Cross-Species Network and Transcript Transfer

Robert Pesch

München 2015

Cross-Species Network and Transcript Transfer

Robert Pesch

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Robert Pesch
geboren in Bonn

München, den 27.10.2015

Erstgutachter: Prof. Dr. Ralf Zimmer
Zweitgutachter: Prof. Dr. Dmitrij Frishman
Tag der mündlichen Prüfung: 05.02.2016

Contents

List of Figures

List of Tables

Summary

Metabolic processes, signal transduction, gene regulation, as well as gene and protein expression are largely controlled by biological networks. High-throughput experiments allow the measurement of a wide range of cellular states and interactions. However, networks are often not known in detail for specific biological systems and conditions. Gene and protein annotations are often transferred from model organisms to the species of interest. Therefore, the question arises whether biological networks can be transferred between species or whether they are specific for individual contexts. In this thesis, the following aspects are investigated: (i) the conservation and (ii) the cross-species transfer of eukaryotic protein-interaction and gene regulatory (transcription factor- target) networks, as well as (iii) the conservation of alternatively spliced variants.

In the simplest case, interactions can be transferred between species, based solely on the sequence similarity of the orthologous genes. However, such a transfer often results either in the transfer of only a few interactions (medium/high sequence similarity threshold) or in the transfer of many speculative interactions (low sequence similarity threshold). Thus, advanced network transfer approaches also consider the annotations of orthologous genes involved in the interaction transfer, as well as features derived from the network structure, in order to enable a reliable interaction transfer, even between phylogenetically very distant species. In this work, such an approach for the transfer of protein interactions is presented (**COIN**). COIN uses a sophisticated machine-learning model in order to label transferred interactions as either correctly transferred (conserved) or as incorrectly transferred (not conserved).

The comparison and the cross-species transfer of regulatory networks is more difficult than the transfer of protein interaction networks, as a huge fraction of the known regulations is only described in the (not machine-readable) scientific literature. In addition, compared to protein interactions, only a few conserved regulations are known, and regulatory elements appear to be strongly context-specific. In this work, the cross-species analysis of regulatory interaction networks is enabled with software tools and databases for global (**ConReg**) and thousands of context-specific (**CroCo**) regulatory interactions that are derived and integrated from the scientific literature, binding site predictions and experimental data.

Genes and their protein products are the main players in biological networks. However, to date, the aspect is neglected that a gene can encode different proteins. These alternative proteins can differ strongly from each other with respect to their molecular structure, function and their role in networks. The identification of conserved and species-specific splice variants and the integration of variants in network models will allow a more complete cross-species transfer and comparison of biological networks. With **ISAR** we support the cross-species transfer and comparison of alternative variants by introducing a gene-structure aware (i.e. exon-intron structure aware) multiple sequence alignment approach for variants from orthologous and paralogous genes.

The methods presented here and the appropriate databases allow the cross-species transfer of biological networks, the comparison of thousands of context-specific networks, and the cross-species comparison of alternatively spliced variants. Thus, they can be used as a starting point for the understanding of regulatory and signaling mechanisms in many biological systems.

Zusammenfassung

In biologischen Systemen werden Stoffwechselprozesse, Signalübertragungen sowie die Regulation von Gen- und Proteinexpression maßgeblich durch biologische Netzwerke gesteuert. Hochdurchsatz-Experimente ermöglichen die Messung einer Vielzahl von zellulären Zuständen und Wechselwirkungen. Allerdings sind für die meisten Systeme und Kontexte biologische Netzwerke nach wie vor unbekannt. Gen- und Proteinannotationen werden häufig von Modellorganismen übernommen. Demnach stellt sich die Frage, ob auch biologische Netzwerke und damit die systemischen Eigenschaften ähnlich sind und übertragen werden können. In dieser Arbeit wird: (i) Die Konservierung und (ii) die artenübergreifende Übertragung von eukaryotischen Protein-Interaktions- und regulatorischen (Transkriptionsfaktor-Zielgen) Netzwerken, sowie (iii) die Konservierung von Spleißvarianten untersucht.

Interaktionen können im einfachsten Fall nur auf Basis der Sequenzähnlichkeit zwischen orthologen Genen übertragen werden. Allerdings führt eine solche Übertragung oft dazu, dass nur sehr wenige Interaktionen übertragen werden können (hoher bis mittlerer Sequenzschwellwert) oder dass ein Großteil der übertragenden Interaktionen sehr spekulativ ist (niedriger Sequenzschwellwert). Verbesserte Methoden berücksichtigen deswegen zusätzlich noch die Annotationen der Orthologen, Eigenschaften der Interaktionspartner sowie die Netzwerkstruktur und können somit auch Interaktionen auf phylogenetisch weit entfernte Arten (zuverlässig) übertragen. In dieser Arbeit wird ein solcher Ansatz für die Übertragung von Protein-Interaktionen vorgestellt (**COIN**). COIN verwendet Verfahren des maschinellen Lernens, um Interaktionen als richtig (konserviert) oder als falsch übertragend (nicht konserviert) zu klassifizieren.

Der Vergleich und die artenübergreifende Übertragung von regulatorischen Interaktionen ist im Vergleich zu Protein-Interaktionen schwieriger, da ein Großteil der bekannten Regulationen nur in der (nicht maschinenlesbaren) wissenschaftlichen Literatur beschrieben ist. Zudem sind im Vergleich zu Protein-Interaktionen nur wenige konservierte Regulationen bekannt und regulatorische Elemente scheinen stark kontextabhängig zu sein. In dieser Arbeit wird die artenübergreifende Analyse von regulatorischen Netzwerken mit Softwarewerkzeugen und Datenbanken für globale (**ConReg**) und kontextspezifische (**CroCo**) regulatorische Interaktionen ermöglicht. Regulationen wurden dafür aus Vorhersagen, experimentellen Daten und aus der wissenschaftlichen Literatur abgeleitet und integriert.

Grundbaustein für viele biologische Netzwerke sind Gene und deren Proteinprodukte. Bisherige Netzwerkmodelle vernachlässigen allerdings meist den Aspekt, dass ein Gen verschiedene Proteine kodieren kann, die sich von der Funktion, der Proteinstruktur und der

Rolle in Netzwerken stark voneinander unterscheiden können. Die Identifizierung von konservierten und artspezifischen Proteinprodukten und deren Integration in Netzwerkmodelle würde einen vollständigeren Übertrag und Vergleich von Netzwerken ermöglichen. In dieser Arbeit wird der artenübergreifende Vergleich von Proteinprodukten mit einem multiplen Sequenzalignmentverfahren für alternative Varianten von paralogen und orthologen Genen unterstützt, unter Berücksichtigung der bekannten Exon-Intron-Grenzen (**ISAR**).

Die in dieser Arbeit vorgestellten Verfahren, Datenbanken und Softwarewerkzeuge ermöglichen die Übertragung von biologischen Netzwerken, den Vergleich von tausenden kontextspezifischen Netzwerken und den artenübergreifenden Vergleich von alternativen Varianten. Sie können damit die Ausgangsbasis für ein Verständnis von Kommunikations- und Regulationsmechanismen in vielen biologischen Systemen bilden.

Chapter 1

Introduction

In biological systems, genes, proteins, enzymes, and compounds influence and interact with each other in complex networks (Barabási and Oltvai, 2004). Such networks can be modeled (Karlebach and Shamir, 2008), visualized (Pavlopoulos et al., 2008), and compared (Sharan and Ideker, 2006) using appropriate approaches. Advanced network models like Petri-nets also allow for precise mathematical modeling and the simulation of biological systems (see e.g. Reddy et al. (1993); Küffner et al. (2000); Koch et al. (2005)). Networks (on a small scale) are intuitive representations of complex systems. They have been successfully used for the prediction of protein function, the study of regulatory dynamics, and the explanation of experimental data (Mitra et al., 2013). Thus, networks are commonly used in systems biology to serve as frameworks for data integration and interpretation.

High-throughput techniques allow the measuring of a wide range of cellular states and interactions. Protein interactions can, for example, be measured using Yeast-Two-Hybrid (Fields and Song, 1989) and Co-ImmunoPrecipitation systems (Co-IP) (Kaboord and Perr, 2008). Various high-throughput Next Generation Sequencing (NGS) techniques like RNA sequencing (RNA-seq) and Chromatin ImmunoPrecipitation sequencing (ChIP-seq) allow measuring the expression of transcripts and the bindings of proteins to the DNA on a genome wide level (Furey, 2012). Projects like ENCODE (ENCODE Project Consortium, 2012b), modENCODE (Celniker et al., 2009), the TCGA Gene Cancer Atlas (Cancer Genome Atlas Research Network, 2008), and Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015) apply such NGS techniques on thousands of samples, under diverse biological conditions, and provide resources of regulatory data for human, fly, and worm. Researchers are now able to combine this data and model biological systems for various species and conditions in detail.

Networks are an abstract representation of (high-throughput) experimental data. A regulatory and protein-interaction network can be represented in different levels of detail. A simplified regulatory network includes genes and directed (binary) edges that represent the regulatory effects of genes, whereas a simplified protein interaction network model consists of proteins and edges representing bindings between proteins. Such representations can be constructed using experimental-type specific workflows (for network definitions from ChIP, protein interaction and digital genomic footprinting data see e.g. Pollack and Iyer (2002);

Kim et al. (2005); Berggård et al. (2007); Neph et al. (2012a)). Further, more advanced network models also include the interaction effect (e.g. activation or repression), context information (e.g. in which tissue/cell-line an interaction occurs), transitions (e.g. which proteins form a protein complex), and the involved gene products. The gene products can differ due to the processes of alternative splicing, alternative transcription, and post-translational modifications (Kelemen et al., 2013; Pal et al., 2011; Khoury et al., 2011). Alternatively spliced variants may exclude, include or replace certain amino-acids and thus can affect protein interfaces and DNA-binding domains (Resch et al., 2004; Kelemen et al., 2013). Compared to another gene product (often the wild-type defined as the gene product with strongest expression), this can result in different binding-affinities and the loss and gain of interaction partners (Ellis et al., 2012; Buljan et al., 2012, 2013). Furthermore, so called non-trivial spliced isoform (Birzele et al., 2008), isoforms which differ in essential parts from the native structure, may even result in distinct protein structures which expose different residues to the protein surface.

Global/static network modeling allows a first and simple representation of networks. Such networks can, for example, be derived from one standardized laboratory condition or be computationally predicted using binding site predictions. A first step towards the analysis of dynamic changes in such networks offers the integration of context-specific data, e.g. gene expression data from different cell-lines, and tissues (Ideker and Krogan, 2012). Thereby, a given global network can be filtered in order to identify the 'active' elements (for example, edges involving not expressed genes can be removed). Thousands of context-specific data sets for many diverse experimental settings are publicly available in repositories like SRA (Leinonen et al., 2011), GEO (Edgar et al., 2002), and ArrayExpress (Kolesnikov et al., 2015), which can be utilized to construct, model, and compare various context-specific biological networks. Thus, differential and context-specific network analysis is now becoming a prevalent tool, as it enables the identification of new interactions, complexes, and pathways which would be obscured in a global network (Ideker and Krogan, 2012). This is of high interest as it is now understood that gene regulation and expression is highly context-specific (Thurman et al., 2012; Gerstein et al., 2012; Neph et al., 2012b,a). Furthermore, different network types can also be integrated into a combined (global, or context-specific) network, thereby, providing a more complete view on cellular dynamics and allowing the analysis, interpretation, and predictions of various aspects (see e.g. Hwang et al. (2005); Chen and Rajewsky (2007); Pesch et al. (2008); Warde-Farley et al. (2010)).

Many terabyte of context-specific experimental data has been generated to measure various aspects of biological systems. But experimental measurements are often labor intensive, expensive, and, sometimes, not feasible for a certain system due to certain technical problems and ethical conflicts. Therefore, model organisms are frequently used to study cellular systems (Fields and Johnston, 2005). Only for these organisms sufficient data is available to model meaningful (sub)-systems. Furthermore, there are many aspects that have only been selectively studied (even for model organisms). For example, only for the (comparable) simple unicellular bakers' yeast nearly the complete protein interactome is measured (Stumpf et al., 2008). For other (model) organisms such as human and fly, the measured binary protein interaction networks have by far not reached their estimated sizes (Hart et al., 2006). A common practice in bioinformatics is the transfer of data between (closely) related species (Bork et al., 1998; Matthews et al., 2001; Yu et al., 2004). This practice is based on the observation that genes that stem from the same common ancestor (orthologous genes) often possess a similar function, even though, at great evolutionary distance, there are cases where the function of orthologous genes differs (Koonin, 2005). For almost all human genes a strongly conserved gene in mouse can be identified (Waterston et al., 2002). Thus, a transfer of functional descriptions and Gene Ontology (Gene Ontology Consortium, 2015) annotations between such closely related species appears to be plausible. The transfer (interpolation) of information can also be applied to biological networks allowing the enrichment of network models and the identification of conserved and species-specific interactions and sub-networks.

The gene and protein sequence (dis)similarity between species only partially explains the species divergences (noted already long before the rise of NGS methods by King and Wilson in 1975). As proteins interact with each other (already small) differences in the coding and non-coding area of a genome can have drastic (phenotypic) effects (Romero et al., 2012; Villar et al., 2014). Already single amino acid substitutions in the binding region of a transcription factor can lead to different binding affinities and, thus, to different regulations (Ihmels et al., 2005; Alon, 2007; Villar et al., 2014). In contrast, remarkable similarities in regulatory mechanisms for several essential regulatory sub-networks, even between phylogenetically very distant species, have been observed — for example, the heart specification kernel in fly and vertebrates (see Figure 1.1) shares many conservations (Davidson, 2006). A first approach for the transfer of interactions can be based on the sequence similarity between orthologous genes. But it remains unclear: (i) whether the (orthologous) transcription factor still recognizes the binding motif in the promoter region of the target gene, and (ii) in which conditions the orthologous regulation occurs in the target species. Furthermore, there is a general agreement that alternative splicing — a quite species-specific process (Merkin et al., 2012; Barbosa-Morais et al., 2012) — affects interactions (Resch et al., 2004; Ellis et al., 2012; Buljan et al., 2012, 2013).

Problem Identification and Contribution

In this thesis, the three aspects: (i) **cross-species transfer** and conservation of biological networks, (ii) **context-specific comparison of networks**, and (iii) the **conservation of alternatively spliced variants** are investigated:

Cross-Species Transfer: Protein-protein interaction networks are typically simplified as they are binary, global, and undirected, but (compared to other network types) many protein interaction networks have been experimentally identified for eukaryotic model organisms like yeast, fly, mouse, and human. Such networks are deposited in structured databases and can be used for further research. Indeed, even though these networks are simplified, they have been successfully used for many research questions; for example for the prediction of protein

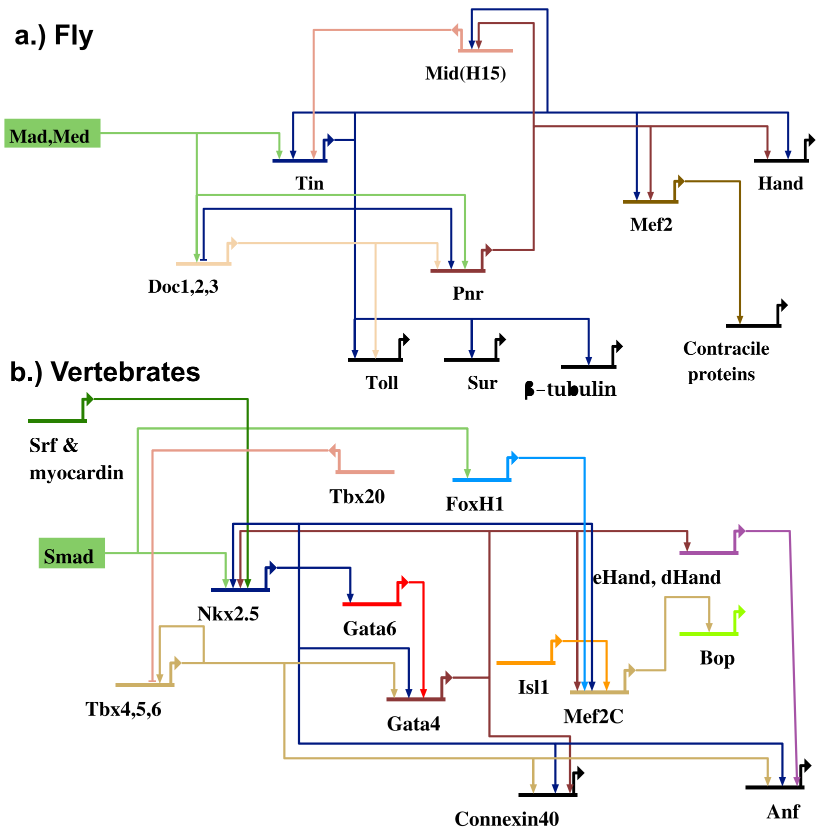


Figure 1.1. Pan-bilaterian kernel for heart specification in fly and vertebrates; adapted from Davidson (2006). (Nearly) all animals with bilateral symmetry (i.e. animals having a front and a back end, as well as an upside and downside) have a heart, even though the organ is structured very differently in different clades. The regulatory sub-network responsible for the specification of the heart progenitor field in (**a**) fly and (**b**) vertebrates shares many similarities (orthologous genes are colored similarly). For example, the auto-regulation of Tin (Nfk.2.5) and the regulation of Tin and Mef2 (Mef2C) and Pnr (Gata4) are conserved.

functions and the interpretation of experimental data (Mitra et al., 2013). Thus, many analyses can benefit from (more or less) complete protein-interaction networks. We present an approach for the cross-species transfer of global protein interaction networks and apply this approach for the enrichment of the interactome for many eukaryotic species (**COIN**, Chapter 3).

Context-Specific Comparison of Networks: Similar to protein-protein interaction networks, regulatory (transcription factor- target) networks can be treated as global and binary networks. Thus, again the question arises to which extent these networks are conserved and can be transferred between species. In contrast to protein-protein interactions, there exists no comprehensive repository of regulatory networks for many eukaryotic species. Furthermore, the conservation of regulatory networks remains mostly speculative. Regulatory interactions are often *only* described in the scientific literature. Furthermore, regulatory elements are strongly context-specific. Projects like ENCODE, TCGA and the Epigenomic Roadmap provide resources of context-specific regulatory raw-data, which in turn allow the definition and analysis of more realistic networks (compared to global networks). Thus, approaches are needed to collect, integrate, and to infer regulatory interactions from diverse data sources in order to conduct cross-species and cross-context regulatory network comparisons. We present such approaches, data repositories, and software tools for the analysis and comparison of global (**ConReg**, Chapter 4) and context-specific regulatory networks (**CroCo**, Chapter 5).

Conservation of Alternatively Spliced Variants: Genes and proteins are the main entities of biological networks. Alternative gene products produced via alternative splicing and alternative transcription can affect interaction networks (Resch et al., 2004; Ellis et al., 2012; Buljan et al., 2012, 2013). Therefore, a realistic network model should integrate alternative variants. The identification of conserved spliced variants and the integration of this information with cross-species network transfers will allow for a more realistic transfer and comparison of biological networks. The first step (identification of conserved spliced variants) is addressed in this thesis with a novel gene, isoform, and exon-intron structure aware multiple sequence alignment approach based on partially ordered graphs (**ISAR**, Chapter 6).

Chapter 2

Background

The transfer of eukaryotic protein interaction and regulatory networks and the cross-species comparison of alternative isoforms are addressed within this work. In the following chapter, a general description of these networks, the evolutionary relationships of genes and the effects of alternative isoforms on networks is provided. The cross-species transfer of biological networks is based on the evolutionary relationships of genes, therefore also the definition and identification of orthologs and paralogs are briefly discussed. Moreover, a short literature review of the influence on the structure of biological networks by: (i) alternative splicing, (ii) alternative transcription, and (iii) post-translational-modifications (PTM) is provided.

2.1 Biological Networks

In biological systems genes, proteins, and (drug) compounds interact with each other. Systems biology studies the often complex interactions between those entities with the ultimate goal of understanding how these interactions cause the observed changes in a system (Ideker et al., 2001). Different network types like protein-protein interaction, gene regulatory (including transcription factor- target), and signaling networks can be modeled and even combined in order to get more complete views on a system. Surprisingly, the general architecture and topology of such biological networks appears to share many organizational properties such as scale-free, small world and high average clustering coefficient with numerous non-biological networks (Albert, 2005). Furthermore, small recurring building blocks, so called network motifs, could be identified in regulatory (Shen-Orr et al., 2002; Alon, 2007) and protein interaction networks (Yeger-Lotem et al., 2004). In the following protein interaction and regulatory networks are briefly introduced:

Protein-Protein Interaction Networks: Proteins are the cell's building blocks carrying out most of the function within a cell (Alberts et al., 2008). But proteins rarely act alone (Berggård et al., 2007; Rao et al., 2014). Indeed, they interact (bind) together in order to perform, or to participate in various essential molecular processes like signal transduction, DNA replication, muscular contraction, and transcription. The physical binding of proteins can be measured with different methods such as Yeast-Two-Hybrid, Tandem Affinity Purification

(TAP), protein microarrays, or directly derived from known molecular three-dimensional structures (see Rao et al. (2014); Phizicky and Fields (1995) for reviews on protein interaction detection methods). These methods have different advantages and disadvantages. For example, a high-throughput method such as Yeast-Two-Hybrid allows the measurement of many interactions simultaneously, but high false positive rates are reported (Rhodes et al., 2005). Protein interactions derived from three-dimensional structures allow precise investigation of the protein interfaces, but structural identifications of protein complexes are still labor intensive and not always feasible. Individual interactions derived from different methods can be combined into a protein-protein interaction network. Databases like iRefIndex (Turinsky et al., 2010), BIND (Bader et al., 2001) and BioGRID (Chatr-Aryamontri et al., 2013) provide such experimentally derived networks for many species with information about the experimental methods used to measure the interactions, literature references and confidence values.

Regulatory Networks: A gene regulatory network describes how molecular entities interact with each other in order to control the abundance of gene products, and subsequently specific cell functions (Karlebach and Shamir, 2008). Transcription Factor (TF) - Target Genes (TG) interactions represent the majority of such relations. The transcription of the TG is mediated by the physical interaction between TFs and cis-acting regulatory elements in the promoter region of the target genes (Janky et al., 2009). Transcription factor binding sites (TFBS) in the promoter region of a TG can be experimentally identified with high-throughput techniques like ChIP-seq, DNaseI Footprinting (Furey, 2012) and inferred from gene expression data (Karlebach and Shamir, 2008). Furthermore, bindings can be predicted purely computationally using Position Weight Matrices (PWM) of TFs (see Stormo (2013) for a review on computational TFBS prediction approaches). Experimentally derived regulations are more realistic than computationally predicted regulations as they represent regulations that have been actually observed in a specific system, but currently no experimental technique allows the measurement of all TF bindings in a system/genome at once. In contrast, computational TFBS predictions can be quickly computed for all factors with associated PWM, but these predictions are typically very speculative and do obscure the strong context-specificity of transcription factor bindings. Compared to protein interactions, currently no comprehensive repository of regulatory interactions is available for many eukaryotic species. Only some species-specific databases like REDfly (Gallo et al., 2011) and YEASTRACT (Teixeira et al., 2006) provide a collection of experimentally derived regulatory and manually curated interactions for fly and yeast, respectively. Furthermore, resources like ORegAnno (Griffith et al., 2008) and TRANSFAC (Matys et al., 2006) provide some manually curated regulatory information for selected model organisms.

2.2 Gene Conservation

The cross-species transfer of biological networks is commonly based on the evolutionary relationship of genes between species. The availability of sequenced and annotated genomes enables the reconstruction of the evolutionary history of genes, the identification of conver-

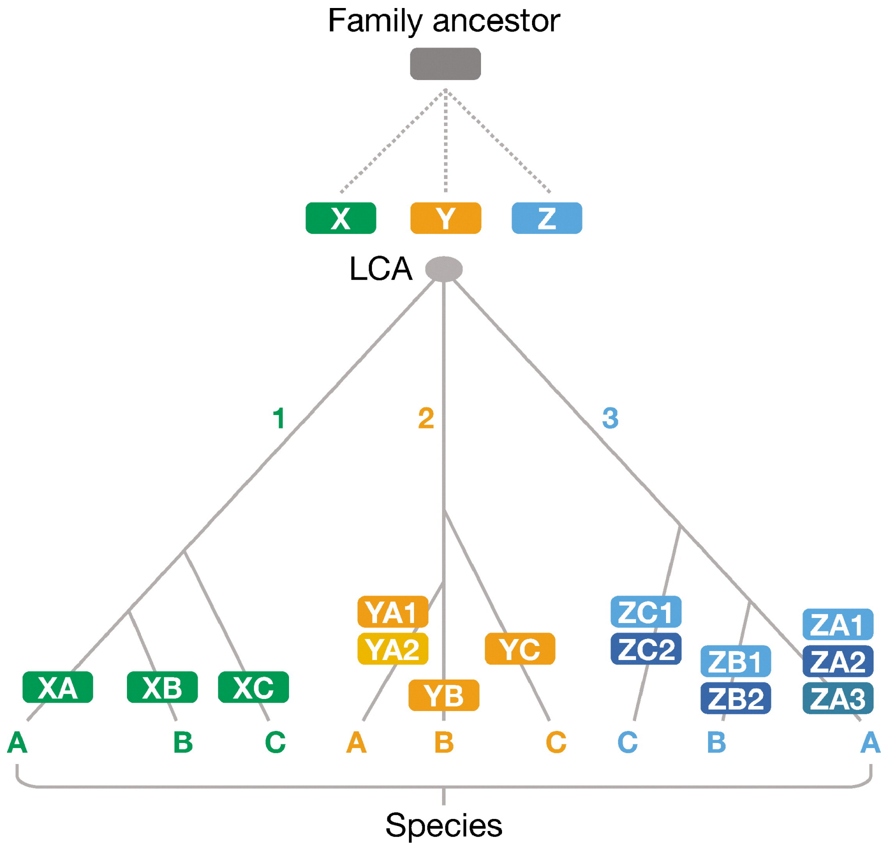


Figure 2.1. Hypothetical gene tree illustrating orthologous and paralogous relations of three genes in three species; taken from Koonin (2005). A common ancestor has three paralogous genes X, Y and Z. The different branches (1, 2 and 3) show hypothetical evolutionary events of these genes. Due to the gene-duplication in the common ancestor, the genes in the different branches in species A, B and C are all out-paralogous to each other (XA, XB, XC to YA1, YA2, YB, YC, and so forth). In branch 1 the ancestor gene X is *only* specialized in species A, B, C. Subsequently, genes XA, XB and XC are orthologous. In branch 2 a lineage-specific duplication of gene Y occurs in species A. According to the ortholog definition the in-paralogs in species A (YA1 and YA2) are still both orthologous to YB and YC. The situation in branch 3 is similar (as the duplication is lineage-specific) and thus the genes ZC1-ZC2 etc. are collectively orthologous (co-ortholog).

sations and species- and lineage-specific adaptions, and thereby the transfer of interactions between species. Gene and protein products are often well conserved between species, but the corresponding genomes often undergo quite complex rearrangements (see for example Figure 2.2 for the mapping of human and mouse orthologous genes on the respective genomes). The evolutionary history of genes can be described with various evolutionary events. Koonin (2005) listed the following events that allow the description of the evolutionary history of related genes (according to their relative occurrence): (i) gene specialization, (ii) gene duplication, (iii) gene loss, (iv) horizontal gene transfer, (v) gene rearrangement including fusion and fission of genes. The history of evolutionary related genes can indeed be quite complex and composed of many such events. A further complicating matter is that the genome of the common ancestors is typically not preserved and thus, cannot be used for the evolutionary event reconstruction.

Depending on the series of evolutionary events, genes in different species are called: (i) **homologous** (genes, sharing a common origin), (ii) **orthologous** (genes, which arise via specifications from a single ancestor gene in the least common ancestor), or (iii) **paralogous** (genes, which arise via gene duplications) (Koonin, 2005). Homology is the most general term, which can be used to describe the evolutionary relationship between genes, independent, of the series of evolutionary events. The orthologous definition appears to be well-defined (given the relationship with the common ancestor), whereas the paralogous definition is imprecise. The paralogous definition does not define whether the duplication is lineage-specific, or has occurred in an ancestor. A series of comments on the importance and common misunderstandings of paralogs and orthologs by Petsko (2001), Koonin (2001) and Jensen (2001) highlighted that precise and further definitions of evolutionary relationships are needed. Indeed, the definition of ortholog and paralog can be further divided into (adapted from Koonin (2005)): co-ortholog (two or more genes in one species are orthologous to a group of genes in another species, due to lineage-specific gene duplications), pseudoparalog (genes which appear orthologous due to lineage-specific gene loss), out-paralog (gene duplications preceding a specialization event), and in-paralog (gene duplications subsequent to a specialization event). Ortholog relations are not necessarily one-to-one relations. In fact, they can be rather complicated. See for example Figure 2.1, which depicts the relationships of three genes X,Y,Z from a common ancestor in three different species A,B,C. In this example, the genes in the right branch are co-orthologous as the evolutionary event in the common ancestor of the three species is a gene specification event (even though gene ZC1-ZC2, ZB1-ZB2, ZA1-ZA3 are in-paralogous).

The Quest for Ortholog consortia (Gabaldón et al., 2009) lists (currently) over 40 different ortholog databases: each using (slightly) different methods and parameters, and including different sets of species. Such databases rely on different methods like reciprocal BLAST best-hit results, graph-based methods that cluster orthologs, or tree-based methods. As the evolutionary history between species often remains unknown, no comprehensive and often only indirect evaluation of the quality of such detection methods can be performed. Typical criteria used to benchmark the quality of such approaches are based on the functional similarity of the identified orthologs, or the overlap with manually curated ortholog sets (Altenhoff and Dessimoz, 2009).

The computational reconstruction of the evolutionary events including the classification of genes in the different orthologous and paralogous classes enables the comparison of species and the transfer of information between them. However, the functional entities in biological systems are transcripts and proteins; therefore, the definitions of the previously described evolutionary event classes could be adapted and extended to these entities as well (Zambelli et al., 2010). With ISAR (see Chapter 6) we present a general approach for the cross-species

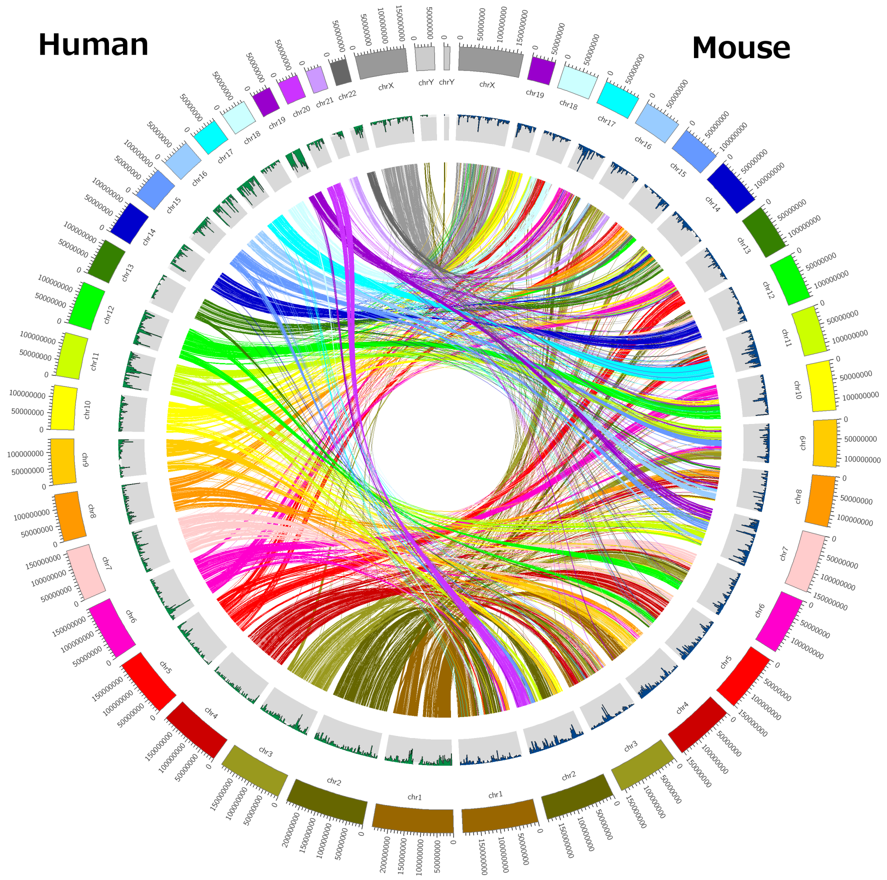


Figure 2.2. Rearrangements of the human and mouse genome. The left side shows the human genome with chromosomes 1-22 and X, Y. The right side shows the mouse genome with chromosomes 1-19 and X, Y. The distribution of genes on the different chromosomes is shown in green and blue for the two species. The lines between the two genomes indicate the mappings of the positions of the orthologous genes. The line color is according to the chromosomal location of the human gene. Some chromosomes like the X chromosome appear to be well conserved between the two species with only inter-chromosomal rearrangements (top of the figure; grey edges), whereas other chromosomes are subject to more complicated rearrangements. The figure has been created using Circos (Krzywinski et al., 2009) and gene and ortholog data from the ENSEMBL database (Flicek et al., 2014).

alignment and transfer of genes, transcripts, isoforms, exons and introns.

2.3 Alternative Gene Products and Biological Networks

The processes of alternative splicing, alternative transcription and post-translational protein modification are ubiquitous in almost the complete eukaryotic domain. They affect almost all genes and thereby increase the diversity of gene products (Pal et al., 2011; Merkin et al.,

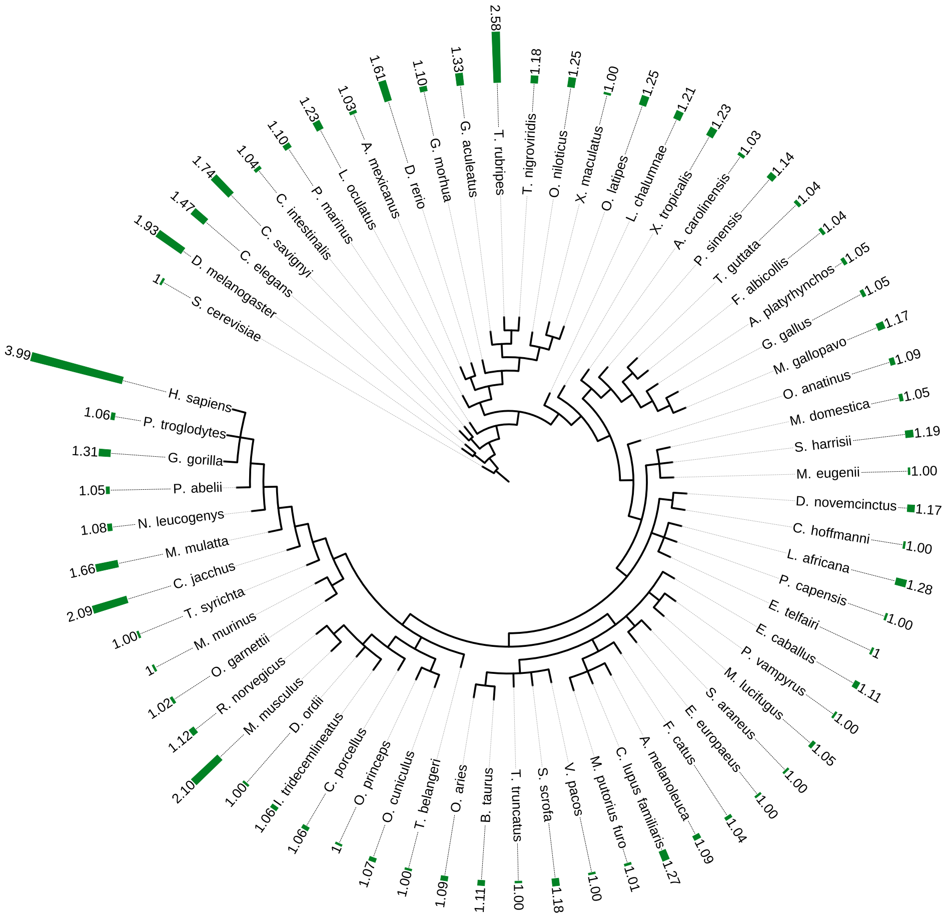


Figure 2.3. Average number of isoforms per gene for 66 species contained in the ENSMBL database (Flicek et al., 2014). For human most isoforms are annotated (on average 4). Also for mouse and fish and a few other species, on average more than two isoforms per gene are known. But for many of the other species only around one isoform per gene on average is annotated.

2012; Barbosa-Morais et al., 2012). With the sequencing of the human genome it became clear that the human genome has *only* between 19,000 and 22,000 (protein-coding) genes (Ezkurdia et al., 2014; Harrow et al., 2012; Flicek et al., 2014), which is just around four times more genes than in baker's yeast, a single-celled eukaryotic organism. But from these 19,000 to 22,000 (protein-coding) genes in human many more different (protein-coding) transcripts can be generated (Harrow et al., 2012; Flicek et al., 2014), whereas in yeast typically only one transcript for each gene is produced. And finally due to alternative splicing — also a process, which is only rarely used by yeast — and post translational modifications, maybe more than a million different proteins can be produced in human (Jensen, 2004). In the following, the mechanisms of alternative splicing, alternative transcription and post translational modification are defined and discussed with respect to their known impact on biological networks.

Alternative Splicing and Alternative Transcription: Transcription describes the process of *converting* DNA segments from a gene to pre-mRNA via the RNA polymerase. In eukaryotes this pre-mRNA consists of coding regions (exons) and non-coding regions (introns). Via the spliceosome, a large molecular machine, different parts of the pre-mRNA are joined together (Alberts et al., 2008). The splicing process can generate a range of different variants by including different regions in the final mRNA (alternative splicing). A gene may have several alternative transcription start sites giving rise to different pre-mRNAs (alternative transcription). These alternative transcripts can in turn again undergo alternative splicing. Alternative spliced isoforms (defined via alternative splicing and alternative transcription) can be found in eukaryotes ranging from yeast to human (Kim et al., 2007; Keren et al., 2010; Grützmann et al., 2014). However, different mechanisms (exon definition vs. intron definition) and different prevalence and types of alternative splicing can be observed among species (see Ast (2004); Keren et al. (2010) for reviews on the evolution of alternative splicing).

In humans 95 % of the multi-exon genes undergo alternative splicing (Pan et al., 2008). Many of these alternative products do have specific functions in specific contexts, but the regulation and function of most of these products still remain unknown. A comprehensive literature review of the functions of alternative spliced isoforms showed that splicing can affect DNA binding domains (Kelemen et al., 2013). Furthermore, preliminary analyses of the structure of spliced isoforms revealed that splicing often affects regions on the surface, within coil regions, and disordered regions (Wang et al., 2005; Romero et al., 2006; Buljan et al., 2013). It was also observed that splicing can affect structurally well-conserved regions of the corresponding protein family (Birzele et al., 2008). The latter highlights that alternative products (when folded) may have a distinct protein structure, which differs strongly from the native structure.

As alternative splicing often affects disordered regions (Buljan et al., 2013) and disorder is common in protein complexes (Fong et al., 2009), one can speculate that alternative splicing has an impact on protein interactions. Several analyses have been conducted in order to investigate the effects of alternative splicing on protein interactions using linear motif predictions, known three-dimensional molecular structures and specialized experimental protein-interaction detection methods (Offman et al., 2004; Resch et al., 2004; Ellis et al., 2012; Buljan et al., 2012, 2013; Colantoni et al., 2013). Resch et al. (2004) identified many sequence domain motifs — including some well-known protein-interaction domain motifs — that are more frequently affected by alternative splicing than other sequence motifs. Furthermore, Ellis et al. (2012) and Buljan et al. (2013) were able to extend the analysis of spliced isoforms to tissue-specificity and protein interaction networks. Ellis et al. (2012) observed that the inclusion/skipping of neural cell specific exons — exons that are regulated by a neural cell specific splicing factors — rewires the protein-protein interaction network (see Figure 2.4). In addition, Buljan et al.; Ellis et al. found that proteins harboring tissue-specific exons tend to occupy central positions in interaction networks. In contrast to the previously summarized studies, preliminary analysis based on molecular-structured protein complexes and confirmed spliced variants do not show a significant removal of protein-protein interaction surfaces (Offman et al., 2004; Colantoni et al., 2013). Offman et al. (2004) analyzed 42 alternatively spliced isoforms in 21 amino acid chains which participate in structurally resolved interaction complexes. Several examples could be identified where alternative splicing almost completely removes protein interaction regions in the considered protein interaction set. The authors tested the hypothesis that alternative splicing is correlated with contact regions in protein-protein interactions. Based on that limited data set no statistical correlation between positions of alternative splicing and protein interaction interfaces could be found. A more recent study by Colantoni et al. (2013) of 431 heterodimeric and 763 homodimeric protein interfaces derived from known protein structures revealed that (in the considered sets) protein interfaces are in general avoided by alternative splicing. Similar to Offman et al. (2004), the authors identify only few examples where an alternative isoform affects the protein interaction surface (see for example Figure 2.5). The authors give some explanations for their observation. For example, protein interactions derived from known protein complexes are biased as disordered regions (which are often subject to alternative splicing) are often not resolved in crystallized regions (Colantoni et al., 2013). Furthermore, so called non-trivial spliced isoform (Birzele et al., 2008), isoforms which differ in essential parts from the native structure, may result in changes of the spatial structure of the interaction domain to a degree that prevents interactions (Offman et al., 2004).

Even though the sequencing of RNA fragments (RNA-seq) is an established technology, the assembly of complete transcripts from high-throughput RNA-seq data is currently still difficult (Steijger et al., 2013). Subsequently, the annotation of genomes is mainly based on completely sequenced cDNA (for the species of interest). Therefore, for most eukaryotes the annotation of spliced variants is still sparse. See for example Figure 2.3 for the number of isoforms per gene for some eukaryotes extracted from the ENSEMBL database (Flicek et al., 2014). Surprisingly, even for species, which are phylogenetically close to *well* annotated species, as for example for chimpanzee, the annotation of isoforms is sparse in current databases. Thus, a cross-species transfer of transcripts and isoforms could be used for the completion of the transcript and protein annotation.

Post-translational Protein Modification: Post-translational modifications (PTM) are (reversible, or irreversible) chemical modifications at the C-, N- termini, or on the amino acid side chains of a protein, which allow to modify amino-acid properties 'on the fly' (Prabakaran et al., 2012). There are over 300 different types of PTMs (Zhao and Jensen, 2009) including very common modifications like phosphorylation, glycosylation and acetylation (addition of phosphate, glycan and acetyl to a protein). PTMs can act as functional switches for proteins. They may: activate, deactivate, and influence the cellular location (Seo and Lee, 2004), or dynamically alter interaction partner preferences for proteins (Woodsmith and Stelzl, 2014). Advances in mass-spectrometry have resulted in a drastic increase in PTM identification (Beltrao et al., 2012). Current annotation pipelines already identified and categorized over 85,000 experimentally and over 230,000 manually curated PTMs (Prabakaran et al., 2012). Proteins are often modified by several PTM types simultaneously (Duan and Walther, 2015). For example, the well-studied p53 tumor suppressor protein is affected by three PTM-types, which influence its stability and function (Brooks and Gu, 2003). Although

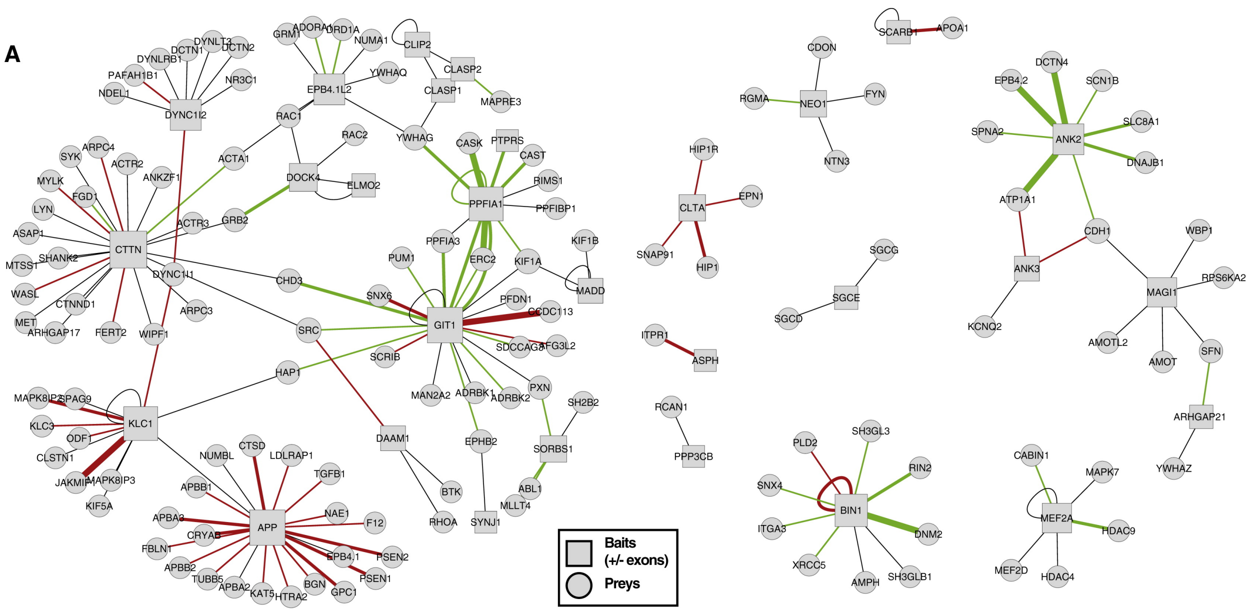


Figure 2.4. Experimentally derived PPI networks for mouse genes containing nSR100-regulated exons with and without nSR-100-regulated alternative exons; taken from Ellis et al. (2012). Green edges and red edges represent interactions that are promoted and inhibited, respectively, whereas gray edges represent unaffected interactions by the inclusion of nSR-100-regulated alternative exons. In a first step, 31 genes containing nSR100/SRRM4 splicing regulated exons were identified, i.e. genes harboring exons that are included in the presence/ absence of nSR100/SRRM4. After that, protein interactions were measured using a co-immunoprecipitation procedure once with and once without nSR100/SRRM4 regulated exons for the previously identified genes.

many PTMs have been studied, still little information is available concerning their function. By linking known protein interfaces with phosphorylation, ubiquitylation and acetylation many PTMs that regulate interactions could be identified (Xin and Radivojac, 2012; Beltrao et al., 2012).

Thus, alternative variants can differ drastically with respect to sequence, molecular structure, molecular function, and role in biological networks. Therefore, alternative products should be considered for network analysis and the cross-species network transfer. However, alternative products are currently not included in biological networks, as experimental data often does not allow the discrimination between different spliced isoforms.

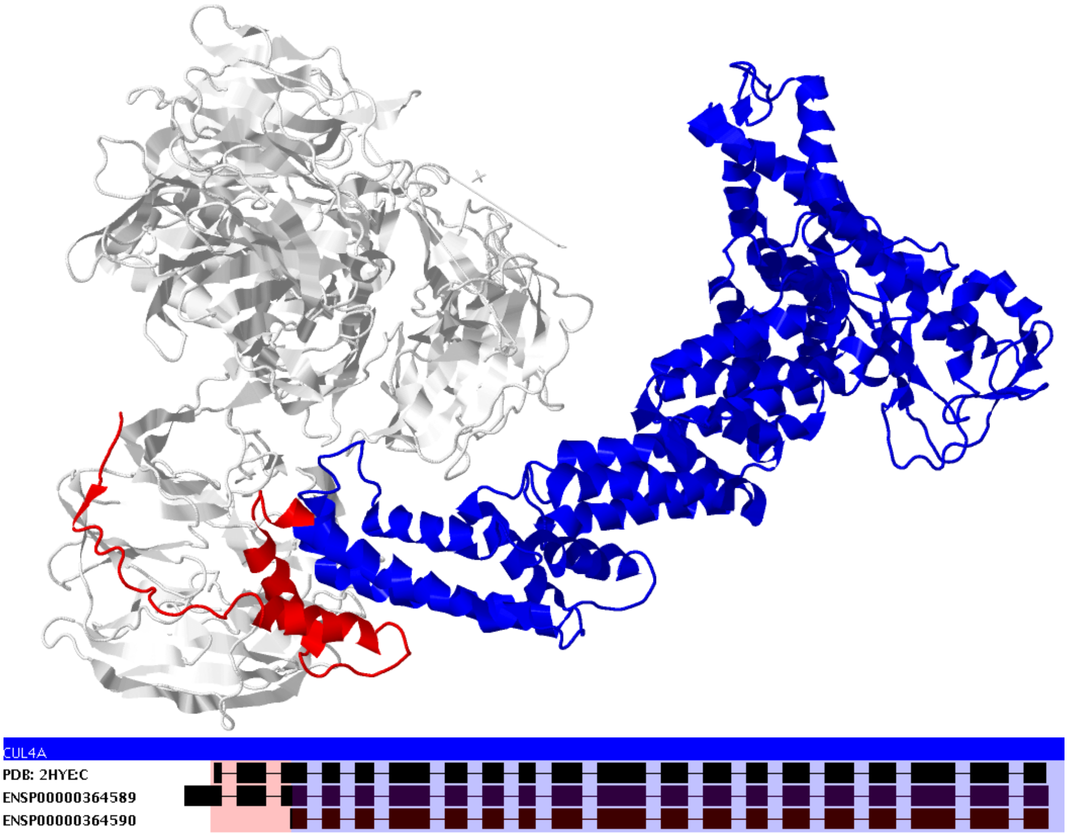


Figure 2.5. Protein interaction of DDB1 (white) and Cul4A (blue/red) affected by an alternative isoform of Cul4A; adapted from Colantoni et al. (2013). Cul4A has an alternative isoform where a huge fraction of the protein interface residues are missing (red). At the top, the protein structure of the interacting proteins is shown (PDB: 2HYE). Below that, the gene structure, i.e. exon-intron structure of the Cul4A wild type, an alternative isoform of Cul4A and the covered region of the Cul4A gene in the PDB are shown. The region that is missing in the variant is highlighted red in the gene structure and the protein complex.

Chapter 3

Protein Interaction Transfer

Abstract: Protein interaction networks are important for the understanding of regulatory mechanisms, for the explanation of experimental data and for the prediction of protein functions. Unfortunately, most interaction data is available only for model organisms. As a possible remedy, the transfer of interactions to organisms of interest using orthologs is common practice, but it is not clear if and when interactions can be transferred from one organism to another and, thus, the confidence in the derived interactions is low. Here, we propose to use a rich set of features to train Random Forests in order to score transferred interactions.

We evaluated the transfer from a range of eukaryotic organisms to *S. cerevisiae* using orthologs. Directly transferred interactions to *S. cerevisiae* are on average only 24 % consistent with the current *S. cerevisiae* interaction network. When, in addition, the interaction type is also transferred, even only 11 % of physical interactions and 15 % of genetic interactions are consistent. By using commonly applied filter approaches the transfer precision can be improved, but at the cost of a large decrease in the number of transferred interactions.

Our Random Forest approach uses various features derived from both the target and the source network as well as the ortholog annotations to assign confidence values to transferred interactions. Thereby, we could increase the average transfer consistency to 85 %, while still transferring almost 70 % of all correctly transferable interactions. If, in addition, the interaction type is transferred we could achieve a transfer consistency of 72 % and 68 % for physical and genetic interactions, respectively.

We tested our approach for the transfer of interactions to other species and showed that our approach outperforms competing methods for the transfer of interactions to species where no experimental knowledge is available. Finally, we applied our predictor to score transferred interactions to 83 target species. We were able to extend the interactomes of *B. taurus*, *M. musculus* and *G. gallus* with over 40,000 reliable interactions.

Our transferred interaction networks are publicly available via our web interface, which allows to inspect and download transferred interaction sets of different sizes, for various species, and at specified expected precision levels.

Publication: The content of this chapter was presented at the German Conference on Bioinformatics 2011 in Munich and is published in PLOS One (Pesch and Zimmer, 2013). Here, I reformatted the text and included the supplement in the corresponding sections. Furthermore, parts of the results and methods are described in a BIOspektrum article (Pesch and Zimmer, 2014).

My contribution: I developed the method and the web interface, carried out the evaluation and drafted the chapter.

Contribution of co-authors: Ralf Zimmer supervised the work and helped to revise the manuscript.

3.1 Introduction

Using high-throughput screening techniques such as Yeast-Two-Hybrid screens, mass spectrometry and protein microarrays large amounts of protein interaction data can be obtained. A protein interaction consists of proteins which bind permanent or transient together in order to carry biological functions. Interaction networks have for example been used to study regulatory networks, to explain experimental data or to predict the functions of proteins (Zhang, 2009). Researchers can query protein interactions from databases like IntAct (Kerrien et al., 2012) and BioGrid (Chatr-Aryamontri et al., 2013). This databases include interactions derived from large-scale experiments, from literature curations, from user submissions, and interactions from protein structures. The current protein interaction networks are mostly derived from high-throughput experiments and hypothesis-driven low-throughput experiments applied to particular gene sets of interest (Sambourg and Thierry-Mieg, 2010).

The experimental identification of interactions is a time consuming and costly process, so that high-throughput experiments have mostly been conducted on model organisms such as *S. cerevisiae* (Gavin et al., 2002), *H. sapiens* (Ewing et al., 2007), *A. thaliana* (Ehlert et al., 2006) and *D. melanogaster* (Uetz and Pankratz, 2004). The interaction networks for other species are still extremely sparse (see Table 3.1). Furthermore, all experimental protein interaction detection methods have different weaknesses and biases (Michaut et al., 2008). For example false positive rates up to 50 % are reported for Yeast-Two-Hybrid screens (Rhodes et al., 2005), literature curations do often not agree (Turinsky et al., 2010), and data from Tandem Affinity Purification (TAP) requires involved data processing in order to infer physical protein interactions (Berggård et al., 2007; Friedel and Zimmer, 2009).

Numerous computational approaches have been developed to predict protein interactions in order to enrich the interactome of species of interest. In particular, knowledge from other (model) organisms can be used to predict protein interactions for a specific target organism. But link attachments, link detachments, gene duplications and gene losses lead to (evolutionary) changes in protein interaction structures (Berg et al., 2004). Gene duplications lead also to the duplication of interactions and again nucleotide substitutions can lead to a network rewiring.

Matthews et al. (2001) introduced the term interolog — an orthologous gene pair interacting in at least one species. Many methods transfer interaction data employing such interologs (Gandhi et al., 2006; Bork et al., 2004; De Bodt et al., 2009; Michaut et al., 2008; Yu et al., 2004). Matthews et al. was able to experimentally validate between 16 % to 32 % of transferred protein interactions from *S. cerevisiae* to *C. elegans* with different ortholog identification techniques. Several features are commonly used to increase the reliability of interaction transfers via interologs. The simplest approach is to require a certain interolog quality, e.g. a minimum bootstrap score for orthologs from the InParanoid database (Gandhi et al., 2006) or a minimum sequence similarity between orthologs in order to transfer an interaction. Yu et al. (2004) showed that protein interactions can be safely transferred if the joint sequence identity between the orthologs involved in the transfer is larger than 80 %. More advanced filter approaches use thresholds for the Gene Ontology (GO) (Ashburner et al., 2000) annotation similarity, domain similarity, gene expression correlation or other features of the interologs (De Bodt et al., 2009; Michaut et al., 2008; Wiles et al., 2010; Garcia-Garcia et al., 2012; Gallone et al., 2011). To achieve a specified performance, random protein pairs are compared with known protein interaction partners to define thresholds for the different features. Besides the inference of protein interaction from interologs, various other approaches try to predict interactions using structural properties (Tuncbag et al., 2011), network topology information (Pao-Yang Chen, 2008), and protein domain information (Luo et al., 2011). The *STRING* database follows a different approach to score interactions by combining information from experiments, databases, text-mining and transfer information (Szklarczyk et al., 2011).

Lewis et al. (2012) claimed that the transfer consistency cannot easily be improved. Furthermore, they showed that the evolutionary change of interactions is too high to allow the direct transfer of interactions for phylogenetically distant species unless a strict definition of homology is used. In contrast van Dam and Snel (2008) showed that protein complexes are highly conserved even between *H. sapiens* and *S. cerevisiae*. All network transfer studies rely on homologies which can be identified with different ortholog detection methods like simple bidirectional BLAST best hit results, graph-based methods that cluster orthologs, or tree-based methods. Benchmarks of orthologs detection methods have shown that there is no best method for ortholog detection (Altenhoff and Dessimoz, 2009). It is obvious that with conservative ortholog detection approaches only relatively few interactions can be transferred, but that these interactions are more likely conserved, whereas with cluster based and tree based methods groups of orthologs are produced which allow the transfer of more interactions. Thus, the usage of ortholog identification approaches, the choice of experimental data (only physical interaction derived from Yeast-Two-Hybrid studies, or more relaxed interaction data which includes interactions from TAP or Co-ImmunoPrecipitation experiments, or even protein complexes) and the approaches used to deal with the incompleteness of current networks result in different estimated protein interaction conservation rates.

In this paper, new features and successfully used features in the literature are exploited to train Random-Forests-Filters (*RFF*) for the reliable transfer of interactions to even phylogenetically distant species. The *RFF* models are trained with interactions transferred from various eukaryotic species to *S. cerevisiae* using all available interactions from an integrated database and orthologs from cluster based approaches. We train the models on yeast for the only reason that the *S. cerevisiae* network is assumed to be the most complete one, which allows to distinguish correct and incorrect transfers in the learning phase. Another assumption we make is that the learned *RFF* models can be used for other species as well. This is reasonable as the models learn the important features (e.g. sequence similarity, orthology, network properties, functional similarities) and their appropriate weightings, which will hold in a species-independent way (there are no particular *S. cerevisiae* specific features or parameters). The transfer performance on *S. cerevisiae* is taken as an estimate for the expected performance on other species, especially for phylogenetically closer ones. We applied the trained *RFF* predictor to transfer interactions on a large scale in-between various eukaryotic species. This increases the available reliable interactions for non-model organisms manyfold without inflicting too many false positives. The transferred networks are publicly available

Table 3.1. Overview of protein interaction networks extracted from the iRefIndex (Turinsky et al., 2010) database for the ten eukaryotic model species with the largest protein interaction networks. Besides the total number of protein interactions, the number of physical, genetic and interactions with an unknown interaction type is given. Only the interaction network of *S. cerevisiae*, *H. sapiens*, *D. melanogaster* and *S. pombe* have more than 2 interaction per gene (*S. cerevisiae* peaks with 28.19).

Species	Genes	Interactions				
		Physical	Genetic	Other	Total	Avg. number of interactions per gene (total)
S. cerevisiae	6,328	55,767	104,926	17,674	178,367	28.19
H. sapiens	28,383	43,412	71	20,992	64,475	2.27
D. melanogaster	14,321	19,088	2,118	17,265	38,471	2.69
S. pombe	4,958	1,943	9,665	804	12,412	2.5
C. elegans	20,184	5,483	1,785	4,208	11,476	0.57
M. musculus	24,865	3,513	3	2,596	6,112	0.25
A. thaliana	26,496	5,048	67	937	6,052	0.23
P. falciparum	5,503	2,215	0	4	2,219	0.4
R. norvegicus	24,770	804	0	867	1,671	0.07
D. rerio	24,352	173	11	13	197	0.01

at our web interface. Compared to competing approaches to predict protein interactions we integrate a wide range of features and, instead of using fixed thresholds, employ a systematic and conservative *RFF* approach with an associated performance estimate for the (distant) transfer to *S. cerevisiae.*

3.2 Materials and Methods

3.2.1 Data Sources

We use iRefIndex (Turinsky et al., 2010), an integrated interaction database, for our study. iRefIndex integrates interaction data for multiple species in a common format from the 13 different interaction databases: BIND (Bader et al., 2001), BioGRID (Chatr-Aryamontri et al., 2013), CORUM (Ruepp et al., 2010), DIP (Xenarios et al., 2000), HPRD (Keshava Prasad et al., 2009), InnateDB (Lynn et al., 2008), IntAct (Kerrien et al., 2012), MatrixDB (Chautard et al., 2011), MINT (Chatr-aryamontri et al., 2007), MPact (Güldener et al., 2006), MPIDB (Goll et al., 2008), MPPI (Pagel et al., 2005) and OPHID (Brown and Jurisica, 2005). All these databases include experimental validated data extracted from differ-

Table 3.2. List of data sources used for this study.

Database	Version	Download date	Used for
UniProt	N.A	July, 2011	Features, mapping, external references
KEGG	N.A	July, 2011	Features
OMA	2011	July, 2011	Orthologs
InParanoid	7	February, 2011	Orthologs
HomologGene	65	July, 2011	Orthologs
iRefIndex	8	July, 2011	Protein interaction data
STRING	9	May, 2011	Transferred human protein interaction network for comparison
InteroPorc	N.A	May, 2011	Transferred human protein interaction network for comparison
eggNog	3.0	January, 2013	Orthologs
TreeFam	7	January, 2013	Orthologs
EnsemblCompara	N.A	January, 2013	Orthologs

ent sources, besides OPHID which also makes use of transferred interactions. Therefore, we excluded interactions from OPHID for our study. Furthermore, iRefIndex includes binary interactions (physical and genetic) and few protein complexes. We transfer binary interactions from iRefIndex (physical, genetic and other interaction types including ambiguous or interactions without type annotation) to target species using publicly available ortholog mappings. Orthologs are obtained from the Orthologs Matrix Project (OMA) (Schneider et al., 2007), InParanoid (Remm et al., 2001) and HomoloGene (Sayers et al., 2009). These databases are used due to their evaluation results in Altenhoff et al. (2011) and the coverage of ortholog mappings for various eukaryotic species. The interaction partners and orthologs are mapped to UniProt (The UniProt Consortium, 2011) as a common reference to obtain annotations including GO terms, synonyms and mappings to external databases (see Table 3.2 for an overview of the used data sources). We consider all eukaryotes species for which we could transfer at least one interaction given the interaction and ortholog databases. Thus, we consider 83 out of the approximate 166 (until January 2013) fully sequenced eukaryotes for the subsequent analysis.

3.2.2 Interaction Transfer

Protein interaction networks are modeled as graphs PPI $= (P, I)$ consisting of a set of proteins P and interactions $I \subseteq P \times P$. Given an interaction network $PPI^i = (P^i, I^i)$, a target protein set P^j and an ortholog mapping $O : P^i \rightarrow P^j$, a *directly interolog based transferred interaction network* consists of
$PPI^j = (P^j, I^j)$ with $(p^j_x, p^j_y) \in I^j \iff (p^i_x, p^i_y) \in I^i \wedge p^j_x = O(p^i_x) \wedge p^j_y = O(p^i_y)$. Trans-

ferred interactions (p_k^j, p_c^j) can be scored and filtered to obtain a *filtered interolog based transferred interaction network*. In our case, a trained Random Forest Filter (*RFF*) model is used for the scoring of interactions. Its performance is estimated via the interaction transfer to *S. cerevisiae.*

3.2.3 Random-Forest-Filter

For the scoring of transferred interactions we use Random Forests (RF) from the WEKA (Hall et al., 2009) machine learning framework. Random Forests predict the outcome class (correct, incorrect) of an instance (transferred interaction) by using a voting procedure on several learned decision trees with different feature sets. Random Forests have shown good evaluation results on similar learning tasks (Caruana and Niculescu-Mizil, 2006) and are considered more robust against noise than other ensemble machine learning methods (Breiman, 2001). RF rely on two parameters, the number of trees to learn and the number of features to consider. We determine these parameters via a grid search. In addition to the output class label, the WEKA Random Forest implementation provides a score value between 0 (low confidence) and 1 (high confidence), which we use as score value for transferred interactions.

3.2.4 Features

As features we use the protein annotations of the interacting partners in the source and the target network and of the orthologs from which an interaction is transferred. The features can be classified into four categories: 1.) Features modeling Gene Ontology similarities (Gene Ontology), 2.) features derived from the network structure (Network), 3.) features describing the similarity between orthologs (Orthologs) and 4.) general features (General).

Gene Ontology

GO similarity: We compute the semantic GO similarity for two proteins based on Resnik (Resnik, 1999) information content measure

$$\mathrm{IC}(go_i) = -\log\left(\frac{\mathrm{Freq}(\mathrm{go}_i)}{\mathrm{Freq}(\mathrm{go}_{\mathrm{root}})}\right), \tag{3.1}$$

with Freq as the number of proteins annotated with a given term go_i, or its descendant terms in the GO tree. For two GO terms g_k, g_l, we define the semantic GO term similarity as the IC for their common ancestor. And for two proteins p_i, $p_j \in P$ we define the semantic GO similarity as the maximum of all combination of GO annotations for the two proteins. Formally defined as

$$\mathrm{GOSim}(p_i, p_j) = \max_{go_k \in \mathrm{GO}(p_i), go_l \in \mathrm{GO}(p_j)} \mathrm{IC}\left(\mathrm{commonAncestor}(go_k, go_l)\right). \tag{3.2}$$

Given that measure, the semantic similarity is computed for the interaction partners in the source and target network and the orthologs. Besides a global semantic GO similarity, one feature is modeled for each of the GO categories cellular component, biological process and molecular function (indicated with C, B, and M behind the feature name in the following) to take the different types individually into account.

Network

Network overlap: The overlap of the neighborhood proteins for a given pair of proteins in the source and target network. For this purpose the Jaccard Index is computed for the direct neighbors of the interacting proteins with the equation

$$J(p_i, p_j) = \frac{n(p_i) \cap n(p_j)}{n(p_i) \cup n(p_j)}, \tag{3.3}$$

where $n(p_j)$ and $n(p_i)$ are the adjacent proteins in the protein interaction network.
Network GO similarity: The average semantic GO similarity between the pair-wise neighbors of the interaction partners in the networks computed with the equation

$$\text{AVGSim}(p_i, p_j) = \underset{p_k \in n(p_i), p_l \in n(p_k)}{\text{avg}} \text{GOSim}(p_k, p_l). \tag{3.4}$$

General

Source interaction database: The source database from which an interaction is extracted as provided as additional information in the used integrated protein interaction database.
Edge support: The number of PubMed abstracts given as evidence for the source interaction.
Source interaction type: The source interaction type (physical, genetic or other) is used as discrete feature value. For this purpose the molecular interaction type (Côté et al., 2006) is used.
Total support: The number of times an interaction is transferred from all other networks to the target network as suggested by Mika and Rost (2006) for confidence scoring.
Gene expression correlation coefficient: Given a gene expression time series for two genes the Pearson correlation coefficient is computed for the putative interacting partners in the target network with the equations

$$\text{Cor}(X, Y) = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} \sum\limits_{i=1}^{n} \sqrt{\sum\limits_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{3.5}$$

where X and Y represent the expression values for the respective genes.

Ortholog

Sequence similarity: The sequence identity of the orthologs.
Harmonic sequence similarity: The harmonic mean of the sequence identities of the

orthologs.
Synonym similarity (Token score): From the orthologs the function of the proteins is extracted from the textual description using UniProt by tokenizing, stemming and filtering stop words and to general words resulting in a set of tokens which are descriptive for the proteins. Based on these function terms we define the similarity for two proteins p_i and p_j from the set of all protein P as

$$\mathrm{TSim}(p_i, p_j) = -\log\left(\frac{|\{p_l \in P | \mathrm{Tokens}(p_i) \cap \mathrm{Tokens}(p_j)\} \subseteq \mathrm{Tokens}(p_l)|}{|P|}\right), \tag{3.6}$$

where Tokens(p_i) and Tokens(p_j) are the function terms of the proteins p_i and p_j.
Domain/Family similarity: The InterPro and PFAM annotations from UniProt are used to compute the domain/family similarity of the orthologs. For two proteins we define the domain/family similarity as

$$\mathrm{DFSim}(p_i, p_j) = -\log\left(\frac{|\{p_l \in P | \mathrm{DFam}(p_i) \cap \mathrm{DFam}(p_j)\} \subseteq \mathrm{DFam}(p_l)|}{|P|}\right), \tag{3.7}$$

where DFam(p_i) and DFam(p_j) are the domain and family annotations.
KEGG Pathway score: Boolean indicator whether the orthologs are involved in the same pathways or in different pathways.
Ortholog source: The database from which the orthologs used for the transfer are extracted.
Ortholog score: Given two orthologs g_i and g_j we define the ortholog score as

$$\mathrm{OScore}(g_i, g_j) = is_i \times is_j \times bs_i \times bs_j, \tag{3.8}$$

where is is the inparalog score and bs the bootstrap score provided by InParanoid for each gene in an orthologous gene cluster.
Ortholog support: The number of times the same ortholog relation between two genes can be found in the different ortholog databases.
Phylogenetic distance: The distance of the source and the target species in a phylogenetic tree provided by Schneider et al. (2007).
Transitive ortholog: The idea behind this feature is that more conserved orthologs can be traced from a source species to a target species along the phylogenetic tree. For this purpose a phylogenetic tree covering all species with ortholog mappings is used. Given such a tree, a path from a source to a target species is computed by:

1. searching the shortest path between the two species and
2. searching the closest leaf nodes for all inner nodes on the shortest path.

The result is a list of species which are "between" the target and the source species. An ortholog is defined as transitively consistent if a direct ortholog between the source and the target species can also be reached when going along the pairwise ortholog mappings on the estimated path.

In the case that a feature cannot be computed because of missing annotation data, the feature is replaced with a missing value indicator. Features are derived from different sources. Thus, in the rest of this article we indicate with (T), (S) and (O) after the feature name whether a feature is modeled between the protein pair in the target network, the source network, or between the orthologs, respectively.

3.2.5 Evaluation Measures

To assess the quality of the learned models we compute the

$$\text{Precision (s)} = \frac{\text{\#Correctly transferred interactions with score} \geq \text{s}}{\text{\#All transferred interactions with score} \geq \text{s}}, \tag{3.9}$$

$$\text{Relative Recall (s)} = \frac{\text{\#Correctly transferred interactions with score} \geq \text{s}}{\text{\#All correctly transferable interactions}}, \tag{3.10}$$

and

$$\text{Regular Recall (s)} = \frac{\text{\#Correctly transferred interactions with score} \geq \text{s}}{\text{\#All experimentally validated interactions in the target species}} \tag{3.11}$$

for a given score value assigned by the learned model. A precision of 1.0 for a given score threshold s is obtained when all transferred interactions with a score value $\geq s$ can be found in the experimentally validated network. The relative recall is 1.0 when all correctly transferable interactions using the available ortholog relations are also transferred after the filtering i.e. all transferred interactions have a score value $\geq s$. We mostly use the relative recall instead of the regular recall in order to assess the recall with respect to a direct transfer. As overall measure for different score thresholds the area under the precision (relative) recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) are computed (Davis and Goadrich, 2006). Furthermore, the Information Gain (IG) i.e. the reduction of entropy of the data set given information about a feature (Mitchell, 1997), is computed to estimate the impact of the different features.

Formally, for a data set D and feature F the IG is defined as

$$\text{IG}(D, F) = \text{E}(D) - \sum_{v \in \text{Values}(F)} \frac{|D_{F=v}|}{|D|} \text{E}(D_{F=v}), \tag{3.12}$$

where $D_{F=v}$ is the set of instances in D with value v for feature F. $\text{E}(D)$ is defined as

$$\text{E}(D) = -p_{\text{positive}} \log(p_{\text{positive}}) - p_{\text{negative}} \log(p_{\text{negative}}), \tag{3.13}$$

where p_{positive} and p_{negative} is the proportion of D belonging to the class of correctly (consistently) and incorrectly (inconsistently) transferred interactions, respectively.

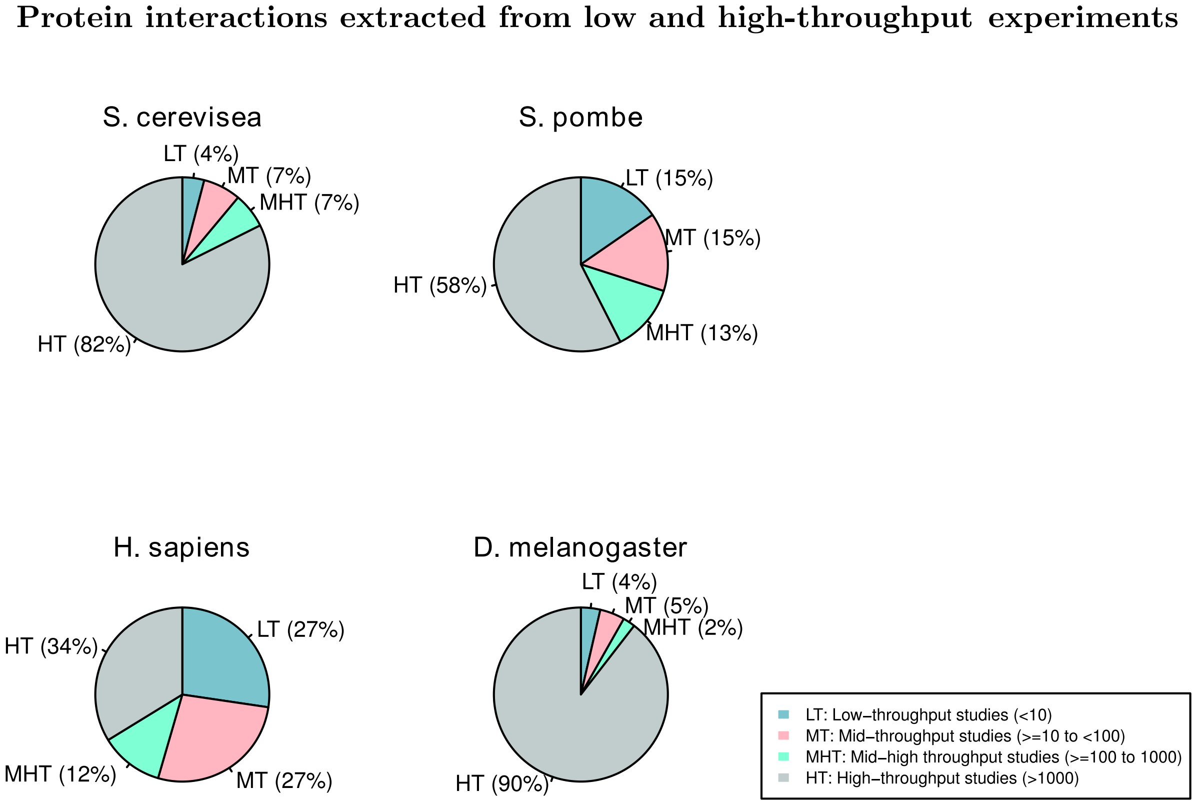


Figure 3.1. Protein interactions for *S. cerevisiae*, *H. sapiens*, *D. melanogaster* and *S. pombe* from iRefIndex (Turinsky et al., 2010) are classified into the categories: derived from low-throughput studies (detected in studies which report between 1 and 10 interactions), derived from mid-throughput studies (detected in studies which report between 10 and 100 interactions), derived from mid-high throughput (detected in studies which report between 100-1000 interactions) and derived from high-throughput studies (detected in studies which report $\geq$ 1000 interactions). See discussion in the main text.

3.3 Results

3.3.1 Current Protein Interaction Networks

Table 3.1 gives an overview of the protein interaction networks derived from the integrated interaction database iRefIndex having the largest number of interactions.

Over 78 % of the *S. cerevisiae* and over 90 % of the *D. melanogaster* interactions stem from high-throughput studies where over 1,000 interactions are reported, whereas for *H. sapiens* only 43 % of the interactions stem from high-throughput studies (see Figure 3.1). Furthermore, most interactions for *S. cerevisiae* are detected with genetic interference and affinity chromatography technology methods like Co-Immunoprecipitation or Tandem Affinity Purification, whereas for *D. melanogaster* most interactions are detected within one high-throughput Yeast Two Hybrid screen (see Figure 3.1 and 3.2).

The total number of interactions consists of physical interactions, genetic interactions and other protein interactions (no interaction type or ambiguous annotations).

With about 180,000 interactions the by far largest eukaryotic interaction network is available for *S. cerevisiae*. The majority of interactions are genetic interactions. When we only consider physical interactions the *S. cerevisiae* interaction network is still the largest. Especially in comparison with the second largest protein interaction network from *H. sapiens* it becomes clear how sparse the networks for the other species still are in current databases. The *H. sapiens* network has fewer physical interactions, but more than four times more genes in the network as compared to *S. cerevisiae*.

Furthermore, only the *S. cerevisiae* network consists of only one connected component. It has been estimated that the complete *S. cerevisiae* network has between 37,800 and 75,500 protein interactions (Hart et al., 2006). Actually, 55,767 physical interactions are contained in iRefIndex for *S. cerevisiae*. Therefore, for the following, we assume that the *S. cerevisiae* network is almost complete and, thus, we use the *S. cerevisiae* network to evaluate the performance of a protein interaction transfer.

It can be expected that more complex organisms also have a more complex network. The number of genes (and maybe also the number of proteins) is not dramatically different and, thus, most likely the number of interactions is different. Therefore, the extremely small coverage of even the best investigated model organisms is apparent. For all other non-model organisms the number of available interactions are neglegible.

3.3.2 Interaction Transfer

Experimental Settings

We transfer interactions from all eukaryotic species with interaction data used in this study to *S. cerevisiae* to train our models. The *S. cerevisiae* interaction network is assumed to be almost complete and possible false negatives in the gold standard are ignored. True positives are defined as transferred interactions, which can be found in the *S. cerevisiae* network, and false positives as transferred interactions, which cannot be found in the network.

Three experimental settings are considered to evaluate our approach:

All interaction setting (*AllI*): All interactions are transferred to *S. cerevisiae* and only the occurrence of the transferred interactions in the gold standard is checked.

Physical interaction setting (*PhyI*): Only physical interactions are transferred to *S. cerevisiae* and in addition to the occurrence of the interactions also the agreement of the interaction type is checked.

Genetic interaction setting (*GenI*): The same as the previous setting, but with genetic interactions.

In total 19,785 interactions from all eukaryotic species considered in this study can be transferred to *S. cerevisiae*. For *AllI* 4,745 interactions can be found in the gold standard and the other 15,040 interactions are used as negative set. The physical, *PhyI*, setting consists of

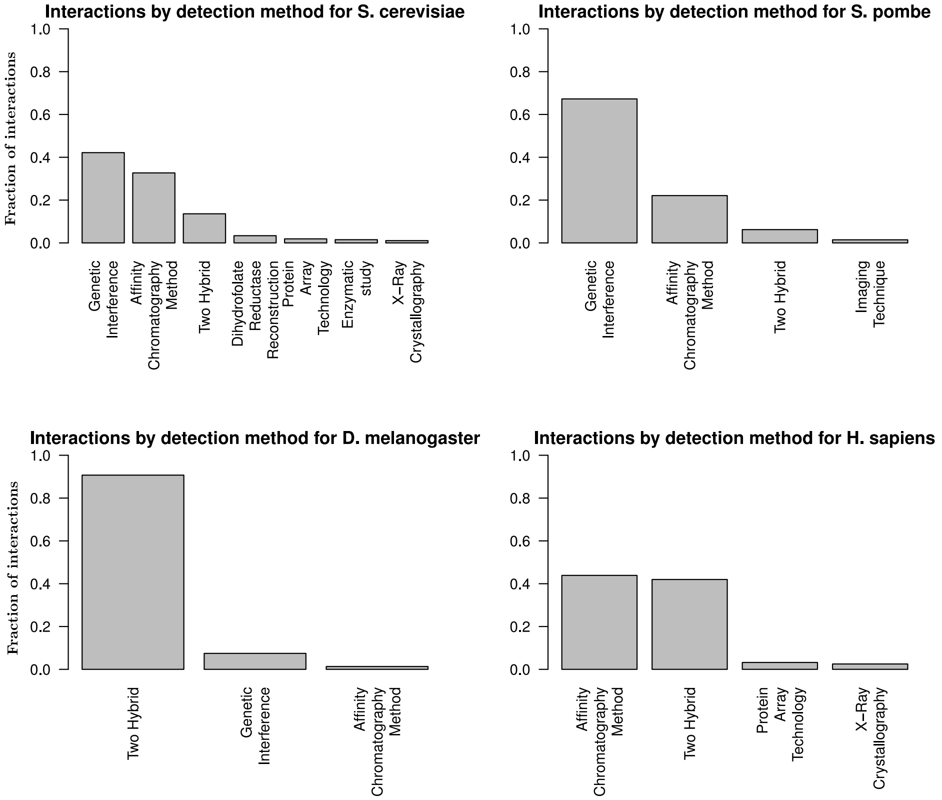


Figure 3.2. Overview of known protein-protein interactions by experimental detection method for *S. cerevisiae*, *H. sapiens*, *D. melanogaster* and *S. pombe* derived from iRefIndex (Turinsky et al., 2010). See discussion in the main text.

1,019 correctly transferred interactions and 8,174 incorrectly transferred interactions. The genetic, *GenI*, setting consists of 901 correctly and 5,300 incorrectly transferred interactions. The remaining 4,391 transferred interactions have an unknown, other, or an ambiguous interaction type.

The features are modeled for the protein pairs involved in the transfer. In total four proteins are considered for each transfer (two proteins from the source network and two proteins from the target network). The features are defined between the different protein pairings in the target network, in the source network and between the orthologs. In total 20 different features types are modeled where for the features used for the orthologs one feature for each of the two orthologs pairs involved in the transfer is created. E.g. for the global GO similarity one feature is modeled between the interaction partners in the source network, one feature is modeled between the interaction partners in the target network and two features are modeled between the orthologs involved in the transfer. For the gene expression feature the compiled gene expression experiment set from Bhardwaj and Lu (2005) which includes normalized intensity values from different cellular states and biological conditions is used.

Six feature sets are constructed for the training of the Random-Forest-Filter (*RFF*) in order to compare the performance and to estimate the feature contribution. This includes

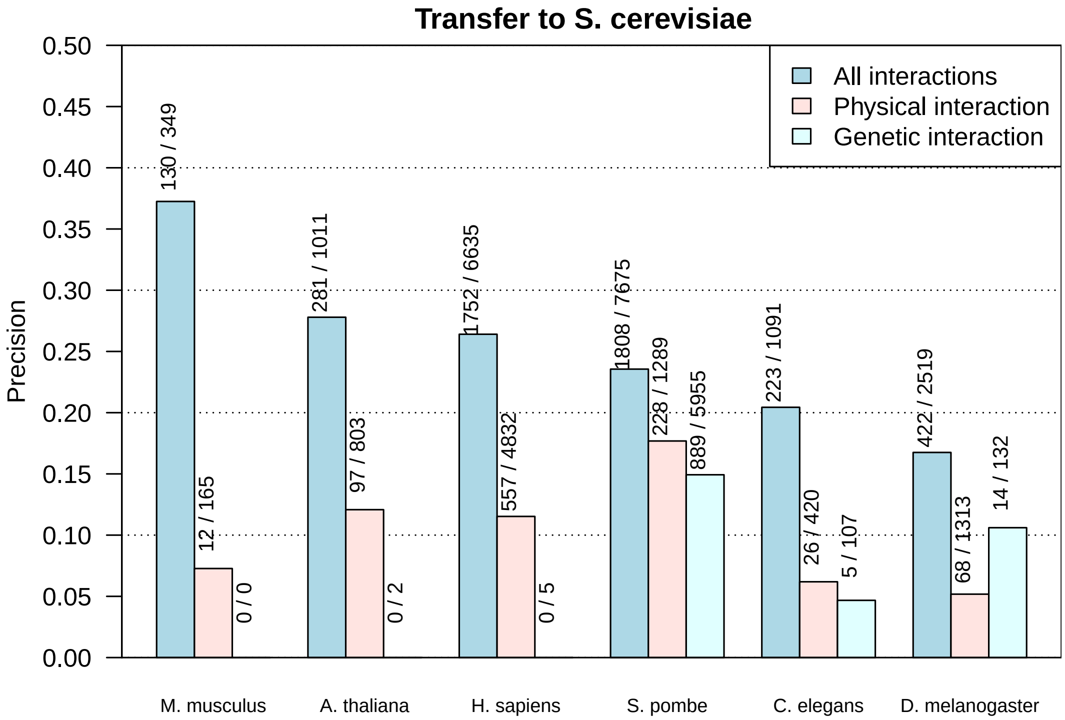


Figure 3.3. Overview of conserved protein-protein interactions, for six eukaryotic species having the largest protein interaction networks. For the all interaction setting (*allI*) only the occurrence of a transferred interaction in the *S. cerevisiae* network is required, whereas for the genetic (*GenI*) and physical (*PhyI*) interaction setting also the exact type of the transferred interaction is checked. In addition to the precisions, the number of total transferred interactions and consistent interactions for each species and type is shown on top of the corresponding bar. Most interactions can be transferred from *S. pombe* to *S. cerevisiae*. There the transfer precision is highest for physical and genetic interactions. For the *allI* setting the highest transfer precision is observed for *M. musculus* to *S. cerevisiae*. This is due to the small number of interactions, which are mostly involved in conserved biological processes like DNA replication and chromosome organization.

two main sets, one in which all features are considered and one setting where only features are used which can be assumed to be available for most of the species. Hence, features containing information about the network structure and the gene expression correlation are excluded in the reduced feature set. The other four feature sets (Network, Gene Ontology, General and Orthologs) consists only of the features from the respective category. In Table 3.3 the composition of the different feature sets and protein pairings is given.

Table 3.3. The table lists the modeled features in the categories "Network", "GO", "General" and "Ortholog". In addition, the involved proteins for the feature are listed (proteins in the target network, proteins in the source network, orthologs). Also the configuration of the full (target network needs to be available) and the reduced feature set (used in real prediction filtering) is shown. For example the feature *GO Global* is modeled employing the interaction partners in the source network (**SP**), in the target network (**TP**) and between the orthologs (**OP**). Furthermore, the *GO Global* feature is included in the GO (**GF**), the reduced (**RF**) and the full feature set (**FF**).

Feature	Feature pairing			Feature set					
	SP	**TP**	**OP**	**NF**	**GF**	**GEF**	**OF**	**RF**	**FF**
Network overlap	✓	✓		✓					✓
GO Network	✓	✓		✓					✓
GO Global	✓	✓	✓		✓			✓	✓
GO (B) similarity (GO Biological process)	✓	✓	✓		✓			✓	✓
GO (C) similarity (GO Cellular component)	✓	✓	✓		✓			✓	✓
GO (M) similarity (GO Molecular function)	✓	✓	✓		✓			✓	✓
Source interaction database	✓					✓		✓	✓
Edge support	✓					✓		✓	✓
Source interaction type	✓					✓		✓	✓
Total support		✓				✓		✓	✓
Gene expression correlation		✓				✓			✓
Sequence identity			✓				✓	✓	✓
Token similarity			✓				✓	✓	✓
Domain similarity			✓				✓	✓	✓
KEGG pathway score			✓				✓	✓	✓
Ortholog source			✓				✓	✓	✓
Ortholog score			✓				✓	✓	✓
Ortholog support			✓				✓	✓	✓
Transitive ortholog			✓				✓	✓	✓
Phylogenetic distance			✓				✓	✓	✓

SP=Source network pairing; **TP**= Target network pairing; **OP**=Ortholog pairing; **NF**=Network feature set; **GF**=GO feature net; **GEF**=General feature set; **OF**=Ortholog feature set; **RF**=Reduced feature set; **FF**=Full feature set.

Direct protein interaction transfer

Using the previously described interaction database and ortholog mappings, interactions are directly transferred to *S. cerevisiae*. In Figure 3.3 the precisions of the interaction transfers from six interaction networks using the previously introduced experimental settings are shown.

We use orthologs from the well established cluster based ortholog detection approaches InParanoid, OMA and HomologGene. Orthologs from these databases result in higher transfer consistencies than orthologs from tree based approaches like EnsemblCompara (see Figure 3.4).

The overall precision of an interaction transfer from the different species to *S. cerevisiae* for *AllI* is 0.24, whereas for *GenI* and *PhyI* the transfer precision is only 0.11 and 0.15, respectively. With 4,745, 1,019 and 901 correctly transferred interactions, 3%, 2% and 1% of the *S. cerevisiae* network can be predicted for the respective experimental settings *AllI*, *PhyI* and *GenI*. The highest transfer precision of physical and genetic interactions can be achieved with a transfer from *S. pombe* (the phylogenetically closest species in our tree with experimentally validated interaction data).

Given complete interaction data for all species it would be expected that the highest precision would be achieved with a transfer from the phylogenetic closest species. But since the interaction data is incomplete and interologs of *S. cerevisiae* might be used as prior knowledge for the interaction discovery, some phylogenetically more distant species show higher interaction transfer precisions than phylogenetically closer species. Most notable is the performance of a transfer from *M. musculus* to *S. cerevisiae* with an unusually high precision of 0.36 in the *AllI* setting. A GO overrepresentation analysis (DAVID, Huang et al. (2009)) of the proteins involved in the transfer from *M. musculus* to *S. cerevisiae* exhibits that some highly conserved biological processes are overrepresented (like DNA-dependent DNA replication, pre-replicative complex assembly, DNA replication initiation and chromosome organization), which might explain the high precision of the interaction transfer. By looking at the transfer precisions for each biological process it can be seen that for these overrepresented biological processes the transfer precision from *M. musculus* to *S. cerevisiae* is almost the same as the transfer precision from *H. sapiens* to *S. cerevisiae*. E.g. 102 transferred interactions from *H. sapiens* to *S. cerevisiae* are associated with the biological process DNA-dependent DNA replication from which 64 are consistent, for the pre-replicative complex assembly process 30 out of 43 and for the DNA replication initiation process 31 out of 45 are consistent.

For phylogenetically distant species ortholog clusters consist of more than two genes which results in 1:n or even n:m mappings. Thus, with a direct transfer a single source interaction can be inferred between different genes in the target network. For *H. sapiens* and *S. cerevisiae* are for example on average 1.9 *H. sapiens* genes and 1.18 *S. cerevisiae* genes in one cluster, whereas for *H. sapiens* and *M. musculus* the cluster contain 1.05 and 1.01 genes, respectively.

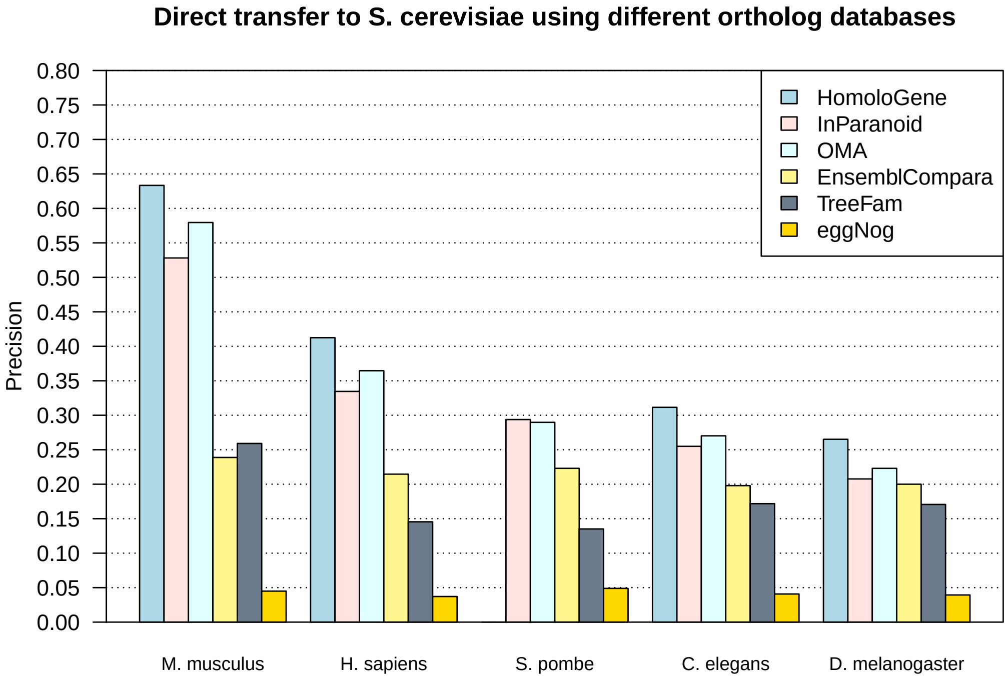


Figure 3.4. Transfer consistencies of a protein interaction transfer from *M. musculus*, *H. sapiens*, *S. pombe*, *C. elegans* and *D. melanogaster* to *S. cerevisiae* using orthologs from OMA, InParanoid, HomoloGene, EnsemblCompara, TreeFam and eggNog for the all interaction setting (allI). See discussion in the main text.

Transfer filter

We train our Random-Forest-Filters (*RFF*) to score directly transferred interactions and to identify possible conservations.

In Figure 3.5 the precision-(relative) recall curves of the Random-Forest-Filters (*RFF*) trained with the full and reduced feature sets and the three experimental settings *AllI*, *PhyI* and *GenI* using a 10-fold cross validation are shown. A simple interaction filter using the harmonic sequence similarity between the orthologs and a filter based on the InParanoid ortholog bootstrap score are evaluated as baseline comparisons.

The *RFF* trained with the full feature set in the *AllI* setting achieves the highest $AUPRC$ score of 0.86 and an $AUPRC$ score of 0.82 with the reduced feature set. When in addition to the occurrence of an interaction also an interaction type agreement is required, the performance drops significantly. Physical interactions can be classified with an $AUPRC$ score of 0.60 and of 0.58 with the *RFF* trained with the full and reduced feature set, respectively. Genetic interactions can be classified with $AUPRC$ score of 0.60 and 0.48.

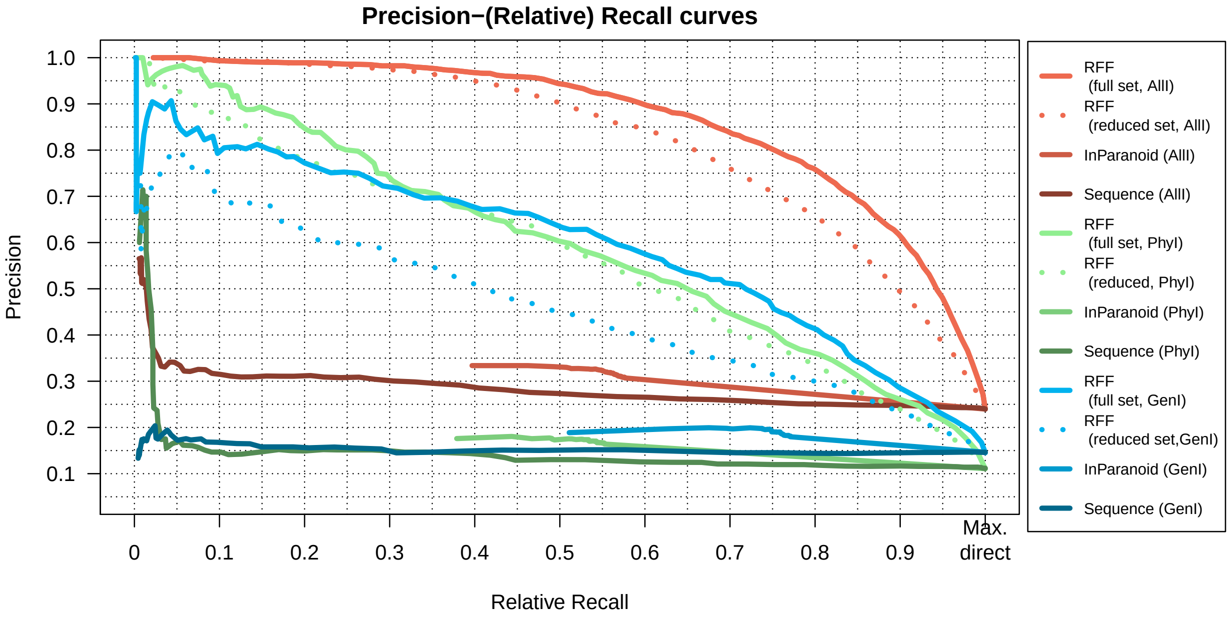


Figure 3.5. Precision - (relative)Recall curves for the *RFF* (Random-Forest-Filter) trained with the reduced feature set, the full feature set and different experimental setting (only physical interactions (*PhyI*), only genetic interactions (*GenI*) and all interactions (*AllI*)) using 10-fold cross validation. Interactions are transferred from all eukaryotic species with interaction data to *S. cerevisiae* and filtered with the respective approaches. In addition, the precision and relative recall is given for a simple sequence similarity filter and a filter based on the InParanoid ortholog bootstrap score. The *RFF* for *AllI* trained with the full (red) and reduced (red dotted) feature set perform best. The reduced feature set performs somewhat worse than the full feature set. For the more strict *PhyI* and *GenI* settings in which also the type of an interaction is transferred, the performance drops in comparison to *AllI*. By comparing the different feature sets it can be seen that for physical interactions (green, green dotted) almost the same performance for the full and reduced feature set can be reached, whereas for genetic interactions (blue, blue dotted) a clear difference in the performance can be observed. But again, for these two settings a huge improvement of *RFF* to the baseline filters based on sequence similarity and ortholog scores can be observed.

Using a maximum InParanoid ortholog bootstrap score of 1.00, a transfer precision of 0.33 for *AllI* can be reached resulting in an $AUPRC$ of 0.30. For physical and genetic interactions the precision of a direct transfer can barely be improved resulting in an $AUPRC$ of 0.15 and 0.18, respectively.

A high threshold has to be used for the sequence similarity filter in order to increase the transfer precision resulting in a low $AUPRC$ score of 0.28 for *AllI*. Even lower are the $AUPRC$s for *PhyI* and *GenI*. This can be explained with the low sequence similarities of the orthologs used for the transfer, which ranges between 33 % and 38 % on average for the different species. For the full feature set the *RFF* for *AllI* yields a precision of 0.85 and a relative recall of 0.69 (regular recall of 0.02) with a typical score threshold of 0.5. With the same score threshold for *PhyI* a precision of 0.72 and a relative recall of 0.33 (regular recall of 0.01) can be reached, whereas for *GenI* a slightly lower precision of 0.68, but a higher relative recall of 0.40 is observed (0.003 regular recall) .

In general, the predictors for *AllI* achieve a better performance than the predictors for the more strict setting in which also the interaction type has been transferred and predicted. This is plausible as for *AllI* the gold standard is larger and as with a direct transfer a consistency of 25 % can be reached already. For the different feature sets (full and reduced) a small drop in the *AllI* and *PhyI* setting and a large drop for the *GenI* setting is observed.

In the following we show examples of transferred physical interactions which receive high and low score values by *RFF*. On one hand, the transferred interaction between LST8 and TOR2 from WAT1 and TOR2 (in *S. pombe*) and also the transferred interaction between SMX3 and LSM5 from SmF and CG6610 (in *D. melanogaster*) get a comparable high score of ≥ 0.90. For the first, but not for the second example also an interaction is known between the orthologs in *S. cerevisiae*. But for the second example, both orthologs (SMX3 and LSM5) carry the Sm domain and the interaction between orthologs of SmF and CG6610 have been found in *S. pombe* and *H. sapiens*, which suggests that SMX3 and LSM5 indeed interact, but that they are not included in the *S. cerevisiae* gold standard. On the other hand, the transferred interaction between CRZ1 and HAT2 from Sp3 and RBBP4 (in *H. sapiens*, identified within a low-throughput study (Zhang and Dufau, 2002)) and the transferred interaction between ARP6 and RPS1A from Actr13E and RpS3A (in *D. melanogaster* which was identified in a Yeast Two Hybrid screen (Uetz and Pankratz, 2004)) gets a score of $\leq$ 0.05. Both transferred interactions are not in the *S. cerevisiae* gold standard, therefore, they are filtered correctly. Due to the low-throughput experiment, which was used to discover the interaction between Sp3 and RBBP4 it can be assumed that this interaction indeed exists for *H. sapiens*, but not in *S. cerevisiae*. In contrast, the interaction between Actr13E and RpS3A could also be false positive due to the high-throughput Yeast Two Hybrid screen which was used to identify the interaction in *D. melanogaster*. In Figure 3.6 the transferred interactions together with their assigned *RFF* scores and their feature values in comparison to the feature distributions of correctly and incorrectly transferred interactions are shown.

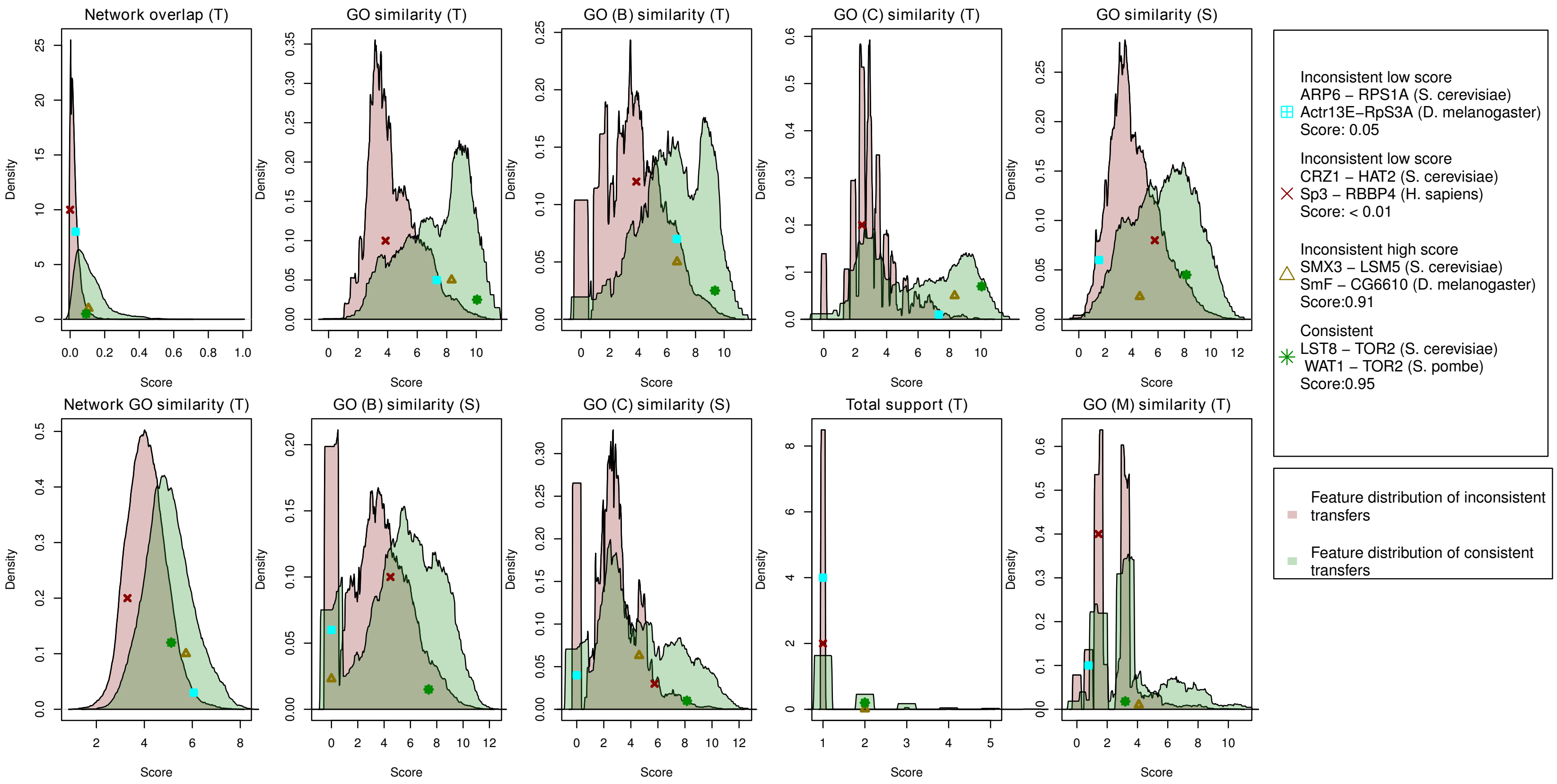


Figure 3.6. Four examples of transferred interactions which get high and low scores by *RFFs* are shown including their individual feature values for the most important transfer features in comparison to the overall feature distribution of conserved (consistent transfers) and not conserved (inconsistent transfers) interactions (see discussion in the main text).

Feature impact

To estimate the contribution of each feature to the performance of *RFF*, the Information Gain (IG) is computed for the different experimental settings (Figure 3.7 d). The IG for the different features differs among the experimental settings, but the sorting of the features according to their IG value is similar. The strongest feature is the network overlap in the target network (Network overlap (T)). But also the GO features yield a high IG. The combined GO features have higher IGs than the category-wise GO features for biological processes, cellular components and molecular functions. This can be explained by the fact that more GO terms are considered for the global semantic GO similarity, so less often a missing value indicator is assigned. From the individual GO term types, the biological processes category has the highest IG. Biological processes have also been identified by De Bodt et al. (2009) as a strong feature to define thresholds for an interaction transfer filter. From the Ortholog features the synonym similarity (token score) and the ortholog score feature contributes most to the prediction.

In contrast, the gene expression correlation, which was used in other studies for the prediction of protein interaction, has a rather low IG. For the two features with highest IG (Network overlap and GO similarity in the target network) also the score distributions of correctly and incorrectly transferred interactions for *AllI* are shown in Figure 3.7. Clearly, the fraction of correctly to incorrectly transferred interactions increases with the score value for these two features. For a feature like the harmonic sequence similarity, which has a low IG, only a small difference in the characteristics of the distribution can be observed, which explains the weak performance of filters based on sequence similarity.

In Table 3.4 the performance of the different individual feature sets (Network, Gene Ontology, General, Ortholog, Reduced set and Full set) is summarized in addition to the filters based on the sequence similarity and the InParanoid bootstrap score. For the GO features the highest feature category-wise $AUPRC$ score can be reached for *AllI* and *PhyI*. For *PhyI* a similar $AUPRC$ score can be achieved with the ortholog features. Using a combination of all introduced features an up to 0.08 higher $AUPRC$ score can be obtained for the different settings. For *GenI* the highest category-wise $AUPRC$ score can be reached with the network features, which is also higher than the score for the reduced feature set. This explains the performance drop for the reduced feature set for this *GenI* setting.

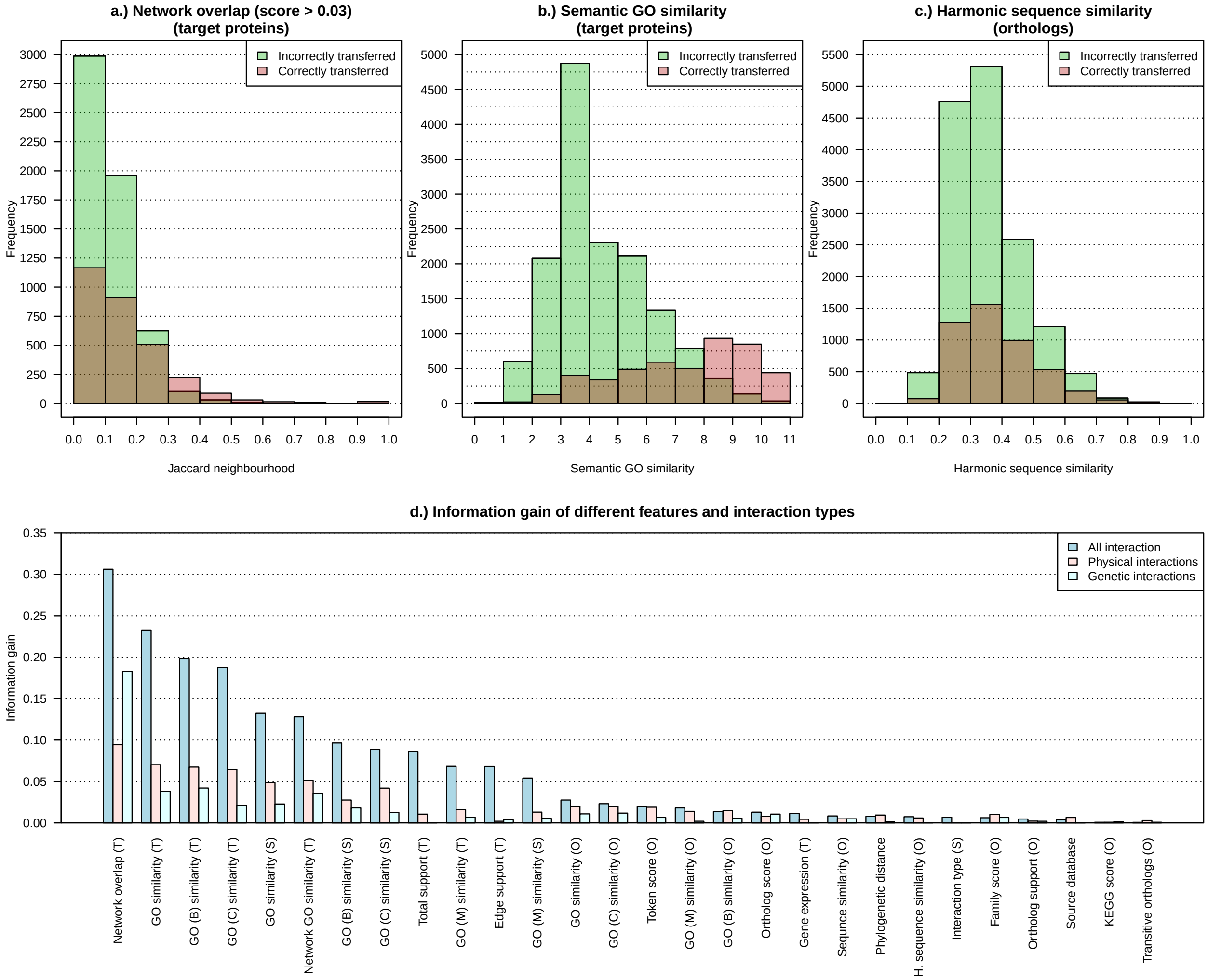
Generalizability

A general transfer approach should be able to achieve a similar performance for the interaction transfer to other species. Since the interaction networks for other species are currently too sparse (see Table 3.1) *RFFs* can not be learned and evaluated for individual species except for *S. cerevisiae*. Therefore, we investigate the applicability of the *RFF* fitted for the interaction transfer to *S. cerevisiae* for the transfer of interactions to other eukaryotic species. It has to be expected:

1. that the *RFF* scores transferred interactions between phylogenetically closer species higher than transferred interactions between phylogenetically distant species,

Table 3.4. Performance of the interaction transfer to *S. cerevisiae* with: the *RFF* (Random-Forest-Filter) trained with different feature sets using a 10-fold cross validation, InParanoid ortholog filter and the sequence similarity filter for different experimental settings. Interactions are transferred from all eukaryotic species with interaction data to *S. cerevisiae.* For each experimental setting and feature set the area under precision recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) are computed. From the individual feature sets the *RFF* trained with the GO and Network feature set perform best for the *AllI* and *GenI* setting. Whereas for physical interactions the performance for the Network features are lower than for the GO and Ortholog feature set.

Method	Experimental setting	Feature set	AUPRC	AUROC
RFF	All	Full	0.86	0.94
RFF	All	Reduced	0.82	0.91
RFF	All	Network	0.79	0.90
RFF	All	GO	0.79	0.89
RFF	All	Ortholog	0.62	0.82
RFF	All	General	0.50	0.68
InParanoid	All	-	0.30	0.59
Sequence	All	-	0.28	0.55
RFF	Physical	Full	0.60	0.89
RFF	Physical	Reduced	0.58	0.88
RFF	Physical	GO	0.50	0.85
RFF	Physical	Network	0.42	0.84
RFF	Physical	Ortholog	0.48	0.82
RFF	Physical	General	0.19	0.62
InParanoid	Physical	-	0.15	0.61
Sequence	Physical	-	0.14	0.55
RFF	Genetic	Full	0.60	0.87
RFF	Genetic	Reduced	0.47	0.82
RFF	Genetic	Network	0.51	0.86
RFF	Genetic	GO	0.45	0.80
RFF	Genetic	Ortholog	0.35	0.75
RFF	Genetic	General	0.19	0.53
InParanoid	Genetic	-	0.18	0.60
Sequence	Genetic	-	0.15	0.51


a.) Network overlap (score > 0.03) (target proteins)

b.) Semantic GO similarity (target proteins)

c.) Harmonic sequence similarity (orthologs)

d.) Information gain of different features and interaction types

Figure 3.7. Histogram of the score values for correctly (red) and incorrectly (green) transferred interactions (without interaction type) for the features: **a.)** Network overlap, **b.)** Semantic GO similarity and **c.)** Harmonic sequence similarity. **d.)** Information Gain of the individual features and experimental settings. For ortholog protein features the average Information Gain of the two orthologous partners is shown. For the features **a.)** Network overlap and especially for **b.)** Semantic GO similarity a different distribution for correctly and incorrectly transferred interactions can be observed resulting in a large Information Gain of these features. In contrast, for the harmonic sequence similarity feature only a small difference in the distributions can be observed, which explains the small Information Gain and the filter performance based only on sequence similarity.

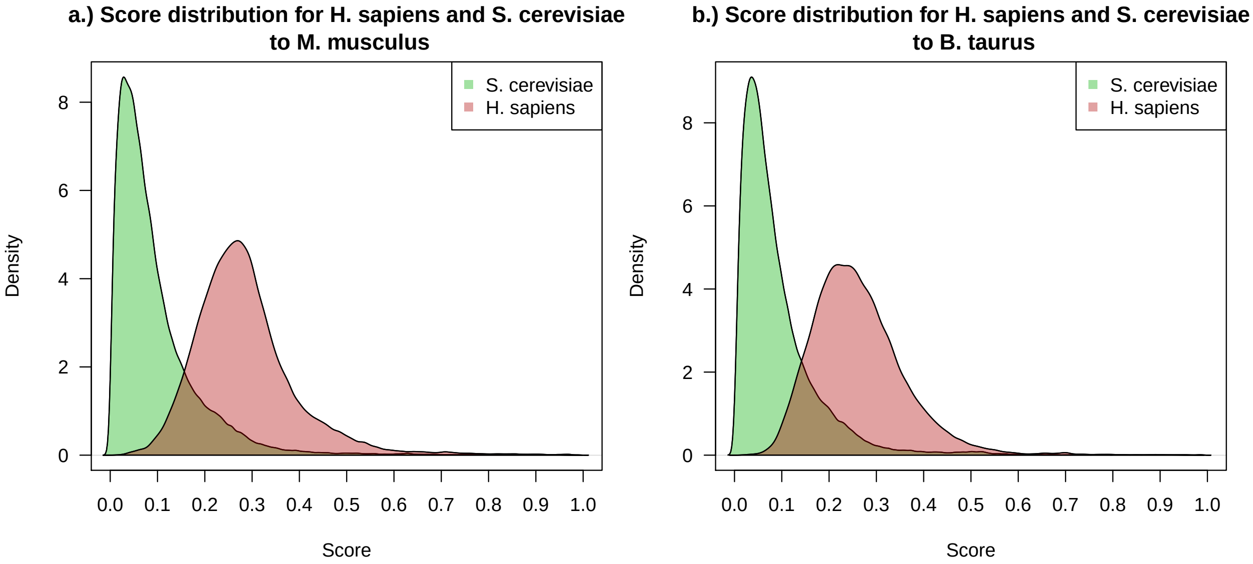


Figure 3.8. Score distributions for transferred interactions with *RFF* from *S. cerevisiae* and *H. sapiens* to the two target species (**a.**) *M. musculus* and (**b.**) *B. taurus*. Transferred interactions from *S. cerevisiae* get significantly lower score values than transferred interactions from *H. sapiens* to both target species. With a low score threshold of 0.2 almost all interactions from *H. sapiens* are transferred to the two target species, whereas a huge fraction of the transferred *S. cerevisiae* interactions is filtered out.

2. that according to their importance, the ranking of features is similar for the interaction transfer to different species even though the networks are to incomplete to train a model and

3. that a comparable performance with competing transfer approaches should be achieved when the *RFF* is applied for the transfer of interactions to other species.

In the following we investigate these three points.

Transfer scores: We use the *RFF* with the reduced feature set trained with transferred interactions to *S. cerevisiae* to transfer protein interactions from the two largest interaction networks *H. sapiens* and *S. cerevisiae* to both *M. musculus* and *B. taurus* and analysed the score distributions. For physical and genetic interactions in the source network, the predictor trained with the respective interaction type (*PhyI* and *GenI*) is used and for interactions with a different type the predictor trained with all data is applied (*AllI*). As expected, the scores for transferred interactions from the phylogenetically closer species, in this case *H. sapiens*, are higher than the scores from the more distant species as shown in Figure 3.8. The score distribution of transferred interactions from *S. cerevisiae* to *M. musculus* and *B. taurus* are very similar with a median score of 0.07 for both distributions. This is comparable to the transfer of interactions to *S. cerevisiae*, where a median score between 0.03

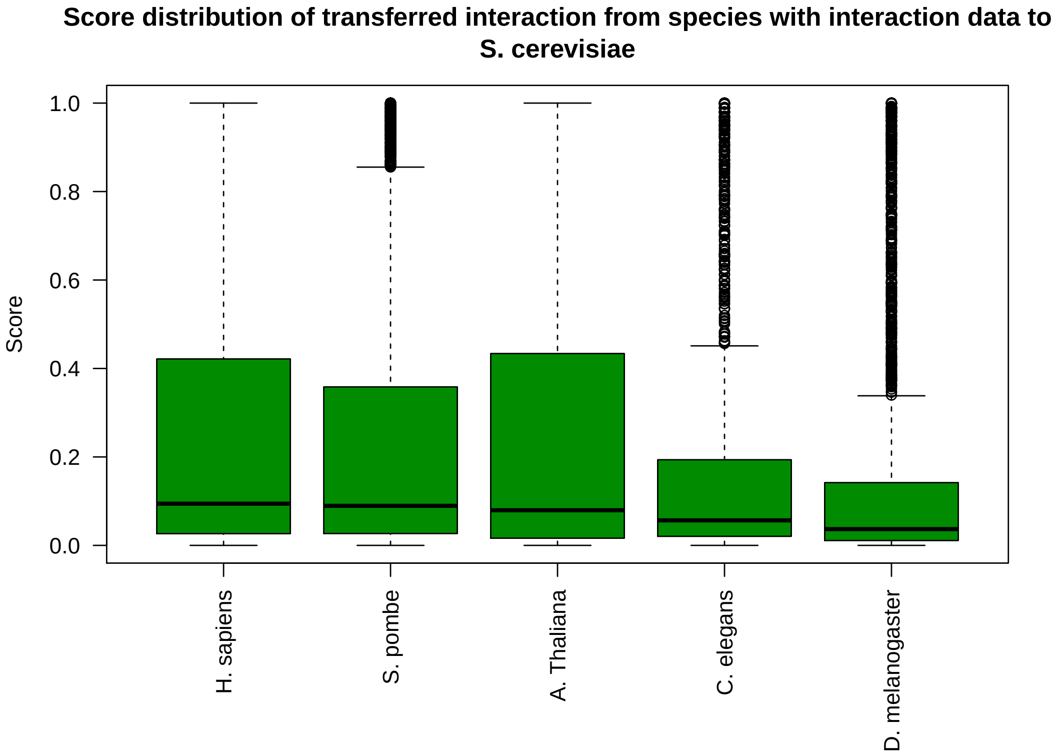


Figure 3.9. The average transfer scores for an interaction transfer from *M. muscles*, *H. sapiens*, *S. pombe*, *A. thaliana*, *C. elegans* and *D. melanogaster* to *S. cerevisiae* using RFFs in a cross-validation setting (see discussion in the main text).

and 0.09 can be observed (see Figure 3.9). In comparison, for the transfer of interactions between phylogenetically closer species, a median score of 0.27 and 0.25 can be observed for the transfer of interactions from *H. sapiens* to *M. musculus* and *B. taurus*, respectively. Thus, as expected with higher score thresholds more interactions can be transferred from *H. sapiens* to *M. musculus* as compared to *H. sapiens* to *B. taurus*. On the other hand, from *S. cerevisiae* almost the same number of interactions is transferred to the two species *B. taurus* and *M. musculus* with different score thresholds.

Cross-species feature importances ranking: We transfer all available interactions to *S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *S. pombe* and *C. elegans* and compute the Information Gain (IG) of each modeled feature given the observed consistently and inconsistently transferred interactions for the respective species. We observe that the similarity of the feature ranking decreases with the IG i.e. that those features which are important for the classifier are consistently ranked high and the ranking of those feature which are not that beneficial to our classifier differ more. For example the Network overlap feature is ranked first for all considered species expect for *C. elegans*. Also the feature which models the GO

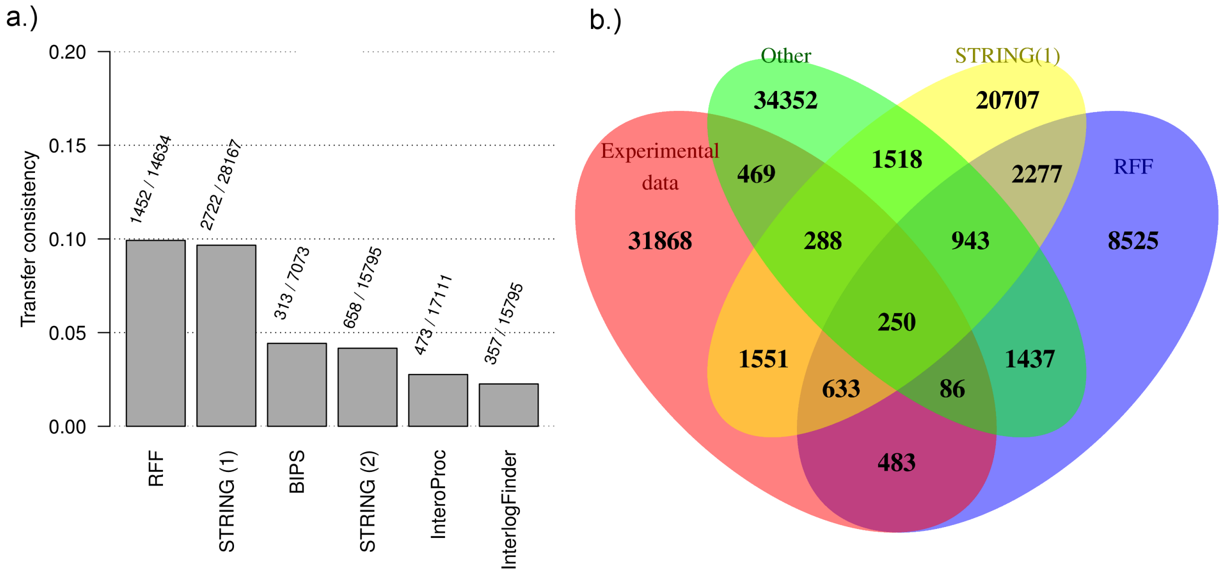


Figure 3.10. Comparison of interaction transfer sets from various methods for *H. sapiens* with known *H. sapiens* interactions from iRefIndex. We compare interaction sets from *STRING* (Szklarczyk et al., 2011), *InteroPORC* (Michaut et al., 2008), *InterologFinder* (Wiles et al., 2010), *BIPS:BIANA* (Garcia-Garcia et al., 2012) and our Random-Forest-Filter (*RFF*). From the *STRING* database only interactions with interaction transfer information from other species and a combined score over 0.7 are included (*STRING(1)*). The combined score uses information from all information sources including knowledge on experimental interactions for the respective species (direct evidence). Therefore, an additional interaction set is created where the combined *STRING* score is recomputed excluding the scores from the direct evidence of databases, experiments and text-mining (*STRING(2)*). In general, the intersections between the different sets and the known interactions are small. **a.)** With the *RFF* and with *STRING(1)* 10 % of the predicted interactions can be found in the experimental data. The modified *STRING(2)* interaction set is 43 % smaller and only 4 % of the predicted interactions are consistent with the experimental data showing a clear performance advantage of the *RFF* for species with no experimentally determined interactions. **b.)** We compare the interaction sets of *RFF*, *STRING(1)*, a combined set of unique interactions from *InteroPORC*, *InterologFinder* and *BIPS:BIANA* and a set of known *H. sapiens* interactions. With the *RFF* 42 % of predicted interactions can also be found in one of the other sets.

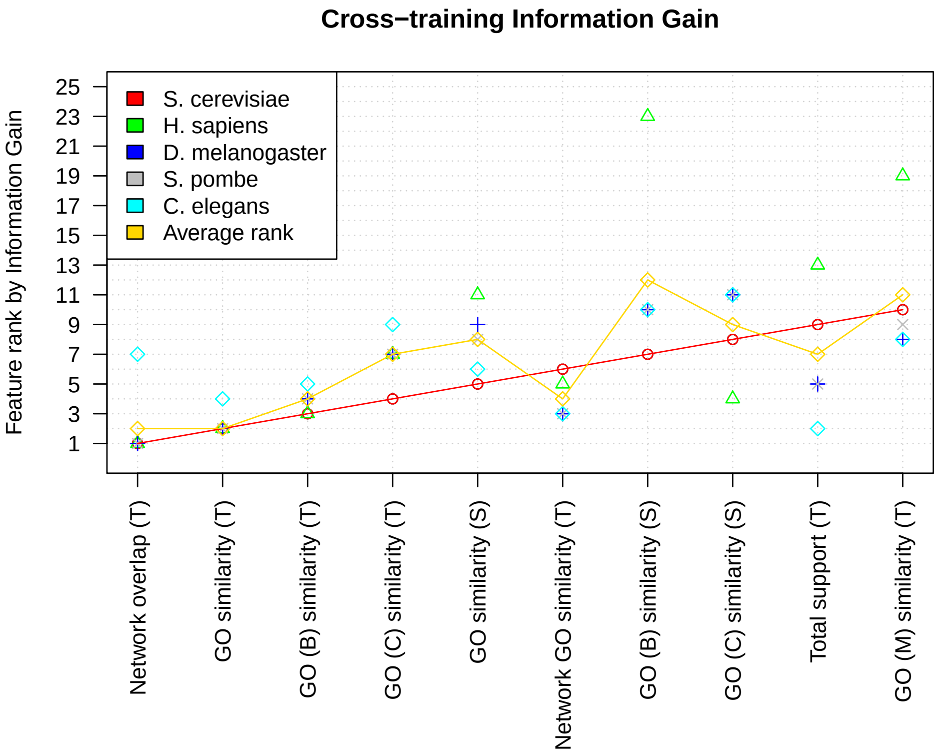


Figure 3.11. The features employed for the protein interaction transfer are ranked according their information gain for different experimental settings: transfer of interactions to *S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *S. pombe* and *C. elegans*. The information gain ranking (importance ranking) is quite similar especially for the most important features, whereas the ranking of the less important feature varies more.

Similarity between the target interactions is ranked second by all considered species expect for *C. elegans*. In Figure 3.11 we show the ranking of the features according to their IG for the different species. As reference we use the ten features with highest IG for the interaction transfer to *S. cerevisiae*.

Comparison with other interaction transfer methods: Most protein interaction transfer methods predict interologs for *H. sapiens* and, in addition, quite many experimentally validated interactions are available for human. Therefore, this network is chosen to evaluate the intersections of predicted protein interactions from different data sets and a set of experimentally discovered physical protein interactions.

Transferred interactions from *InteroPORC* (Michaut et al., 2008), the *STRING* database (Szklarczyk et al., 2011), *InterologFinder* (Wiles et al., 2010), *BIPS:BIANA* (Garcia-Garcia et al., 2012) and interactions predicted with *RFF* are used for the comparison. In order to compare the sets, the protein identifiers are mapped to UniProt/Swissprot identifiers. The following prediction sets are constructed using the publicly available transferred networks from the considered approaches for *H. sapiens*:

STRING(1) high confidence interactions with at least one evidence of an interaction transfer from another species (interactions with a combined score below 0.7 are excluded);

STRING(2) The combined score of *STRING* incorporates evidence from many sources including experimental knowledge for the respective species (direct evidence). Thus, transferred interactions with also direct evidence are scored higher, which biases the *STRING* set for this comparison. Therefore, an additional *STRING* interaction set is created where the combined score is recomputed without the scores for the direct evidence from databases, experiments and text-mining using the equation for the combined score (Szklarczyk et al., 2011). Again for this set a combined score threshold of 0.7 is used to filter interactions.

InteroPORC all transferred interactions;

InterologFinder 15,795 transferred interactions with highest score (the score threshold is set so that the same number of interaction as in the *STRING(2)* set are predicted);

BIPS:BIANA all transferred interactions in the online available precomputed prediction set with domain interactions or shared GO terms;

RFF The *RFF* for physical source interactions (*PhyI*) trained with the reduced feature set and with transferred interactions to *S. cerevisiae* is used. All transferred interactions with a score ≥ 0.18 for the transfer to *H. sapiens* from all species considered in the study are used. The score value is experimentally chosen to yield roughly the same number of transferred interactions as *STRING(2)*.

In the entire *InteroPORC* prediction set 17,111 physical interactions and in the selected *BIPS:BIANA* set 7,073 interactions are included. With 28,155 links between proteins the interaction set from the *STRING(1)* is the largest, the *STRING(2)* set is only slightly larger (15,795 interactions) than the set from *RFF* which includes 14,634 predicted physical interactions. 35,628 experimentally validated physical interactions are taken from iRefIndex (7,784 interactions are excluded because the proteins are only mappable to UniProt/TrEMBL).

In Figure 3.10 the consistency with experimentally validated interactions (**a.**) and the intersections between different *H. sapiens* protein interaction sets are shown (**b.**).

In general, the intersections between the sets are small. The highest consistency of 10 % between the predicted interaction sets and the experimental set can be reached with the *RFF* and with the *STRING(1)* interaction set.

From the *STRING(2)* and *BIPS:BIANA* interaction set 4 % and from the *InteroPORC* and *InterologFinder* around 3 % of the predicted interactions are consistent with the experimental data. In total 42 % of predicted interactions with the *RFF* can be found in at least one other set whereas for the *STRING(1)* set only 26 %, for the *BIPS:BIANA* set 18 %, for the *InteroPORC* set 17 % and for the *InterologFinder* set 10 % can be found in another interaction set. Besides *BIPS:BIANA* all methods transfer interactions from all available interactions in public available databases. But *BIPS:BIANA* explicitly excludes interactions from Tandem Affinity Purification experiments which explains the rather small interaction set. In comparison to *STRING(2)*, *BIPS:BIANA*, *InteroPORC* and *InterologFinder* a clear performance gain of our *RFF* approach can be observed. Furthermore, *RFF* cannot be outperformed by *STRING(1)* even with the integration of experimental knowledge (which is not available for most species) via the combined score. Thus, for species without experimental knowledge but also for model organisms with experimental protein interactions a performance advantage of our approach in comparison to *STRING* can be expected.

3.3.3 Enriched Protein Interaction Networks

As shown above via the comparison with other state-of-the-art methods our *RFF* approach has a decent performance for the transfer of interactions to species without experimental interaction data. Therefore, we use our approach to obtain as comprehensive as possible interaction networks for various eukaryotic species. For this we use all available experimental interaction data for all 83 eukaryotic species for the transfer to all other eukaryotic species whenever ortholog mappings of appropriate quality are available. We employ three *RFFs* trained on *S. cerevisiae*: *RFF, PhyI* for physical source interactions, *RFF, GenI* for genetic source interactions and *RFF, AllI* for interactions for the remaining interactions including interactions without annotated interaction type. The same score threshold of 0.18 is used for all models.

With a *direct* interaction transfer the interactome of 83 eukaryotic species can be extended from currently 321,808 interactions to 5,751,775 interactions. With the *RFF* 1,248,609 pairwise interactions can be transferred (i.e. more than 78,% of transferred interactions are filtered out as possible false positive). An overview of the resulting interactomes is shown in Figure 3.12 using the Interactive Tree Of Life (Letunic and Bork, 2011) (only species are shown for which at least 50 % of the genes have associated GO annotations). For higher vertebrates of interest such as the farm animals *B. taurus*, *M. musculus* and *G. gallus* each interaction network can be enriched with over 40,000 interactions. After that, the resulting interactomes have a decent coverage of more than 2 interactions per gene on average. Still, with our method for some species only few interactions can be transferred. Examples are plants like *O. sativa* or *V. vinifera* with an average of 0.35 interactions per gene. The reason for the low coverage in these cases is the small number of available orthologs in the ortholog databases.

It is clear that for the large scale interaction transfer with our *RFF* method the limitations are: (i) the availability of ortholog relations, (ii) the mappings of the orthologs to UniProt entries, (iii) and the annotations of the UniProt entries. This implies that for some species only few interactions can be transferred. Of course, *RFF* will profit from the expected improvements of protein annotations, ortholog mappings and further experimental protein interactions.

The transferred interaction networks for the 83 species are available on our web service and can be inspected and downloaded. The user can specify score thresholds corresponding

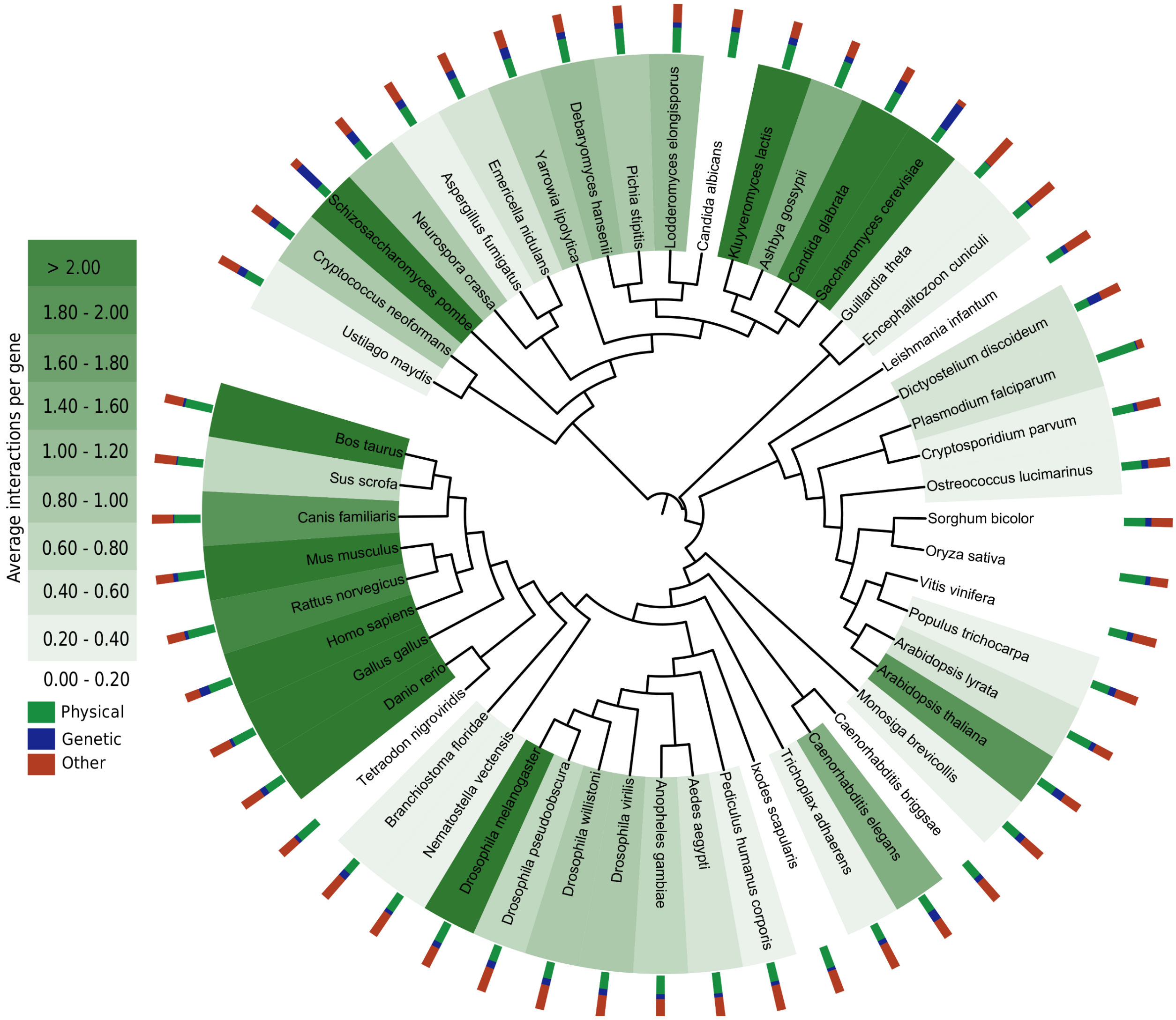
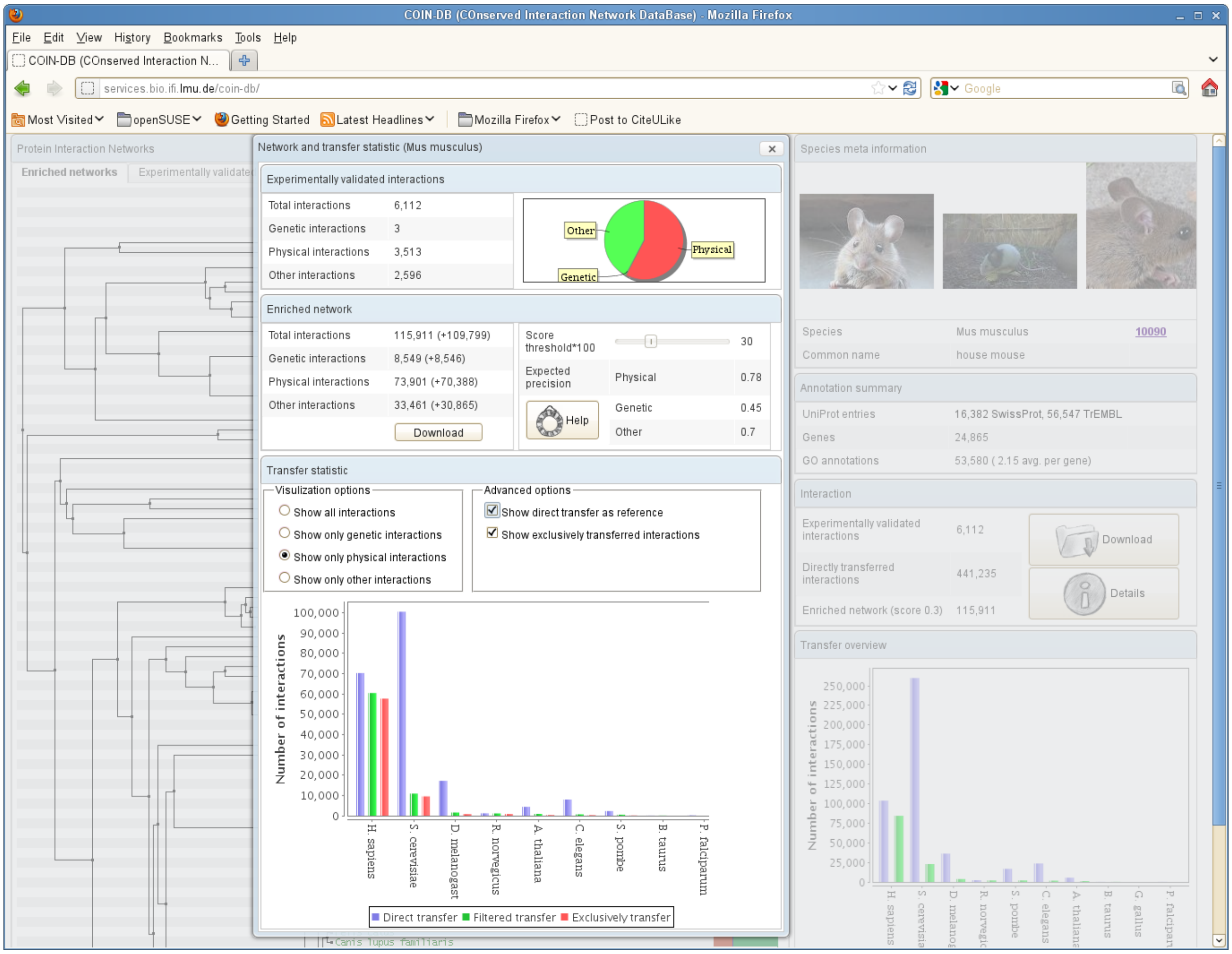


Figure 3.12. The interactomes of 83 eukaryotic species can be increased from currently 321,808 interactions to 1,076,996 pair-wise interactions using the *RFFs* with reduced feature set with a score threshold of 0.18. In the figure the enriched protein interactomes are shown for all species where at least GO annotations for half of the genes are available. Interactions are transferred from all eukaryotic species to all other species with available ortholog mappings. The color of the species nodes indicates the average number of interactions per gene and the associated bar chart indicates the fraction of physical interactions (green), genetic interactions (blue) and other interaction types (red) in the enriched interaction networks for the respective species. For species with rich annotation information including *M. musculus* and *B. taurus* over 40,000 interactions can be transferred resulting in an average number of interactions per gene larger than 2. For species with sparse annotation information and few ortholog references only a small number of interactions can be transferred. For example for the plants *O. sativa* and *V. vinifera* only 0.35 interactions per gene on average can be obtained.



Network and transfer statistic (Mus musculus)

Experimentally validated interactions

Total interactions	6,112
Genetic interactions	3
Physical interactions	3,513
Other interactions	2,596

Enriched network

Total interactions	115,911 (+109,799)
Genetic interactions	8,549 (+8,546)
Physical interactions	73,901 (+70,388)
Other interactions	33,461 (+30,865)

Score threshold*100		30
Expected precision	Physical	0.78
	Genetic	0.45
	Other	0.7

Transfer statistic

Visulization options
○ Show all interactions
○ Show only genetic interactions
● Show only physical interactions
○ Show only other interactions

Advanced options
☑ Show direct transfer as reference
☑ Show exclusively transferred interactions

🟦 Direct transfer 🟩 Filtered transfer 🟥 Exclusively transfer

Species meta information

Species	Mus musculus	10090
Common name	house mouse	

Annotation summary

UniProt entries	16,382 SwissProt, 56,547 TrEMBL
Genes	24,865
GO annotations	53,580 (2.15 avg. per gene)

Interaction

Experimentally validated interactions	6,112
Directly transferred interactions	441,235
Enriched network (score 0.3)	115,911

Transfer overview

Figure 3.13. Screenshot of the web interface for the transferred and scored protein interactions. Transferred and experimentally validated interactions can be downloaded for 83 eukaryotic species for user specified score thresholds. For species of interest the transfer profiles can be inspected in detail including the number of interactions for the different interaction types, the number of uniquely transferred interactions, and the expected performance of the transfer.

to the expected transfer precision of our models. In Figure 3.13 the web interface including the 'transfer statistics view' for *M. musculus* is outlined as an example.

3.4 Discussion and Conclusion

Years after high-throughput screening techniques for the identification of protein interactions were introduced most interaction data still is available for only a few model organisms, in particular for *S. cerevisiae*. Transferring protein interactions works best between phylogenetically close species, but already between the two yeast species *S. pombe* and *S. cerevisiae* only a consistency of 36 % for transferred physical interactions can be observed. The transfer consistency between more distant species is of course much lower. The transfer consistency is also lower for genetic interactions between the two yeasts, which might be due to the incompleteness of the *S. cerevisiae* genetic interaction network.

We observed that for only 3 % of the *S. cerevisiae* interactions evidence of conservation between orthologs in different species could be found. In order to improve the transfer quality and to be able to also consider interactions from phylogenetically distant species, e.g. from *S. cerevisiae* to *M. musculus*, we introduced a new method using Random Forests (Random-Forest-Filter *RFF*) to score and filter transferred interactions.

We trained the models with transferred interaction data from eukaryotic organisms to *S. cerevisiae*. We did the training on yeast, as the *S. cerevisiae* network is currently the largest eukaryotic interaction network and for most of the proteins in the network curated functional annotations are available. We evaluated the models with different feature sets and experimental settings and compared the models with commonly applied filter approaches e.g. using the sequence similarity and the InParanoid bootstrap score. We showed that for the task of transferring interactions to *S. cerevisiae* our approach performs better than commonly applied filter approaches. Based on these results we assume that the performance of the transfer to *S. cerevisiae* is a lower bound for the performance of the method for the transfer between phylogenetically closer species.

But still, our observed results are limited with respect to different aspects:

1. Possible false negatives in the *S. cerevisiae* network result in lower transfer consistencies, whereas false positives in the *S. cerevisiae* network may result in an overestimation of the consistency.

2. Our method makes use of interaction data from various sources like Yeast two Hybrid, or Tandem Affinity Purification and thus includes measured-binary and measured-predicted binary interactions. We only address the interaction transfer on a general level and currently only consider binary-interactions. Our method will benefit from further discrimination of protein interactions e.g. discrimination between transient or permanent protein interactions, or the pre-identification of conserved protein complexes. And thus, stronger claims on the conservation rate and also a more complete interaction transfer will be possible.

3. Low-throughput experiments are commonly hypothesis-driven (Sambourg and Thierry-Mieg, 2010) and involve proteins of particular interest to the researcher performing the experiments. These low-throughput experiments can also be based on the observation that a conservation in a particular species exists, which could lead to an overestimation of the consistency and to overfitting.

4. The ortholog and protein annotations quality have a direct influence on our models. For example KEGG pathway information, or gene ontology and synonym annotations are themselves often inferred using homology information (directly or indirectly). For

example the KEGG databases transfers pathway information from well studied species based on manually defined ortholog groups. It is obvious that with solely transferred annotations our approach can not improve the prediction performance.

5. We fitted our model for the transfer to *S. cerevisiae* only. Due to these reasons, we can not give an accurate estimation on the performance for the protein interaction transfer to species except for *S. cerevisiae.*

But we could show that our approach can be applied for the transfer of interactions to species beyond *S. cerevisiae* as well. On one hand, we tested the generalizability of *RFF* with transferred interactions to *H. sapiens*, *M. musculus* and *B. taurus.* We showed that (as expected) transferred interactions from phylogenetically closer species get higher scores than transferred interactions from phylogenetic more distant species. Furthermore, we showed that those features which are most beneficial for the classification of interaction for the transfer to *S. cerevisiae* are also most beneficial for the classification of interactions for other species. On the other hand, we compared different protein interaction approaches. We showed for *H. sapiens* that with our approach the highest consistency of transferred interactions can be observed and that 42 % of transferred interactions can be explained with high confidence relations extracted from *STRING*, *InteroPORC*, *InterologFinder*, *BIPS:BIANA* or the available experimental interactions. Furthermore, in an experimental setting where we recomputed the *STRING* combined edge score for *H. sapiens* to mimic a species without experimental knowledge, we showed that *RFF* predicts almost the same number of interaction as *STRING*, but with our approach more than twice as many interactions are consistent with the available experimental protein interaction network.

Chapter 4

Regulatory Network Transfer and Conservation

Abstract: Transcription factors play a fundamental role in cellular regulation by binding to promoter regions of target genes in order to control their gene expression. Transcription factors - target gene networks are widely used as representations of regulatory mechanisms, e.g. for modeling the cellular response to input signals and perturbations.

As the experimental identification of regulatory interactions is time consuming and expensive, one tries to use knowledge from related species when studying an organism of interest. Here, we present ConReg, an interactive web application to store regulatory relations for various species and to investigate their level of conservation in related species. Currently, ConReg contains data for eight model organisms. The regulatory relations stored in publicly available databases cover only a small fraction both of the actual interactions and also of the regulatory relations described in the scientific literature. Therefore, we included regulatory relations extracted from PubMed and PubMedCentral using sophisticated text-mining approaches and from binding site predictions.

We applied ConReg for the investigation of conserved regulatory motifs in *D. melanogaster*. From the 471 regulatory relations in *REDfly* our system was able to identify 66 confirmed conserved regulations in at least one vertebrate model organism (*H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*). The conserved network contains among others the well studied motifs for eye-development and the pan-bilaterian kernel for heart specification.

ConReg is publicly available and can be used to analyze and visualize regulatory networks and their conservation among eight model organisms. It also provides direct links to annotations including literature references to possible conserved regulatory relations.

Publication: ConReg is published in the OpenAccess Series in Informatics (OASIcs) (Pesch et al., 2012) and has been presented at the German Conference on Bioinformatics 2012 in Jena. Here, I adapted the layout, restructured the sections and included Table 4.2 and Figure 4.6.

My contribution: I implemented the software systems, defined the networks, performed the conservation analysis and drafted the manuscripts.

Contribution of co-authors: Ralf Zimmer supervised the work and helped drafting the published manuscripts. Matthias Böck provided the transcription factor binding site predictions used for the ConReg system and helped drafting the manuscript.

4.1 Introduction

The physical regulatory relationships of an organism can be described by gene regulatory networks (GRNs). Transcription factors (TFs) and their respective targets (TGs) define the majority of these regulations. GRNs can describe systems on the scale of a few genes, a particular pathway or even on the whole set of available genes for an organism. The inference of these GRNs is generally done from experimental data sets, like gene expression data and additional prior information from available databases (Hecker et al., 2009). Even though more and more high-throughput methods for the identification of TF-TG relations have been recently developed, the data of experimentally validated relations is still sparse for higher multi-cellular organisms (Rottger et al., 2012). Therefore, transferring knowledge using orthologs from related species is typically done when studying an organism of interest. Several approaches have been already proposed which are capable of transferring physical protein-protein interactions even between phylogenetically distant species, like from *S. cerevisiae* to *A. thaliana* or *C. elegans* (Yu et al., 2004). The conservation of a regulatory relation requires, that at least the involved TF and the TG have to be conserved and that the TF binds to the promoter region of the TG in two or more organisms. Depending on the number of organisms in which the regulatory relation is conserved and the evolutionary distance, relations between different organisms can be transferred with a certain confidence (Baumbach, 2010; Sharma et al., 2011; Taher et al., 2011).

Different methods have already been proposed and used for the transfer of regulatory networks from one organism to another. A well-known example is KEGG, which transfers confirmed regulatory relations to (non-model) organisms based on ortholog definitions (Kanehisa et al., 2012). For bacteria more advanced approaches have been successfully applied, which additionally incorporate conserved information of the binding site (Baumbach, 2010) and subfamily classifications (Sharma et al., 2011). Similar approaches exist for eukaryotes, which also make use of conserved transcription factor binding sites (Taher et al., 2011). Taher et al. (2011) claimed that 88 % of the orthologs between *H. sapiens* and *D. rerio* retain their regulatory mechanisms. Nevertheless, it remains to be controversial to which extend regulatory relations can be directly transferred between organisms (Baumbach, 2010). There are some well-known regulatory motifs, which appear to be conserved among a group of quite distant species, supporting the transfer of conserved regulatory relations. A famous example is the conservation of regulatory relations for the development of the eye in *D. melanogaster* and vertebrates. It was shown that in *M. musculus* and other vertebrates *Pax-6*, the ortholog of the *eyeless (ey)* gene — one of the central TFs controlling the eye development in *D. melanogaster* — shares an extensive sequence identity and is even capable of inducing ectopic eyes in *D. melanogaster* (Wawersik and Maas, 2000). Also other motifs, like the pan-bilaterian kernel for heart specification (Davidson, 2006) or regulation of apoptosis regulation in *D. melanogaster* and vertebrates (Zhai et al., 2012) appear to be conserved. Studies revealing the similarity and the conservation of regulatory sub-networks have been conducted for different species as well, like *MAP kinase expression* in *C. elegans* and *H. sapiens* (Lee et al., 2007) or *Toll-like receptor 4* regulated genes (Schroder et al., 2012).

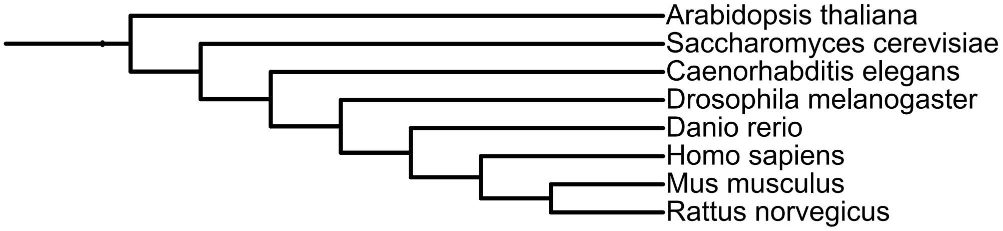


Figure 4.1. The eight species considered in our study and the associated taxonomic tree as extracted from NCBI. Currently, ConReg contains six animal model organisms (*H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*) as well as *S. cerevisiae* and the model plant *A. thaliana*.

In the following we present ConReg, an interactive web application to investigate regulatory relations as well as evidence for their conservation in other eukaryotic model organisms. For that purpose, we collected regulatory data for eight model organisms (*H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae*). The data was obtained from general and species-specific regulatory databases, from text-mining approaches applied to PubMed abstracts and PubMedCentral full text publications and from transcription factor binding site predictions (TFBS).

4.2 Materials and Methods

Data Sources

We collected regulatory relations for *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae* (see Figure 4.1 for the taxonomic tree of these species). Regulatory relations were extracted from the multi-species curated databases TRANSFAC (Version 9.3) (Matys et al., 2006) and ORegAnno (Griffith et al., 2008). Species-specific relations were extracted from YEASTRACT (Teixeira et al., 2006), REDfly (Gallo et al., 2011) and AtRegNet (Palaniswamy et al., 2006) and from curated pathways from Biocarta and NCI-Pathway (Schaefer et al., 2009). Transcription factors were collected from these regulatory databases and the DBD database (Kummerfeld and Teichmann, 2006). For the transfer of relations, we used orthologs from InParanoid (Remm et al., 2001), EnsemblCompara (Vilella et al., 2009) and OMA (Schneider et al., 2007). These databases were used due to their evaluation results in (Altenhoff et al., 2011; Hulsen et al., 2006) and the coverage of ortholog mappings for all considered species. All genes were mapped to ENSEMBL to obtain unique genomic locations and the associated annotations. Relations involving genes which could not be mapped were not considered.

4.2.1 Regulatory Relation Extraction from the Scientific Literature

Abstracts from PubMed (20,766,340 abstracts) and the corresponding full text publications from the PubMedCentral open access subset (389,322 documents) were used to search for

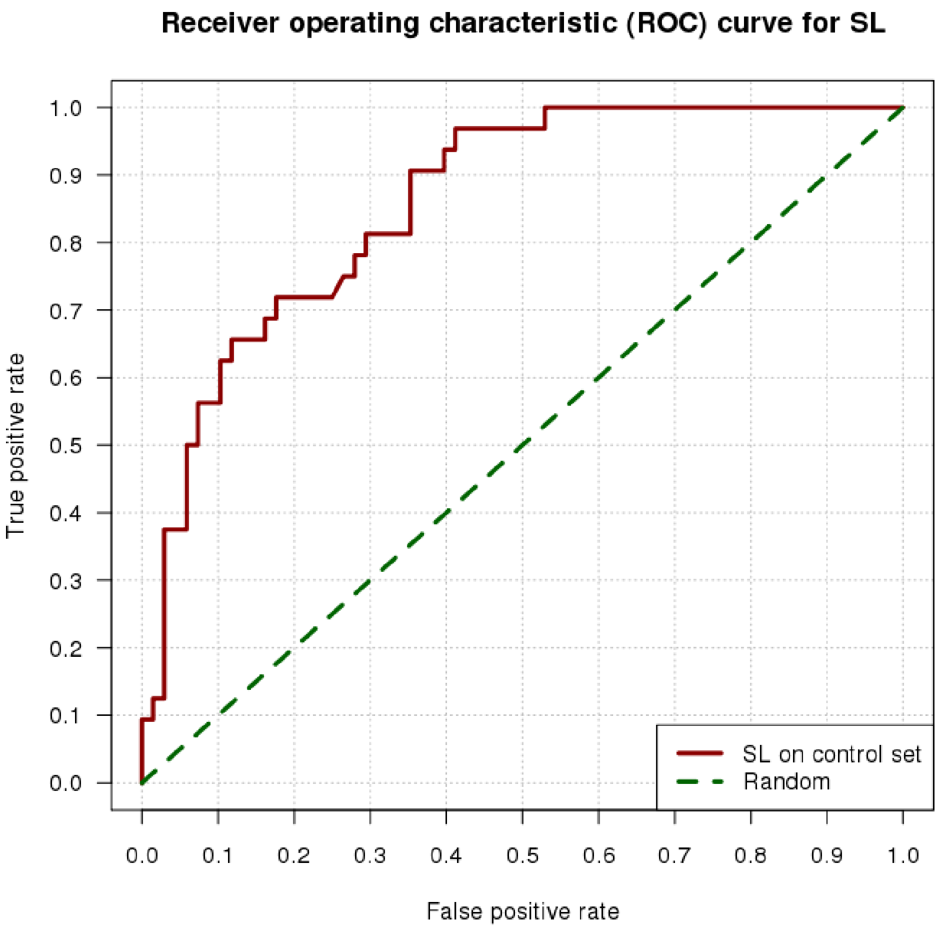


Figure 4.2. Receiver operating characteristic (ROC) curve for the final shallow linguistics (SL) SVM model which was used for the identification of regulatory relations. The evaluation set consists of 100 examples including 33 positive regulatory relations and 67 negative regulatory relations. On this control set the model could reach an area under the receiver operating characteristic (ROC) of 0.85 and an area under the precision-recall curve of 0.72. Furthermore, with a typical used probability threshold of 0.5 for the SVM, a precision of 0.56 and a recall of 0.75 could be reached.

regulatory relations in textual descriptions. In order to find relations in unstructured descriptions two tasks have to be accomplished: the named entity recognition (NER) of gene names and the correct identification of relations between genes. For example, consider the following sentence: *"There is evidence that the expression of Six3 is regulated by Pax6."*, (Manuel et al., 2008). To infer a regulatory relation, the gene names *Six3* and *Pax6* need to be found and the regulatory relation between *Pax6* → *Six3* has to be identified. We used syngrep (Csaba, 2008), a dictionary based NER tool, for the gene name recognition and the mapping of gene names to identifiers. Dictionaries were compiled by combining gene names, aliases and synonyms from UniProt, ENSEMBL, HGNC, MGI, RGD, Tair and FlyBase. Regulatory relations between genes were initially identified with a simple Tri-occurrence approach. For this approach, a relation was assumed between all pairs of genes which were found in a sentence, if a keyword indicating a regulatory relation was found and at least one gene is annotated as a TF. For this task, we defined a list of keywords, which are supposed to indicate regulatory interactions, like *regulates*, *represses*, or *down regulates*. Such a Tri-occurrence approach provides a good recall, but produces also many false positives. Therefore, we also used the following more sophisticated relation extraction approaches to filter the discovered relations found with the Tri-occurrence approach:

RelEx (Fundel et al., 2007): RelEx is a rule based relation extraction tool using dependency parse trees to find relations.

SL (Giuliano et al., 2006): SL is a shallow linguistics SVM kernel for the identification of relations. Since no model was available for the identification of regulatory relations with this kernel, we used the simple margin active learning (Tong et al., 2001) approach to train a SVM model with probability estimations. A set consisting of 175 positive and negative relations was used to learn an initial model. This model was refined by applying the learned predictor on 10,000 randomly selected relations found by the Tri-occurrence approach and used the 100 instances which were closest to the separating margin of the SVM for the further training. The model was iteratively refined with this approach until no further performance improvement on a control set consisting of 100 examples (including 33 positive regulatory relations) could be observed. A final area under the receiver operating characteristic (ROC) of 0.85 and an area under the precision-recall curve of 0.72 could be reached on the control set. Furthermore, with a typical used probability threshold of 0.5 for the SVM, a precision of 0.56 and a recall of 0.75 could be reached (see Figure 4.2 for the ROC curve of the final predictor).

We decided to use RelEx and SL due to their good performance on the task of identifying undirected protein-protein interactions on different corpora (Tikk et al., 2010). The Tri-occurrence and the SL kernel approach predict only undirected relations. We used our list of TFs to set the direction from the transcription factor to the non-factor. In the case that both genes are non-factors or both are factors the relations were omitted by default in our system. Gene names between closely related species can highly overlap. Therefore, we identified the species context in each abstract using a pre-defined set of possible names for the different species.

4.2.2 Transcription Factor Binding Site Predictions

The promoter sequence for each gene was extracted using RSAT (Thomas-Chollier et al., 2011). The same promoter size of 1 kilo base pairs upstream of the transcription start site was chosen for all species. Binding motifs for the different TFs were taken from TRANSFAC (Matys et al., 2006) and JASPAR (Mathelier et al., 2014). The matching of these motifs to the promoter sequences was predicted with the R package cureos (Westermann et al., 2008). We used an empirically chosen threshold of 16 on the TFBS scores to filter out insignificant binding sites.

4.2.3 ConReg System Design

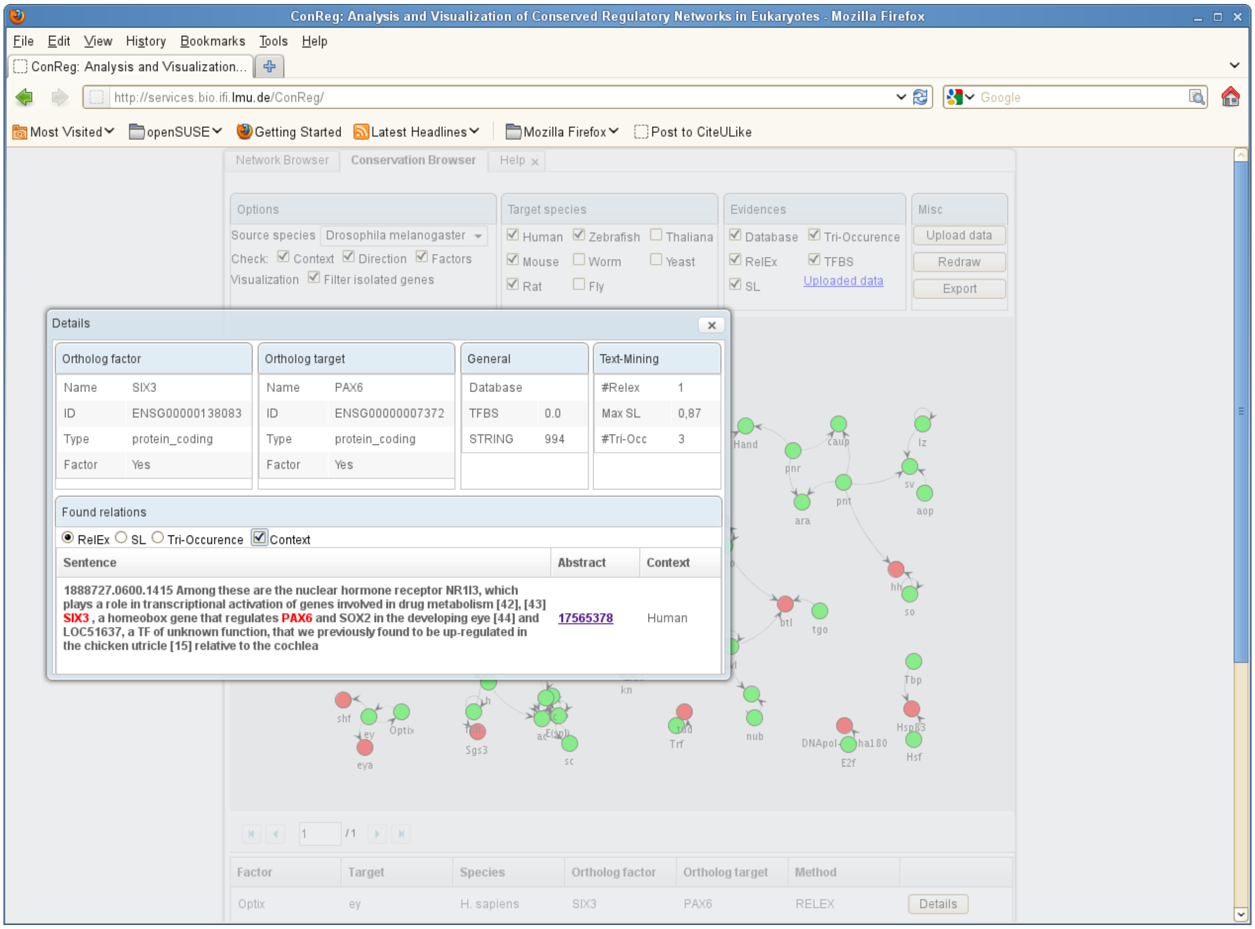
ConReg was developed as object oriented Java application using the open-source Ajax Web application framework ZK. The underlying data was unified in a structured MySQL database.

4.3 Discovery of Conserved Regulatory Relations with ConReg

With our system conserved regulatory relations for a source species can be discovered in a target species if these relations can be found between the respective orthologs in both species (by default we do not require conservation of the transcription factor binding site in the two species). ConReg searches regulatory relations of a source species which were extracted from regulatory databases, in the specified target species. Several types of evidence for regulatory relations of the target species can be considered based on the user's selections. Currently, regulatory information from publicly available databases like TRANSFAC (Matys et al., 2006) or REDfly (Gallo et al., 2011), relations found with text-mining approaches (RelEx (Fundel et al., 2007), SL (Giuliano et al., 2006), Tri-occurrence) in PubMed and PubMed-Central and TFBS predictions can be selected (see Materials and Methods for details). The *Conservation Browser*, where the entire predicted conserved network is shown and the *Motif Finder* with which the user can search for conservations in a defined subset of genes are the two main features of ConReg. For both features, the user can select the source species, regulatory data sources for the target species and further constraints for the text-mining approaches. Our system shows the conserved regulatory network as well as details for each found conserved relation. This includes information about the orthologs, where our text-mining approach found regulatory relations in the literature, the TFBS predictions and the protein-protein interaction score from the STRING database (version 9) (Szklarczyk et al., 2011). For further analysis the networks can be exported as tab separated file and used in advanced network analysis tools. An example can be seen in Figure 4.3, which shows a screenshot of the *Conservation Browser* with conserved relations for *D. melanogaster* as source species. The detail view window in the front shows information about the regulatory relations in the target species including an example of a regulatory relation which was discovered by RelEx between *Pax6* and *Six3* in *H. sapiens.*

4.3.1 Regulatory Data

For the source and target species, data from different data sources is available in our system. Table 4.1 gives an overview of the collected data in ConReg. For *S. cerevisiae* and *A. thaliana*, processed data from genome-wide chromatin immunoprecipitation experiments for



Options
Source species: Drosophila melanogaster
Check: ☑ Context ☑ Direction ☑ Factors
Visualization ☑ Filter isolated genes

Target species
☑ Human ☑ Zebrafish ☐ Thaliana
☑ Mouse ☐ Worm ☐ Yeast
☑ Rat ☐ Fly

Evidences
☑ Database ☑ Tri-Occurence
☑ RelEx ☑ TFBS
☑ SL Uploaded data

Misc
Upload data
Redraw
Export

Details

Ortholog factor	
Name	SIX3
ID	ENSG00000138083
Type	protein_coding
Factor	Yes

Ortholog target	
Name	PAX6
ID	ENSG00000007372
Type	protein_coding
Factor	Yes

General	
Database	
TFBS	0.0
STRING	994

Text-Mining	
#Relex	1
Max SL	0,87
#Tri-Occ	3

Found relations
● RelEx ○ SL ○ Tri-Occurence ☑ Context

Sentence	Abstract	Context
1888727.0600.1415 Among these are the nuclear hormone receptor NR1I3, which plays a role in transcriptional activation of genes involved in drug metabolism [42], [43] SIX3 , a homeobox gene that regulates PAX6 and SOX2 in the developing eye [44] and LOC51637, a TF of unknown function, that we previously found to be up-regulated in the chicken utricle [15] relative to the cochlea	17565378	Human

1 /1

Factor	Target	Species	Ortholog factor	Ortholog target	Method	
Optix	ey	H. sapiens	SIX3	PAX6	RELEX	Details

Figure 4.3. Screenshot of ConReg for the interactive discovery of conserved regulatory relations for a source species in user selected target species. Conservation of regulatory relations for a species can be interactively discovered using ConReg. The system allows searching for conservation in all provided species with different prediction methods for the target species, whereas for the source species only reliable relations from databases are considered. For an identified conservation of a regulatory relation details such as text-mining results and binding site predictions can be visualized.

some TFs was additionally available. This explains why more regulatory relations were found for these two species. For the other species only very few relations could be extracted which emphasizes the need of text-mining approaches to get a more complete view on the currently discovered regulatory networks. For instance, for *D. rerio* no relations were found in the databases, but 16,219 putative relations were found using text-mining. Nevertheless, we assume that data extracted from databases is reliable and use this data as source data for the discovery of conservations, whereas also the predicted relations are considered for the conservation search in the target species.

For our prediction methods, most relations were found with the Tri-occurrence text-

Table 4.1. Overview of the model organisms from our database with the number of genes, number of predicted and known transcription factors and the number of factors with position weight matrices (PWMs). In addition, the number of regulatory relations collected from databases and relations which were extracted from the scientific literature by using different text-mining approaches (RelEx (Fundel et al., 2007), SL (Giuliano et al., 2006) and Tri-occurrence) and from transcription factor binding site predictions (TFBS) are shown. Most regulatory relations could be found for *S. cerevisiae* and *A. thaliana* which originate mostly from genome-wide chromatin immunoprecipitation experiments. The numbers of found text-mining relations and of predicted binding sites is quite different for the model organisms.

Species	Genes	**TF**	**PWM**	**T**	**D**	**R**	**SL**	**TO**	**TFBS**
H. sapiens	21,673	1,416	300	6.5	3,230	20,391	29,422	103,511	220,245
M. musculus	23,497	1,431	276	6.1	932	10,682	15,616	51,729	130,456
R. norvegicus	22,503	1,181	20	5.2	321	5.950	8,905	33,857	3,050
D. rerio	21,322	1,081	0	5.9	0	2,930	4,322	16,219	0
D. melanogaster	14,076	570	139	4.0	471	2,433	3,802	11,635	6,054
C. elegans	19,992	688	6	3.2	128	102	149	385	1,308
A. thaliana	26,207	1,235	32	3.4	11,284	926	1,460	5,073	8,282
S. cerevisiae	5,884	233	170	4	29,716	812	1,446	4,036	6,075

TF=Transcription factors; **PWM**=Position weight matrices; **T**=$\frac{100\times \mathrm{TF}}{\mathrm{Gene}}$; **D**=Database relations; **R**=RelEx relations; **SL**=Shallow linguistics SVM kernel relations; **TO**=Tri-occurrence relations; **TFBS**=Transcription factor binding site predictions

mining approach and the TFBS predictions, but probably with a large number of false positives. Unfortunately, for some organisms the number of position weight matrices (PWM) for the search of TFBS is very limited. For *D. rerio* no PWMs were available and for *C. elegans* and *A. thaliana* only six and 20 PWMs could be found in the public domain. This explains the comparably low number of binding site predictions for these three species. The Tri-occurrence approach was used as pre-filter for the more sophisticated relation extraction approaches RelEx and SL. By comparing the relations found with RelEx to known relations extracted from databases a small overlap can be observed. For example for *H. sapiens* 22 % of the database relations could also be found with RelEx. A similar consistency could be observed for *R. norvegicus*, *M. musculus* and *D. melanogaster*. For SL 21 % of the known relations from *H. sapiens* could be detected. By combining the two state-of-the-art relation extraction approaches RelEx and SL, this rate could be increased to 28 % for *H. sapiens*. For those species with regulatory data from genome-wide chromatin experiments (*S. cerevisiae* and *A. thaliana*), this fraction is much lower as can be seen on the number of found regulatory text-mining relations. Furthermore, for *A. thaliana* and *C. elegans* only 34,729 and 31,325 species relevant abstracts could be found, whereas for *H. sapiens* and *M. musculus* 13,053,996 and 1,121,698 abstracts could be used. This explains the quite small number of relations for *A. thaliana* and *C. elegans* extracted with our text-mining approaches.

Most of the TFBS predictions could neither be confirmed with databases knowledge nor with the text-mining results. For example for *S. cerevisiae* 84 % of the predictions were unique for this method. The number and quality of TFBS predictions strongly depends on the available PWMs and their quality. Short PWMs for example produce many hits, but only with low scores which were not considered for the predicted relations in ConReg.

The currently available regulatory data is distributed in many different databases and stems from different data sources like manual literature curations or genome-wide chromatin immunoprecipitation experiments. The collection and integration of data from different sources and organisms is a difficult task and needs to be continued to make the most out of the available knowledge.

4.3.2 ConReg for the Discovery of Conserved Relations in Fruit Fly

We used *D. melanogaster* as source species to outline the usability of our system to find conserved regulatory relations for the 471 documented regulatory relations in REDfly (Gallo et al., 2011). We selected as target species the vertebrates *H. sapiens*, *M. musculus*, *R. norvegicus* and *D. rerio* and used all available data sources for these species. *D. melanogaster* is phylogenetically distant from the other species, but several conserved motifs are described in the literature as already mentioned in the introduction. We checked for the predicted conserved relations if we could confirm them in the target species. We assume that relations extracted from databases are correct and manually checked the relations found with our Tri-occurrence approach by reading the provided literature reported for each found relation. The Tri-occurrence relations are a super set of the relations extracted with our other text-mining approaches so that the performance for these approaches could also be checked, whereas the TFBS predictions were compared to the relations found in the databases and with the text-mining approaches.

The entire predicted conserved network is shown in Figure 4.4. Manual annotations where we could confirm a conserved regulatory relation between the orthologs in at least one vertebrate are shown as red edges. The conserved *D. melanogaster* network also contains the well-studied motifs for eye-development (*Optix*, *ey*, *eya* and *shf*, see Figure 4.6 and Table 4.2) and conservations for the pan-bilaterian kernel for heart specification, including the genes *Tin*, *Mef2* and *Mad*.

Only seven conserved relations, involving nine different genes could be identified with target relations extracted from databases. From these seven relations four were auto-regulations and the others were isolated edges. By using only the knowledge from databases, not even the well-studied conserved motifs between *D. melanogaster* and the other organisms could be rediscovered. With our Tri-occurrence approach 132 possible conservations could be found from which we could confirm 66 relations in at least one species (50 %). We compared the different methods to each other with respect to the number of predicted and confirmed relations. Furthermore, we compared the intersections of the predicted conserved regulatory relations from the different approaches (see Figure 4.5). All of the 66 found conserved regulatory relations found with RelEx could be confirmed or were also found by SL or the binding site predictions. With SL six additional conservations could be found. In addition, 124 pos-

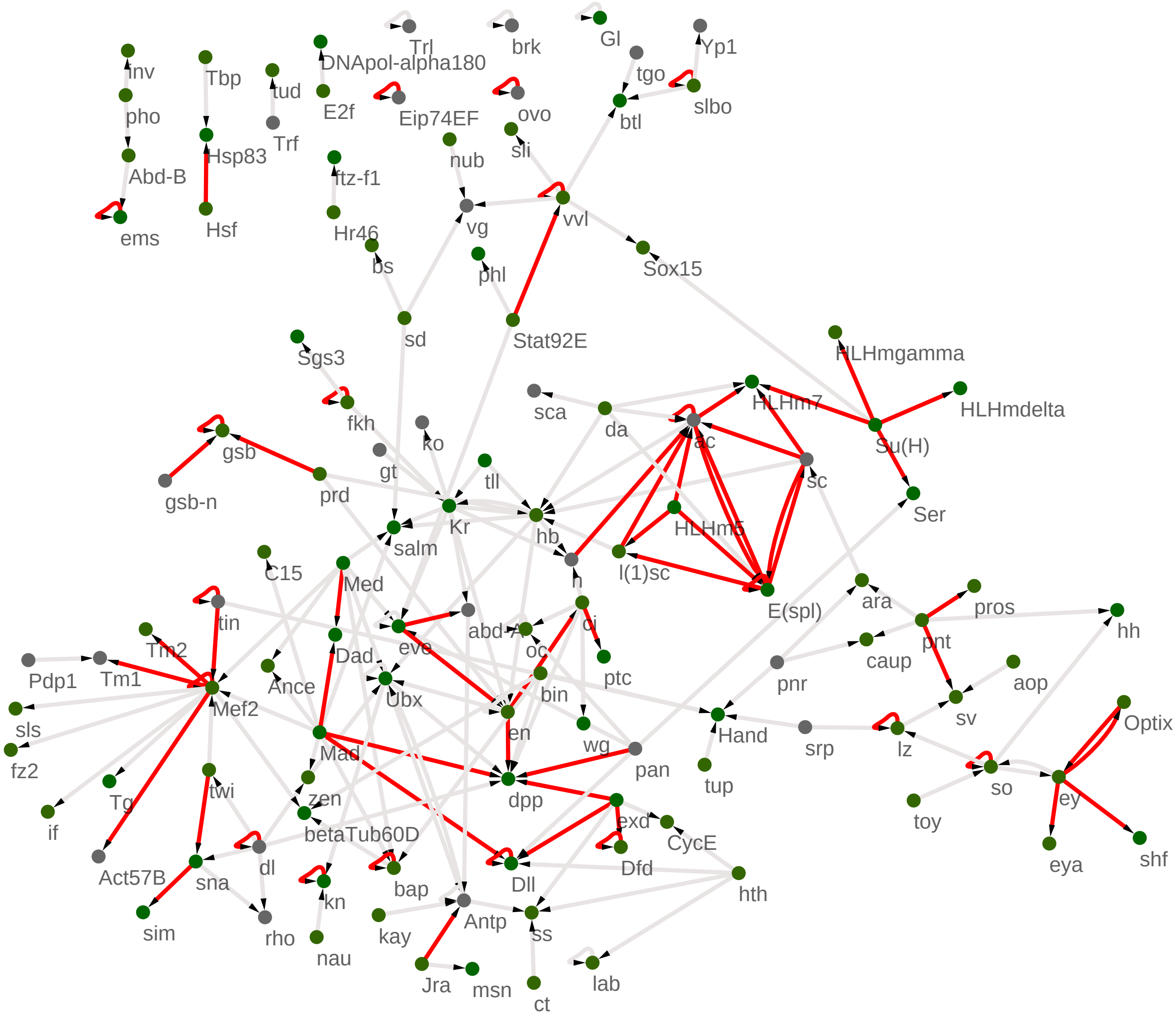


Figure 4.4. Network of conserved regulatory relations from the 471 documented regulatory relations in REDfly for *D. melanogaster* in at least another vertebrate. The gray edges represent all relations where we could find a possible conservation. Red edges represent edges where we could confirm the relations between the orthologs in vertebrates using the literature references provided by ConReg. In green, we highlighted the nodes where at least two ortholog identification approaches found an ortholog mapping to another vertebrate for the respective gene. The conserved network contains, among others, the well studied motifs for eye-development and the pan-bilaterian kernel for heart specification.

sible conserved relations were discovered with the TFBS predictions. 33 of these relations could be found with a different method including 25 confirmed Tri-occurrence relations. We note, that with RelEx the best relation extraction performance could be achieved with 57 out of 67 confirmed conserved relations (85 %). With SL a comparable performance could be reached with 59 out of 76 confirmed conserved relations (78 %).

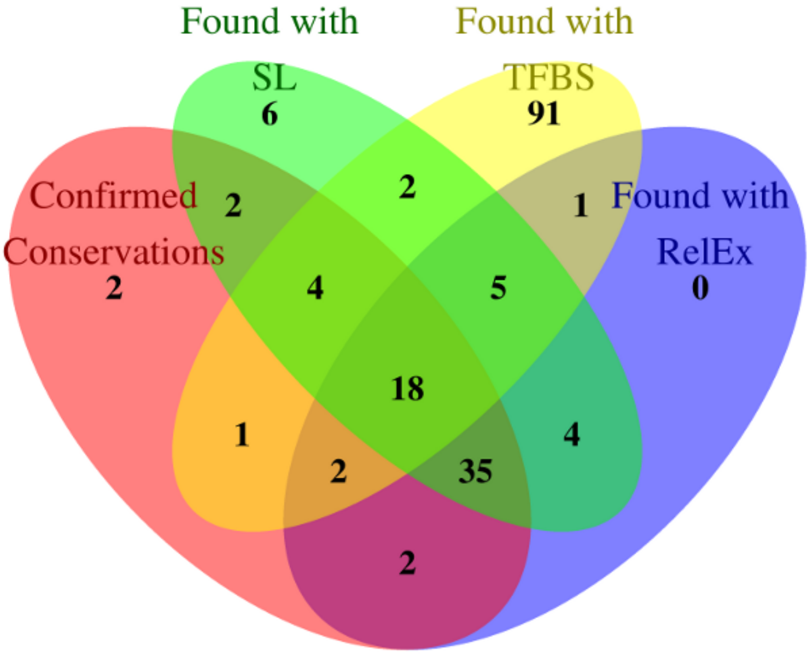


Figure 4.5. Venn-Diagram of found regulatory conservations between the 471 documented regulatory relations in REDfly for *D. melanogaster* and vertebrates. Confirmed conservations are regulatory relations which could be transferred from databases, or were correctly identified with our Tri-occurrence approach for at least one vertebrate (relations were manually checked by reading the corresponding literature). All relations found with RelEx could also be found with another method, whereas most of the TFBS predictions were not reported with our text-mining approaches.

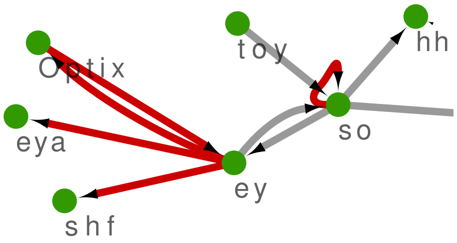


Figure 4.6. *D. melanogaster* eye development sub-network with conservation evidence (red edges could be confirmed with the scientific literature see Table 4.2).

Factor	Target	Species	Ortholog factor	Ortholog target	Sentence	Type	PMID
ey	eya	Rat/Mouse	Pax6	Eya1/Eya2	"Previous studies have suggested that Pax6 directly or indirectly regulates expression of DNA-binding transcription factors Six3, Sox2, Pitx3, Prox1, Sox1, and c-Maf [7], Sox11 [69] as well as transcriptional co-activators Eya1, Eya2 [70] and a co-repressor Dach1 [71] during early stages of lens development, i.e. lens placode and lens vesicle formation."	Fulltext	19132093
ey	Optix	Fish	Pax6a	Six3a	"However, we tested their function because of a previous study by Anders Fjose's lab that demonstrated that a possible Pax6.1 binding site on module F and a putative Brn3b binding site on module E are important for regulating six3a [33]"	Fulltext	20346166
		Worm	Vab-3	Ceh-32	"Our results suggest that VAB-3 acts upstream of ceh-32 during head morphogenesis and directly induces ceh-32"	Abstract	11476572
		Mouse	Pax6	Six6	"We found that the Lhx2 and Pax6 transcription factors operate in a concerted manner during retinal development to promote transcriptional activation of the Six6 homeobox-gene in primitive and mature retinal progenitors"	Abstract	19146846

Factor	Target	Species	Ortholog factor	Ortholog target	Sentence	Type	PMID
ey	shf	Rat/Mouse	Pax6	Wif1	"Promoters from the chitinase 3-like 3, Wnt inhibitory factor 1, and fms-related tyrosine kinase 1/soluble VEGF receptor genes were upregulated five-, seven-, and threefold, respectively, by Pax6 in transfected COS7 cells."	Abstract	21447684
Optix	ey	Human/ Mouse	Six3	Pax6	"Six3 activation of Pax6 expression is essential for mammalian lens induction and specification"	Abstract	17066077
so	so	Human/Mouse	Six1	Six1	"Positive autoregulation of Six1 is achieved through the regulation of Six protein-binding sites."	Abstract	21447684

Table 4.2. The table lists author statements for potential regulatory interactions.

Comparison to Alternative Tools

Different tools also focus on the identification of conserved relations in eukaryotes. For example with the UCSC Genome Browser (Fujita et al., 2011) bindings from ORegAnno or other genome-wide chromatin immunoprecipitation experiments can be mapped on the genome and information of conservations on the DNA level for different species can be displayed. Also the Genomatix suite[1] allows uploading experimental data for further analyses and for searching for conservations. Compared to prokaryotic genomes, eukaryotic genomes are rich in non-coding sequences of unknown functions and promoters can lay several kilobases upstream from the transcription start site. Nevertheless, different approaches have been introduced to search for conserved binding site predictions (Loots and Ovcharenko, 2004; Berezikov et al., 2005). Furthermore, for microbial gene regulatory networks different platforms exists for the storage and web-based analysis as reviewed by Baumbach et al. (2009).

In comparison to these tools, ConReg focuses on eukaryotes and provides detailed information of putative conservations. The user is enabled to interactively explore the conservations in the underlying processed and unified data. ConReg does not only rely on the knowledge from databases or predicted binding sites, but also strongly uses information extracted from the literature which are currently only used to a minor extend by other tools.

4.4 Conclusion

We presented ConReg a novel interactive online system for the discovery of conserved regulatory relations in eight eukaryotic model organisms. Our system allows searching for regulatory conservations among all possible sets of target species and gives rich information details for possible conserved relations. We collected regulatory relations from structured databases, via text-mining from unstructured textual descriptions and from binding site predictions. We observed the incompleteness of regulatory relations in databases which are not even sufficient for the discovery of well-known conserved motifs. With the integration of information from state-of-the-art text-mining approaches and binding site predictions, several conserved motifs could be found using *D. melanogaster* as source species. We were able for *D. melanogaster* to identify conserved regulations for 14% (66 out of 461) of the relations from *REDfly* in at least one vertebrate. But still it remains unknown to which extend regulatory relations are conserved since only few regulatory relations are experimentally confirmed for eukaryotes.

For our selected show case we noticed that even with the simple Tri-occurrence text-mining approach 50 % of the identified regulatory relations are correctly identified when also experimentally validated regulatory relations between orthologs were known. Thus, with the integration of additional background knowledge the relation extraction could be significantly increased.

[1] http://www.genomatix.de

We designed our system so that further information sources can easily be added. In particular, we are planning to incorporate further information from the increasing number of available chromatin immunoprecipitation experiments into ConReg. Furthermore, we are going to provide a Cytoscape (Smoot et al., 2011) plug-in to access the data for follow-up analyses in addition to our web interface.

Chapter 5

Context-Specific Regulatory Network Framework

Abstract: The ENCODE, mouseENCODE and modENCODE projects have published various genome-wide measurements for various human, mouse, fly, and worm cell lines. More such data have been made available by the TCGA and Epigenomics Roadmap consortia. From these measurements a wide range of global and context-specific functional features and annotations can be derived. The analysis of these large data sets and the derived features, in particular the differential analysis of two or more sets across conditions or even across compendia is cumbersome and difficult.
Many of these context-specific regulatory features can be modeled as Transcription Factor (TF) - Target Gene (TG) networks. Such networks provide intuitive views on the ENCODE data and allow the comparative analysis of replicates, different contexts, different cell lines, different cell types, and different species. But cross-species and cross-condition comparative analysis on many and large networks still requires time consuming manual work. This applies in particular for the identification of conserved and context-specific interactions with currently available network analysis and visualization software solutions.

The Cross-species Conservation framework (CroCo) enables comparative network analysis on both standard conventional global networks and on context-specific regulatory networks derived from thousands of ENCODE regulatory experiments. CroCo provides both a network repository and ontology of pre-computed networks as well as a software tool suite to efficiently conduct networks analysis. The networks in the repository are derived from all ENCODE regulatory ChIP-ChIP, ChIP-seq and open chromatin experiments (DNase-seq, DGF and FAIR-seq), the scientific literature, binding site predictions and curated databases. The CroCo tool suite includes a web interface for network property queries, a plug-in for connecting the network repository with Cytoscape and an Application Programming Interface (API) to support the development of tailor-made analysis workflows. Applications of the CroCo framework range from simple evidence look-up for user-defined regulatory interactions to the identification of conserved sub-networks in diverse cell lines, conditions, or even species.

CroCo adds an intuitive unifying view on the data from the ENCODE projects via a

comprehensive repository of derived context-specific regulatory networks and enables flexible cross-context, cross-species and cross-compendia comparison by a basis set of analysis tools.

Publication: The CroCo system is briefly described in a BioSpektrum article (Pesch and Zimmer, 2014). A manuscript describing the details of the CroCo framework is in preparation.

My contribution: I implemented the CroCo systems, defined the networks, performed the conservation analysis and drafted the manuscripts.

Contribution of co-authors: Ralf Zimmer supervised the work and helped drafting the published manuscripts. Madox Sesen implemented a prototype of the regulatory sub-network overlap functionality in the *croco-web* application as a student helper.

5.1 Introduction

In September 2012, the ENCODE project (ENCODE Project Consortium, 2012b) published functional annotations for over 80 % of the human genome. Thousands of genome-wide measurements and features have been made publicly available. Also the mouseENCODE project (Mouse ENCODE Consortium, 2012), started together with the ENCODE project, and the modENCODE project (Celniker et al., 2009) provide similar information for the genomes of mouse, fly and worm. Diverse experimental methods have been applied in ENCODE projects to obtain (genome-wide) functional genome annotation for hundreds of cell lines and conditions. Fortunately, the experimental methods employed by the ENCODE project are standardized and follow common guidelines (Landt et al., 2012; ENCODE Project Consortium, 2012a). This enables integration and combination of various data sets and eases their comparison. Exploiting this huge data repository many different aspects can be investigated, like the comparative and cell-specific analysis or regulatory elements (Boyle et al., 2014; Neph et al., 2012b).

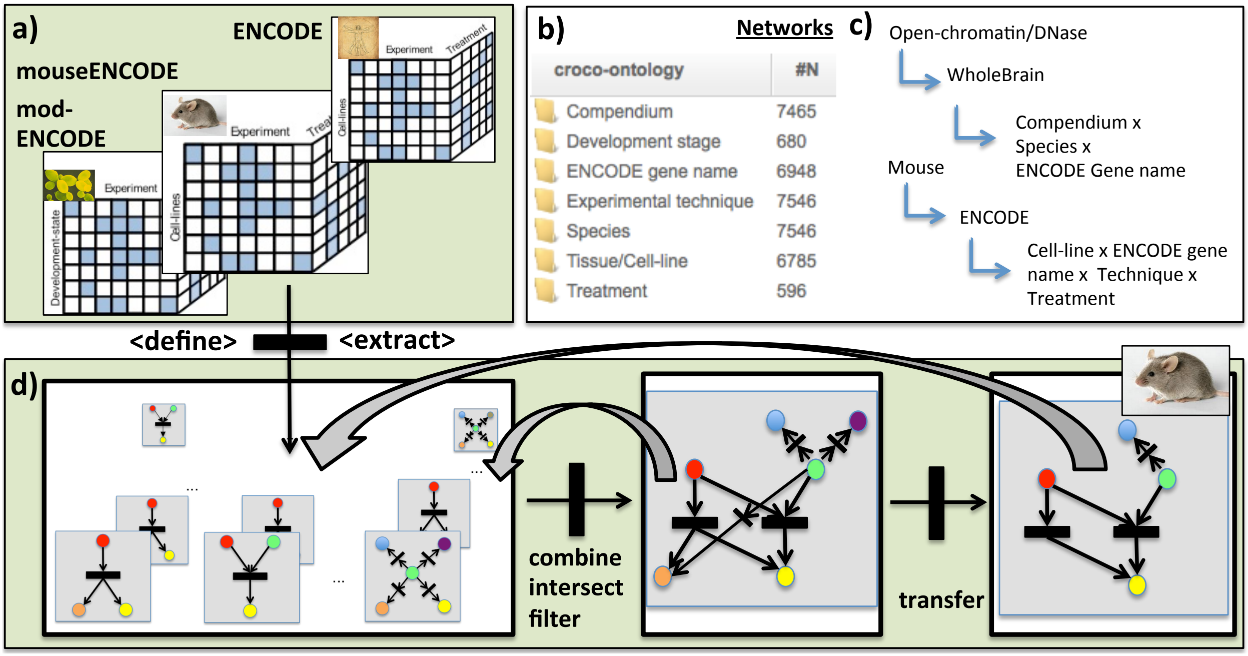
With the ENCODE data a huge amount of regulatory data is provided. A standard approach to analyze, interpret and visualize the underlying (context-specific) mechanisms of cellular systems is via modeling of Transcription Factor (TF) - Target Gene (TG) regulatory networks (Karlebach and Shamir, 2008). Regulatory network models (TF-TG networks) are used in various contexts for generating and validating new biological hypotheses or for explaining experimental data (Küffner et al., 2005; Van Landeghem et al., 2013; Faro et al., 2012; Pesch et al., 2012). For example, we previously used regulatory networks from fly and vertebrates mined from the scientific literature in combination with binding site predictions to identify conserved regulatory sub-networks between them (see Chapter 4 and Pesch et al. (2012)).

For various research questions either **Context-specific networks** (network that represent a specific state of a system), or **Global networks** (networks that represent context-independent features of a system, i.e. features and interactions collected and combined from several states and contexts) are used. Context-specific networks, for example, allow the study of differential bindings for a specific factor, while a global network could summaries all possible binding regions across contexts. Context-specific (regulatory) networks can be derived from Chromatin immunoprecipitation sequencing (Johnson et al., 2007) (ChIP-seq), or open chromatin experiments such as (i) DNase I hypersensitive sites sequencing (Boyle et al., 2008) (DNase-seq), (ii) Digital Genomic Footprinting (DGF), or (iii) Formaldehyde-Assisted Isolation of Regulatory Elements (FAIR-seq). In contrast to those context-specific networks can standard conventional global networks be derived from binding site predictions, from merging various condition-specific networks, or from text-mining the scientific literature.

Many context-specific networks for many system states can be modeled with the data provided by the ENCODE projects. Initial analyses revealed that regulatory elements are highly context-specific and complex (Neph et al., 2012b; Gerstein et al., 2012; Thurman et al., 2012). Thus, only a fraction of regulations can be observed across many different cell lines, or in any individual cell-line. This implies that it is essential to consider the context for (cross-species) analysis of regulatory networks. But currently no comprehensive network repository exists for condition specific regulatory networks. Thus, for the (cross-species) network analysis the raw data needs to be manually gathered, processed and networks need to be constructed. The construction of TF-TG networks from experimental binding data requires, for example, the identification of binding sites, and the prediction of possible targets for the DNA binding protein in the respective context. With ChIP-seq experiments the binding of a protein to the DNA is measured directly making the inference for regulatory targets for the ChIP-ed factor possible for all genes with bindings within the promoter region. This approach was for instance used by Kim et al. (2008) for several transcription factors in mouse embryonic stem cells in order to induce cell type-specific regulatory sub-networks. Advanced experimental techniques and computational predictions like the combination of open chromatin data and transcription factor specific Position Weight Matrices (PWM) allow the construction of networks for many factors at once. For example, Neph et al. (2012b) combined Digital Genomic Footprinting (DGF) (Hesselberth et al., 2009) of DNase I cleavages from 41 cell lines and tissues with transcription factor specific PWM to infer TF-TF relations on a genome-wide scale for 475 transcription factors at once. Apart from the methods required for the network construction the currently available software support for the analysis of hundreds of networks as derived from context-specific ENCODE data is limited.

Cross-species Conservation (CroCo) framework

With CroCo we present a **repository of pre-computed regulatory networks** and a user-friendly tool suite for the efficient analysis of various aspects of both global and context-specific networks derived from the ENCODE data sets and further data compendia. The representation of the data as regulatory networks with a common set of nodes provides a uniform handling of the available information derived from thousands data sets of heterogeneous types, various experimental techniques and from different sources. A **common set of nodes** (CN) is maintained via appropriate mappings of the respective measured objects. Thus, every context-specific network derived from individual or sets of primary data is a set of edges defined of the CN. The uniform set of nodes CN allows simple means to combine networks in a straightforward and easy to interpret way. Due to standardized procedures followed in most large scale data compendia the mappings between the measured objects are obvious, but there are also complications imposed by e.g. definitions of genes and gene or isoform structures or by the incorporation of additional data sets not using standardized procedures. A challenging mapping between species can be obtained from orthology mappings to allow mapping of networks across species. Ideally this establishes a one-to-one correspondence between objects measured in different species thereby again realizing a common node set between species. Unfortunately, due to parallels and weak homology the situation is not as clear leading to quite some n:m relations. Apart from that any othology mapping allows to transfer networks from one to another species in order to generate or validate regulatory hypotheses **across species**. In the ENCODE, TCGA, and Epigenomics Roadmap compendia, the data sets are classified with respect to various criteria defined in the associated metadata. This induces a multi-dimensional organization of the available data set into what is called the



croco-ontology	#N
Compendium	7465
Development stage	680
ENCODE gene name	6948
Experimental technique	7546
Species	7546
Tissue/Cell-line	6785
Treatment	596

Figure 5.1. CroCo provides a uniform view on compendia of genome-wide measurements along with global networks derived from structured databases and further resources. **a)** Datasets in these compendia are classified into a high-dimensional data cube along the dimension listed in **b)**. **c)** Each dimension can be navigated via ontologies in any order. **d)** CroCo uses default or user-defined procedures to define and extract networks resulting in a high-dimensional cube of networks structured along the same dimensions. These networks can be filtered, merged, and combined in various ways to produce new networks. Moreover, networks can be transferred between species via orthology mappings of the network nodes. This enables a prediction of regulatory interactions from one or a set of species to closely related species. Via combination and transfer operations new networks are defined thereby enabling a flexible construction of user-specified networks from the compendia.

data cube in the following (see also Figure 5.1). Typical classification criteria and dimensions in this 7-dimensional data cube are: Compendium x Development stage x ChIP-Factor x Experimental technique x Species x Tissue/Cell-line x Treatment (C x DS x CF x ET x S x TC x T). CoCo systematically exploits this intuitive structure for representing, browsing, and handling the available data sets and the associated networks in the software. Thereby, CroCo GUI allows navigation through the data and networks in an intuitive way according to the known classes imposed by the compendia and the implied data cube. Moreover, we introduced more convenience for the user by allowing to navigate the dimensions of the data cube in any order starting with an arbitrary dimension and continuing subsequently along any other of the remaining dimensions (**multidimensional browsing**). Thereby, CroCo provides an intuitive overview of all available data sets and tries to ease the search and selection of particular individual data sets. Moreover, we systematically employ **ontologies** to structure any dimension of the data cube. These ontologies are either provided by the metadata of the data compendia or are derived from additional information. Ontologies can also be provided by the user to structure the data according to personal preferences or to classifications derived from previous analyses.

The analysis of those networks is supported via a client side Cytoscape plug-in and a server-side web-application. The client-side application is suited for in detail downstream analysis. Networks are accessed via a publicly available web-interface, which supports basic network operation such as union, intersection, merging as well as (cross-species) transfer of many networks. The web-application gives an initial view on the networks available in CroCo. For example, networks can be browsed, downloaded or compared according to the out-degree of a transcription factor, or the overlap with previously identified annotated sub-networks. Furthermore, regulatory evidence can be visualized including the exact binding positions of a transcription factor within the promoter of target gene, literature evidence and information from structured database.

5.2 Materials and Methods

5.2.1 System Architecture

CroCo consists of five components: (i) a network repository *croco-repo*, (ii) an Application Programming Interface (API) *croco-api*, (ii) a Cytoscape plug-in *croco-cyto*, (iv) a web application *croco-web*, and (v) a web-service for remote access to the central repository. In Figure 5.2 the interplay of the different components is shown.

The network repository (*croco-repo*) is the central component of the CroCo system. It consists of more than 7,500 pre-computed global and context-specific networks for human, mouse, fly and worm together with gene annotations and ortholog mappings. Via the combination of the publicly available web-service and the *croco-api* the server-side data repository can be accessed from the client-side. We structured the *croco-api* into: (i) a repository query layer, (ii) a network operation layer, and (iii) the network construction workflows, which derives networks from the raw ENCODE data. The query layer provides a low level set of operations to access the *croco-repo*. Examples of such queries are: list networks in the repository, read a network, or retrieve the metadata and the construction parameters for a specific network. The *croco-repo* can either be accessed via a direct database connection using the Structured Query Language (SQL) or via the Hypertext Transfer Protocol (HTTP). The web-services exposes the query operations on a web-server and tunnels the requests to a server side *croco-api* instance with direct access to the *croco-repo*. On top of this API we offer components for conducting network analyses. With *croco-web* we offer a web interface, which allows — without the need of installing additional client-side software — to query and compare network statistics or to look-up evidence for specific TF-TG relations via standard web browsers. For downstream network analyses we developed a plug-in for the bioinformatics network tool Cytoscape (Cline et al., 2007) in order to access the network repository and to perform common operations.

The components have been implemented in Java in combination with MySQL for the

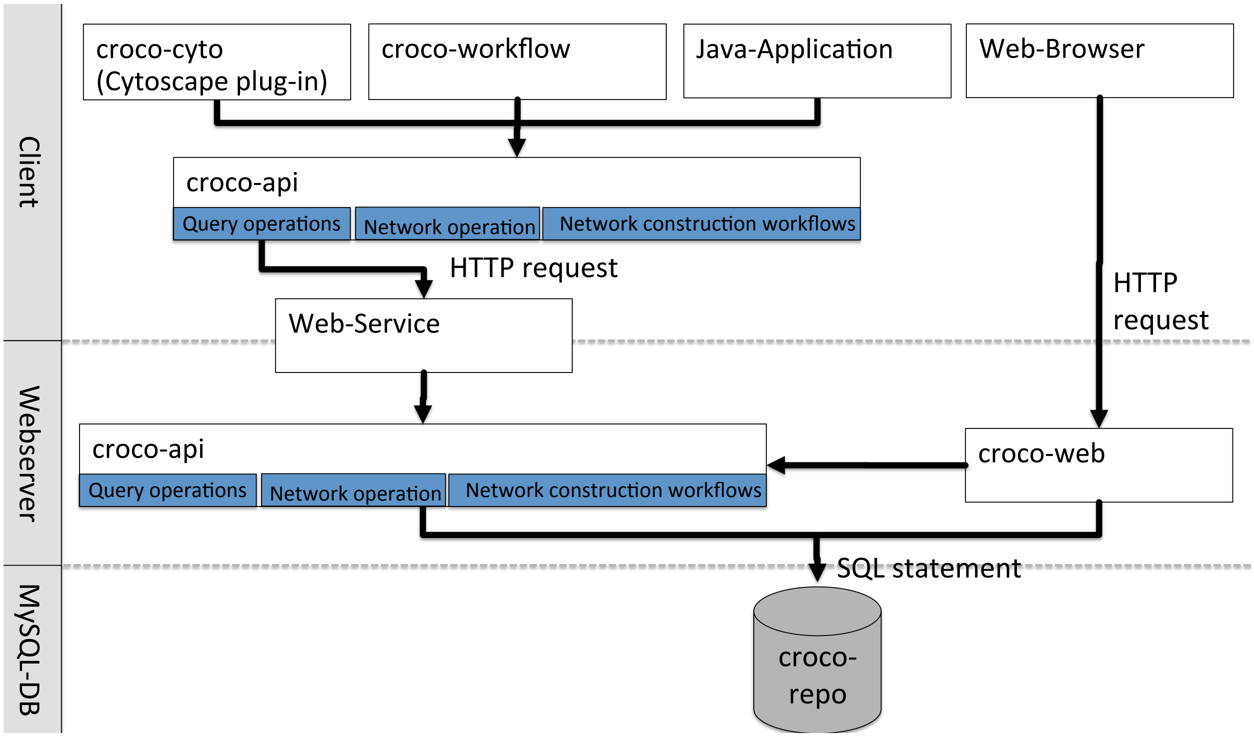


Figure 5.2. The CroCo framework consists of a data repository (*croco-repo*), an Application Programming Interface (API) (*croco-api*), an interactive web interface (*croco-web*), and a Cytoscape plug-in (*croco-cyto*). The *croco-repo* is a central database which includes derived condition specific and global networks, ortholog mappings and gene annotations. Via the *croco-web* interface networks can be compared based on several properties such as the number of total interactions, or the number of in-/out-interactions (in-/out-degree) of specific genes. For Cytoscape we developed a plug-in (*croco-cyto*) for downstream analysis. Finally, in order to assist the development of customized workflows, the *croco-api* can be used to integrate CroCo in additional processing and analysis pipelines.

croco-repo and the Open Source Community Edition of the ZKOSS Web Framework for *croco-web*.

5.2.2 Network Definition

The *croco-repo* contains global and context-specific networks, ortholog mappings for 59 eukaryotic species from ENSEMBL Compara (Vilella et al., 2009), and gene annotations for the organisms investigated in an ENCODE project (human, mouse, worm, and fly). The networks are represented as nodes with ENSEMBL gene identifiers serving as common set of nodes (CN) and edges as directed pair of nodes. This simple uniform representation of the networks facilitates the comparison of species-specific networks between different contexts (inter-context) and between species (inter-species). We use the following network definitions to create networks from binding site predictions, ChIP and open chromatin data:

Binding site predicted networks: We use FIMO (Grant et al., 2011) with Position Weight Matrices (PWM) from TRANSFAC (Version 9.3) (Matys et al., 2006), JASPAR Version 2014 (Mathelier et al., 2014), UniPROBE (Robasky and Bulyk, 2011), Wei et al. (2010), Wang et al. (2012) and Chen et al. (2008) to scan for possible binding sites with a p-value threshold of 10^{-5} in the genomes of human, mouse and worm. Regulations are predicted between TF-TG if a PWM hit associated with the TF is located within ± 5 kilo bases of the Transcription Start Sites (TSS) of the TG in human and mouse and 500 base pairs in worm. Furthermore, we construct a high-confidence network by further filtering the binding site predictions with a p-value threshold of 10^{-6}. In addition, conserved TFBS predictions in 12 Drosophila genomes are integrated from Kheradpour et al. (2007) for fruit fly.

ChIP-chip/seq networks: ChIP peaks are provided by the ENCODE projects for thousands of contexts with a median base pair (bp) peak length of 409 for worm, 671 for fly, and 150 for human and mouse. Regulations are inferred between the ChIP-ed protein and all TGs with peaks within ± 5 kilo bases for human and mouse and 500 base pairs for worm and fly of their TSSs.

Open chromatin network

1. We integrate the 41 human pre-computed cell-specific TF-TF networks derived from Digital Genomic Footprinting (DGF) published by Neph et al. (2012b). They used DGF footprints with a length of 6–40 bp and overlapped those footprints with predicted TRANSFAC motif-binding sites using FIMO with a p-value threshold of 10^{-5}. Regulations were inferred between TF-TF if an associated PWM for the first TF is found within a footprint of the second TF.

2. Similar to Neph et al. (2012b) we use open chromatin peaks derived from all ENCODE DGF, DNase and FAIR-seq experiments to predict regulations. The open chromatin peaks have a length of 150 bp. We overlay those peaks with the above mentioned binding site predictions.

In addition, we integrated networks from ConReg, a resource for global regulatory networks (see Chapter 4 and Pesch et al. (2012) for the detailed network construction workflows). ConReg provides the following network types:

1. Curated-database networks: Networks extracted from structured regulatory databases like ORegAnno (Griffith et al., 2008) or REDFly (Gallo et al., 2011).

2. Literature-networks: Networks derived from the scientific literature (PubMed and PubMedCentral) using a text mining approach. Triple occurrences, i.e. sentences with at least two genes and a regulatory keyword are used to generate labeled edges between

the two genes. (Undirected) relations are predicted between all found genes in those sentences, i.e. each unique triple occurrence generates two directed relations. In order to filter the networks and to generate more specific networks, versions of the text mining network are produced using a species filter (species-specific relations) and an approach to generate directed networks.

5.2.3 Network Operations

Since the *croco-repo* contains many and large networks, efficient network operations are crucial to perform network analyses in a user-friendly and interactive way. Thus, the API provides various common network operations optimized to work on the networks from the repository. Each network operation takes as input one or more networks and additional parameters in order to produce a new network. In addition to basic network operations Union, Intersection and Set-Difference the following specific operations are provided:

Orthology Transfer: Transfers a network using orthologs from the *croco-repo* to another species.

Binding Site Ortholog Transfer: Transfers measured and predicted binding sites available as additional information for some networks to other species using genome wide chained BLASTZ alignments (Schwartz et al., 2003) provided by ENCODE for many different species.

Shuffle: Shuffles the edges in a given network, but keeps the same in- and out-degree for the genes.

Gene Set Filter: Creates an induced network only consisting of genes with a particular Gene Ontology (GO) annotation, or genes from a user-defined gene set.

Support Filter: Removes edges which have been observed in less than a user-defined number of times in a merged network.

Binding Site Filter: Filters interactions based on the distance between the TF and TG or the associated p-value of the predicted binding.

The network operations can be chained and hierarchically organized, which results in a top-down processing by automatically retrieving the necessary data i.e. the networks, ortholog mappings and gene name information from the network repository. By combining several network operations typical tasks such as the identification of similarities and differences in networks derived from different cell lines or even from different species can be performed.

Table 5.1. Global networks and context-specific networks derived from the ENCODE data are integrated into the CroCo repository. Database derived networks stem from different curated sources (e.g. TRANSFAC and ORegAnno for mouse), whereas different binding site predicted networks result from different PWM collections and a sensitive and a specific PWM match threshold. For each species four different (filtered) literature derived networks are included in the repository. The majority of networks in the repository is inferred from context-specific ChIP and open chromatin experiments.

Species	Global			Context-specific							Total
	Databases	Literature	Binding site	ChIP				Open chromatin			
	N	**N**	**N**	**CO**	**AB**	**E**	**N**	**CO**	**E**	**N**	**N**
Human	4	4	12	103	192	1,206	1,206	105	248	3,617	4,843
Mouse	2	4	12	28	50	162	162	39	123	1,800	1,980
Worm	2	4	4	15	91	561	561	—	—	–	575
Fly	3	4	22	23	59	119	119	—	—	–	148
Total	11	16	54	169	392	2,048	2,048	144	371	5,417	7,546

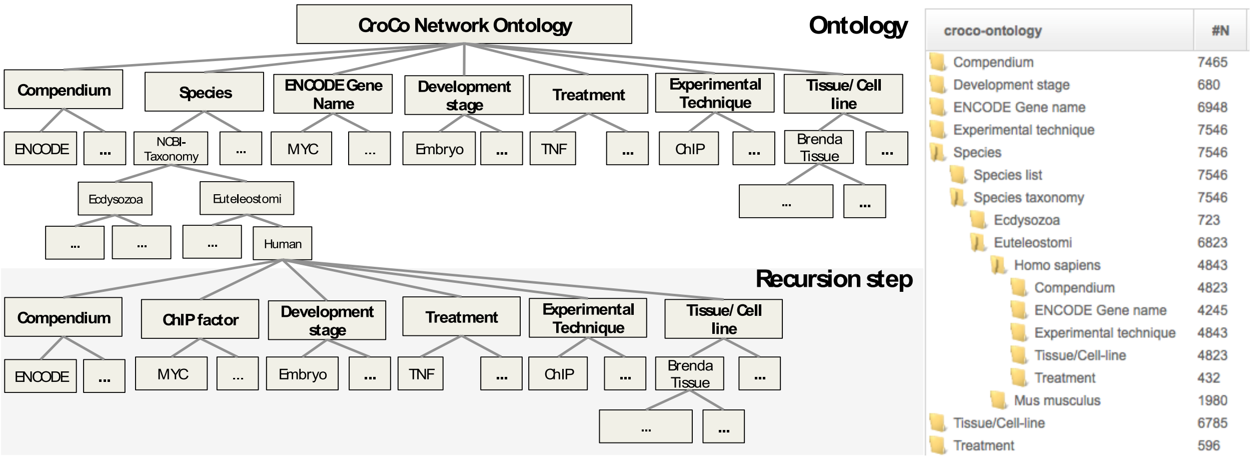
N=Number of networks; **CO**=Number of different contexts (cell lines for human, cell lines or tissues for mouse, development stages for worm and fly); **AB**=Number of different antibodies; **E**=Number of experiments.

5.3 Results

5.3.1 Comprehensive Context-Specific Regulatory Network Repository

croco-repo contains context- and species-specific networks for every ChIP and open chromatin ENCODE data set, which allows the analysis of different tissues, cell lines and replicates. In addition, 81 global networks are integrated in the repository yielding 7,546 networks in total.

The availability of experimental data differs across the considered species in the *croco-repo*. For instance open chromatin derived networks are only available for human and mouse, whereas literature derived and binding site derived networks are available for all species. For human 4,843 networks are contained in the repository; including 4 networks from curated databases, 12 networks predicted with different PWM collection sets, 4 literature derived network, 1,206 ChIP-derived factor-specific networks for 103 different cell lines, and 3,617 open chromatin derived networks. ChIP experiments have been conducted for different conditions and with different antibodies. For example, for the human cell K562, the binding of 116 different TF have been measured, whereas for the human cell line WI-38 only the binding



croco-ontology	#N
Compendium	7465
Development stage	680
ENCODE Gene name	6948
Experimental technique	7546
Species	7546
Species list	7546
Species taxonomy	7546
Ecdysozoa	723
Euteleostomi	6823
Homo sapiens	4843
Compendium	4823
ENCODE Gene name	4245
Experimental technique	4843
Tissue/Cell-line	4823
Treatment	432
Mus musculus	1980
Tissue/Cell-line	6785
Treatment	596

Figure 5.3. The *croco-ontology* is structured according to seven dimensions. Each dimension can be further structured according to dimension-specific ontologies e.g. the Brenda Tissue Ontology for the Tissue/Celll-line dimension. Via the components of the CroCo system the *croco-ontology* can be browsed in a recursive manner. The left figure shows the entire *croco-ontology* and the first recursion step, where all not yet selected dimensions are appended to the 'leave' node 'Human' in the **Species** dimension. The right screenshot shows a specific realization of the recursive browsing in *croco-web*. In *croco-web* and *croco-cyto* only dimensions are visualized, which further separate the data.

of CTCF was investigated. The different human open chromatin networks stem from 207 DNase-seq, 54 DGF, 37 FAIR-seq experiments in combination with the 12 different binding site predicted networks and the 41 networks from Neph et al. (2012b), i.e. (207+54+37) Experiments × 12 binding site networks + 41 = 3,617 different open chromatin networks. In Table 5.1 we summarize the available network for the considered species in the *croco-repo*.

The repository includes for each network the detailed meta data and parameterization, which is used for network construction. For example for text mining networks the sentence from the scientific literature, which supports a particular regulation can be retrieved.

5.3.2 Context-Specific Network Ontology

In order to enable flexible navigation and selection of networks from the CroCo network repository, we organized the data according to various dimensions (annotations). We identified the following dimensions:

Compendium: Data compendia: ENCODE, modENCODE, mouseENCODE

Species: Species with corresponding ENCODE project.

ENCODE gene name: Transcription factors with corresponding ENCODE ChIP-seq experiments.

Development stage: Development stage of a sample (experiment).

Treatment: Treatment of a sample(experiment).

Experimental technique: Experimental technique such as ChIP-seq.

Tissue/Cell line: Tissue/cell-line of a sample.

Each dimension can be further structured according to simple value lists, or even according to specific ontologies. Networks are assigned to node(s) in the dimension-specific value lists and ontologies based on their meta-information. For example, a network derived from a human ChIP-seq experiment performed by ENCODE for CTCF in K562 cells is assigned to: ENCODE in the **Compendium** dimension, Human in the **Species** dimension, CTCF in the **ENCODE gene name** dimension, ChIP-seq in the **Experimental technique** dimension and K562 in the **Tissue/Cell line** dimension. Note, however, that not all networks must be organized according to all dimensions. We build a meta-ontologie (*croco-repo*), which includes the seven dimensions including their categorizations. In the components of the CroCo system, users can start at any of our specified dimensions e.g. **Species** and browse for attributes of interest in the corresponding value list and ontologies (see Figure 5.3). As soon as a leaf node within the specific dimension is reached (e.g. human for the **Species** dimension) the user can select a further dimension to browse the remaining data according to the not yet selected dimensions (note: only those dimensions are shown, which further separate the remaining data). That way it is possible to first select a species and than to select an experimental technique, or vise versa.

5.3.3 Network Conservation Analysis with Cytoscape

The networks contained in the *croco-repo* together with the available operations in the *croco-api* can be accessed directly via *croco-cyto*, a plug-in for the bioinformatics networks analysis tool Cytoscape. Result networks, for example shared conserved networks of the analogous leukemia cell lines in mouse and human for genes involved in the KEGG leukemia pathway can be produced by selecting the desired networks and applying/stacking network operations. *croco-cyto* uses the *croco-api* to retrieve the pre-computed networks, ortholog mappings and gene descriptions from the server-side *croco-repo* via the publicly available web service. Thus, networks of interest can be easily defined, e.g. shared sub-network between cell lines or conserved sub-networks among different species.

In Figure 5.4 we show an example network generated with *croco-cyto* for the analogous leukemia cancer cell-lines MEL and K562 in human and mouse for genes involved in the KEGG Leukemia pathway. The cross-species comparison is archive by using the network transfer operation. The left screenshot shows a unified network of the two species consisting of 597 interactions, and the right screenshot shows the intersected network between the two cell lines and species consisting of 52 consistently observed regulations. Furthermore, the literature networks are used to highlight regulations with additional evidence from the scientific literature.

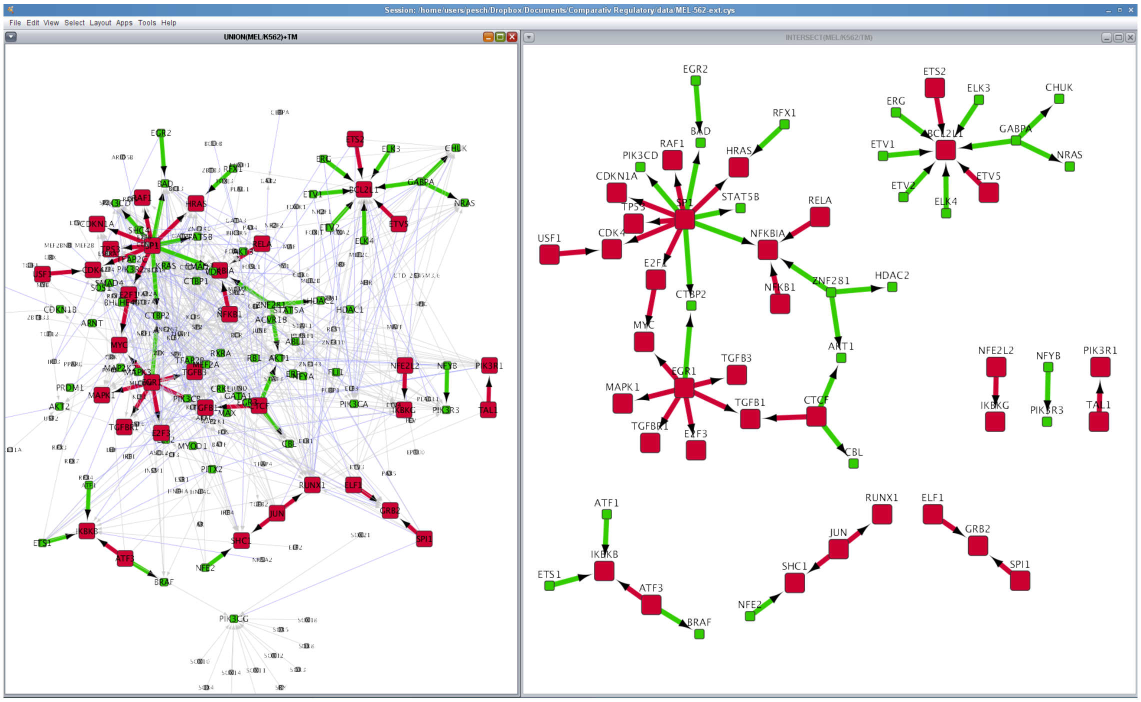
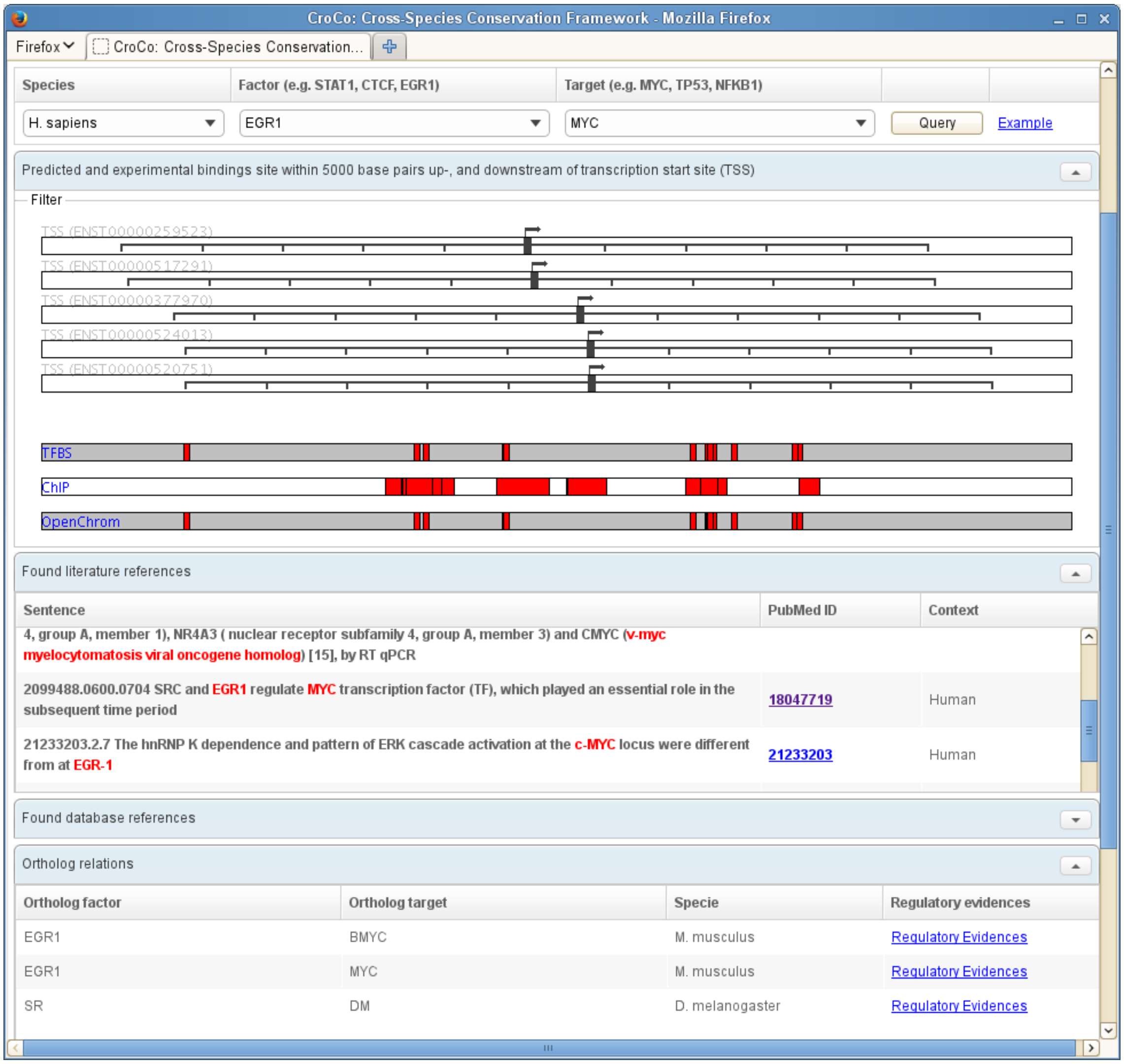


Figure 5.4. The screenshot shows: (i) the union of the orthology-transferred network derived from three MEL mouse experiments and two human K562 networks consisting of 597 interactions (left screenshot), and (ii) the conserved network between the two networks consisting of 52 consistently inferred interactions (right screenshot). The edges are colored according to the available evidence for an interaction. Grey edges represent interactions only inferred from human, blue edges represent interactions only inferred from mouse, green edges represent conserved interactions (inferred from human and mouse), red edges represent conserved interactions with literature evidence.

5.3.4 Network Metric and Evidence Look-up

The web service *croco-web* gives a first view on the networks in the *croco-repo* and enables several network queries without the need of installing additional software. *croco-web* consists of three analysis tools *Evidence-Lookup*, *Geneset Overlap Browser* and *Network/Metric-Browser*. The *Evidence-Lookup* allows to investigate regulatory bindings and literature references from the *croco-repo* for a given TF-TG pair. Figure 5.5 shows an example output of the *Evidence-Lookup* tool for the regulation of the Early growth response protein 1 (EGR1) and the Myc proto-oncogene protein (MYC). For that particular example, several binding site predictions within open chromatin peaks and ChIP bindings are detected within ± 5 kilobase of the five transcription start site (TSS) of MYC. The bindings site can be further filtered by certain criteria like the cell line or tissue. In addition, literature references are given describing regulatory mechanisms between the selected TF and TG and orthologs are provided allowing the investigation of the selected regulation in other species.

The second main feature, is the *Gene-Set Overlap Browser*, which allows to navigate the



Species	Factor (e.g. STAT1, CTCF, EGR1)	Target (e.g. MYC, TP53, NFKB1)
H. sapiens	EGR1	MYC

Predicted and experimental bindings site within 5000 base pairs up-, and downstream of transcription start site (TSS)

Found literature references

Sentence	PubMed ID	Context
4, group A, member 1), NR4A3 (nuclear receptor subfamily 4, group A, member 3) and CMYC (v-myc myelocytomatosis viral oncogene homolog) [15], by RT qPCR		
2099488.0600.0704 SRC and EGR1 regulate MYC transcription factor (TF), which played an essential role in the subsequent time period	18047719	Human
21233203.2.7 The hnRNP K dependence and pattern of ERK cascade activation at the c-MYC locus were different from at EGR-1	21233203	Human

Found database references

Ortholog relations

Ortholog factor	Ortholog target	Specie	Regulatory evidences
EGR1	BMYC	M. musculus	Regulatory Evidences
EGR1	MYC	M. musculus	Regulatory Evidences
SR	DM	D. melanogaster	Regulatory Evidences

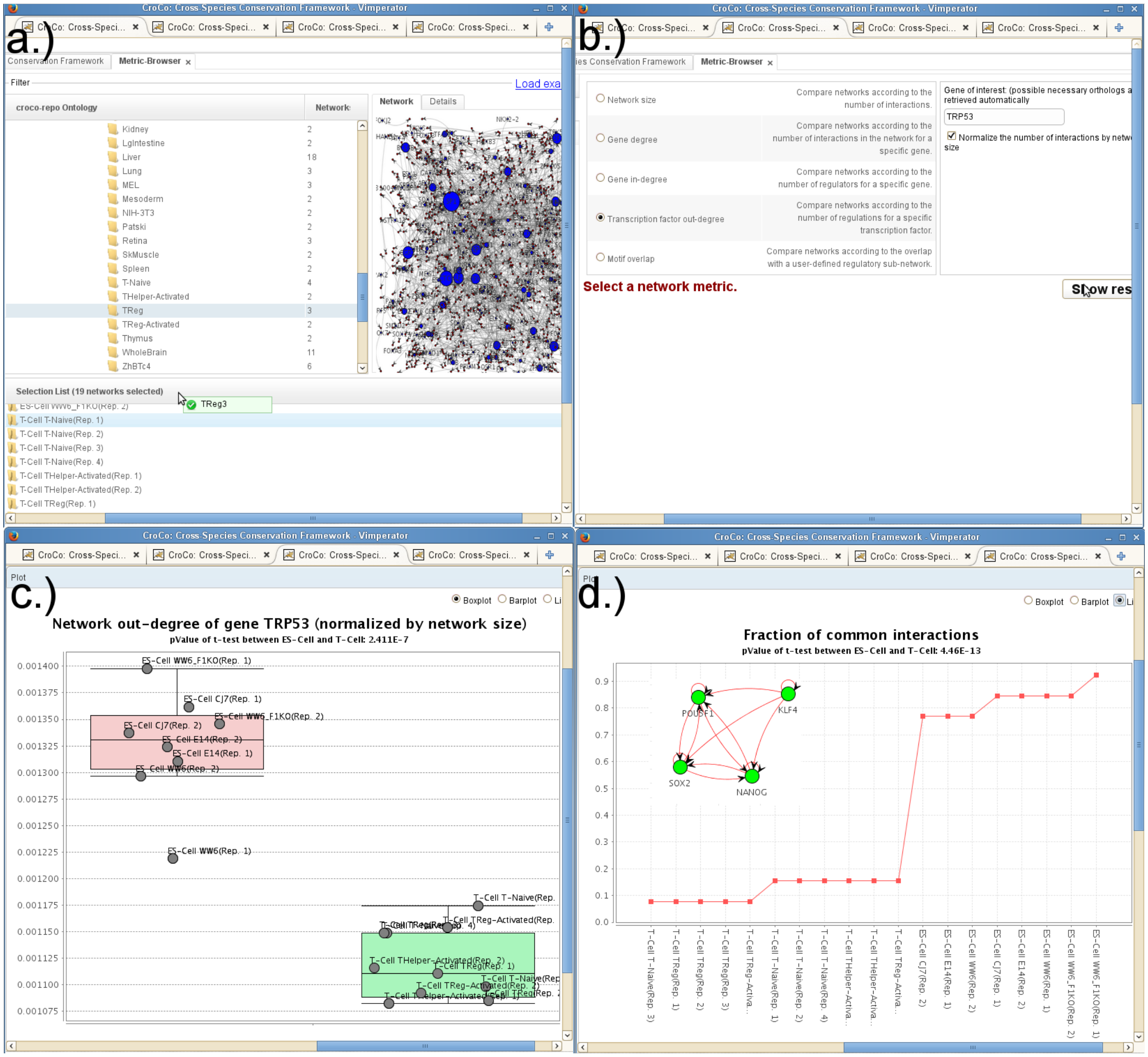
Figure 5.5. The *Evidence-Lookup* shows predicted and experimental binding sites and literature references for a given TF and TG. The screenshot shows the TFBS predicted, ChIP and open chromatin identified binding sites (red rectangles) of EGR1 within ± 5 kilobase of the five annotated transcription start site (TSS) of MYC, the available literature references for that particular regulation, and regulatory evidence between orthologs.

croco-network ontology according to the number of interactions between a user-defined set of genes.

Finally, the *Network/Metric-Browser* allows to perform cross-species comparisons of networks according to several network metrics and to download the networks. With the *Metric-Browser* the

1. network size, represented as the number of inferred interactions,

2. the number of nodes in the derived networks,



○ Network size — Compare networks according to the number of interactions.

○ Gene degree — Compare networks according to the number of interactions in the network for a specific gene.

○ Gene in-degree — Compare networks according to the number of regulators for a specific gene.

● Transcription factor out-degree — Compare networks according to the number of regulations for a specific transcription factor.

○ Motif overlap — Compare networks according to the overlap with a user-defined regulatory sub-network.

Gene of interest (possible necessary orthologs a retrieved automatically

TRP53

☑ Normalize the number of interactions by netw size

Select a network metric.

Figure 5.6. Four embryonic stem cell (ES) and four T-cell networks derived from the mouseENCODE open chromatin experiments with two to four replicates are: (i) compared according to the transcription factor out-degree of the cellular tumor antigen p53 (TRP53), and (ii) intersected with a well studied regulatory sub-network consisting of four pluripotency transcription factors. (a) The network from the *croco-repo* are organized in an ontology and can be selected for network comparisons. In the particular example ES and T-Cells are selected from the repository. (b) Currently, five metrics are available in *croco-web* and can be used to compare previously selected networks. In the example the number of regulatory interactions of the transcription factor TRP53 is selected. The screenshots (c) and (d) show the results produced according to selected metrics: (c) shows the transcriptional activity of TP53 in ES-Cells and T-Cells, and (d) shows the fraction of common interactions of the selected networks and that of a (stem cell related) sub-network.

3. the in-, out- and total-degrees of specific genes,

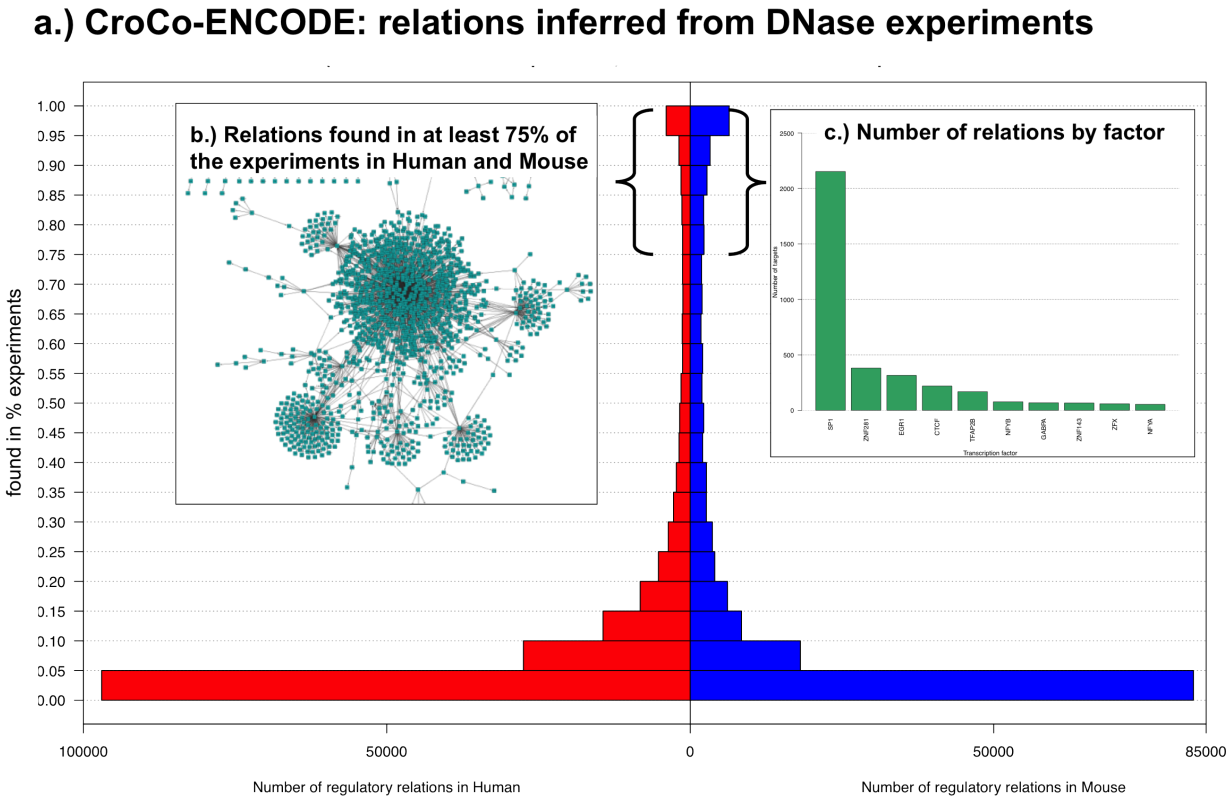
4. the overlap of interactions with a user-defined regulatory sub-network

can be compared between networks. The results of such a network comparison are visualized as boxplots, barplots and lineplots. Any combination of networks from the *croco-repo* can be selected for metric comparisons and if desired the network operation union and intersect can be applied to the selected networks (as the web-application is designed for giving a (fast) first impression on the networks in the *croco-repo*, we limit the complexity of the network operations). Furthermore, selected networks can be organized into groups. This feature supports the structured comparison of sets of networks, for example a collection of stem cells networks can be assigned to one group and a collection of T-Cell networks can be assigned to another group. Furthermore, for the special case that exactly two groups are defined, a t-test between the two groups is performed according to the selected metric value. The required ortholog mappings for the comparison of some features are automatically retrieved. This allows, for example, the comparison of node degrees of specific genes between networks across different species. The regulatory sub-network overlap function allows to investigate the occurrence of regulatory interaction, e.g. regulations between major pluripotency factors, in different contexts. With the *croco-web* tool such regulatory sub-network can either be manually defined, or selected from a pre-defined set of motifs from the *croco-repo* motif repository. As overlap measure for the sub-network overlap metric we use:

$$\text{Overlap}(N, \text{Sub-Network}) = \frac{\text{Number of common interactions between N and Sub-Network}}{\text{Number of interactions in Sub-Network}}.$$

Networks can be ranked according to the overlap of interactions with the sub-network. The feature is inspired by Neph et al. (2012b), who showed that several regulatory sub-networks are highly context-specific.

In Figure 5.6 we use the *Metric-Browser* to compare the transcription factor out-degree and network overlap for several open chromatin experiments derived networks of mice embryonic stem cells (ES-cells) and T-cells. The screenshots shows (a) the steps required for the selection of networks from the *croco-repo*, (b) the selection of a metric and (c,d) results produced for the selected network and comparison feature. For the transcription factor out-degree comparison we selected the Cellular tumor antigen p53 (TRP53). And for the network overlap comparison we selected a regulatory sub-network consisting of four major pluripotency factors with 13 regulatory interactions from Kim et al. (2008). The results produced with the *Metric-Browser* show that the cells cluster, as expected, according to their biological similarity. For example, in T-Cells a lower transcription factor out-degree of TP53 is observed than in ES-cells. Furthermore, a significant overlap with the regulatory sub-networks for the pluripotency factors is only observed in the embryonic stem cells. A selection in a list below the network-overlap statistics allows the visualization of the intersection of the selected sub-network with the select network (not shown in the sceenshot).


a.) CroCo-ENCODE: relations inferred from DNase experiments

b.) Relations found in at least 75% of the experiments in Human and Mouse

c.) Number of relations by factor

Figure 5.7. The histogram shows the cell specificity of the inferred relations from 105 human cell lines (red) and 42 mouse cell lines (blue). In b.), the inferred conserved network between the two species with relations that are observed in at least 75 % of the networks for each species is shown. Furthermore, in c.) the number of relations for the most highly connected transcription factors in the conserved network from the network in (b) is shown.

5.3.5 Cell-Specificity of Regulatory Interactions

As a simple use-case, the cell-specificity of regulatory networks is analyzed, i.e. how often a regulatory interactions is observed in a fraction of networks. We reproduce and extend the analysis by Neph et al. (2012b); we use derived context-specific networks for both human and mouse (228 networks derived from 105 different human cell-lines, and 123 networks derived from 38 different mouse cell-lines/tissues). We observe that the majority of interactions are only observed in a small number of experiments (Figure 5.7a). But also some regulations can be inferred from almost all experiments. This observation is consistent with Neph et al. (2012b), who observed that human TF networks are highly cell selective. The conserved network between human and mouse for relations that could be found in at least 75 % of the experiments for both species is dominated by few transcription factors such as SP1, a known house keeping factor, and CTCF, which is known to be strongly conserved (see Figure 5.7b and c).

As another use-case, we overlap global (literature derived and binding site predicted net-

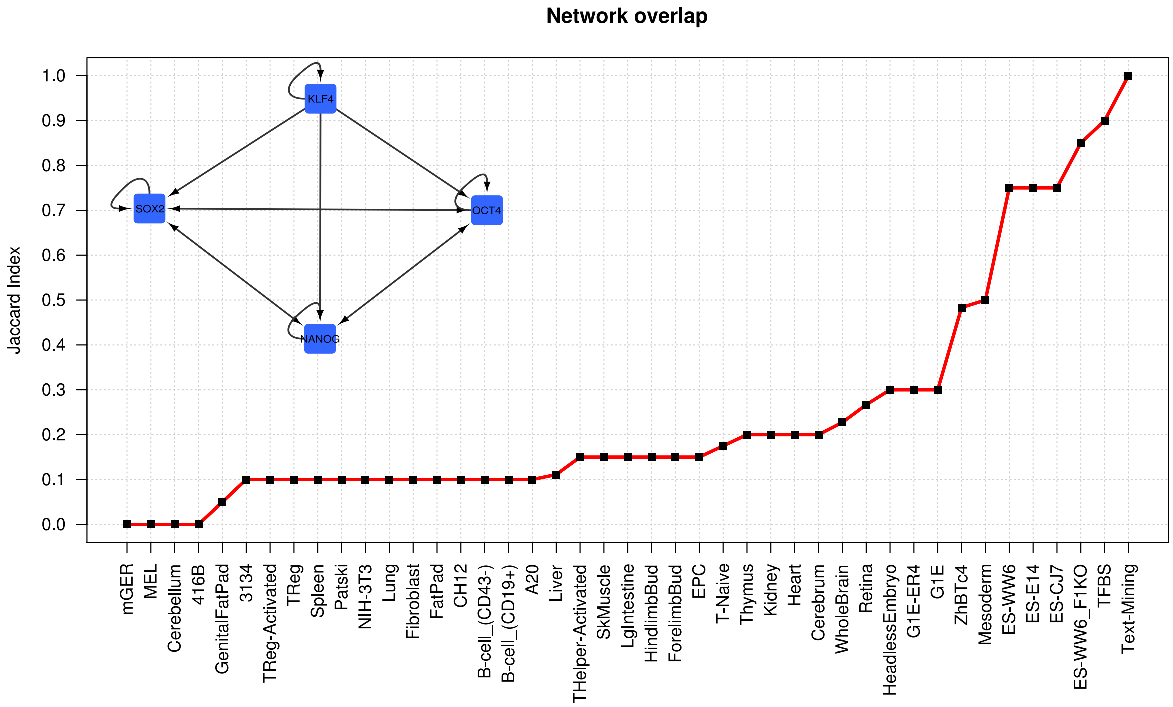


Figure 5.8. Overlap of the regulatory sub-network for major pluripotency factors defined experimentally in mouse ES cells (Kim et al., 2008) with context-species and global regulatory networks from the *croco-repo*.

works) and context-specific networks (DNase-seq derived networks) with an annotated and well characterized regulatory sub-network consisting of interactions between four strongly connected pluripotency factors, which were experimentally validated in mouse embryonic stem cells (ES-cells) (Kim et al., 2008). In Figure 5.8 the overlap of the selected sub-network with the selected global and context-specific networks from the *croco-repo* network repository is shown. In the global literature derived networks evidence for all interactions from the given sub-networks are found. Also 90 % of the interactions from the sub-network are included in the selected computational binding site predicted network. Furthermore, in the context-specific networks derived from different ES cells over 70 % of the interactions from the sub-network can be found. This is of course plausible as the sub-network consists of regulations between pluripotency factors also derived from ES cells. In the majority of the other cell-lines this sub-network is not inferred. In particular no interaction from the given network is found in the two leukemia cell lines MEL and mGER. Analyses like this highlight the strong context-specificity of networks. Also they allow distinguishing between 'housekeeping' sub-networks, i.e. regulatory sub-networks observed in many diverse cell lines, and context-specific sub-networks, i.e. sub-networks only observed in networks derived from specific tissues/cell-lines/conditions.

5.4 Discussion

The different components of CroCo have been designed to: (i) support the identification of possible conservations as well as differences between networks from different species and from different cell lines, (ii) provide a uniform collection of networks for several model organism, and (iii) allow the straightforward navigation through thousands of context-specific networks. CroCo has a particular focus on the ENCODE projects and is tailored to work efficiently with the raw ENCODE data. Use cases for CroCo range from the general comparison of network properties to the validation of hypotheses such as the identification of sub-networks that are conserved or unique in a set of cell lines or species.

CroCo supports the following key functionalities, which we identified to perform the outlined use cases:

ENCODE data: Availability of ENCODE condition-specific regulatory networks in (i) a unified format and a (ii) structured ontology.

Comparative analysis: Fast and efficient network operations on many and large networks for the collection of ENCODE networks to enable comparative network analysis.

Downstream analysis: Downstream network analysis such as network clustering or significant sub-network identification on derived networks from different ENCODE data sets.

Network property look-up: Comparative network property analysis in order to get a first overview on the networks.

Additional information: Query of additional evidence for a particular regulatory relations e.g. query literature evidence for a specific regulation.

Databases like the NCBI SRA and the GEO database (Sayers et al., 2009), modMine (Contrino et al., 2012) and the ENCODE genome browser (Meyer et al., 2013) only provide the raw data itself. Thus, making computational and labour intensive work necessary in order to collect the data and derive networks from them.

In the supplement of the Neph et al. (2012b) publication an interactive web application is presented that allows to visually compare derived regulatory networks for 41 human cell lines. Even though the interactive web site gives a nice overview of the networks, its functionality is limited with respect to comparative analysis and the number of available networks.

Cytoscape in combination with additional plug-ins comes close to implement the needed functionalities provided by CroCo. With the Cytoscape Advanced Network Merge and ID Mapping option networks can be intersected, set-differenced and merged. Available plug-ins allow to perform various downstream analysis including ortholog networks transfers using the Homecat plug-in (Zorzan et al., 2013) and literature queries for protein interactions using the Agilent Literature Search plug-in. Finally, via the Network Analysis feature network properties can be visualized. But there are shortcomings, which limit the usability of currently available Cytoscape plugins for the analysis of many and huge condition-specific regulatory networks. For example, each network needs to be loaded into Cytoscape before network operations can be performed, thus network operations on a large number of networks requires tedious and error-prone manual work. Furthermore, since no comprehensive ENCODE network repository exist, networks need to be manually created, which requires knowledge of the ENCODE data structure and resource intensive computation in order to derive networks.

With CroCo we provide a collection of pre-computed networks (*croco-repo*) derived from ENCODE (ENCODE data) and further external databases organized in an easy to navigate network ontology. Network operations including the transfer of many and large networks from the *croco-repo* can be processed at once. This can be done via the *croco-cyto* plug-in, which makes the repository and the CroCo API functionality available within Cytoscape. Or it can be done directly via the *croco-api* (Comparative analysis; Downstream analysis). Also several network queries including the comparison of network features and the query of regulatory evidence for user-specified TG-TG pairs can be performed directly via *croco-web*, without installing additional software (Network property look-up; Additional information).

The raw data processing workflow, the choice of thresholds and the used data processing tools have an impact on the network model. Currently, the networks in the *croco-repo* are pre-processed, which allows on the one hand side a fast retrial of networks, but limits on the other hand side the network re-definition. In order to provide a higher flexibility, the network construction workflows are included in the *croco-api*, which allows to generate entirely new networks with desired parameters and input data. Additionally, interactions can be filtered using different criteria like the PWM p-value threshold, and the distance to a TSS.

The CroCo system implies several avenues for further research. Possible extensions are the integration of networks from further sources i.e. networks derived from the Roadmap Epigenomics Project (Bernstein et al., 2010), from protein-protein interaction networks, or from proteomics data. Other types of extensions involves the development of approaches for fast, flexible and resource saving redefinition of the networks included in CroCo network repository, i.e. flexible variation of parameters.

5.5 Conclusion

The ENCODE projects (ENCODE, mouseENCODE, modENCODE) and other large-scale compendia such as TCGA and Epigenomics roadmap provide genome-wide annotations for hundreds of cell lines, tissues and treatments using standardized experimental protocols for the model organisms human, mouse, fly and worm. The importance of the ENCODE projects for the scientific community has already been demonstrated in 30 high-impact articles at the end of the second funding phase (Sep 2012) in Nature, Genome Biology and Genome Research. Due to the large amount of data the use and the systematic analysis of ENCODE data is not straightforward as systematic cross-species and cross-condition comparative analysis requires a lot of cumbersome work as well as local storage to download and process the data. Network representations can be employed yielding intuitive abstracted views on the data and allowing the investigation of regulatory mechanisms. Available tools such as Cytoscape contain a wide range of interesting network analysis functionalities. In order to use them in combination with ENCODE data, some time-consuming manual work is required for downloading, preprocessing, and deriving of network models. With the CroCo system systematic analysis of networks is supported via a network repository, which contains thousands of global and context-specific networks. An accompanying software tool suite implements access and network analysis feature to those networks, which can be a starting point for further downstream analyses. The modular design of CroCo and the wide-range of query operations directly via the web application (*croco-web*), via Cytoscape (*croco-cyto*), and via an Application Programming Interface (*croco-api*) provides access to networks and analysis tools to a broad community. Finally, CroCo features the data cube paradigm, which allows for: (i) convenient multi-dimensional navigation in any order of the dimensions, (ii) the ontology-supported browsing of the data cube dimensions, (iii) combination, and (iv) comparison of networks including cross-species transfer of regulatory network models.

Chapter 6

Isoform Structure Alignment Representation

Abstract The structure of eukaryotic genes is complex. Many coding sequences have been observed and are being observed through various experimental techniques. A convenient and comprehensive cross-species representation of genes, their isoforms, and their exon-intron structure is needed for understanding the function(s) and evolution of genes. State-of-the-art Multiple Sequence Alignment (MSA) approaches fail to produce such a representation, as they are unaware of the interrelationships between isoforms, thereby producing misleading alignments. We address this issue by introducing the Isoform Structure Alignment Representation (ISAR), a gene, isoform, and exon-intron structure aware representation of isoforms from sets of orthologous and paralogous genes. An efficient algorithm constructs such a representation from large sets of gene and isoform sequences by successively integrating highly confidence alignments and constraints in the alignment process. The approach is based on partially ordered sets and novel operations to query aligned and not aligned regions, allowing to represent maximal consistent alignments in a sparse graph data structure. We compute a comprehensive collection of 16,066 ISARs containing isoforms of orthologous and paralogous genes from 10 species ranging from yeast to human. An analysis of the gene structure conservation of exon skipping events reveals conserved, lineage- and species-specific alternative splicing events. ISARs allow for the systematic analysis and in detail exploration of the exon-intron structure across large set of phylogenetic taxa and the efficient prediction of new isoforms across phylogenetically distant species. Finally, ISAR is fast enough to be practically applied to very large gene isoform/transcript sets.

Publication: A manuscript of this chapter is in preparation.

My contribution: I developed the ISAR algorithm, performed the event conservation analysis and drafted the manuscript.

Contribution of co-authors: Ralf Zimmer inspired the work by suggesting the recursive dynamic programming approach for the isoform alignment problem and drafted the manuscript. Gergely Csaba implemented the ISAR partial order data structure, and assisted with the event conservation analysis.

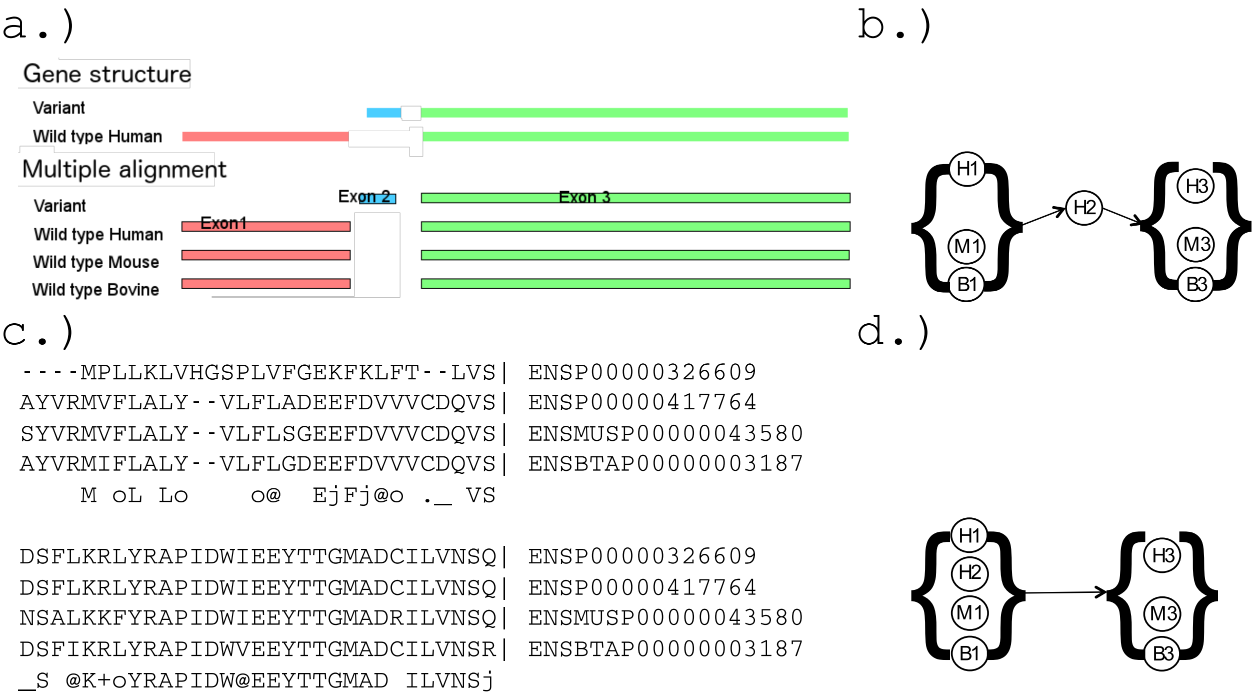
6.1 Introduction

The gene structure of eukaryotic genes appears to be much more complex and complicated as previously thought. Due to the increasing number of sequencing reads from various high-throughput techniques, it can be observed that for almost any gene many gene products are possible and are actually being produced in various expression contexts. Several processes such as alternative transcription and alternative splicing are sources of a high diversity of gene products, called **(alternative) isoforms** (Pal et al., 2011; Kelemen et al., 2013). Many of these products do have specific functions in specific contexts. Alternatively spliced isoforms may lead to different cell specializations, regulations, and differences in the protein-protein-interaction networks in various contexts and/or species (Barbosa-Morais et al., 2012; Ellis et al., 2012; Kelemen et al., 2013). Even highly improbable, non-trivial splicing isoforms yielding very different protein structures with diverse functions are much more likely than expected (Birzele et al., 2008). However, the regulation, function, and evolution of isoforms remain largely unknown (Merkin et al., 2012). As alternative splicing is prevalent in almost the entire eukaryotic domain and strongly affects the regulation of cells, it is important to understand the extent, distribution, and evolution of alternative splicing. Alternative splicing appears to occur during relatively short evolutionary periods of time (a few million years) and, thus, is frequently lineage-specific and well conserved in only a subset of tissues (Barbosa-Morais et al., 2012; Merkin et al., 2012). Nevertheless, conserved exon skippings in fungi and multiple vertebrates — species separated by over one billion years of evolution — could be experimentally identified (Awan et al., 2013). Specific estimations of the conservation of alternative isoforms are subject to the used data set and the protocol to identify conserved isoforms. In order to analyze the conservation of gene structures, isoforms, splicing events, and evolutionary changes of the exon-intron structures of genes, a comprehensive representation of all known isoforms for a group of related genes is needed.

We define such a representation as **Isoform Structure Alignment (ISA)**, which is a multiple alignment (or equivalent) representation of a set of isoform sequences from a set of genes and species, which defines and exhibits all the relationships between the regions of the isoforms implied by the processes of producing alternative gene products (**isoform consistency**). Typically, the set of isoforms to be aligned is large, and thus, the underlying structures are complex. Therefore, highly efficient and practically applicable methods producing interpretable results are required. Even a perfect multiple alignment would not be enough, tools facilitating the understanding of gene and isoform structures are required, for example, tools for visualization, statistical analyses, and exploration.

The ISA problem exhibits a number of characteristics, which demand tailor-made solutions: First, due to the processes of generating (alternative) isoforms, shared regions are often highly similar or even identical if they stem from the same gene or a close paralog. Second, large parts might be missing (gapped out) due to alternative splicing of transcripts and lineage-specific exons. And, third, the relationships in-between the input isoform sequences are often known due to known gene, paralog, ortholog, and phylogenetic annotations derived from genome databases.

Practical solutions to the problem hardly exist. The closest current approaches to the



```
----MPLLKLVHGSPLVFGEKFKLFT--LVS| ENSP00000326609
AYVRMVFLALY--VLFLADEEFDVVVCDQVS| ENSP00000417764
SYVRMVFLALY--VLFLSGEEFDVVVCDQVS| ENSMUSP00000043580
AYVRMIFLALY--VLFLGDEEFDVVVCDQVS| ENSBTAP00000003187
    M oL Lo    o@  EjFj@o ._ VS

DSFLKRLYRAPIDWIEEYTTGMADCILVNSQ| ENSP00000326609
DSFLKRLYRAPIDWIEEYTTGMADCILVNSQ| ENSP00000417764
NSALKKFYRAPIDWIEEYTTGMADRILVNSQ| ENSMUSP00000043580
DSFIKRLYRAPIDWVEEYTTGMADCILVNSR| ENSBTAP00000003187
_S @K+oYRAPIDW@EEYTTGMAD ILVNSj
```

Figure 6.1. Example of a problematic multiple sequence alignment of isoform structures from orthologous of Alpha-1,3/1,6-Mannosyltransferase (ALG) in human, mouse and bovine. In (a) and (b), the gene structure of the two isoforms from the human gene and the expected multiple alignment of the orthologs in human and mouse given the relationship between the alternative human isoforms is shown. Exon 1 and exon 3 are apparent in all three genes with almost identical sequence. Only exon 2 is unique for human. In (a) the expected multiple alignment projected to the exons of the genes is shown, and in (b) the expected alignment as Hasse diagram, a summarized representation of the associated between regions in the alignment, is shown. As multiple alignment methods are not aware of the relationship between the isoforms, an incorrect isoform alignment is typically constructed (c) and (d). Exon 2 is aligned with exon 1 and exon 3 and not as it should be gapped-out.

problem are complex workflows that try to enhance annotation confidence (Yandell and Ence, 2012). A recent approach addressing the ISA problem (at least to some extent) is the gene-structure-aware extension of the Multiple Sequence Alignment (MSA) tool PRRN (Gotoh et al., 2014), which additionally scores for aligned exon boundaries. However, MSA methods (Gotoh et al., 2014; Thompson et al., 1994; Notredame et al., 2000; Edgar, 2004; Löytynoja et al., 2012; Katoh and Standley, 2013) do not guarantee isoform consistency, as they consider each input sequence individually (including the previously mentioned gene-structure-aware method). In Figure 6.1, a symptomatic problem of multiple alignment approaches is apparent for the simple setting of aligning two isoforms from a human gene and one isoform from an orthologous bovine and mouse gene. Because of the peculiarities of scoring and gap penalties, the first and second human mutually exclusive exons are aligned to each other. Although exon 2 is quite short, it considerably distorts the alignment in this region and conceals the correct isoform structure alignment. Such errors are symptomatic and can be expected quite often with currently used alignment approaches as these approaches are not aware of the relationships between the isoforms (see Figure 6.7 for a comprehensive analysis of such conflicts). In order to avoid such problems, alternative isoforms are typically not considered for the MSA computations (Villanueva-Cañas et al., 2013). Furthermore, it remains unclear whether exon 2 is specific for human or just not annotated for the bovine and mouse gene. Spliced aligners such as EXALIGN (Zhang and Gish, 2006) and SPALN2 (Iwata and Gotoh, 2012) align a query sequence to a target genome and are, thereby, theoretically capable of mapping entire gene structures and completing the gene and isoform annotations. But these methods suffer from severe limitations as they transfer isoforms based on the input sequences alone, that is, they do not incorporate (often reliable) available annotations and perform only a pairwise transfer. Therefore, they suffer from multi-species inconsistencies. The same (with even greater impact) holds true for entire genome alignments as integrated in the UCSC Genome Browser (Miller et al., 2007).

Here we present the Isoform Structure Alignment Representation (ISAR) system, a solution for the ISA problem, that is, for the elucidation of relationships within (large) sets of isoforms from orthologous and paralogous genes. The system includes an efficient method for the computation of isoform relations in the form of a partial order representation (the ISAR) generalizing multiple alignments. An ISAR is built such that the genomic annotations are observed, and within these restrictions, the sequence similarity between the input sequences is optimized. We provide a general framework to convert any alignment into an ISAR, so that even the incorrect alignment as shown in Figure 6.1 can be 'repaired'. Finally, the system includes tools to visualize the resulting representations and to query aligned and non-aligned regions. We apply the system for the identification and prediction of conserved spliced events in a comprehensive set of genes from 10 eukaryotic species. Also we analyze the CPU time for the computation of ISARs and compare the gene coverage of computed MSAs with ISAR and state-of-the-art multiple alignment methods.

6.2 Materials and Methods

An Isoform Structure Alignment Representation (ISAR) is built from a given set of genes and isoforms together with their genomic location and their exon-intron structure, and a set of alignment oracles, i.e. state-of-the-art alignment algorithms, which produce suggestions for alignments between the isoforms. The isoforms may stem from the same gene, from paralogos within the same species and from orthologs and their paralogs in different species. ISAR is based on a partial order set (poset) of the sequence positions (which have been proposed for the MSA problem (Lee et al., 2002) and are used by some recent alignment approaches such as PAGAN (Löytynoja et al., 2012)). The advantage of a poset representation is that clearly matching regions can be aligned, whereas uncertain parts of the alignment can remain unaligned. Thus, posets are well suited for representing isoform structure alignments, as it presumably consists of clearly aligned/matched (unified) elements (e.g. conserved exons) and on unaligned (unordered, incomparable) elements otherwise (lineage-specific exons).

Starting with only the total orders of the individual input sequence positions and the relationships between alternative isoforms, orderings and alignment constraints are induced by successively matching positions in different sequences. Alignment suggestions are provided by the alignment oracles, which are applied with respect to the alignment restrictions. These suggestions are ranked and inserted in the poset. In the following, we describe our poset framework, and the algorithm employed for the successive and constrain-based extension of ISARs. Finally, we describe a general approach for the identification and conservation classification of alternative splicing events using the proposed ISAR data structure.

6.2.1 Partially Ordered Sets

Formally, a (strict) poset $PO = (P, <)$ consists of a set of elements (P) and a binary relation ($<$). A poset PO satisfies for all $a, b, c \in P$ the following conditions:

- not $a < a$ (irreflexivity),
- $a < b$ then not $b < a$ (antisymmetry),
- $a < b$ and $b < c$ then $a < c$ (transitivity).

Elements x, y with either $x < y$ or $y < x$ are called *comparable* and *incomparable* otherwise. In our poset realization, each gene is considered as linearly (totally) ordered sets of (local) genomic positions. Alignments introduce matches of genomic positions, which as a result introduce a unified set consisting of the matched positions. Both genomic positions and unified sets of genomic positions are called elements and constitute the base set P of the partial order. The unified sets in P inherit all the $<$-relations from its elements. Thus, only incomparable genomic coordinates can be matched in order to maintain consistency. In the following, we describe the (recursive) alignment process that constructs a chain of posets $(P_i, <_i)$ where both P_i and $<_i$ change due to the matching of further positions.

6.2.2 Partially Ordered Sets Representation and ISAR Construction

For the ISAR, we implement such a partial order representation as a time and memory efficient Directed Acyclic Graph (DAG) data structure based on a minimal edge representation of the partial order. The data structure features constructions functions such as adding isoforms, merging of elements (extending the binary relations), adding of elements (extending the set P) and obtaining the as yet unmatched regions (retrieve incomparable regions) and various output and visualization functions. The binary relations of the posets are represented as edges and the elements of P are represented as nodes in the graph structure, accordingly. A node (element in P) represents (a set of matched) genomic positions.

An ISAR is initialized from a set of n genes with length $1 \ldots \mathsf{L}(G_i)$, i.e. $1 \ldots G_{\mathsf{end}} - G_{\mathsf{start}}$ for genes located on the plus-strand and $1 \ldots G_{\mathsf{start}} - G_{\mathsf{end}}$ for genes located on the minus-strand.

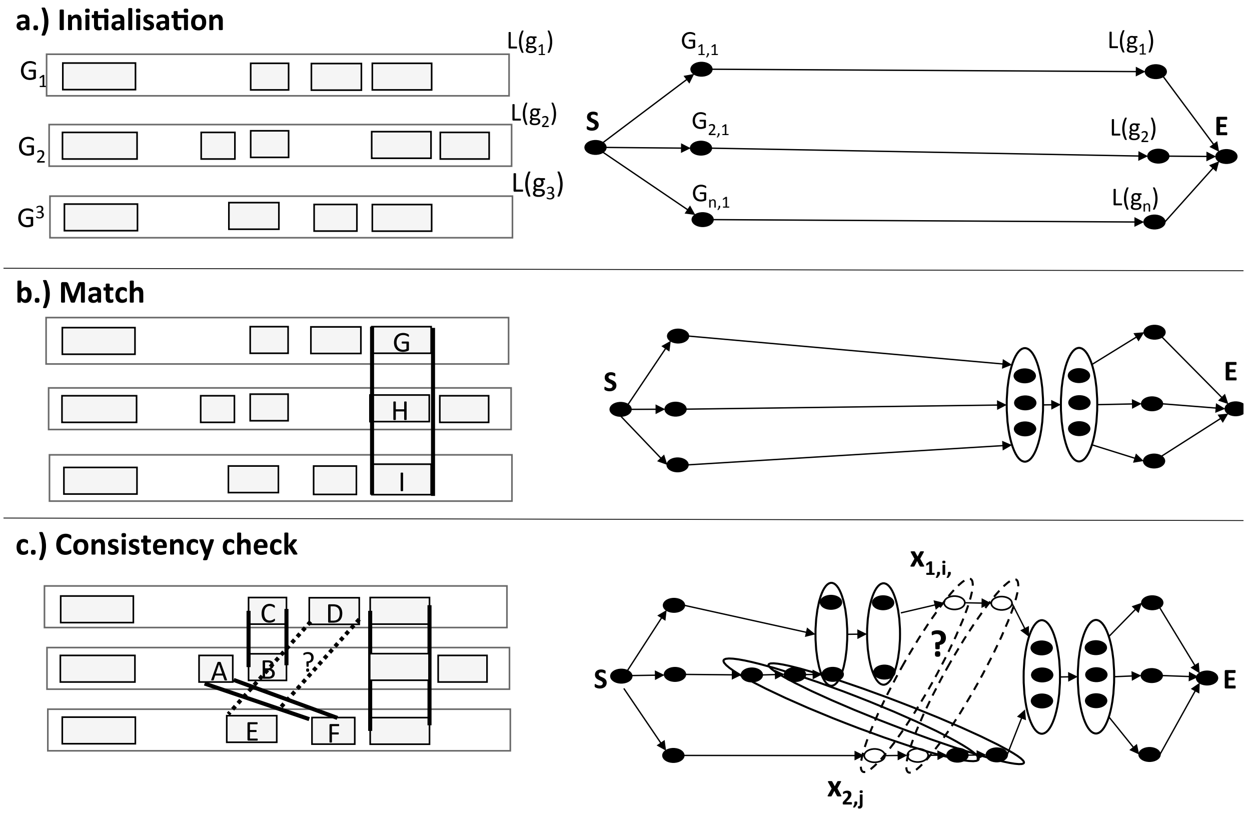


Figure 6.2. Gene structure and internal graph structure for essential ISAR operations. On the left we show matched regions based on the gene structure and on the right we show the corresponding internal ISAR graph structure at the initialization stage and after the matching of regions. Furthermore, we show a schematic overview of the consistency check performed before the matching of regions. (a) At the initialization stage, for each gene the start and end points are added to the ISAR graph structure. (b) For matched regions, for example, exons G, H, and I, a new poset for the start and the end positions of the region is (transitively) extended or newly constructed and integrated in the current ISAR graph structure. As not all matches are allowed, a consistency check is required before regions can be matched in an ISAR. Consider for example an ISAR where exons D and E and exons A and F are matched as depicted in (c). Due to transitive inconsistencies in this graph structure the regions D and E cannot be matched.

In the following, we consider positions within $1 \ldots L(G_i)$ as genomic positions. It is apparent that these local positions can be converted back to the real global coordinates with the G_{start}, G_{end}, strand, and chromosome information. As we implement a sparse representation of the poset, initially only the start (position 1) and end positions (position $L(G_i)$) for each gene are added to P. Furthermore, only a total order of the individual genomic positions is given without any edges between the genes. Thus, for example, for three genes G_1, G_2, G_3 a graph structure such as shown in Figure 6.2a is constructed. An isoform of gene G_i consists of one or multiple exons (regions) within the range 1 to $L(G_i)$.

The ISAR is refined with matched regions between genes derived from the alignment of the isoforms, i.e. alignments between (i) the coding regions and (ii) coding and intronic regions between genes. A matched region may represent any region within the genes such as an exon, an amino-acid, or a nucleotide. Following the sparse representation of posets, we also only add the start and end positions of matched (aligned) regions to the ISAR by default. However, also note that a complete matching of all aligned positions within a regions is possible allowing to produce base-pair and amino-acid precise MSA outputs. For two elements $x_1, x_2 \in P$, the match operation modifies P by adding a new element $\{x_1, x_2\}$ to P, which inherits all previous relations for x_1 and x_2 and removes the individual x_1 and x_2 elements (just to keep the number of edges as small as possible). The effect of extending the ISAR by such a match is shown in Figure 6.2b, where the matched elements (the aligned regions) are unified in a set that inherits all the <-relations from its elements.

The extension of the ISAR by additional matches is in general straightforward, but the new match has to be checked for consistency with the matches already represented in the ISAR. For example, Figure 6.2c shows a hypothetical case of a match, which is inconsistent with other matches; from the figure it is clear that $A < B = C < D$, and $A = F$ holds, now if we match $D = E$, this would imply $A = F < D = E$, i.e. $F < E$ an obvious contradiction to the linear order $E < F$. The consistency check in ISAR checks, whether the start and end points (x_1, x_2) of the regions D and F can be matched.

Given the sets of preceding and succeeding elements for x_1 ($\mathrm{PRE}_1 = (x \in P | x_i < x_1)$, $\mathrm{SUC}_1 = (x \in P | x_1 < x_i)$) and the analogous set PRE_2 and SUC_2 for x_2 (see Figure 6.3), the consistency check can be done via

$$(\mathrm{PRE}_1 \cap \mathrm{SUC}_2 = \emptyset) \wedge (\mathrm{PRE}_2 \cap \mathrm{SUC}_1 = \emptyset) \wedge \\ (x_1 \notin \mathrm{PRE}_2 \cup \mathrm{SUC}_2) \wedge (x_2 \notin \mathrm{PRE}_1 \cup \mathrm{SUC}_1),$$

i.e. checking that the respective PRE and SUC sets are disjunct (the respective sets would be transitively consistent after matching x_1 and x_2) and that the elements x_1, x_2 are not element in $\mathrm{PRE}_2 \cup \mathrm{SUC}_2$ and $\mathrm{PRE}_1 \cup \mathrm{SUC1}_1$ (in which case either x_1 or x_2 would lead to a direct inconsistency), respectively.

6.2.3 Extract Unmatched Regions

The operation to obtain yet unmatched (unordered) regions and the resulting possible consistent matchings between these regions from the ISAR data structure allows the extension of ISARs in an isoform-consistent manner (in the following this operation is called get_unmatched).

The basic mode of extracting yet unordered regions from an ISAR returns all regions between consistent matches in the ISAR (see for example region R_1 in Figure 6.4a). These regions can then be aligned with the alignment oracles, but due to transitive relations induced by matchings with further genes (the dashed lines in Figure 6.4) it is not guaranteed that

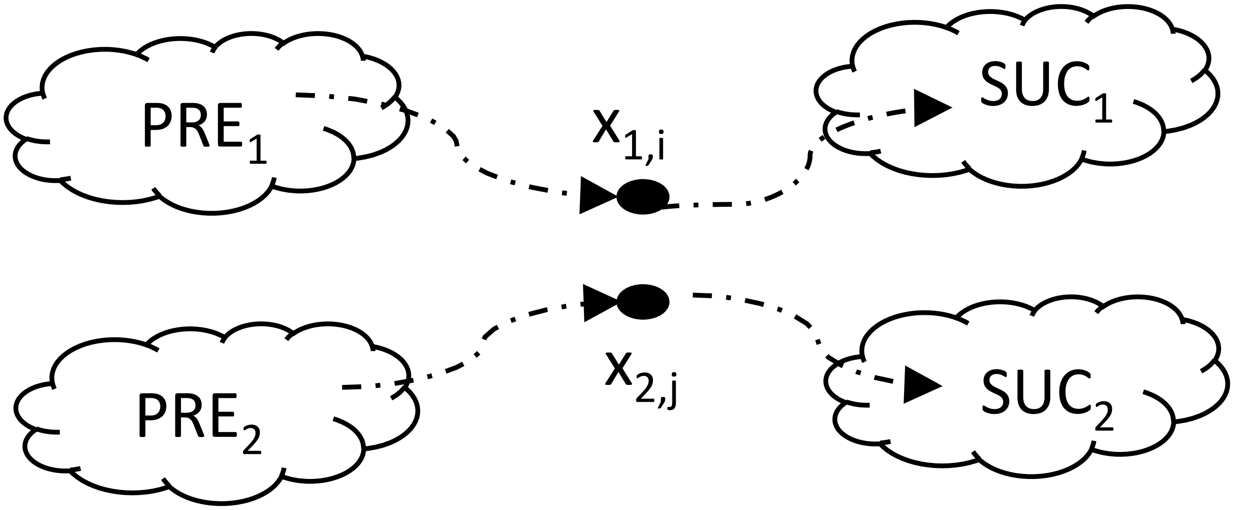


Figure 6.3. Sets considered for the consistency check of a match of two points ($x_{1,i}$, $x_{2,j}$) for genes G_1 and G_2. For the consistency check the set of elements before $x_{1,i}$ and $x_{2,j}$ (PRE_1,PRE_2) and the set of elements after these points (SUC_1,SUC_2) are retrieved from the ISAR data structure. The points can only consistency be matched when no (transitive) relation is conflicted. That is, the PRE and SUC sets are disjunct ($\mathrm{PRE}_1 \cap \mathrm{SUC}_2 = \emptyset) \wedge (\mathrm{PRE}_2 \cap \mathrm{SUC}_1 = \emptyset$) and $x_{1,i}$ and $x_{2,j}$ is not already matched in the others PRE and SUC sets ($x_1 \notin \mathrm{PRE}_2 \cup \mathrm{SUC}_2) \wedge (x_2 \notin \mathrm{PRE}_1 \cup \mathrm{SUC}_1$).

the oracles produce alignments consistent with the already computed ISAR. For example a match between $(G_{1,1}, G_{2,3})$ would be inconsistent as $G_{1,1}$ precedes $G_{2,3}$.

In order to allow for more sensitive matches, we therefore also provide a consistent get_unmatched mode that computes all region pairs from the current ISAR for which any alignment will be consistent with the ISAR. The rationale behind this partitioning is to produce region pairs, which necessarily allow for consistent alignments. Thus, if acceptable and consistent alignments are existent, they will be considered under this mode of operation. Therefore, this mode is used as a second phase in the ISAR algorithm. The query and resulting unmatched regions (R_2, R_3, R_4) for this mode is depicted in Figure 6.4b for two query genes G_1 and G_2. In order to identify all possible consistently matching regions, we first generate all transitive relations between G_1 and G_2 by enumerating all shortest paths in the current ISAR starting at $G_{1,0}$ and ending at $G_{2,n}$ or vice versa. Given such a completed poset graph, consistently matched regions start at an incoming edge (extended to the left up to the next point) in one gene and end at an outgoing edge towards another gene (extended to the right up to the next point). Thus, the possible consistent alignment regions are determined by the actual pattern of in and out-going edges. The result is not symmetric such that the return value is an union of the two calls get_unmatched(G_1, G_2) and get_unmatched(G_2, G_1). The maximal partner region is determined such that no $<$ constraint is violated by an alignment of this region. Thereby, any alignment produced by an

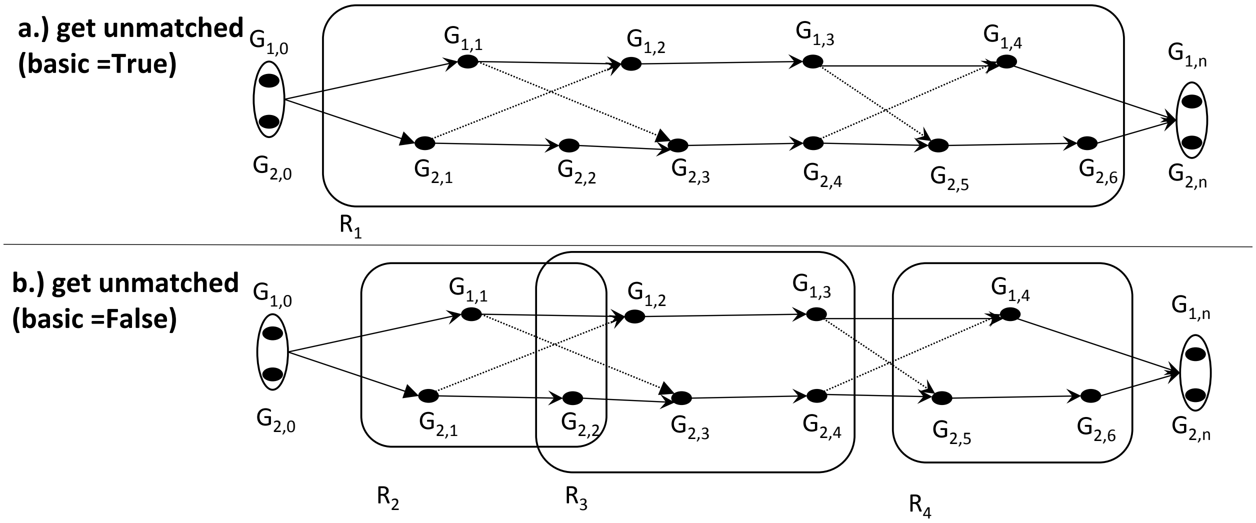


Figure 6.4. get_unmatched (unaligned regions) between two sequences from the ISAR data structure. The figure shows two genes G_1 and G_2 with matched positions at $(G_{1,0}, G_{2,0})$ and $(G_{1,n}, G_{2,n})$. Between these matched positions only transitive relations (dashed lines) for the points between $G_{1,0}/G_{2,0}$ and $G_{1,n}/G_{2,n}$ are given. a.) The basic get_unmatched method just returns the entire regions between matched regions (R_1). But due to the transitive relations not all possible alignments/matchings within this region would be consistent e.g. a matching between $(G_{1,1}, G_{2,3})$ would be inconsistent as $G_{1,1}$ precedes $G_{2,3}$. b.) The sensitive get_unmatched method takes into account the transitive relations and returns instead three (partly overlapping) unmatched regions R_2, R_3 and R_4, where all alignments will be (individually, i.e. for each region) consistent with the current ISAR.

oracle for this region pair will be consistent. This guarantees that even if all the oracles are applied to the whole unmatched region and only produce inconsistent alignments, consistent alignments for certain sub-regions are obtained nevertheless (any alignment in this region is consistent, whether it is worthwhile to be included depends on the actual quality and score of it). Of course, to determine these consistent region pairs requires effort, but also help to maximize the consistent matches (avoid premature stop).

6.2.4 ISAR Algorithm and Alignment Oracles

The ISAR system is based on a simple approach (similar to the Recursive Dynamic Programming (RDP) approach proposed for multiple protein threading (Thiele et al., 1999)): starting from a set of genes with isoform annotations from different species but also from paralogs from the same species, several oracles are applied to generate initial (highly confident) alignments. These alignments dynamically define regions and imply mappings of (some of these) regions from (a subset of) the isoforms, thereby introducing further constraints for the remaining alignment. Every mapping of these regions partitions the original isoform

Listing 6.1. The ISAR algorithm. The algorithm constructs a ISAR from a set of isoforms employing interchangeable oracles to generate candidate alignment regions.

```
 1  /* init*/
 2  isar := new ISAR;
 3  R := False;
 4  SoG := set of genes with isoform annotations;
 5  for all (SoG s)
 6          /* insert returns true if consistent */
 7          R := R or insert(isar, s);
 8
 9  /* phase I+II: use simple or sensitive unmatched_regions */
10  for basic in {true, false} do
11          for oracle in {oracles} do
12                  while (R) do
13                  begin
14                  /* step 1: alignment suggestions */
15                          U = get_unmatched(isar, basic);
16                          A := oracle(U, SoG);
17                  /* step 2: define regions */
18                          B := partition(A);
19                          BS := score(B);
20                          SBS := sort(BS);
21                  /* step 3: modify ISAR */
22                          R := False;
23                          for all (SBS b)
24                                  R := R or insert(isar, b);
25                  end
26          end
27  end
```

alignment problem into respective sub-problems, which can be obtained from the ISAR and are handled recursively with the same procedure until no more confident region alignments can be found for the remaining sub-problem instances. These remain as unmatched and unordered regions.

With the previously introduced poset data structure and the implemented query operations, this procedure can be easily realized. The ISAR algorithm (see pseudo code in Listing 6.1) initializes the ISAR structure (i.e. the poset data structure) and then recursively applies the following three steps: (i) alignment suggestions (line 15-16), (ii) region definition (line 18-20), and (iii) ISAR modification (line 22-25). The ISAR is initialized with a set of genes harboring isoform annotations as shown in Figure 6.2a. As previously described, only the start and end positions of each gene are inserted into the date structure, initially. The alignment suggestions step consists of a get_unmatched call and subsequent alignment oracle calls to generate suggestions for the yet unmatched regions. The get_unmatched call provides ranges of possible mapped regions for all genes including exonic and intronic regions. As initially no region is matched between genes, the first get_unmatched call just returns for each pair of genes G_i and G_j the entire gene regions, i.e (1,1)-(L(G_i),L(G_j)). After that, the alignment oracles are applied to compute alignment suggestions between the relevant unmatched regions. The second step (region definition) partitions the alignment suggestions from the oracles into small regions representing aligned (parts of) exons. These regions are then scored, filtered, and sorted according to a quality score. Finally, in the last step (ISAR modification), the ISAR is sequentially extended by matched regions with a prioritization according to the selected strategy. Steps (i) - (iii) are repeated as long as new matches can still be identified and inserted. The different steps allow for a range of variants influencing the sensitivity/specificity of matches and/or the prioritization of inconsistent solutions. To allow for more sensitivity, two different modi for obtaining the as yet unmatched regions from the ISAR are applied: the basic method returns the maximal unmatched regions between two matches, and the sensitive method partitions these regions into sub-region pairs for which any new alignment will be consistent. By inserting the region pairs with the highest score first, the ISAR is extended such that inconsistent edges are automatically discarded (not inserted into the ISAR). Thereby, an ISAR as shown in Figure 6.1a and b is constructed instead of the isoform inconsistent one in Figure 6.1c and d. So even 'wrong' (inconsistent) alignments can often be converted to an ISAR and, thereby, corrected with the proposed algorithm.

The oracles are interchangeable components of the ISAR systems that are used to generate alignment suggestions of yet unmatched regions. The alignment suggestions are (filtered and) partitioned into regions (function partion in Listing 6.1), and sequentially inserted into the current ISAR according to a prioritization strategy (function sort in Listing 6.1). Partitioning, filtering, and the order of insertion of alignments/matches into ISAR are subject to different parameters, which can be modified for the ISAR computation. Here, the pairwise alignments produced by the oracles are mapped to the exon structure of the isoforms and partitioned into regions according to the exon annotations so that a region corresponds to an alignment of a (part) of one exon in each sequence. Many aligned regions emerge from each oracle iterations, which may include regions that are inconsistent with the current ISAR (see e.g. Figure 6.1c and d). We rank these regions based on their harmonic mean of the sequence identity and the normalized length in order to add the most reliable regions first into the ISAR and filter the inconsistent alignments. We make use of the following oracles:

MSA initialization oracle: Multiple Sequence Alignments (MSAs) are computed using one representative isoform per gene with PRRN (Gotoh, 1996; Gotoh et al., 2014) using the PRRN gene-structure-aware feature by providing exon annotations for the isoforms. We apply a phylogeny-aware selection of representative isoforms. That is, we traverse the phylogenetic tree and select the isoforms for each species along the given tree, which maximize the sequence similarity. This allows to select a *core* set of isoforms, i.e. the strongest related genes in a given set of genes. For each remaining gene (genes not represented in this tree traversal *core* set), we select the isoform with maximum sequence similarity to any isoform in the *core* set.

Free-shift pairwise oracle: Pairwise free-shift alignments with the Dayhoff matrix and affine gap costs are computed for unmatched regions of the current ISAR. For two genes G_i and G_j the oracle first intersects the unmatched region with the isoform annotations of the two genes in order to derive coding sequences. In the next step, the actual pairwise alignments between the coding sequences in the different genes are computed. Note that, because of alternative exon usage, 3' and 5', intron retention and alternative frame usage, many coding sequences can be defined from an unmatched region.

Spliced alignment oracle: Similar to the pairwise sequence alignment, SPALN2 (Iwata and Gotoh, 2012) is used with cross-species settings to compute spliced alignments in order to infer (not yet) annotated exons in intronic regions. The unmatched regions are intersected in one gene with the isoform annotations and aligned to the target (unmapped) region in the other gene.

6.2.5 Query of Conserved Alternative Splicing Events

ISARs computed as previously described can be used for a wide range of analysis like the prediction of new isoforms, the identification of orthologous and paralogous spliced events and isoforms, and the evolutionary study of exon-intron changes. Here, we describe a systematic approach for the definition of (alternative) splicing events and for the identification and classification of conserved events using the ISAR data structure. We define an alternative splicing event as a tuple of donor (d) and acceptor (a) sites (genomic locations), which are exclusively used in one or the other isoform. Formally, a Splice Event ($\mathsf{SE} = ((d_1, a_1), (d_2, a_2))$) is any pair of donor (d) and acceptor (a) of an isoform, represented as the d/a genomic position within the receptive gene. With this definition Alternative Splicing Events $\mathsf{ASE} = (\mathsf{SE}_1, \mathsf{SE}_2)$ are tuples of overlapping SEs between isoforms of a particular gene. A Transferred Alternative Splicing Event $\mathsf{TASE} = (\mathsf{m}(g_i, g_j, \mathsf{SE}_1), \mathsf{m}(g_i, g_j, \mathsf{SE}_2))$ for a target gene G_j is the projected (mapped) ASE defined for a gene G_i using the projection m, the mapping of positions from one gene to another gene through a look-up in the poset elements P. In order to conduct a comprehensive analysis of alternative splicing events, we classify the TASE according to the following criteria for a source gene G_1 and target gene G_2 (see also Figure 6.5 for a graphical representation of the classification):

Annotated (A): The TASE corresponds to an G_2-annotated ASE.

Gene Structure supported ASE (GS): The mapped donor/acceptor (d/a) sites in G_2 correspond to annotated d/a sites.

Predicted Gene Structure supported ASE (P_GS): The mapped d/a sites correspond to sites flanking novel predicted exons.

Supported by a Novel Intron (NI): The mapped d/a sites lie within an annotated exon.

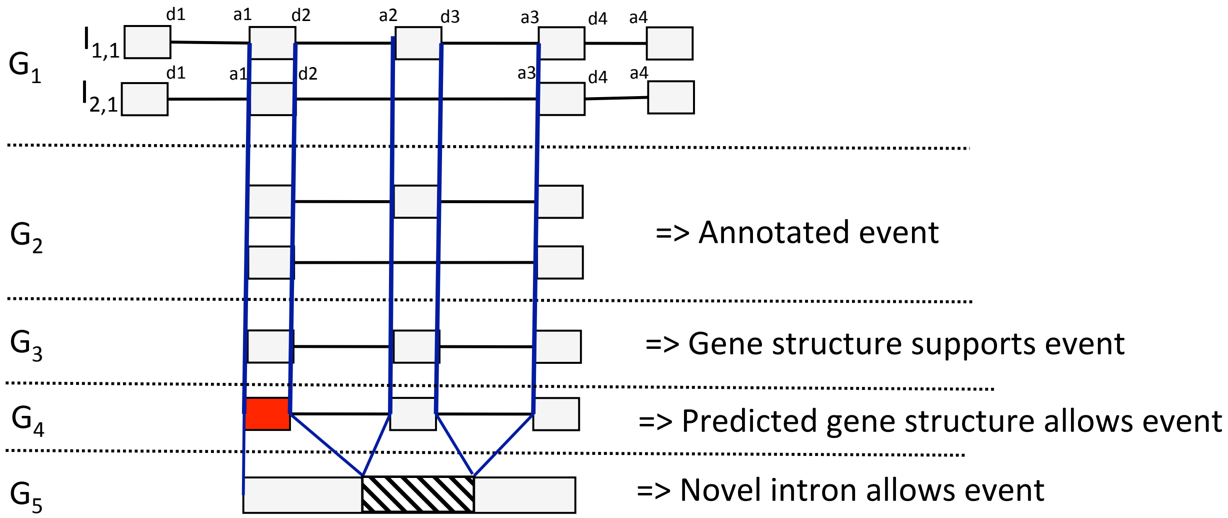

G1
I1,1
I2,1
d1
a1
d2
a2
d3
a3
d4
a4
G2
=> Annotated event
G3
=> Gene structure supports event
G4
=> Predicted gene structure allows event
G5
=> Novel intron allows event

Figure 6.5. Schema for the exon skipping conservation classification. Exon skipping events are defined between alternative isoforms. We show the classification of an event (derived from gene G_1 harboring two isoforms) as: (i) annotated, (ii) predicted, (iii) gene structure supported, and (iv) supported by novel intron in genes G_2, G_3, G_4 and G_5, respectively. In the figure each row corresponds to an isoform, each gray box correspondences to an exon and the blue lines indicate aligned positions. For gene G_2 an orthologous exon skipping event is depicted as for both genes G_1 and G_2 one isoform is observed where the same exon (the second exon) is skipped in one isoform. In gene G_3, the event is not annotated, but the gene structure allows the event, as the exon-intron boundaries are well aligned for this event. Gene G_4 also supports the event, but the boundary of one exon had to be inferred, i.e. the first exon is not annotated, but it could be predicted in the intronic region. Finally, for gene G_5 the event is classified as predicted by novel intron as the event positions align to exonic regions.

For **P_GS** and **NI** we, conservatively, only accept predictions that result in canonical splice sites (donor-acceptor = GT-AG), or fully conserved non-canonical splice site, i.e. the donor and acceptor sides are the same for the ASE and the TASE.

6.3 Results

The ISAR approach has been implemented as a practically applicable tool for a very large isoform and transcript sets. It is accompanied by analysis tools, which allow for the visualization and query of the sometimes surprisingly complicated gene structures across various species. Its intended use is also for the analysis of these structures for the forthcoming new genomic, meta-genomic, and transcriptomic sequencing data sets. ISAR also allows for the systematic visualization, analysis, and in detail exploration of splicing events across large set of phylogenetic taxa. In the following, we build ISARs for a wide range of eukaryotes in order

Table 6.1. Total number of genes and isoforms in the ENSEMBL database and in the computed ISARs. Genes and isoforms from ten species derived from the ENSEMBL database (Flicek et al., 2014) are clustered into Ortholog Gene Groups (OGG) based on their sequence identity resulting in some genes which are not included in an ortholog cluster and are therefore not included in an ISAR.

Species	**Common name**	**PG**	**G_OGG**	**I_ISAR**
S. cerevisiae	Baker's yeast	6,692	1,704 (25%)	1,704
S. pombe	Fission yeast	5,143	1,379 (27%)	1,379
C. elegans	Worm	20,541	3,896 (19%)	5,575
D. melanogaster	Fly	13,937	4,525 (32%)	7,167
T. nigroviridis	Pufferfish	19,602	16,004 (82%)	19,048
G. gallus	Chicken	15,508	13,381 (86%)	14,072
B. taurus	Cow	19,994	19,411 (97%)	21,415
M. musclus	Mouse	23,119	20,266 (88%)	43,517
M. mulatta	Rhesus monkey	21,905	19,940 (91%)	33,254
H. sapiens	Human	23,393	19,493 (83%)	82,533
$\sum$		169,834	119,999 (71%)	229,664

PG=Protein coding genes; **G_OGG**=Genes in OGG; **I_ISAR**=Isoforms in ISARs

to demonstrate the large-scale applicability of our approach. We analyze the CPU time for the ISAR computation and compare the gene coverage of the computed ISAR alignments with alignments computed with different MSA tools. Finally, we employ the constructed ISARs for the identification of conserved exon skipping events.

6.3.1 Experimental Settings

In order to apply the ISAR approach on meaningful sets of isoforms, we use the gene definition from ENSEMBL (v.75) and cluster genes from 10 selected species based on ortholog and paralogs information from ENSEMBL Compara (Flicek et al., 2014; Vilella et al., 2009) in order to define Ortholog Gene Groups (OGGs) for all relevant genes (see Table 6.1). We define an OGG as a set of $n \geq 2$ (orthologous or paralogous) genes, where each gene has at least one isoform with amino acid sequence identity of $\geq 40\,\%$ to another isoform from a different gene in the same OGG, i.e. we apply a single-linkage clustering of the orthologs. Finally, we compute for each of the 16,066 OGGs one ISAR containing all isoforms for the corresponding genes. In total, 119,999 genes with 229,664 isoforms are contained in the computed ISARs. ISARs are computed with the PRRN MSA (Gotoh et al., 2014), free-shift pairwise, and the SPALN2 spliced aligner oracle (Iwata and Gotoh, 2012). The MSA oracle is used together with the free-shift pairwise oracle in order to generate alignment candidates between the annotated isoforms. Only after these oracles are applied, the spliced aligner

Table 6.2. CPU time for the individual steps needed for the ISAR computation. The CPU time is divided into the time needed for the computation of alignment suggestions with the oracles, and the time needed for maintaining the (poset) ISAR data structure.

Type	Operation	Time (h)
Oracles	PRRN Multiple sequence alignment	8.52
Oracles	Pairwise alignments	1.38
Oracles	SPALN2 Spliced alignments	107.89
Data structure	Update unmatched (basic=true)	0.16
Data structure	Update unmatched (basic=false)	1.45
Data structure	Insert regions (consistency checks)	2.18

oracle is applied to infer new exons in intronic regions in order to complete the isoform alignment and gene annotation. As previously described, the MSA oracle is only used for the computation of a MSA between one represented isoform for each gene in an OGG. Thus, this oracle provides only at the first iteration alignment suggestions. The alignment suggestions are partitioned, filtered, and sorted based on their normalized sequence identity and the length. We require that a region must have a minimum sequence identity of 40 % between the (sub-)exon sequences. Additionally at least 10 amino acid matches must be in the region, or at least 50 % of one of the corresponding exons must be covered in the alignment.

6.3.2 ISAR Computational Time

The entire computation of the 16,066 ISARs takes 121.58 hours. Most of the CPU time is spent for the computation of alignment suggestions using the oracles. The computation of alignments with the MSA oracle and the pairwise aligner takes 8.52 and 1.38 hours, respectively. The by far most time-consuming operation is the computation of spliced-alignments, i.e the inference of new exons with SPALN2 that takes 107.89 hours. In addition to that, the overhead of maintaining the ISAR data structure including the query of unmapped regions, the consistency checks and the insertion of mapped regions is comparable small and takes (only) 3.79 hours.

6.3.3 Alignment Gene Coverage

A correct multiple alignment of genes with all known alternative isoforms reveals conserved, species, and lineage-specific coding regions (inserted amino-acids, exons, and parts of exons). And thus, allows (besides other) the identification of conserved and species-specific exons and the transfer of isoforms across species. But, state-of-the-art MSA methods make it difficult to perform such identifications as they are unaware of the interrelationships between isoforms, tend to produce compact alignments, and are, thereby, often misleading. In order to further

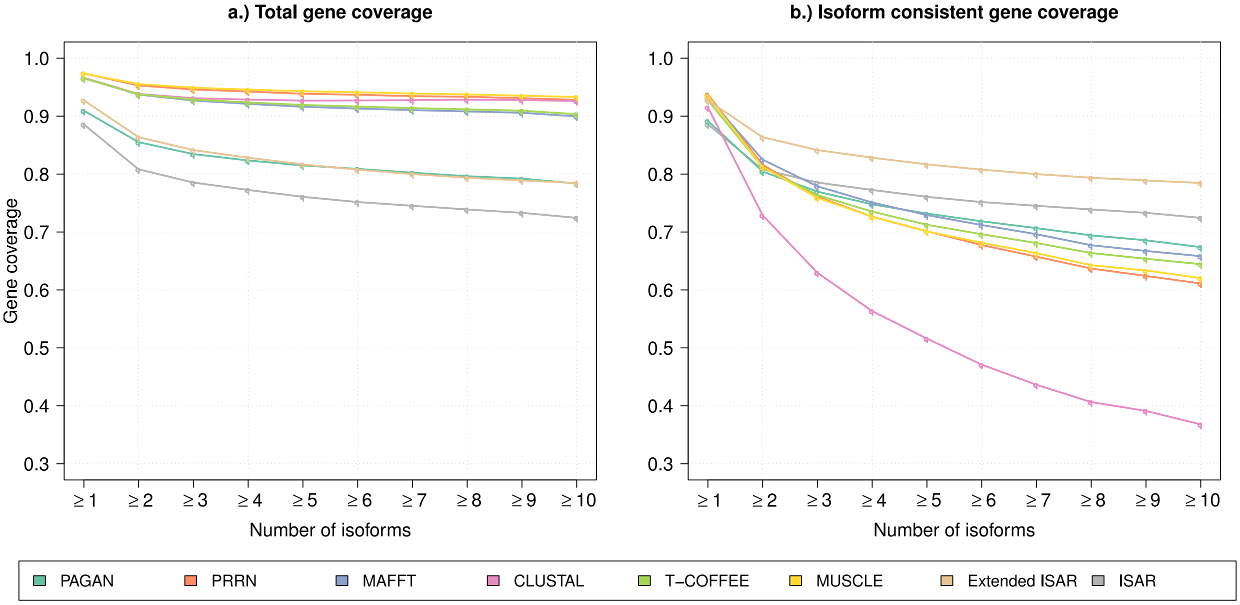


Figure 6.6. Gene coverage of aligned Ortholog Gene Groups with different MSA methods. We show the gene coverage of 16,066 MSAs computed with different MSA methods, ISAR, and extended ISAR for groups of orthologous genes. For each pair-wise gene combination, the alignment coverage defined as the number of aligned amino-acids divided by the total number of amino-acids in all coding regions for the respective gene is depicted. In (a) we show the total gene alignment coverage including conflicting alignment positions. MSA methods are not aware of the interrelationships between isoforms and typically tend to produce compact alignments, thus a high gene coverage is achieved (> 0.9) for methods such as CLUSTAL, T-Coffee, MUSCLE, PRRN, and MAFFT. However, a huge fraction of the aligned positions is wrong, i.e. isoform inconsistent. In (b) we show the correct gene coverage for the considered methods by excluding isoform inconsistent regions from the alignment. Due to the multiple isoform consistent alignment, ISAR is able to produce a more complete alignment.

asses this aspect, we compute MSAs for the previously defined OGGs with CLUSTALW (Thompson et al., 1994), PAGAN (Löytynoja et al., 2012), PRRN (Gotoh et al., 2014), MAFFT (Katoh and Standley, 2013), MUSCLE (Edgar, 2004), and T-Coffee (Notredame et al., 2000) using standard parameters. We compare the MSA methods based on the

$$\text{Gene Coverage}(G_1, G_2) = \frac{\text{Aligned amino-acids of}\, G_1 \,\text{in}\, G_2}{\text{Total amino-acids in}\, G_1}$$

between all pairwise genes contained in these 16,066 OGGs. In addition, we compute a maximum extended ISAR by including speculative alignments in the ISAR computations, that is, we omit the filter step after the definition of the regions. Compared to ISAR and PAGAN, the considered MSA methods produce in general a higher and very uniform total gene coverage of over 90 %, even for the genes with a complex gene structure, i.e. for the genes with many different alternative isoforms (Figure 6.6a). However, when we correct the alignments by removing isoform inconsistent regions, it becomes apparent that this uniform gene coverage produced by most MSA methods is only an artifact (Figure 6.6b). Indeed, ISAR and the Extended ISAR produce MSAs with significantly higher gene coverage. This is expected as the MSAs from gene and isoform unaware methods do have many conflicts (see also Figure 6.7).

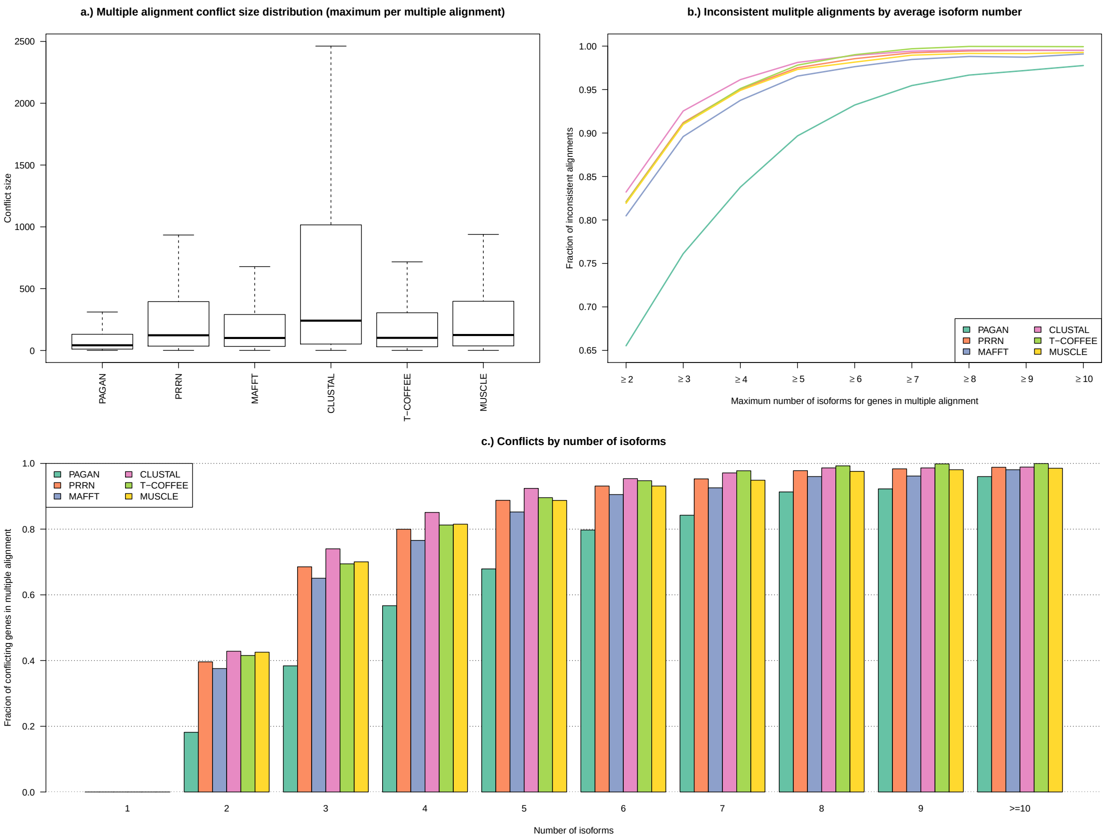


Figure 6.7. We compare the number of isoform-conflicts in 16,066 multiple sequence alignments computed with CLUSTAL, MAFFT, PPRRN, T-Coffee, PAGAN and MUSCLE for ortholog gene groups containing 229,664 isoforms stemming from 119,999 genes. We evaluate the number of isoform conflicts for alternative isoforms in the 16,066 MSAs. We define a conflict as an aligned position of amino-acids from the same genes but located at different genomic positions. The size of a conflict is the number of conflicting position between two isoforms. In **(a)**, the distribution of the sizes of such conflicts for the different MSA methods is shown. In **(b)**, the fraction of MSAs with (at least 10) conflicting amino-acid positions by the maximum number of isoforms per gene in the 16,066 MSAs is shown. Furthermore, in **(c)**, we show the number of conflicts for each gene contained in the MSAs by the number of isoforms (again with a minimum conflict size of 10 amino-acids). PAGAN performs best in our evaluation, but still a huge fraction of 65 % of the MSAs having at least one alternative isoform are inconsistent.

6.3.4 Isoform Structure Alignment Representation

Most isoforms are defined for human and mouse (82,533 and 43,517 isoforms, respectively). For the other species, only comparable few isoforms per gene are annotated, even though the prevalence of alternative splicing, i.e. the prevalence of alternative isoforms, is likely the same for all mammals. Since ISAR gives a complete mapping of all isoforms (i.e. also a complete mapping of the gene structure), a transfer of the isoforms is possible by projecting the splice sites to other genes.

Consider, for example, the (reduced) Hasse diagram and exon-intron mapped MSA visualization of the ISAR for the Ras-related protein Rab-1A ortholog group consisting of 17 genes and 30 isoforms in Figure 6.8. In the Hasse diagram, the matchings of the different exons or parts of exons are represented by the blue lines, whereas in the exon-intron mapped MSA visualization, a complete alignment of the isoforms based on the exons is shown. The first gene in this representation is the query Rab-1A gene from human. This gene consists of eight annotated (black) and one inferred (red) exons. These eight different exons are used in different combinations to produce six different isoforms of this gene in human. In addition, the gene has a close human paralog with two isoforms. This particular example shows a multiple alignment/Hasse diagram for a relatively small and clearly aligned set of isoforms, which indicates that the overall picture can be quite complicated.

This kind of representation enables the formulation of hypotheses of the evolution of gene and isoform structures. In particular, splicing events can be compared and transferred between species, allowing the identification of *orthologous* and *paralogous* splicing events and the prediction of additional isoforms by projecting splicing events between species and genes. Consider, for example, the first and second human isoforms in Figure 6.8. The third exon of the second isoform is skipped in the first isoform. The boundaries of this skipped exon are mapped to other genes as well and thereby isoforms that lack this specific exon can be predicted for some of the other genes. As a simple step towards this direction, we investigate the conservation of exon skipping events using the computed ISARs.

6.3.5 Conservation of Exon Skippings

Alternative splicing events (ASEs) like the previously described exon skipping event for two isoforms of the human Rab-1A gene can be inferred and classified across the genes contained in the ISARs. In Figure 6.9 examples of a transferred and classified alternative splicing event are shown. The input exon skipping event from human is classified as gene structure conserved (**GS**) in worm as the splice event in worm is unknown, but all mapped splice sites are annotated, i.e. no alternative isoform exists with the mapped acceptor and donor site (Figure 6.9b). In fly, however, one acceptor site is mapped within an exon and a canonical acceptor pattern (AG) is observed downstream (Figure 6.9c). Thus, the gene structure is considered to be conserved with respect to the analyzed ASE, and the event is predicted as conserved through a novel intron (**NI**) for fly.

As shown in Figure 6.9 the conservation levels of ASEs across the taxonomy tree can exhibit complex situations. In the sample shown, besides the bony vertebrates (Euteleostomi)

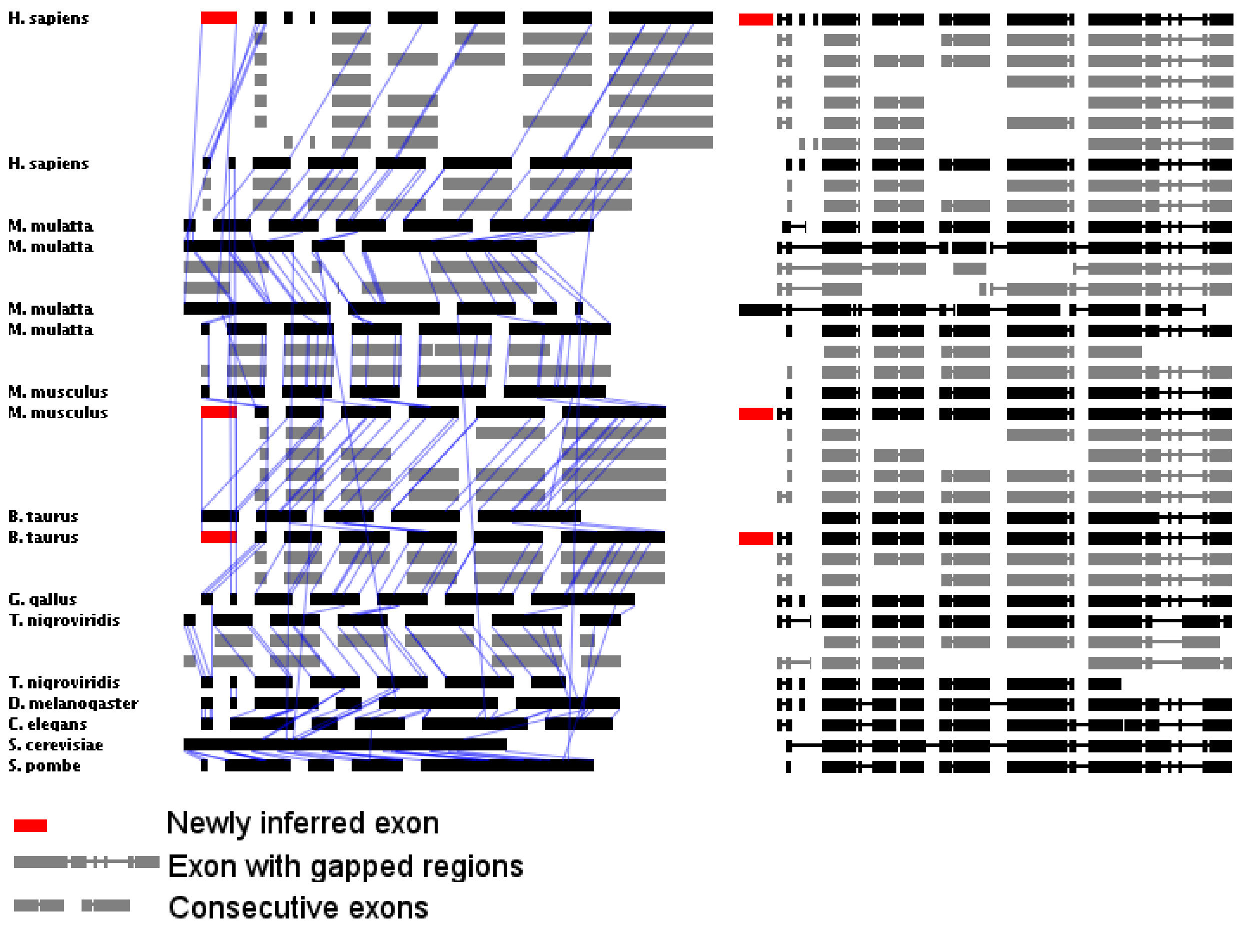


Figure 6.8. Multiple sequence alignment of 17 genes with 30 isoforms from 10 species for the RAB1 gene family. For each gene, we show the gene structure as the union of all exons for all isoforms (first isoform for a gene labeled with the species name), and in the associated rows, we show the isoforms (when more than one is annotated). We show two visualizations (print options of the ISAR): on the left, the ISAR as a partial order graph (Hasse diagram); on the right, the implied ISAR multiple alignment. Vertical blue lines indicate matched positions in the ISAR partial order graph. Red exons indicate inferred exons from other genes.

also protostome animals (Ecdysozoa) support the ASE on the **GS** level using worm as evidence. The most likely explanation in terms of intron loss/gain events for this case would be that the (unlikely) gain of an intron at that very position occurred once in a common ancestor and that the intron afterwards has been removed in sub-trees, for example, for the fly lineage.

Thus, to highlight this, for an ASE and an inner taxonomy node, the maximum conservation level observed in two different branches is assigned in the tree visualization. In our case, we assign **GS** to protostome animals as any leaf from the bony vertebrates and worm is **GS**. The event is also classified as **GS** in fungi (S. pombe) and, thus, also classified as **GS**

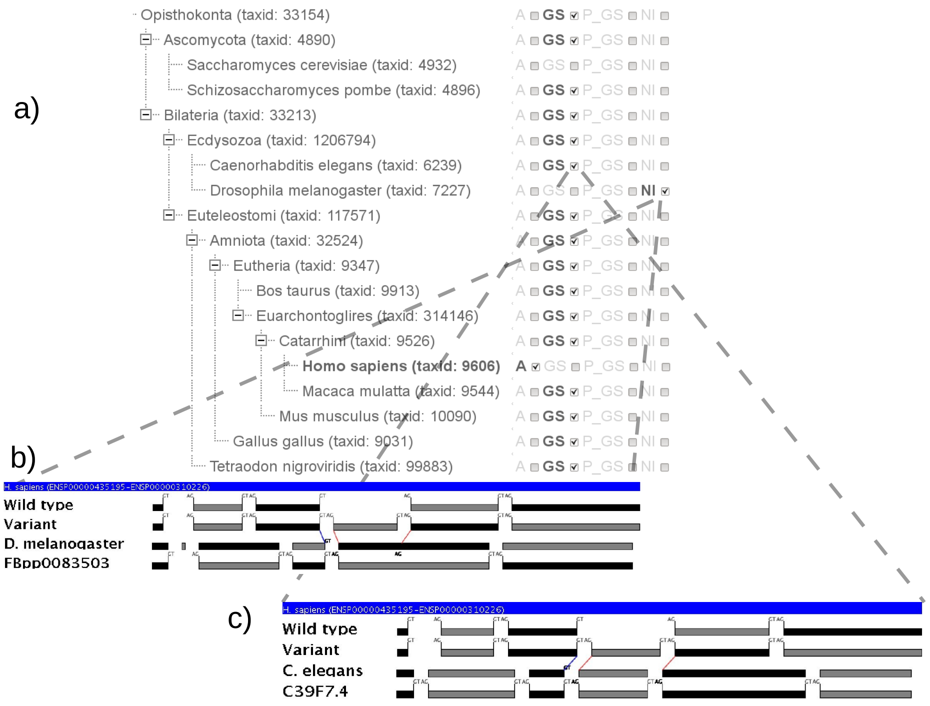

a)
Opisthokonta (taxid: 33154) A GS P_GS NI
Ascomycota (taxid: 4890) A GS P_GS NI
Saccharomyces cerevisiae (taxid: 4932) A GS P_GS NI
Schizosaccharomyces pombe (taxid: 4896) A GS P_GS NI
Bilateria (taxid: 33213) A GS P_GS NI
Ecdysozoa (taxid: 1206794) A GS P_GS NI
Caenorhabditis elegans (taxid: 6239) A GS P_GS NI
Drosophila melanogaster (taxid: 7227) A GS P_GS NI
Euteleostomi (taxid: 117571) A GS P_GS NI
Amniota (taxid: 32524) A GS P_GS NI
Eutheria (taxid: 9347) A GS P_GS NI
Bos taurus (taxid: 9913) A GS P_GS NI
Euarchontoglires (taxid: 314146) A GS P_GS NI
Catarrhini (taxid: 9526) A GS P_GS NI
Homo sapiens (taxid: 9606) A GS P_GS NI
Macaca mulatta (taxid: 9544) A GS P_GS NI
Mus musculus (taxid: 10090) A GS P_GS NI
Gallus gallus (taxid: 9031) A GS P_GS NI
Tetraodon nigroviridis (taxid: 99883) A GS P_GS NI
b)
H. sapiens (ENSP00000435195-ENSP00000310226)
Wild type
Variant
D. melanogaster
FBpp0083503
c)
H. sapiens (ENSP00000435195-ENSP00000310226)
Wild type
Variant
C. elegans
C39F7.4

Figure 6.9. Classification of a human Alternative Splicing Event (AES). (a) For the analyzed set of isoforms and species the classification of the event is shown along the phylogenetic tree. Classifications for inner nodes of the tree are inferred according to the most likely explanation (see main text). (b, c) For both worm and fly the evidence for the classification is also shown in form of the respective alignment of the ASE as extracted from the respective ISARs. For the particular examples in (b) and (c) the alignment of the start and the end positions of the AES is shown via the red and blue lines between the source isoforms (wildtype/ variant) to the ortholog genes in fly and worm.

for Ascomycota as well as for Bilateria and consequently also for Opisthokonta.

We identify exon skipping events between all genes in the 16,066 computed ISARs for the 10 selected eukaryotic species and classify their conservation level. In total, we consider 25,788 ASEs. Most events are defined for human, mouse, and rhesus monkey with 14,205, 5,346, and 4,266 events, respectively. We estimate for each species and inner taxonomy node the maximum classification of the transferred splicing event (TASE). In general, we observe that most events seem plausible for the higher mammals such as human and mouse (see Figure 6.10 for the classification of the human events). This is not very surprising as most events are derived from these two closely related species. For example, 2,344 of the 14,205 (17 %) events defined for human are also classified as annotated for mouse, i.e. there exists clear evidence of an orthologous ASE. Additionally, 8,654 (61 %) of the human events are supported by the mouse gene structure. That is, the splicing event could be mapped to known donor and acceptor sites. Only a small number of 102 and 184 of these events are

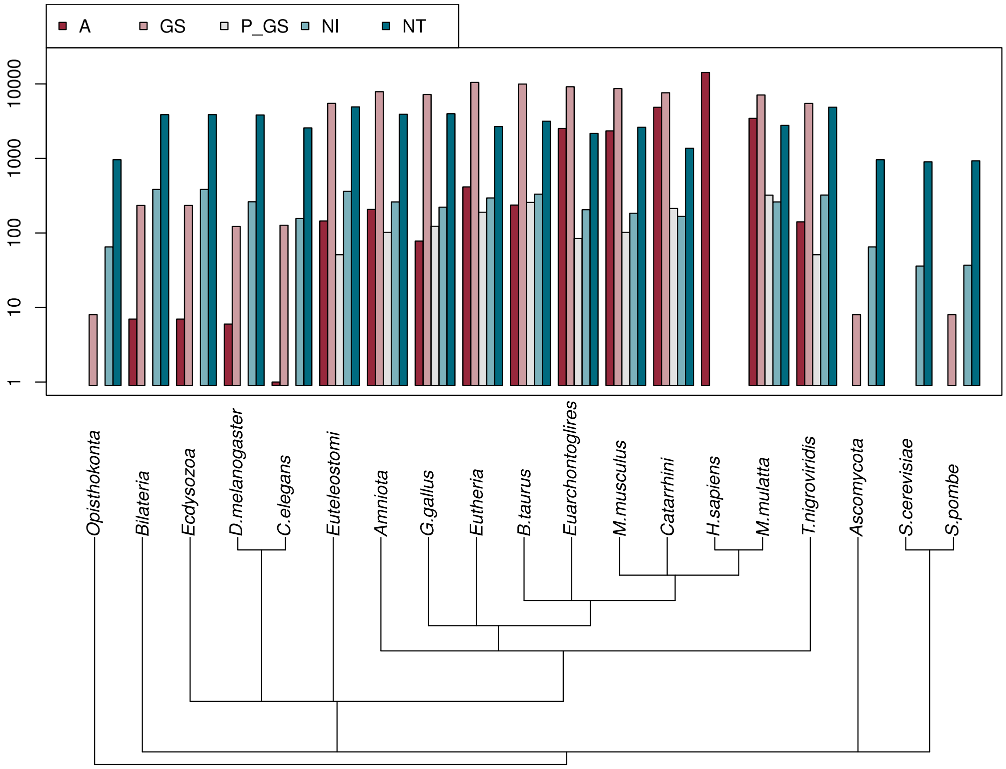
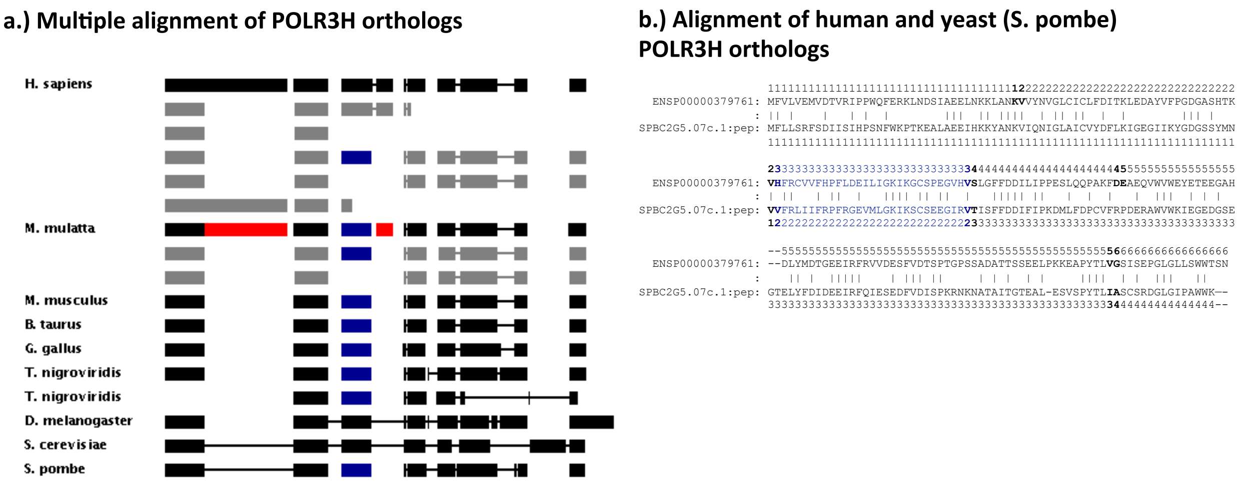


Figure 6.10. Classification of human alternative splicing events along the phylogenetic tree. We classified 25,789 exon skipping events derived from the alternative isoforms in the 16,066 computed ISARs containing 247,960 isoforms for 10 species. Here, we show the classification of the 14,205 exon skipping events annotated for human. These events are transferred to the genes/species contained in ISAR and classified as: (**A**) annotated, (**GS**) gene structure supported, (**NI**) supported by novel intron, (**P_GS**) supported by predicted gene structure, and (**NT**) the gene for which an event is defined is conserved (i.e. for the respective species, an orthologous gene is in the same OGG), but the event could not be transferred. For each specie we show the highest classification class for all its genes and for the inner-node we perform a maximum parsimony classification (see main text). Most events are annotated, or possible for the considered mammals. Surprisingly, also several instances of gene structure supported events in fission yeast (S. pombe) could be identified.

classified as P_GS and NI, respectively. Furthermore, 2,619 (18 %) of these events could not be transferred with the given ISARs. Surprisingly, we could also identify conserved events between phylogenetically distant species. In total, we classified 12 events as gene structure conserved for fission yeast (with 10 events conserved from human to yeast). In Figure 6.11 we show such an example for orthologs of the human DNA-directed RNA polymerase III subunit RPC8 (POLR3H) (another example is shown in the previously discussed ISAR for RAB1 in Figure 6.8). In human and rhesus monkey, an exon skipping is annotated (the blue exon is skipped in one isoform). The acceptor and donor sides — the boundaries of this event — are well aligned to other species (except for fly and worm) and thereby plausible for



```
                   111111111111111111111111111111111111122222222222222222222222222222222
  ENSP00000379761: MFVLVEMVDTVRIPPWQFERKLNDSIAEELNKKLANKVVYNVGLCICLFDITKLEDAYVFPGDGASHTK
                 : || |    |   | |  |        |||  || ||||  | || ||  |  |        ||| |
SPBC2G5.07c.1:pep: MFLLSRFSDIISIHPSNFWKPTKEALAEEIHKKYANKVIQNIGLAICVYDFLKIGEGIIKYGDGSSYMN
                   111111111111111111111111111111111111111111111111111111111111111111111

                   233333333333333333333333333334444444444444444444445555555555555555
  ENSP00000379761: VHFRCVVFHPFLDEILIGKIKGCSPEGVHVSLGFFDDILIPPESLQQPAKFDEAEQVWVWEYETEEGAH
                 : | ||   | ||  |   |||| || ||  |   ||||| ||   |  |  |   |  |||  | | |
SPBC2G5.07c.1:pep: VVFRLIIFRPFRGEVMLGKIKSCSEEGIRVTISFFDDIFIPKDMLFDPCVFRPDERAWVWKIEGEDGSE
                   122222222222222222222222222222333333333333333333333333333333333333333

                   --55555555555555555555555555555555555555555555555666666666666666666
  ENSP00000379761: --DLYMDTGEEIRFRVVDESFVDTSPTGPSSADATTSSEELPKKEAPYTLVGSISEPGLGLLSWWTSN
                 :    || |  |||||    | |||  ||     | | |  | |     ||||  | |  |||   ||
SPBC2G5.07c.1:pep: GTELYFDIDEEIRFQIESEDFVDISPKRNKNATAITGTEAL-ESVSPYTLIASCSRDGLGIPAWWK--
                   33333333333333333333333333333333333333333333333333344444444444444--
```

Figure 6.11. Possible conserved exon skipping in multiple species for orthologous genes of the human DNA-directed RNA polymerase III subunit RPC8 (POLR3H). (a) The blue exon is skipped in a human and rhesus monkey isoform. The boundaries of this exon skipping are well conserved across the phylogenetic tree (with exceptions for the fly and worm group). In (b) the implied amino-acid sequence alignment derived from the ISAR between the human wild type (the isoform with the highlighted exon) and the S. pombe isoform is shown. The rows on top and bottom of the isoforms indicates the exon number for each amino-acid in the respective gene.

many species.

Our analysis reveals that many currently not annotated events are conserved with respect to the gene structure. Thus, ISARs can be used for the systematic cross-species analysis of spliced isoforms, the prediction of new isoforms (e.g. via the transfer of gene structure conserved events), and the cross-species analysis of exon-intron structure changes.

6.4 Discussion

Given a set of genes with isoform annotations from several species, i.e. sequence variants of genes, paralogs and orthologs, our goal is to represent this set such that all the relations between parts of the sequences are exhibited and easy to access. Such a representation has many practical applications ranging from the study of the evolution of isoforms, and spliced events (as for example addressed in this paper) to the (cross-)species interpretation of sequencing data-sets. We assume that the genomic positions and the exon-intron structure of genes and its isoforms are known, which is typically the case for all isoforms stemming from sequenced genomes. Subsequently, the relations between alternative isoforms are known. Characteristic for a set of isoforms is that the sequences are very similar, in the case of alternative isoforms even identical, in large parts, but that other (sometimes also large) parts are simply missing in some or many of the sequences.
]

Thus, the ISA problem demands tailor-made solutions. We adapt partial order sets (posets) as well as the concept of Recursive Dynamic Programming (RDP) for the representation of isoform consistent alignments. We extend the poset data structure with unique query operations like the get_unmatched operation, which enables the constraint based extension of alignments. That is, this operation allows to align highly reliable regions first and then to successively extent the alignment considering the already introduced constraints. As a consequence, the ISAR graph structure can be consistently (i.e. consistent with the current ISAR and the isoform/gene structure) extended with new alignments, e.g. newly identified isoforms from next generation sequencing experiments can be inserted into an ISAR. Furthermore, regions, which are not conserved, can remain unaligned in the alignment representation. This is of particular interest for the problem of isoform alignment as alternative products often exhibit species and lineage-specific, or not yet annotated exons in related genes.

Thus, the advantages of ISAR are manifold; the loosely coupled alignment oracles can be easily exchanged enabling a flexible choice of alignment tools for the generation of alignment suggestions used for the ISAR construction. This includes, that any pre-computed alignment can be converted to an ISA representation and, thereby, isoform inconsistent regions are corrected in the given alignment. As ISAR is aware of the gene and isoform structure only alternative (i.e. not yet aligned) regions of isoforms need to be matched. Thus, alignment extension, i.e. the insertions of further isoforms and genes to an existing ISAR, can be easily performed. Also, typical post-processing steps, e.g. the filtering of suspicious aligned positions, are not necessary as only alignment regions with a certain quality are considered (e.g. region with a certain length and sequence identity). Finally, the alignment can be nicely represented by a partial ordering of these regions (which along their isoform sequence are of course totally ordered), where some sets of regions are mapped to each other.

6.4.1 Multiple Sequence Alignments

One obvious solution for the ISA problem would be to solely rely on the computation of an optimal Multiple Sequence Alignment (MSA) of the set of isoforms. Already, over 100 different MSA methods have been published (Kemena and Notredame, 2009). But (to the best of our knowledge) the available MSA methods fail for our goal of exhibiting and respecting the gene and isoform structures. Moreover, as often (but in particular here), the choice of an appropriate scoring function is not obvious. For example for the very large alternative regions (gaps) and the large identical regions in isoform alignments the scoring does not really help much. Computing optimal multiple alignments of many long sequences is not easy and computationally expensive. Thus, heuristics and approximations are often used for the multiple alignment problem. Another solution might be to rely on pairwise alignments, which can be computed efficiently, and then construct a consistent multiple alignment from these pairwise alignments. There are many approaches based on this idea: progressive alignments with guide trees, phylogeny reconstruction methods, profile based methods, and HMMs. Iterative profile based methods approximating multiple alignments try to remedy the consistency problem, but have problems with the profile definition and the appropriate scoring of profile alignments, which might not perfectly fit to the context of aligning a large set of isoforms. In practice, we observe examples where the mentioned problems lead to suboptimal solutions, which induce inconsistent and biologically misleading interpretations as for example shown in Figure 6.1. Furthermore, different reading-frames for genomic positions (frame-shifts) makes it for many cases impossible to represent the ISA problem as simple alignment problem of many amino-acid sequences (typically used as input for MSA programs). To avoid all these problems it is common practice that only one isoform per gene is selected (typically the longest isoform, or the set of isoforms that are most similar to each other) for currently available multiple sequence alignments tools (Villanueva-Cañas et al., 2013), i.e. the entire ISA problem is currently ignored.

Another possible solution for the ISA problem would be to perform a multiple DNA aligning of all (coding) regions. But this results in a huge information loss as the scoring is only based on nucleotides rather than amino-acids. Furthermore, the integration of predicted exons from orthologous and paralogous genes in the computed alignment cannot be automatically performed with available MSA programs (as we did with the spliced-alignment oracles). The predicted exons complete the gene and isoform annotation as they highlight not yet annotated isoform and regions which are still conserved (to some extend), but not anymore used by a species.

6.4.2 ISAR Mapped Regions

We construct a multiple alignment representation of all the sequences exhibiting all the mappings between the alternative isoform structures. With respect to the elements, it is clear that the basic elements are the letters/nucleotides which are totally ordered along the isoform sequences. Apart from that, no other <-relations are implied at the beginning. By introducing aligned matches additional <-relations are induced (inherited via the matching). The best representation (partitioning) of isoform sequences with mapped and not mapped regions is defined by the ultimate alignment. In principle, any base position can establish the start or end of a region, but of course dealing with nucleotide elements is inconvenient as their number is large, thus, typically elements can also be chains of letters (defined by the first and last letter in a region). The size of these regions is defined by the extent of the reliable matches between isoform sequences. Typically, isoform sequences are partitioned into a relatively small number of these regions (comparable in size and number with the exons of the gene), which reduce the computational effort and makes the resulting structure and solution/alignment much more comprehensible.

Another choice to be made concerns the representation of the transitive <-relation. In principle, one can try to represent all the pairs (x,y) for which $x < y$ holds such that all $<$ queries can be answered immediately in constant time (i.e. directly in one computational step). The other extreme is to represent as few edges as possible such that still all true $x < y$ can be derived (representation with the minimal number of edges). In the latter case, an edge (x, y) is represented in the ISAR if and only if $x < y$ and there is no z such that $x < z < y$.

Here, we choose the ISAR to be built from a poset with: (i) as large as possible elements, and (ii) as few edges for the <-relation as possible. This makes the representation as sparse (and we think as interpretable) as possible. This comes at the cost of computing transitive <-relations between some elements if necessary. On the other hand, having fewer edges can significantly reduce the effort for consistency checks (e.g. for the case that additional matches have to be introduced into the ISAR).

6.4.3 Selection of Regions

Given a scored list of aligned regions the ISAR algorithm builds an as large (maximal) as possible ISAR in a greedy fashion, i.e. the ISAR is extended by matched regions that are ordered by the region score until no further extension is possible. Note, that this procedure is heuristic and neither guarantees optimality nor a relative performance factor (approximation quality guarantee). It is of course possible that not all matches can consistently be satisfied in the ISAR at the same time. In this case, maximal consistent posets will be produced. Of course, 'maximal' needs to be defined and there are several options for the objective function, e.g. the number of matches, the sum of match weights (alignment scores), and the accumulated p-values of all represented matches. Here we adopt a straightforward approach by first initializing the ISARs with highly reliable alignments and then successively extend the ISAR with further alignments ranked by their sequence identity and length. We think, the greedy approach is sufficient for problem instances arising in practical problem instances. Thus, we choose this simple and highly efficient strategy. Another reason for the greedy strategy is that it can easily be extended to produce *all* maximal consistent ISARs. Moreover, remaining alignments not consistent with the ISAR can be obtained for special treatment and subsequent analysis.

6.5 Conclusion

ISAR is a new poset based approach for the isoform structure alignment problem. The representation on its own is no multiple sequence alignment method, but a general framework to integrate alignments from state-of-the-art tools in a gene and isoform consistent manner with the ability to recursively divide the alignment task into sub-problems. The ISAR algorithm constructs a data structure representing the gene structures and their mappings for a (possibly large) set of input (alternative) isoform sequences. The data structure allows to output a multiple alignment of the set of isoform sequences. As the alignment can be large and complicated, a graphical visualization of the alignment based on the exons and introns can be more appropriate. Thus, the core of the ISAR algorithm and also its main result is a graph representation (ISAR) of (partial) isoform alignments, which are all consistent with the annotated gene structures.

The ISAR algorithm can be quite flexibly customized and extended with new/additional oracles and optimization strategies. Moreover, a detailed analysis of splicing events between smaller subsets of sequences can be conducted. ISAR is fast enough to allow for genome-wide analyses, e.g. for the investigation of all human genes together with its orthologs and paralogs across large taxonomies in order to statistically analyze splicing patterns. Thus, ISARs of many genes facilitate the analysis of conserved spliced patterns, the transfer of isoforms across species, and the study of gene structure evolution.

Chapter 7

Conclusion and Outlook

In this thesis, the cross-species transfer and the context-specific analysis of networks as well as the identification of conserved alternative isoforms were addressed. In the following section the main findings will be summarized.

Protein-Protein Interaction Transfer: Global binary protein-protein interaction networks are available for some eukaryotic model organisms. Such networks have been successfully used for the prediction of protein functions and the interpretation of experimental data. But still the interactome for most species is sparse (especially for non-model organisms).

With **COIN**, we enabled the cross-species protein-protein interaction network transfer (see Chapter 3 and Pesch and Zimmer (2013)). COIN combines diverse novel features from orthologous genes and the network structure in order to score the likelihood of a conserved interaction in a given target species. This approach outperforms competing methods for the transfer of interactions to species where no or only little experimental data is available. The sets of transferred interactions for 83 eukaryotic species can be interactively filtered and downloaded via a web-service. Thereby, reliable protein-protein interaction networks are made available for many species.

Cross-Species and Cross-Context Regulatory Networks: It remains unclear to which extent regulatory networks (transcription factor-target networks) are conserved between species. Some well-studied regulatory sub-networks suggest remarkable cross-species similarities of regulatory mechanisms. With the **ConReg** system, we present a comprehensive collection of global regulatory networks for eukaryotic model organisms and a system to query conserved regulatory (sub)-networks (see Chapter 4 and Pesch et al. (2012); Pesch and Zimmer (2014)). For ConReg networks were derived, integrated, and constructed from the scientific literature, curated databases and computationally binding site predictions. We have successfully applied the system for the identification of many conserved regulations in fly and vertebrates.

Regulatory interactions are strongly context-specific. Therefore, besides the identification of conserved interactions in a target species (as done with COIN and ConReg) the context of an interaction should also be considered. Projects like ENCODE, mouseENCODE, and modENCODE represent a rich resource of context-specific regulatory data for hundreds of different cell-lines.

We systematically derived regulatory networks from this data, mapped the network entities to a common node set, and developed **CroCo**, a novel context-specific regulatory network framework (see Chapter 5 and Pesch and Zimmer (2014)). This framework allows performing various cross-context and cross-species network comparisons via the integration of orthology information and feature-rich network analysis tools. Thereby, context-specific networks can be transferred between species (similar to COIN). Flexible browsing and aggregation of networks of interest is enabled via the organization of networks into ontologies according to their meta-data. Thus, CroCo adds a unifying network-oriented view on the data from the ENCODE projects and provides several ways to compare networks in a cross-species and cross-context manner.

Conservation of Alternatively Spliced Variants: Network models typically neglect alternative splicing, as experimental data often does not allow the discrimination between different spliced isoforms. But alternative splicing can have (drastic) effects on the protein structure, the protein function and subsequently on the networks. The analysis of alternative splicing induced effects on the cross-species network transfer requires (besides others) the identification of conserved, lineage- and species-specific isoforms. Correct Multiple Sequence Alignments (MSA) allow the identification of such conserved spliced isoforms, but state-of-the-art MSA methods produce inconsistent alignments as they ignore the interrelationships between different alternative isoforms. With ISAR we introduced an isoform-consistent multiple sequence alignment approach for the alignment of isoforms from orthologous and paralogous genes (see Chapter 6). We employed ISAR for the representation of hundreds of thousands of isoforms from ten species ranging from human to yeast. Using these ISARs, we were able to identify conserved spliced events between phylogenetically distant species.

ISAR allows to identify similar spliced isoforms in different genes and species, and also to perform a cross-species isoform transfer. Thereby, the evolution and origin of (alternatively) spliced isoforms can be studied.

Perspectives for Future Research

The approaches, data repositories, and software applications devolved in this thesis offer multiple avenues for further research:

Comprehensive Cross-Species Networks Comparison: GEO, SRA, and ArrayExpress provide tens of thousands of further transcription factor-binding site experiments for diverse species and experimental settings. The integration of this data into the presented regulatory network repositories will provide a more complete view on the experimentally identified binding sites. Furthermore, the user-defined specification and extraction of networks via custom procedures could be integrated in order to account for different parameters for the network definition. The on-demand comparison and overlap of user-defined networks with networks from the same species and transferred networks could be supported in order to enable the comprehensive, cross-species, differential, and interactive analysis of context-specific networks.

Isoform Structure Representation and Protein Interactions: Protein interfaces derived from structurally resolved interacting proteins — e.g. from the PDBePISA database (Krissinel and Henrick, 2007) — provide a set of protein interactions with precise interaction region positions on the isoforms. Such a dataset can be used together with ISAR in order to study the conservation of protein interactions and the alternative splicing induced effects on interaction regions, simultaneously.

Furthermore, gene and protein expression data can be mapped to the genes and isoforms represented in the ISARs in order to perform cross-species expression analysis and to check the cross-species transferred isoforms in the experimental data.

Outlook

The definition of networks from (high-throughput) data can be used as starting point for the understanding of regulatory mechanisms. But the construction of networks requires some simplified assumptions and subsequently may not yet capture the entire complexity of regulatory mechanisms. This includes that the current network models are mostly binary and gene centered (e.g. only one single representative isoform for each gene is considered). An integrated (regulatory) network model for a species could include: genes, transcripts, proteins, protein complexes, protein modifications, histone modifications, and microRNAs with precise (context and conservation) annotations of the entities and interactions. Furthermore, the user-defined extensions and modifications should be supported in order to account for future data sets. Such an integrative network will be more realistic than current network models and be straightforward to transfer between species as all information is at hand. Sophisticated network analysis tools will be required to handle such a potentially very huge and complex network. The analysis tools and networks presented within this work could be used as building blocks to construct such an integrated network model.

Conclusion

Systems biology seeks to achieve a comprehensive understanding of interactions in biological systems with the ultimate goal of understanding how these interactions are responsible for the observed changes in a system. An overwhelming amount of genomic and proteomic data is generated in various huge consortia projects and presented to the scientific community, but still for many contexts, systems and species either little or no experimental data is available. Furthermore, it appears that *everything* in a cell is context-dependent: chromatin conformation, open chromatin regions, RNA splicing, and thus subsequently, also the gene expression and the protein interaction, and regulatory networks. Networks are an abstract representation of experimental data. They have been successfully used to study regulatory dynamics, to predict protein function, and to interpret experimental data. The transfer of networks using orthologous and paralogous genes allows the prediction of networks even for species without experiment data. However, networks (especially derived from high-throughput experiments) often have hundreds of thousands of interactions and are difficult to be interpreted and compared. With ConReg, CroCo, and COIN comprehensive network resources and software tools for the differential and context-specific network analysis and the cross-species network transfer are presented. In addition, the ISAR approach allows to take a step towards the cross-species analysis of the effects of alternative splicing on networks, by providing evolutionary relationships between isoforms of orthologous and paralogous genes. Thus, the approaches presented in this work can be used as a starting point for the understanding of (species-specific and conserved) regulatory and signaling mechanisms in many biological systems.

Bibliography

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci*, 118:4947–4957.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walte, P. (2008). *Molecular Biology of the Cell*. Garland Science, 5th edition. ISBN: 9780815341055.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461.

Altenhoff, A. M. and Dessimoz, C. (2009). Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Comput Biol*, 5(1):e1000262.

Altenhoff, A. M., Schneider, A., Gonnet, G. H., and Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res*, 39(Database issue):D289–D294.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29.

Ast, G. (2004). How did alternative splicing evolve? *Nat Rev Genet*, 5(10):773–782.

Awan, A. R., Manfredo, A., and Pleiss, J. A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc Natl Acad Sci USA*, 110(31):12762–12767.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242–245.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593.

Baumbach, J. (2010). On the power and limits of evolutionary conservation–unraveling bacterial gene regulatory networks. *Nucleic Acids Res*, 38(22):7877–7884.

Baumbach, J., Tauch, A., and Rahmann, S. (2009). Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Brief Bioinform*, 10(1):75–83.

Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell*, 150(2):413–425.

Berezikov, E., Guryev, V., and Cuppen, E. (2005). CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res*, 33(Web Server issue):W447–W450.

Berg, J., Lässig, M., and Wagner, A. (2004). Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol*, 4:51.

Berggård, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16):2833–2842.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, 28(10):1045–1048.

Bhardwaj, N. and Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738.

Birzele, F., Csaba, G., and Zimmer, R. (2008). Alternative splicing and protein structure evolution. *Nucleic Acids Res*, 36(2):550–558.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J Mol Biol*, 283(4):707–725.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14(3):292–299.

Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L. W., Janette, J., Jiang, L., Kasper, D., Kawli, T., Kheradpour, P., Kundaje, A., Li, J. J., Ma, L., Niu, W., Rehm, E. J., Rozowsky, J., Slattery, M., Spokony, R., Terrell, R., Vafeados, D., Wang, D., Weisdepp, P., Wu, Y.-C., Xie, D., Yan, K.-K., Feingold, E. A., Good, P. J., Pazin, M. J., Huang, H., Bickel, P. J., Brenner, S. E., Reinke, V., Waterston, R. H., Gerstein, M., White, K. P., Kellis, M., and Snyder, M. (2014). Comparative analysis of regulatory information and circuits across distant species. *Nature*, 512(7515):453–456.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brooks, C. L. and Gu, W. (2003). Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr Opin Cell Biol*, 15(2):164–171.

Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082.

Buljan, M., Chalancon, G., Dunker, A. K., Bateman, A., Balaji, S., Fuxreiter, M., and Babu, M. M. (2013). Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr Opin Struct Biol*, 23(3):443–450.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., and Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*, 46(6):871–883.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 161–168, New York, NY, USA. ACM.

Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K. P., Waterston, R. H., and modENCODE Consortium (2009). Unlocking the secrets of the genome. *Nature*, 459(7249):927–930.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res*, 41(D1):D816–D823.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–D574.

Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N., and Ricard-Blum, S. (2011). MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res*, 39(Database issue):D235–D240.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and micrornas. *Nat Rev Genet*, 8(2):93–103.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. (2007). Integration of biological networks and gene expression data using cytoscape. *Nat Protoc*, 2(10):2366–2382.

Colantoni, A., Bianchi, V., Gherardini, P. F., Tomba, G. S., Ausiello, G., Helmer-Citterich, M., and Ferrè, F. (2013). Alternative splicing tends to avoid partial removals of protein-protein interaction sites. *BMC Genomics*, 14:379.

Contrino, S., Smith, R. N., Butano, D., Carr, A., Hu, F., Lyne, R., Rutherford, K., Kalderimis, A., Sullivan, J., Carbon, S., Kephart, E. T., Lloyd, P., Stinson, E. O., Washington, N. L., Perry, M. D., Ruzanov, P., Zha, Z., Lewis, S. E., Stein, L. D., and Micklem, G. (2012). modMine: flexible access to modENCODE data. *Nucleic Acids Res*, 40(Database issue):D1082–D1088.

Côté, R. G., Jones, P., Apweiler, R., and Hermjakob, H. (2006). The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97.

Csaba, G. (2008). syngrep - Fast synonym-based named entity recognition. Master's thesis, LMU-Munich.

Davidson, E. H. (2006). *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, 1st edition. ISBN: 9780120885633.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

De Bodt, S., Proost, S., Vandepoele, K., Rouze, P., and Van de Peer, Y. (2009). Predicting protein-protein interactions in arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, 10:288.

Duan, G. and Walther, D. (2015). The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput Biol*, 11(2):e1004049.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797.

Ehlert, A., Weltmeier, F., Wang, X., Mayer, C. S., Smeekens, S., Vicente-Carbajosa, J., and Dröge-Laser, W. (2006). Two-hybrid protein-protein interaction analysis in arabidopsis protoplasts: establishment of a heterodimerization map of group C and group S bZIP transcription factors. *Plant J*, 46(5):890–900.

Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J. A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P. M., Wrana, J. L., and Blencowe, B. J. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell*, 46(6):884–892.

ENCODE Project Consortium (2012a). Data standards. Retrieved October 27, 2015, from https://genome.ucsc.edu/ENCODE/dataStandards.html.

ENCODE Project Consortium (2012b). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, 3:89.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet*, 23(22):5866–5878.

Faro, A., Giordano, D., and Spampinato, C. (2012). Combining literature text mining with microarray data: advances for system biology modeling. *Brief Bioinform*, 13(1):61–82.

Fields, S. and Johnston, M. (2005). Whither model organism research? *Science*, 307(5717):1885–1886.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J. P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Res*, 42(Database issue):D749–D755.

Fong, J. H., Shoemaker, B. A., Garbuzynskiy, S. O., Lobanov, M. Y., Galzitskaya, O. V., and Panchenko, A. R. (2009). Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. *PLoS Comput Biol*, 5(3):e1000316.

Friedel, C. C. and Zimmer, R. (2009). Identifying the topology of protein complexes from affinity purification assays. *Bioinformatics*, 25(16):2140–2146.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(Database issue):D876–D882.

Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx–relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Furey, T. S. (2012). Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. *Nat Rev Genet*, 13(12):840–852.

Gabaldón, T., Dessimoz, C., Huxley-Jones, J., Vilella, A. J., Sonnhammer, E. L., and Lewis, S. (2009). Joining forces in the quest for orthologs. *Genome Biol*, 10(9):403.

Gallo, S. M., Gerrard, D. T., Miner, D., Simich, M., Des Soye, B., Bergman, C. M., and Halfon, M. S. (2011). REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in drosophila. *Nucleic Acids Res*, 39(Database issue):D118–D123.

Gallone, G., Simpson, T. I., Armstrong, J. D., and Jarman, A. P. (2011). Bio::Homology::InterologWalk–a perl module to build putative protein-protein interaction networks through interolog mapping. *BMC Bioinformatics*, 12:289.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., and Oliva, B. (2012). BIPS: BIANA interolog prediction server. a tool for protein-protein interaction inference. *Nucleic Acids Res*, 40(Web Server issue):W147–W151.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

Gene Ontology Consortium (2015). Gene ontology consortium: going forward. *Nucleic Acids Res*, 43(Database issue):D1049–D1056.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., and Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100.

Giuliano, C., Lavelli, A., and Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *In Proc. EACL 2006*, pages 401–408.

Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T., and Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics*, 24(15):1743–1744.

Gotoh, O. (1996). Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol*, 264(4):823–838.

Gotoh, O., Morita, M., and Nelson, D. R. (2014). Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics*, 15:189.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Van Loo, P., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I. J., Robertson, G., Wadelius, C., De Bleser, P., Vlieghe, D., Halfon, M. S., Wasserman, W., Hardison, R., Bergman, C. M., Jones, S. J. M., and , O. R. A. C. (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*, 36(Database issue):D107–D113.

Grützmann, K., Szafranski, K., Pohl, M., Voigt, K., Petzold, A., and Schuster, S. (2014). Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. *DNA Res*, 21(1):27–39.

Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.-W., and Stümpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):D436–D441.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22(9):1760–1774.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models — A review. *Biosystems*, 96(1):86–103.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4):283–289.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13.

Hulsen, T., Huynen, M. A., de Vlieg, J., and Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*, 7(4):R31.

Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., de Atauri, P., Aitchison, J. D., Hood, L., Siegel, A. F., and Bolouri, H. (2005). A data integration methodology for systems biology. *Proc Natl Acad Sci USA*, 102(48):17296–17301.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Mol Syst Biol*, 8:565.

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., and Barkai, N. (2005). Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, 309(5736):938–940.

Iwata, H. and Gotoh, O. (2012). Benchmarking spliced alignment programs including spaln2, an extended version of spaln that incorporates additional species-specific features. *Nucleic Acids Res*, 40(20):e161.

Janky, R., Helden, J. v., and Babu, M. M. (2009). Investigating transcriptional regulation: from analysis of complex networks to discovery of cis-regulatory elements. *Methods*, 48(3):277–286.

Jensen, O. N. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol*, 8(1):33–41.

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biol*, 2(8):interactions1002.1–1002.3.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502.

Kaboord, B. and Perr, M. (2008). Isolation of proteins and protein complexes by immunoprecipitation. *Methods Mol Biol*, 424:349–364.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114.

Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*, 9(10):770–780.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. *Gene*, 514(1):1–30.

Kemena, C. and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*, 11(5):345–355.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40(Database issue):D841–D846.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database — 2009 update. *Nucleic Acids Res*, 37(Database issue):D767–D772.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 drosophila genomes. *Genome Res*, 17(12):1919–1931.

Khoury, G. A., Baliban, R. C., and Floudas, C. A. (2011). Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep*, 1:90.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, 35(1):125–131.

Kim, J., Bhinge, A. A., Morgan, X. C., and Iyer, V. R. (2005). Mapping dna-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat Methods*, 2(1):47–53.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Koch, I., Junker, B. H., and Heiner, M. (2005). Application of petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, 21(7):1219–1226.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Megy, K., Pilicheva, E., Rustici, G., Tikhonov, A., Parkinson, H., Petryszak, R., Sarkans, U., and Brazma, A. (2015). Arrayexpress update–simplifying data submissions. *Nucleic Acids Res*, 43(Database issue):D1113–D1116.

Koonin, E. V. (2001). An apology for orthologs - or brave new memes. *Genome Biol*, 2(4):comment1005.1–1005.2.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39:309–338.

Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, 372(3):774–797.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res*, 19(9):1639–1645.

Kummerfeld, S. K. and Teichmann, S. A. (2006). DBD: a transcription factor prediction database. *Nucleic Acids Res*, 34(Database issue):D74–D81.

Küffner, R., Fundel, K., and Zimmer, R. (2005). Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics*, 21(Suppl 2):ii259–ii267.

Küffner, R., Zimmer, R., and Lengauer, T. (2000). Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16(9):825–836.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). Chip-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, 22(9):1813–1831.

Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464.

Lee, M.-H., Hook, B., Pan, G., Kershner, A. M., Merritt, C., Seydoux, G., Thomson, J. A., Wickens, M., and Kimble, J. (2007). Conserved regulation of MAP kinase expression by PUF RNA-binding proteins. *PLoS Genet*, 3(12):e233.

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res*, 39(Database issue):D19–D21.

Letunic, I. and Bork, P. (2011). Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*, 39(Web Server issue):W475–W478.

Lewis, A. C. F., Jones, N. S., Porter, M. A., and Deane, C. M. (2012). What evidence is there for the homology of protein-protein interactions? *PLoS Comput Biol*, 8(9):e1002645.

Loots, G. G. and Ovcharenko, I. (2004). rvista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):W217–W221.

Luo, Q., Pagel, P., Vilne, B., and Frishman, D. (2011). DIMA 3.0: Domain interaction map. *Nucleic Acids Res*, 39(Database issue):D724–D729.

Lynn, D. J., Winsor, G. L., Chan, C., Richard, N., Laird, M. R., Barsky, A., Gardy, J. L., Roche, F. M., Chan, T. H. W., Shah, N., Lo, R., Naseer, M., Que, J., Yau, M., Acab, M., Tulpan, D., Whiteside, M. D., Chikatamarla, A., Mah, B., Munzner, T., Hokamp, K., Hancock, R. E. W., and Brinkman, F. S. L. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol Syst Biol*, 4:218.

Löytynoja, A., Vilella, A. J., and Goldman, N. (2012). Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, 28(13):1684–1691.

Manuel, M., Pratt, T., Liu, M., Jeffery, G., and Price, D. J. (2008). Overexpression of Pax6 results in microphthalmia, retinal dysplasia and defective retinal ganglion cell axon guidance. *BMC Dev Biol*, 8:59.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 42(Database issue):D142–D147.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12):2120–2126.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110.

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R. A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B. M., Fujita, P. A., Dreszer, T. R., Diekhans, M., Cline, M. S., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2013).

The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*, 41(Database issue):D64–D69.

Michaut, M., Kerrien, S., Montecchi-Palazzi, L., Chauvat, F., Cassier-Chauvat, C., Aude, J.-C., Legrain, P., and Hermjakob, H. (2008). InteroPORC: automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24(14):1625–1631.

Mika, S. and Rost, B. (2006). Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol*, 2(7):e79.

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S. L., Nekrutenko, A., Giardine, B., Harris, R. S., Tyekucheva, S., Diekhans, M., Pringle, T. H., Murphy, W. J., Lesk, A., Weinstock, G. M., Lindblad-Toh, K., Gibbs, R. A., Lander, E. S., Siepel, A., Haussler, D., and Kent, W. J. (2007). 28-way vertebrate alignment and conservation track in the ucsc genome browser. *Genome Res*, 17(12):1797–1808.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Education, 1st edition. ISBN: 9780070428072.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet*, 14(10):719–732.

Mouse ENCODE Consortium (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol*, 13(8):418.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012a). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R. S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M. J., Akey, J. M., Bender, M. A., Groudine, M., Kaul, R., and Stamatoyannopoulos, J. A. (2012b). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217.

Offman, M. N., Nurtdinov, R. N., Gelfand, M. S., and Frishman, D. (2004). No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions. *BMC Bioinformatics*, 5:41.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834.

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L. C., Dahmane, N., and Davuluri, R. V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res*, 21(8):1260–1272.

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006). AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol*, 140(3):818–829.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415.

Pao-Yang Chen, Charlotte M. Deane, G. R. (2008). Predicting and validating protein interactions using network structure. *PLoS Comput Biol*, 4(7):e1000118.

Pavlopoulos, G. A., Wegener, A.-L., and Schneider, R. (2008). A survey of visualization tools for biological network analysis. *BioData Min*, 1:12.

Pesch, R., Böck, M., and Zimmer, R. (2012). ConReg: Analysis and Visualization of Conserved Regulatory Networks in Eukaryotes. In Böcker, S., Hufsky, F., Scheubert, K., Schleicher, J., and Schuster, S., editors, *German Conference on Bioinformatics 2012*, volume 26 of *OpenAccess Series in Informatics (OASIcs)*, pages 69–81, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Pesch, R., Lysenko, A., Hindle, M., Hassani-Pak, K., Thiele, R., Rawlings, C., Köhler, J., and Taubert, J. (2008). Graph-based sequence annotation using a data integration approach. *J Integr Bioinform*, 5(2):94.

Pesch, R. and Zimmer, R. (2013). Complementing the eukaryotic protein interactome. *PLoS One*, 8(6):e66635.

Pesch, R. and Zimmer, R. (2014). To be, or not to be: konservierte eukaryotische Regulationsnetzwerke? *BIOspektrum*, 20(5):514–516.

Petsko, G. A. (2001). Homologuephobia. *Genome Biol*, 2:comment1002.1.

Phizicky, E. M. and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94–123.

Pollack, J. R. and Iyer, V. R. (2002). Characterizing the physical genome. *Nat Genet*, 32:515–521.

Prabakaran, S., Lippens, G., Steen, H., and Gunawardena, J. (2012). Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip Rev Syst Biol Med*, 4(6):565–583.

Rao, V. S., Srinivas, K., Sujini, G. N., and Kumar, G. N. S. (2014). Protein-protein interaction detection: methods and analysis. *Int J Proteomics*, 2014:147648.

Reddy, V. N., Mavrovouniotis, M. L., and Liebman, M. N. (1993). Petri net representations in metabolic pathways. *Proc Int Conf Intell Syst Mol Biol*, 1:328–336.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052.

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., and Lee, C. (2004). Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res*, 3(1):76–83.

Resnik, P. (1999). Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–959.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Robasky, K. and Bulyk, M. L. (2011). UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, 39(Database issue):D124–D128.

Romero, I. G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet*, 13(7):505–516.

Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A. K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci USA*, 103(22):8390–8395.

Rottger, R., Ruckert, U., Taubert, J., and Baumbach, J. (2012). How little do we actually know? – on the size of gene regulatory networks. *IEEE/ACM Trans Comput Biol Bioinform*, 9(5):1293–1300.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 38(Database issue):D497–D501.

Sambourg, L. and Thierry-Mieg, N. (2010). New insights into protein-protein interaction data lead to increased estimates of the s. cerevisiae interactome size. *BMC Bioinformatics*, 11:605.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 37(Database issue):D5–15.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the pathway interaction database. *Nucleic Acids Res*, 37(Suppl 1):D674–D679.

Schneider, A., Dessimoz, C., and Gonnet, G. H. (2007). OMA browser–exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 23(16):2180–2182.

Schroder, K., Irvine, K. M., Taylor, M. S., Bokil, N. J., Le Cao, K.-A., Masterman, K.-A., Labzin, L. I., Semple, C. A., Kapetanovic, R., Fairbairn, L., Akalin, A., Faulkner, G. J., Baillie, J. K., Gongora, M., Daub, C. O., Kawaji, H., McLachlan, G. J., Goldman, N., Grimmond, S. M., Carninci, P., Suzuki, H., Hayashizaki, Y., Lenhard, B., Hume, D. A., and Sweet, M. J. (2012). Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc Natl Acad Sci USA*, 109(16):E944–E953.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107.

Seo, J. and Lee, K.-J. (2004). Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J Biochem Mol Biol*, 37(1):35–44.

Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433.

Sharma, R., Evans, P. A., and Bhavsar, V. C. (2011). Regulatory link mapping between organisms. *BMC Syst Biol*, 5(Suppl 1):S4.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–68.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., Bertone, P., and , R. G. A. S. P. C. (2013). Assessment of transcript reconstruction methods for rna-seq. *Nat Methods*, 10(12):1177–1184.

Stormo, G. D. (2013). Modeling the specificity of protein-dna interactions. *Quant Biol*, 1(2):115–130.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci USA*, 105(19):6959–6964.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–D568.

Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S., and Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res*, 21(7):1139–1149.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., Alenquer, M., Freitas, A. T., Oliveira, A. L., and Sá-Correia, I. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res*, 34(Database issue):D446–D451.

The UniProt Consortium (2011). Ongoing and future developments at the universal protein resource. *Nucleic Acids Res*, 39(Suppl 1):D214–D219.

Thiele, R., Zimmer, R., and Thomas, L. (1999). Protein threading by recursive dynamic programming. *J Mol Biol.*, 290(3):757–779.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D., and van Helden, J. (2011). RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res*, 39(Web Server issue):W86–W91.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). 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(22):4673–4680.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6:e1000837.

Tong, S., Koller, D., and Kaelbling, P. (2001). Support Vector Machine Active Learning with Applications to Text Classification. In *Journal of Machine Learning Research*, pages 999–1006.

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism. *Nat Protoc*, 6(9):1341–1354.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford)*, 2010:baq026.

Uetz, P. and Pankratz, M. J. (2004). Protein interaction maps on the fly. *Nat Biotechnol*, 22(1):43–44.

van Dam, T. J. P. and Snel, B. (2008). Protein complex evolution does not involve extensive network rewiring. *PLoS Comput Biol*, 4(7):e1000132.

Van Landeghem, S., De Bodt, S., Drebert, Z. J., Inzé, D., and Van de Peer, Y. (2013). The potential of text mining in data integration and network biology for plant research: a case study on arabidopsis. *Plant Cell*, 25(3):794–807.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2):327–335.

Villanueva-Cañas, J. L., Laurie, S., and Albà, M. M. (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol Evol*, 5(2):457–467.

Villar, D., Flicek, P., and Odom, D. T. (2014). Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet*, 15(4):221–233.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22(9):1798–1812.

Wang, P., Yan, B., Guo, J.-T., Hicks, C., and Xu, Y. (2005). Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc Natl Acad Sci USA*, 102(52):18920–18925.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. (2010). The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, 38(Web Server issue):W214–W220.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C., and Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Wawersik, S. and Maas, R. L. (2000). Vertebrate eye development as modeled in Drosophila. *Hum Mol Genet*, 9(6):917–925.

Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H. G., Ukkonen, E., Hughes, T. R., Bulyk, M. L., and Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J*, 29(13):2147–2160.

Westermann, F., Muth, D., Benner, A., Bauer, T., Henrich, K.-O., Oberthuer, A., Brors, B., Beissbarth, T., Vandesompele, J., Pattyn, F., Hero, B., König, R., Fischer, M., and Schwab, M. (2008). Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol*, 9(10):R150.

Wiles, A. M., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B., and Bishop, A. J. R. (2010). Building and analyzing protein interactome networks by cross-species comparisons. *BMC Syst Biol*, 4:36.

Woodsmith, J. and Stelzl, U. (2014). Studying post-translational modifications with protein interaction networks. *Curr Opin Struct Biol*, 24:34–44.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289–291.

Xin, F. and Radivojac, P. (2012). Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics*, 28(22):2905–2913.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*, 13(5):329–342.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci USA*, 101(16):5934–5939.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome Res*, 14(6):1107–1118.

Zambelli, F., Pavesi, G., Gissi, C., Horner, D. S., and Pesole, G. (2010). Assessment of orthologous splicing isoforms in human and mouse orthologous genes. *BMC Genomics*, 11:534.

Zhai, Z., Ha, N., Papagiannouli, F., Hamacher-Brady, A., Brady, N., Sorge, S., Bezdan, D., and Lohmann, I. (2012). Antagonistic regulation of apoptosis and differentiation by the cut transcription factor represents a tumor-suppressing mechanism in drosophila. *PLoS Genet*, 8(3):e1002582.

Zhang, A. (2009). *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, 1st edition. ISBN: 9780521888950.

Zhang, M. and Gish, W. (2006). Improved spliced alignment from an information theoretic approach. *Bioinformatics*, 22(1):13–20.

Zhang, Y. and Dufau, M. L. (2002). Silencing of transcription of the human luteinizing hormone receptor gene by histone deacetylase-mSin3A complex. *J Biol Chem*, 277(36):33431–33438.

Zhao, Y. and Jensen, O. N. (2009). Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics*, 9(20):4632–4641.

Zorzan, S., Lorenzetto, E., Ettorre, M., Pontelli, V., Laudanna, C., and Buffelli, M. (2013). Homecat: consensus homologs mapping for interspecific knowledge transfer and functional genomic data integration. *Bioinformatics*, 29(12):1574–1576.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Name, Vorname

....

Ort, Datum Unterschrift Doktorand/in

Formular 3.2