on the state and the type of the cells and changes dynamically with time. Alternative splicing of protein-encoding genes further complicates the understanding of cellular processes. Alternative splicing is a process that leads to the expression of different transcript variants from a single gene that, in the case of protein-encoding genes, may lead to different protein variants obtained from a single gene. While it was long known that alternative splicing can increase protein diversity, it was frequently believed that only few genes undergo alternative splicing. With the sequencing of the human genome and transcriptome, however, it became evident that there is a large discrepancy between the number of proteins and the number of protein-encoding genes. Today it is known that more than 90% of all human genes undergo alternative splicing (Wang et al. (2008); Pan et al. (2008)). Moreover, the complexity of the proteome is increased by post-translational modifications such as phosphorylation or glycosylation. Many proteins undergo such modifications, which may alter the functions of a protein (Rogers and Foster (2009)). These modifications are often reversible and depend on the biological conditions in a cell.

Proteins participate in all kinds of biological processes. To understand the cellular dynamics, detailed knowledge on the proteins, for example, on their presence in different tissues or cellular states, the functions of different protein variants, or their abundance,

is crucial. For decades, researchers in molecular biology have focused on the analysis of individual proteins, for instance, elucidating their functions or their structures. In more recent years, genes that are expressed in a certain tissue or a particular cellular state have been identified by measuring all transcripts expressed by the cells under inspection. The knowledge on expressed genes was then used as an indicator for the presence of their protein products in a specific tissue or condition (Bossi and Lehner (2009); Lehner and Fraser (2004)). Commonly used experimental methods for measuring gene expression levels include microarray platforms or, in recent studies, newly developed next-generation sequencing methods. While gene expression levels are a good indicator for the presence or absence of the encoded proteins in the cell, they do not provide a fine-grained view on the proteins, for instance, the presence of particular protein variants, post-translational modifications of the proteins or the protein abundance. However, these details are important for understanding biological processes and even subtle differences in the proteome can have a large impact on the cellular dynamics.

Proteins usually perform their biological functions in concert by forming pairwise interactions or molecular complexes. Dynamic changes in the proteome, for instance, the context-dependent expression of specific protein isoforms or post-translational modifications, however, complicate the identification of their interactions in a specific cellular state. First attempts to unravel the entirety of all protein interactions that are present in the cells, namely the interactome, were based on high-throughput methods such as the experimental yeast two-hybrid screen (Rual et al. (2005)). Although many screens have been performed to date, leading to the detection of numerous protein-protein interactions, a large number of interactions still remains to be discovered and the results of these screens suffer from a high false positive rate (Venkatesan et al. (2009); Deane et al. (2002)). In addition, interactions that result from large-scale screens usually lack a description of their biological context, for example, the assignment of a particular protein isoform to the detected interaction or the tissues, in which the interaction occurs in vivo.

The high-throughput interactome studies have resulted in large protein interaction networks. However, such networks represent a static picture of the detected interactions and do not take the cellular dynamics into account. While a protein interaction may occur in one cellular state, it may not in another. This fact is not reflected by current interactome representations. Initially, interactome research focused on topological network properties, mostly based on graph theory, which do not take dynamic aspects into account (Albert et al. (2000); Jeong et al. (2001)). More recently, researchers started to incorporate context-dependency such as structural and temporal constraints into their interactome studies (Han et al. (2004); Nooren and Thornton (2003b); Bossi and Lehner

(2009)). While this is a first step towards understanding the dynamics of biological processes, this research field is still emerging and dynamic aspects need to be studied in more detail.

In summary, many advances in the detection and analysis of proteins and their interactions have been achieved in the last years, for instance, the identification of tens of thousands of protein interactions from large-scale screens and the discovery of many novel transcript and protein isoforms. We are, however, still far from understanding the biological processes of the cells. In this thesis, we aim at providing new insights into the context-dependency of molecular networks. To this end, we analyzed different spatial and temporal aspects involved with molecular interaction networks as well as their biological relevance. Furthermore, we devised new methods and software for the analysis of context-dependent aspects of the interactome.

1.2 Overview

In this thesis, we develop new analysis approaches to relate protein and domain interactions to the biological context in which they are present. Our first analysis was focused on protein interactions with solved three-dimensional structures. To this end, we developed an approach to the detection of protein collisions that identifies interactions that may be inhibited due to structural collisions of the interacting proteins. We evaluated all structurally solved multi-interface proteins that have the potential of binding to multiple interaction partners simultaneously and put the detected collisions into a biological context.

In addition, we studied the tissue-specific formation of protein interactions and protein complexes as well as their functional properties. In an initial study, we compared the performance of different microarray platforms to next-generation RNA-sequencing technologies with regard to their ability to reliably detect gene expression. Furthermore, we analyzed the impact of inaccurate results on functional studies such as those investigating protein interaction and protein complex formation. We showed that, by performing next-generation sequencing, one is able to detect gene expression even at low levels. We used such data to perform a functional analysis of protein interactions, protein domains, and protein complexes regarding their tissue-specific expression. Based on these analyses, we identified proteins that are only expressed in a specific biological context, which leads to the presence or absence of protein interactions and protein complexes. In addition, we identified biological processes and functions that are preferentially modified in different biological contexts.

Finally, we developed a software framework called DomainGraph, which facilitates the analysis of protein architectures and interactions in the context of naturally occurring protein variants produced by alternative splicing of genes. The software is targeted at both bioinformaticians and biologists. Using DomainGraph, researchers can visually analyze the effects of alternative splicing on proteins and their functional subunits. Furthermore, the potential gain or loss of protein interactions due to alternative splicing in domain-coding regions can be investigated. We applied DomainGraph to two publicly available datasets and identified protein interactions and pathways that can be altered by alternative splicing.

Over the course of preparing this thesis, we published ten papers, which are listed in Appendix B. Seven of these are first-author papers, which are the basis for the thesis and are described in more detail in the following chapters. Part of this work has been financially supported by the Max Planck Society, the German National Genome Research Network (NGFN), and the DFG-funded Cluster of Excellence for Multimodal Computing and Interaction. The research visit to the University of California at Berkeley was additionally supported by the Boehringer Ingelheim Fonds, Foundation for Basic Research in Medicine.

1.3 Outline

The remainder of this thesis is divided into six chapters followed by a list of references and an appendix. Chapter 2 introduces the biological background on genes, proteins, domains, and their interactions and describes different types of alternative splicing events that increase protein diversity. Furthermore, publicly available biological data sources and their scope and limitations are presented. This chapter also provides information on state-of-the-art technologies for measuring gene and exon expression together with methods to process data produced by the respective technologies.

In Chapter 3, we focus on structural characteristics that can prevent the formation of protein interactions. We describe a structural approach to identify protein-protein interactions that may be inhibited due to protein structure collisions in three-dimensional space.

Chapters 4 and 5 concentrate on the context-dependent formation of protein interactions based on tissue-specific gene expression. In particular, Chapter 4 compares different technologies used to measure gene expression. Similarities and differences in gene expression detection results are identified and the functional implications on tissue-specific protein interactions and protein complexes are evaluated. In Chapter 5, we make use of gene ex-

pression estimates obtained with next-generation RNA-sequencing to analyze functional implications of tissue-specific gene expression. In detail, we study the tissue-specific occurrence of protein interactions, protein domains, and protein complexes, and we present biological functions related to tissue specificity.

Chapter 6 examines the impact of alternative splicing on protein and domain interaction networks. We describe the new software DomainGraph, a Cytoscape plugin for the visual analysis of alternative splicing events based on microarray data. Specifically, DomainGraph has been designed for downstream analyses, namely, studying the biological effects of alternative splicing on proteins, protein domains, interaction networks, and pathways.

Chapter 7 summarizes the results and findings of the presented analyses and discusses possibilities for future research directions.

Appendix A provides additional information on the tissue specificity of human diseases and of protein complexes, as identified in the study described in Chapter 5. Finally, Appendix B lists all publications that have been accomplished in the course of this PhD thesis.

Chapter 2

Background

2.1 Proteome and Interactome

In the following, we introduce the biological terms used throughout this thesis. We introduce gene expression and primarily focus on alternative splicing, a process that contributes to the generation of transcript and protein variants from a single gene. Furthermore, we present structural and functional properties of proteins and their interactions.

2.1.1 Gene Expression and Alternative Splicing

Gene expression is a biological process that occurs in every cell to transform the genetic information stored in the DNA into functional gene products. To date, about 30,000 human genes have been identified (Bonetta (2010)). The majority of these genes, about 22,000, are protein-encoding, while the remaining 8,000 genes encode functional non-coding RNAs (ncRNAs) such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), or small RNAs like microRNAs (miRNAs) (Amaral et al. (2008)). Both proteins and non-coding RNAs are essential for the functioning of the cells. However, since the remainder of the thesis mainly deals with proteins and their interactions, we will focus on protein-encoding genes and their expressed products in the following.

The main steps in the gene expression process of protein-encoding genes consist of transcription and translation (*Figure 2.1*). Gene transcription includes the transformation of the genomic information into a mature transcript (the messenger RNA, mRNA), which is converted into a functional protein during translation.

The majority of protein-encoding genes, more than 90%, are composed of alternating stretches of coding and non-coding sequences, that is, exons and introns, respectively. These genes are called multi-exon genes, while the others are single-exon genes. A gene is first transcribed into the primary RNA transcript (heterogeneous nuclear RNA, hnRNA),

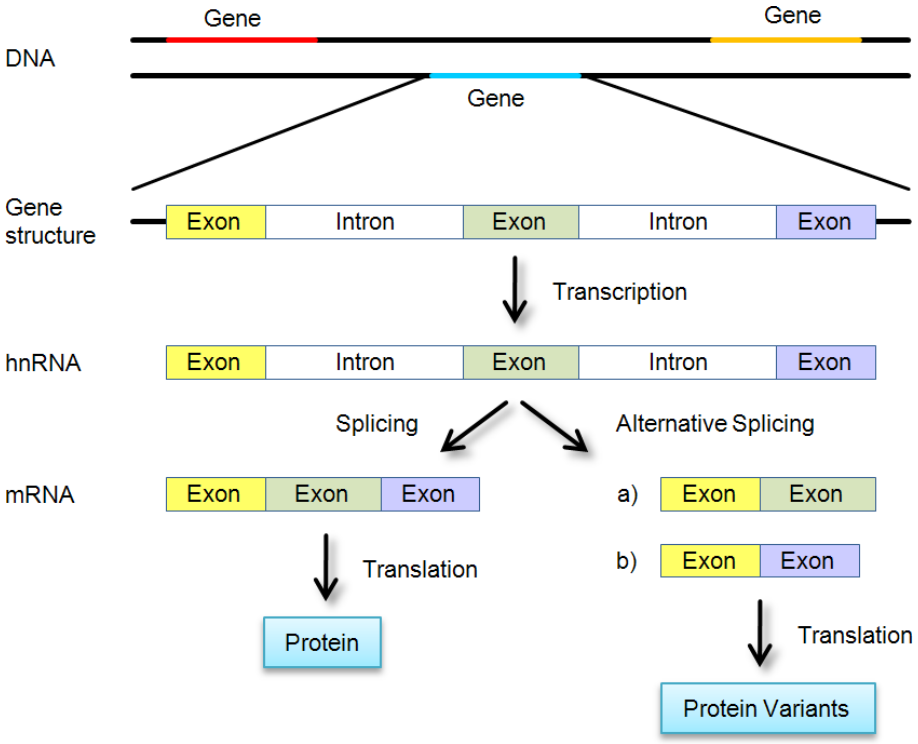


Figure 2.1: *The main steps in the gene expression process for protein-encoding genes. The genes (red, blue, orange) are located on the DNA. As demonstrated for the blue gene, the typical gene structure consists of exon and intron regions. The gene is first transcribed into the premature hnRNA, which contains all exons and introns. Next, the introns are spliced out, which can result in one mRNA (splicing), or in different mRNA variants (alternative splicing). Lastly, the mRNA is translated into a protein.*

which contains all exons and introns of the gene. In case of multi-exon genes, the introns are subsequently spliced out of the hnRNA, resulting in the mRNA. In eukaryotic cells, RNA splicing is a complex process that, for about 98% of the multi-exon genes, leads to multiple mRNA variants per gene. The process of generating different mRNA variants from a single gene is commonly known as alternative splicing (Blencowe (2006)). Here, different combinations of exons are concatenated, leading to transcript and thus protein variants from a single gene. Although the presence of these variants usually depends on the cellular condition or the tissue, the regulation of alternative splicing and the functions of the variants are still poorly understood (Fagnani et al. (2007)).

Alternative splicing events can be divided into several types. The most frequent type

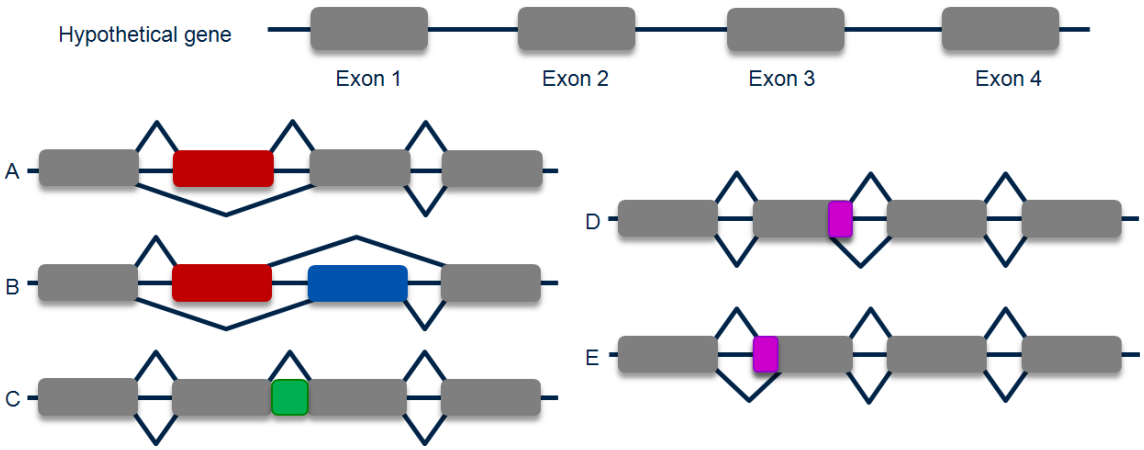


Figure 2.2: *Types of alternative splicing events. The colored boxes indicate exons, the black triangles represent the inclusion and exclusion of DNA regions.* ***(A)*** *shows a cassette exon (red),* ***(B)*** *highlights two mutually exclusive exons (red, blue), and* ***(C)*** *displays a retained intron (green).* ***(D)*** *and* ***(E)*** *show alternative 3' and 5' splice sites, respectively (pink).*

is exon skipping, which includes or excludes an alternative exon, also called a cassette exon, into or from the mRNA (*Figure 2.2* **A**) (Sammeth et al. (2008)). Mutually exclusive exons comprise two or more exons that do not co-occur in the same mRNA (*Figure 2.2* **B**). Instead, each mRNA contains exactly one of the mutually exclusive exons. Another type of alternative splicing is intron retention, which includes an otherwise non-coding genomic sequence into an mRNA transcript (*Figure 2.2* **C**). Furthermore, exon boundaries are not always clearly delimited, resulting in alternative 3'- and 5' splice sites (*Figure 2.2* **D, E**).

Besides alternative splicing, other mechanisms exist that can increase the diversity of mRNA variants: alternative promoters usually result in alternative first exons while alternative polyadenylation sites lead to alternative last exons. Although the outcome is related to alternative splicing and may result in the alternative inclusion of different start and end exons in the mRNAs, they are technically not considered alternative splicing events (Matlin et al. (2005)).

2.1.2 Structural Properties of Proteins

The final products of many genes are proteins, macromolecules that are essential for many cellular processes. They fulfill diverse functions, acting as enzymes, signaling molecules, or structural modules (Lodish et al. (2004)). The folding of a protein into its three-

dimensional structure determines the specific functions of the protein.

The primary protein structure corresponds to the amino acid sequence produced in the process of translating an mRNA into a protein. The amino acid chain, also called polypeptide chain, varies in length and determines the protein fold. The primary structure of the protein forms specific local substructures, the α-helices and β-sheets, which are mainly driven by hydrogen bonds and are known as the secondary structure elements of a protein. In addition to hydrogen bonds, other chemical interactions, such as Van-der-Waals forces, salt bridges, and disulfide bonds, stabilize the folding of the polypeptide chain into an often, but not always, unique three-dimensional structure. If the structure is unique, it is referred to as the tertiary structure of the protein, otherwise the protein is considered disordered (Dyson and Wright (2005)). In soluble globular proteins, hydrophilic amino acids are preferentially located at the protein surface, and hydrophobic amino acids tend to be buried inside the protein core. Other proteins such as transmembrane proteins include special hydrophobic regions to fit into the cell membrane. While many proteins are fully functional once they are folded into their tertiary structures, other proteins consist of more than one polypeptide chain. Here, each of these chains is called a subunit and the permanent aggregation of two or more polypeptide chains into a functional multi-subunit protein is referred to as the quaternary structure. In contrast, a protein complex is an assembly of multiple proteins that can dissociate from the complex. Protein complexes may consist of multiple copies of the same protein, a homomultimeric protein complex, or of different proteins, a heteromultimeric complex.

Modularity is intrinsic to proteins and their structures are best characterized by their subunits, the protein domains (Sonnhammer and Kahn (1994)). A domain is a structural unit of a protein that is usually able to fold independently from the rest of the protein. The length of a protein domain is usually between 50 and 200 amino acids (Chothia and Gough (2009)), folding into a compact and globular three-dimensional structure. Domains are frequently regarded as autonomous in their biological functions but, in multi-domain proteins for instance, they may also act in concert to fulfill a specific function. The vast majority of eukaryotic proteins, more than 80%, are composed of several domains (multi-domain proteins), while most of the remaining proteins, except for disordered ones, contain only one domain (single-domain proteins). *Figure 2.3* shows part of the three-dimensional structure of the *growth hormone binding protein* (GHBP), which is a multi-domain transmembrane protein. Since transmembrane regions are only stable when incorporated into the cell membrane, only the extracellular region of GHBP could be crystallized. This region consists of two domains, the *erythropoietin receptor, ligand binding* domain and the *fibronectin type III* domain, which are connected by a short linker sequence.

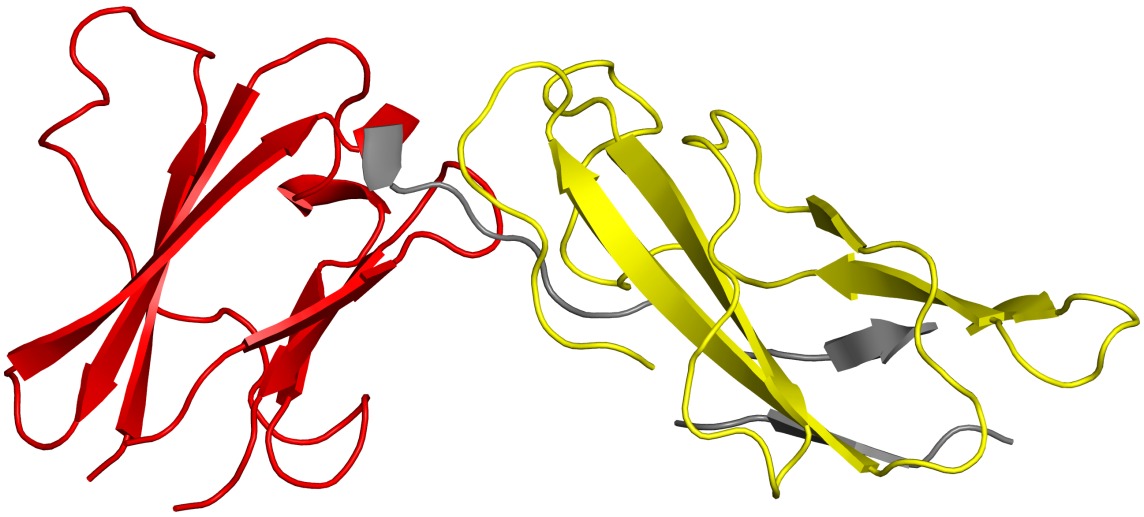


Figure 2.3: *Cartoon representation of the 'growth hormone binding protein' structure (extracellular region, PDB identifier 1hwg, chain B). The protein contains two Pfam domains: the 'erythropoietin receptor, ligand binding domain' (EpoR_lig-bind, PF09067) shown in red and the 'fibronectin type III domain' (fn3, PF00041) in yellow. Protein regions not assigned to domains are colored in gray.*

A characteristic feature of domains is their repeated appearance in different proteins and species (Chothia (1992)). In the course of molecular evolution, nature duplicated existing domains and re-used them in other proteins. Protein structures are thus limited in their modular domain composition, and functional classifications of proteins are frequently based on their domain architecture. While domains are commonly accepted as the structural and evolutionary unit for protein classification, there are different approaches to the definition of a domain, which are primarily structure- and sequence-based approaches. A prominent example for a structural approach is the 'Structural Classification of Proteins' (SCOP) database (Murzin et al. (1995)). SCOP is based on solved protein structures and the domains are manually assigned to the SCOP classification. SCOP employs a hierarchical classification scheme through which users can navigate in a top-down fashion. The topmost level in this classification is *class* (similar secondary structure composition), followed by *fold* (similar secondary structure composition in the same topological order), *superfamily* (homologs with low sequence similarity), *family* (closely related homologs) and finally leads to the structural domains of a single protein in the PDB. In contrast, sequence-based approaches such as Pfam apply sequence profiling and identify domains that are shared by multiple proteins based on their sequence conservation (Finn et al. (2010)). Unlike SCOP, Pfam does not necessarily take three-dimensional information into

account, although sequence conservation does not always reflect the structural or evolutionary relationships. Nevertheless, many molecular databases incorporate Pfam domain annotations and a more detailed description of the Pfam domain annotation method can be found in Chapter 2.1.4.

Domains are often associated with specific biological functions such as catalytic or binding activities. Variations in the domain sequence can result in differences of its three-dimensional structure and can thus impact the domain function (*Figure 2.4*). Protein isoforms, resulting from alternative splicing events during gene expression, are known to vary with respect to their functionalities and isoforms originating from the same gene may even have opposing functions (Stamm et al. (2005)). Alternative splicing can, for example, lead to the expression of long and short protein isoforms, resulting in the gain or loss of a complete protein domain (Resch et al. (2004)). However, it may also alter the sequence of a domain by the inclusion of an alternative exon, resulting in structural differences and the potential gain or loss of important functional residues (Salomonis et al. (2009)).

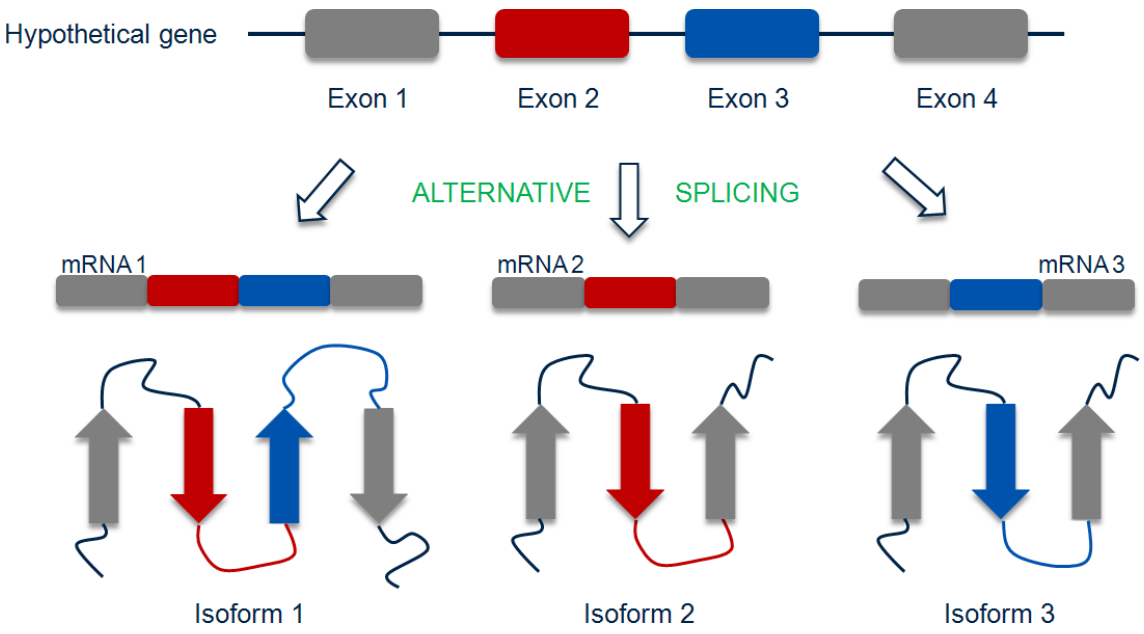


Figure 2.4: *Effects of alternative splicing on the protein structure. A hypothetical gene with four exons is shown. For the gene, three mRNA isoforms are depicted together with their secondary protein structures as ribbon diagrams. Alternative exons are highlighted in red and blue, and structural differences can be seen in the ribbon diagrams.*

2.1.3 Protein Interactions

Most biological processes are carried out by groups of proteins acting in concert, forming pairwise interactions or multimeric complexes. The interactions between proteins can be characterized as obligate or non-obligate. In an obligate interaction, the participating proteins are not stable on their own and can thus not be found as single structures in vivo. Proteins involved in non-obligate protein interactions, however, are stable and functional as single molecules. Obligate interactions are usually very stable, while non-obligate interactions can be further classified into permanent or transient based on their ability to associate and dissociate in vivo. For instance, signaling processes involve many

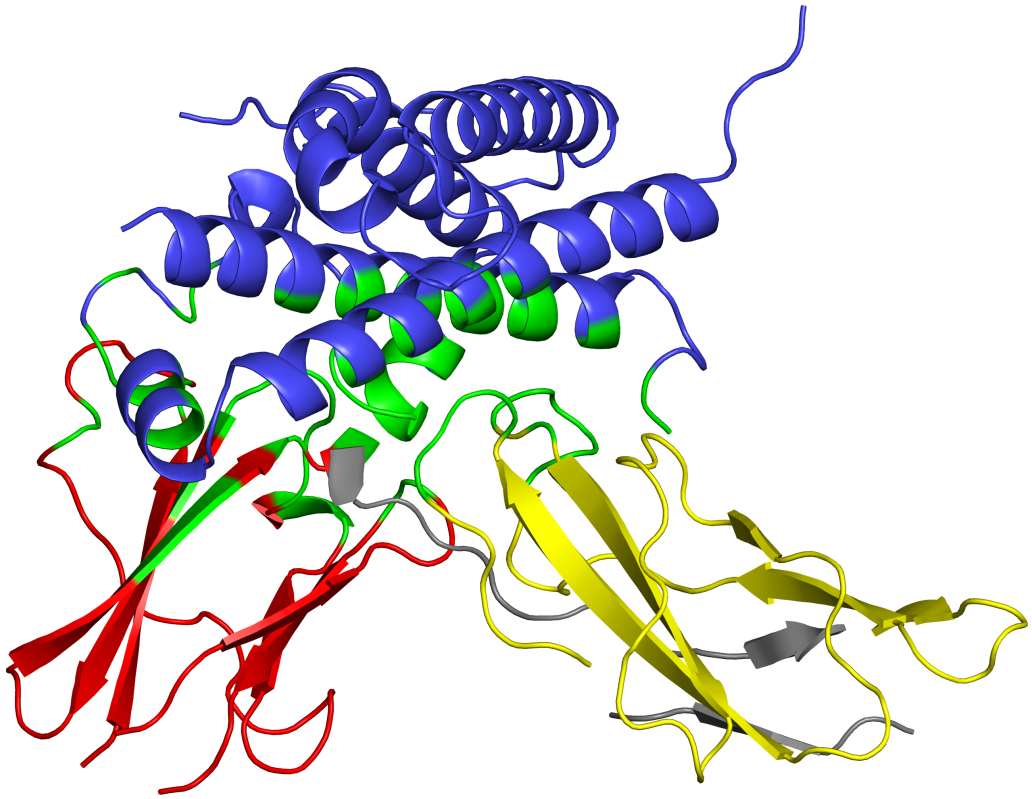


Figure 2.5: *Crystal structure of the 'growth hormone' (GH) in interaction with the 'growth hormone binding protein' (GHBP) in cartoon representation (PDB id 1hwg, chains A and B). The GH contains the 'Somatotropin hormone family domain' (Hormone_1, PF00103) shown in blue. The colors and domains of the GHBP are described in Figure 2.3. The backbone of the interface residues of the protein interaction are highlighted in green.*

non-obligate protein interactions, both transient and permanent (Nooren and Thornton (2003a)). A signaling cascade can be triggered by molecules interacting with a receptor located at the cell surface. In *G-protein-coupled receptor* (GPCR) pathways, for example, the receptor stimulation induces a conformational change of the receptor on the intracellular side. The permanent G-protein complex associated to the receptor is phosphorylated, which results in the dissociation of its α-subunit. The dissociated subunits subsequently start the intracellular signaling cascade via transient interactions with other proteins to communicate the signal towards the nucleus of the cell (Brivanlou and Darnell (2002)).

As described before, proteins are composed of domains and protein interactions are often mediated by domain interactions (Itzhaki et al. (2010)). In particular, many protein domains are able to bind to other domains or peptides permanently or transiently. These domain interactions are known as inter-molecular interactions occurring between two proteins. Besides protein interactions, domain interactions often occur within a single protein, intra-molecular interactions, to stabilize the three-dimensional fold. A domain interaction is established by chemical, mostly noncovalent, bonds between specific amino acids located at the domain surface. The residues involved in the molecular binding are called the binding residues and form the binding interface. *Figure 2.5* shows the protein interaction between the GHBP introduced in *Figure 2.3* and the *growth hormone* (GH). The positions of the interface residues are highlighted and reveal that both domains of the GHBP are involved in the interaction with the domain of the GH. The figure also points out that the binding residues are close to each other in three-dimensional space. However, this does not necessarily imply that they are neighboring residues in the sequence of the protein.

2.1.4 Molecular Databases

Many databases exist, which focus on specific biological aspects, such as genome annotation, gene expression, protein structure, interactions, and pathways. Information from several databases has to be retrieved and integrated in order to obtain a detailed view of a particular gene or protein and its biological functions. In the following, we present an overview about the databases that are most relevant for the analyses described in this thesis.

Ensembl

Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute (Hubbard et al. (2009)). Ensembl is a genome browser, which is focused on the sequences

of large genomes that have become available over the last years. In the process of the sequencing of the human genome, it became obvious that the amount of sequences was too large to manually annotate genes. The Ensembl project was first launched to enable the automatic annotation of the human genome. Over the last years, as more genomes have been sequenced, the database has steadily been expanded and includes more than 30 vertebrate genomes today. The goal of Ensembl is to annotate genes in genome sequences in an automated pipeline, the *genebuild* process, which combines both known genes and predictions of novel genes at high accuracy. Apart from gene annotations, the genebuild pipeline outputs known and putative transcripts together with their underlying genomic exon and intron structures. Furthermore, Ensembl stores information on the resulting protein isoforms as well as on their functional subunits such as protein domains. The transcript, exon, and protein annotations are a main advantage of Ensembl, as they allow researchers to extensively study alternative splicing and their effects on gene products at a genome-wide scale.

With the sequencing of more and more species and the integration of their annotations, Ensembl has now established a basis for the comparative analysis of species, Ensembl Compara. Compara affords the structural comparison of genes from different species, for example, to identify conserved gene regions. A comparative analysis of exon structures can reveal conserved and novel exons and identify exons related to novel gene functions.

UniProt

The UniProt knowledgebase (UniProtKB) is a central resource for high-quality protein sequences and annotations (Apweiler et al. (2010)). Today, UniProtKB consists of three components: UniProt, which stores all protein records with biological annotations, UniParc, which contains a non-redundant set of all currently available protein sequences from a large number of protein databases, and UniRef, which groups proteins according to their sequence similarities and provides representative sets based on different sequence similarity thresholds.

The core of UniProtKB is the UniProt resource, which itself is divided into two components. SwissProt contains all proteins with manually curated annotations, while TrEMBL stores computationally annotated proteins waiting for their manual curation. The protein records in UniProt are annotated according to a standardized set of attributes. These attributes include general information such as the gene name, species of origin, and protein length. Furthermore, functional subunits and posttranscriptional modifications of the proteins are maintained in the database as well as alternative protein isoforms with evidence

at the protein level. In addition, UniProt assigns Gene Ontology (GO) terms and links to PubMed citations underlying the manual assignments (Ashburner et al. (2000)). Aside of their manually curated protein annotations, UniProt provides cross-links to many other biological data sources at external websites. These cross-links include protein structure, protein interaction, and pathway repositories, as well as gene expression, protein domain, and phylogenomic databases.

Protein Data Bank

The Protein Data Bank (PDB) is the central repository for storing structural data of biological macromolecules (Berman et al. (2000)). The PDB currently contains more than 68,000 three-dimensional structures of proteins, nucleic acids, and complexes thereof. The majority of PDB structures (almost 90%) have been resolved with X-ray crystallography, about 10% by NMR spectroscopy, and the remaining with methods like electron microscopy. Both X-ray crystallography and NMR spectroscopy provide three-dimensional coordinates of the molecules at the atom level. X-ray crystallography requires a crystal structure of the molecule, which is exposed to the X-rays causing their diffraction. From the X-ray diffraction pattern, the spatial positions of the atoms can be determined, resulting in the three-dimensional structure of the molecule. However, many proteins contain flexible regions that cannot be crystallized and analyzed with X-ray crystallography. In this case, NMR spectroscopy allows for determining structures in solution. The molecules are exposed to a magnetic field to obtain the atomic coordinates. Although NMR spectroscopy supports resolving flexible structures, it can only handle small molecules up to the size of a small protein domain.

Although the PDB currently stores more than 63,000 protein structures, it contains a lot of redundancy and the number of unique proteins is much lower. Certain proteins and protein complexes have been extensively studied by many researchers and their structures have been solved and deposited multiple times. When clustering all protein chains in the PDB according to their sequence similarity, less than 40,000 clusters are identified using a 100% sequence similarity threshold and the number of clusters reduces to less than 30,000 using a 95% sequence similarity threshold. Hemoglobin, for example, is a very well-studied protein complex, and the PDB currently stores about 500 different hemoglobin structures. Apart from the native molecule, such studies often introduce mutations to the protein to reveal the function of certain residues, and thus not all structures deposited in the PDB correspond to in-vivo molecules and conformations.

Apart from the three-dimensional structures, the PDB database provides additional

information. UniProt sequences are linked to the PDB files as well as to SCOP and Pfam domains, GO terms, and publications. These annotations can be used to analyze functional characteristics of the molecules. For instance, molecular contacts between Pfam domains found in PDB structures have been studied extensively to discover the basis of protein interactions (Finn et al. (2005); Stein et al. (2005)).

Pfam

Motivated by the growing number of available protein sequences, Pfam was developed to provide the framework for an automatic assignment of proteins to families with similar biological functions. Pfam is a widely-used database of protein domain families and classifies proteins based on their sequence similarity (Finn et al. (2010)). Pfam identifies conserved protein subsequences, which correspond to protein domains, and clusters the proteins into families accordingly.

The identification of conserved protein domains is based on multiple sequence alignments and hidden Markov models (HMMs). In the Pfam database, two types of families exist: manually curated Pfam-A domain families and automatically generated Pfam-B domain families. Currently, more than 10,000 Pfam-A families have been derived, covering about 75% of all protein sequences stored in UniProt.

To obtain the high-quality Pfam-A domain families, a small set of representative protein sequences known to share a protein domain is collected. From these representatives, a multiple sequence alignment (seed alignment) is constructed and manually curated. From the seed alignment, an HMM is built for each protein domain family. Finally, the HMMs are applied to all protein sequences available in UniProt in order to identify so far unknown members of the respective domain family.

In the latest Pfam version, about 25% of all protein sequences could not be assigned to any Pfam-A domain family. These sequences are automatically clustered and the resulting clusters correspond to Pfam-B domain families, of which about 140,000 are contained in the current Pfam database.

Protein and Domain Interaction Databases

To date, numerous protein-protein interaction databases have been created. The largest of them is IntAct, which is developed and maintained at EMBL-EBI (Hermjakob et al. (2004)). IntAct comprises a collection of approximately 230,000 literature-derived and experimentally detected protein interactions. Experimental protein interaction datasets

are usually deposited by the authors and further curated by EBI staff members. IntAct provides additional information for each stored protein interaction such as the experimental technique and evidence of physical binding. While large-scale studies based on yeast two-hybrid screens result in physical protein-protein interactions, other experimental techniques such as tandem affinity purification report groups of associated proteins rather than the physical binding.

Other widely-used protein interaction databases include BioGRID (Stark et al. (2006)), HPRD (Peri et al. (2004)), and DIP (Xenarios et al. (2000)). BioGRID is a database that provides access to manually curated and experimentally derived interactions. Unlike the other databases, BioGRID does not only store protein-protein interactions, but also contains information on genetic interactions.

HPRD is a web-based database that provides information on proteins and their features, such as protein interactions, protein domains, post-translational modifications, and subcellular localization. All information stored in HPRD is manually curated and, therefore, the database is much smaller than IntAct and BioGRID, containing information on about 30,000 proteins forming approximately 40,000 protein-protein interactions.

DIP is one of the oldest databases providing information on protein interactions. The developers of DIP aimed at integrating the knowledge on protein interactions, which was scattered in diverse scientific literature, into a single and easily accessible repository. Recently, the DIP developers augmented the set of manually curated entries with protein interactions obtained from high-throughput experiments and now store about 70,000 protein-protein interactions.

As described in Chapter 2.1.3, domain-domain interactions often underlie the formation of protein-protein interactions. Therefore, domain interactions have been studied extensively in order to discover protein domains that are likely to interact and the results of the different methods have been stored in various databases. Two well known approaches to identify interacting domains are iPfam and 3did (Finn et al. (2005); Stein et al. (2005)). Both are based on three-dimensional structures contained in the PDB. They map Pfam domains to the protein structures and compute the molecular contacts between domain pairs that are close to each other in the structure. For both methods, the domain-domain interaction results are provided in databases that are available via comprehensive web servers.

All other techniques for detecting domain-domain interactions are based on computational methods. The most prominent one led to the InterDom database, which contains the largest number of domain-domain interactions (Ng et al. (2003)). The InterDom ap-

proach integrates different data sources, ranging from protein interactions to scientific literature mining to domain fusion events, in order to infer the most probable domain interactions. Other prediction methods are based on a variety of statistical approaches. Several algorithms, such as DPEA, LLZ, and IPPRI, employ maximum likelihood estimations to predict an interaction probability for all pairs of domains (Riley et al. (2005); Liu et al. (2005); Schelhorn et al. (2008)). Other strategies such as DIMA and RCDP are based on phylogenetic profiling (Pagel et al. (2008); Jothi et al. (2006)), the RDFF method makes use of random forests (Chen and Liu (2005)), and LP employs linear programming (Guimaraes et al. (2006)). Both the structure- and prediction-based methods result in datasets containing pairwise domain interactions based on Pfam-A identifiers. While the structure-based approaches provide the domain interactions as they are contained in PDB files, the datasets of predicted interactions additionally assign scores to the putative domain interactions to provide the user with a confidence measure.

Not all of the domain interaction prediction results are accessible via web services, and comparing and evaluating the results of different methods can be tedious. Therefore, several databases have been developed that integrate domain interactions obtained from various methods and data sources. The most comprehensive one is DASMI (Blankenburg et al. (2009)), which stores about 20 different datasets and allows for a straightforward analysis of protein and domain interactions. Furthermore, DASMI enables users to upload and share new domain interaction datasets via a DAS software architecture. Another well known domain-domain interaction database is DOMINE (Raghavachari et al. (2008)), which currently integrates 15 of the domain interaction datasets. Like DASMI, DOMINE enables the user to easily access the data via a web service.

Pathway Databases

In addition to the well known KEGG database (Kanehisa (2002)), there are other widely-used pathway databases such as Reactome and WikiPathways (Matthews et al. (2009); Pico et al. (2008)). Although Reactome and WikiPathways both share the main goal of making pathway-related data publicly available, they differ considerably in their implementation.

Reactome was developed to provide manually curated human pathway data and currently stores approximately 1,000 pathways. New pathways are included into Reactome based on the scientific board members, who decide on which topic or scientific field to focus. In this process, which is comparable to the editorial announcement of a special issue of a journal, Reactome staff work together with independent researchers to create human

pathway models. In addition to the manually curated human pathways, Reactome stores inferred pathway data from other species. To obtain the inferred pathways, orthologs of the molecules involved in human pathways are identified for other species. Subsequently, the pathway reactions from human are transferred to the inferred pathway.

The WikiPathways resource follows a different implementation concept and has been designed in the style of Wikipedia. WikiPathways aims at allowing all researchers not only to download, but also to edit and curate the available pathway data. All changes made to a pathway are tracked and incorrect edits to a pathway can be reversed by other users, a concept that has proven to work well for Wikipedia. While the data quality may theoretically suffer from inappropriate editing, WikiPathways is a very specialized website and data curation is usually performed by experts in the field. Furthermore, curation events take place every few months to check recent changes in the data to ensure the high data quality. Unlike Reactome, there are no topic restrictions, and scientists can include any pathway from any species. Therefore, pathway availability and species coverage depend solely on the research community, their research area of interest, and their willingness to share curated data. Although WikiPathways is a fairly new database, it already contains more than 1,000 pathways in total, covering 19 different species.

2.2 Technologies for Measuring Gene and Exon Expression

The most common and established technique for measuring gene and exon expression are microarrays. Microarrays are plates to which probes in the form of oligonucleotides are synthesized. Affymetrix microarray probes are 25mers, which are designed to exclusively match a specific genomic region. Transcript samples extracted from a cell can be mounted onto the chip. Sequences complementary to a probe hybridize, emitting a signal that corresponds to the gene expression level. Different types of microarrays exist that are used for the detection of gene and exon expression. For gene expression analyses, 3' microarrays are the most conventional platforms that measure the presence of the 3' end of transcripts in a given sample. Microarrays developed more recently are whole-transcript arrays, which contain probes matching all parts of the transcripts. Whole-transcript microarrays are more accurate than 3' arrays since their probe density is much higher and the consequences of alternative splicing events modifying the 3' end of transcripts are less pronounced. The arrays can also be used to measure exon expression since their probes are distributed across all exons of the gene.

Recently, next-generation sequencing methods have been used to measure gene expression at the transcript level. This very promising new technology makes use of massively

parallel sequencing and allows for sequencing complete transcriptomes. High-throughput sequencing methods result in millions of reads, which allow for the detection of gene and exon expression at high accuracy as well as for the reconstruction of specific transcript isoforms. However, next-generation sequencing is very expensive with approximately 20,000 USD per sequencing run, making microarrays a less accurate, yet comparatively cheap, platform.

In the following, microarray and high-throughput RNA-sequencing technologies are detailed as well as methods to obtain gene and exon expression estimates.

2.2.1 Affymetrix 3' IVT Array

The Affymetrix 3' IVT Array is a microarray that has been designed for gene expression detection by measuring the 3' ends of expressed transcripts. It is a whole-genome array, which is available for about thirty different species. A well known representative of this microarray series is the HG-U133A array, which has been designed for the analysis of gene expression in human cells and has been employed in large-scale studies such as the Novartis Gene Atlas (Su et al. (2004)). The microarray probes have been designed to perfectly match the 3' ends of the transcripts and are grouped into probesets. The HG-U133A microarray contains approximately 22,000 probesets, and at most two of them target the same gene. As can be seen in *Figure 2.6*, transcript expression can be measured only if the expressed transcript isoform contains the exons at the 3' end for which the probes have been designed. Gene expression results are thus dependent on the inclusion or exclusion of the 3' end in the predominant transcript isoforms present in the experiment.

The probes contained on a 3' microarray follow a perfect-match/mismatch probe model. A perfect-match probe is designed to perfectly match the transcript sequence. The corresponding mismatch probe is identical with the perfect-match probe except for a one base substitution in the middle of the probe sequence. While preceding array designs required 16-18 perfect-match/mismatch probe pairs per probeset, the 3' Array contains a reduced number of 11 probe pairs per probeset while maintaining the detection accuracy. The decreased number of probes per probeset is rendered possible due to an improved probe selection method (Affymetrix (2010d)). Ideally, a perfect-match probe would provide an exact measure of the transcript expression level and the corresponding mismatch probe would show no expression signal at all. However, this ideal scenario is rarely achieved, for example, due to cross-hybridization events. Therefore, mismatch signals are treated as noise, and the perfect-match probes are adjusted according to the noise level.

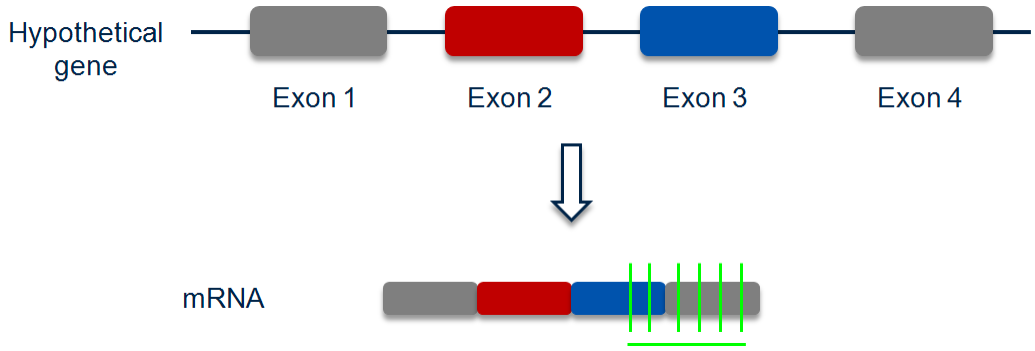


Figure 2.6: *Design of the conventional Affymetrix 3' microarray. A hypothetical gene consisting of four exons is shown. Constitutive exons are shown in gray, alternative exons in blue and red. The probes are designed to match the 3' end of the transcript, without taking potential alternative splicing events into account. The array probes are shown as vertical green lines. The horizontal green line indicates the grouping of the probes into a probeset.*

Processing of 3' IVT Array Data

The MAS5.0 algorithm can be used to infer the presence or absence of a gene in an experiment (Affymetrix (2010e)). More precisely, MAS5.0 computes a detection call for each probeset, which can be a presence call (P-call), marginal call (M-call), or absence call (A-call). While a P-call indicates that the transcript is expressed in the analyzed microarray experiment, those with an assigned A-call could not be detected above background, and M-calls highlight the uncertainty of the detection. MAS5.0 is included in several programs such as the Affymetrix Expression Console (EC) (Affymetrix (2010c)).

MAS5.0 first compares all probe pairs in a probeset and generates a discrimination score (DS) based on the intensity of the perfect-match (PM) and mismatch (MM) probe intensities. For the i-th probe in a probeset, the DS is calculated as follows:

$$DS_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

The discrimination score reflects whether a perfect-match/mismatch pair equally hybridizes to the given sample or not. A discrimination score close to zero indicates that the intensities of the probes are very similar and thus not valuable, whereas a score close to 1 highlights the difference of the probe intensities. To compute the detection p-value for a probeset, the Wilcoxon Signed Rank Test is employed. The Wilcoxon Test first ranks the discrimination scores of the probes in a probeset according to their distance to a

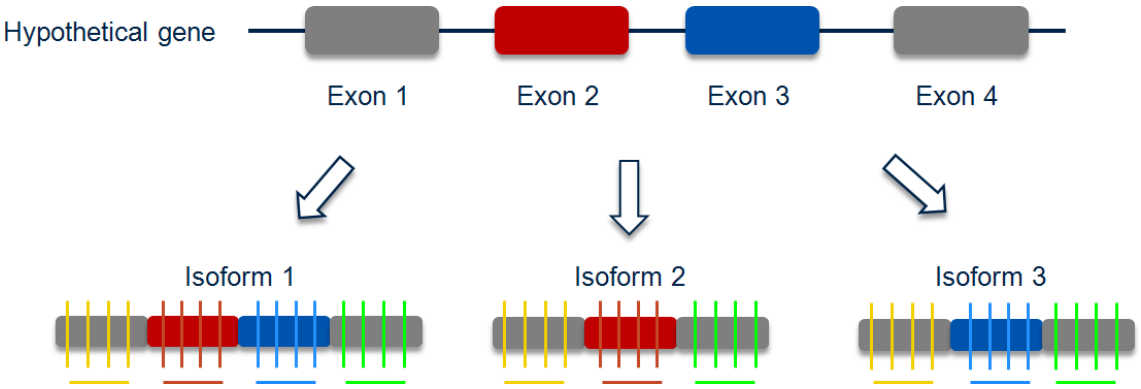


Figure 2.7: *Design of the whole-transcript Affymetrix Exon Array. A hypothetical gene with four exons encoding for three transcript isoforms is shown. The array probes are designed for all exons and are displayed as vertical lines together with the corresponding exons. The probes are grouped into probesets shown as horizontal lines below the exons. The gray exons are constitutive exons, red and blue represent cassette exons.*

pre-defined threshold parameter τ. Based on the sum of the positive ranks, the Wilcoxon Test assigns a p-value to the probeset. From this p-value, the corresponding detection call is inferred by introducing two significance levels α_1 and α_2 as boundaries: a P-call is assigned if the detection p-value is smaller than α_1. An M-call is assigned if the detection p-value lies between α_1 and α_2, and an A-call is assigned otherwise.

2.2.2 Affymetrix Exon Array

The Affymetrix Exon Array is a whole-transcript microarray, which has been developed to measure transcript expression at the exon level for the identification of alternative splicing events. The Exon Array is available for human, mouse, and rat. For this platform, the microarray probes have been designed to uniquely match all known and predicted exons with four probes per exon on average. The probes are grouped into probesets, usually containing four probes, and each exon is covered by at least one probeset. The Exon Array is a high-density microarray and contains about 1.4 million probesets for human, 1.2 million for mouse, and 1.0 million for rat. *Figure 2.7* shows a hypothetical gene consisting of four exons that encodes for three different transcript isoforms. Since the Exon Array probes are designed for all exons, the microarray is able to measure the expression of all transcripts independent of the predominant isoform.

While the Affymetrix 3' Array contains pairs of perfect-match/mismatch probes, the

Exon Array probes have been designed as perfect-match probes only. To compensate for the lack of mismatch probes, a collection of background probes is placed on the Exon Array. These background probes are not designed to match a specific perfect-match probe, but instead contain the same G-C content as the perfect-match probes for which they were designed. The G-C content of the background probes may thus vary from 0 to 25 bases, since all probes are oligonucleotides of length 25. The control probes are binned based on their G-C content and about 1,000 probes were designed for each bin. The median expression of the background probes with a certain G-C content is used for the background correction of the perfect-match probes. One major advantage of the background correction of the Exon Array compared to the perfect-match/mismatch probes model is the space efficiency, which is a fundamental requirement for high-density microarrays. An analysis of the G-C background correction compared to the perfect-match/mismatch model showed that the error rates are comparable and thus the G-C background is suitable for the detection of probeset signals (Affymetrix (2010b)).

Processing of Exon Array Data

Statistical methods such as the MAS5.0 algorithm cannot be applied to Exon Array data, since these methods require a perfect-match/mismatch probe model for the background correction of the probe signals. Instead, a frequently used method for the processing of Exon Array data is the 'robust multi-chip analysis' (RMA) method (Irizarry et al. (2003)), which is included in programs such as APT (Affymetrix (2010a)) and AltAnalyze (Salomonis et al. (2009)). RMA does not require a perfect-match/mismatch probe model and allows for a background correction based on control probes with a certain G-C content. By default, RMA outputs probeset expression signals that usually correspond to exon expression signals. However, RMA can also be used to infer gene expression levels by combining all probes that map to a certain gene into a single meta-probeset and summarizing their expression signals. After RMA is completed, the 'detection above background' (DABG) method computes a p-value for each probeset reflecting the reliability of the probeset signal (Okoniewski and Miller (2008)).

Identification of Alternative Splicing Events

The **Affymetrix Power Tools (APT)** are a collection of command line scripts for the statistical analysis of Affymetrix GeneChip microarray data including Exon Array data (Affymetrix (2010a)). One of the built-in methods, apt-probeset-summarize, enables users to run an RMA summarization on their raw Affymetrix CEL files to obtain probeset

expression values. Furthermore, APT includes the DABG method, which outputs the corresponding p-value for each probeset. The probeset expression values and their p-values can be used to infer occurrences of alternative splicing in a *single-array analysis*. In this type of analysis, probesets detected above background are regarded as present, while unreliable probesets are regarded as absent, similar to the calls computed by MAS5.0. Probeset presence and absence is then used to identify exons that potentially undergo alternative splicing.

AltAnalyze is a program that has been developed for the statistical analysis of raw Affymetrix CEL files (Salomonis et al. (2009)). It allows for a *comparative analysis* of pairs of biological groups, namely, the experimental group and the control group. These groups represent two biological conditions, such as healthy and diseased cells or different developmental stages of the cells, in order to identify probesets that are differentially expressed in the two groups. AltAnalyze includes APT and uses the apt-probeset-summarize method for the processing of the raw Exon Array CEL files. After the raw data have been processed, probesets, which are differentially expressed in the two biological conditions, can be identified. A well-established method for the detection of differentially expressed probesets is the Splicing Index method (Srinivasan et al. (2005)), which is employed in AltAnalyze. The Splicing Index method performs a pairwise comparison of the probesets in the two biological groups to identify exon-level fold changes. The Splicing Index (SI) is calculated as the log ratio of the normalized intensities (NI) of the expressed probesets:

$$SI(probeset_i) = \log_2 \left(\frac{NI(probeset_i)_{group1}}{NI(probeset_i)_{group2}} \right)$$

Here, the normalized intensity of a probeset is calculated as the probeset intensity, i.e. the probeset expression signal, normalized by the expression level of the corresponding gene, i.e. the gene expression signal of the gene to which the probeset belongs. The gene expression levels are based solely on constitutive exons so that gene expression levels are independent of putative alternative splicing events:

$$NI(probeset_i) = \frac{probeset\ intensity}{expression\ level\ of\ gene}$$

To remove false positive results, AltAnalyze performs a two-tailed t-test to calculate an associated p-value. The output of AltAnalyze is a text file containing all probesets with their SI values and corresponding p-values. The default p-value threshold is set to 0.05 as recommended by Affymetrix, and probesets with a p-value below the threshold are regarded as significantly up- or down-regulated in one of the two biological groups.

2.2.3 Next-Generation Sequencing

Next-generation RNA-sequencing is a novel technique that allows for sequencing whole transcriptomes in a high-throughput fashion and can be employed to measure expression at the transcript and exon level. *Figure 2.8* shows an overview of a typical RNA-sequencing experiment. The transcriptome is extracted from a sample and fragmented into pieces of lengths between 200 and 500 bases. These fragments are then amplified and converted into cDNA using RT-PCR, and primers are added to the cDNA sequences. These cDNA fragments are sequenced in a high-throughput fashion, resulting in the sequencing reads. There exist two different sequencing approaches: the single-end sequencing technology, which sequences the cDNA starting from one cDNA primer, and paired-end sequencing technologies, which start sequencing from both primers of the cDNA. While the cDNA fragments can vary in length, the read lengths are pre-defined for a sequencing run. Today, most published transcriptome analyses are based on the Illumina Genome Analyzer and contain about 40 million reads per sample, each of length 32 bases (Pan et al. (2008); Wang et al. (2008)). However, the latest RNA-sequencing technologies can already sequence reads of up to 150 bases and even the sequencing of complete transcripts might become feasible in the next years.

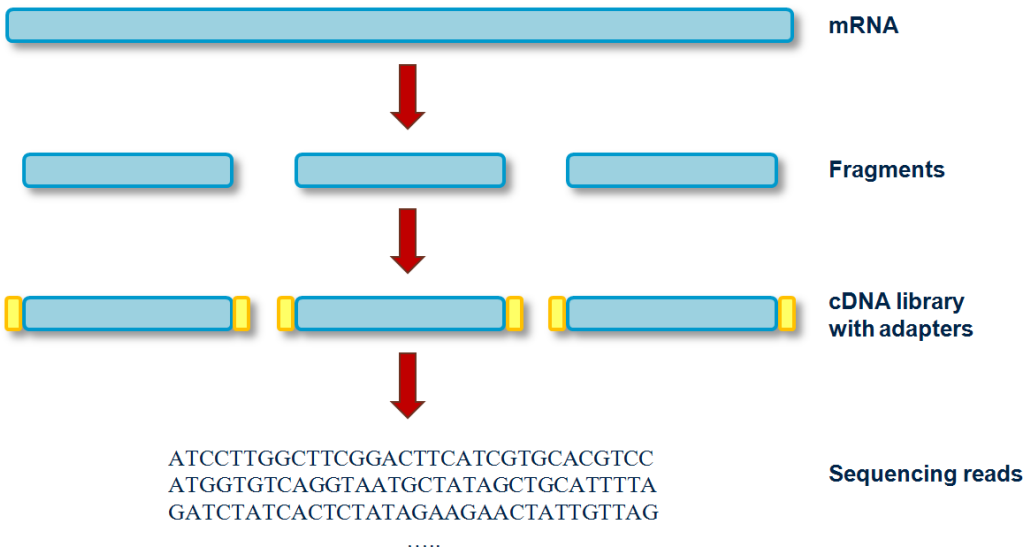


Figure 2.8: *Overview of the next-generation RNA-sequencing technology. First, the transcriptome is fragmented. Second, the fragments are converted into cDNA to which adapter sequences are added. Third, the cDNA library is sequenced, resulting in the sequencing reads.*

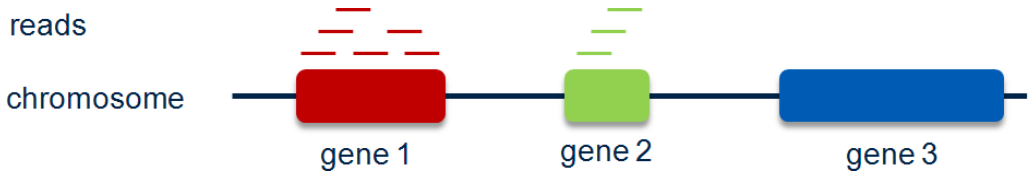


Figure 2.9: *Obtaining gene expression estimates from RNA-sequencing data. Three hypothetical genes are shown together with the RNA-sequencing reads that are uniquely mapped to the respective genes.*

Estimating Expression Levels

Gene expression levels can be estimated from RNA-sequencing reads by aligning them to a reference genome and counting the reads that were mapped to genes (*Figure 2.9*). Estimates of gene expression levels are usually given in a unit called 'reads per kilobase of transcript model per million mapped reads' (RPKM), which was first introduced by Mortazavi and colleagues (Mortazavi et al. (2008)). The RPKM measure is a simple metric, which counts the number of reads that are mapped to a gene, and normalizes the count by the overall length of the gene times the total number of reads that could be mapped to the genome. A recent analysis of gene expression estimates obtained from RNA-sequencing data showed that genes with an RPKM value above 0.3 should be considered expressed (Ramskold et al. (2009)).

Besides gene expression, RNA-sequencing is also useful for the identification and analysis of alternative transcripts produced by a single gene. Conventional microarray technologies such as the Affymetrix Exon Array are limited in their ability to identify alternative transcript isoforms. The Exon Array does not contain exon junction probes, which makes it challenging to infer the correct combinations, in which the exons are connected and expressed. Furthermore, the array probes are pre-defined and do not allow for identifying novel alternative exons. In contrast, RNA-sequencing reads comprise both exon and exon junction reads, which may originate from both known and unknown exon sequences. This combination of reads allows for reconstructing the transcripts present in a given sample and for identifying novel exons and exon junctions. Recent analyses using RNA-sequencing data obtained from human tissue samples identified new splice junctions in about 20% and alternative splicing events in about 95% of all human multi-exon genes (Pan et al. (2008)).

Chapter 3

Structural Aspects of Protein Interactions

Protein-protein interactions take place at defined interfaces. One protein may bind two or more proteins at the same time via different interfaces. So far it has been commonly accepted that non-overlapping interfaces allow a given protein to bind other proteins simultaneously. However, even if multiple non-overlapping interfaces exist, there is still the possibility that collisions in three-dimensional space occur and prevent simultaneous protein binding. In the following, we explore all currently known structures of protein interactions to investigate if three-dimensional collisions are a biologically relevant mechanism for inhibiting protein interactions. A paper describing the results of our analysis has been submitted recently (Emig et al. (2011)).

3.1 Introduction

Most molecular processes involve interactions between proteins. The physical contact between protein interaction partners is formed at defined binding interfaces, and one protein may bind various interaction partners at the same interface or at different interfaces. Due to the increasing number of protein structures available in the PDB (Berman et al. (2000)), systematic protein interaction studies that integrate structural information have become more and more attractive (Aloy and Russell (2006); Devos and Russell (2007); Kiel et al. (2008); Keskin et al. (2008)).

It has been a commonly accepted assumption that a protein containing multiple, non-overlapping interfaces can always interact simultaneously with other proteins. As part of a large-scale structural analysis of a protein interaction network in yeast, Kim and colleagues presumed that the number of simultaneous interactions, in which a protein can participate, is determined by the number of its non-overlapping binding interfaces (Kim et al. (2006)). To this end, the authors gave a structure-based definition of single- and multi-interface proteins and found differences in expression profiles and evolutionary

rates. Subsequently, Kim *et al.* investigated the role of disorder in structural networks and discovered that disordered interface regions are more common in single-interface than in multi-interface proteins (Kim et al. (2008)). Other network-based studies also included structural information into the analyses to increase the informative value of a given network or the reliability of methods predicting protein interactions (Campagna et al. (2008); Aloy et al. (2004)). All of them were based on the assumption that two or more proteins can always interact simultaneously with another protein if the interactions take place at different interfaces.

Further protein network analyses concentrated on various aspects of single- and multi-interface proteins, ranging from protein interaction partners to interface specificity and interaction motifs. For instance, Keskin and Nussinov studied multi-specific interfaces known to bind proteins with different structures (Keskin and Nussinov (2007)). They primarily focused on the ability of the same binding interface to form interactions with different proteins and identified key residues potentially responsible for binding. In a related study, Humphris and Kortemme analyzed restrictions imposed on the protein sequences for permitting multiple binding partners and predicted residues essential for the respective interactions (Humphris and Kortemme (2007)). Aragues and colleagues analyzed hub proteins, i.e., highly connected proteins, in the context of interaction motifs (iMotifs) (Aragues et al. (2007)) and compared their results to those previously found by Kim *et al.* (Kim et al. (2006)). The iMotif approach is based on the idea that proteins sharing interaction partners most likely interact with them via the same binding sites. Clustering proteins according to their interaction partners showed that the number of identified iMotifs correlated with the number of protein interfaces in the work by Kim *et al.* (Kim et al. (2006)). Aragues and coworkers also found that the essentiality of a gene and the gene conservation correlate better with the number of these iMotifs than with the absolute number of interactions. Furthermore, Tuncbag *et al.* presented a concept introducing the time dimension into the analysis of protein interaction networks using protein structures and interface information, which was utilized for the characterization of interactions in the p53 pathway (Tuncbag et al. (2009)). This work highlights the fact that the formation of simultaneous protein interactions depends on various factors including temporal aspects, which should be considered in the analysis of protein interaction networks.

To our knowledge, however, the above-described basic assumption has never been investigated that simultaneous interactions at different interfaces are always spatially possible. In detail, two or more binding partners R and S of a protein P might collide in three-dimensional space, which would prevent the simultaneous interaction of R and S with P

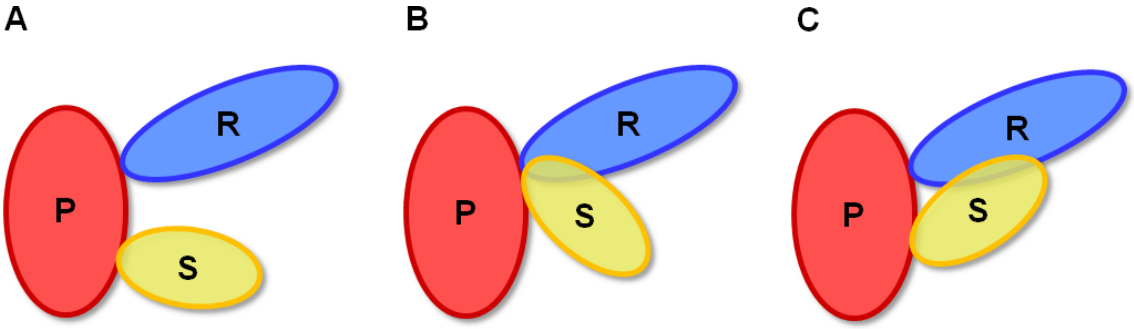


Figure 3.1: *Schematic overview of the structurally possible and impossible interactions between three proteins P, R, and S.* ***(A)*** *The three proteins P, R, and S interact simultaneously via two distinct binding interfaces at P.* ***(B)*** *R and S cannot interact with P at the same time due to the overlapping binding interface at P.* ***(C)*** *Although R and S interact with P via separate binding interfaces, their simultaneous interaction with P is prevented by a collision of R and S.*

even though the binding sites are non-overlapping (*Figure 3.1*). Therefore, we developed a structure collision approach for interactions between protein structure chains in the PDB to examine spatial conflicts between interaction partners.

3.2 Protein Structure Collisions

In this study, we investigated whether a protein P can simultaneously bind two different proteins R and S at distinct binding interfaces. We refer to protein P as the primary protein, while its interaction partners R and S are the secondary proteins. In principle, we regarded all known protein structures that contain an interaction between proteins P and R in one structure and between proteins P and S in another structure, requiring that R and S were bound to P at different interfaces. After the two primary proteins P of the pairwise protein interactions *(P, R)* and *(P, S)* were superimposed, a collision detection method was applied to identify structure collisions between otherwise simultaneously possible interactions of the three proteins (see *Figure 3.1* for simultaneously possible and impossible interactions).

3.2.1 Identification of Colliding Protein Interaction Pairs

The generation of the results proceeded in four main steps (see *Figure 3.2*). First, we identified all potential pairs of primary proteins, that is, all pairwise combinations of

protein chains with identical UniProtKB accession numbers that were contained in at least two PDB structures and could serve as the primary proteins P of the interaction pair *(P, R)* and *(P, S)*. In this first step, we did not consider the interaction partners of the primary proteins but aimed at selecting those proteins with multiple occurrences in the PDB. We found 4,832 proteins that were contained in at least two PDB files out of a total of 17,213 PDB files. This resulted in 1,145,086 possible combinations of potential primary proteins since many PDB files contain multiple copies of the same protein and the number of possible combinations of primary proteins grows quadratically with the number of proteins. Second, to obtain the interaction pairs, we selected those primary proteins that interact with at least two different secondary proteins. When examining the primary proteins and their respective secondary proteins, we detected 2,309,561 interaction pairs with different secondary proteins according to their UniProtKB accession numbers. Third, we compared the interface residues forming the interactions *(P, R)* and *(P, S)* in order to remove those interaction pairs with overlapping interfaces. After this step, 551,944 interaction pairs with distinct interfaces remained, which could be assigned to 6,691 PDB structures. Finally, all these interaction pairs were used as input for the collision detection method, and the volume overlap of the secondary proteins was computed for each interaction pair.

3.2.2 Distinct Protein Binding Interfaces

In detail, we first retrieved all protein structure files from the PDB (Berman et al. (2000)). In case of NMR entries we used the representative protein structure, which is provided in the corresponding PDB file. We identified the binding interface residues between all pairs of interacting protein structure chains by means of the SPPIDER web service (http://sppider.cchmc.org/) (Porollo and Meller (2007)). SPPIDER takes the PDB structure of a protein complex as input and identifies all binding interface residues between pairs of protein chains based on the change of the relative solvent accessibility (RSA) in the unbound protein chains and the complex.

Then we annotated all PDB chains with UniProtKB accession numbers using the mapping provided by PDBSWS (Martin (2005)). We used the resulting annotations to identify pairs of protein interactions consisting of *(P, R)* and *(P, S)*, where the UniProtKB accession numbers of the primary proteins P were identical for both interactions while the UniProtKB accession numbers of the secondary proteins R and S were different.

We compared the binding interface residues of each protein interaction pair to find pairs with overlapping or distinct interfaces. The binding interfaces of P in the interaction pair *(P, R)* and *(P, S)* were defined to be distinct if all interface residues in *(P, R)* were different

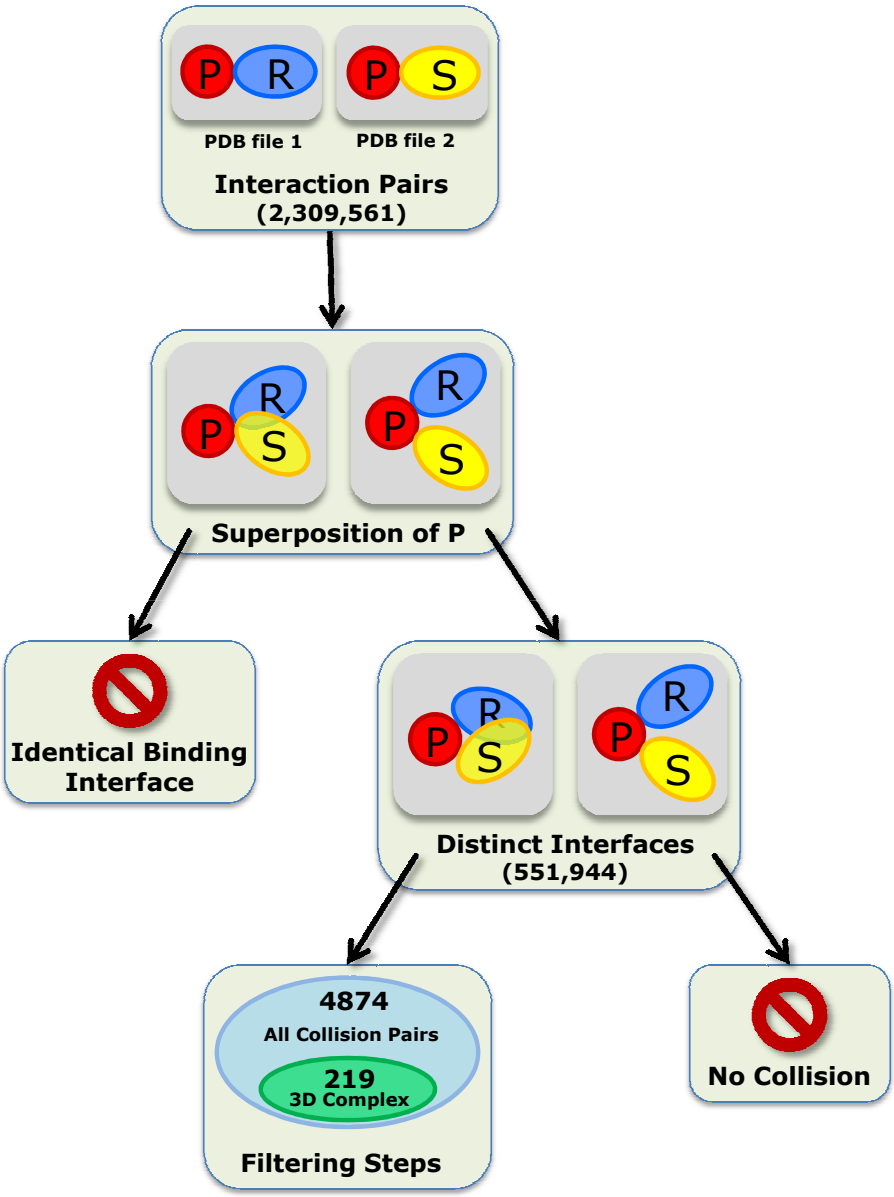


Figure 3.2: *Overview and results of our structure collision approach. The flow chart illustrates the necessary steps for identifying collisions of interacting proteins in three-dimensional space. Additionally, the number of interaction pairs remaining after each step is presented.*

from those in *(P, S)*, analogous to the previous study by Kim *et al.* (Kim et al. (2006)). If at least one interface residue was involved in both interactions, we regarded the interfaces as overlapping and the simultaneous interaction of the three proteins as impossible. To

ensure that the proteins can really establish a functional interaction, we considered only those interaction pairs *(P, R)* and *(P, S)* whose number of interface residues for each interaction was at least five residues.

After all pairs of interactions *(P, R)* and *(P, S)* that met the described criteria were identified, the primary proteins were superimposed and tested for collisions between the secondary proteins. Even if the UniProtKB accession numbers of two PDB chains are identical, the actual structure may not contain the complete protein because certain protein regions might not have been structurally determined. Therefore, the primary proteins P had to be aligned with each other to identify their corresponding PDB residues for computing the transformation matrix of the superposition. The alignments were performed using ClustalW (Thompson et al. (1994)), and the resultant files were parsed to extract the matching PDB residues.

3.2.3 Collision Detection Methods

The pairs of interactions together with the alignments for the superposition served as the input for the collision detection methods, where the execution of the collision detection was performed by Oliver Sander. To quantify the extent of the collision between the two secondary proteins, we computed the volume of the overlap of the secondary proteins after superimposing the primary proteins. C_α atoms of the corresponding residues in the primary proteins were superimposed by a rigid-body transformation (translation and rotation) to minimize the RMSD between corresponding C_α atoms. The rotation was determined by Kearsley's quaternion method (Kearsley (1989)), posing the minimization as an eigenvalue problem, which is solved by a singular value decomposition. After optimal rigid-body superimposition of the primary proteins, the overlap volume of the secondary proteins was computed as the difference between the sum of the individual volumes of the secondary proteins and the volume of the union of the secondary proteins. For the computation of the molecular volumes, we calculated the solvent excluded volume with MSMS by Sanner *et al.* (Sanner et al. (1996)). To confirm the results of this collision detection method, we alternatively computed the volume within the solvent accessible surface using ALPHAVOL (Liang et al. (1998)). Using these two complementary methods and measures, we filtered out 1,235 cases with numeric irregularities or instabilities and kept only those results in our dataset that were consistently identified by both collision detection methods. However, 1,067 of these 1,235 interaction pairs that we filtered out had a high RMSD value for the superimposed primary proteins (above 7 Å). Additionally, 174 of the 1,235 cases revealed a poor alignment quality of the primary proteins, consisting of

less than 30 residues. These factors may have had an influence on the observed numeric irregularities and instabilities, but due to the high RMSD values and low number of aligned residues we would have removed these interaction pairs from further analyses anyway.

3.2.4 Collision Constraints

We defined a collision to occur if both collision detection methods (MSMS and ALPHAVOL) consistently reported an overlap of the secondary proteins of at least 2,000 Å^3. Based on this definition, we identified 12,772 interaction pairs with colliding secondary proteins. As can be seen in *Figure 3.3*, the correlation of the overlap values produced by the two applied collision detection methods is 0.98, indicating a high reliability of the detected overlaps. The results were further refined and collisions were only retained if the RMSD of the superposition of the primary proteins was less than 7 Å to avoid false positives due to improper superposition. We also excluded results where the sequence lengths of the primary proteins differ by more than 15 residues in order to avoid large structural differences between the primary proteins. Additionally, we required the alignment of the two primary proteins to cover at least 30 amino acids in order to remove interaction pairs where the primary proteins corresponded to small fragments of a full-length protein.

These constraints reduced the number of colliding interaction pairs to 4,874 with an average RMSD of 1.23 Å and average overlap results of 2,659 Å^3 (MSMS) and 7,049 Å^3 (ALPHAVOL). The results were contained in 244 PDB structures, and 37 different primary proteins as well as 86 different secondary proteins participated in the interactions. These numbers show that many collisions of interaction pairs consisting of *(P, R)*, *(P, S)* involved the same proteins. However, the collisions were not evenly distributed among these primary and secondary proteins. Instead, we found 3,777 of the 4,874 collisions to occur in the interaction pair consisting of hemoglobin α-subunits as primary proteins, which were interacting with another α- and a β-subunit serving as the secondary proteins. The over-representation of hemoglobin likely results from a bias in available PDB protein structures towards certain well-studied protein complexes. We also observed that, in 98% of the 4,874 interaction pairs, both the primary and the secondary protein chains comprise single SCOP domains (Murzin et al. (1995)). Therefore, almost all collisions occur between single structural units of the participating proteins. One of the exceptions is illustrated in *Figure 3.5*, where the extracellular domain of the growth hormone receptor contains two SCOP domains and the collision involves both of them.

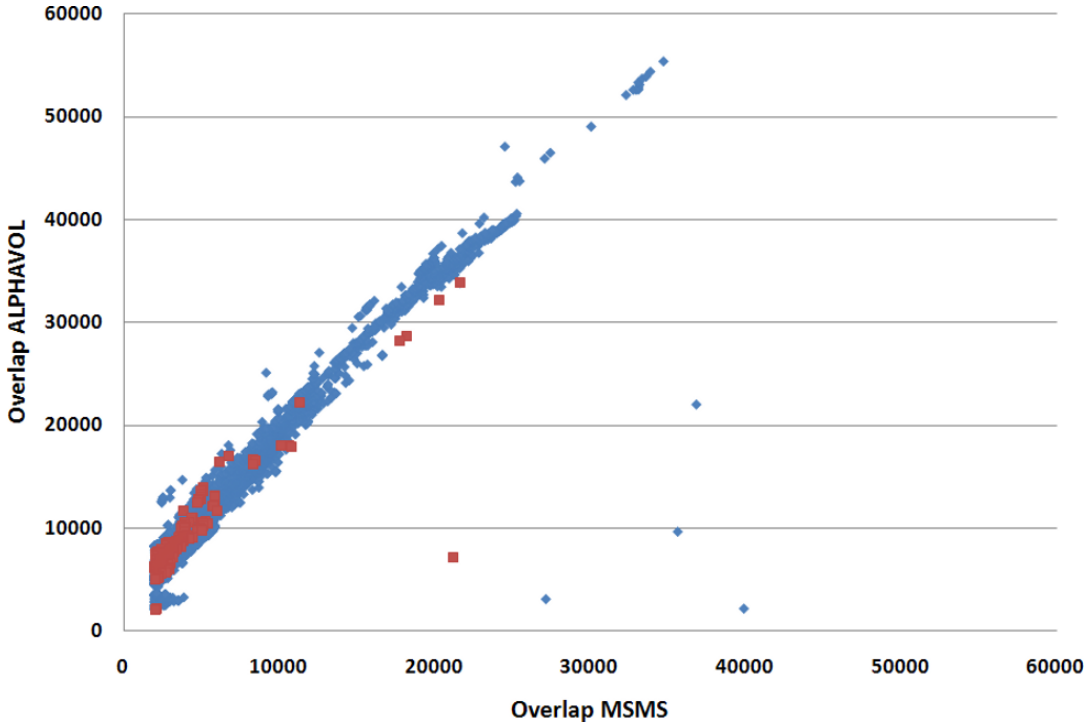


Figure 3.3: *Correlation of the results generated by the two collision detection methods MSMS and ALPHAVOL. All overlap values detected by the two methods are shown in blue. The correlation of all results (12,772 protein interaction pairs) is 0.98. The filtered results (4,874 protein interaction pairs, described in Chapter 3.3) are shown in red. For the filtered results, the correlation between the two methods slightly decreases to 0.90.*

3.3 Analysis of Binding Interfaces

Since protein interactions are often formed by domain-domain interactions, we studied the binding interfaces of the detected interaction pairs in more detail. We were particularly interested in the Pfam-A domains that form part of the two interaction interfaces of the primary proteins, because most of the currently available domain interaction databases contain Pfam-A domains. Our analysis revealed that, for most of our results (4,807 colliding interaction pairs, about 98%), the interface residues of the primary proteins could not be exclusively assigned to a single Pfam-A domain-coding region (Finn et al. (2010)). Instead, the interface residues belonged either to unstructured protein parts shared between one domain and additional unstructured parts of the primary proteins or shared between more than one domain and unstructured parts. This is particularly interesting since it

Pfam ID	Pfam Name	GO Terms	Instances
PF00142	Fer4_NifH	ATP binding; oxidoreductase activity	13
PF00160	Pro_isomerase	peptidyl-prolyl cis-trans isomerase activity; protein folding	10
PF05739	SNARE	-	6
PF02921	UCR_TM	ubiquinol-cytochrome-c reductase activity	4
PF02331	P35	caspase inhibitor activity; anti-apoptosis	3
PF02866	Ldh_1_C	oxidoreductase activity	2
PF00405	Transferrin	extracellular region; ferric iron binding; cellular iron ion homeostasis; iron ion transport	2
PF00607	Gag_p24	viral reproduction	1
PF00993	MHC_II_alpha	membrane; MHC class II protein complex; antigen processing and presentation; immune response	1

Table 3.1: *Pfam domains with unique interfaces. The table lists all Pfam domains participating in the 42 interaction pairs together with their GO annotations and the number of instances per domain.*

is assumed that domain-domain interactions often underlie protein-protein interactions and binding residues outside domain regions are not considered in domain interaction databases. A possible explanation for this observation is that most of the colliding interaction pairs identified in this study involve mutated proteins or non-natural protein interactions, which can have an impact on the computation of the interface residues. However, it is also possible that residues outside a domain region are involved in protein binding, for example, to further stabilize the interaction. In the collision results, we found only 42 interaction pairs consisting of *(P, R)*, *(P, S)* where the interface residues of both primary proteins P could exclusively be assigned to the same single domain-coding region. The latter regions included 9 different Pfam-A domain families occurring in up to 13 interaction pairs, of which 5 domain families participate in enzymatic activities (*Table 3.1*).

3.4 Selecting Biological Interactions

In order to avoid false positive results due to crystal packing effects, we used the database 3D Complex (Levy et al. (2006)) to identify protein interactions that are reported as truly interacting. We kept only those results in which both protein interactions *(P, R)* and *(P, S)* are contained in 3D Complex. This reduced the number of colliding interaction pairs to 219, corresponding to less than 5% of all collisions we identified before. Although

the number of collisions dramatically decreased, the remaining 219 interaction pairs were regarded as true positives and were analyzed in more detail. While the majority of these interaction pairs involved single SCOP domains, 5 of these remaining collisions included multi-domain secondary proteins containing two SCOP domains with the collisions always occurring in the binding domains. Most of the biological interaction pairs, i.e., 184, involved interactions between hemoglobin protein chains. Again, we found hemoglobin to be over-represented in our results as the result of a bias in currently available PDB structures. For the other colliding interaction pairs, the number of instances was below ten. A manual investigation of the 219 interaction pairs, however, revealed that all of the detected collisions occur as a consequence of non-natural structural conformations due to artificially constructed protein interactions.

Although we only found a very limited number of biological interaction pairs, this is mostly due to the previously mentioned fact that protein interactions may occur repeatedly in PDB files and that, while we identified 4,874 colliding interaction pairs, they only involved 37 and 86 primary and secondary proteins, respectively. Furthermore, the low number of biological interaction pairs indicates that many protein interactions contained in the PDB database may not occur in vivo and network-based studies incorporating structural information should consider this fact.

3.5 Examples of Structure Collisions

In the following, we present three examples of colliding protein interaction pairs (*Figure 3.4* to *Figure 3.6*). The selection and biological discussion of the collision examples has been performed in collaboration with Gabriele Mayr.

Figure 3.4 shows the superposition of Rac1 protein chains (primary proteins) that are in complex with an Arfaptin fragment or crystallized as a Rac1 trimer (secondary proteins). Regarding the superposition of the Rac1 protein chains, 177 residues were aligned and the RMSD of the superimposed primary proteins is 1.99 Å. The overlap between the secondary proteins is approx. 2,215 $Å^3$ according to MSMS and approx. 5,368 $Å^3$ according to ALPHAVOL. Rac1 is a hub protein that forms part of more than 70 complexes in the PDB and participates in well over 200 different pairwise protein interactions (see BioMyn database at http://www.biomyn.de) (Ramírez and Albrecht (2010)). Arfaptin functions as an effector of Rac1 (Tarricone et al. (2001)). One chain of the Rac1 trimer collides with the Arfaptin fragment. However, Rac1 trimerization was experimentally triggered by unnatural high levels of zinc that do not occur in living cells (Prehna and Stebbins (2007)). Therefore, this trimer complex is not expected to exist in vivo.

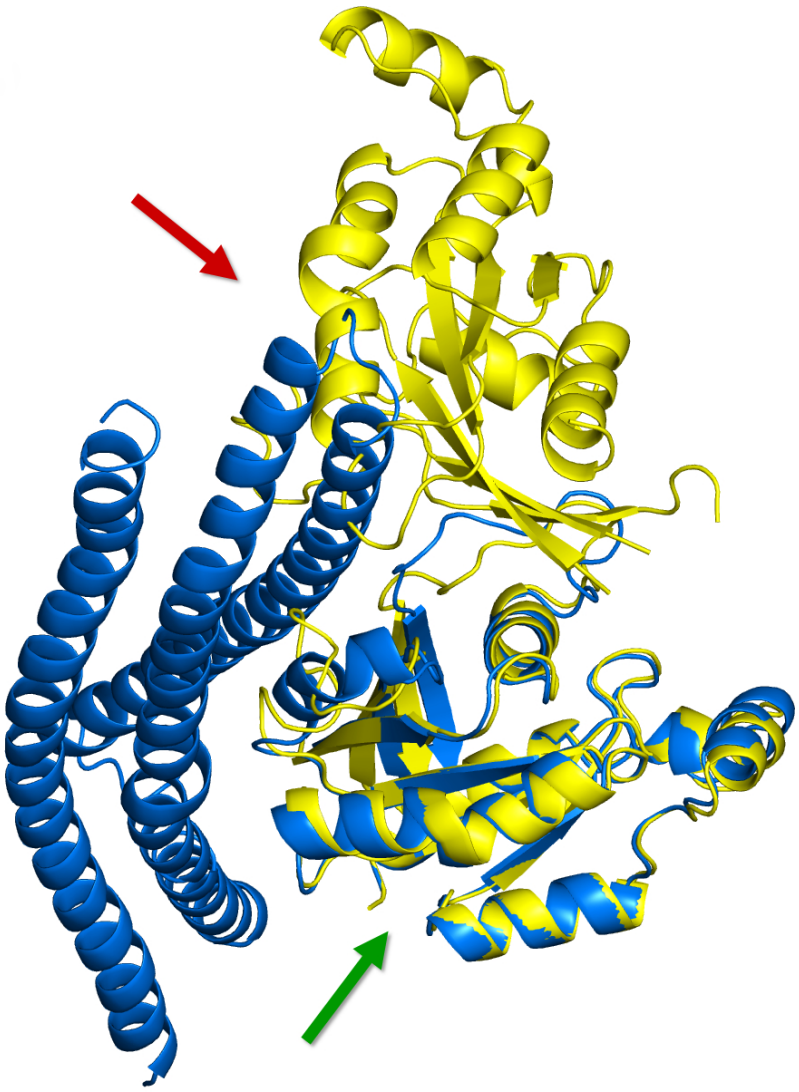


Figure 3.4: *Collision of the secondary proteins Arfaptin (PDB 1i4d chain A, blue) and Rac1-GDP (PDB 2p2l chain B, yellow), which both interact with the primary protein Rac1-GDP (PDB 1i4d chain D, blue, and 2p2l chain C, yellow). The primary proteins were superimposed (green arrow) and colliding regions are marked with a red arrow. Only colliding protein chains are shown.*

Figure 3.5 visualizes the superposition of the primary proteins cyclophilin A, which are in complex with a mutated HIV-1 capsid protein in one PDB structure and with a calcineurin B subunit in the other structure. 164 of the residues of the cyclophilin A chains could be aligned, resulting in a very precise superposition with an RMSD of 0.61

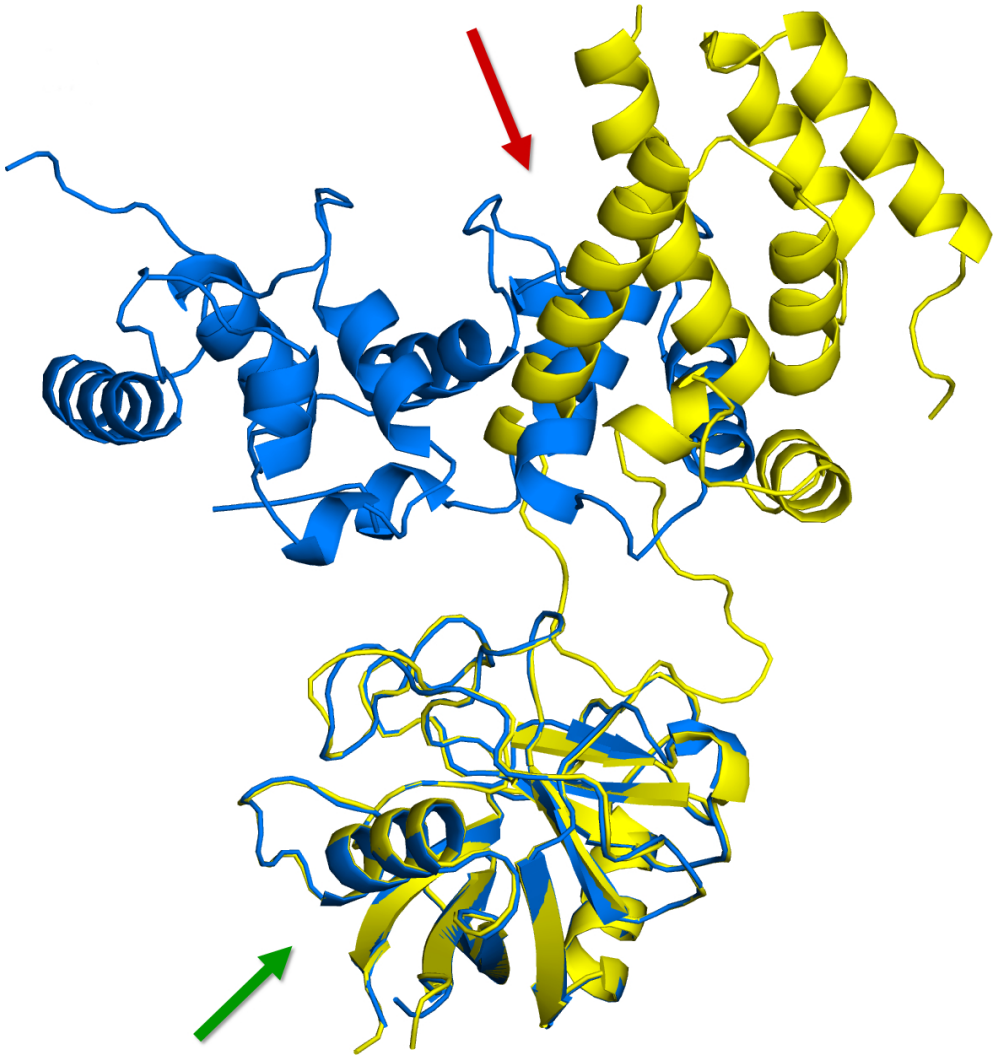


Figure 3.5: *Collision of the secondary proteins calcineurin B subunit isoform 1 (PDB 1mf8 chain B, blue) and HIV-1 capsid protein (PDB 1m9x chain D, yellow) when simultaneously interacting with the primary protein cyclophilin A (PDB 1mf8 chain C, blue, and 1m9x chain A, yellow). The primary proteins were superimposed (green arrow) and colliding regions are marked (red arrow). Only colliding protein chains are shown.*

Å. The detected collision is larger than in the previous example, with approx. 2,807 $Å^3$ reported by MSMS and approx. 5,995 $Å^3$ by ALPHAVOL. Cyclophilins are enzymes involved in diverse functions including protein folding, transport, and signaling (Howard et al. (2003)). They possess both sequence-specific binding and proline *cis-trans* isomerase

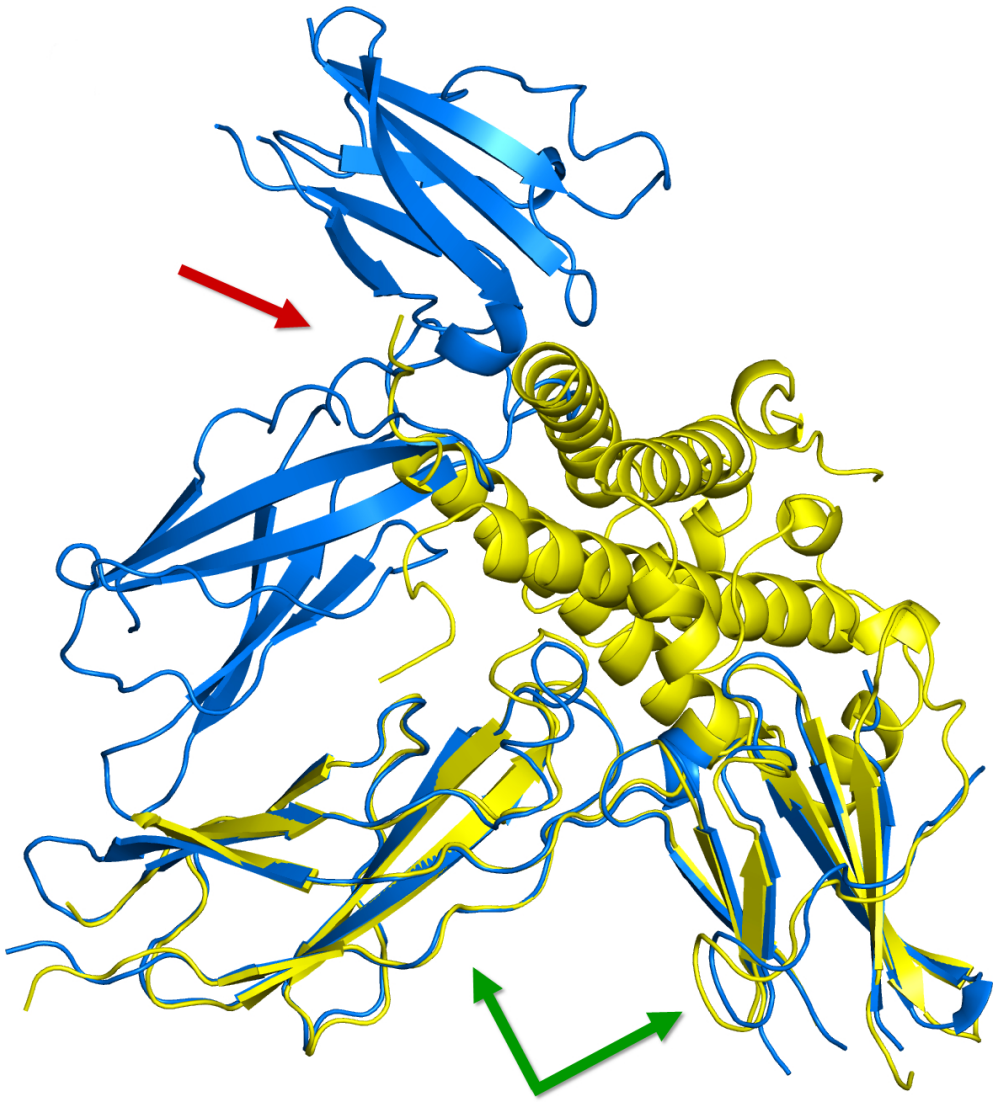


Figure 3.6: *Collision of the secondary proteins growth hormone receptor (PDB 1hwg chain B, blue) and growth hormone (PDB 1a22 chain A, yellow) when simultaneously interacting with the primary protein soluble growth hormone receptor (PDB 1hwg chain C, blue, 1a22 chain B, yellow). The primary proteins were superimposed (green arrow) and colliding regions are marked (red arrow). Only colliding protein chains are shown.*

activities. Cyclophilin A binds the HIV-1 capsid protein and facilitates virus replication. Calcineurin B participates in signaling for T-cell activation. The interaction between cyclophilin A and calcineurin B is part of a ternary complex with the immunosuppressive

drug cyclosporin A. The latter binds to cyclophilin A, enabling both the binding and the inhibition of calcineurin B and is thus an artificial construct (Jin and Harrison (2002)).

Figure 3.6 shows a collision between a growth hormone receptor (GHR) and a growth hormone (GH), which are both crystallized in interaction with the primary protein GHR. GHR was aligned with an RMSD of 1.65 Å ranging over 186 residues. A collision was detected between the second GHR from the dimer with the GH chain from the monomer, and MSMS reported approx. 2,330 $Å^3$ and ALPHAVOL approx. 6,194 $Å^3$. The active signaling complex has a stoichiometry of one GH molecule bound to two copies of its receptor (Sundstrom et al. (1996)). The detected collision originates from the artificial construct of a GHR monomer in complex with GH (PDB 1a22), which does not exist in vivo (Clackson et al. (1998)).

3.6 Conclusions

Our structure collision approach enabled the discovery of several cases of protein interaction pairs with colliding protein structures. We did not detect biologically relevant three-dimensional collisions of simultaneously possible protein interactions, but our analysis was limited by the low number of structurally determined protein complexes in the PDB. The identified collisions usually occurred between structures of experimentally modified proteins, which were crystallized in order to study alternative conformations or quaternary structures of the proteins. Nevertheless, our approach revealed several interesting occurrences of structural collisions.

Therefore, it is still important for future studies of protein interaction networks that separate binding interfaces might not imply simultaneously possible protein interactions. The functional implications of spatially colliding interaction partners can be manifold and similar to those of overlapping or identical binding sites such as the temporal control or inhibition of protein binding. In particular, structure collisions might constitute an essential mechanism of regulating transient protein interactions as occurring in signaling processes (Nooren and Thornton (2003a)). Here, collisions might involve adaptor and scaffold proteins and their interaction partners. These proteins frequently have a greater number of interaction partners than binding interfaces (Ramírez and Albrecht (2010)). Thus the combination of proteins that bind simultaneously to another protein at a specific time point or cellular location needs to be well-defined (Bhattacharyya et al. (2006)). Regulatory mechanisms different from the number of binding interfaces are needed for understanding the binding of specific combinations of proteins.

Finally, aside from the lack of structural data, there might be other reasons for not observing biologically relevant collisions in our study. For instance, PDB structures often consist of single protein domains as independently folded structural units rather than of the complete proteins. Therefore, different domains from a multi-domain protein can be found in multiple PDB structure files. Modeling structural linkers between the domains is still a very difficult task and cannot be performed at large scale yet. Consequently, we might have missed collisions between protein chains that bind the same protein in separate domains. Further issues are the existence of disordered regions and allosteric effects (Tsai et al. (2009); Goodey and Benkovic (2008)), i.e., the flexible nature of proteins, which might promote or prevent collisions. However, the required flexibility data on minor and major structural movements have not been available yet for such large-scale analyses as performed by us as well as other researchers. When more comprehensive structural datasets of protein complexes will have become available, further work might shed light on the presence and functional relevance of naturally occurring structure collisions.

Chapter 4

Comparison of Detection Methods for Tissue-Specific Gene Expression

Besides the three-dimensional structure of proteins, their co-occurrence in different tissues is necessary for the formation of protein interactions. Accurate knowledge of the tissue-specific gene and protein expression is important for understanding biological processes. As outlined in Chapter 2, there are different technologies available to measure gene expression. The technologies, however, are very diverse and it is commonly believed that they are not equally precise in their results. However, the quantitative differences of gene expression results have not yet been investigated and the impact on the outcome of functional analyses is still unknown. Therefore, we performed a quantitative analysis of gene expression data obtained from microarrays and high-throughput RNA-sequencing. Furthermore, we analyzed to what extent the different technologies influence the identification of tissue-specific proteins, interactions, and protein complexes. From the outcome of these analysis we identify the technology that is most suited for studying biological processes and functions. The results of this study have been presented and published at the International Workshop on Computational Systems Biology, WCSB 2010 (Emig et al. (2010a)).

4.1 Introduction

It is important for human systems biology and medicine to understand the tissue specificity of expressed genes and their products involved in cellular processes and diseases. Over the last years, many studies were based on the freely available Novartis Gene Atlas to investigate the tissue specificity of human gene expression and its biological impact on protein expression and protein interaction networks (Bossi and Lehner (2009); Lehner and Fraser (2004)). The Gene Atlas consists of comprehensive gene expression datasets

for a wide variety of tissues and cell lines obtained from a conventional Affymetrix 3' microarray, the HG-U133A, and a custom Affymetrix 3' microarray, the GNF1H (Su et al. (2004)). However, these data were already published in 2004, and the microarrays employed to obtain the data were 3' arrays of low probe density and specifically designed to measure genes that were annotated to the human genome at that time. This raises the question whether these relatively old datasets should still be regarded as a reliable source for studying the tissue-specific expression of human genes. A more recently released microarray is the Affymetrix Exon Tiling Array, which has been developed to measure exon expression rather than gene expression (Clark et al. (2007)). Its probe density per gene is much larger than that of the microarray technology used to generate the Gene Atlas. Furthermore, the advent of next-generation sequencing methods allows for further technological advances in the accuracy of transcriptome measurements (Ramskold et al. (2009)).

In the following, we explore three tissue-dependent gene expression datasets produced by microarray technologies and high-throughput RNA-sequencing. We first study the detection sensitivities of the technologies and compare the measured gene expression datasets. In addition, we investigate protein interactions to identify tissue-specific and universally occurring interactions. Last, we utilize the gene expression data for the identification and comparison of tissue-specific protein complexes and analyze to what extent functional implications of tissue specificity depend on the applied expression detection method.

4.2 Data Sources and Preprocessing

4.2.1 Gene and Protein Data

All analyses are based on the Ensembl database, version 52 (Hubbard et al. (2009)). We unified the gene and protein identifiers of all data sources by mapping them to Ensembl gene identifiers via Ensembl BioMart (Smedley et al. (2009)).

We obtained a human protein interaction network consisting of 80,922 interactions between 10,229 proteins from a recently published study (Bossi and Lehner (2009)). The protein interactions had been compiled from more than 20 publicly available data sources and were required to have experimental evidence of physical interaction. We mapped all proteins to Ensembl gene identifiers and kept a protein interaction only if both interacting partners could be mapped. Furthermore, we required the genes to have gene expression estimates assigned in all analyzed datasets. This reduced the original interaction network to 60,760 interactions between 8,413 proteins.

Human protein complexes were obtained from PDB and CORUM (downloaded July 2009) (Berman et al. (2000); Ruepp et al. (2010)). We mapped all complex members to Ensembl gene identifiers. We kept only those complexes for which all proteins could be mapped and had gene expression estimates available in all analyzed expression datasets. We also required the complexes to be composed of at least three different proteins and removed duplicates contained in the CORUM and PDB data. This resulted in 572 distinct protein complexes.

4.2.2 Expression Data

We downloaded the raw Novartis Gene Atlas data from NCBI's Gene Expression Omnibus (GEO; accession number GSE1133) together with probeset-to-gene annotations for the HG-U133A and the GNF1H microarrays. The data contain samples for 79 human tissues and cell lines, with two replicates for each tissue and cell line. For the Affymetrix Exon Array, we downloaded sample data for 11 tissues as provided by Affymetrix, with three biological replicates for each tissue. RNA-sequencing data for 15 tissues and cell lines was contained in the supplementary data of a study by Wang and colleagues (Wang et al. (2008)). Five human tissue samples were contained in all three expression datasets and were used for the following analyses: heart, liver, testis, skeletal muscle, and cerebellum.

4.2.3 Probeset to Gene Mapping

We mapped the probesets for the arrays to Ensembl genes using all identifiers that were annotated in the probeset-to-gene mappings. For the Affymetrix HG-U133A array, we were able to map 21,778 probesets to 12,489 Ensembl genes, out of which 12,448 encode proteins. For the GNF1H array, we were able to map 8,875 probesets to 6,086 Ensembl genes, out of which 5,943 encode proteins. The Gene Atlas data is based on the combination of both microarrays and consists of a total of 16,989 distinct protein-coding genes.

For the Exon Array, we mapped the probesets to Ensembl genes according to the genomic coordinates of the probesets as given by the respective NetAffx annotations (Cheng et al. (2004)). Altogether, the probesets could be mapped to 20,444 protein-coding genes.

4.2.4 Gene Detection Calls

The raw Novartis Gene Atlas data were normalized using the Affymetrix Expression Console software. For the normalization of the samples we applied the MAS5.0 algorithm with default parameters, as recommended by Affymetrix (see Chapter 2.2.1 for details of

the MAS5.0 algorithm). The resulting detection calls for the probesets, which are automatically derived by MAS5.0, were then used to identify gene expression in the respective tissue samples. While the presence call of a probeset suggests that the gene is expressed, an absence call indicates that the gene could not be reliably detected. In case of a marginal calls, however, the probeset signal is somewhere in between very reliable (presence call) and not reliable (absence call) and the statistical software itself cannot reliably predict whether the corresponding gene is expressed or not. In the following, we treated marginal calls as an indicator for gene expression. We regarded a probeset as being present in a tissue if it was present in at least one of the two replicates. We would actually expect that a probeset is either present in both replicates or in none. However, there are numerous factors that can influence the outcome of a microarray experiment, for instance, slight differences in the experimental procedure. Since the absence call of a probeset is simply an indicator for the unreliability of the probeset and not for the absence of the gene expression, we assumed that one reliable detection per tissue was acceptable for this study. If more than one probeset mapped to a gene, we required at least one of these probesets to be present for gene expression. Present and absent probesets that all map to a single gene may, for example, occur as a result of alternative splicing. Since 3' microarrays do not take alternatively spliced transcripts into account, it is possible that the predominant transcript isoform in a tissue does not match all probesets and therefore, we did not require a presence call for all probesets.

The raw Exon Array data were processed using AltAnalyze with default parameters as recommended by the authors (see Chapter 2.2.2). AltAnalyze computed a detection p-value for every Ensembl gene in each of the three replicates per tissue. The p-values were derived using the detection above background (DABG) method, which is integrated in AltAnalyze and which is the standard procedure for computing presence and absence calls for Exon Arrays. We obtained gene presence and absence calls by taking the median of the three p-values for every gene in each sample and set the presence p-value threshold to 0.05, which is the recommended procedure for DABG p-values (Clark et al. (2007)).

Gene expression estimates (RPKM values) for the RNA-sequencing data were obtained from the study by Wang and colleagues (Wang et al. (2008)). We chose a very conservative expression threshold and treated all genes having an RPKM value ≥ 1 as present and all others as absent (Ramskold et al. (2009)). In contrast to the other tissues with a single sample each, six different samples were available for cerebellum. To obtain a single RPKM value per gene in cerebellum, we took the mean of these expression estimates and regarded genes as expressed if their mean RPKM values were ≥ 1.

4.2.5 Comparison of Detection Calls

Although the three datasets contain many tissue and cell line samples, the overlap is rather small consisting of five tissues only. We therefore defined a tissue-specific gene to be expressed in exactly one of these five tissues without any tolerance interval.

The gene presence and absence calls amount to a binary classification of gene expression that does not take expression levels into account. Therefore, we used the Matthews correlation coefficient (MCC) to compute pairwise correlations between the datasets according to their binary gene expression results (Matthews (1975)). The MCC is based on the true/false positives and negatives in the pairwise comparison of the datasets and can be computed as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

In this study, TP is the number of true positives, i.e. genes classified as expressed in both datasets. TN is the number of true negatives, i.e. genes that are not expressed according to both datasets. FP is the number of false positives, FN the number of false negatives, corresponding to the number of genes detected as expressed in the one dataset but not expressed in the other, respectively.

4.3 Results of the Gene Expression Analysis

We first extracted all protein-coding genes that were contained in all three expression datasets in order to compare their presence and absence calls (i.e. expression detected or not). This resulted in a total of 14,718 Ensembl genes. We find that RNA-sequencing results and Exon Array data have a comparatively high agreement in their presence and absence calls, while the Novartis Gene Atlas results show inverse calls for many genes (*Figure 4.1*). More precisely, the correlation between the RNA-sequencing results and the Exon Array data is clearly higher than the correlation of either of these datasets to the Gene Atlas data (based on the MCC). On average, the correlation between RNA-sequencing results and Exon Array data is 0.56, with a maximum of 0.61 in liver and a minimum of 0.44 in testis. The average correlation between the Gene Atlas and RNA-sequencing data is 0.27 and between the Gene Atlas and Exon Array data 0.28. The respective maximal correlations are found in liver (0.31) and in testis (0.32), while the minimal correlations, 0.18 and 0.20, are both detected in muscle.

RNA-sequencing appears to be the most sensitive method for detecting gene expression. *Figure 4.2* shows that, for each tissue except cerebellum, the number of expressed genes

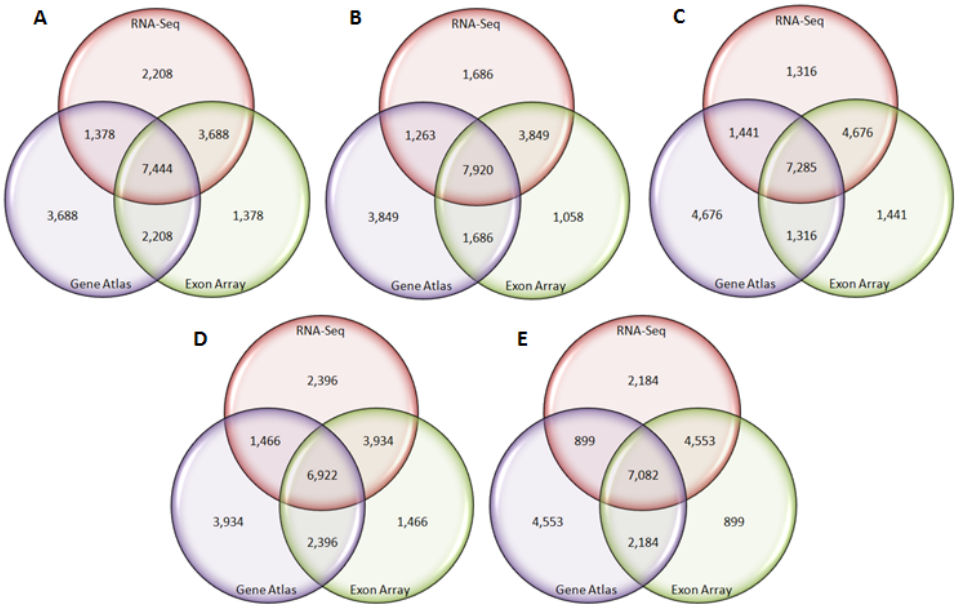


Figure 4.1: *Gene presence and absence calls in the datasets according to the three technologies. Only genes contained in all datasets were taken into account (14,718 genes total). Venn diagrams are shown for:* ***(A)*** *heart,* ***(B)*** *liver,* ***(C)*** *cerebellum,* ***(D)*** *testis, and* ***(E)*** *skeletal muscle.*

is highest when using RNA-sequencing data, a finding that is in agreement with a recent study by Ramskold *et al.* (Ramskold et al. (2009)). As could be seen from the correlation of the presence and absence calls above, the Gene Atlas arrays are not able to detect many of the genes found expressed according to the Exon Array and RNA-sequencing technologies. The number of tissue-specific genes however, i.e. the genes expressed in exactly one of the five tissues, is low for all methods. The fewest tissue-specific genes are detected in skeletal muscle and the highest number is found in testis.

We also compared the genes that were found to be expressed according to the different methods. We observed a high agreement of genes with presence calls for RNA-sequencing and Exon Arrays, with the lowest agreement (37%) in skeletal muscle and the highest (56%) in cerebellum. The Gene Atlas, however, is not able to detect many of these genes and, on average, shows a low agreement with the other datasets.

A closer look at the tissue specificity of expressed genes reveals that the gene expression detection results vary significantly between the datasets and across tissues (*Figure 4.3*). While RNA-sequencing detects more than 6,000 genes (41% of all shared genes) to be expressed in all tissues, the Exon Array identifies only about 4,500 genes (31%) and the

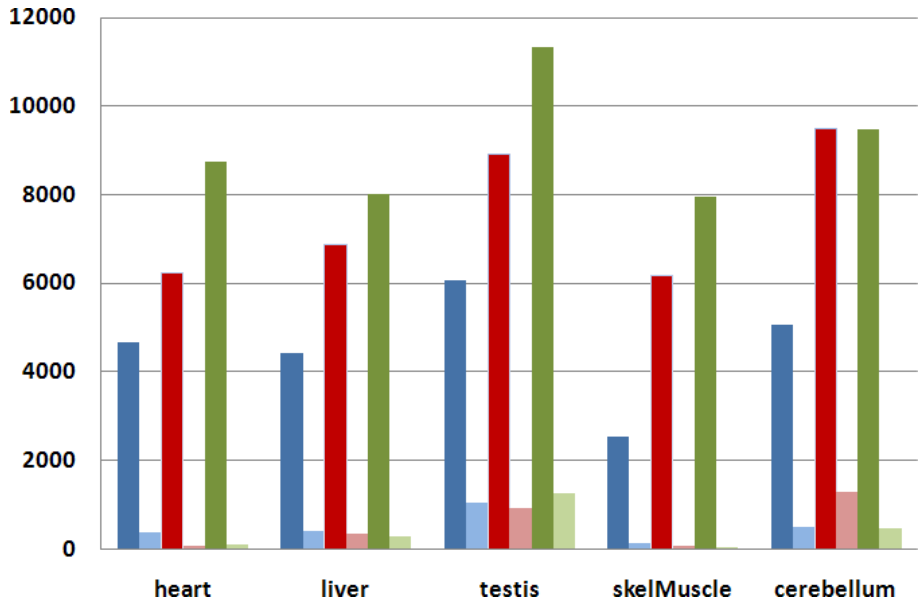


Figure 4.2: *Histogram with the frequencies of all expressed genes and the tissue-specific fraction (only expressed in the respective tissue) in each tissue as detected by the Gene Atlas (2 left bars, blue), Exon Array (2 middle bars, red), and RNA-sequencing (2 right bars, green). The first bars always show the total number of expressed genes and the second ones the number of tissue-specific genes.*

Gene Atlas finds only about 1,500 genes (10%) to be expressed in all tissues. For genes expressed in at most four tissues, the numbers are very similar for all datasets. The reverse can be observed for those genes not expressed in any of the five tissues: RNA-sequencing identifies the lowest number of absent genes (about 2,100), while the Gene Atlas is not able to detect expression for more than 6,000 genes.

These results demonstrate clearly that fewer genes are tissue-specific than previously thought and that tissue expression studies will need to be re-examined using the novel RNA-sequencing technology. Obviously, microarrays are less sensitive with respect to detecting gene expression than are RNA-sequencing techniques. Statistical methods used for normalizing microarray data rarely can distinguish between very low gene expression and experimental noise. Therefore, it is likely that low expression is mistakenly reported as noise and thus the respective gene is regarded as not expressed. Furthermore, technical problems such as cross-hybridization events can lead to biased gene expression results. RNA-sequencing methods however, which are simply based on read-to-gene mappings, can more accurately detect genes even at very low expression levels.

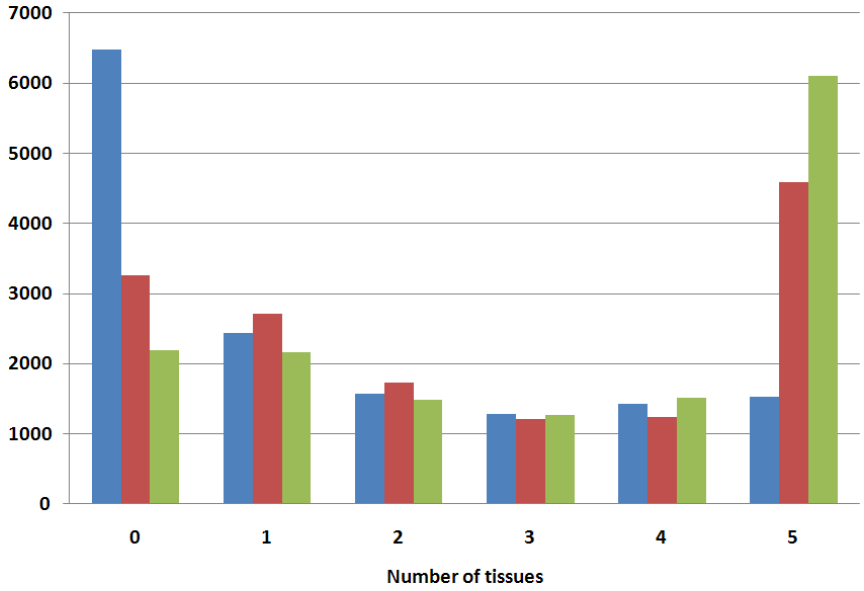


Figure 4.3: *Histogram with the frequencies of expressed genes in the respective number of tissues. Only genes contained in all datasets are taken into account. Both RNA-sequencing (green) and the Exon Array (red) show the highest number of expressed genes in all 5 tissues while the Gene Atlas (red) identifies a comparatively low number of genes to be expressed in all genes. All datasets show similar expression rates for 1 to 4 tissues. The number of genes absent in all tissues is by far the highest using the Gene Atlas, while RNA-sequencing and the Exon Array report much fewer absent genes.*

We also compared the detection sensitivity of RNA-sequencing and the microarrays based on the RPKM values reported in the RNA-sequencing results. For each tissue, we first extracted all of the 14,718 genes contained in all three datasets that have an RPKM value of at least 1 in the RNA-sequencing results. We found that RNA-sequencing detects a high number of genes expressed at low levels (with a log2-transformed RPKM value of below 2). Next, we investigated the fraction of these genes that are also detected as expressed by the microarray methods and annotated the respective RPKM values to them. We observed that, for all tissue samples, the Exon Array identifies a greater number of genes expressed at low levels (with a low RPKM value according to the RNA-sequencing data) than the Gene Atlas (*Figure 4.4*). This suggests that the Exon Array is better suited to distinguish between low gene expression and noise than the Gene Atlas arrays, which is likely due to the high probe density of the Exon Array.

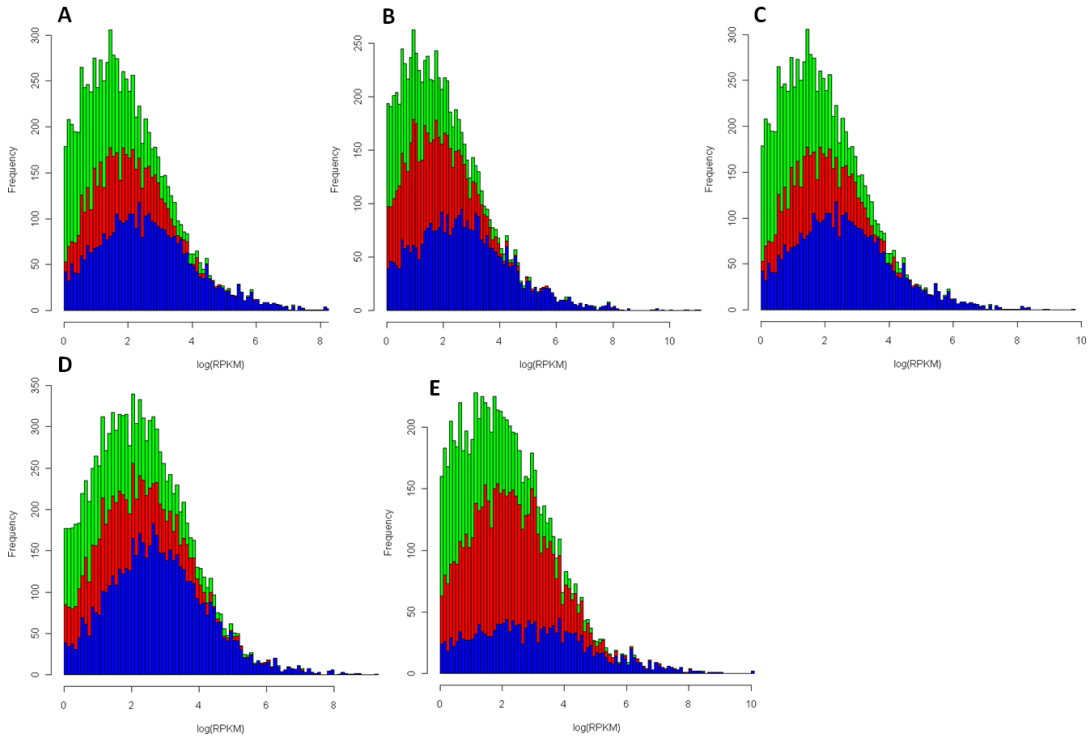


Figure 4.4: *Microarray gene presence related to RNA-sequencing RPKM gene expression estimates. Histograms are plotted for expressed genes according to the RNA-sequencing values. The RPKM distribution for present genes according to the RNA-sequencing results is shown in green, for the Exon Array in red and for the Gene Atlas in blue.* ***(A)*** *represents heart,* ***(B)*** *liver,* ***(C)*** *cerebellum,* ***(D)*** *testis, and* ***(E)*** *skeletal muscle.*

4.4 Results of the Protein Interaction Analysis

Gene expression of protein-coding genes usually leads to the production of proteins in the cells. Therefore, proteins and their interactions can only occur in a certain tissue if the corresponding genes are expressed. Functional analyses of proteins and interactions are thus highly dependent on accurate gene expression estimates. We re-examined a recent study by Bossi and Lehner regarding the tissue specificity of physical protein interactions, which was based on the Novartis Gene Atlas data (Bossi and Lehner (2009); Su et al. (2004)). In this study, a high number of tissue-specific protein interactions was reported, which mainly occurred due to the interaction of a tissue-specific protein with a universally expressed protein. Using the Gene Atlas data, we are able to reproduce these findings. However, *Figure 4.5* shows that the number of protein interactions occurring in the tissues

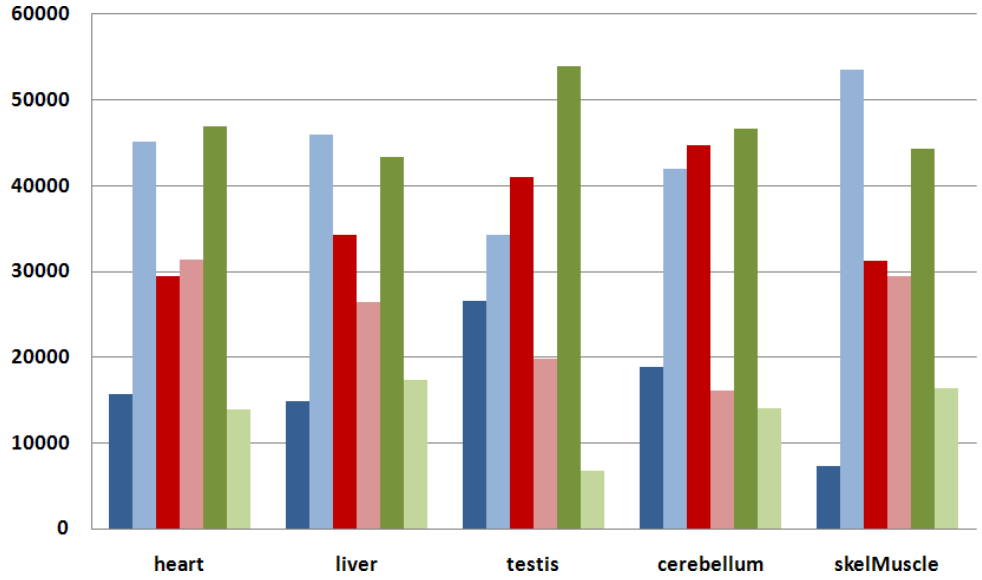


Figure 4.5: *Histogram of the numbers of present and absent protein interactions in each tissue. For each tissue, the two leftmost bars show presence and absence according to the Gene Atlas (blue), the third and fourth bars according to the Exon Array (red), and the two rightmost bars according to RNA-sequencing (green).*

rapidly grows when applying the Exon Array or RNA-sequencing data. It is especially noteworthy, that for both the Exon Array data and the RNA-sequencing results, we usually find many more present than absent protein interactions in the tissues. Comparing the number of absent protein interactions to present ones based on the Gene Atlas data, we always find higher numbers of absent protein interactions. This finding suggests that fewer protein interactions are tissue-specific than assumed previously, and, according to the Exon Array and RNA-sequencing results, relatively few protein interactions are involved in tissue specificity.

4.5 Results of the Protein Complex Analysis

Additionally, we investigated to what extent microarrays and RNA-sequencing are able to detect the expression of protein complexes in different tissues. We distinguished between completely expressed complexes (all involved genes are co-expressed), partially expressed complexes (at least one of the involved genes is not expressed, but we require the partial complex to consist of at least two expressed proteins), and completely absent complexes (at most one involved gene is expressed). As shown in *Figure 4.6*, RNA-sequencing appears

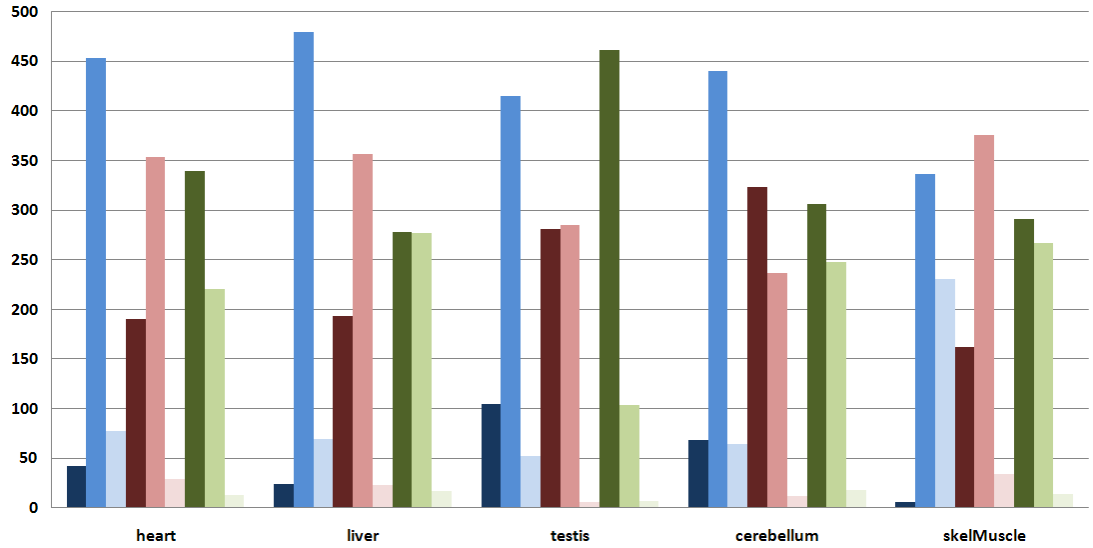


Figure 4.6: *Expression of protein complexes in each tissue according to the Gene Atlas (three leftmost bars), Exon Array (three middle bars), and RNA-sequencing (three rightmost bars). The respective three bars are ordered according to complete expression, partial expression, and complete absence of the protein complexes.*

to be the most sensitive method, and the highest number of completely expressed protein complexes is found in all tissues using this technology. In contrast, the Exon Array identifies fewer completely expressed complexes, and the Gene Atlas hardly detects any complexes as completely expressed. Since the detection sensitivity of the Gene Atlas has been shown to be the lowest, we expected to find few completely expressed complexes according to these data. However, the detection rate for protein complexes is even lower than we expected, with only 0.01% in skeletal muscle (compared to $\approx$ 51% using RNA-sequencing). Conversely, it is interesting that the number of completely absent complexes is low for all methods, suggesting that most of them contain highly expressed gene products detectable by all methods.

To compare the expression measurements of the microarrays and RNA-sequencing, we computed their correlation regarding the detection of protein complexes. For each complex, we calculated the percentages that were found to be expressed according to the different measurement methods. A percentage of 0% indicated that the complex is completely absent, while a percentage of 100% suggested that the complex is completely present. Everything in between 0% and 100% was defined to be a partially expressed complex. As for the previously computed correlation of the gene expression results, the correlation regarding the expression of protein complexes is clearly higher for RNA-sequencing and the

Exon Array than for any of these technologies correlated to the Gene Atlas. On average, the correlation between RNA-sequencing and the Exon Array is 0.66, with a maximum of 0.73 in muscle and a minimum of 0.48 in testis. For the Gene Atlas and RNA-sequencing as well as for the Gene Atlas and the Exon Array, the average correlation is 0.31 in both cases, with a minimum of 0.23 in muscle and a maximum of 0.39 in cerebellum, and a minimum of 0.26 in cerebellum and a maximum of 0.36 in testis, respectively.

4.6 Conclusions

Our quantitative analysis of gene expression detection technologies confirms the widespread belief that the results are highly dependent on the applied technologies and that the new RNA-sequencing technology outperform microarrays. We found that, using RNA-sequencing technologies, a considerably larger number of genes is widely expressed than previously thought and that many of the detected genes are expressed at low levels. Using the widely-used, yet low-density, 3' microarrays we were not able to detect many of these genes. However, it is remarkable that the Exon Array results correlate well with the RNA-sequencing results, which suggests that the high probe density of this microarray is partially able to identify low gene expression.

We additionally integrated the gene expression results obtained from the different technologies with protein interactions and protein complexes to discover to what extent the differences in gene expression detection have an impact on the outcome of functional analyses. We found that, when using the 3' microarrays, the number of protein interactions and complexes detected in each tissue is very low and many interactions and complexes are classified as being highly tissue-specific. Using RNA-sequencing results, however, we found a considerably larger number of expressed protein interactions and complexes per tissue and we also classified much fewer as tissue-specific. These results indicate that previous functional analyses that are based on 3' microarrays need to be re-considered. While these studies revealed that a large number of proteins and interactions is tissue-specific, the results are likely to be biased towards highly expressed genes and thus cannot provide precise insights into tissue-specific elements and functions. In the following chapter, we therefore investigate tissue-specific functions and processes based on gene expression results obtained from high-throughput RNA-sequencing to accurately identify tissue-specific elements to better understand their biological meaning.

Chapter 5

Functional Implications of Tissue-Specific Gene Expression

As described in the previous chapter, high-throughput RNA-sequencing methods can accurately measure gene expression even at low levels. Therefore, analyses based on these data can provide more precise information on the context-dependency of biological processes than those based on microarray data. This chapter describes a study, in which we integrate human gene expression data based on RNA-sequencing with protein interactions, protein domains, and protein complexes to identify tissue-specific characteristics and functions. This study has been the focus of a manuscript, which we published recently (Emig and Albrecht (2011)).

5.1 Introduction

Proteins are involved in almost all biological processes, and their interactions are essential for the survival of cells. Tissue-specific gene expression can result in the presence or absence of certain protein interactions and complexes, leading to profound functional differences of biological processes between the tissues. Accurate knowledge of tissue-specific genes and proteins is of great importance for understanding the biological functions and determining biomarkers and drug targets (Vasmatzis et al. (2007)). Therefore, we examined the functional implications of tissue-specific gene and protein expression based on data generated by RNA-sequencing for 15 human tissues and cell lines.

In the last years, many research projects were performed to identify universally expressed and tissue-specific genes and their products. Lehner and Fraser discovered tissue-specific mammalian protein domains and compared their cellular functions with those of universally expressed protein domains (Lehner and Fraser (2004)). Another work by Bossi and Lehner focused on the tissue-specific occurrence of human protein interactions (Bossi

and Lehner (2009)). Based on gene expression data from microarray experiments, they showed that many protein interactions are tissue-specific. They also found that universally expressed proteins frequently interact with tissue-specific ones.

Other studies concentrated on the identification and analysis of disease-related genes and protein complexes by examining the tissue specificity of known disease genes and their features such as evolutionary conservation and functional annotation (Tu et al. (2006); Lage et al. (2008)). Eisenberg and Levanon analyzed the genomic structure of universally expressed genes, also called housekeeping genes. They revealed that it is more compact than the structure of tissue-specific genes, i.e., the number of exons, introns and untranslated regions is lower and the regions are shorter in length (Eisenberg and Levanon (2003)). The results of follow-up research by She and colleagues suggested that CpG islands are enriched in transcription start sites of universally expressed genes (She et al. (2009)). In addition, Farré and colleagues observed that the sequence conservation of promoters of universally expressed genes is significantly lower than of tissue-specific ones (Farre et al. (2007)).

Most of the above-mentioned analyses on tissue-specific gene expression and structural and functional properties of genes and their products were based on microarray experiments. One of the most extensive experiments resulted in the Novartis Gene Atlas (Su et al. (2004)), which provides gene expression patterns for 79 human tissues and cell lines. However, statistical methods for analyzing microarray data are limited in their ability to distinguish between low gene expression and experimental noise. Thus, expression profiles are error-prone especially for genes expressed at low levels, giving rise to the misclassification of genes regarding their tissue specificity. Furthermore, technical issues such as cross-hybridization can further bias the results. This was confirmed in a recent study by Zhu and colleagues who compared various gene expression datasets. They concluded that our current knowledge on universally expressed genes is quite deficient and that their number is considerably under-estimated (Zhu et al. (2008)). Recently, these findings were further supported by the first gene expression analysis based on next-generation RNA-sequencing data (Ramskold et al. (2009)). Here, the authors discovered that the number of universally expressed genes is much higher than estimated in previous microarray-based studies.

In the following, we present our analysis of the tissue occurrence of protein interactions, domains, and complexes as well as of transcript isoforms. We identify tissue-specific elements and biological functions and investigate to which extent former findings based on microarrays are still in agreement with new results based on RNA-sequencing.

5.2 Materials and Methods

5.2.1 Gene Expression Estimates

We obtained gene expression estimates (RPKM values) for 10 human tissues and 5 human cell lines from the publicly available supplementary data of an alternative splicing study by Wang and colleagues (Wang et al. (2008)). As described in Chapter 2.2.3, the RPKM value is a measure of the number of RNA-sequencing reads mapped to the constitutive exons of a certain gene and reflects whether a gene is transcribed in the tissue or not (Mortazavi et al. (2008)). For each of the tissues and cell lines, the data contained about 20 million reads, of which about 60% could be uniquely mapped to the reference genome. We defined a gene to be expressed in a certain tissue or cell line if the respective RPKM value was at least 0.3. This is a reasonable gene expression threshold as has been shown by Ramskold *et al.*, who examined different RPKM gene expression thresholds in an extensive gene expression study using these RNA-sequencing data (Ramskold et al. (2009)).

5.2.2 Proteins, Domains, and Complexes

The protein interaction network was obtained from a study by Bossi and Lehner (Bossi and Lehner (2009)). We mapped the gene expression estimates onto the dataset of 80,923 physical protein-protein interactions and retained only those interactions where we had expression values for both interacting proteins. This reduced the dataset to a protein interaction network of 63,815 interactions involving 8,805 human proteins. Drugs and drug targets were taken from the supplementary data provided in the study by Yildirim *et al.* (Yildirim et al. (2007)).

We obtained the Pfam domain annotations from Ensembl with 17,289 genes encoding at least one domain. For 3,908 of these genes, we did not have expression data and thus excluded them from further analyses. This provided 13,381 genes encoding a total of 3,330 different Pfam domains. For each of the remaining Pfam domains, we also computed the percentage of genes encoding the respective Pfam domain that had expression values annotated. To increase the reliability of the analysis, we excluded domains that had expression values for fewer than 75% of the encoding genes annotated. This reduced the number of remaining Pfam domains to 2,840.

Protein complexes were obtained from CORUM and PDB structures, downloaded in August 2009 (Berman et al. (2000); Ruepp et al. (2010)). CORUM usually does not provide information on the stoichiometry of the complexes and only reports the names of the co-complexed proteins. Therefore, we disregarded the stoichiometry of the complexes

for both CORUM and PDB and required at least 3 different proteins to be contained in a complex. We excluded binary complexes because they represent the physical interaction of two proteins, which we regard as a protein-protein interaction as in the protein interaction network described above.

The complexes were given by UniProt identifiers and mapped to Ensembl gene identifiers using the available Ensembl annotations in BioMart (Smedley et al. (2009)). We retained only complexes if all co-complexed proteins could be mapped to Ensembl gene identifiers. Furthermore, we required the presence of expression values for all proteins in a complex, which yielded 648 complexes.

5.2.3 Tissue-Specific Gene Expression

We used two alternative definitions of universal and tissue-specific gene expression. The first definition is based on the mRNA presence and absence in the tissues and cell lines (P/A definition, PAD). We defined a gene to be universal if expression was detected in at least 14 of the 15 tissues and cell lines, accounting for potentially inaccurate read-to-genome mappings due to noisy and incomplete data. Accordingly, tissue-specific genes are expressed in at most two tissues and cell lines.

The second definition combines mRNA presence and absence with tissue-specific over-expression (Peak definition, PKD). As for PAD, we regarded a gene as tissue-specific if gene expression was detected in at most two tissues and cell lines. For genes expressed in at least three tissues and cell lines, we applied an over-expression analysis adapted from a study by Winter and colleagues (Winter et al. (2004)) and computed the multinomial distribution of the expression values. For each gene, we checked whether the expression levels were equal in the 15 tissues and cell lines or whether the gene was over-expressed in one or two of them. To this end, we computed the tissue specificity value $TS(g)_t$ of a gene g expressed in tissue t using the following formula:

$$TS(g)_t = \frac{RPKM_t}{\sum_{\hat{t} \in tissues} RPKM_{\hat{t}}}$$

$TS(g)_t$ describes the contribution of the gene expression level ($RPKM_t$) of gene g in tissue t to the sum of the gene expression levels of g in all 15 tissues and cell lines. As defined by Winter *et al.*, we regarded a gene to be tissue-specific if its maximum tissue specificity value $max(TS(g)_t)$ was above 0.4 and the corresponding $RPKM_t$ value was above the mean expression level in the specific tissue or cell line. The second requirement was necessary to ensure that genes expressed at low levels in most tissues and cell lines and at a slightly higher level in another tissue would not be identified as outliers since

moderate differences in mRNA levels are negligible. Genes expressed in at least 14 tissues and cell lines with $max(TS(g)_t)$ below 0.4 (i.e., without a clear over-expression in at least one tissue or cell line) or with an $RPKM_t$ expression value below the mean were defined as universal.

5.2.4 Comparing Definitions for Tissue Specificity

In the following, we use both alternative definitions of tissue-specific gene expression, PAD and PKD, and compare potential differences in the biological results. However, concerning PKD based on gene over-expression, it is to note that gene expression levels do not necessarily reflect protein abundances (Maier et al. (2009); Vogel et al. (2010)). Furthermore, the alternative definitions of tissue-specific gene expression naturally lead to different classifications of genes. Genes identified as tissue-specific according to PAD are a subset of those detected with PKD. Similarly, genes classified as universally expressed with PKD are a subset of those found using PAD. This means that genes above the RPKM threshold but with varying expression levels in all tissues and cell lines will be classified as universally expressed by PAD, but they might be classified as tissue-specific based on PKD.

5.2.5 Tissue Specificity of Proteins, Interactions, and Domains

We inferred the tissue specificity of a protein from the tissue specificity of the corresponding protein-encoding gene. We defined proteins to be universal or universally expressed if the encoding gene was expressed in at least 14 of the 15 tissues and cell lines. Tissue-specific proteins are defined to be expressed in 0 to 2 tissues and cell lines.

Based on the tissue specificity of proteins, we defined a protein interaction to be universal (or universally expressed) if the interacting proteins were co-expressed in at least 14 of the 15 tissues and cell lines. An interaction of two universal proteins does not lead to a universal protein interaction according to our definition if the universal proteins are not co-expressed in the required number of tissues and cell lines. We defined a tissue-specific interaction to occur in at most two tissues and cell lines. The interacting proteins do not need to be tissue-specific themselves, but they have to co-occur in at most two tissues and cell lines.

Protein domain expression was determined by the genes encoding the respective Pfam domain. For each domain, we averaged the number of tissues in which the genes encoding the domain are expressed. We defined a domain to be universal if the average number of tissues and cell lines was greater than 13, and to be tissue-specific if it was less than 3.

5.2.6 Tissue Specificity of Protein Complexes

In case of protein complexes, we combined the tissue specificity with the completeness of a complex. We first classified a protein complex according to its completeness in a specific tissue or cell line, i.e. the fraction of expressed co-complexed proteins. If all co-complexed proteins were expressed, the complex was regarded as *fully expressed* in this tissue or cell line. If at least two of the co-complexed proteins, but not all of them, were expressed, we defined the complex to be *partially expressed* in this tissue or cell line. Protein complexes with less than two expressed proteins were called *absent* in this tissue or cell line.

From the completeness of a complex in each of the 15 tissues and cell lines, we inferred the tissue specificity by counting the number of tissues and cell lines, in which the complex was fully expressed. We defined a universal protein complex to be fully expressed in at least 14 tissues and cell lines, and tissue-specific complexes to be fully expressed in at most two tissues and cell lines.

5.2.7 Quantifying Tissue Similarities

We performed pairwise comparisons of the tissues and cell lines based on their gene expression profiles. Let t_1 and t_2 be two tissues, and let TSG_1 and TSG_2 be the sets of tissue-specific genes expressed in the respective tissues and cell lines. Then, the pairwise similarity $simG$ is computed by counting the number of shared tissue-specific genes normalized by the size of the smaller set as suggested in the study by Ramírez and colleagues (Ramírez et al. (2007)):

$$simG(t_1, t_2) = \frac{|TSG_1 \cap TSG_2|}{min\left(|TSG_1|, |TSG_2|\right)}$$

Furthermore, we computed pairwise tissue similarities based on their protein interaction profiles. Let $TSPPI_1$ and $TSPPI_2$ be the sets of tissue-specific protein interactions that occur in the respective tissues and cell lines. Then, the pairwise tissue similarity $simPPI$ can be computed by counting the number of shared tissue-specific protein interactions normalized by the size of the smaller set:

$$simPPI(t_1, t_2) = \frac{|TSPPI_1 \cap TSPPI_2|}{min\left(|TSPPI_1|, |TSPPI_2|\right)}$$

5.2.8 Gene Ontology Enrichments

We computed the protein GO term enrichments for molecular function and cellular component using the web-based tool GOrilla (Eden et al. (2009)) with a p-value threshold of

10^{-4} and default parameters otherwise. The domain GO terms were obtained from Pfam release 24 (Finn et al. (2010)) and the enrichments were computed with the R-package topGO using default parameters (Alexa et al. (2006)).

5.2.9 Interaction Degree

The interaction degree of a protein describes the number of interactions that some protein forms. The interaction degree can be obtained from a network graph by counting all adjacent edges of a protein node. In the following, we distinguish between the *upper-bound interaction degree* and the *expressed-interaction degree*. The upper-bound interaction degree describes the number of edges adjacent to a protein node in the static protein interaction network. The expressed-interaction degree is context-dependent and incorporates information on the protein expression. Here, only edges that represent an interaction between two expressed proteins contribute to the interaction degree of a protein node.

5.3 Protein Interaction Analysis

In the first part of this work, we analyzed the protein interaction network in order to identify universal and tissue-specific protein interactions. Furthermore, we investigated several functional characteristics of proteins forming tissue-specific and universal interactions. Of course, the analyses are dependent on the definition of tissue specificity. The use of PKD results in an increase in the number of tissue-specific genes ($\approx$ 1,300 additional protein-encoding genes) compared to PAD. These additional tissue-specific protein-encoding genes identified by PKD lead to 6,233 additional tissue-specific protein interactions.

5.3.1 Tissue Specificity

We first analyzed the tissue specificity of 63,815 human protein interactions. In contrast to the previous study by Bossi and Lehner based on the Novartis Gene Atlas (Bossi and Lehner (2009); Su et al. (2004)), we find many protein interactions ($\approx$ 73% for PAD and $\approx$ 69% for PKD) to occur universally, and the number of tissue-specific protein interactions to be surprisingly low (less than 5% for PAD and $\approx$ 14% for PKD, *Table 5.1* top). However, in agreement with the study by Bossi and Lehner, we see that, for both PAD and PKD, the tissue specificity of the protein interactions is often determined by one tissue-specific protein interacting with a universal one, while few interactions are formed by two tissue-specific proteins (*Table 5.1* middle). Yet again contrary to the study by Bossi and Lehner, we observe that the majority of universally expressed proteins ($\approx$ 83% for PAD

and ≈57% for PKD) in the protein interaction network do not interact with tissue-specific proteins at all (*Table 5.1* bottom). In summary, our results indicate that by far fewer protein interactions are tissue-specific than previously thought, and tissue diversity seems to involve only few tissue-specific protein interactions.

Tissue specificity of protein interactions	**Number of protein interactions**
universal	46,291 (PAD); 44,006 (PKD)
tissue-specific	2,965 (PAD); 9,198 (PKD)
other	14,559 (PAD); 10,611 (PKD)

Types of interacting proteins	**Number of protein interactions**
both universal	46,393 (PAD); 44,091 (PKD)
universal with tissue-specific	1,574 (PAD); 5,971 (PKD)
both tissue-specific	158 (PAD); 974 (PKD)
other	15,690 (PAD); 12,779 (PKD)

Number of tissue-specific interaction partners	**Number of universal proteins**
0	4,517 (PAD); 2,976 (PKD)
1	650 (PAD); 1,154 (PKD)
2	170 (PAD); 443 (PKD)
3	62 (PAD); 210 (PKD)
4 - 16	68 (PAD); 441 (PKD)
17 - 39	0 (PAD); 24 (PKD)

Table 5.1: *The top part shows the number of protein interactions classified by their tissue specificity. The middle part demonstrates the tissue specificity of proteins involved in the given protein interactions. Many tissue-specific protein interactions are formed by a tissue-specific protein interacting with a universal protein, while interactions of two tissue-specific proteins are very rare. The bottom part gives the number of tissue-specific proteins a universal protein interacts with. This shows that the majority of universal proteins do not interact with tissue-specific proteins.*

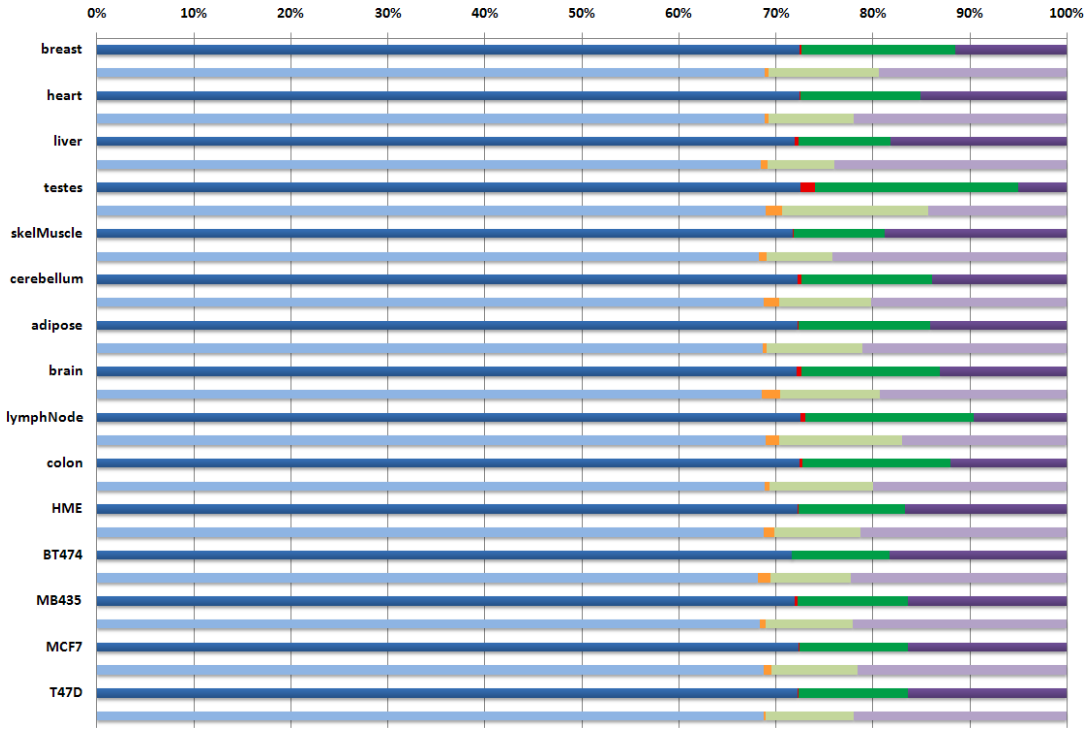


Figure 5.1: *For each of the 15 tissues and cell lines, the respective percentage of universal protein interactions is shown in blue, of specific ones in red, of absent ones in purple, and of the remaining ones in green. For each tissue, the distributions according to PAD and PKD are shown. The PAD results are shown in dark colors, and the corresponding results for PKD are shown in light colors below.*

5.3.2 Protein Interactions across Tissues

The number of tissue-specific protein interactions is very low throughout all tissues and cell lines (*Figure 5.1*). Although the PAD and PKD results vary with respect to their fraction of tissue-specific protein interactions, testis always contains a comparatively high number of tissue-specific interactions, with 977 interactions for PAD and 1,040 interactions for PKD. Furthermore, for both PAD and PKD we find a relatively high number of tissue-specific interactions in brain, cerebellum, and lymph node. The highest number of interactions that are absent in a tissue or cell line due to missing gene expression are found in skeletal muscle for both PAD and PKD (11,979 using PAD and 15,456 using PKD), followed by the breast cancer cell line BT474 and liver tissue. Overall, our results suggest that, in addition to tissue-specific interactions, the absence of certain interactions might be important as well to achieve tissue diversity.

5.3.3 Tissue Similarities

We compared the tissues and cell lines regarding tissue-specific gene expression and protein interactions by computing pairwise similarities based on the number of shared tissue-specific genes and protein interactions. The pairwise similarities change depending on whether gene expression or protein interactions are analyzed. Regarding tissue-specific gene expression profiles, cerebellum and brain show the highest similarity for both PAD and PKD, while the similarities between all other tissues are equally low (*Figure 5.2* **A**, **B**). Interestingly, the investigation of tissue-specific protein interactions reveals high similarities between testis and MB435 in addition to brain and cerebellum as can be seen in *Figure 5.2* **C**, **D**. It has been shown recently that, contrary to the general opinion, MB435 is not a breast cancer cell line (like the other cancer cell lines in this study), but is in fact a melanoma cell line (Chambers (2009)). Interestingly, we find that at the protein interaction level, MB435 is distinguishable from the other cell lines in this study, confirming that there are substantial differences between the breast cancer cell lines and this melanoma cell line.

The same trends are found for both PAD and PKD, although about 1,400 more tissue-specific genes are identified with PKD than with PAD. These findings highlight that, even though the overall similarities between tissues and cell lines are low when considering all tissue-specific genes, only a fraction of the encoded proteins are actually involved in tissue-specific protein-protein interactions.

5.3.4 Enrichments of Gene Ontology Terms

We next analyzed the functions of tissue-specific proteins interacting with universal ones. Our analysis of the enrichments of Gene Ontology terms regarding molecular function reveals that, according to both PAD and PKD, tissue-specific proteins are mainly involved in transporter and receptor activities, which is in agreement with previous observations (Lehner and Fraser (2004); Winter et al. (2004)). Furthermore, tissue-specific proteins are active in the immune system and, according to PKD, are involved in structural activities of the cytoskeleton (*Table 5.2*). When computing enriched GO terms for cellular component, we primarily find extracellular and membrane regions for both PAD and PKD (*Table 5.3*). These findings imply that many tissue-specific cellular processes are induced by the stimulation and activation of receptors, causing tissue-specific signaling cascades. Such signaling cascades can then result in tissue-specific gene expression and protein interactions.

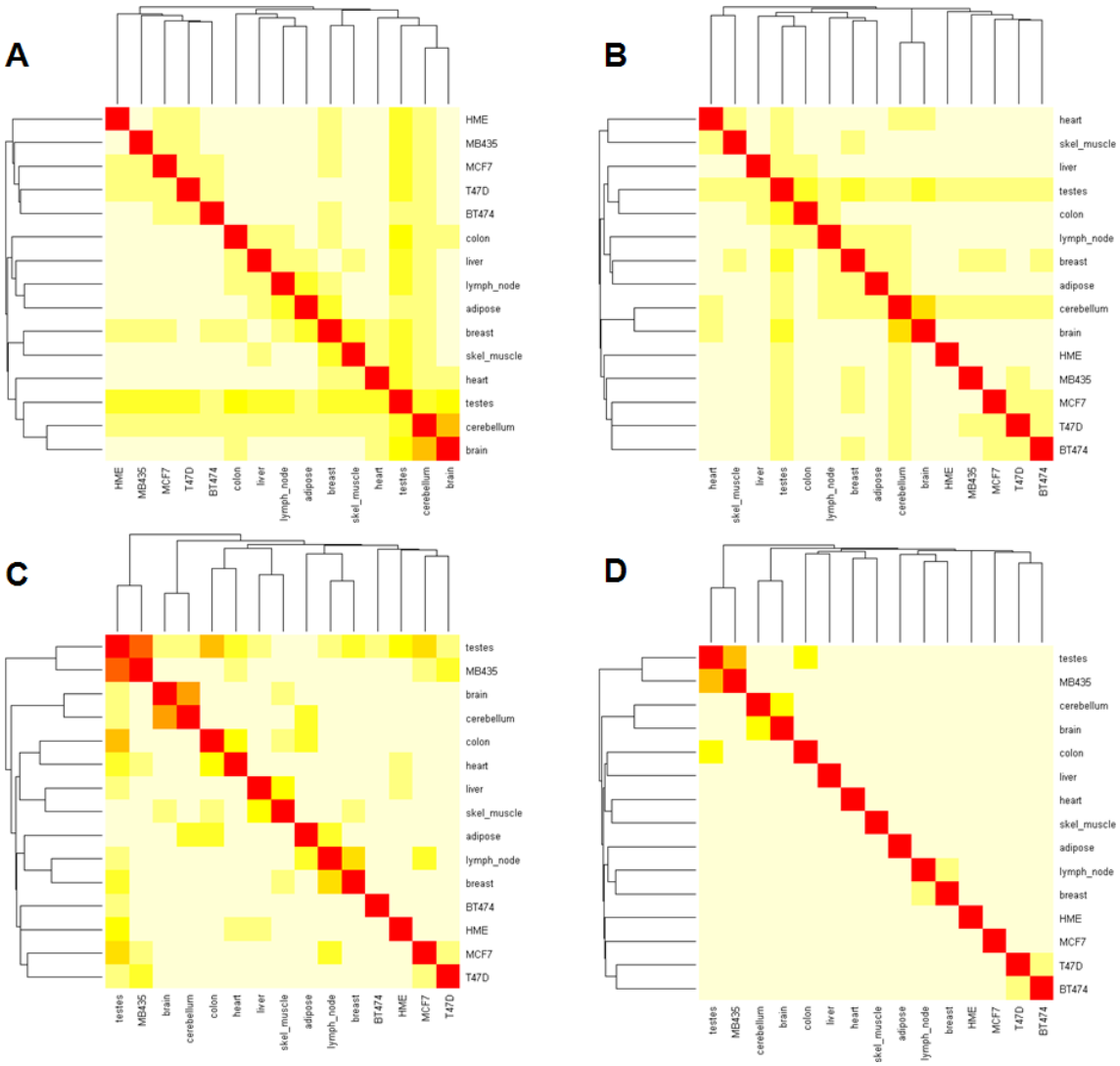


Figure 5.2: ***(A)***, ***(B)*** *The heat maps show the pairwise similarities of tissues and cell lines according to their expression profiles of tissue-specific genes.* ***(C)***, ***(D)*** *The heat map shows the pairwise similarities of tissues and cell lines according to the presence of tissue-specific protein interactions.* ***(A)*** *and* ***(C)*** *are based on PAD,* ***(B)*** *and* ***(D)*** *on PKD. For both heat maps, the color range is from white (completely different) to red (identical).*

5.3.5 Characterization of Drug Targets

Drug targets need to be highly specific for a certain disease to avoid undesirable side effects. When comparing the tissue specificity of protein targets of FDA-approved and

GO Terms - Molecular Function (PAD)	P-Value
G-protein coupled receptor activity	$1.34e^{-10}$
transmembrane receptor activity	$1.48e^{-10}$
receptor activity	$1.59e^{-10}$
cytokine activity	$1.75e^{-10}$
cytokine receptor binding	$1.19e^{-8}$

GO Terms- Molecular Function (PKD)	P-Value
transporter activity	$6.07e^{-11}$
transmembrane transporter activity	$1.65e^{-10}$
substrate-specific transmembrane transporter activity	$5.07e^{-10}$
structural constituent of muscle	$7.16e^{-10}$
substrate-specific channel activity	$4.49e^{-9}$

Table 5.2: *The most significant GO term enrichments for molecular function based on PAD and PKD are listed together with their p-values. The enrichments are computed using GOrilla for the tissue-specific proteins that interact with at least one universal protein.*

experimental drugs, we observed that many more experimental drugs target universal proteins than FDA-approved drugs (for both PAD and PKD). The targets of FDA-approved drugs show a multimodal distribution regarding the number of tissues in which the targets are expressed. The two highest peaks are detected for tissue-specific and universal targets (*Figure 5.3* **A**). Targets of experimental drugs, however, are frequently universal proteins according to both PAD and PKD. Using PKD, we additionally find a large number of tissue-specific drug targets corresponding to over-expressed genes (*Figure 5.3* **B**). This observation is very likely due to the fact that universal proteins are often involved in certain diseases such as cancer that occur in different tissues, while other diseases are associated with very tissue-specific proteins and processes (see *Table A.1*).

5.3.6 Alternative Splicing and Tissue Specificity

We investigated the number of transcript isoforms encoded by the different genes and additionally combined the results with gene expression data to find out whether the number of transcripts depends on the gene expression level. When using PAD, tissue-specific genes encode fewer transcript isoforms on average than universally expressed genes, and

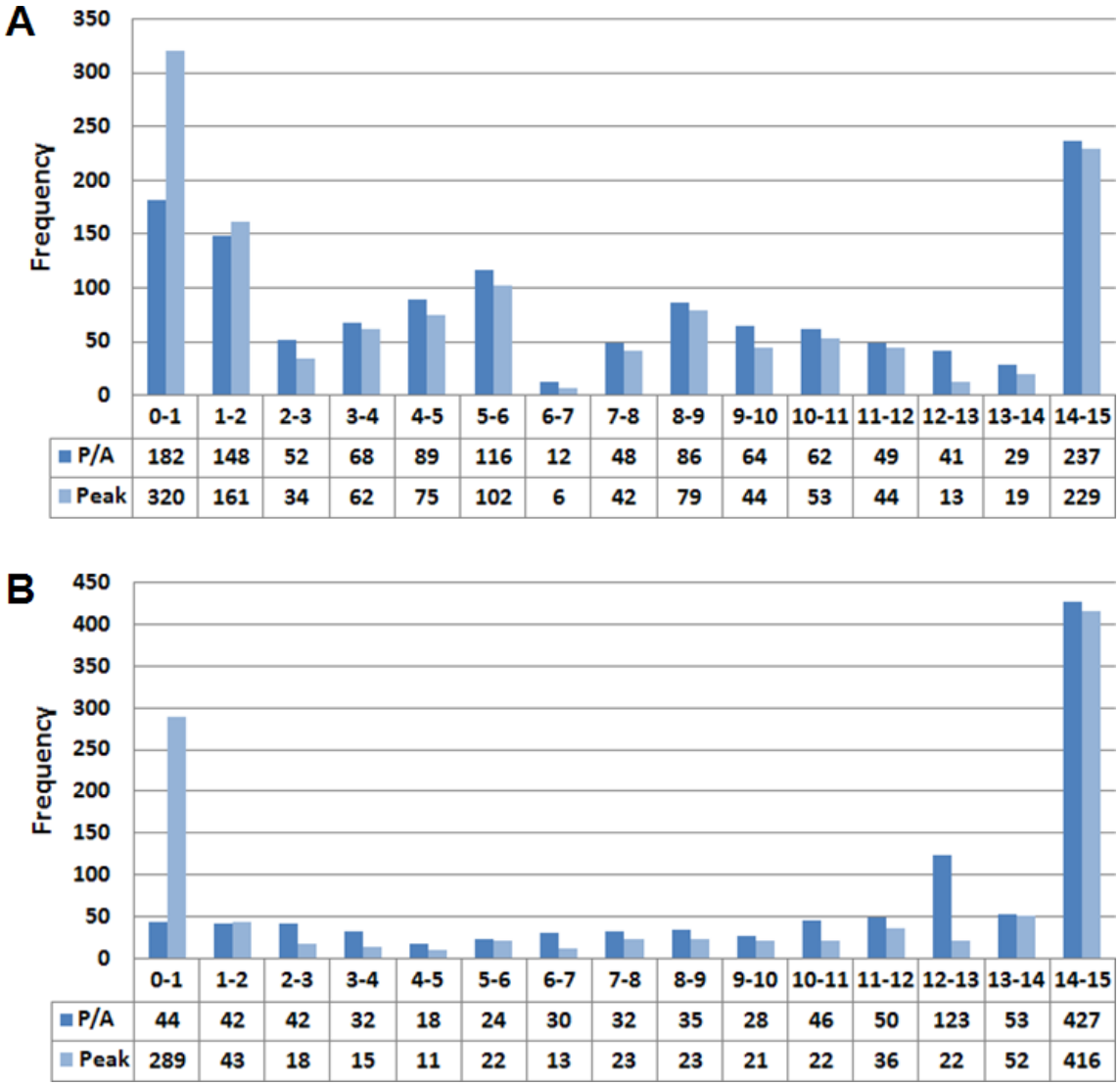
GO Terms - Cellular Component (PAD)	P-Value
extracellular region	$1.61e^{-22}$
intrinsic to plasma membrane	$1.99e^{-15}$
extracellular space	$2.45e^{-15}$
integral to plasma membrane	$4.38e^{-15}$
extracellular region part	$8.87e^{-14}$

GO Terms- Cellular Component (PKD)	P-Value
plasma membrane part	$1.21e^{-27}$
plasma membrane	$1.14e^{-20}$
extracellular region	$1.3e^{-20}$
membrane part	$3.91e^{-17}$
intrinsic to membrane	$1.29e^{-16}$

Table 5.3: *The most significant GO term enrichments for cellular component based on PAD and PKD are listed together with their p-values. The enrichments are computed using GOrilla for the tissue-specific proteins that interact with at least one universal protein.*

the more tissues a gene is expressed in, the more transcript isoforms it encodes (*Figure 5.4* **A**). We find this trend on a genome-wide basis as well as for the subset of genes encoding for proteins involved in the protein interaction network. Interestingly, for PKD, we find a high number of transcript isoforms for tissue-specific genes that are over-expressed in exactly one tissue. This results from genes that are classified as widely expressed by PAD, but are over-expressed in a single tissue and accordingly classified as tissue-specific by PKD. The comparatively high number of transcript isoforms for such tissue-specific genes suggests that specific isoforms of the over-expressed genes may be essential in the respective tissue. It may also be that such genes are actually expressed in multiple tissues as detected by PAD and encode a variety of transcript isoforms, which allow them to adapt to different tissue environments. Here, different transcript isoforms may contain different functional motifs, for instance, miRNA binding sites or protein interaction domains.

The correlation between the expression level of the genes and the transcript isoforms is low for both PAD and PKD, especially when considering only genes encoding proteins involved in the protein interaction network (Pearson correlation 0.07 (PAD) and 0.04 (PKD) in contrast to 0.51 (PAD) and 0.27 (PKD) for all genes; *Figure 5.4* **B**). This suggests that the detection of transcript isoforms does not depend on transcript abundance. We also



	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15
P/A	182	148	52	68	89	116	12	48	86	64	62	49	41	29	237
Peak	320	161	34	62	75	102	6	42	79	44	53	44	13	19	229

	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15
P/A	44	42	42	32	18	24	30	32	35	28	46	50	123	53	427
Peak	289	43	18	15	11	22	13	23	23	21	22	36	22	52	416

Figure 5.3: ***(A)*** *The number of tissues, in which protein targets of FDA-approved drugs are expressed.* ***(B)*** *The number of tissues, in which targets of experimental drugs are expressed. Dark blue bars show the results using PAD, light blue bars refer to results using PKD.*

suppose that the lack of correlation can be due to the fact that the used Ensembl database stores both experimentally verified and computationally derived transcript isoforms. Nevertheless, our results indicate that alternative splicing might be an important mechanism for protein isoforms to function in different environments and to enlarge the repertoire of available interaction partners.

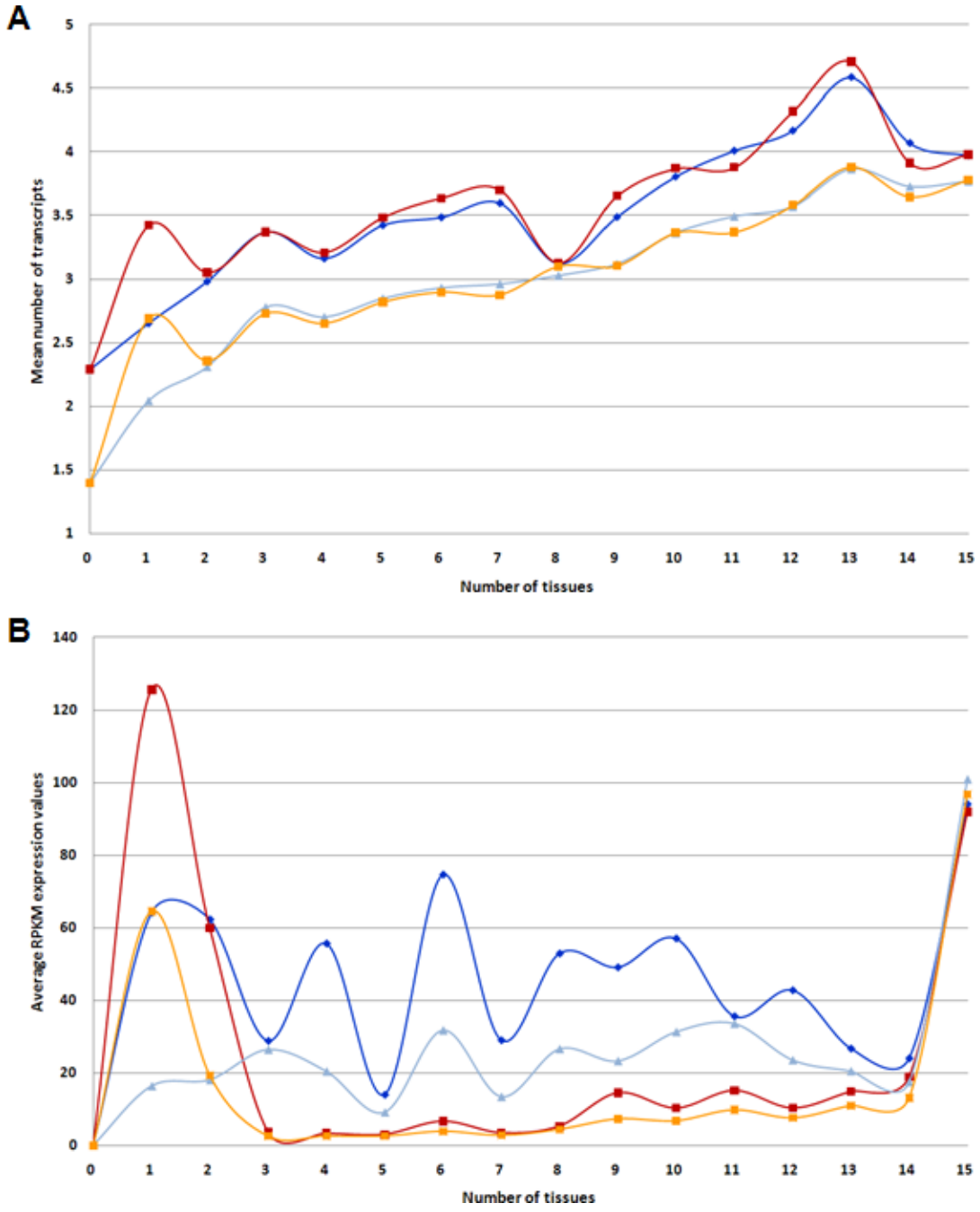


Figure 5.4: ***(A)*** *Tissue specificity related to alternative splicing. The importance of splicing is illustrated by plotting the average number of transcript isoforms produced in the respective number of tissues.* ***(B)*** *Plot of the average RPKM expression values for genes expressed in the respective number of tissues. The dark blue and light blue lines correspond to results found with PAD. The dark red and light red lines depict the results with PKD. Dark colors correspond to genes encoding for network proteins, light colors to all genes.*

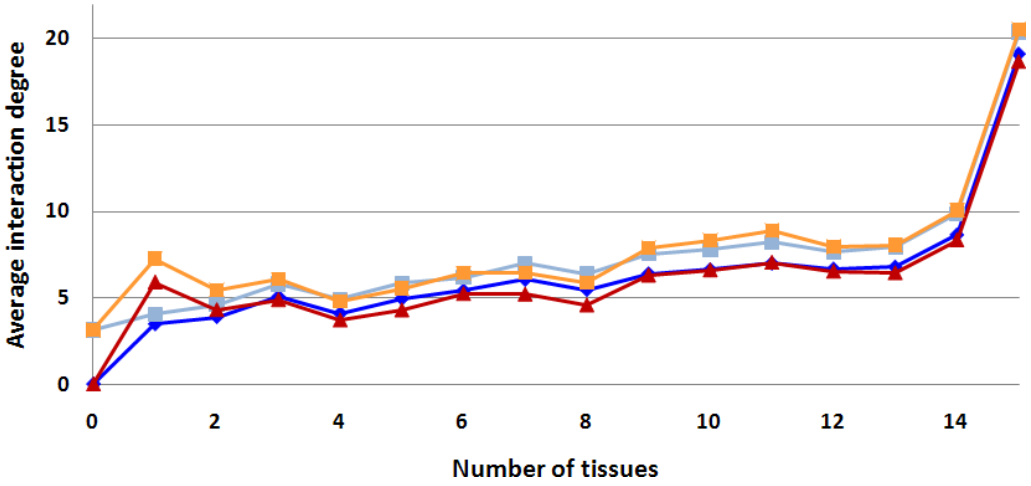


Figure 5.5: *Plot of the average interaction degrees in the respective number of tissues. The interaction degree of a protein describes the number of its interactions. The dark blue and red lines depict the average of the expressed-interaction degrees for PAD and PKD, respectively. The light blue and red lines show the upper bounds of the average interaction degrees as given by the protein interaction network (independent of the tissue expression data) for PAD and PKD.*

Furthermore, we analyzed the interaction degrees in the protein network. We examined all proteins expressed in a certain number of tissues, computed their expressed-interaction degrees according to the number of expressed interaction partners and compared them to the upper bound of the interaction degrees as given by the static interaction network (see Chapter 5.2.9). With respect to both PAD and PKD, we see the same increase of the interaction degrees with the number of tissues as we saw for the number of transcript isoforms (Spearman correlation coefficients for expressed-interaction degree are 0.98 (PAD) and 0.84 (PKD), for the upper bound of the interaction degree 0.98 (PAD) and 0.86 (PKD), *Figure 5.5*), an observation that is in agreement with the study by Bossi and Lehner (Bossi and Lehner (2009)). This suggests that alternative splicing is an important mechanism for increasing the number of protein isoforms from universally expressed genes, which may be an important means for functionally diverse interactions in different tissues.

5.3.7 Combining Definitions of Tissue Specificity

While we identify tissue-specific proteins according to their RPKM values when applying PAD, PKD results point to over-expression in certain tissues. This is exemplified further by means of the STAT1 protein, a well-known signal transducer and activator of transcription

(Liu et al. (1998)). This protein is involved in 76 protein interactions in our protein interaction network. According to PAD, STAT1 is universally expressed with an average RPKM value of 19.7, while STAT1 is classified as tissue-specific in the cancer cell line MCF7 when using PKD, since the expression in MCF7 is about six-fold higher than in the other tissues and cell lines. This comparison demonstrates that applying PAD and PKD in combination can provide additional insights. Here, the combined results of PAD and PKD suggest that STAT1 is universally required for signal transduction and transcription activation and that the specific over-expression of STAT1 in MCF7 is characteristic of this cell line.

5.4 Protein Domain Analysis

In the second part of this study, we investigated functional characteristics of protein domains with respect to their tissue-specific occurrence. The following analyses are not restricted to domains occurring in the interacting proteins of the network analyzed in the first part. Instead, we considered all domains contained in human proteins to discover additional tissue-specific domain functions.

5.4.1 Tissue Specificity

We performed a proteome-wide analysis of Pfam domains to identify universal and tissue-specific domain families. We find many domains to be universally expressed according to both PAD and PKD (1,527 ($\approx$ 54%) for PAD; 1,428 ($\approx$50%) for PKD). However, in case of both PAD and PKD, our results also identify a remarkably large number of domains that are neither universal nor tissue-specific (1,209 ($\approx$ 43%) for PAD; 1,204, ($\approx$ 42%) for PKD). Only 104 (PAD) and 204 (PKD) domains are tissue-specific.

64 ($\approx$ 62%) of the 104 domains that are tissue-specific according to PAD occur in proteins contained in the interaction network. The remaining 40 domains ($\approx$ 38%), however, do not occur in any of the network proteins. In comparison, 147 ($\approx$ 72%) of the 204 tissue-specific domains found with PKD are contained in proteins of the interaction network, while only 57 ($\approx$ 28%) are not.

The proportions of universal domains significantly differ from those of the tissue-specific domains. In case of PAD, 1,265 domains ($\approx$ 83%) are contained in interacting proteins and only 262 ($\approx$ 17%) are not (two-tailed p-value 0.001, Fisher's exact test). Similarly, in case of PKD, 1,175 domains ($\approx$ 82%) are included in interacting proteins, while 253 ($\approx$ 18%) are not (two-tailed p-value 0.129, Fisher's exact test). To sum up, the results

GO Terms - Molecular Function (PAD)	P-Value
receptor binding	$1.10e^{-9}$
cytokine receptor binding	$4.66e^{-8}$
growth factor receptor binding	$1.53e^{-7}$
hormone activity	$1.16e^{-5}$
growth factor activity	$1.33e^{-5}$

GO Terms- Molecular Function (PKD)	P-Value
receptor binding	2.00e-$^{-8}$
growth factor receptor binding	$1.29e^{-7}$
cytokine receptor binding	$3.22e^{-7}$
hormone activity	$7.66e^{-6}$
growth factor activity	0.000401

Table 5.4: *The most significant GO term enrichments for molecular function based on PAD and PKD are listed together with their p-values. The enrichments are computed using topGO for the tissue-specific domains.*

from both PAD and PKD suggest that tissue-specific domains are not necessarily involved in protein-protein interactions, but might also fulfill other biological functions such as DNA-binding.

5.4.2 Enrichments of Gene Ontology Terms

We examined the GO functions of all Pfam domains classified as tissue-specific. The GO term enrichment analysis shows a very similar outcome for both PAD and PKD. The results for molecular function reveal that tissue-specific domains are highly enriched in receptor binding functions (*Table 5.4*). Our analysis also confirms the previous observation that growth factor binding domains are very tissue-specific (Lehner and Fraser (2004)). Furthermore, we find several other less pronounced enrichments, depending on whether PAD or PKD is used, such as symporters, DNA-binding, amino acid binding, lipid binding, and enzymatic activities.

When computing the GO term enrichment for cellular component, we always find DNA- and chromosome-related terms to be enriched in addition to extracellular region (*Table 5.5*), which is in agreement with previous studies (Lehner and Fraser (2004)). Apparently, many tissue-specific domains play an important role in the nucleus and are probably

GO Terms - Cellular Component (PAD)	P-Value
extracellular region	$3.28e^{-17}$
protein-DNA complex	$2.58e^{-5}$
chromatin	$2.58e^{-5}$
chromosomal part	0.000958
chromosome	0.001439

GO Terms- Cellular Component (PKD)	P-Value
extracellular region	$< 1.00e^{-20}$
protein-DNA complex	0.000183
chromatin	0.000183
chromosomal part	0.008503
chromosome	0.016816

Table 5.5: *The most significant GO term enrichments for cellular component based on PAD and PKD are listed together with their p-values. The enrichments are computed using topGO for the tissue-specific domains.*

responsible for transcriptional control. In brief, many tissue-specific proteins form either protein-protein interactions or protein-DNA interactions, the latter of which are not represented by the protein interaction network used in the first part of this work.

5.5 Protein Complex Analysis

The last part of this study concentrates on multimeric complexes to identify tissue-specific and universal protein complexes as well as tissue-specific proteins that might control the formation of a complex in a given tissue. In particular, we analyze the tissue specificity of protein complexes and their assembly.

5.5.1 Tissue Specificity

To investigate the occurrence of protein complexes in the different tissues, we mapped the gene expression data from RNA-sequencing results onto 648 known protein complexes. Surprisingly, we find a large number of universal complexes ($\approx 58\%$ and $\approx 51\%$ according to PAD and PKD, respectively), i.e., complexes that are fully expressed in more than 13 tissues and cell lines. Comparatively few of them are highly tissue-specific ($\approx 4\%$ and $\approx$

21% according to PAD and PKD, respectively), and the remaining complexes are fully expressed in a medium number (3-13) of tissues and cell lines. The complexes included in our study have a minimum size of 3 and a maximum size of 18, but the size is not correlated with the completeness of the expressed complex (Pearson correlation coefficient -0.02 for PAD and -0.19 for PKD). For example, the largest complex in our dataset, the HCF-1 complex (Wysocka et al. (2003)), is fully expressed in all 15 tissues and cell lines according to both PAD and PKD. HCF-1 acts as a transcriptional regulator, and our data suggest that this complex is universally required.

According to both PAD and PKD, the highest number of fully expressed complexes is found in testis, while the lowest number occurs in liver (*Figure 5.6*). In particular, our results also indicate that the complete absence of a complex is very rare among all tissues and cell lines because at least some parts of a protein complex are usually expressed. This supports the hypothesis that cells always maintain partial complexes, which are activated by expressing the missing proteins at appropriate time points (de Lichtenberg et al. (2005)). It also supports the notion of core complexes and attachment proteins, where tissue-specific attachment proteins can alter the function of a complex (Gavin et al. (2006)). Another explanation might be that proteins observed in partially expressed complexes perform multiple biological functions in cells and are needed even in the absence of the complete complex.

5.5.2 Regulation of Tissue Specificity

We identified 28 and 139 tissue-specific protein complexes in our data when applying PAD and PKD, respectively (*Figure 5.7*, see *Table A.2* for the list of all tissue-specific complexes). Interestingly, the results vary considerably for the two definitions. Using PAD yields the most tissue-specific complexes in cerebellum and none or very few in the cell lines, while the PKD results give the highest number of tissue-specific complexes for the HME cell line and slightly less for cerebellum. In contrast to PAD, when using PKD we also identify a high number of tissue-specific complexes in most of the other cell lines. This discrepancy in the results suggests that the cell lines, which are all cancer cell lines except for HME, differ from normal tissues in their gene expression profile regarding the over-expression of certain genes. Gene over-expression is a typical property of cancer cell lines, however, the over-expression results of the HME cell line show that non-cancerogenous cell lines may exhibit similar gene expression profiles as well. The use of PAD only does not allow for the detection of over-expressed genes and in this case applying PKD can help to identify abnormal over-expression of otherwise universal complexes.

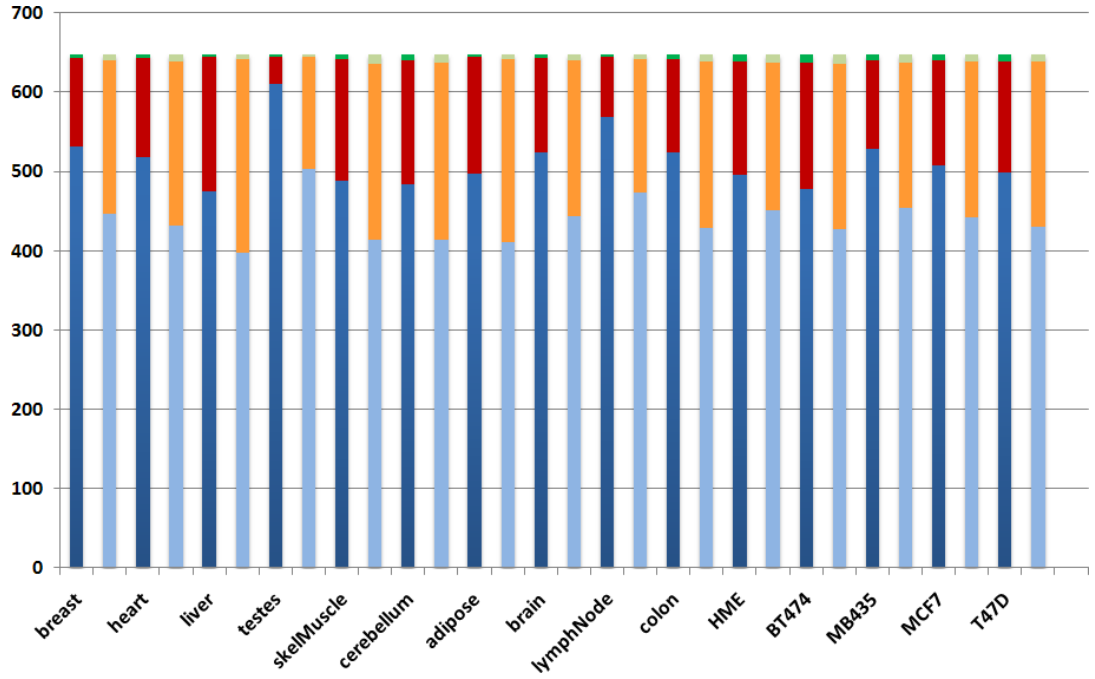


Figure 5.6: *Distribution of fully expressed, partially expressed, and absent complexes in the respective tissues. The dark-colored bars represent the results according to PAD, the light-colored bars according to PKD. The number of fully expressed protein complexes in the respective tissue is shown in blue, the number of partially expressed complexes in red, and the number of absent complexes in green.*

Only 17 of the 28 tissue-specific complexes identified by PAD are fully expressed in at least one of the tissues and cell lines. According to PKD, 113 of the 139 tissue-specific complexes are fully expressed in some tissue or cell line. One interesting observation is that 17 of the 28 complexes identified using PAD (12 of the 17 fully expressed ones) and 123 of the 139 complexes identified using PKD (106 of the 113 fully expressed ones) include universal proteins. Based on PAD, 25 of the complexes (14 of the 17 fully expressed ones) involve one or more tissue-specific proteins, which may be important for regulating the assembly and functioning of the complex. Similar results are found by PKD with 138 of the complexes (112 of the 113 expressed ones) containing tissue-specific proteins. Interestingly, though many of the co-complexed proteins are universal, the formation of the complexes appears to be controlled by very few tissue-specific proteins or even only a single protein.

5.5.3 Tissue Specificity of a SNARE Complex

SNARE complexes are essential for the exocytosis of transport vesicles by mediating the fusion of vesicles and the membrane. Many variations of SNARE complexes exist, and one particular SNARE complex (CORUM identifier 1137) in our study is known to be involved in synaptic transport (Reim et al. (2005)). This complex consists of 7 proteins. According to the PAD results, 3 of them are universally expressed, 3 are neither tissue-specific nor universal, and 1 protein, Complexin-4, is expressed in cerebellum, but no other tissues and cell lines in our study. When applying PKD, we find 3 of the proteins to be universal and 1 to be neither universal nor tissue-specific. In this case, the three proteins Complexin-4, Complexin-3, and SNAP-25 are tissue-specific. Depending on the used definition of tissue specificity, it appears that either Complexin-4 is the sole protein that controls the tissue-specific assembly of this complex or the control is performed together with Complexin-3 and SNAP-25. Strikingly, Complexin-3 and SNAP-25 are found to be over-expressed in cerebellum only. Functional annotations of the proteins suggest that SNAP-25 is contained in the SNARE core complex, while Complexin-3 and Complexin-4 regulate late steps of

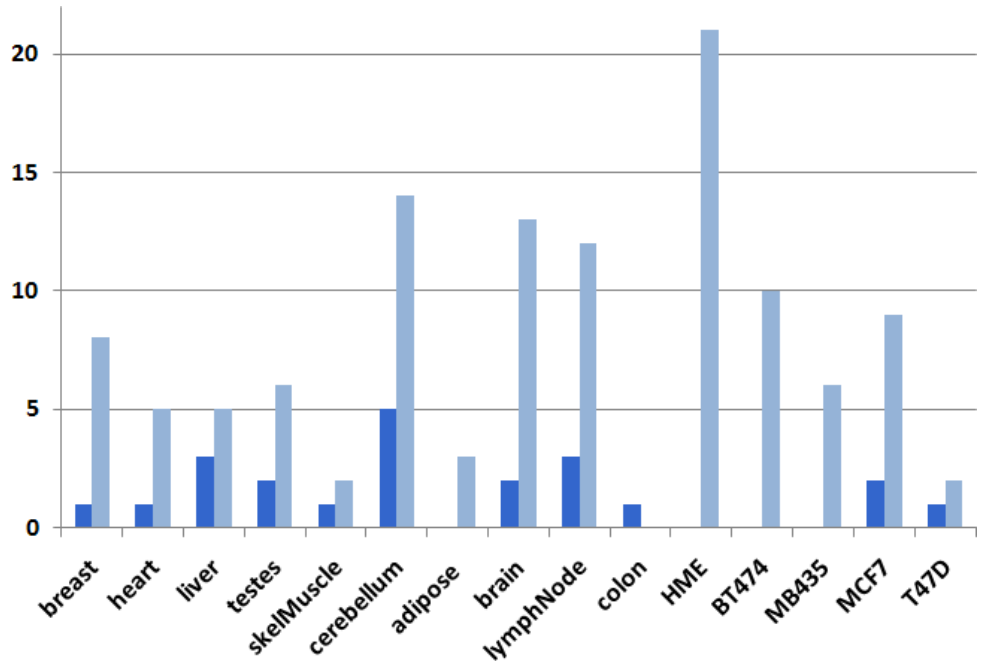


Figure 5.7: *Occurrence of tissue-specific complexes in different tissues and cell lines. The histogram presents the number of specific protein complexes occurring in the tissues and cell lines. The dark blue bars shows the results found using PAD, the light blue according to PKD.*

the vesicle exocytosis. The over-expression of Complexin-3 and Complexin-4 according to PKD could thus be an indicator for a temporal aspect in the tissue-specific complex formation as suggested in general by de Lichtenberg and colleagues (de Lichtenberg et al. (2005)). This example demonstrates that combining results obtained from both definitions of tissue specificity helps identifying tissue-specific as well as time-specific proteins.

5.5.4 Transcriptional Regulation Achieved by a Single Protein

The NCOR-SIN3-HDAC-HESX1 complex (CORUM identifier 3167) functions as transcriptional regulator. According to both PAD and PKD, 5 of the 6 co-complexed proteins are universally expressed, while HESX1 is the only protein expressed in a highly tissue-specific manner in pituitary organogenesis (Dasen et al. (2001)). Correspondingly, HESX1 is not found to be expressed in any of the tissues and cell lines studied by us. HESX1 directs its transcriptional regulation complex to promoter regions that have to be enhanced or silenced at a particular developmental stage. Strikingly, we detected several sub-complexes of NCOR-SIN3-HDAC-HESX1 that are widely expressed in all tissues and cell lines. This suggests that these sub-complexes act as universal transcriptional regulators, and the presence of one additional protein transforms the universal complex into a complex with very specific function. This example points to the possibility that several transcriptional complexes, both the tissue-specific one containing the HESX1 protein and the universally expressed sub-complexes, are present and required in pituitary organogenesis. Alternatively, it is possible that the complex is specifically formed to associate with HESX1 during pituitary organogenesis. This would suggest that HESX1 is a highly tissue-specific attachment protein, which alters the function of the protein complex, while the rest constitutes the complex core.

5.6 Conclusions

Next-generation RNA-sequencing is able to measure gene expression even at low levels. Our functional analyses based on these expression data indicate that substantially fewer protein interactions, protein domains, and protein complexes are responsible for tissue specificity than estimated in previous microarray-based studies. Some tissue-specific functions identified by us agree well with former findings, and our results are more informative and accurate due to the drastically increased detection sensitivity of RNA-sequencing. In particular, we found a remarkably low number of protein interactions to be tissue-specific, many of which are involved in transporter activities or receptor-activated signaling pro-

cesses.

Furthermore, we observed a considerably increased number of transcript isoforms for universally expressed genes. This suggests that the encoded protein isoforms are necessary for different environments and increase the number of possible interactions. We also found universal domains to form protein-protein interactions more frequently than tissue-specific domains, the latter of which are additionally often involved in binding functions such as DNA-interactions. Therefore, transporter activities and receptor activation as well as transcriptional regulation seem to be important factors, besides alternative splicing, for tissue specificity. Moreover, our results suggest that many known protein complexes are widely expressed regardless of their size, and their tissue-specific assembly is often controlled by few tissue-specific proteins.

As with previous studies, our findings rely on the currently available biological datasets and, while the number of tissues and cell lines in this study is still very limited, we assume that the number of sequenced tissues will increase over the next years. Apart from the increase in available tissue samples, we also expect datasets to be made available that contain time-series data. This will allow for analyzing temporal aspects of protein interactions beyond examining the over-expression of genes. Furthermore, many protein interactions in human cells are still unknown to date (Venkatesan et al. (2009)). Therefore, additional biological functions as well as tissue- and time-dependent interactions remain to be discovered in the future.

Our results regarding alternative splicing and the interaction degree of proteins suggest that alternative splicing may be an important means for proteins to increase their interaction repertoire. Proteins that are classified as hub proteins in a static protein interaction network may actually not be single entities but rather be a multitude of protein variants, each of which makes a contribution to the large number of observed protein interaction partners in the currently known interactome. In Chapter 7, we discuss possibilities for research in this field in more detail.

Chapter 6

The Impact of Alternative Splicing on Biological Processes

Alternative splicing is an important molecular mechanism for increasing the protein diversity in eukaryotic cells. However, its functional effects are largely unknown. Alternative splicing events that alter the protein structure and the domain composition can be responsible for the formation of protein interactions. As described in the previous chapter, gene co-expression in a cell is the basis for the potential formation of protein interactions. However, alternative splicing can modify gene products in a way that prevents certain protein interactions. Discovering the occurrence of alternative splicing events and studying protein isoforms are thus important for understanding the effects on biological processes. The identification of alternative splicing events has become feasible using Affymetrix Exon Arrays and, more recently, using RNA-sequencing technologies. In this chapter, we describe the versatile Cytoscape plugin DomainGraph that supports the visual analysis of genes, pathways, and interaction networks and their integration with exon expression data obtained from the Exon Array. DomainGraph, the integration of biological processes with exon expression data, and related aspects such as biological network layouts were the focus of several published papers (Emig et al. (2010b, 2008a,b,c)). These publications are the basis for the following chapter.

6.1 Introduction

Alternative splicing is an important biological mechanism for producing a great variety of eukaryotic protein isoforms from a comparatively small number of genes. Recent studies indicate that about 98% of all multi-exon genes, which is 92-94% of all human genes, undergo alternative splicing (Wang et al. (2008)). A large number of alternatively spliced genes and their protein products are identified by exon tiling microarrays, such as

the Affymetrix Exon Array (Clark et al. (2007)), as well as by deep sequencing of transcriptomes (Sultan et al. (2008); Tang et al. (2009)). Important functional implications of alternative splicing have been demonstrated for selected genes (Stamm et al. (2005); Resch et al. (2004)), but not yet for the large majority of splicing events discovered for many mammalian genes. Splice variants of a gene might differ in the composition of encoded functional regions, such as protein domains and other sequence motifs. Resch and colleagues showed in a number of case studies that several domains, which are responsible for mediating protein-protein interactions, can be inactivated or removed by alternative splicing events (Resch et al. (2004)). Furthermore, protein translation can be prevented by the introduction of a premature stop codon that leads to a transcript encoding for a truncated, non-functional protein. A cellular control mechanism, the nonsense mediated decay, is able to identify such erroneous transcripts and they are degraded without translation (McGlincy and Smith (2008)). The functional impact of alternative splicing can thus be profound, ranging from the gain or loss of specific molecular interactions to changes of pathway dynamics (Leeman and Gilmore (2008)). Recently, it was found that alternative splicing can also be a means of regulating the inclusion of microRNA (miRNA) binding sites into transcript isoforms as an important means of controlling protein expression (Duursma et al. (2008)). While most alternative splicing events are intended by nature, it is known that disrupting the control of alternative splicing can be the cause of diseases (Orengo and Cooper (2007)). Alternative splicing events modifying the sequence of a protein and thus also its functions nondeliberately can be related to diseases such as cancer. For instance, Gardina and coworkers compared samples of healthy and colon cancer tissues and identified a number of differentially expressed genes and novel splicing events that might result in disease-causing protein isoforms (Gardina et al. (2006)). In addition to alternative splicing, post-transcriptional modifications resulting from alternative promoter selection and alternative polyadenylation sites are other critical modes of transcript regulation that may effect protein composition and expression (Millevoi and Vagner (2010); Mayr and Bartel (2009)).

Several stand-alone programs, web services, and Bioconductor packages have been developed to aid in the analysis of Affymetrix Exon Array data and to increase the accuracy and reliability of alternative exon detection. Whereas the majority of currently available tools are principally focused on statistical methods for alternative exon detection, only few report the absolute positions of regulated probesets within transcripts and exons or their positions relative to other regulated probesets. Users thus have to map regulated probesets to the corresponding genomic regions themselves to reveal the potential effects on protein domains and other functional regions, which can be a very cumbersome pro-

cedure. Furthermore, alternative splicing events influencing each other can be identified when analyzing the relative positions of regulated probesets. In addition, none of the currently available programs indicate whether there is prior evidence for alternative splicing or alternative promoter activity in regulated exons and how such events might alter the protein composition in terms of protein domains, motifs or other important sequence elements. The programs easyExon (Chang et al. (2008)) and Expression Console (EC) (Affymetrix (2010c)) add a few biological annotations to their expression statistics such as a probeset-to-gene mapping including the corresponding GO terms. However, users have to perform advanced analyses manually. Other programs, such as the Affymetrix Power Tools (APT) (Affymetrix (2010a)), MADS (Xing et al. (2008)), Exonmap (Yates et al. (2008)), and FIRMA (Purdom et al. (2008)), concentrate on statistical computations only. They do not provide an easy-to-use graphical interface that guides the user through the analysis, and they require prior knowledge of statistical programming languages like R (Okoniewski and Miller (2008)). Web services, such as ExonMiner (Numata et al. (2008)), do not depend on additional tools or prior programming knowledge for the statistical analysis, but require users to upload their potentially confidential microarray data. In summary, few of the described tools provide methods for downstream interpretation of the experimental data, and none of them evaluates the effects of alternative splicing on biological functions that result from the protein domain composition, miRNA binding site inclusion, and modified pathway and interaction dynamics.

Therefore, we developed a software tool called DomainGraph that focuses on the biological effects of alternative splicing events and supports the analysis of exon expression data in the context of interaction networks, pathways, protein domains, and miRNA binding sites. The software allows for exploring the functional impact of alternative splicing and other modes of transcript regulation in human, mouse, and rat. *Table 6.1* provides an overview of the main functionalities of DomainGraph compared to other programs and highlights its unique features.

6.2 Alternative Splicing Analysis

There are two different approaches to the analysis of Exon Array data.

The first approach comprises a *comparative analysis* between two biological groups - an experimental group and a control group - in order to identify probesets that are up- or down-regulated in one of the two biological conditions. A comparative analysis is useful for comparing healthy and diseased tissues, for example, such that the experimental group would correspond to diseased cells and the control group to healthy cells. In this type of

	DomainGraph	AltAnalyze	APT	EC	MADS/JETTA	easyExon	ExonMiner	X:MAP	Exonmap	FIRMA
graphical user interface	X	X		X	X	X	X	X		
cross-platform software	X	X	X		X	JSE6	X	X	X	X
local storage of confidential data	X	X	X	X	X	X		N/A	X	X
CEL file summarization		X	X	X	X	X	X		X	X
gene expression summarization		X	X	X	X					
splicing-index fold		X			X	X			X	
probeset-exon annotation	X	X						X	X	
probeset-gene annotation	X	X		X	X	X	X	X	X	
pathway/GO-annotation	X	X		X		X				
probeset-transcript visualization	X				X			X	X	
probeset-domain visualization	X									
probeset-miRNA-BS visualization	X									
probeset-pathway visualization	X									
web archive export of results	X									
protein-domain composition analysis	X	X								
miRNA-BS analysis	X	X								
probeset-splicing annotation	X	X								
probeset-expression graphs					X	X	X	X		
dependencies / accessory applications	Cytoscape	APT			R|GeneBASE	APT			R|xmap	R|aroma

Table 6.1: *Comparison of DomainGraph to other non-commercial tools. The main functionalities of DomainGraph are compared to programs providing statistical and/or visual methods for the analysis of Affymetrix Exon Array data. Dependencies or accessory applications are indicated in the last row. Unique features of DomainGraph, mainly biological analyses, are highlighted in blue. Abbreviations are as follows: APT = Affymetrix Power Tools, EC = Affymetrix Expression Console, GO = Gene Ontology, miRNA-BS = microRNA binding sites.*

analysis, only those genes are taken into account that are expressed in both biological groups. A well-established measure for such a comparative analysis is the Splicing Index, which reports significantly up- or down-regulated probesets (Srinivasan et al. (2005)). As described in Chapter 2.2.2, the Splicing Index values of the probesets can be computed using AltAnalyze (Salomonis et al. (2009)). The probeset statistics produced by AltAnalyze are specifically designed for input into DomainGraph. DomainGraph annotates the significantly up- or down-regulated probesets with gene and pathway information and facilitates investigating potential functional implications of differentially regulated probesets.

The other approach comprises a *single-array analysis*. Unlike in a comparative analysis, only one biological group is statistically processed and analyzed. The goal is to identify probesets that are detected above background (present call), and those that cannot be reliably detected (absent call). From these probeset presence and absence calls it is possible to infer, which protein regions, domains, and exons are present. Thus, potential alternative splicing events and their effects on protein and domain interactions can be derived and evaluated. For the analysis of single arrays, pre-processing programs such as APT are employed (see Chapter 2.2.2). The pre-processing includes the computation of the probeset expression and probeset p-values. Subsequently, these data can be imported into DomainGraph in order to analyze occurrences of alternative splicing and their effects on specific interaction networks.

6.3 Software Development

Our software tool DomainGraph works as a plugin in the open-source network visualization software Cytoscape (Shannon et al. (2003)). The latest release of DomainGraph enables users to perform both comparative and single-array analyses of Exon Array experiments. To this end, DomainGraph includes a mapping between Exon Array probesets and Ensembl genes, transcripts, exons, proteins, and Pfam domains. An overview of the software and the different analysis options is shown in *Figure 6.1*.

For the comparative analysis option (*Figure 6.1*, Option 1), DomainGraph directly loads and analyzes alternative exon statistics computed with AltAnalyze. This enables users to analyze their data without prior knowledge of genes or pathways potentially affected by alternative splicing. If a probeset shows significant up- or down-regulation according to the results of AltAnalyze, biological annotations such as gene symbols, pathways obtained from WikiPathway and Reactome, alternative splicing annotations, and miRNA binding sites are automatically displayed in a tabular view. Biological data can be selected from the table and visualized along with potential effects of alternative splicing on pathways,

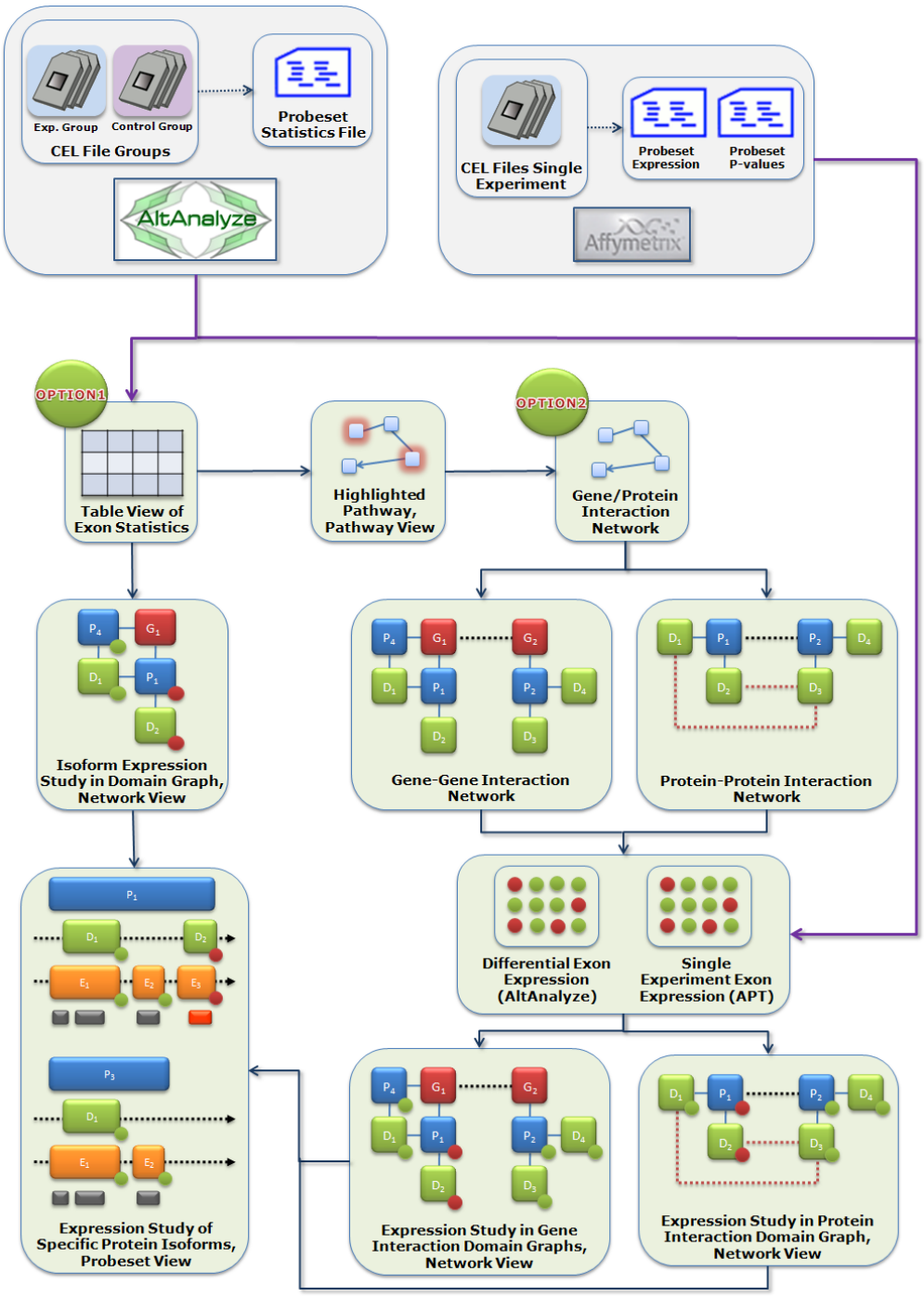


Figure 6.1: *Overview of DomainGraph. DomainGraph supports the visual analysis of statistics pre-processed by AltAnalyze (differential expression of probesets in two biological groups). Additionally, single arrays can be analyzed based on the probeset expression in a single biological group. Purple edges indicate the input points of the AltAnalyze/APT results into DomainGraph. DomainGraph provides a table view, pathway view, network view, and probeset view, which are explained in detail in Chapters 6.3.1 and 6.3.2.*

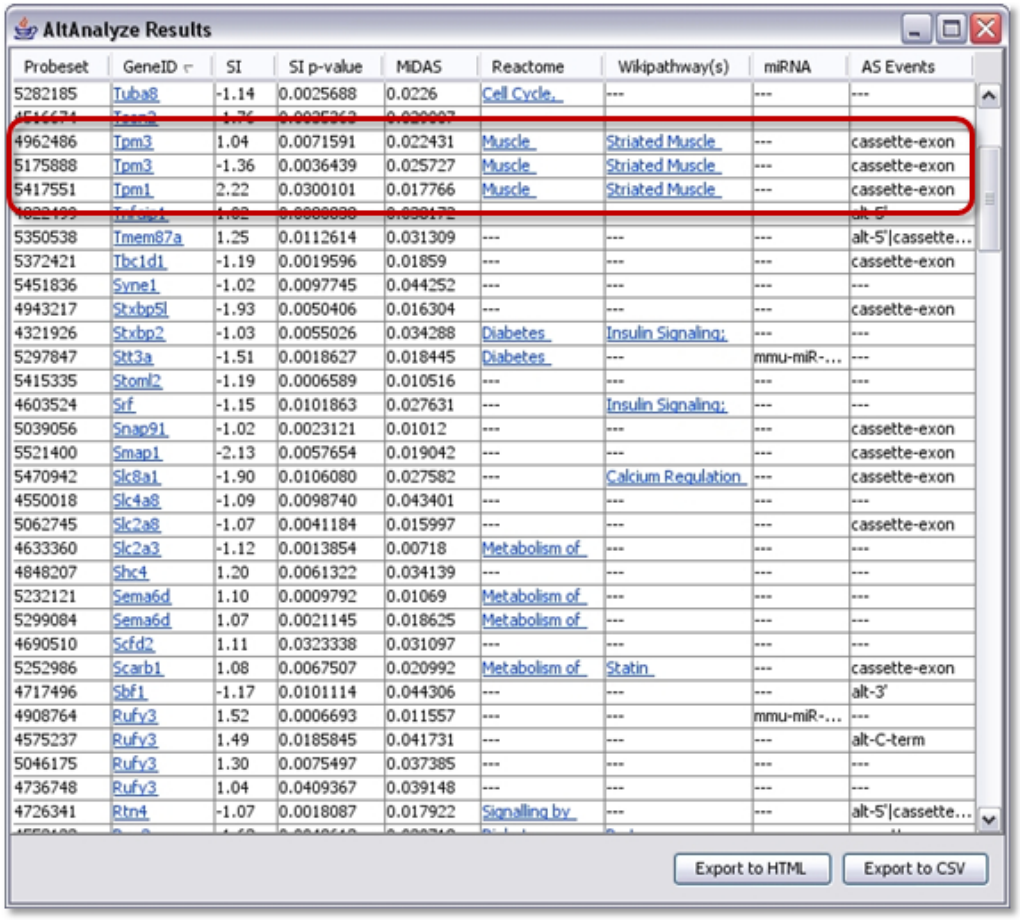
genes, transcripts, exons, protein isoforms, protein domains, and miRNA binding sites.

In addition, DomainGraph supports the analysis of particular gene, protein, and domain interaction networks (*Figure 6.1*, Option 2) and their integration with statistics produced by AltAnalyze, as well as the integration with single-array data. Putative alternative splicing events are highlighted in the networks and can be comprehensively visualized and evaluated at different levels of granularity ranging from network- to exon-level perspectives.

6.3.1 Comparative Analysis of Exon Array Data

The most direct way to evaluate alternative exon statistics computed by AltAnalyze is to view significantly up- and down-regulated probesets in DomainGraph. After importing the AltAnalyze statistics file into DomainGraph, the user is automatically provided with a *table view* containing the AltAnalyze results with information on gene symbols, Reactome and WikiPathway pathway occurrences, miRNA binding site disruption, and alternative splicing annotations for each probeset identified as differentially expressed by AltAnalyze (*Figure 6.2*). Gene and pathway annotations immediately provide an overview of the biological context in which the regulation event occurs. Furthermore, the user can directly obtain a general overview about the up- and down-regulated probesets mapping to putative miRNA binding sites and the genes they belong to. Additionally, several types of alternative exons are annotated in the table, e.g. cassette exons and alternative splice sites (see Chapter 6.6).

The selection of a gene in the table will display a *probeset view* of all gene-encoded protein isoforms together with constituent Pfam domains, corresponding mRNA transcripts and exon structures, Affymetrix Exon Array probesets, and miRNA binding sites (*Figure 6.3*). DomainGraph does not predict new protein isoforms or transcripts, but integrates all information on curated and computationally derived isoforms as stored in the Ensembl database. The *probeset view* enables users to directly compare and analyze alternative exon expression between different protein isoforms produced by the same gene. In this view, probesets are colored according to their differential expression, pointing users to probesets with a significant up- or down-regulation in one of the biological groups, and thus to the corresponding exons, transcripts, and protein isoforms. *Figure 6.3* **A** shows two isoforms of the mouse gene *Tropomyosin 1 (Tpm1)*, which is listed in the *table view* as being alternatively regulated. The *probeset view* reveals that exon 20 is a cassette exon, which is down-regulated in the experimental group. *Figure 6.3* **B** displays two isoforms of the gene *Tropomyosin 3 (Tpm3)*, for which two probesets are alternatively regulated according to the *table view*. A visual investigation highlights that exons 11 and 12 are



Probeset	GeneID	SI	SI p-value	MiDAS	Reactome	Wikipathway(s)	miRNA	AS Events
5282185	Tuba8	-1.14	0.0025688	0.0226	Cell Cycle,	---	---	---
4962486	Tpm3	1.04	0.0071591	0.022431	Muscle	Striated Muscle	---	cassette-exon
5175888	Tpm3	-1.36	0.0036439	0.025727	Muscle	Striated Muscle	---	cassette-exon
5417551	Tpm1	2.22	0.0300101	0.017766	Muscle	Striated Muscle	---	cassette-exon
5350538	Tmem87a	1.25	0.0112614	0.031309	---	---	---	alt-5'|cassette...
5372421	Tbc1d1	-1.19	0.0019596	0.01859	---	---	---	cassette-exon
5451836	Syne1	-1.02	0.0097745	0.044252	---	---	---	---
4943217	Stxbp5l	-1.93	0.0050406	0.016304	---	---	---	cassette-exon
4321926	Stxbp2	-1.03	0.0055026	0.034288	Diabetes	Insulin Signaling;	---	---
5297847	Stt3a	-1.51	0.0018627	0.018445	Diabetes	---	mmu-miR-...	---
5415335	Stoml2	-1.19	0.0006589	0.010516	---	---	---	---
4603524	Srf	-1.15	0.0101863	0.027631	---	Insulin Signaling;	---	---
5039056	Snap91	-1.02	0.0023121	0.01012	---	---	---	cassette-exon
5521400	Smap1	-2.13	0.0057654	0.019042	---	---	---	cassette-exon
5470942	Slc8a1	-1.90	0.0106080	0.027582	---	Calcium Regulation	---	cassette-exon
4550018	Slc4a8	-1.09	0.0098740	0.043401	---	---	---	---
5062745	Slc2a8	-1.07	0.0041184	0.015997	---	---	---	cassette-exon
4633360	Slc2a3	-1.12	0.0013854	0.00718	Metabolism of	---	---	---
4848207	Shc4	1.20	0.0061322	0.034139	---	---	---	---
5232121	Sema6d	1.10	0.0009792	0.01069	Metabolism of	---	---	---
5299084	Sema6d	1.07	0.0021145	0.018625	Metabolism of	---	---	---
4690510	Scfd2	1.11	0.0323338	0.031097	---	---	---	---
5252986	Scarb1	1.08	0.0067507	0.020992	Metabolism of	Statin	---	cassette-exon
4717496	Sbf1	-1.17	0.0101114	0.044306	---	---	---	alt-3'
4908764	Rufy3	1.52	0.0006693	0.011557	---	---	mmu-miR-...	---
4575237	Rufy3	1.49	0.0185845	0.041731	---	---	---	alt-C-term
5046175	Rufy3	1.30	0.0075497	0.037385	---	---	---	---
4736748	Rufy3	1.04	0.0409367	0.039148	---	---	---	---
4726341	Rtn4	-1.07	0.0018087	0.017922	Signalling by	---	---	alt-5'|cassette...

Figure 6.2: *DomainGraph table view for AltAnalyze alternative exon statistics. The table view contains all differentially expressed probesets with biological annotations. Gene and pathway annotations are clickable and lead to the pathway, network, and probeset views. Differentially regulated probesets for the genes Tpm1 and Tpm3 with their statistical and biological annotations are highlighted.*

mutually exclusive exons (i.e. they do not occur in any transcript together). This fact is also emphasized by the simultaneous up-regulation of exon 11 and down-regulation of exon 12 in the experimental group. Additionally, a *single-gene network view* with the gene and all known Ensembl protein isoforms and their domain compositions is shown (*Figure 6.4*).

Furthermore, users can select Reactome or WikiPathways annotations from the table to load and visualize pathways of interest (*pathway view*). These pathways are automatically overlaid with the AltAnalyze probeset statistics, and all network nodes associated with differentially expressed probesets are highlighted to facilitate the identification of potentially

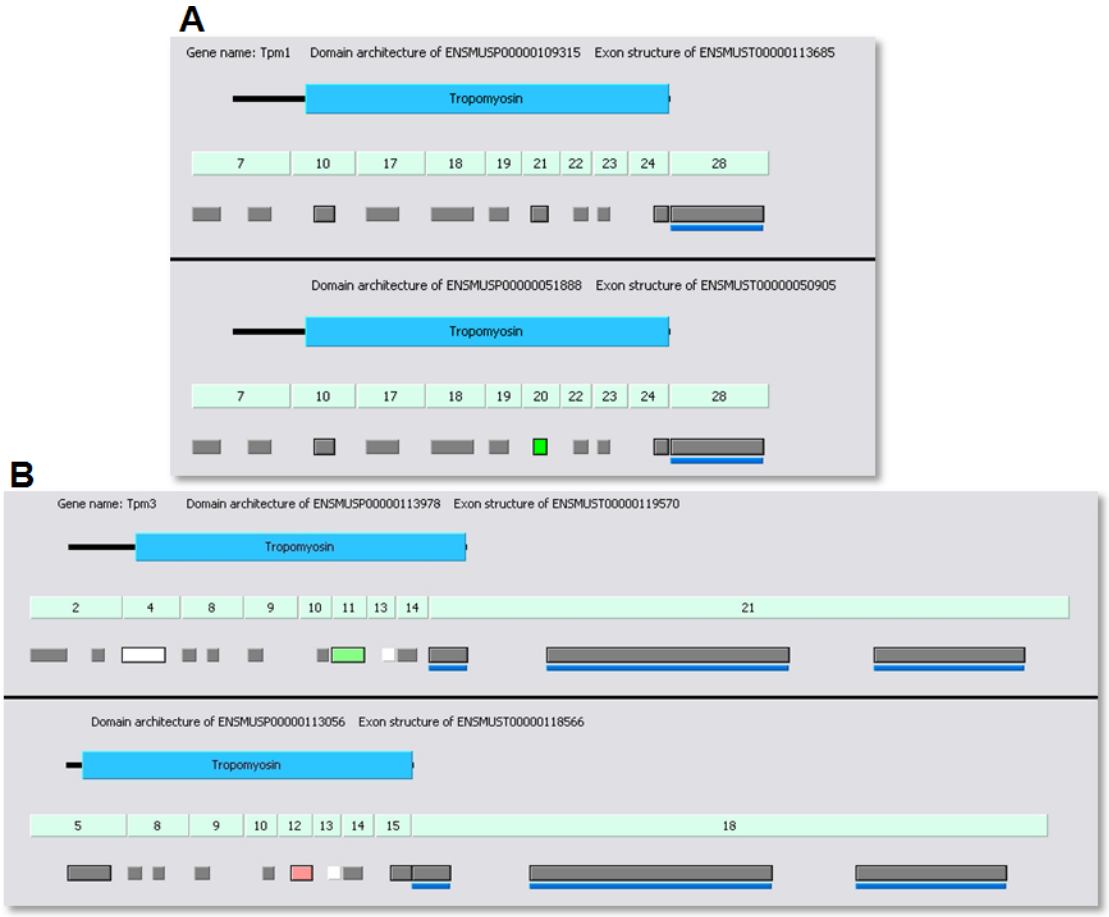


Figure 6.3: *DomainGraph probeset view for AltAnalyze alternative exon statistics.* ***(A)*** *and* ***(B)*** *each display two protein isoforms (black lines) with constituent domains (blue rectangles) and mRNA transcripts (light green, subdivided into exons; identical numbers correspond to the same exons, including 3' and 5' UTRs) produced by Tpm1 and Tpm3, respectively. Probesets are shown below the mRNA transcripts (white probesets did not meet the significance threshold in AltAnalyze; gray boxes for no differential expression among the groups; green boxes for decreased exon inclusion in the experimental group; red boxes for increased exon inclusion in the experimental group; black frames around gray boxes for alternative splicing annotation). Potential miRNA binding sites are drawn as blue lines below probesets.*

modified pathways (*Figure 6.5*). The *table*, *pathway*, and *probeset views* can be exported as an HTML web archive, which can be used to publish the data for all affected genes on a web server (*Figure 6.2*). The web archive includes the table as well as graphics for all

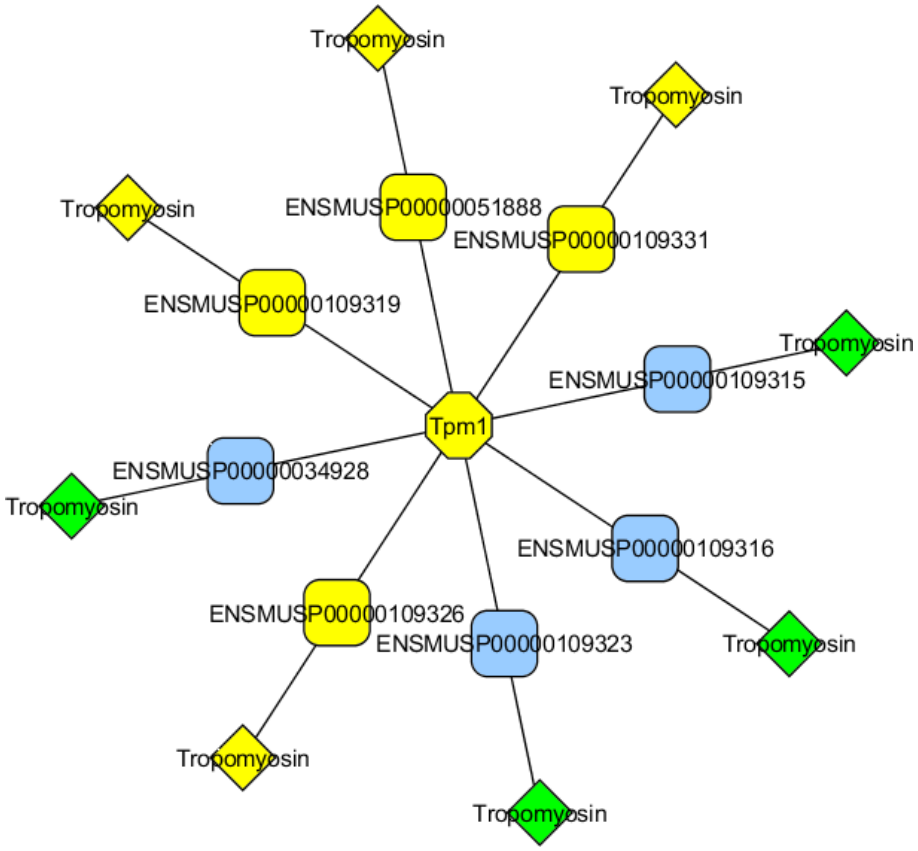


Figure 6.4: *DomainGraph single-gene network view for AltAnalyze alternative exon statistics. The single-gene network view depicts known protein isoforms (rectangle nodes) encoded by Tpm1 together with their constituent domains (diamond nodes). Yellow nodes indicate the node overlaps with regulated probesets.*

alternatively regulated genes and the annotated WikiPathways and Reactome pathways.

6.3.2 Network Analysis

If a user is interested in a particular interaction network or pathway, statistical results obtained from both comparative and single-array analyses can be integrated in order to evaluate protein isoforms or putative protein domain interactions and disruptions thereof. To this end, the user can import either gene or protein interactions into Cytoscape from a flat file or by using other Cytoscape plugins. Interactions can also be obtained from external pathway resources, such as WikiPathways and Reactome. DomainGraph supports using both gene identifiers (Ensembl or Entrez) and protein identifiers (Ensembl or

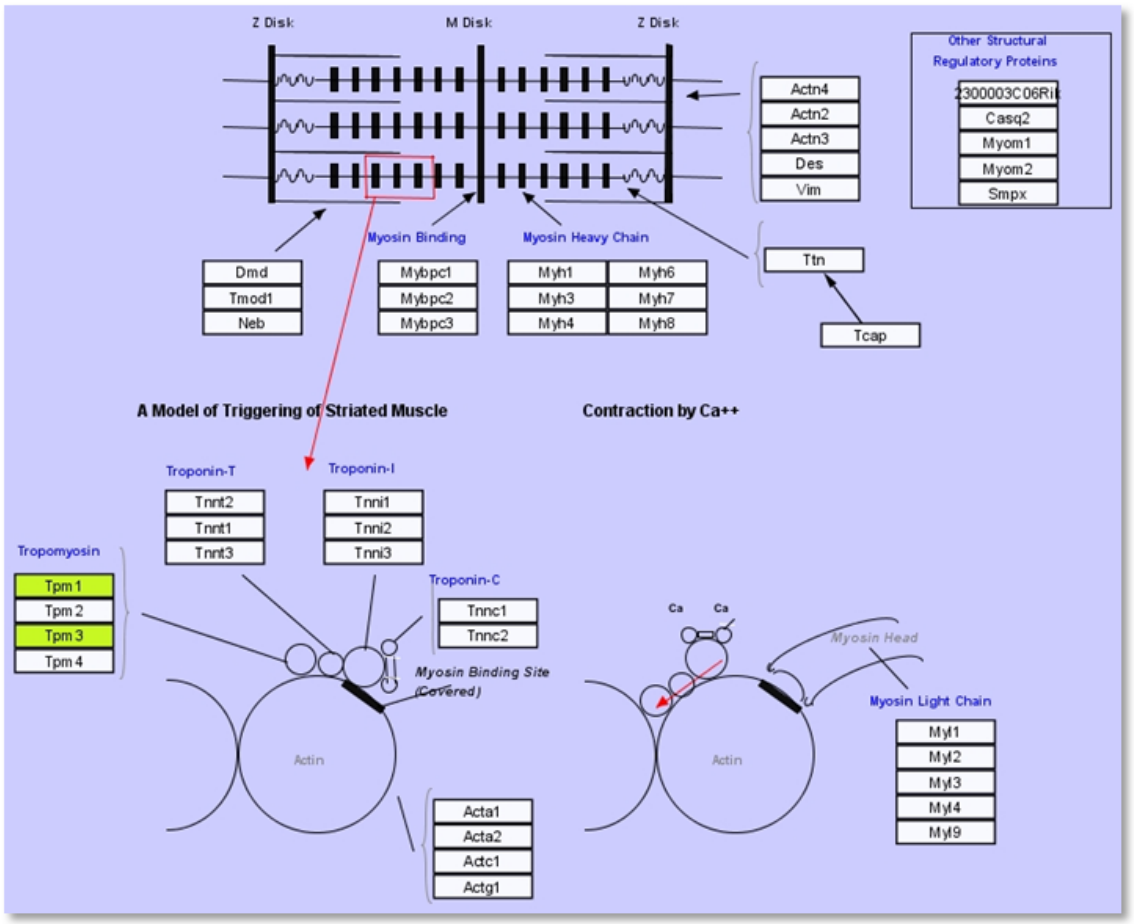


Figure 6.5: *DomainGraph pathway view for AltAnalyze alternative exon statistics. The pathway view shows the Striated Muscle Contraction pathway from WikiPathways; light green colored gene boxes indicate the presence of alternative exons.*

UniProt), since many protein interaction databases provide gene identifiers only and do not specify the protein isoforms involved in the interactions.

When importing a gene interaction network (i.e. the network is given by gene identifiers), the focus lies on the encoded protein isoforms and their domain compositions. The imported genes are visualized as gene nodes and all protein isoforms and their domains are extracted from the embedded DomainGraph database and are automatically added to the gene interaction network. This allows for comparing all protein isoforms regarding their composition to identify those isoforms potentially affected by alternative splicing and those remaining unchanged (*Figure 6.6* **A**).

In contrast, when importing protein interactions, the focus lies on the underlying domain

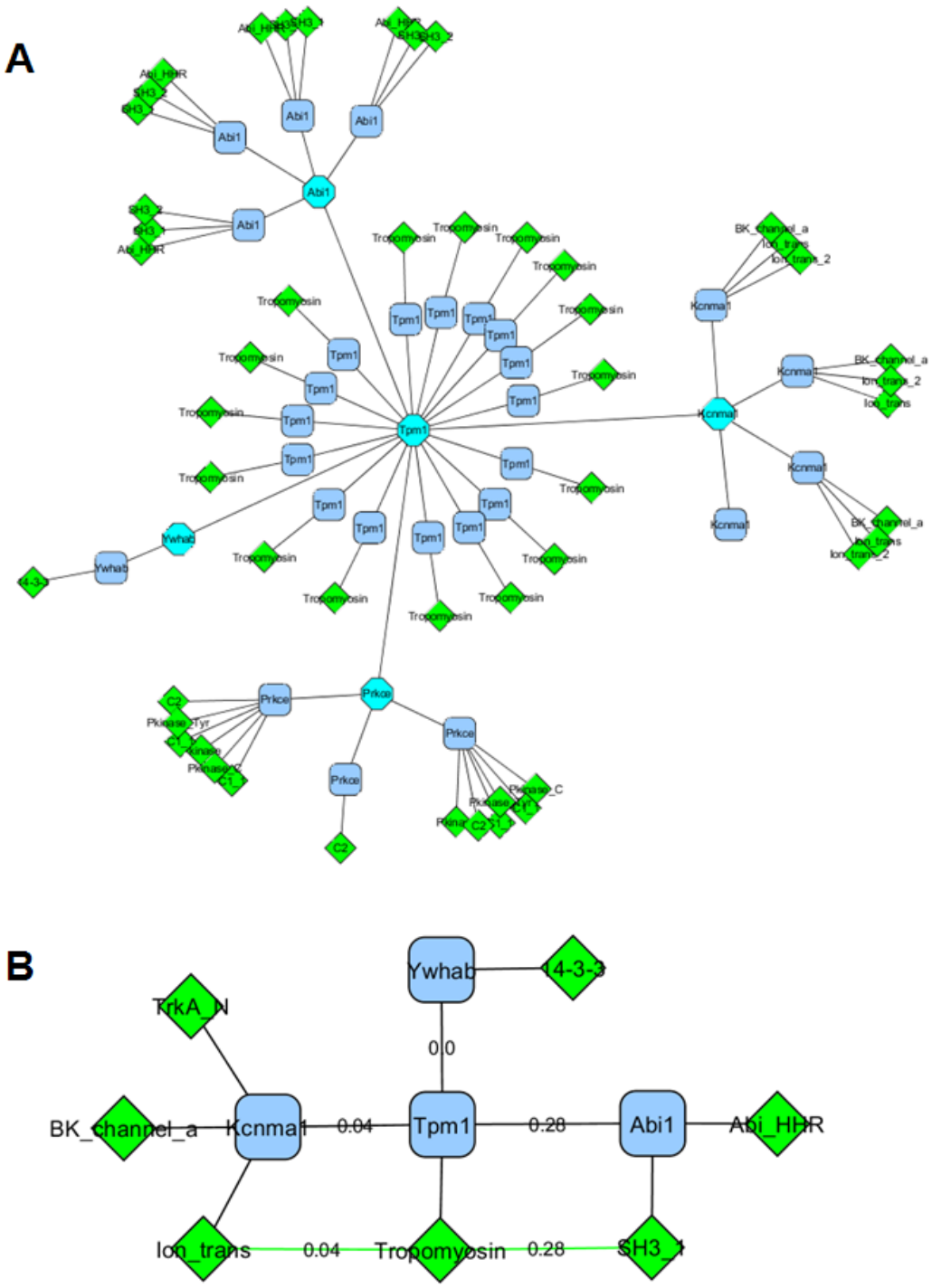


Figure 6.6: ***(A)*** *Gene interaction network: genes (turquoise nodes) with encoded protein isoforms (blue rectangles) and constituent domains (green diamonds) as created by DomainGraph.* ***(B)*** *Protein and domain interaction network created by DomainGraph. Protein nodes are shown as blue rectangles, domain nodes as green diamonds. Domain interactions (green edges) are derived by InterDom. Protein and domain interaction edges are labeled with the corresponding confidence scores.*

interactions of specific protein isoforms, and domain interactions potentially disrupted by alternative splicing can be readily identified. Domain-domain interactions are automatically extracted from the DomainGraph database and the domains and their interactions are added to the network. The user can select domain interactions from twelve different sources: iPfam and 3did were derived from structural data (Finn et al. (2005); Stein et al. (2005)), and the other ten were obtained from various interaction prediction methods (Ng et al. (2003); Liu et al. (2005); Riley et al. (2005); Pagel et al. (2008); Lee et al. (2006); Chen and Liu (2005); Jothi et al. (2006); Schelhorn et al. (2008); Wang et al. (2007); Bjorkholm and Sonnhammer (2009)). The protein and domain interaction edges are annotated with confidence scores if provided by the user-selected domain interaction dataset. For that purpose, the domain interaction edges are labeled with their corresponding confidence scores in a first step. The protein interaction edges are then annotated with the maximum of the confidence scores provided by the underlying domain interactions. If a user-imported protein interaction cannot be traced to any underlying domain interaction, the confidence score 0 is assigned to the protein interaction edge, indicating the uncertainty of the protein interaction (*Figure 6.6* **B**). Annotating the given protein interactions with confidence scores is especially useful if the protein interactions originate from high-throughput methods like yeast-two-hybrid, which are known to be error-prone and may easily contain as many as 50% false positives (Deane et al. (2002)).

Overlaying a Particular Network with Results of Comparative Analysis

The AltAnalyze probeset statistics file can be used for both the comparative analysis of complete Exon Array datasets as described in Chapter 6.3.1 and for investigating a particular network. Once the gene or protein interaction network has been created, AltAnalyze data can be integrated into DomainGraph. Genes, proteins, and domains associated with differentially expressed probesets are automatically highlighted in yellow (*Figure 6.4*, see Chapter 6.6 for details on the implementation). By double-clicking on a gene or protein, the *probeset view* is displayed. Clicking on a gene shows all isoforms encoded by the gene together, while clicking on a protein restricts the view to the respective isoform. Just like in the analysis of a complete dataset, the *probeset view* highlights the differentially expressed probesets (*Figure 6.3*).

Overlaying a Particular Network with Results of Single-Array Analysis

For the single-array analysis, a pre-processed expression data file, and a p-value data file are needed. Both files can easily be generated by applying the 'apt-probeset-summarize'

method included in APT (see Chapter 2.2.2). The expression data file can be directly retrieved from the raw Affymetrix CEL files, for instance, using the provided RMA or PLIER methods for normalization and background correction. The processed expression data file then contains a list of probeset identifiers together with their respective expression values. The p-value data file can also be obtained from the CEL files using the DABG method of APT. This method assigns a p-value to each probeset, which can be seen as a presence or absence call for the respective probeset. The default threshold for the presence of a probeset is set to 0.05, but can be modified by the user.

After pre-processed Exon Array data have been integrated into the network, occurrences and effects of alternative splicing events as well as gene presence and absence are highlighted. To enable the user to notice these occurrences and their effects on the interaction network, the coloring of nodes is adapted accordingly (*Figure 6.7*, **top**). While 'normally' expressed genes, proteins, and domains are colored as introduced in *Figure 6.6*, missing gene expression, alternative splicing events, and indirect effects thereof are visually highlighted. To this end, genes, proteins, and domains, for which no evidence of gene expression was found, are grayed out. To highlight occurrences of alternative splicing in domain-coding regions, domain nodes are colored pink if the domain is partly or completely missing due to alternative splicing events. Since an alternatively spliced domain may not form domain-domain interactions, potential interaction partners are indirectly affected because they would interact with this domain. Therefore, such a domain node is colored orange to point to the loss of a domain-domain interaction due to an alternative splicing event.

The identification of missing gene expression and of alternatively spliced domains according to the expression data is computed from the presence and absence calls of the probesets. The default p-value threshold for the presence or absence call of a probeset is set to 0.05, which is the recommended threshold by Affymetrix. The probesets are first mapped to their corresponding exons. An exon is regarded as expressed if at least 50% of the assigned probesets are expressed. Next, all probesets mapped to the respective protein are counted to determine if the gene should be treated as expressed. By default, we require at least 50% of the assigned probesets to be detected above background and assume the gene is not expressed otherwise. To identify domain expression and domains affected by alternative splicing, the exons are mapped to the corresponding domains. We define a domain to be present if more than 75% of the exons that form the domain are expressed. Otherwise, the domain is regarded as spliced out if the protein itself is expressed. The parameters applied to the integrated Affymetrix expression data are set by default. However, before integrating own expression data, the user can easily customize the default

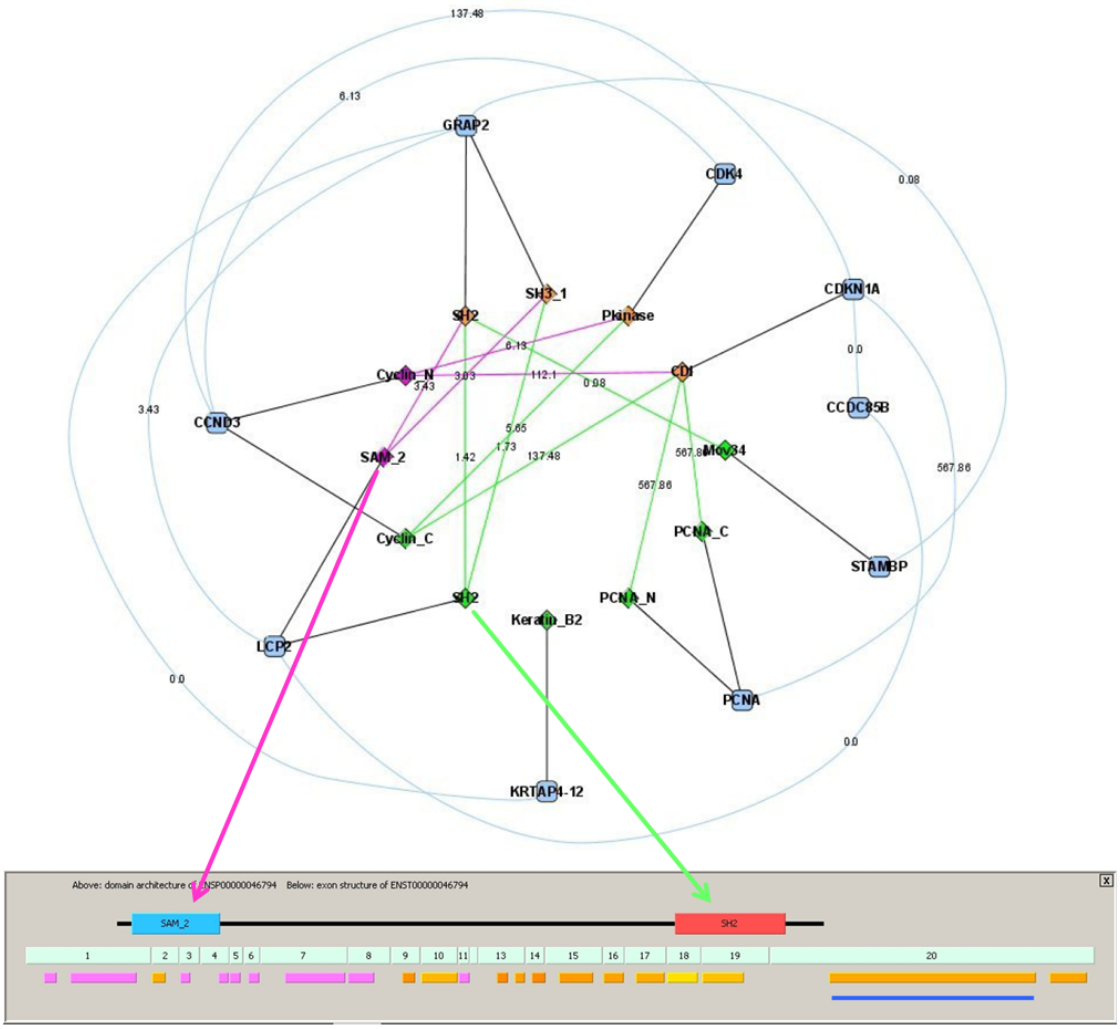


Figure 6.7: ***Top****: Protein and domain interaction network created by DomainGraph. Pink domain nodes indicate the loss of the domain due to alternative splicing events. Orange nodes represent indirectly affected domains.* ***Bottom****: The probeset view shows the LCP2 protein with integrated expression data. Present probesets are colored according to their expression level using a color gradient from yellow (low expression) to red (high expression). Absent probesets are uniformly colored pink.*

values in an options dialog provided by DomainGraph.

The probesets in the *probeset view* of a single-array analysis are colored according to the imported expression and p-value data. A color gradient for the expression strength is applied to all probesets that are present according to the p-value data, while absent probesets are uniformly colored pink. If the proteins in the interaction network are identi-

fied with UniProtKB accession numbers, they are mapped to the corresponding Ensembl proteins first because the Affymetrix mapping is provided for Ensembl identifiers only (see Chapter 6.6 for details). *Figure 6.7* shows a sample protein and domain interaction network with human Exon Array data for testis integrated together with the *probeset view* for the lymphocyte cytosolic protein 2 (LCP2). As can be seen from the pink coloring of the domain node, the SAM_2 domain is regarded as spliced out according to the imported testis data. A closer look at the *probeset view* reveals that the majority of probesets covering this domain-coding region are absent (pink). However, all probesets covering the region of the SH2 domain of LCP2 show expression according to the imported p-value and expression data, and the SH2 domain is thus considered as expressed and the domain node colored green.

6.3.3 Additional Software Features

We support different variations for the *network view* to adjust the visualization. Switching between the views is possible at any time.

For protein interaction networks, we provide three different views: The most detailed

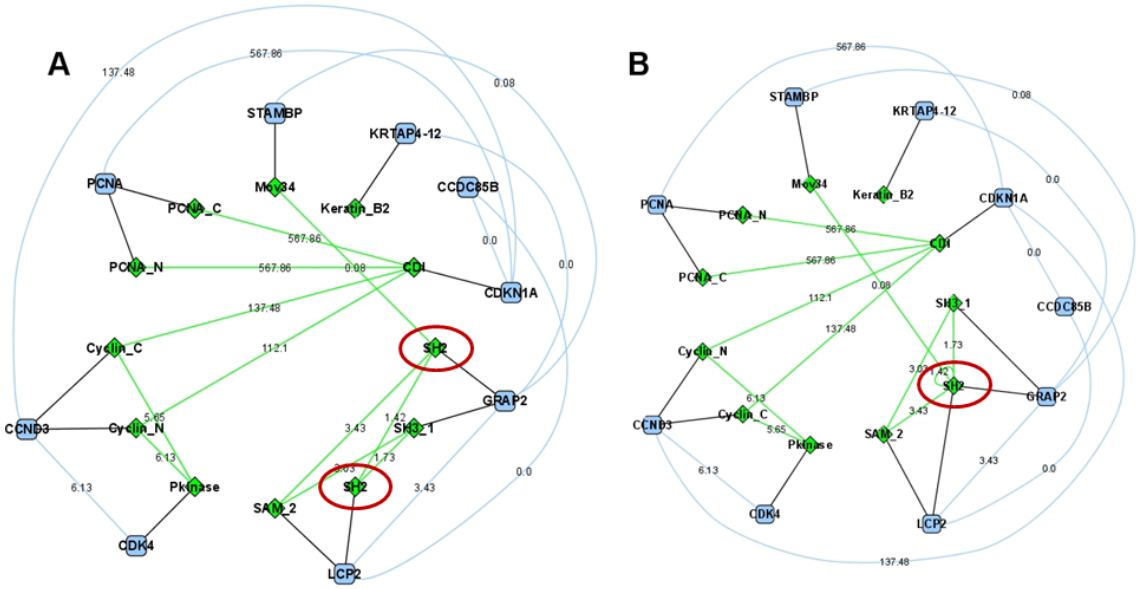


Figure 6.8: ***(A)*** *The extended view and* ***(B)*** *the compact view of a protein and domain interaction network as created by DomainGraph. The SH2 domains (in red circles) of the proteins LCP2 and GRAP2 are displayed separately in the extended view and are merged into a single meta-node in the compact view.*

view is the *extended view*, in which all domain instances for each protein are shown separately. The *compact view* reduces the number of nodes and edges in the network by merging all domains of the same family into a single meta-node and by linking all proteins containing this domain to the meta-node. This *compact view* is especially useful if a large number of proteins contains domains of the same family (*Figure 6.8*). The third view is the *protein interaction network view*, in which only the protein interaction network is displayed. The user can then select the proteins of interest and add the underlying domain-domain interactions for these proteins. This view is useful for exploring the input protein interaction network in a step-by-step fashion without losing track of the relevant data.

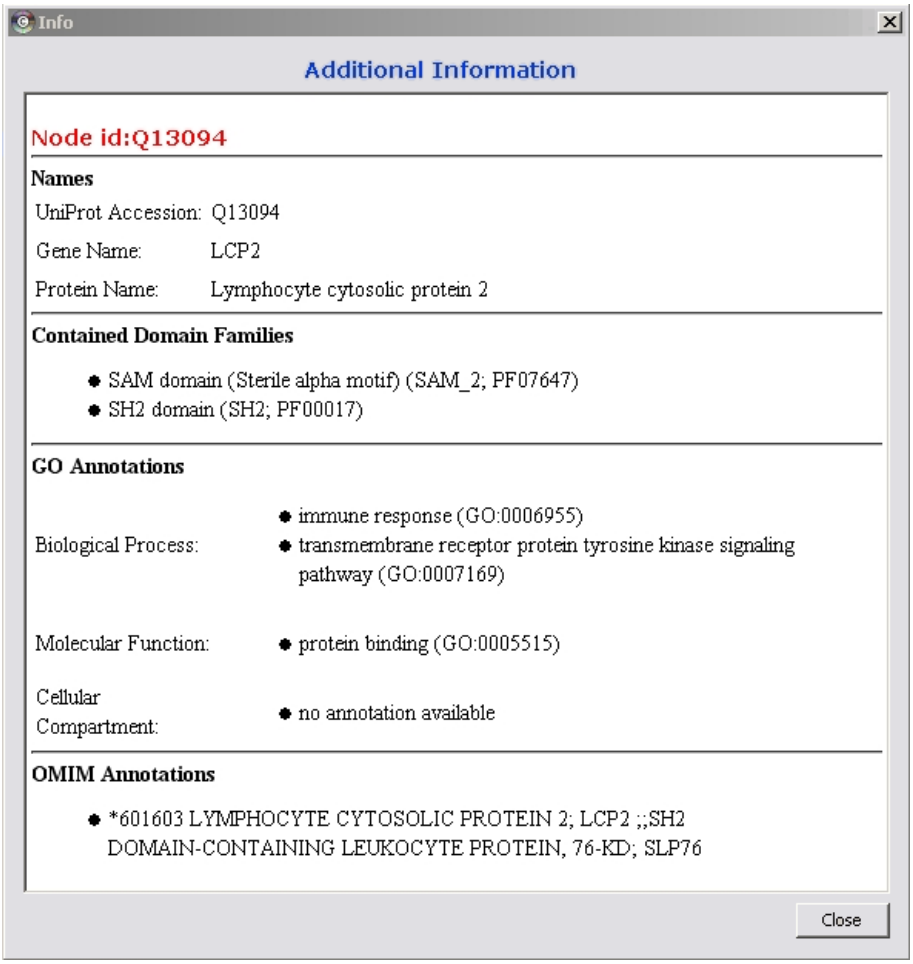


Figure 6.9: *Textual information on the LCP2 protein, including the protein name, constituent domains, Gene Ontology terms for biological process, molecular function, and cellular component, and OMIM disease annotations.*

Similarly, for gene interaction networks, we provide two different views: The *extended view* displays the genes together with all their known protein isoforms and their constituent domains. The other is the *gene interaction network view*, which only shows the user-imported gene interactions to begin with. Users can then select genes of interest and explore their protein isoforms and contained domains in a step-wise manner.

In addition, textual information like Gene Ontology and OMIM disease annotation is available for genes, proteins, and domains to obtain a quick overview of the biological functions and disease associations of alternatively regulated genes (*Figure 6.9*). The gene, protein, and domain nodes are also linked to their source databases via a node context menu such that the user can easily retrieve additional available external information.

Tooltips provide additional information and are available for the gene, protein, and domain nodes in the *network view* as well as for the domains, exons, probesets, and miRNA binding sites in the *probeset view*. For example, regarding the domains and exons in the *probeset view*, information on their lengths and on the start and end positions of a domain within an exon is available. In a comparative analysis, probeset tooltips provide additional information such as the Splicing Index fold change of the probesets together with their p-values, alternative splicing annotations, and cross-hybridization types. The latter indicates if a probeset matches one or several genomic locations (see Chapter 6.6 for details). For single-array analyses, the tooltips of the probesets display information such as the length of the probeset, the expression strength, and the DABG p-value.

6.4 Software Applications

6.4.1 Analysis of a Splicing Factor Knockdown Dataset

As an exemplary application of a comparative analysis using DomainGraph in combination with AltAnalyze, we chose a previously described Exon Array dataset from an experiment, in which a splicing factor, the polypyrimidine tract binding protein (PTB), was knocked down in a mouse neuroblastoma cell line using PTB short-hairpin RNA (shRNA) (Xing et al. (2008)). The corresponding Affymetrix CEL files were downloaded from the Gene Expression Omnibus (GEO, accession number GSE11344). The Exon Array samples were divided into a PTB-shRNA group (the experimental group) and an empty-vector treated group (the control group), each consisting of three biological replicates, and were statistically processed by AltAnalyze using default parameters. Out of 110,092 core probesets, i.e. probesets that are mapped to Ensembl exons, this analysis yielded 205 alternative probesets corresponding to 150 unique genes. 30 of these alternative probesets were also

identified in the previous analysis and 27 of them were confirmed by RT-PCR (Xing et al. (2008)). According to the results of AltAnalyze, the majority of the alternatively expressed probesets, 144 in total, are predicted to directly or indirectly alter the composition of functional protein regions such as protein domains, while 11 probesets overlap with putative miRNA binding sites.

AltAnalyze probeset statistics for this dataset were then imported into DomainGraph for a general analysis of the dataset with all alternatively expressed probesets displayed in the *table view* of DomainGraph (*Figure 6.2*). As can be seen from the table, there are two alternatively regulated probesets assigned to the gene *Tropomysin 3 (Tpm3)* and one assigned to *Tropomyosin 1 (Tpm1)*. Tropomyosin, along with Actin and Troponin, is an important factor for muscle contraction and several alternative transcripts of the Tropomyosin genes are known (Lees-Miller and Helfman (1991)). The *probeset view* of two protein isoforms and domain compositions together with the mRNA transcripts and miRNA binding sites of *Tpm3* are shown in *Figure 6.3*. As can be seen, the two probesets that have been identified as differentially expressed align to the mutually exclusive exons 11 and 12, respectively. A literature search reveals that the up-regulated isoform containing exon 12 uniquely associates with the Golgi apparatus, while the down-regulated isoform associates with stress fibers (Percival et al. (2004)). Selecting the WikiPathway *Striated Muscle Contraction*, which is annotated to *Tpm3*, illustrates that the *Tpm3*-interacting gene, *Tpm1*, is the only other gene in the pathway likely to undergo alternative splicing (*Figure 6.5*). The regulation of alternative splicing by *PTB* may thus impact interactions between distinct tropomyosin genes and this observation can serve as a starting point for biologists to analyze the interaction of the specific *Tpm1* and *Tpm3* isoforms in more detail to elucidate their biological functions.

6.4.2 Comparison of Human Embryonic Stem Cells and Cardiac Precursors

In a second application example, we performed a comparative analysis using a previously described Exon Array dataset (GEO accession GSE13297) containing data for human embryonic stem cells and cardiac precursors (Salomonis et al. (2009)). The main goal of the described analysis was the identification of alternative splicing events potentially involved in the development of stem cells into cardiac precursors. The Affymetrix CEL files were first processed in AltAnalyze using default parameters, with cardiac precursors treated as experimental group and human embryonic stem cells as control group. AltAnalyze found 187,569 core probesets remaining after all filtering steps, with 4,660 of them significantly up- or down-regulated in one of the biological groups.

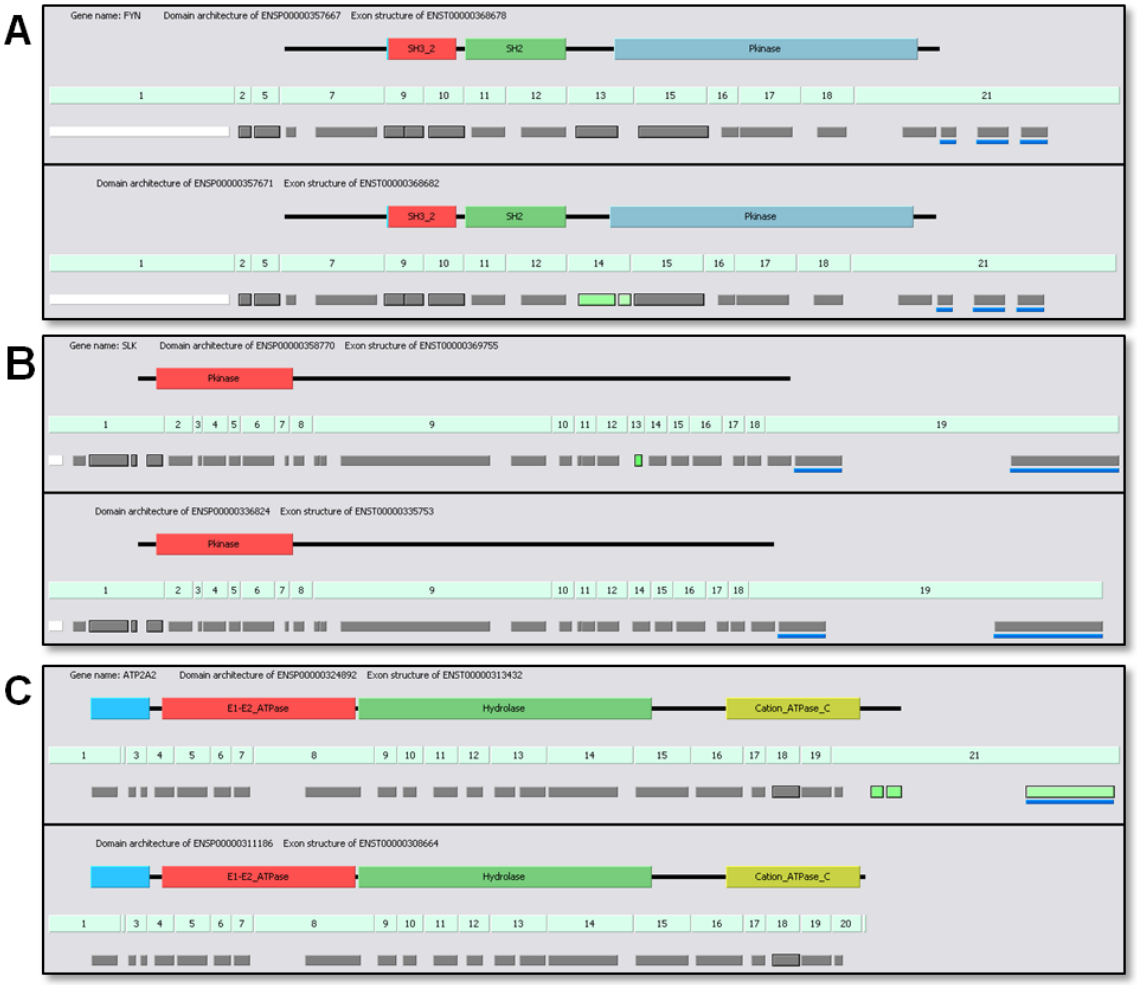


Figure 6.10: *Differentially expressed exons and their functional impact.* ***(A)*** *shows the mutually exclusive exons 13 and 14 overlapping the Pkinase domain region of FYN.* ***(B)*** *displays the alternative regulation of a cassette exon in a disordered protein region encoded by SLK.* ***(C)*** *shows the alternative regulation of an exon containing a putative miRNA binding site in ATP2A2.*

Importing the AltAnalyze results into DomainGraph, we found the effects of alternative probeset expression to be diverse in terms of the apparent mechanism of action and its functional impact. Alternative exon expression of the gene products described in the following has also been verified experimentally (Salomonis et al. (2009)). *Figure 6.10* **A-B** shows the tyrosine-protein kinase *FYN* and the serine/threonine-protein kinase *SLK*. The *probeset view* for two transcripts of *FYN* reveals that the exon 13 and 14 are mutually exclusive and exon 14 is down-regulated in the cardiac precursor group. These exons

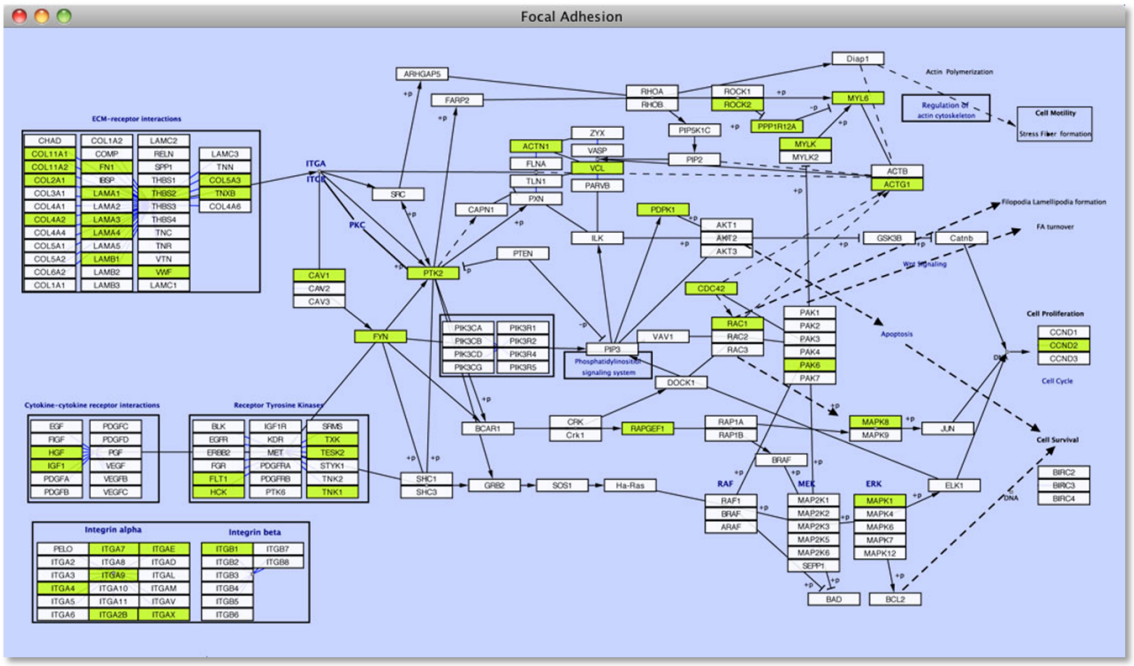


Figure 6.11: *The 'Focal Adhesion' pathway retrieved from WikiPathways is displayed with light green colored gene boxes indicating the presence of alternative exons for that gene.*

partly overlap with the Pkinase domain region and the exchange of the two exons may thus have an effect on the functioning of the domain. The *probeset view* of the two isoforms encoded by *SLK* demonstrates the down-regulation of the cassette exon 13 in the cardiac precursors. Although this exon does not overlap a domain-coding region, but falls within a disordered region of the protein, it may still have a functional impact on the expressed protein variant. Disordered regions may contain short linear motifs or functional residues, which fulfill particular functions and are known to play an important role in protein interactions (Stein and Aloy (2010)). The web service NetPhos predicts the occurrence of a serine phosphorylation site within this cassette exon (Blom et al. (1999)). To confirm this prediction, we additionally checked the protein using the 'Eukaryotic Linear Motif resource' (ELM) (Puntervoll et al. (2003)). Like NetPhos, ELM predicts this phosphorylation site, suggesting that alternative splicing may alter protein functions by modifications of domains and disordered regions. Finally, *Figure 6.10* **C** shows two protein isoforms of *ATP2A2*. As can be seen from the probeset coloring, the longer isoform is down-regulated in cardiac precursors. This region includes a predicted miRNA binding site at the 3'UTR of the mRNA, which may be responsible for mRNA stability.

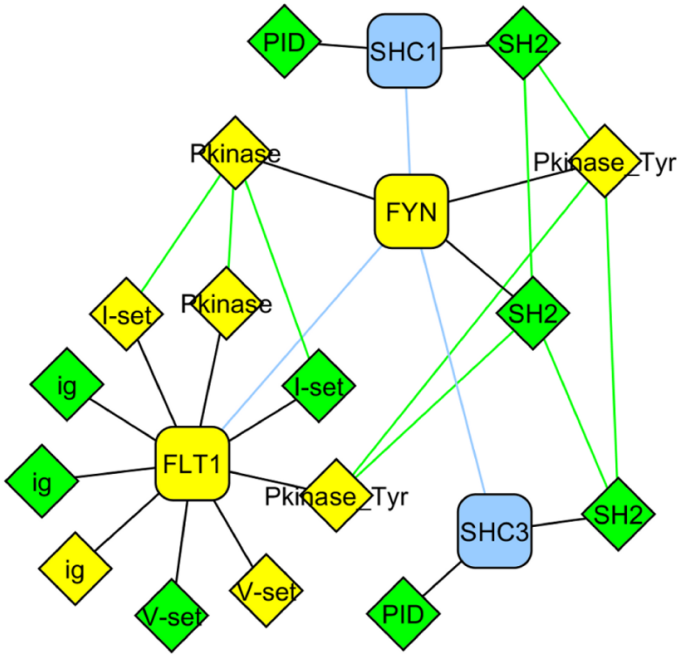


Figure 6.12: *The impact of alternative exon expression on protein and domain interactions. Four genes of the 'Focal Adhesion' WikiPathway are shown. FYN and FLT1 as well as some of their constituent domains are highlighted in yellow to indicate occurrences of regulated probesets. Domain interactions (obtained from iPfam) for one protein isoform per gene are shown. Protein isoforms are depicted as rectangle nodes and constituent domains as diamond nodes. Protein interactions are drawn as blue edges and domain interactions as green edges.*

How these alternative exons affect larger biological processes can be assessed further by examining the interactions between these alternative genes in biological pathways or by examining protein and domain interactions between these genes. The DomainGraph annotations reveal that *FYN* is involved in a number of pathways, four Reactome and eleven WikiPathway pathways. One of them is the *Focal Adhesion* WikiPathway, which is automatically overlaid with the AltAnalyze probeset statistics (*Figure 6.11*). The pathway shows that, besides *FYN*, there are several other proteins that may be functionally modified due to alternatively regulated exons. A subset of the proteins participating in this pathway were imported into DomainGraph to evaluate the potential effects of alternative splicing on their interactions. DomainGraph automatically adds putative domain interactions to the network and highlights potentially affected domain interactions (*Figure 6.12*). This interaction network specifically demonstrates that alternative exon inclusion within

the domain of both binding partners has the potential to significantly alter interactions in the *Focal Adhesion* WikiPathway.

6.5 Laying Out Protein and Domain Networks

For the visualization of protein interaction networks, several algorithms already exist that consider biological information for layout computation. For example, the approach by Ho *et al.* visualizes biological data integrated with protein networks and supports protein complex information to group proteins that are members of a single complex into a common region of the layout space (Ho et al. (2008)). Another idea is the animated visualization of typed interaction networks, by which a subnetwork of proteins that support a specific, user-selected type of interaction is displayed using a force-directed algorithm, while the remaining proteins are placed on a circle around the drawing of the subnetwork (Friedrich and Schreiber (2003)). In order to reduce the number of crossings in the layout, Kato *et al.* add a crossing cost penalty to a grid-based layout algorithm, which can also deal with placement constraints to model subcellular localization information (Kato et al. (2005)).

Other methods try to overcome the problem of edge crossings by using three-dimensional visualizations. However, when such visualizations are displayed on a plane such as provided by a standard monitor, the problem of intersection and overlap occurs again, since nodes and edges that are perfectly laid out in three-dimensional space may appear on top of each other when visualized in a plane. Additionally, three-dimensional visualizations often pose orientation problems for the users, as the third dimension can only be comprehended by rotating the network (Ho et al. (2008); Han and Ju (2003)). Li and Kurata present a grid-based layout algorithm that allows for clustering the objects in the drawing based on their membership in functional modules (Li and Kurata (2005)).

6.5.1 Layout Requirements

Even though existing graph drawing algorithms like the automatic layouts provided by Cytoscape may be a good starting point for the development of methods for visualizing biological networks, they need to be adapted to reflect specific biological information in the computed drawing. Moreover, applying generic approaches to dense biological networks often leads to cluttered layouts with a large number of node and edge intersections. *Figure 6.13* shows two examples of a protein and domain interaction network created by DomainGraph using two Cytoscape standard algorithms for grid and force-directed layout. As can be seen, the results may be far away from biologically meaningful layouts that

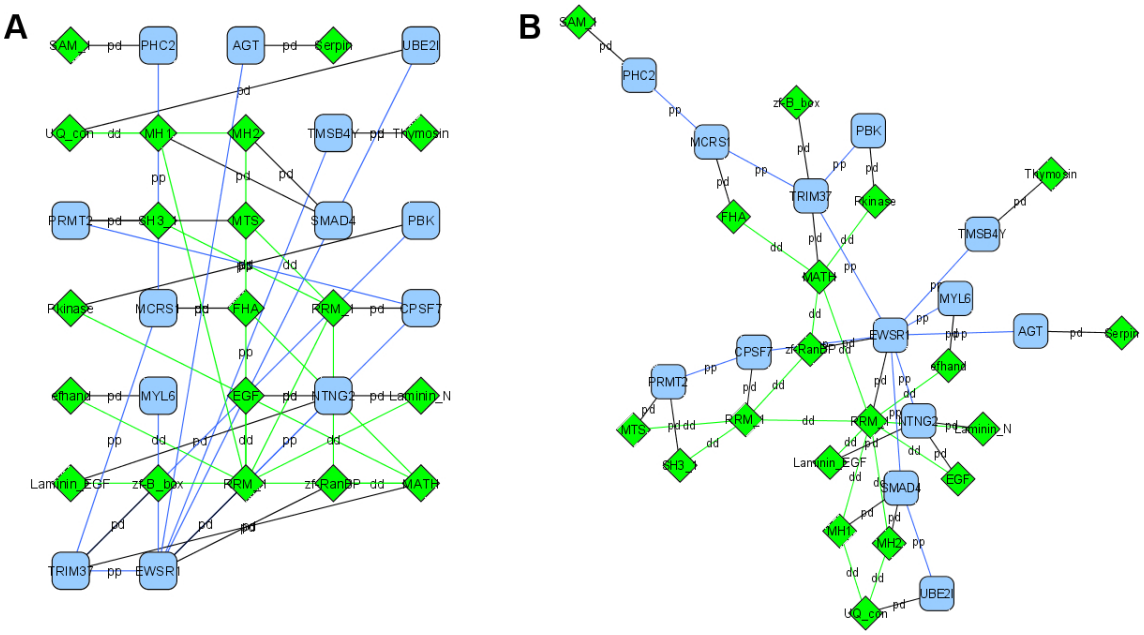


Figure 6.13: *Protein and domain interaction network created by DomainGraph.* ***(A)*** *shows a grid layout representation and* ***(B)*** *a force-directed layout of the same interaction network. Protein nodes are visualized as blue squares and domains nodes as green diamonds. Blue edges represent protein interactions, green domain interactions, and black protein-domain linker.*

could help the biologist to visually investigate and evaluate their experimental data.

In the case of a protein and domain interaction network, the layout should allow for a clear visual distinction between protein and domain nodes and the corresponding distinct interaction types. Although DomainGraph provides a graphical representation for the protein and domain nodes that separates them by their visual appearance, the application of generic network layout algorithms as described may make it difficult to quickly identify the topology of the protein and domain interactions. We therefore developed a layout method that takes into account the object semantics, i.e. the node and edge types, and emphasizes the corresponding topological characteristics.

An additional problem for the identification of protein and domain interaction partners and of structural interaction patterns are edge crossings and also overlapping nodes. A good layout thus has to provide a clear visual separation of the nodes without occlusion, a minimization of node-edge and edge-edge crossings, a maximization of the angular resolution, and it should allow for easy identification of the different node and edge types. In addition, protein nodes should be placed close to the domains they comprise.

6.5.2 Layout Design

Some of the layout aesthetics and optimization goals described in Chapter 6.5.1 compete with each other and therefore cannot be met simultaneously. As a compromise, we decided to implement an algorithm that combines elements of circular, radial, and layered layout algorithms, which we call *RadialLayout*. The layout has been designed in collaboration with Karsten Klein from the Technical University in Dortmund (TU Dortmund). Anne Kunert (TU Dortmund) implemented the layout algorithm in the course of her diploma thesis, which was conducted at the Max Planck Institute.

In order to allow for a clear separation of protein and domain nodes, they are placed on separate concentric circles around a common center. The domain nodes are placed on the inner circle such that the domain interactions in the focus of the user's interest are located in the center of the drawing and proteins are placed on the outer circle as close as possible to their constituent domains (see *Figure 6.7* and *Figure 6.8*). The use of separate circles for the two node types allows for the clear distinction between proteins and domains as well as between the interaction and linkage edge types in the drawing. A further separation of the domain nodes is achieved by highlighting the occurrences of alternative splicing events and grouping the domain nodes on the domain circle with regard to their node type, i.e. affected by alternative splicing or not (*Figure 6.7*). This facilitates visually assessing the influence of alternative splicing events on the functioning of a certain protein and its interactions.

In order to reveal the interaction topology, the protein and domain nodes need to be positioned in a way that minimizes the number of edge crossings since they complicate the identification of corresponding connected nodes. This problem is similar to the crossing reduction in the bilayer crossing minimization, which has an important application in hierarchical graph drawing. The order of the nodes on each of the two circles has to be fixed such that the number of crossings is minimized. The bilayer edge crossing minimization problem is NP-hard, but it can be solved to optimality for small instances in acceptable time, and heuristics have been developed that also perform very well on larger instances (Jünger and Mutzel (1997)). For DomainGraph, we decided to implement a heuristic in order to be able to layout larger instances in reasonable time as well.

The basic idea of our approach is to start with an initial permutation of the nodes on both circles. Then the order on one circle is fixed and the algorithm tries to find an order on the second circle with a decreased number of crossings. In our implementation, this is achieved by using an adaption of the so-called barycenter heuristic (Sugiyama et al. (1981)). This heuristic places each node at the barycenter of its neighbors on the second

circle. The crossing reduction is iteratively performed with alternating roles of the two circles until the number of crossings cannot be reduced any further. The computation of the number of edge crossings in each step can be a bottleneck with regard to the running time (Waddle and Malhotra (1999)). Therefore, we use an implementation of the bilayer cross counting method. This approach proved to be very fast in experiments (Barth et al. (2002)) and has $O(|E| \, log \, |V_{small}|)$ asymptotic running time where V_{small} is the smaller set of protein and domain nodes. In case expression data are integrated, the set of domain nodes is additionally split into subsets according to their node types, and the crossing reduction step starts with a permutation that reflects the grouping of the domains into domain types. Our implementation of the barycenter heuristic is then applied to arrange the nodes in each subtype group, minimizing the crossings in the drawing.

The inner circle area is reserved for domain-interaction edges, while protein-interaction edges and protein-domain linkers that cross domain nodes on the inner circle or in the inner circle area are routed as curved splines around the protein circle to further reduce crossings and to facilitate the visual recognition of the domain interactions.

In summary, the main requirements met by RadialLayout are the visual separation of protein and domain nodes, the focus on domain interactions and the impact of alternative splicing, the avoidance of object occlusion, and the minimization of edge crossings in the drawing.

6.6 Implementation Details

DomainGraph is written in Java and runs on Windows, Unix, and Mac OS. The program is designed as a plugin for the free, open-source network visualization software Cytoscape (Shannon et al. (2003)). For users wishing to perform a comparative analysis, DomainGraph can be installed and run either consecutively with or separately from AltAnalyze on the user's computer. The downloadable AltAnalyze package includes both Cytoscape and DomainGraph, enabling users to run the complete software workflow without separate installation of the programs. Users can thus immediately continue analyzing potential functional implications after the statistical analysis has finished. For users who prefer to run AltAnalyze and DomainGraph separately, DomainGraph is included in the Cytoscape Plugin Manager and can be downloaded and installed directly from within Cytoscape. In case of a single-array analysis, a statistical pre-processing program such as APT needs to be downloaded in addition to DomainGraph.

Embedded Database

DomainGraph relies on a locally installed database based on annotation files provided by Affymetrix and the corresponding builds of the Ensembl database (Hubbard et al. (2009)). The database is updated whenever new annotation files are made available by Affymetrix, and can be downloaded from within the program. Currently, there are three different database versions available and users can up- and downgrade their version whenever necessary. The databases contain all necessary gene and protein data for the analysis and visualization of Affymetrix Exon Array data. The data are stored in flat files and are hosted at the Max Planck Institute for Informatics. Upon user requests, the files are automatically downloaded and installed locally on the user's computer. These files are then processed by DomainGraph and read into an embedded Apache Derby database (http://db.apache.org/derby), which is employed by the program. Users may also import their Exon Array data into this database, such that the data can readily be used any time.

Mapping Affymetrix Probesets to Exons and Domains

We obtained the probeset genome coordinates from the appropriate NetAffx releases (Cheng et al. (2004). These genome coordinates were first mapped to the genome coordinates of the exons as provided by Ensembl. Then, the relative positions of the domains within the coding sequences of the transcripts were computed according to the Pfam domain coordinates given by Ensembl.

Mapping Genes and Proteins to Pathways

The embedded DomainGraph database contains a mapping between Ensembl gene identifiers and WikiPathways as well as a mapping between UniProt accessions and Reactome pathways. Identifier mappings as provided by BioMart are used to uniformly map WikiPathway annotations to Ensembl genes, while UniProt accessions for Reactome pathways are readily annotated. WikiPathways are available for human, mouse, and rat and can be loaded via the DomainGraph *table view*. DomainGraph makes use of the Cytoscape plugin *GPML-Plugin* for loading WikiPathways data. Reactome provides stable pathway information for human, which can be loaded via the *table view*. For mouse and rat, only predicted results are available, for which no stable links are contained in Reactome and thus loading these pathways is not possible. Here, Reactome annotations are provided as an overview only.

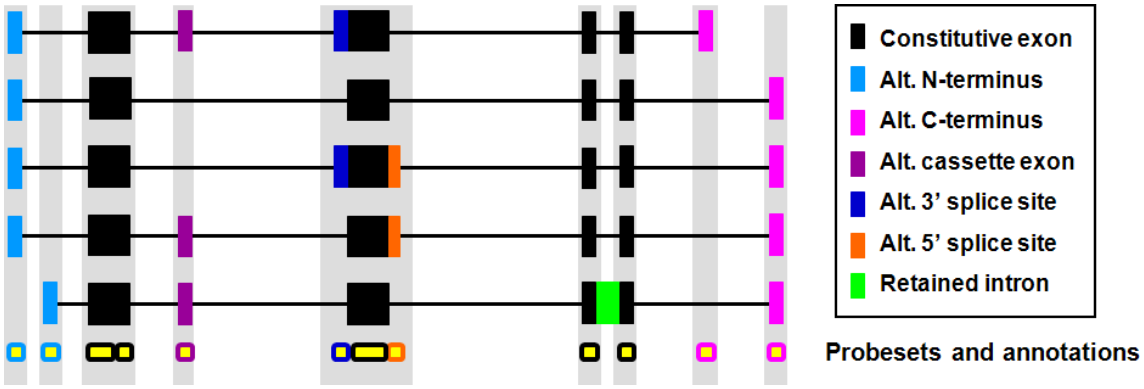


Figure 6.14: *Assignment of alternative splicing annotations to probesets. The figure depicts five hypothetical transcripts of a sample gene. Alternative splicing annotations are shown by the coloring scheme of the exons. Exon Array probesets are displayed below the transcripts and are assigned to the exons according to their genomic positions. The probesets are framed using the color of the corresponding exon regions and assigned the respective exon annotations.*

Node Coloring in Pathway and Network View

Pathway or interaction network nodes that are annotated with gene or protein identifiers found in the DomainGraph database are highlighted if they are associated with differentially expressed probesets. Since probeset associations are available for Ensembl identifiers only, node identifiers different from Ensembl identifiers are first mapped to Ensembl via the mapping contained in the embedded DomainGraph database.

Probeset View and Annotations

The *probeset view* is available only for Ensembl identifiers since probeset-to-exon mappings are provided only for Ensembl. Therefore, if network annotations are given with UniProt or Entrez Gene identifiers, they are mapped to the appropriate Ensembl identifiers in the *probeset view* and the Ensembl product is displayed. Tooltips provide additional information, e.g., the Splicing Index, expression value, DABG p-value, alternative splicing annotations, and cross-hybridization types of probesets.

Alternative splicing annotations are assigned to probesets to reflect whether a probeset matches a constitutive or alternative exon region (*Figure 6.14*). The annotations allow users to distinguish between alternative splicing events that are known to occur and those that are not. This information can be an indicator for the reliability of an alternative

splicing event based on a specific up- or down-regulated probeset, especially if the p-value computed by AltAnalyze is close to the significance threshold.

The cross-hybridization types of the probesets are taken from NetAffx and include the following three types: 'unique' (all probes in the probeset perfectly match exactly one genomic region), 'similar' (all probes in the probeset perfectly match more than one sequence), and 'mixed' (some probes in the probeset perfectly or partially match more than one sequence). The *probeset view* is enabled by selecting a gene in the *table view* (when starting with AltAnalyze probeset statistics) or by double-clicking on a gene or protein node (when starting with a gene or protein interaction network).

6.7 Conclusions

The DomainGraph plugin is a powerful tool for the visual analysis of genes, proteins, interactions, and pathways, which is amenable even to biologists with limited knowledge and experience of bioinformatics and programming. The program is well received by the community and to date, DomainGraph has more than 600 registered users and over 2,000 downloads via the Cytoscape website. DomainGraph provides a convenient and straightforward way of identifying and studying alternative splicing events occurring in mammalian genomes at large scale. The software supports comparative analyses to identify similarities and differences between different biological groups and enables single-array analyses to distinguish present and absent probesets in the same biological condition. For both the comparative and the single-array analysis, the gain or loss of protein and domain interactions due to alternative splicing within domain-coding regions is highlighted in interaction networks and can easily be identified by the user. Visualizing proteins and domains together with the exon structures and probeset annotations illustrates the precise sequence positions of alternative splicing events. Additional information like GO and OMIM annotations for genes, proteins, and domains may direct the user to the impact of alternative splicing on cellular processes and to protein functions disturbed in some disease.

Two application cases for the comparative analysis have been demonstrated. In the first analysis, we studied the effects of a knockdown of the splicing factor PTB in mouse and presented examples of alternatively regulated exons and their potential biological impact. In the second analysis we compared different developmental stages of human cells (human embryonic stem cells and cardiac precursors) to identify alternative gene products that are responsible for the differentiation of stem cells into cardiac progenitor cells.

The development and application of the RadialLayout algorithm has shown to be a step forward from conventional layout algorithms. It is specifically designed for the visualization of protein and domain networks that have exon expression data integrated and allows for a straightforward identification of alternative splicing events. Minimization of the number of edge crossings aims at providing the best possible overview of the network, its topology, and interactions. In the future, DomainGraph may be extended to incorporate data by other high-throughput exon and splicing detection methods, for example, deep-sequencing techniques.

Chapter 7

Conclusions

In the following, the projects and findings achieved in the course of this thesis are summarized. Furthermore, possible extensions of the performed projects and directions for future research in the field of dynamic interaction networks are described.

7.1 Summarizing Remarks

In this thesis, we focused on different biological aspects that lead to the context-dependent formation of protein interaction networks. While many studies have been performed in the last years aiming at the discovery of complete proteomes and interactomes, they only represent a static picture of the molecular networks. The ultimate goal of proteome and interactome research, however, is the understanding of the dynamics in cellular processes. A first step towards this goal is the analysis of context-dependency in molecular networks such as temporal and spatial properties.

In the first project detailed in Chapter 3, we focused on the structural proteome and interactome as stored in the PDB. We investigated the common assumption that multi-interface proteins can always interact with multiple interaction partners at the same time. Our main goal was the identification of collisions in three-dimensional space that prevents an otherwise feasible simultaneous interaction. Due to the very limited number of structurally solved protein interactions, our analysis yielded few biological protein interactions with collisions, which were mainly caused by mutated protein chains. However, with a growing number of solved structures, protein interactions with integrated three-dimensional information will likely lead to new insights.

Next, we focused on the tissue-specific co-occurrence of proteins based on the corresponding gene expression, and the effects on the formation of protein interactions. In order to achieve this goal, we first needed to identify the technology with the highest measurement accuracy for the identification of co-occurring proteins. As described in

Chapter 4, we conducted a study comparing the gene expression measurements of a well-established, yet low-density, microarray to those of a more recently developed high-density microarray and to those produced by a novel next-generation sequencing technology. Here, we demonstrated that next-generation sequencing is able to detect gene expression even at very low levels, while microarray techniques are less sensitive and the statistical methods cannot reliably distinguish between low expression and noise. The findings of this quantitative analysis support the wide-spread belief that next-generation sequencing is the most suitable technology for studying the functional implications of tissue-specific gene expression. As described in Chapter 5, we therefore used next-generation sequencing data of 15 human tissues and cell lines to study all currently known protein interactions, protein domains, and protein complexes in more detail. Our analysis revealed that only few protein interactions are tissue-specific in the tissues and cell lines we investigated, while the majority of protein interactions occur universally. Our functional analysis of the tissue-specific protein interactions showed that they are mainly involved in receptor and transporter activities. Regarding the protein domains, we found only few of them to be tissue-specific and those are preferentially involved in receptor, transporter, and DNA-binding activities. Furthermore, we analyzed protein complexes and confirmed that the assembly of protein complexes is often controlled by one or few co-complexed proteins, which supports the concept of core and attachment proteins. Overall, our results indicate that the number of tissue-specific interactions, domains, and complexes is limited and that they mostly participate in well-defined biological functions such as regulatory processes. These findings are, however, limited by the currently available data, and future studies that include a larger number of diverse tissues may likely identify additional biological functions that occur in specific tissues only.

Finally, in Chapter 6 we addressed the current lack of software tools that are targeted at the analysis of existing protein variants produced by a single gene and at a better understanding of how these variants relate to different interaction partners and biological functions. While a number of programs exist for the statistical analysis of alternative splicing, none of them includes downstream analyses to assess the biological effects. To overcome these limitations, we developed the software DomainGraph for the analysis of protein and domain interaction networks in the context of alternative splicing. DomainGraph is a user-friendly software tool that guides the user through the analysis via graphical user interfaces and is thus even amenable to users with limited knowledge of programming and statistical methods. We described user options to identify genes, proteins, domains, and miRNA binding sites affected by alternative splicing together with their graphical representation. Additionally, we demonstrated the analysis of protein and domain interaction

networks and the detection of potentially disrupted interactions due to alternative splicing. We presented two sample application cases for DomainGraph: a comparison of human embryonic stem cells and cardiac progenitor cells, and a comparison of a mouse splicing factor knockdown dataset and untreated cells. In these case studies, we identified protein isoforms and pathways that are potentially disturbed due to alternative splicing. Such findings can serve as a starting point for wet-lab biologists to analyze the implications of protein isoforms in detail.

In summary, we aimed at the integration of molecular networks with temporal and spacial aspects to provide a better understanding of the biological machineries of the cells. Focusing on several such aspects, we demonstrated that the consideration of context-dependent properties provides a much deeper knowledge of molecular networks, which is an important step towards the ultimate goal of comprehending the dynamics of cellular processes.

7.2 Perspectives

Currently available biological databases such as IntAct, BioGRID, and HPRD store tens of thousands of human protein-protein interactions. Yet, the currently known interactome is still far from complete (Venkatesan et al. (2009)) and the databases usually do not provide any additional information necessary to relate the interactions to a biological context. For instance, information on spatial and temporal aspects such as the simultaneous binding of proteins, the tissue specificity of a protein interaction, and the actual protein isoforms that were used for the experimental detection is often neglected.

The next-generation RNA-sequencing dataset that we used in Chapter 5 to identify tissue-specific elements and functions contained gene expression data for only 15 human tissues and cell lines. In the next years, however, as high-throughput sequencing will become affordable for more researchers, we expect an increase in the number of sequenced tissues. As a result, tissue-specific gene and protein expression will be known for a large number of tissues and information on the tissue-specific occurrence of protein interactions can be included in the available databases. High-throughput sequencing data can also be utilized to investigate expression differences between different tissues or different biological conditions. For instance, protein interactions with distinct expression patterns for healthy and diseased cells can be identified and may serve as biomarkers or drug targets. In addition, protein interactions that are only active at a specific time point, such as in a specific developmental stage of a cell, can be detected and functionally characterized. Therefore, our relatively small-scale analyses presented in Chapter 5 can be repeated

at large scale to provide an accurate picture of the dynamic processes in different cells. Furthermore, next-generation sequencing data may also be utilized to identify tissue-specific functional motifs of proteins such as short linear motifs and microRNA binding sites, which may provide new insights into the tissue-specific proteomes and interactomes.

Apart from detecting gene expression, the advent of high-throughput RNA-sequencing also enables identifying novel protein variants produced by alternative splicing. As has been demonstrated in several extensive studies (Wang et al. (2008); Pan et al. (2008)), alternative splicing frequently occurs in higher eukaryotes and many protein isoforms still remain to be uncovered. With the sequencing of more tissues, next-generation sequencing will ultimately allow for the construction of an alternative splicing map. Such a map will incorporate information on the mechanisms and factors that are involved in the alternative splicing process and responsible for the maturation of alternatively spliced transcripts. Moreover, this map will include the occurrences and abundance levels of functional protein isoforms in different tissues and conditions, and their functions in biological processes.

A first step towards the construction of such an alternative splicing map includes the better understanding of the functional implications of alternative splicing, which can be addressed by extending the DomainGraph software. As described in Chapter 6, DomainGraph has been designed for downstream analyses of alternative splicing based on microarray data. Therefore, the software needs to be made compatible with high-throughput RNA-sequencing data in order to allow researchers to combine biological analyses with these novel data. One possible way is the adaptation of DomainGraph to processing files given in wiggle (WIG) format, a file format that has been developed for continuous data such as transcriptome data. A WIG file contains an expression value for each genomic position based on the number of sequencing reads assigned to the respective position. Thus, the expression values can be automatically mapped to all known Ensembl exons and transcripts, enabling the inference of their expression levels. Since this software extension still restricts the alternative splicing analysis to known Ensembl transcripts, the outcome is similar, yet more accurate, to the one using microarray data. To overcome this limitation, a more advanced version of DomainGraph has to support the identification and visualization of novel transcripts based on next-generation sequencing data. The re-construction of novel transcripts from sequencing reads can be achieved by combining the position-specific expression levels with information on all putative splice junctions to infer new transcript sequence assemblies.

A further step towards an alternative splicing map is the analysis of alternative splicing events in functional regions of protein isoforms, such as in protein domains, short linear motifs, and miRNA binding sites. Previous studies have dealt with the comparison of

protein isoforms regarding their domain compositions. These studies primarily focused on the identification of protein domains that can be included or excluded from protein isoforms due to premature stop codons introduced by alternative splicing events (Resch et al. (2004); Loraine et al. (2003)). In addition to alternative splicing, premature stop codons can result from single nucleotide polymorphisms (SNPs) occurring in the coding region of a gene. Analyzing the relationship between SNPs, alternative splicing, and the gain or loss of functional subunits in the protein variants is therefore important for a better understanding of the protein functions.

Apart from the loss of functional protein regions, we observed that alternative splicing events frequently do not alter the inclusion or exclusion of a complete domains but affect only small parts of a domain region (see *Figure 6.3* for example). Since the alternative sequence region is often small compared to the whole protein domain, functional annotations such as SCOP and Pfam domains usually remain unchanged. Such domain annotations can easily lead to wrong conclusions regarding the precise functions of these isoforms since they suggest the same functions for all isoforms containing the domain. In reality, however, the alternative splicing event may alter important functional residues, such as phosphorylation sites or interface residues, that are responsible for the formation of protein-protein interactions.

Studying such alterations of interface residues may help to answer the question how certain proteins can bind to a multitude of other proteins. Previous studies mainly focused on hub proteins, i.e., proteins with an exceptionally high number of protein interaction partners, and their functional properties (Han and Ju (2003); Zotenko et al. (2008); Tsai et al. (2009)). However, none of them investigated protein isoforms with small sequence modifications in the domain-coding regions. These variations can have a strong influence on the formation of protein interactions and alternative splicing may be an efficient way to increase the interaction repertoire of proteins.

In addition to alternative splicing, alterations of interface residues can also occur as a result of non-synonymous SNPs. With the launching of next-generation sequencing technologies, the identification of SNPs occurring in coding regions of the genome has been simplified using exome sequencing. Exome sequencing is a type of DNA-sequencing that restricts the sequencing to the exonic regions of the genome, which corresponds to as little as 1% of the human genome. While exome sequencing is usually applied in genome-wide association studies to identify disease-related SNPs, the analyses can also be extended to the identification of SNPs that may alter the binding interface of a protein interaction.

In conclusion, future research will need to integrate aspects describing the biological context, such as gene expression, protein isoforms, protein structures, and SNPs, into

protein interaction studies to understand the functional and topological characteristics. Incorporating such information will ultimately lead to understanding not only the static properties of molecular networks but also the dynamics of cellular processes.

Bibliography

Affymetrix. Affymetrix Power Tools, 2010a. http://www.affymetrix.com/partners_progs/programs/developer/tools/powertools.affx.

Affymetrix. Exon background correction, 2010b. http://media.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf.

Affymetrix. Expression Console, 2010c. http://www.affymetrix.com/products_ services/software/specific/expression_console_software.affx.

Affymetrix. Array design for the GeneChip human genome U133 set, 2010d. http://media.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf.

Affymetrix. Statistical algorithms description document, 2010e. http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

R. Albert, H. Jeong, and A. L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22:1600–1607, 2006.

P. Aloy and R. B. Russell. Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, 7:188–197, 2006.

P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. B. Russell. Structure-based assembly of protein complexes in yeast. *Science*, 303:2026–2029, 2004.

P. P. Amaral, M. E. Dinger, T. R. Mercer, and J. S. Mattick. The eukaryotic genome as an RNA machine. *Science*, 319:1787–1789, 2008.

R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, D. Barrell, B. Bely, M. Bingley, D. Binns, L. Bower, P. Browne, W. M. Chan,

E. Dimmer, R. Eberhardt, A. Fedotov, R. Foulger, J. Garavelli, R. Huntley, J. Jacobsen, M. Kleen, K. Laiho, R. Leinonen, D. Legge, Q. Lin, W. Liu, J. Luo, S. Orchard, S. Patient, D. Poggioli, M. Pruess, M. Corbett, G. di Martino, M. Donnelly, P. van Rensburg, A. Bairoch, L. Bougueleret, I. Xenarios, S. Altairac, A. Auchincloss, G. Argoud-Puy, K. Axelsen, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, L. Bollondi, E. Boutet, S. B. Quintaje, L. Breuza, A. Bridge, E. deCastro, L. Ciapina, D. Coral, E. Coudert, I. Cusin, G. Delbard, M. Doche, D. Dornevil, P. D. Roggli, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, S. Gehant, N. Farriol-Mathis, S. Ferro, E. Gasteiger, A. Gateau, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hulo, J. James, S. Jimenez, F. Jungo, T. Kappler, G. Keller, C. Lachaize, L. Lane-Guermonprez, P. Langendijk-Genevaux, V. Lara, P. Lemercier, D. Lieberherr, T. de Oliveira Lima, V. Mangold, X. Martin, P. Masson, M. Moinat, A. Morgat, A. Mottaz, S. Paesano, I. Pedruzzi, S. Pilbout, V. Pillet, S. Poux, M. Pozzato, N. Redaschi, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, E. Stanley, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, L. Yip, L. Zuletta, C. Wu, C. Arighi, L. Arminski, W. Barker, C. Chen, Y. Chen, Z. Z. Hu, H. Huang, R. Mazumder, P. McGarvey, D. A. Natale, J. Nchoutmboube, N. Petrova, N. Subramanian, B. E. Suzek, U. Ugochukwu, S. Vasudevan, C. R. Vinayaka, L. S. Yeh, and J. Zhang. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38: D142–148, 2010.

R. Aragues, A. Sali, J. Bonet, M. A. Marti-Renom, and B. Oliva. Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput. Biol.*, 3:1761–1771, 2007.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, 2000.

Wilhelm Barth, Michael Jünger, and Petra Mutzel. Simple and efficient bilayer cross counting. In *GD '02: Revised Papers from the 10th International Symposium on Graph Drawing*, pages 130–141. Springer-Verlag, 2002. ISBN 3-540-00158-1.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

R. P. Bhattacharyya, A. Remenyi, B. J. Yeh, and W. A. Lim. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu. Rev. Biochem.*, 75:655–680, 2006.

P. Bjorkholm and E. L. Sonnhammer. Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics*, 25:3020–3025, 2009.

H. Blankenburg, R. D. Finn, A. Prlic, A. M. Jenkinson, F. Ramírez, D. Emig, S. E. Schelhorn, J. Büch, T. Lengauer, and M. Albrecht. DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25:1321–1328, 2009.

B. J. Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126:37–47, 2006.

N. Blom, S. Gammeltoft, and S. Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, 294:1351–1362, 1999.

L. Bonetta. Protein-protein interactions: Interactome under construction. *Nature*, 468: 851–854, 2010.

A. Bossi and B. Lehner. Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, 5:260, 2009.

A. H. Brivanlou and J. E. Darnell. Signal transduction and the control of gene expression. *Science*, 295:813–818, 2002.

A. Campagna, L. Serrano, and C. Kiel. Shaping dots and lines: adding modularity into protein interaction networks using structural information. *FEBS Lett.*, 582:1231–1236, 2008.

A. F. Chambers. MDA-MB-435 and M14 cell lines: identical but not M14 melanoma? *Cancer Res.*, 69:5292–5293, 2009.

T. Y. Chang, Y. Y. Li, C. H. Jen, T. P. Yang, C. H. Lin, M. T. Hsu, and H. W. Wang. easyExon–a Java-based GUI tool for processing and visualization of Affymetrix exon array data. *BMC Bioinformatics*, 9:432, 2008.

X. W. Chen and M. Liu. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21:4394–4400, 2005.

J. Cheng, S. Sun, A. Tracy, E. Hubbell, J. Morris, V. Valmeekam, A. Kimbrough, M. S. Cline, G. Liu, R. Shigeta, D. Kulp, and M. A. Siani-Rose. NetAffx Gene Ontology

Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, 20:1462–1463, 2004.

C. Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 357: 543–544, 1992.

C. Chothia and J. Gough. Genomic and structural aspects of protein evolution. *Biochem. J.*, 419:15–28, 2009.

T. Clackson, M. H. Ultsch, J. A. Wells, and A. M. de Vos. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.*, 277:1111–1128, 1998.

T. A. Clark, A. C. Schweitzer, T. X. Chen, M. K. Staples, G. Lu, H. Wang, A. Williams, and J. E. Blume. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, 8:R64, 2007.

J. S. Dasen, J. P. Barbera, T. S. Herman, S. O. Connell, L. Olson, B. Ju, J. Tollkuhn, S. H. Baek, D. W. Rose, and M. G. Rosenfeld. Temporal regulation of a paired-like homeodomain repressor/TLE corepressor complex and a related activator is required for pituitary organogenesis. *Genes Dev.*, 15:3193–3207, 2001.

U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307:724–727, 2005.

C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, 1:349–356, 2002.

D. Devos and R. B. Russell. A more complete, complexed and structured interactome. *Curr. Opin. Struct. Biol.*, 17:370–377, 2007.

A. M. Duursma, M. Kedde, M. Schrier, C. le Sage, and R. Agami. miR-148 targets human DNMT3b protein coding region. *RNA*, 14:872–877, 2008.

H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6:197–208, 2005.

E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10: 48, 2009.

E. Eisenberg and E. Y. Levanon. Human housekeeping genes are compact. *Trends Genet.*, 19:362–365, 2003.

D. Emig and M. Albrecht. Tissue-specific proteins and functional implications. *In press, Journal of Proteome Research*, 2011.

D. Emig, M. S. Cline, K. Klein, A. Kunert, P. Mutzel, T. Lengauer, and M. Albrecht. Integrative visual analysis of the effects of alternative splicing on protein domain interaction networks. *J Integr Bioinform*, 5, 2008a.

D. Emig, M. S. Cline, T. Lengauer, and M. Albrecht. Integrating expression data with domain interaction networks. *Bioinformatics*, 24:2546–2548, 2008b.

D. Emig, K. Klein, A. Kunert, P. Mutzel, and M. Albrecht. Visualizing domain interaction networks and the impact of alternative splicing events. In *Proceedings of the 5th International Conference on BioMedical Visualization, Information Visualization in Medical and Biomedical Informatics, MediVis 2008*, pages 36–43, 2008c.

D. Emig, T. Kacprowski, and M. Albrecht. Measuring and analyzing tissue specificity of human genes and protein complexes. In *Proceedings of the 7th International Workshop on Computational Systems Biology (WCSB)*, pages 27–30, 2010a.

D. Emig, N. Salomonis, J. Baumbach, T. Lengauer, B. R. Conklin, and M. Albrecht. AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.*, 38 (Web Server issue):W755–762, 2010b.

D. Emig, O. Sander, G. Mayr, and M. Albrecht. Structure collisions between interacting proteins. *In revision, PLoS ONE*, 2011.

M. Fagnani, Y. Barash, J. Y. Ip, C. Misquitta, Q. Pan, A. L. Saltzman, O. Shai, L. Lee, A. Rozenhek, N. Mohammad, S. Willaime-Morawek, T. Babak, W. Zhang, T. R. Hughes, D. van der Kooy, B. J. Frey, and B. J. Blencowe. Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol.*, 8:R108, 2007.

D. Farre, N. Bellora, L. Mularoni, X. Messeguer, and M. M. Alba. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.*, 8:R140, 2007.

R. D. Finn, M. Marshall, and A. Bateman. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21:410–412, 2005.

R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res.*, 38:D211–222, 2010.

C. Friedrich and F. Schreiber. Visualisation and navigation methods for typed protein-protein interaction networks. *Appl. Bioinformatics*, 2(3 Suppl):19–24, 2003.

P. J. Gardina, T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams, and Y. Turpaz. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7:325, 2006.

A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.

N. M. Goodey and S. J. Benkovic. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.*, 4:474–482, 2008.

K. S. Guimaraes, R. Jothi, E. Zotenko, and T. M. Przytycka. Predicting domain-domain interactions using a parsimony approach. *Genome Biol.*, 7:R104, 2006.

J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

K. Han and B.H. Ju. A fast layout algorithm for protein interaction networks. *Bioinformatics*, 19(15):1882–1888, 2003.

H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, 32:D452–455, 2004.

E. Ho, R. Webber, and M. R. Wilkins. Interactive three-dimensional visualization and contextual analysis of protein interaction networks. *Journal of Proteome Research*, 7 (1):104–112, 2008.

B. R. Howard, F. F. Vajdos, S. Li, W. I. Sundquist, and C. P. Hill. Structural insights into the catalytic mechanism of cyclophilin A. *Nat. Struct. Biol.*, 10:475–481, 2003.

T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. Ensembl 2009. *Nucleic Acids Res.*, 37:D690–697, 2009.

E. L. Humphris and T. Kortemme. Design of multi-specificity in protein interfaces. *PLoS Comput. Biol.*, 3:e164, 2007.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

Z. Itzhaki, E. Akiva, and H. Margalit. Preferential use of protein domain pairs as interaction mediators: order and transitivity. *Bioinformatics*, 26:2564–2570, 2010.

H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

L. Jin and S. C. Harrison. Crystal structure of human calcineurin complexed with cyclosporin A and human cyclophilin. *Proc. Natl. Acad. Sci. U.S.A.*, 99:13522–13526, 2002.

R. Jothi, P. F. Cherukuri, A. Tasneem, and T. M. Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J. Mol. Biol.*, 362:861–875, 2006.

M. Jünger and P. Mutzel. 2-layer straightline crossing minimization: Performance of exact and heuristic algorithms. *J. Graph Algorithms and Applications*, 1(1):1–24, 1997.

M. Kanehisa. The KEGG database. *Novartis Found. Symp.*, 247:91–101, 2002.

M. Kato, M. Nagasaki, A. Doi, and S. Miyano. Automatic drawing of biological networks using cross cost and subbcomponent data. *Genome Informatics*, 16(2):22–31, 2005.

S. K. Kearsley. On the orthogonal transformation used for structural comparisons. *Acta Cryst*, Chapter 2:208–210, 1989.

O. Keskin and R. Nussinov. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, 15:341–354, 2007.

O. Keskin, A. Gursoy, B. Ma, and R. Nussinov. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.*, 108:1225–1244, 2008.

C. Kiel, P. Beltrao, and L. Serrano. Analyzing protein interaction networks using structural information. *Annu. Rev. Biochem.*, 77:415–441, 2008.

P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314:1938–1941, 2006.

P. M. Kim, A. Sboner, Y. Xia, and M. Gerstein. The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.*, 4:179, 2008.

K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 105:20870–20875, 2008.

H. Lee, M. Deng, F. Sun, and T. Chen. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, 7:269, 2006.

J. R. Leeman and T. D. Gilmore. Alternative splicing in the NF-kappaB signaling pathway. *Gene*, 423:97–107, 2008.

J. P. Lees-Miller and D. M. Helfman. The molecular basis for tropomyosin isoform diversity. *Bioessays*, 13:429–437, 1991.

B. Lehner and A. G. Fraser. Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet.*, 20:468–472, 2004.

E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, 2:e155, 2006.

W. Li and H. Kurata. A grid layout algorithm for automatic drawing of biochemical networks. *Bioinformatics*, 21(9):2036–2042, 2005. ISSN 1367-4803.

J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules I and II. *Proteins: Struct. Funct. Genet.*, 33:1–17,18–29, 1998.

B. Liu, J. Liao, X. Rao, S. A. Kushner, C. D. Chung, D. D. Chang, and K. Shuai. Inhibition of Stat1-mediated gene activation by PIAS1. *Proc. Natl. Acad. Sci. U.S.A.*, 95:10626–10631, 1998.

Y. Liu, N. Liu, and H. Zhao. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21:3279–3285, 2005.

H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipurksy, and Darnell J. *Molecular Cell Biology.* WH Freeman and Company, 2004.

A. E. Loraine, G. A. Helt, M. S. Cline, and M. A. Siani-Rose. Exploring alternative transcript structure in the human genome using blocks and InterPro. *J Bioinform Comput Biol*, 1:289–306, 2003.

T. Maier, M. Guell, and L. Serrano. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.*, 583:3966–3973, 2009.

A. C. Martin. Mapping PDB chains to UniProtKB entries. *Bioinformatics*, 21:4297–4301, 2005.

A. J. Matlin, F. Clark, and C. W. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6:386–398, 2005.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.

L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37: D619–622, 2009.

C. Mayr and D. P. Bartel. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138:673–684, 2009.

N. J. McGlincy and C. W. Smith. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.*, 33:385–393, 2008.

S. Millevoi and S. Vagner. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res.*, 38:2757–2774, 2010.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5:621–628, 2008.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

S. K. Ng, Z. Zhang, S. H. Tan, and K. Lin. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, 31:251–254, 2003.

I. M. Nooren and J. M. Thornton. Diversity of protein-protein interactions. *EMBO J.*, 22:3486–3492, 2003a.

I. M. Nooren and J. M. Thornton. Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.*, 325:991–1018, 2003b.

K. Numata, R. Yoshida, M. Nagasaki, A. Saito, S. Imoto, and S. Miyano. ExonMiner: Web service for analysis of GeneChip Exon array data. *BMC Bioinformatics*, 9:494, 2008.

M. J. Okoniewski and C. J. Miller. Comprehensive analysis of affymetrix exon arrays using BioConductor. *PLoS Comput. Biol.*, 4:e6, 2008.

J. P. Orengo and T. A. Cooper. Alternative splicing in disease. *Adv. Exp. Med. Biol.*, 623: 212–223, 2007.

P. Pagel, M. Oesterheld, O. Tovstukhina, N. Strack, V. Stumpflen, and D. Frishman. DIMA 2.0–predicted and known domain interactions. *Nucleic Acids Res.*, 36:D651–655, 2008.

Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40:1413–1415, 2008.

J. M. Percival, J. A. Hughes, D. L. Brown, G. Schevzov, K. Heimann, B. Vrhovski, N. Bryce, J. L. Stow, and P. W. Gunning. Targeting of a tropomyosin isoform to short microfilaments associated with the Golgi complex. *Mol. Biol. Cell*, 15:268–280, 2004.

S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker,

N. Padma, V. Niranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. Shivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, 32:497–501, 2004.

A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: pathway editing for the people. *PLoS Biol.*, 6:e184, 2008.

A. Porollo and J. Meller. Prediction-based fingerprints of protein-protein interactions. *Proteins*, 66:630–645, 2007.

G. Prehna and C. E. Stebbins. A Rac1-GDP trimer complex binds zinc with tetrahedral and octahedral coordination, displacing magnesium. *Acta Crystallogr. D Biol. Crystallogr.*, 63:628–635, 2007.

P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, and T. J. Gibson. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, 31:3625–3630, 2003.

E. Purdom, K. M. Simpson, M. D. Robinson, J. G. Conboy, A. V. Lapuk, and T. P. Speed. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, 24:1707–1714, 2008.

B. Raghavachari, A. Tasneem, T. M. Przytycka, and R. Jothi. DOMINE: a database of protein domain interactions. *Nucleic Acids Res.*, 36:D656–661, 2008.

F. Ramírez and M. Albrecht. Finding scaffold proteins in interactomes. *Trends Cell Biol.*, 20:2–4, 2010.

F. Ramírez, A. Schlicker, Y. Assenov, T. Lengauer, and M. Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7:2541–2552, 2007.

D. Ramskold, E. T. Wang, C. B. Burge, and R. Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, 5:e1000598, 2009.

K. Reim, H. Wegmeyer, J. H. Brandstatter, M. Xue, C. Rosenmund, T. Dresbach, K. Hofmann, and N. Brose. Structurally and functionally unique complexins at retinal ribbon synapses. *J. Cell Biol.*, 169:669–680, 2005.

A. Resch, Y. Xing, B. Modrek, M. Gorlick, R. Riley, and C. Lee. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome Res.*, 3:76–83, 2004.

R. Riley, C. Lee, C. Sabatti, and D. Eisenberg. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, 6:R89, 2005.

L. D. Rogers and L. J. Foster. Phosphoproteomics–finally fulfilling the promise? *Mol Biosyst*, 5:1122–1129, 2009.

J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H. W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.*, 38:497–501, 2010.

N. Salomonis, B. Nelson, K. Vranizan, A. R. Pico, K. Hanspers, A. Kuchinsky, L. Ta, M. Mercola, and B. R. Conklin. Alternative splicing in the differentiation of human embryonic stem cells into cardiac precursors. *PLoS Comput. Biol.*, 5:e1000553, 2009.

M. Sammeth, S. Foissac, and R. Guigo. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, 4:e1000147, 2008.

M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38:305–320, 1996.

S. E. Schelhorn, T. Lengauer, and M. Albrecht. An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24:35–41, 2008.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, 2003.

X. She, C. A. Rohl, J. C. Castle, A. V. Kulkarni, J. M. Johnson, and R. Chen. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics*, 10:269, 2009.

D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk. BioMart–biological queries made easy. *BMC Genomics*, 10:22, 2009.

E. L. Sonnhammer and D. Kahn. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, 3:482–492, 1994.

K. Srinivasan, L. Shiue, J. D. Hayes, R. Centers, S. Fitzwater, R. Loewen, L. R. Edmondson, J. Bryant, M. Smith, C. Rommelfanger, V. Welch, T. A. Clark, C. W. Sugnet, K. J. Howe, Y. Mandel-Gutfreund, and M. Ares. Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, 37:345–359, 2005.

S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. A. Thanaraj, and H. Soreq. Function of alternative splicing. *Gene*, 344:1–20, 2005.

C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34:D535–539, 2006.

A. Stein and P. Aloy. Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures. *PLoS Comput. Biol.*, 6:e1000789, 2010.

A. Stein, R. B. Russell, and P. Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, 33:D413–417, 2005.

A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.*, 101:6062–6067, 2004.

K. Sugiyama, S. Tagawa, and M. Toda. Methods for visual understanding of hierarchical systems. *IEEE Transact. Syst. Man. Cybern*, SMC-11(2):109–125, 1981.

M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M. L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321:956–960, 2008.

M. Sundstrom, T. Lundqvist, J. Rodin, L. B. Giebel, D. Milligan, and G. Norstedt. Crystal structure of an antagonist mutant of human growth hormone, G120R, in complex with its receptor at 2.9 A resolution. *J. Biol. Chem.*, 271:32197–32203, 1996.

F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6:377–382, 2009.

C. Tarricone, B. Xiao, N. Justin, P. A. Walker, K. Rittinger, S. J. Gamblin, and S. J. Smerdon. The structural basis of Arfaptin-mediated cross-talk between Rac and Arf signalling pathways. *Nature*, 411:215–219, 2001.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

C. J. Tsai, B. Ma, and R. Nussinov. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem. Sci.*, 34:594–600, 2009.

Z. Tu, L. Wang, M. Xu, X. Zhou, T. Chen, and F. Sun. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7:31, 2006.

N. Tuncbag, G. Kar, A. Gursoy, O. Keskin, and R. Nussinov. Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. *Mol Biosyst*, 5:1770–1778, 2009.

G. Vasmatzis, E. W. Klee, D. M. Kube, T. M. Therneau, and F. Kosari. Quantitating tissue specificity of human genes to facilitate biomarker discovery. *Bioinformatics*, 23: 1348–1355, 2007.

K. Venkatesan, J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi, and M. Vidal. An empirical framework for binary interactome mapping. *Nat. Methods*, 6:83–90, 2009.

C. Vogel, R. d. e. S. Abreu, D. Ko, S. Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva. Sequence signatures and mRNA

concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.*, 6:400, 2010.

Vance E. Waddle and Ashok Malhotra. An e log e line crossing algorithm for levelled graphs. In *Proceedings of the 7th International Symposium on Graph Drawing*, pages 59–71. Springer-Verlag, 1999.

E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, 2008.

R. S. Wang, Y. Wang, L. Y. Wu, X. S. Zhang, and L. Chen. Analysis on multi-domain cooperation for predicting protein-protein interactions. *BMC Bioinformatics*, 8:391, 2007.

E. E. Winter, L. Goodstadt, and C. P. Ponting. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.*, 14:54–61, 2004.

J. Wysocka, M. P. Myers, C. D. Laherty, R. N. Eisenman, and W. Herr. Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev.*, 17:896–911, 2003.

I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Res.*, 28:289–291, 2000.

Y. Xing, P. Stoilov, K. Kapur, A. Han, H. Jiang, S. Shen, D. L. Black, and W. H. Wong. MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA*, 14:1470–1479, 2008.

T. Yates, M. J. Okoniewski, and C. J. Miller. X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res.*, 36:D780–786, 2008.

M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi, and M. Vidal. Drug-target network. *Nat. Biotechnol.*, 25:1119–1126, 2007.

J. Zhu, F. He, S. Song, J. Wang, and J. Yu. How many human genes can be defined as housekeeping with current expression data? *BMC Genomics*, 9:172, 2008.

E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.*, 4:e1000140, 2008.

Appendix A

List of the Tissue Specificity of Diseases and Protein Complexes

Table A.1: *OMIM diseases and the average number of tissues for the respective disease genes. Cancer disease genes show a high average number of tissues both for the P/A and the Peak definition. Blue indicates diseases with tissue-specific genes for both P/A and Peak definition. Green represents diseases with tissue-specific genes according to Peak definition only.*

Omim	Avg. tissues P/A	Avg. tissues Peak
Hypocalcemia	0	0
Hypomagnesemia	0	0
Intrinsic factor deficiency	0	0
Low renin hypertension	0	0
Adrenocortical insufficiency	1	1
Alcohol dependence	1	1
Aldosteronism	1	1
Allergic rhinitis	1	1
Antithrombin III deficiency	1	1
Attention-deficit hyperactivity disorder	1	1
Autoimmune thyroid disease	1	1
Chronic infections, due to opsonin defect	1	1
Conjunctivitis, ligneous	1	1
Goiter	1	1
Graves disease	1	1
Meningococcal disease	1	1
Nephrotic syndrome	1	1
Neuroblastoma	1	1
Obsessive-compulsive disorder	1	1

Ovarian hyperstimulation syndrome	1	1
Precocious puberty, male	1	1
Aplastic anemia	1.5	0.5
Diabetes insipidus	1.5	1.5
Hypothyroidism	2	2
Narcolepsy	2	2
Alcoholism	3	3
Aromatase deficiency	3	3
MODY	3.4	3.4
Megaloblastic anemia	3.5	3.5
COPD, rate of decline of lung function in	4	4
Cystic fibrosis	4	4
Keratitis	4	1
Pancreatitis	4	2.5
Anterior segment anomalies and cataract	4.33	4.33
Systemic lupus erythematosus	4.5	4.5
Atrial fibrillation	4.6	2.8
Cataract	4.62	4.38
Hypoglycemia	5	1
Non-Hodgkin lymphoma	5	5
Pneumonitis, desquamative interstitial	5	5
Wilms tumor	5	5
AIDS	5.67	5.67
Asthma	5.8	5.8
Convulsions	6	6
Hypogonadism, hypergonadotropic	6	6
Kaposi sarcoma	6	6
Multiple sclerosis	6	6
Phenylketonuria	6	5.33
Venous thrombosis	6	6
Immunodeficiency	6.17	5
Epidermolytic hyperkeratosis	6.33	1.67
Hemophilia	6.5	6.5
Ataxia	6.75	5.75
Leprosy	7	7
Obesity	7	4.75
Osteoarthritis	7	7
Rheumatoid arthritis	7.29	7.29
Autoimmune disease	7.5	7.5

HIV	7.5	1.5
Tuberculosis	7.5	7.5
Myasthenic syndrome	7.57	4
Epilepsy	7.75	7
Ichthyosis	7.86	4.71
Adrenal hyperplasia, congenital	8	8
Angioedema	8	8
Apnea, postanesthetic	8	8
Cramps, potassium-aggravated	8	1
Infantile spasm syndrome	8	8
Meniere disease	8	8
Ossific. of the post. longitudinal spinal ligaments	8	8
Rapp-Hodgkin syndrome	8	1
Thrombocytopenia	8	4.75
Diabetes mellitus	8.42	6
Hypophosphatemia	8.5	8.5
Platelet defect/deficiency	8.5	8.5
Rhabdomyosarcoma	8.5	8.5
Hypercholesterolemia	8.63	7
Coronary artery disease	8.67	8.67
Dystonia	8.67	7.33
Malaria	8.67	8.67
Factor x deficiency	8.88	4.63
Fibrosis	9	9
Glucocorticoid deficiency	9	1
Lupus erythematosus	9	9
Macular degeneration	9	7
Migraine	9	5
Tyrosinemia	9	6
Schizophrenia	9.17	9.17
Glioblastoma	9.2	6.4
Myocardial infarction	9.25	4.88
Neutropenia	9.25	4.25
Insulin resistance	9.33	9.33
Ventricular tachycardia	9.33	5.67
Cardiomyopathy	9.45	4.82
Cystinuria	9.5	9.5
Gastrointestinal stromal tumor	9.5	9.5
Thrombocythemia	9.5	9.5

Hyperparathyroidism	10	10
Multiple myeloma	10	10
Myelodysplastic syndrome	10	10
Myelomonocytic leukemia, chronic	10	10
Myeloproliferative disorder	10	10
Rett syndrome	10	10
Stroke	10	1
von Willebrand disease	10	1
Parkinson disease	10.45	9.55
Anemia	10.5	10.5
Crohn disease	10.5	10.5
Osteolysis	10.5	10.5
Sarcoidosis	10.5	10.5
Adenomas	10.67	10.67
Myopathy	10.78	7.44
Hypertension	10.82	7.55
Atherosclerosis	11	11
Basal cell carcinoma	11	9.33
Fluorouracil toxicity, sensitivity to	11	11
Insomnia	11	11
Myelogenous leukemia	11.25	11.25
Paget disease	11.33	11.33
HMG-CoA deficiency	11.5	8
Rickets	11.5	11.5
Viral infection	11.5	5
Breast cancer	11.83	11.06
Lymphoma	11.86	11.86
Leukemia	11.88	11.44
Chronic granulomatous disease	12	8
Encephalopathy	12	8
Hypoparathyroidism	12	1
MASS syndrome	12	12
Neuropathy	12	12
Osteoporosis	12	12
Periodontitis	12	1
Dementia	12.2	11.4
Prostate cancer	12.4	12.4
Glaucoma	12.5	9
Iron overload/deficiency	12.5	8

Acyl-CoA dehydrogenase, deficiency of	12.67	8
Alzheimer disease	12.83	9.67
Adenocarcinoma	13	6
Adenoma, periampullary	13	13
Angiotensin I-converting enzyme	13	1
Generalized epilepsy	13	1
Mycobacterial infection	13	9.5
Protein S deficiency	13	13
Amyotrophic lateral sclerosis	13.25	9.75
Colon cancer	13.27	12.8
Pancreatic cancer	13.33	12.11
Squamous cell carcinoma	13.5	13.5
Thyroid carcinoma	13.64	13.64
Intrauterine and postnatal growth retardation	14	14
Major depressive disorder	14	14
Mesothelioma	14	14
Melanoma	14.2	12
Ovarian cancer	14.25	12.5
Cirrhosis	14.33	5.67
Esophageal cancer	14.5	14.5
Polycythemia	14.5	14.5
Renal cell carcinoma	14.67	14.67
Mucopolysaccharidosis	14.86	13
Acromegaly	15	15
Adenosine deaminase deficiency	15	15
B-cell non-Hodgkin lymphoma, high-grade	15	15
Bipolar disorder	15	15
Bladder cancer	15	15
Carnitine deficiency	15	15
CHARGE syndrome	15	15
Darier disease	15	15
Down syndrome	15	15
Emphysema	15	1
Ewing sarcoma	15	15
Fabry disease	15	15
Gaucher disease	15	15
HARP syndrome	15	15
Hemolytic anemia	15	15
Hepatic failure, early onset, and neurol. disorder	15	15

Hyperlipidemia	15	15
Hyperlipoproteinemia	15	1
Lead poisoning	15	15
Lung cancer	15	15
Multiple malignancy syndrome	15	15
Nephropathy	15	15
Osteopetrosis	15	15
Osteosarcoma	15	15
Pheochromocytoma	15	15
Pituitary tumor, invasive	15	15
Polyposis	15	15
Pulmonary hypertension, familial primary	15	15
Retinoblastoma	15	15
Stomach cancer	15	15
Wilson disease	15	15

Table A.2: *The 139 most tissue-specific complexes with their tissue occurrence according to P/A and Peak definition. '—' corresponds to 0 tissues. Complexes colored in blue are detected using both the P/A and the Peak definition.*

Complex Name	Tissues (P/A)	Tissues (Peak)
NCOR-SIN3-HDAC-HESX1 complex	—	—
ITGA5-ITGB1-CAL4A3 complex	—	—
IL12A-IL12B-IL12RB2 complex	—	—
TDT-TDIF2-core-histone complex	—	—
ESCRT-I complex	—	—
IL12A-IL12B-IL12RB1 complex	—	—
cytokine receptor complex	—	—
IL12B-IL12RB1-IL12RB2 complex	—	—
TSG101-VPS37B-VPS28 complex	—	—
TRPC1-TRPC3-TRPC7 complex	—	—
Human Follicle Stimulating Hormone - Receptor complex	—	—
SNARE complex (VAMP2, SNAP25, STX1a, STX3, CPLX1, CPLX3, CPLX4)	cerebellum	cerebellum
MPP4-MPP5-CRB1 complex	cerebellum	cerebellum
Thrombin - central "E" region of fibrin complex	liver	liver
CD20-LCK-FYN-p75/80 complex	lymph node	lymph node
IFNB1-IFNAR1-IFNAR2- complex	MCF7	MCF7
Insulin-like growth factors - IGF binding proteins complex	liver	liver
CRB1-MPP5-INADL complex	cerebellum	cerebellum
CD20-LCK-LYN-FYN-p75/80 complex	lymph node	lymph node
SNARE complex (VAMP2, SNAP25, STX1a, CPLX3, CPLX4)	cerebellum	cerebellum
ULBP1-KLRK1-HCST complex	T47D	—
PCI-PSA-SCG2 complex	colon	—
SLP-76-PLC-gamma-1-ITK complex, alpha-TCR stimulated	lymph node	lymph node
ITGAV-ITGB5-SPP1 complex	breast, heart	—
LSD1-CoREST selectivity in histone H3 recognition	testis, MCF7	testis, MCF7
PICK1-GRIP1-GLUR2 complex	brain, testis	brain, testis
VILIP-1-AChR-alpha-4-AChR-beta-2 complex	brain, cerebellum	brain, cerebellum
Human fibrinogen	liver, skeletal muscle	liver
TRPC1-STIM1-ORAI1 complex	>2	brain
GINS complex	>2	T47D
ITGA6-ITGB4-Laminin10/12 complex	>2	HME
ITGA5-ITGB1-SPP1 complex	>2	brain
ITGA3-ITGB1-CD63 complex	>2	HME

Complexin complex (STX3, CPLX1, CPLX3)	>2	cerebellum
ITGA3-ITGB1-BSG complex	>2	HME
ITGA6-ITGB4-CD9 complex	>2	HME
GPR56-CD81-Galphaq/11-Gbeta complex	>2	MB435
ITGA5-ITGB1-FN-1-NOV complex	>2	MB435
SMCC complex	>2	BT474
PLC-gamma-1-SLP-76-SOS1-LAT complex	>2	lymph node
PCNA complex	>2	HME
BRD4 complex	>2	BT474
ITGA6-ITGB1-CYR61 complex	>2	HME
ITGA6-ITGB4-SHC1-GRB2 complex	>2	HME
HD-RAB8A-OPTN complex	>2	skeletal muscle
MKK4-ARRB2-JNK3 complex	>2	brain
ITGA4-ITGB1-CD53 complex	>2	lymph node
Profilin 2 complex	>2	brain
MMP-9-TIMP-1-LRP complex	>2	breast
ITGAV-ITGB3-EGFR complex	>2	—
P-TEFb-BRD4-TRAP220 complex	>2	BT474
SRC-3 complex	>2	MCF7
ITGAV-ITGB3-SPP1 complex	>2	brain
VEGF transcriptional complex	>2	testes
20S proteasome	>2	BT474
RNA polymerase II complex, chromatin structure modifying	>2	testes
LAT-PLC-gamma-1-p85-GRB2-SOS signaling complex, C305 activated	>2	lymph node
PLC-gamma-2-Lyn-FcR-gamma complex	>2	brain
46kDa domain of human cardiac troponin in the Ca2+ saturated form	>2	heart
ERG-JUN-FOS DNA-protein complex	>2	breast
G protein complex (CACNA1A, GNB1, GNG2)	>2	cerebellum
ITGA9-ITGB1-TNC complex	>2	MB435
Urokinase receptor, urokinase and vitronectin complex	>2	—
SNARE complex (SNAP25, VAMP3, VAMP2, NAPB, STX13)	>2	—
ITGA3-ITGB1-CD151 complex	>2	HME
ULBP3-KLRK1-HCST complex	>2	—
SNARE complex (HGS, SNAP25, STX13)	>2	cerebellum
Sarcoglycan-sarcospan complex SG-SPN	>2	skeletal muscle
SNARE complex (VAMP2, SNAP25, STX1a, CPLX1, CPLX3)	>2	cerebellum
ITGA2-ITGB1-COL6A3 complex	>2	—
PA28-20S proteasome	>2	BT474
ITGAM-ITGB2-CD11 complex	>2	lymph node
Ubiquitin E3 ligase (FBXW7, CUL1, SKP1A, RBX1)	>2	brain

ARF-Mule complex	>2	MB435
ITGA11-ITGB1-COL1A1 complex	>2	breast
Cell cycle kinase complex CDC2	>2	HME
FIB-associated protein complex	>2	—
ITGA1-ITGB1-COL6A3 complex	>2	adipose
FHL2-p53-HIPK2 complex	>2	heart
Cell cycle kinase complex CDK5	>2	HME
ITGA6-ITGB1-CD151 complex	>2	HME
Notch1-p56lck-PI3K complex	>2	—
Human MCAD:ETF E165betaA complex	>2	heart
Gamma-secretase complex (APH1B, PSEN2, PSENEN, NCSTN)	>2	testes
TRAP complex	>2	BT474
NFAT-JUN-FOS DNA-protein complex	>2	breast
Sam68-p85 P13K-IRS-1-IR signaling complex	>2	MCF7
ITGAV-ITGB3-NOV complex	>2	MB435
Polycystin-1-E-cadherin-beta-catenin-Flotillin-2 complex	>2	—
ETS2-FOS-JUN complex	>2	breast
NDC80 kinetochore complex	>2	MB435
Gamma-secretase complex (APH1B, PSEN1, PSENEN, NCSTN)	>2	testes
ITGA9-ITGB1-SPP1 complex	>2	brain
ACTR-p300-PCAF complex	>2	MCF7
RHOA-IP3R-TRPC1 complex	>2	cerebellum
Cell cycle kinase complex CDK2	>2	HME
PLC-gamma-1-LAT-c-CBL complex, OKT3 stimulated	>2	lymph node
CD19-Vav-PI 3-kinase (p85 subunit) complex	>2	lymph node
CoREST-HDAC complex	>2	MCF7
SNARE complex (VAMP2, SNAP25, STX1a, CPLX1)	>2	cerebellum
SMAD3-SMAD4-cJun-cFos complex	>2	breast
ITGB1-RAP1A-PKD1 complex	>2	cerebellum
ITGB5-ITGAV-VTN complex	>2	liver
ITGA7-ITGB1-ITGB1BP3 complex	>2	heart
ULBP2-KLRK1-HCST complex	>2	—
PLC-gamma-2-Syk-LAT-FcR-gamma complex	>2	—
SNARE complex (STX11, VAMP2, SNAP23)	>2	adipose
SNARE complex (VAMP2, SNAP25, STX1a, CPLX2)	>2	cerebellum
LLGL2-PAR-6B-PRKCI complex	>2	MCF7
60S APC containing complex	>2	brain
PA28gamma-20S proteasome	>2	BT474
Polycystin-1-E-cadherin-beta-catenin complex	>2	—
YY1-Notch1-RBP-Jkappa complex	>2	HME
TRAP complex	>2	BT474
RSmad complex	>2	MCF7

MAML1-RBP-Jkappa-Notch1 complex	>2	HME
SNX complex (SNX1a, SNX2, SNX4, EGFR)	>2	HME
ITGB3-ITGAV-VTN complex	>2	liver
AMY-1-S-AKAP84-RII-beta complex	>2	breast
TF-FVIIa-FXa-TFPI complex	>2	—
FOXO1-FHL2-SIRT1 complex	>2	heart
WIP-WASp-actin-myosin-IIa complex	>2	lymph node
TRAP-SMCC mediator complex	>2	BT474
JUND-FOSB-SMAD3-SMAD4 complex	>2	breast
SMCC complex	>2	BT474
ITGA6-ITGB4-LAMA5 complex	>2	HME
Cell cycle kinase complex CDK4	>2	HME
LCK-SLP76-PLC-gamma-1-LAT complex, pervanadate-activated	>2	lymph node
ITGA6-ITGB4-FYN complex	>2	HME
SNARE complex (VAMP2, SNAP25, STX13)	>2	cerebellum
SNX complex (SNX1a, SNX2, SNX4, TFRC)	>2	T47D
ITGA3-ITGB1-THBS1 complex	>2	HME
EGFR-CBL-GRB2 complex	>2	HME
ITGA5-ITGB3-COL6A3 complex	>2	adipose
ITGAV-ITGB5-SPP1 complex	>2	brain
ITGA6-ITGB4-CD151 complex	>2	HME
Interferon-stimulated gene factor 3 transcription complex ISGF3	>2	MCF7
LAT-PLC-gamma-1-p85-GRB2-CBL-VAV-SLP-76 signaling complex, C305 activated	>2	lymph node
ITGAV-ITGB1-SPP1 complex	>2	brain

Appendix B

List of Own Publications

1. **Dorothea Emig**, Oliver Sander, Gabriele Mayr and Mario Albrecht.
   Structure collisions between interacting proteins.
   PLoS ONE, in revision, 2011.
2. **Dorothea Emig** and Mario Albrecht.
   Tissue-specific proteins and functional implications.
   Journal of Proteome Research, in press, 2011.
3. Tobias Wittkop, **Dorothea Emig**, Anke Truss, Mario Albrecht, Sebastian Boecker and Jan Baumbach.
   Comprehensive cluster analysis with Transitivity Clustering.
   Nature Protocols, in press, 2011.
4. Tobias Wittkop, **Dorothea Emig**, Sita J. Lange, Sven Rahmann, Mario Albrecht, John H. Morris, Sebastian Boecker, Jens Stoye and Jan Baumbach.
   Partitioning biological data with transitivity clustering.
   Nature Methods 7(6):419-420, 2010.
5. **Dorothea Emig**, Nathan Salomonis, Jan Baumbach, Thomas Lengauer, Bruce R. Conklin and Mario Albrecht.
   AltAnalyze and DomainGraph: analyzing and visualizing exon expression data.
   Nucleic Acids Research, 38(Web Server issue):W755-W762, 2010.
6. **Dorothea Emig**, Tim Kacprowski and Mario Albrecht.
   Measuring and analyzing tissue specificity of human genes and protein complexes.
   Proceedings of the 7th International Workshop on Computational Systems Biology (WCSB), 51:27-30, 2010.
7. Hagen Blankenburg, Robert D. Finn, Andreas Prlic, Andrew M. Jenkinson, Fidel Ramírez, **Dorothea Emig**, Sven-Eric Schelhorn, Joachim Büch, Thomas Lengauer and Mario Albrecht.

DASMI: exchanging, annotating and assessing molecular interaction data.
Bioinformatics, 25(10):1321-1328, 2009.

8. **Dorothea Emig**, Melissa S. Cline, Thomas Lengauer and Mario Albrecht.
   Integrating expression data with domain interaction networks.
   Bioinformatics, 24(21):2546-2548, 2008.
9. **Dorothea Emig**, Melissa S. Cline, Karsten Klein, Anne Kunert, Petra Mutzel, Thomas Lengauer and Mario Albrecht.
   Integrative visual analysis of the effects of alternative splicing on protein domain interaction networks.
   Journal of Integrative Bioinformatics, 5(2):101.1-15, 2008.
10. **Dorothea Emig**, Karsten Klein, Anne Kunert, Petra Mutzel and Mario Albrecht.
    Visualizing domain interaction networks and the impact of alternative splicing events.
    Proceedings of the 5th International Conference on BioMedical Visualization, Information Visualization in Medical and Biomedical Informatics, MediVis 2008, IEEE Society Press, pp. 36-46, 2008.