Analysis of High-Throughput Data – Protein-Protein Interactions, Protein Complexes and RNA Half-life

Caroline Christina Friedel

München 2008

Analysis of High-Throughput Data – Protein-Protein Interactions, Protein Complexes and RNA Half-life

Caroline Christina Friedel

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

vorgelegt von
Caroline Christina Friedel
aus München

München, den 4.12.2008

Erstgutachter: Prof. Dr. Ralf Zimmer
Zweitgutachter: Prof. Dr. Hans-Werner Mewes
Tag der mündlichen Prüfung: 4.2.2009

Contents

List of Figures

List of Tables

Summary

The development of high-throughput techniques has lead to a paradigm change in biology from the small-scale analysis of individual genes and proteins to a genome-scale analysis of biological systems. Proteins and genes can now be studied in their interaction with each other and the cooperation within multi-subunit protein complexes can be investigated. Moreover, time-dependent dynamics and regulation of these processes and associations can now be explored by monitoring mRNA changes and turnover. The in-depth analysis of these large and complex data sets would not be possible without sophisticated algorithms for integrating different data sources, identifying interesting patterns in the data and addressing the high variability and error rates in biological measurements. In this thesis, we developed such methods for the investigation of protein interactions and complexes and the corresponding regulatory processes.

In the first part, we analyze networks of physical protein-protein interactions measured in large-scale experiments. We show that the topology of the complete interactomes can be confidently extrapolated despite high numbers of missing and wrong interactions from only partial measurements of interaction networks. Furthermore, we find that the structure and stability of protein interaction networks is not only influenced by the degree distribution of the network but also considerably by the suppression or propagation of interactions between highly connected proteins. As analysis of network topology is generally focused on large eukaryotic networks, we developed new methods to analyze smaller networks of intraviral and virus-host interactions. By comparing interactomes of related herpesviral species, we could detect a conserved core of protein interactions and could address the low coverage of the yeast two-hybrid system. In addition, common strategies in the interaction of the viruses with the host cell were identified.

New affinity purification methods now make it possible to directly study associations of proteins in complexes. Due to experimental errors the individual protein complexes have to be predicted with computational methods from these purification results. As previously published methods relied more or less heavily on existing knowledge on complexes, we developed an unsupervised prediction algorithm which is independent from such additional data. Using this approach, high-quality protein complexes can be identified from the raw purification data alone for any species purification experiments are performed. To identify the direct, physical interactions within these predicted complexes and their subcomponent structure, we describe a new approach to extract the highest scoring subnetwork connecting the complex and interactions not explained by alternative paths of indirect in-

teractions. In this way, important interactions within the complexes can be identified and their substructure can be resolved in a straightforward way.

To explore the regulation of proteins and complexes, we analyzed microarray measurements of mRNA abundance, de novo transcription and decay. Based on the relationship between newly transcribed, pre-existing and total RNA, transcript half-life can be estimated for individual genes using a new microarray normalization method and a quality control can be applied. We show that precise measurements of RNA half-life can be obtained from de novo transcription which are of superior accuracy to previously published results from RNA decay. Using such precise measurements, we studied RNA half-lives in human B-cells and mouse fibroblasts to identify conserved patterns governing RNA turnover. Our results show that transcript half-lives are strongly conserved and specifically correlated to gene function. Although transcript half-life is highly similar in protein complexes and families, individual proteins may deviate significantly from the remaining complex subunits or family members to efficiently support the regulation of protein complexes or to create non-redundant roles of functionally similar proteins.

These results illustrate several of the many ways in which high-throughput measurements lead to a better understanding of biological systems. By studying large-scale measurements in this thesis, the structure of protein interaction networks and protein complexes could be better characterized, important interactions and conserved strategies for herpesviral infection could be identified and interesting insights could be gained into the regulation of important biological processes and protein complexes. This was made possible by the development of novel algorithms and analysis approaches which will also be valuable for further research on these topics.

Zusammenfassung

Die Entwicklung von Hochdurchsatzmethoden hat zu einem Paradigmenwechsel in der Biologie weg von der Analyse individueller Gene und Protein hin zu einer genomweiten Analyse ganzer biologischer Systeme geführt. Interaktionen von Proteinen und Genen miteinander und das Zusammenspiel innerhalb von Proteinkomplexen kann nun untersucht werden. Darüber hinaus ermöglichen es Messungen von mRNA Auf- und Abbau die Dynamik und Regulation dieser Prozesse und Interaktionen zu erforschen. Die detaillierte Analyse dieser großen und komplexen Datensätze wäre nicht möglich ohne hochentwickelte Algorithmen, die verschieden Datenquellen integrieren, interessante Muster in den Daten erkennen und mit der hohen Variabilität und den hohen Fehlerraten in biologischen Messungen umgehen können. In dieser Doktorarbeit wurden solche Methoden entwickelt, um Protein-Protein Interaktionen und Proteinkomplexe und die dazugehörigen regulatorischen Prozesse zu untersuchen.

Im ersten Teil analysieren wir Netzwerke aus physikalischen Protein-Protein Interaktionen, die in groß angelegten Messreihen bestimmt wurden. Wir zeigen, dass die Topologie der Netzwerke mit hoher Konfidenz aus nur teilweise gemessenen Interaktionsnetzwerken abgeleitet werden kann trotz einer hohen Anzahl von fehlenden und falschen Interaktionen. Außerdem zeigen wir, dass die Struktur und Stabilität von Protein-Protein Interaktionsnetzwerken nicht nur von der Gradverteilung bestimmt wird sondern auch beträchtlich von der Art wie Interaktionen zwischen stark verbundenen Proteinen entweder unterdrückt oder verstärkt werden. Da die Analyse von Netzwerken sich im Allgemeinen auf große eukaryotische Netzwerke konzentriert, entwickelten wir neue Methoden zur Analyse kleinerer Netzwerke von intraviralen und Virus-Wirt Interaktionen. Durch den Vergleich der Interaktome verwandter Herpesviren konnte ein konservierter Kern aus Protein-Protein Interaktionen identifiziert und die niedrige Sensitivität des Yeast Two-Hybrid Systems adressiert werden. Zusätzlich wurden gemeinsame Strategien in der Interaktion der Viren mit der Wirtszelle bestimmt.

Neue Purifikationsmethoden ermöglichen es direkt die Assoziationen von Proteinen in Komplexen zu untersuchen. Aufgrund experimenteller Fehler sind bioinformatische Methoden nötig um die individuellen Proteinkomplexe aus den Purifikationsergebnissen zu bestimmen. Da bisher publizierte Methoden mehr oder weniger stark auf bekanntes Wissen über Proteinkomplexe zurückgreifen, entwickelten wir einen Algorithmus, der unabhängig von solch zusätzlichen Daten ist. Mithilfe dieses Ansatzes können wir Proteinkomplexe mit hoher Genauigkeit aus den Rohdaten für jede Spezies vorhersagen, für die Purifika-

tionsexperimente durchgeführt werden. Außerdem beschreiben wir eine neue Methode, um die direkten, physikalischen Interaktionen innerhalb dieser vorhergesagten Komplexe und ihre Substruktur zu bestimmen. Dabei wird das Subnetzwerk mit dem höchsten Gewicht identifiziert, welches den Komplex verbindet. Anschließend wird das Netzwerk um die Interaktionen erweitert, welche nicht durch alternative Pfade von indirekten Interaktionen erklärt werden können. Auf diese Weise können wichtige Interaktionen innerhalb der Komplexe vorhergesagt werden und ihre Substruktur auf einfache und intuitive Weise bestimmt werden.

Um die Regulation von Proteinen und Komplexen zu erforschen, analysierten wir Messungen zur Gesamt-mRNA-Menge in einer Zelle sowie der Transkription und des Abbaus von RNA. Basierend auf der Beziehung zwischen den verschiedenen RNA Fraktionen können Halbwertszeiten für individuelle Transkripte mithilfe einer neuen Microarray Normalisierungsmethode berechnet werden und eine Qualitätskontrolle angewendet werden. Wir zeigen, dass präzise RNA Halbwertszeiten aus Messungen der RNA Transkription bestimmt werden können. Diese haben eine höhere Genauigkeit als bisher publizierte Ergebnisse, welche auf Messungen des RNA Abbaus basieren. Mithilfe solch präziser Messungen wurden RNA Halbwertszeiten in menschlichen B-Zellen und murinen Fibroblasten bestimmt, um konservierte Muster zu identifizieren, welche die Geschwindigkeit des RNA Auf- und Abbaus beeinflussen. Unsere Ergebnisse zeigen, dass mRNA Halbwertszeiten stark konserviert sind und spezifisch mit der Funktion der entsprechenden Gene korreliert sind. Obwohl mRNA Halbwertszeiten innerhalb von Proteinkomplexen und Familien sehr ähnlich sind, können einzelne Proteine signifikant von den anderen Untereinheiten im Komplex oder den anderen Familienmitgliedern abweichen. Dies ermöglicht einerseits die effiziente Regulation von Proteinkomplexen und andererseits nicht-redundante Funktionen funktionell sehr ähnlicher Proteine.

Unsere Ergebnisse zeigen einige der vielen Wege in denen Hochdurchsatzmessungen zu einem besseren Verständnis biologischer Systeme führen. Durch die Untersuchung von groß angelegten Messungen konnte die Struktur von Protein-Protein Interaktionsnetzwerken und Proteinkomplexen besser charakterisiert werden, wichtige Interaktionen und konservierte Strategien für die Herpesvirusinfektion identifiziert werden und interessante Erkenntnisse zur Regulation wichtiger biologischer Prozesse and Proteinkomplexe gewonnen werden. Dies wurde ermöglicht durch die Entwicklung neuartiger Algorithmen und Analyseansätze, welche auch für die weitere Forschung auf diesen Gebieten wertvoll sein werden.

Chapter 1

Motivation and overview

1.1 High-throughput data in bioinformatics

Before the development of current high-throughput technologies, genes, proteins and interactions between them were studied one at a time and the results of these small-scale experiments could still be evaluated manually. Even at this slow rate, however, biological information accumulated fast and, thus, databases were established to store, manage and organize the available data (e.g. GenBank (Benton, 1990), PDB (Bernstein *et al.*, 1977), Swiss-Prot (Bairoch and Boeckmann, 1991)).

Advances in experimental techniques eventually made it possible to identify several genes at a time by sequencing expressed sequence tags (ESTs) (Adams *et al.*, 1991). At the beginning of the 1990s, when the sequencing of the complete human genome was still a decade away, gene identification via EST sequencing was seen as the most promising way to increase the number of known genes rapidly (Adams *et al.*, 1991, 1992). Even then, the amount of data generated could no longer be evaluated without the use of efficient computer programs.

Improvements of sequencing technologies and new sequencing strategies such as whole genome shotgun (Fleischmann *et al.*, 1995) and massively parallel sequencing eventually made it possible to sequence complete genomes and in particular the human genome (Fleischmann *et al.*, 1995; Goffeau *et al.*, 1996; Blattner *et al.*, 1997; C. elegans Sequencing Consortium, 1998; Adams *et al.*, 2000; Lander *et al.*, 2001; Venter *et al.*, 2001; Levy *et al.*, 2007). Here, the importance and necessity of bioinformatics became apparent as efficient algorithms were developed for assembling genomes from short sequence reads, predicting the location of genes in the sequence and annotating and analyzing the genome sequences (Mewes *et al.*, 1997; Frishman *et al.*, 2001; Hubbard *et al.*, 2002).

While the sequencing of the first human genomes cost several hundred million dollars, next generation sequencing technologies (Genome sequencer by 454/Roche, Solexa sequencer by Illumina, the SOLiD system by Applied Biosystems) now allow the sequencing of individual genomes at less then US$1 million, a fraction of the original costs (Wheeler *et al.*, 2008; Wang *et al.*, 2008b; Bentley *et al.*, 2008). As prices are dropping even further,

this opens up new applications of massively parallel sequencing for the identification of protein-DNA interactions (Mardis, 2007; Robertson *et al.*, 2007), the discovery of miRNAs (Friedländer *et al.*, 2008) and alternative splicing (Wang *et al.*, 2008a), transcriptome profiling (Wang *et al.*, 2008c) and many more.

Since the sequencing of the first genomes the focus has shifted from a small-scale analysis of individual genes or proteins to a systems level study of biological processes. New high-throughput methods now make it possible to quantify RNA abundance for thousands of genes at the same time (Schena *et al.*, 1995; Lockhart *et al.*, 1996; Shalon *et al.*, 1996; Wodicka *et al.*, 1997; Velculescu *et al.*, 1997) and to determine protein-protein interactions (Fields and Song, 1989; Fromont-Racine *et al.*, 1997; Bartel *et al.*, 1996; Uetz *et al.*, 2000; Ito *et al.*, 2001) and protein complexes (Rigaut *et al.*, 1999; Ho *et al.*, 2002; Gavin *et al.*, 2002, 2006; Krogan *et al.*, 2006) on a genome scale. Thus, while the function of individual genes and proteins was previously examined separately, it can now be investigated in the interaction and cooperation with other proteins and within protein complexes.

This increases the level of detail that biological processes can be studied at but also the complexity. The number of possible binary interactions is quadratic in the genome size and the number of possible protein complexes is even larger. In extreme cases, protein complexes can contain more than 100 protein subunits (e.g. the spliceosome). As a consequence, these interactions and complexes cannot be studied without the development of efficient and sophisticated algorithms for analyzing the large-scale data sets and extracting new biological information from them. Furthermore, as high-throughput methods often sacrifice accuracy to obtain higher coverage and speed, these algorithms have to consider and deal with high error rates in the data.

While the mapping of protein-protein interactions provides a blueprint for the architecture of biological systems, the time-dependent development of interactions and protein complexes and their regulation is to a large degree unknown. Although the regulation of complete protein complexes cannot be studied yet on a large scale, the precise study of the regulation of individual genes can lead to new insights into the regulation of these complex biological systems.

In this thesis (see Figure 1.1), we developed new methods for the analysis of high-throughput measurements of protein-protein interactions and associations of proteins in multi-subunit complexes. Furthermore, by studying RNA turnover determined from simultaneous measurements of RNA decay and transcription, we investigated the dynamics and regulation of biological processes and important protein complexes. Our results show that, despite high measurement errors, novel and interesting insights into biological systems and regulatory processes can be obtained from large-scale high-throughput data.

1.2 Thesis outline

In **part I** we analyze protein-protein interaction (PPI) networks obtained from different large-scale experiments, in particular physical interactions identified with the so-called yeast two-hybrid (Y2H) system (see section 2.1 for an introduction).

- In **chapter 2**, we provide a brief overview on network analysis. We describe large-scale experimental methods for identifying protein interaction networks, commonly analyzed network properties and algorithms for network generation and evolution.

- As high-throughput methods are error-prone, we analyze in **chapter 3** the influence of both false negative and false positive interactions on the most commonly analyzed network property: the degree distribution which characterizes the network topology. Generally, interaction data from incomplete large-scale experiments is used to infer the topology of the complete interactome. These partial networks often show a scale-free behavior with only a few proteins having many and the majority having only few connections. Recently, Han *et al.* (2005) suggested that this scale-free nature may not actually reflect the topology of the complete interactome but might be due to measurement errors of large-scale experiments. By analyzing the effects of such measurement errors on an additional network property, the clustering coefficient, we show both analytically and in simulations that the topology of the complete interaction networks can be extrapolated from incomplete and imperfect experimental networks (Friedel and Zimmer, 2006a,b). Although the correct topology of the interactome may not be inferred beyond any reasonable doubt from the interaction networks available, a number of topologies can nevertheless be excluded with high confidence.

- In **chapter 4**, we analyze correlations between the number of interactions (degree) of interacting proteins and the effect of such degree correlations on network structure and stability. Based on the analysis of one large-scale interaction network for yeast, it has been suggested that interactions between highly connected nodes (hubs) are avoided in order to suppress the propagation of perturbations in the network (Maslov and Sneppen, 2002b). More recent studies could not confirm this observation for high-confidence interaction sets and, thus, it was attributed to distorting effects of high-throughput experiments. To understand the role and prevalence of degree correlations in PPI networks, we analyze experimentally derived interaction networks and investigate how degree correlations influence the structure and resilience of these networks (Friedel and Zimmer, 2007). We find that network structure and stability is highly influenced by the way interactions between hubs are either suppressed or promoted. Nevertheless, we can show that in real PPI networks interactions between hubs are not avoided and, accordingly, that high-throughput methods do not seem to introduce a bias towards a certain type of degree correlations. Although structural properties and stability of networks can be modified considerably by degree correlations, this possibility does not actually seem to be used in biological PPI networks. A possible explanation for this may be the structural disadvantages of different types of degree correlations that we identify in our simulations.

- In **chapter 5**, we study protein-protein interaction networks of herpesviruses determined experimentally with the Y2H system by the group of Jürgen Haas, our collaborators at the Max von Pettenkofer-Institut of the LMU München (Fossum

et al., 2008; Dong *et al.*, 2008). As herpesviral genomes are very small, screens of all possible intraviral interactions are feasible and have been performed. A comparative analysis of five intraviral networks indicates that protein-protein interactions are highly conserved between different herpesvirus species. Thus, we can identify a common core of conserved protein interactions in herpesviruses. Furthermore, we analyze protein interactions between herpesviruses and their human host. Our results show that cellular targets of herpesviral proteins, in particular of conserved proteins, are highly connected and central to the host interactome but are not enriched for specific functional categories or biological processes.

Many of the direct, physical protein interactions we analyzed in the first part are formed within larger associations of proteins in complexes. These protein complexes are important components of biological systems and perform a wide range of functions. Recent developments of high-throughput technology now make it possible to directly identify associations of proteins in complexes using tandem affinity purification (TAP). However, as protein complexes can only be purified incompletely from the cell, their actual conformation has to be identified with computational methods from the purification results. In **part II**, we focus on the prediction and analysis of protein complexes from such genome-scale measurements.

- In **chapter 6**, we describe an approach to predict protein complexes (Friedel *et al.*, 2008a) from large-scale measurements obtained with the tandem affinity purification (TAP) method. Contrary to previous approaches, our algorithm does not depend on the availability of data on known complexes. We calculate intuitive interaction scores more accurate than all other published scoring methods and predict complexes with the same quality as the best supervised predictions. Differences between the best previous predictions and the complexes predicted by our approach are analyzed in detail to illustrate the advantages of our method. The results show that meaningful and reliable complexes can be determined from the purification experiments alone. Thus, this method can be applied to large-scale TAP experiments for any species even if few complexes are already known.

- When predicting protein complexes as sets of proteins, the modular substructure and the physical interactions within protein complexes are not considered. In **chapter 7**, we present an approach for identifying the direct physical interactions and the subcomponent structure of protein complexes predicted from affinity purification results. Our algorithm (Friedel and Zimmer, 2008) calculates the union of all maximum spanning trees from scoring networks for each protein complex to extract important interactions and subsequently extends this network by interactions not accounted for by alternative indirect paths. By applying this approach to the complexes predicted in chapter 6, we show that our algorithm identifies experimentally derived physical interactions with higher accuracy than baseline approaches and that the subcomponent structure of the complexes can be resolved more satisfactorily. In this way, we obtain networks of direct, physical interactions which were studied in part I.

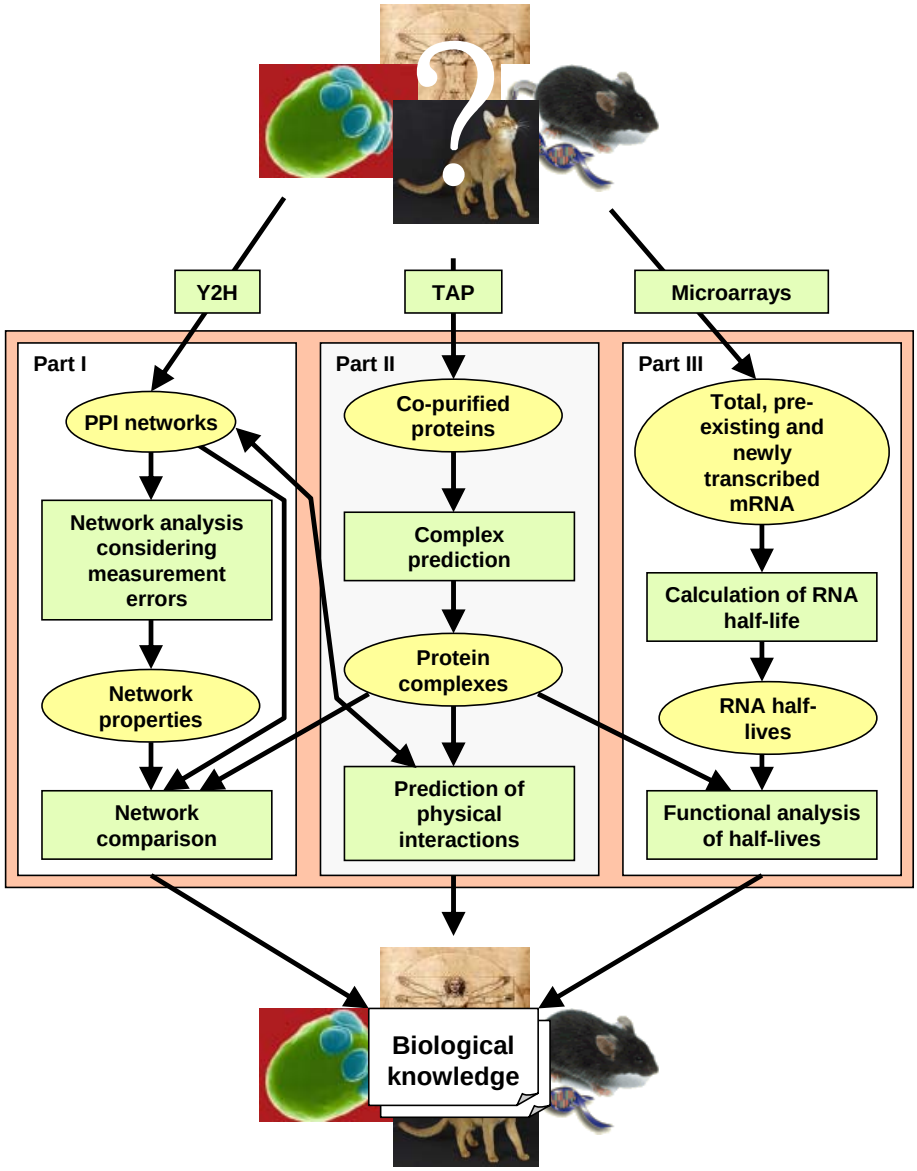


Figure 1.1: Thesis outline and connections between the different parts of the thesis. Pictures for human, mouse, yeast and cat were taken from the Ensembl website (`http://www.ensembl.org`).

To explore the time-dependent dynamics and regulation of protein interactions, complexes and biological systems in general, gene expression measurements are analyzed in **part III**. In standard gene expression analysis, mRNA abundance is quantified on a global scale using microarray technology and changes of this total abundance between different conditions are evaluated. As these approaches cannot distinguish between changes in RNA synthesis and decay, a novel high-throughput system has been developed recently by Lars Dölken

from the Max von Pettenkofer-Institut to simultaneously measure synthesis and decay in a single experimental setting (Dölken *et al.*, 2008). In **part III**, we describe methods we developed in collaboration with Lars Dölken for analyzing this new type of data and study the conservation of RNA turnover and its implications for gene function and regulation of protein complexes and important biological processes.

- In **chapter 8**, we describe an algorithm for calculating RNA turnover rates in terms of RNA half-life from measurements of de novo transcription and decay. We developed a new and intuitive method for microarray normalization based on the relationship between total, newly transcribed and pre-existing RNA (Dölken *et al.*, 2008). Using this approach, high quality probe sets can be selected for genes represented several times on the array and a quality control can be applied. Furthermore, we show that half-life measurements from de novo transcription are more precise than measurements from RNA decay and that differentially regulated genes can be identified more reliably by taking into account RNA half-life.

- In **chapter 9**, we analyze the conservation and function of RNA half-life in human B-cells and mouse fibroblasts. We find that RNA half-lives are highly conserved across species and cell lines and specifically correlated to gene function (Friedel *et al.*, 2008b). Our results indicate that while RNA half-life is relatively homogenous within protein complexes and families, significant deviations can occur to efficiently support the regulation of protein complexes and to create differential regulatory patterns in protein families. Thus, analysis of RNA half-life can lead to new insights on the regulation and dynamics of cellular processes and protein complexes.

In the final **chapter 10**, we discuss the conclusions from this work and directions for future research.

Part I

Protein-protein interactions

Chapter 2

A short introduction to network analysis

2.1 Large-scale experimental methods

Since protein-protein interactions are of fundamental importance for all processes taking place in a cell, great efforts have been devoted to the systematic identification of protein interactions for a number of organisms. To generate large-scale protein interaction maps, two methods are commonly used: (i) yeast two-hybrid (Y2H) (Fields and Song, 1989) and (ii) affinity purification techniques, e.g. Co-immunoprecipitation (Co-IP)(Ho *et al.*, 2002) or tandem affinity purification (TAP)(Rigaut *et al.*, 1999) followed by mass spectrometry.

The Y2H method is based on the Gal4 transcription factor which can initiate the transcription of the LacZ gene (see Figure 2.1). The Gal4 protein consists of two domains: a binding domain (BD) which binds to the upstream activating sequence (UAS) and an activating domain (AD) which induces transcription upon binding to the UAS. As the two domains are modular, they do not have to be combined in the same protein to initiate transcription but it is sufficient if they are in close proximity. This fact is used by fusing a target protein, the so-called bait protein to the binding domain and a potential interaction partner of the bait, a so-called prey to the activating protein. If the two proteins interact, binding and activating domain are in contact via the bait and prey proteins and the reporter gene is expressed.

While the Y2H system identifies direct physical interactions, affinity purification can also identify indirect interactions via other proteins. For this purpose, bait proteins are purified from cell extracts with affinity chromatography. Proteins interacting directly or indirectly with the bait are then co-purified and identified later with mass spectrometry. For co-immunoprecipitation (Co-IP), a protein of interest and its interaction partners are purified from the cell lysate using an antibody specific to this protein. Depending on the experimental set-up, Co-IP can also be used to verify physical interactions. As a consequence, Co-IP experiments are often performed as a control in large-scale Y2H experiments to estimate specificity (e.g. Rual *et al.*, 2005; Stelzl *et al.*, 2005).

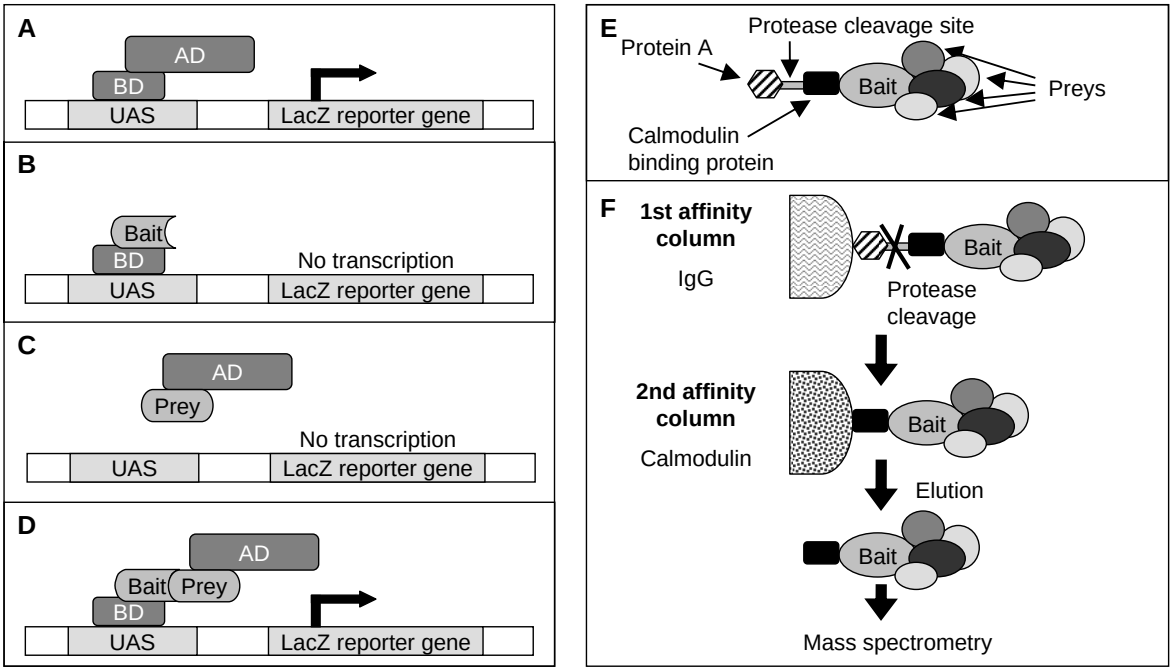


Figure 2.1: Experimental methods for determining protein-protein interactions. (**A-D**) Outline of the yeast-two hybrid (Y2H) method (Fields and Song, 1989). (**A**) The GAL4 protein consists of a binding domain (BD) and an activating domain (AD) and can initiate transcription of the LacZ reporter gene. Fusion proteins of the binding domain and a bait protein (**B**) as well as the activating domain and a prey protein (**C**) are prepared. Neither of those can induce transcription on its own. (**D**) The interaction of bait and prey results in transcription of the reporter gene. (**E-F**) Outline of tandem affinity purification (TAP) (Rigaut *et al.*, 1999). (**E**) Structure of the TAP tag. (**F**) Outline of the tandem affinity purification procedure.

For tandem affinity purification (TAP), the target protein (bait) is fused to the so-called TAP tag which consists of protein A and a calmodulin binding protein separated by a protease cleavage site (Figure 2.1). The bait is then purified using two consecutive affinity columns. In the first one, immunoglobulin G (IgG) beads bind to the protein A part of the tag such that the bait protein and interacting proteins are retained. Possible contaminants are then washed off. Afterwards, the bait protein is cut off from the IgG column by protease cleavage and purified with a second Calmodulin column. The calmodulin binding protein binds to the calmodulin beads in the presence of Ca^{2+}. The protease and additional contaminants are then washed off. Finally, the bait protein and interacting prey proteins are eluted with EDTA and the preys are identified with mass spectrometry.

The Y2H method has been applied on a genome-scale to yeast (Uetz *et al.*, 2000; Ito *et al.*, 2001; Yu *et al.*, 2008), *Drosophila* (Giot *et al.*, 2003), *C. elegans* (Li *et al.*, 2004), *P. falciparum* (LaCount *et al.*, 2005) and human (Rual *et al.*, 2005; Stelzl *et al.*, 2005).

Large-scale affinity purification experiments have so far only been performed in yeast (Ho *et al.*, 2002; Gavin *et al.*, 2002, 2006; Krogan *et al.*, 2006).

Both methods are prone to spurious interactions (false positives) due to self-activators (Y2H), protein contaminants (affinity purification) or non-specific interactions. Based on expression data and information about paralogs, the fraction of wrong high-throughput interactions has been estimated for the first large-scale experiments at 30-50% (Deane *et al.*, 2002). In the more recent human Y2H experiments, more than 60% of interactions could be verified with additional Co-IP measurements (Rual *et al.*, 2005; Stelzl *et al.*, 2005). In addition to false positives, high-throughput experiments are characterized by a large fraction of false negatives, i.e. correct interactions that are missed in the experiment. Accordingly, only small overlaps can be observed between interaction maps for the same species but determined in different experiments and with different methods (Bader and Hogue, 2002; Bader *et al.*, 2004; Yu *et al.*, 2008).

2.2 Network properties

A protein-protein interaction (PPI) network can be described as an undirected graph $G = (V, E)$ with a set of nodes V and a set of edges E. The nodes in G then correspond to interacting proteins and two nodes u and v are connected by an edge (u, v) if and only if they interact. If u and v interact, they are denoted as neighbors. Depending on the experimental method, interactions may be either direct physical interactions (Y2H) or include also indirect interactions via other proteins (affinity purification).

Interactions can be extracted from affinity purification results using either the spoke or matrix model. In the spoke model, only interactions between the bait and its prey proteins are included, while in the matrix model interactions between preys of the same bait are also included. The spoke model is more accurate but coverage of the matrix model is higher (Bader *et al.*, 2004). Networks derived with the spoke or matrix model are unweighted. Alternatively, interaction weights can be calculated with different scoring methods (see chapter 6).

Generally, the first step in analyzing experimental networks is an analysis of their structural properties. The most important network properties are described in the following.

2.2.1 Degree distribution

The most commonly analyzed network characteristic is the *degree distribution* (Dorogovtsev and Mendes, 2002; Albert and Barabási, 2002; Newman, 2003). Often, degree distribution is used synonymously with network topology. The degree k_v of a node v is the number of its edges and the average degree $\bar{k}$ is calculated as $2|E|/|V|$. The degree distribution describes the probability of a node v having degree k (see Figure 2.2 for examples):

$$P(k) = \frac{|\{v \in V | k_v = k\}|}{|V|}. \tag{2.1}$$

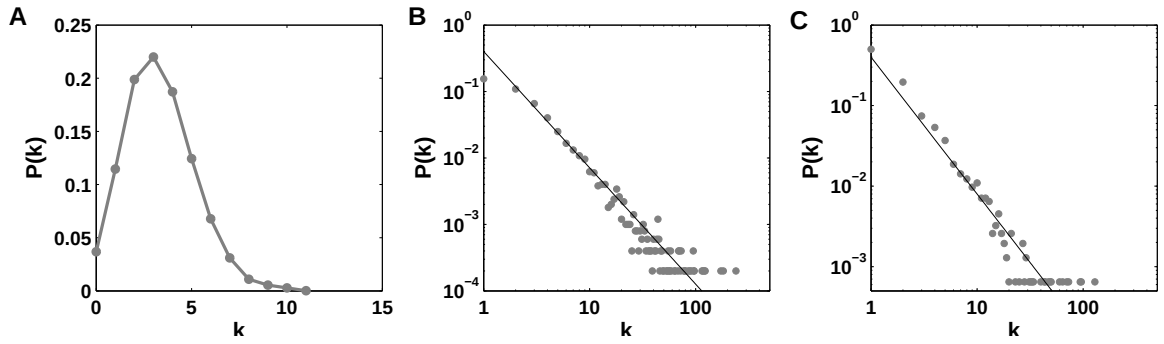


Figure 2.2: Degree distributions for a random graph network with average degree 5 (**A**), a power-law network with $\gamma = 1.75$ (**B**) and the human interaction network identified by Rual *et al.* (2005) (**C**). Black lines in **B** and **C** indicate the fit to the power-law distribution.

As we will see later, an important characteristic of the degree distribution is its asymmetry, i.e. its skewness. Although there exist several alternative definitions of skewness, the one most commonly used is

$$skewness = \frac{\sum_{v \in V}(k_v - \bar{k})^3}{(|V| - 1)s^3} \tag{2.2}$$

where s is the sample standard deviation of the degree distribution. For symmetric distributions the skewness is close to zero whereas for left-tailed distributions it is negative and for right-tailed distribution, such as e.g. power-law distributions (see below), it is positive.

Random graphs

Random graphs were introduced by Erdős and Rényi (1959) as a model of random networks. In random graphs the number of nodes n is fixed and an edge between two nodes is included with a constant probability p. The degree distribution is then binomial and for large n it can be approximated by a Poisson distribution (Bollobás, 2001):

$$P(k) = e^{-\bar{k}}\bar{k}^k/k!. \tag{2.3}$$

The average degree in this case is $\bar{k} = p(n-1)$.

Scale-free networks

In random graphs, most nodes have a degree k around the average degree $\bar{k}$ and the probability of extremely large degrees is effectively 0. In protein-protein interaction networks, however, most nodes have a small degree, i.e. most proteins interact with few other proteins, but a small fraction of nodes (so-called hubs) have connections to many other nodes in the network (Figure 2.2).

This topology of PPI networks is generally described as scale-free (Jeong *et al.*, 2001; Wagner, 2001; Yook *et al.*, 2004) and is common to many networks from various domains (Dorogovtsev and Mendes, 2002; Albert and Barabási, 2002; Newman, 2003). Scale-free networks are characterized by a power-law degree distribution with $P(k) \propto k^{-\gamma}$ for a constant γ (see Figure 2.2). Power-law functions are called scale-free because the scaling of k by a constant factor c leads only to a proportionate scaling of the power function:

$$P(ck) \propto (ck)^{-\gamma} = c^{-\gamma}k^{-\gamma} \propto P(k). \tag{2.4}$$

If plotted in logarithmic scales, the power-law function is a straight line and can be easily identified. Generally, however, the degree distribution of protein-protein interaction network is not a perfect straight line in a log-log plot and may be more accurately described by a generalized power law ($P(k) \propto (k+\alpha)^{-\gamma}$), by an exponential cut-off ($P(k) \propto k^{-\gamma}e^{-k/\beta}$ for $\beta \geq 1$) which limits the maximum possible degree or a combination of both. It has also been suggested that at least some protein-protein interaction networks are not scale-free at all but may be better modeled by an exponential distribution ($P(k) = \lambda e^{-\lambda k}$) (Tanaka *et al.*, 2005) or a random geometric model (Przulj *et al.*, 2004).

2.2.2 Clustering coefficient

The *clustering coefficient* quantifies the probability that two vertices which are connected to the same node are also connected. Accordingly, the clustering coefficient C_v of a node v in a network is defined as (Watts and Strogatz, 1998):

$$\begin{aligned} C_v &= P\big((u,w) \in E \mid (u,v) \in E \wedge (v,w) \in E\big) \\ &= \frac{P\big((u,w) \in E \wedge (u,v) \in E \wedge (v,w) \in E\big)}{P\big((u,v) \in E \wedge (v,w) \in E\big)} \\ &=: \frac{P(\bigtriangledown \in E)}{P(\vee \in E)}. \end{aligned} \tag{2.5}$$

More simply, the clustering coefficient of a node is calculated as

$$C_v = \frac{2|E_v|}{k_v(k_v-1)}. \tag{2.6}$$

where E_v are the edges between neighbors of v. Since the clustering coefficient is only defined for nodes with at least two neighbors, the clustering coefficient C of the complete network is defined as the average clustering coefficient of all nodes with degree at least 2. The clustering coefficient of random graphs is p and significantly lower than the clustering coefficient of scale-free networks (see chapter 3).

2.2.3 Characteristic path length

Characteristic path length L is defined as the average shortest path length between all pairs of nodes (Watts and Strogatz, 1998). The shortest path length d_{ij} between two nodes v_i

and v_j in an undirected and unweighted network is calculated as:

$$d_{ij} = \begin{cases} min|(v_i \rightarrow \cdots \rightarrow v_j)| & \text{if a path } (v_i \rightarrow \cdots \rightarrow v_j) \text{ exists} \\ \infty & \text{otherwise} \end{cases} \tag{2.7}$$

For calculation of characteristic path length, only pairs are considered for which $d_{ij} < \infty$, i.e. which are actually connected by at least one path in the network. L can be calculated efficiently in $O(|V||E|)$ for an unweighted network $G = (V, E)$ by breadth-first searches starting from each node on the unweighted interaction graph. For weighted networks Dijkstra's algorithm can be used to calculate weighted shortest paths (Cormen *et al.*, 2000). Sometimes, the maximum shortest path length between any pair of notes is also analyzed which is denoted as the *diameter* of the graph.

2.2.4 Small-world effect

For random graphs, characteristic path length scales logarithmically with the size of the network (Chung and Lu, 2002; Albert and Barabási, 2002):

$$L_{rand} \sim log(|V|)/log(\bar{k}). \tag{2.8}$$

Thus, it is relatively small even for large networks. Protein-protein interaction networks, as most real-life networks, also show small characteristic path length but contrary to random graphs, have relatively high clustering coefficients (Han *et al.*, 2005). This combination of small characteristic path length and large clustering coefficients is generally denoted as the small-world effect (Watts and Strogatz, 1998).

2.2.5 Error and attack tolerance

An additional network property which is often analyzed is the tolerance of the network to deletions of selected nodes. If nodes are removed randomly irrespective of their degree, this is denoted as error or random failure. Contrary to that, a selective deletion of the most-connected nodes, i.e. the hubs, is generally described as targeted attack. The influence of the degree distribution on error and attack tolerance was analyzed by Albert *et al.* (2000) who used characteristic path length as a measure of network interconnectness. Removing nodes from the network increases characteristic path length until the network becomes disconnected and communication between nodes is no longer possible. Thus, the slower the increase of characteristic path length upon removal of nodes, the less vulnerable is the network.

Albert *et al.* (2000) found that the degree distribution has a severe effect on network resilience to errors or attack (see Figure 2.3). Random graphs are vulnerable to random failure and targeted attack in a similar way. Contrary to that, scale-free networks are very resilient to random failure but on the other hand much more vulnerable to a targeted deletion of hubs. As protein-protein interaction networks show a scale-free topology, this

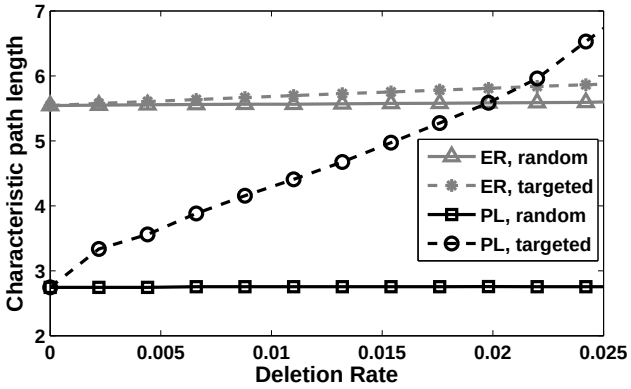


Figure 2.3: Error and attack tolerance in random graph and scale-free networks. Development of characteristic path length with increasing deletion rates is shown for the random graph (ER) and power-law network (PL) from Figure 2.2. Nodes are either deleted without regard to their degree (random) or the highest connected nodes are deleted selectively (targeted).

suggests that highly interactive proteins are more essential than less connected ones. Indeed, it has been shown for large-scale yeast interaction networks that highly connected hubs are more likely to be essential and conserved in eukaryotes than less connected proteins (Jeong *et al.*, 2001; Wuchty, 2004), although this has been questioned recently (Coulomb *et al.*, 2005). Furthermore, recent studies have proposed that the enrichment of essential proteins among hubs may not be due to the central role of hubs in the network, but instead a consequence of their involvement in essential protein-protein interactions (He and Zhang, 2006) or complex biological modules (Zotenko *et al.*, 2008).

2.2.6 Betweenness centrality

The degree of a node is one possible way to quantify the centrality of a node based on local connectivity. An alternative measure based on both local and global measures, is the betweenness centrality (Freeman, 1977). The betweenness centrality $C_B(v)$ of a node v is defined as the fraction of shortest paths between any pair of nodes s and t which also contain v.

Let σ_{st} be the number of shortest pairs betwen s and t and $\sigma_{st}(v)$ the number of these shortest paths which pass v. Then, we have that

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.9}$$

To control for network size, $C_B(v)$ is divided by the total number of node pairs. Betweenness centrality for all nodes in unweighted networks can be calculated efficiently in $O(|V||E|)$ using the algorithm by Brandes (2001).

2.3 Reference networks

Clustering coefficients and characteristic path length depend both on the degree distribution and the size of the network. In order to assess if the clustering coefficient in a network is high and characteristic path length is small, the network is generally compared against random graphs of the same size and the same number of edges. However, in many cases it is also interesting if topological properties differ significantly from random networks with the same degree distribution as the network considered or other degree distributions. In the following, we describe how reference networks can be created with the same distribution as in a specific network or with arbitrary degree distributions.

2.3.1 Rewired reference networks

Random networks with the same degree distribution as a given network (the "null model") can be easily obtained by randomly rewiring edges many times such that the degree distribution is preserved (Maslov and Sneppen, 2002b). Here, rewiring consists in deleting two random edges (u, v) and (w, x) and replacing them by two edges (u, x) and (w, v) if these do not exist already. Here, further restrictions can be imposed if not all edges are possible, e.g. if the network under consideration was created in a large-scale experiment with different sets of baits and preys. In this case, edges between two preys may not be allowed. This edge-swapping strategy is repeated a sufficiently large number of times to completely randomize the network.

2.3.2 Networks with arbitrary degree distributions

Random networks with a given degree distribution can be generated using the method described by Chung and Lu (2002). For this purpose, an expected degree sequence $\mathbf{w} = (w_1, \ldots, w_n)$ is drawn first from the chosen distribution for the n nodes. An edge is created between nodes v_i and v_j with probability $p_{ij} = w_i w_j \rho$ with $\rho = (1/\sum_i w_i)$. The expected degree of node v_i is then $\sum_j p_{ij} = w_i$.

2.4 Network evolution

To understand how real networks have evolved to have specific network properties, such as the ubiquitous scale-free topology, several models have been developed which simulate network growth by the stepwise addition of nodes and edges. The objective behind this is to identify evolutionary processes which lead to the specific structural features of real networks. In the following, we describe *preferential attachment*, the standard model for the development of scale-free networks by Barabási and Albert (1999), as well as specific models for the evolution of protein-protein interaction networks by gene duplication.

2.4.1 Preferential attachment

The model of Barabási and Albert (1999) is based on the work of Simon (1955) who showed that power-law distributions arise naturally from certain stochastic processes in which "the rich get richer". It is similar to the *cumulative advantage* model described by Price (1976) but more well-known.

The network generation algorithm of the Barabási-Albert model consists of two major parts:

1. Growth: At each time step, a new node is added with m edges to nodes already contained in the network. At the beginning the network consists of a small number (m_0) of nodes.

2. Preferential attachment: The new node is attached to an existing node i in the network with probability proportional to the degree k_i of i:

$$p_i = \frac{k_i}{\sum_j k_j}. \tag{2.10}$$

It has been shown that this model generates power-law networks with a power coefficient γ=3 for all values of m. Several generalizations of this model have been proposed to better explain the structure of real networks (e.g. Krapivsky *et al.*, 2000; Krapivsky and Redner, 2001; Dorogovtsev *et al.*, 2000; Dorogovtsev and Mendes, 2001; for a more extensive list refer to Table III in Albert and Barabási, 2002).

2.4.2 Duplication models

Although preferential attachment has become the standard model for the development of scale-free networks it not always evident how it emerges. For instance, although protein-protein interaction networks grow due to evolutionary processes, it is not clear why these networks should grow according to preferential attachment. Here, gene duplications have been proposed as a major mechanism of PPI network growth which lead to a scale-free topology (Qian *et al.*, 2001; Kim *et al.*, 2002; Solé *et al.*, 2002; Bhan *et al.*, 2002; Vázquez *et al.*, 2003; Chung *et al.*, 2003; Berg *et al.*, 2004; Ispolatov *et al.*, 2005; Middendorf *et al.*, 2005). An alternative explanation has been proposed by Eisenberg and Levanon (2003) who suggested that highly connected proteins are under selective pressure to become even more connectable.

It is straightforward that copying mechanisms such as gene duplication constitute some type of preferential attachment. Highly connected nodes have many interaction partners and as a consequence the propability is high that one of their interaction partners is copied and that they consequently gain an interaction. Several duplication mechanisms have been described which explain the scale-free topology of protein interaction networks more or less well. Apart from gene duplications, they generally also describe mutational mechanisms for gaining and loosing existing interactions. In particular, it is important that interaction loss is modelled as the average degree would otherwise increase rapidly.

In the following chapters, we used a combination of the growth models by Vázquez *et al.* (2003) and Solé *et al.* (2002) to simulate the evolution of real protein interaction networks. It generates highly skewed networks which are similar to power-law networks. Starting from a small random graph, the network grows by duplication until a certain size is reached. At each duplication step a node is added to the network according to the following rules:

i) Duplication: A randomly selected node v is copied to create a new node v' with connections to all neighbors of v. The nodes v and v' are connected with probability p.

ii) Divergence: Each interaction (u, v') is deleted with probability q.

By increasing p the clustering coefficients of the network can be raised. However, there are limitations to the clustering coefficients which can be achieved with this approach without drastically diverging from a power-law topology at small degree values.

Chapter 3

Inferring the topology of protein-protein interaction networks

3.1 Introduction

Although protein-protein interaction (PPI) networks identified in large-scale experiments are characterized by large false positive and false negative rates, the topology of these experimental networks has been thoroughly investigated (e.g. Jeong *et al.*, 2001; Wagner, 2001; Eisenberg and Levanon, 2003; Wuchty, 2004; Yook *et al.*, 2004; Coulomb *et al.*, 2005; see also chapter 2). Here, all studies implicitly assumed that the topology of the complete interactome can be inferred from measured PPI networks containing only a fraction of proteins and interactions. Recently, this assumption has been called into question (Stumpf *et al.*, 2005; Han *et al.*, 2005). Based on mathematical modeling, Stumpf *et al.* (2005) showed that, unlike for random graph and exponential topologies, random sampling from scale-free networks has a distorting effect on the topology of subnetworks. Conversely, these results imply that the scale-free topology of the PPI networks is unlikely to result from random graphs or exponential networks by the random sampling approach postulated by Stumpf *et al.*, which selects only a fraction of nodes and all edges between these nodes.

Since such a random sampling procedure does not accurately reflect the impact of large-scale experimental methods, Han *et al.* (2005) defined a different limited sampling procedure which emulates the effect of the yeast two-hybrid (Y2H) approach. Based on simulations they argue that such a limited sampling can lead to an apparent scale-free topology in the sampled networks regardless of the original topology. They conclude that, while a scale-free topology appears to be more likely than the other models considered, these other topologies cannot be safely excluded based on the degree distribution alone given the currently available interaction data. This implies that it may not be possible to draw any conclusions from the partial protein-protein interaction networks available to the true structure of the interactome.

In this context, we proposed (Friedel and Zimmer, 2006a,b) that apart from the degree distribution and the related network statistics discussed by Han *et al.* (power-law coeffi-

cient, goodness of fit to power-law distribution, size of the largest component and average degree of the network), other characteristics of the network might help to further assess the likelihood of different topology models and exclude at least some of them. One such characteristic is the average clustering coefficient, i.e. the "cliquishness" of the network. Here, we analyzed the effect of the sampling procedure described by Han and co-workers on the clustering coefficient analytically in addition to simulations.

Both our analytical and simulation results suggest that random sampling with a limited coverage of proteins and interactions always leads to lower clustering in the resulting sub-network compared to the original network. As a consequence, in this setting the clustering coefficients of protein interaction networks derived by Y2H can be considered as a lower bound on the clustering coefficients of the original networks and network topologies with significantly lower clustering coefficients than observed can be ruled out.

We furthermore extended the model of Han *et al.* by additionally adding spurious interactions to the sampled networks and analyzed the effects of these false positive interactions both analytically and with simulations. Although false positive interactions were just considered as another sampling artifact by Han *et al.*, we observed that the average clustering coefficient of a network is affected differently by false positive interactions than by false negative interactions.

In our model, interactions are added using a preferential attachment model (Barabási and Albert, 1999, see also section 2.4.1) and, accordingly, false positive interactions alone can increase the skewness of the theoretical networks and, thus, their similarity to scale-free networks. Our findings show that although clustering coefficients of networks can be increased by wrong interactions for some network topologies, the degree to which they can be increased depends strongly on the degree of randomness of clustering coefficients and the degree distribution of the original topology. As a consequence, several topologies remain unlikely and can be excluded with high confidence.

3.2 Methods

3.2.1 Modeling yeast-two hybrid experiments

Protein-protein interaction networks are modeled as undirected and unweighted graphs as described in chapter 2. The type of interactions included in the network differ between Y2H (only physical interaction) and affinity purification (both physical and indirect interactions via other proteins). Since these differences make it difficult to define a comprehensive model for both experimental methods, the sampling procedure described by Han *et al.* (2005) simulates only the Y2H approach.

Although many topological properties can be analyzed, we concentrate on two of them, the degree distribution and the average clustering coefficient as described in chapter 2. In the following, a network is defined as randomly clustered if clustering coefficients are hardly changed by rewiring the network (see section 2.3.1). Consequently, a network is clustered less than randomly if clustering coefficients are increased by rewiring or more

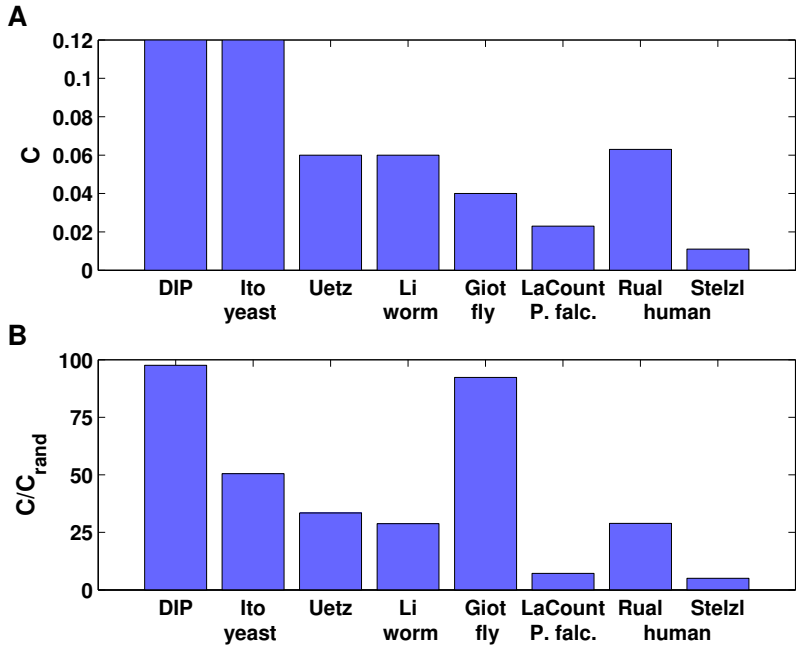


Figure 3.1: Clustering coefficients in large-scale Y2H interaction networks. Clustering coefficients (**A**) and the ratio to clustering coefficients of random graphs (Erdős and Rényi, 1959) with the same size (**B**) are shown for the following interaction networks: yeast interactions from DIP (Xenarios *et al.*, 2002) and the Y2H studies by Ito *et al.* (2001) and Uetz *et al.* (2000); *C. elegans* interactions by Li *et al.* (2004); *Drosophila* interactions by Giot *et al.* (2003); *P. falciparum* interactions by LaCount *et al.* (2005); and human interactions from the studies of Rual *et al.* (2005) and Stelzl *et al.* (2005). Only high-confidence interactions were considered for the Ito, Li and Giot data set and self-edges were ignored for the calculation of clustering coefficients.

than randomly if they are decreased. We will see examples for all three cases later on.

Figure 3.1 **A** shows the average clustering coefficients for a number of high-throughput Y2H data sets (Ito *et al.*, 2001; Uetz *et al.*, 2000; Li *et al.*, 2004; Giot *et al.*, 2003; Rual *et al.*, 2005; Stelzl *et al.*, 2005; LaCount *et al.*, 2005). Here, only high-confidence interactions were considered for the data sets of Ito *et al.* (2001), Li *et al.* (2004) and Giot *et al.* (2003). For comparison purposes, the same characteristics are given for the yeast protein-protein interaction network from DIP (Xenarios *et al.*, 2002) (version of April 2nd, 2006) which contains high-throughput data as well as interactions determined with other experimental methods. Although the clustering coefficients of some of the partial networks appear to be rather small, they are in most cases at least one order of magnitude higher than clustering coefficients of random graphs with the same number of nodes and edges (see Figure 3.1 **B**).

3.2.2 Missing interactions

The sampling procedure described by Han *et al.* (2005) simulates the effect of the Y2H method under the assumption that interactions may be missed in the process but no wrong interactions are obtained. It is determined uniquely by two parameters: bait coverage (denoted by β) and edge coverage (denoted by ε). Bait coverage specifies the selective effect of choosing only a fraction of the proteome as baits in a large-scale Y2H experiment, whereas edge coverage determines the fraction of true interactions which can actually be resolved for a bait. Accordingly, a network is sampled from the original network as follows. A fraction β of nodes is selected as baits and then for each bait a fraction ε of its interactions. Edges connecting two baits are selected with higher probability $2\varepsilon - \varepsilon^2 = \varepsilon(2-\varepsilon)$. The sampled network then contains the bait nodes as well as non-bait nodes which are connected to a bait via a sampled edge. In the following, the latter ones are referred to as preys. The resulting network is referred to as $G^1 = (V^1, E^1)$ and the set of baits is called B. The resulting degree of a node v and its clustering coefficient are consequently referred to as k_v^1 and C_v^1. The average degree of the network and the average clustering coefficient are denoted by $\bar{k}^1$ and C^1.

3.2.3 Spurious interactions

Since false positive interactions may affect both the degree distribution and the clustering coefficient, we extended the simple sampling model to include also wrong interactions. For this purpose, a second step is added after the first sampling step in which false positive interactions are simulated. False positive interactions can be added between each bait and any other node u even if no interaction of u was sampled in the first step. We add an interaction between a bait v and any other node u with a specific probability $\omega(v, u)$ and the resulting network is denoted as $G^2 = (V^2, E^2)$.

The probability $\omega(v, u)$ can be defined in different ways. In the first case, the probability of adding an edge between v and u depends neither on the degree of v or u, i.e. is constant for all pairs of nodes. Since random graphs (Erdős and Rényi (1959), see also section 2.2.1) are created in a similar way, this process is denoted as random attachment. In the second case $\omega(v, u)$ does only depend on the degree of the bait v but is constant for all its possible neighbors u. We denote this behavior as semi-preferential attachment, since new edges will be attached preferentially to baits with high degree. The last possible scenario involves preferential attachment for both v and u.

Since preferential attachment is most likely to change the degree distribution towards a power-law distribution (Barabási and Albert, 1999), our model is based on such a scenario. For this purpose, we use an adaption of the method described by Chung and Lu (2002) for creating random graphs with a given degree distribution (see section 2.3.2). Accordingly, $\omega(v, u)$ is defined as

$$\omega(v, u) = \theta \frac{(k_v + \iota)(k_u + \iota)}{\sum_{w \in V}(k_w + \iota)}. \tag{3.1}$$

Note that k_v denotes the degree of node v in the original network. Thus, the number

of wrong interactions a protein obtains depends on the number of true interactions it forms. This is based on the assumption that highly interactive proteins are more prone to spurious interactions than proteins which form only a few but very specific interactions. The parameter θ controls the false positive rate, whereas ι is used as a pseudo-count to guarantee that singular nodes, i.e. nodes with degree zero, can also obtain wrong interactions. We have that $0 \leq \iota < \infty$ and the larger ι the smaller is the influence of the actual degree values of v and u on the probability $\omega(v, u)$. For our purposes, ι was set to 1.

3.3 Results

3.3.1 Analytical results

In the following, theoretical derivations are given which describe the influence of the complete model on the clustering coefficient of networks. For simplification, we address the effect of limited sampling, i.e. missing interactions, and false positives, i.e. spurious interactions, separately from each other.

Missing interactions

In this section, we analyze the effect of limited sampling on the clustering coefficient of a node and the complete network. We show that both limited bait coverage and limited edge coverage leads to a reduction in clustering coefficients and therefore that limited sampling as a whole lowers the clustering coefficient.

The clustering coefficient (see section 2.2.2) of a node v after sampling can be formulated as:

$$\begin{aligned} C_v^1 &= \frac{P\big((u,w) \in E^1 \wedge (u,v) \in E^1 \wedge (v,w) \in E^1\big)}{P\big((u,v) \in E^1 \wedge (v,w) \in E^1\big)} =: \frac{P\big(\bigtriangledown \in E^1\big)}{P\big(\vee \in E^1\big)} \\ &= \frac{P\big(\bigtriangledown \in E^1 \big| \bigtriangledown \in E\big)}{P\big(\vee \in E^1 \big| \vee \in E\big)} \frac{P\big(\bigtriangledown \in E\big)}{P\big(\vee \in E\big)} \\ &= \frac{P\big(\bigtriangledown \in E^1 \big| \bigtriangledown \in E\big)}{P\big(\vee \in E^1 \big| \vee \in E\big)} C_v. \end{aligned} \tag{3.2}$$

Thus, the clustering coefficient of v depends on its original clustering coefficient and the probabilities $P\big(\bigtriangledown \in E^1 \big| \bigtriangledown \in E\big)$ and $P\big(\vee \in E^1 \big| \vee \in E\big)$.

To examine the full impact of sampling on the clustering coefficient of a node v, we have to differentiate between baits and preys. First, let v be a prey. In this case, two edges from v to two neighbors u and w can only be conserved if both u and w are chosen as baits. If at least one of them is not a bait, the corresponding edge to v is missed. However, if both nodes are baits, the two edges connecting them to v are each selected with probability ε (see Figure 3.2 **A**), since they connect a bait to a prey. The joint probability that both

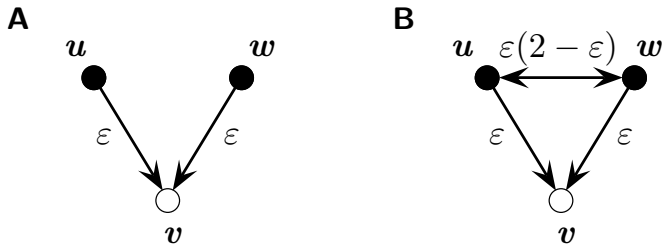


Figure 3.2: Effect of limited sampling on clustering coefficients of prey proteins. Probabilities for selecting edges in the limited sampling step if the node v considered is a prey. Here, baits are indicated by black nodes and preys by white notes. The arrows at the end of edges indicate the bait and prey relationship for this edge and edges are directed from bait to prey. Accordingly, edges between baits have arrows at both ends.

edges are kept can be expressed by the product of the individual probabilities since they are independent. As a consequence, we have that

$$P\big(\vee \in E^1 \big| \vee \in E\big) = \beta^2\varepsilon^2. \tag{3.3}$$

If u and w are connected in G, the corresponding edge again can only be selected if both nodes are baits. If this is true, the probability that this edge is conserved is then $\varepsilon(2-\varepsilon)$, since it connects two baits (see Figure 3.2 **B**). Accordingly, we have that

$$P\big(\bigtriangledown \in E^1 \big| \bigtriangledown \in E\big) = \beta^2\varepsilon(2-\varepsilon)\varepsilon^2 \tag{3.4}$$

and

$$C_v^1 = \varepsilon(2-\varepsilon)C_v \leq C_v. \tag{3.5}$$

We thus observe that the clustering coefficient of a prey is only affected by limited edge coverage. If $\varepsilon = 1$, the expected clustering coefficient after sampling is approximately the same as before sampling regardless of the value of bait coverage.

Second, let v be a bait. In this case, the edges (u, v) and (v, w) can be conserved no matter if the nodes u and w are baits or preys. If both nodes are baits (see Figure 3.3 **A**), each edge is selected with probability $\varepsilon(2-\varepsilon)$. If only one of them is a bait (Figure 3.3 **B** and **C**), one edge is selected with probability ε and the other one with probability $\varepsilon(2-\varepsilon)$. If both are preys (Figure 3.3 **D**), both edges are only selected with probability ε. Thus, we observe that

$$P\big(\vee \in E^1 \big| \vee \in E\big) = \beta^2\varepsilon^2(2-\varepsilon)^2 + 2\beta(1-\beta)\varepsilon^2(2-\varepsilon) + (1-\beta)^2\varepsilon^2. \tag{3.6}$$

On the other hand, a triangle between u, v and w can only be conserved if at least one of the two nodes u or w is also a bait. The probabilities for selecting edges (u, v) or (v, w) are in these cases the same as above. The third edge (u, w) is then selected with

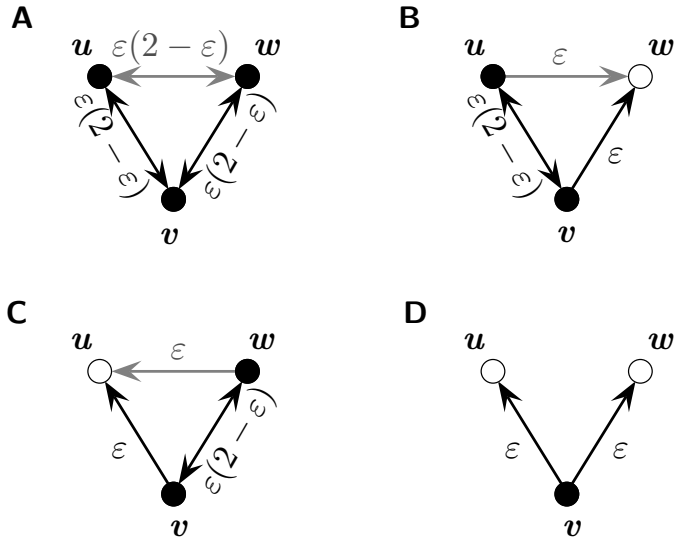


Figure 3.3: Effect of limited sampling on clustering coefficients of bait proteins. The probabilities for selecting edges are shown for the case that v is a bait. The notation is the same as in Figure 3.2. For each possible bait-prey combination of u and w, the probabilities are shown separately. The edge completing the triangle and the corresponding probabilities for selecting this edge are shown in grey.

probability $\varepsilon(2-\varepsilon)$ if both nodes are baits and with probability ε if only one of the two nodes is a bait (see also Figure 3.3). Accordingly, we have that

$$P\big(\bigtriangledown \in E^1 \big| \bigtriangledown \in E\big) = \beta^2\varepsilon^3(2-\varepsilon)^3 + 2\beta(1-\beta)\varepsilon^3(2-\varepsilon). \tag{3.7}$$

By inserting equations (3.6) and (3.7) into (3.2) we obtain that

$$C_v^1 = \varepsilon(2-\varepsilon)\lambda C_v \tag{3.8}$$

with

$$\lambda := \frac{\beta^2(2-\varepsilon)^2 + 2\beta(1-\beta)}{\beta^2(2-\varepsilon)^2 + 2\beta(1-\beta)(2-\varepsilon) + (1-\beta)^2}. \tag{3.9}$$

It is easy to see that $\lambda \leq 1$ since

$$\begin{aligned} &\beta^2(2-\varepsilon)^2 + 2\beta(1-\beta) \\ &\quad\leq \beta^2(2-\varepsilon)^2 + 2\beta(1-\beta)(2-\varepsilon) \\ &\quad\leq \beta^2(2-\varepsilon)^2 + 2\beta(1-\beta)(2-\varepsilon) + (1-\beta)^2. \end{aligned} \tag{3.10}$$

As a consequence we have that $C_v^1 \leq \varepsilon(2-\varepsilon)C_v$ and in particular that $C_v^1 < \varepsilon(2-\varepsilon)C_v$ if either $\beta < 1$ or $\varepsilon < 1$. This shows that both limited bait coverage as well as limited edge

coverage lower the clustering coefficients of baits. Since G^1 contains at least one bait, we can conclude that $C_v^1 < C_v$ if bait or edge coverage is limited.

The sampling procedure described by Han *et al.* (2005) corresponds to an experimental setting in which only a small set of proteins is chosen as baits and then subsequently screened against a much larger set of preys. This set-up is often used when due to a large genome size an exhaustive search for all possible protein pairs is infeasible (Legrain and Selig, 2000). An alternative approach consists in doing such exhaustive pairwise screens only for a subset of the proteome (see e.g. in the screen of human interactions by Rual *et al.*, 2005). We can easily reduce this scenario to the one considered here if we set G as the subgraph of the original network containing only the bait nodes and all edges between these nodes. Thus, we only need to consider the additional effect of this reduction. It can be shown that clustering coefficients of nodes selected for the screen remain approximately constant and, hence, that the average clustering coefficient of the subgraph G is approximately the same as in the original network. As a consequence, the matrix screen is reduced to a simple case of our model with $\beta = 1$ and we have that $C^1 = \varepsilon(2-\varepsilon)C$ with C the original clustering coefficient of the complete network.

Spurious interactions

In the first step discussed above, the possibility of additional spurious interactions is ignored and accordingly the probability is zero that edges which have not been part of the original network occur in the sampled network. However, since exactly this happens in the second step, we have that

$$P(\diamond \in E^2) = P(\diamond \in E^2 | \diamond \in E^1)P(\diamond \in E^1) + P(\diamond \in E^2 | \diamond \notin E^1)P(\diamond \notin E^1). \tag{3.11}$$

with $\diamond \in \{\vee, \bigtriangledown\}$.

In general, the resulting clustering coefficient is difficult to determine theoretically since C^2 cannot be given relative to C^1 as in the previous step. Therefore, we determine an approximation for the clustering coefficient only for the simple case that $\beta = 1$ and $\varepsilon = 0$, i.e. all proteins are selected as baits and none of the true edges are found. Thus, we see that

$$C_v^2 = \frac{P(\bigtriangledown \in E^2 | \bigtriangledown \in E^1) \cdot 0 + P(\bigtriangledown \in E^2 | \bigtriangledown \notin E^1) \cdot 1}{P(\vee \in E^2 | \vee \in E^1) \cdot 0 + P(\vee \in E^2 | \vee \notin E^1) \cdot 1} = \frac{P(\bigtriangledown \in E^2 | \bigtriangledown \notin E^1)}{P(\vee \in E^2 | \vee \notin E^1)}. \tag{3.12}$$

We furthermore assume that the probability that two nodes are connected is independent of the probability that any other two nodes are connected. In general, this is not the case for the preferential attachment model since the assumption holds only if all possible false positive edges are equally likely and, thus, if all nodes have approximately the same degree. Nevertheless, as we will see later, the resulting assumption is still useful for assessing the impact of false positives on clustering in networks.

Based on this assumption we have that

$$P(\vee \in E^2 | \vee \notin E^1) = \sum_{\substack{u \in V \\ u \neq v}} \sum_{\substack{w \in V \\ w \neq v,u}} \Big[P((u,v) \in E^2 | (u,v) \notin E^1) \cdot P((v,w) \in E^2 | (v,w) \notin E^1) \Big]$$
$$\approx \sum_{u,w \in V} \Big[P((u,v) \in E^2 | (u,v) \notin E^1) \cdot P((v,w) \in E^2 | (v,w) \notin E^1) \Big]. \tag{3.13}$$

$P(\bigtriangledown \in E^2 | \bigtriangledown \notin E^1)$ can be rewritten similarly. Since all nodes have been selected as baits we have for each pair u and v that $P((u,v) \in E^2 | (u,v) \notin E^1) = \omega(u,v)(2 - \omega(u,v)) \approx 2\omega(u,v)$. Hence, equations (3.12), (3.13) and (3.1) result in

$$\begin{aligned} C_v^2 &\approx \frac{\sum_{u,w \in V} 2\omega(u,v) 2\omega(v,w) 2\omega(u,w)}{\sum_{u,w \in V} 2\omega(u,v) 2\omega(v,w)} \\ &= \frac{2\theta}{\sum_{u \in V}(k_u + \iota)} \frac{\sum_{u,w \in V}(k_u + \iota)^2 (k_w + \iota)^2}{\sum_{u,w \in V}(k_u + \iota)(k_w + \iota)} \\ &= \frac{2\theta}{\sum_{u \in V}(k_u + \iota)} \frac{\sum_{u \in V}(k_u + \iota)^2 \sum_{w \in V}(k_w + \iota)^2}{\sum_{u \in V}(k_u + \iota) \sum_{w \in V}(k_w + \iota)} \\ &= 2\theta \frac{\big(\sum_{u \in V}(k_u + \iota)^2\big)^2}{\big(\sum_{u \in V}(k_u + \iota)\big)^3} := 2\theta\xi \end{aligned} \tag{3.14}$$

As a consequence, we have that $C^2 \approx 2\theta\xi$.

Note that $\sum_{u \in V}(k_u + \iota) = 2|E| + \iota|V|$ is independent of the degree distribution whereas $\sum_{u \in V}(k_u + \iota)^2$ depends strongly on it. It is minimal if all nodes have the same average degree and maximal if all edges connect only one node to itself and the remaining nodes are singular, i.e. without connections. Accordingly, for networks with approximately the same number of nodes and edges, ξ is highly correlated with the skewness (see section 2.2.1) of the degree distribution (see also Figure 3.10).

The skewness of a network also allows us to assess how strongly the independence assumption is violated. As mentioned before, this assumption is only valid if the edge probabilities are independent and this is only the case if all nodes have approximately the same degree. As a consequence, the more skewed a network, the more is this assumption violated and the more does the observed clustering coefficient deviate from the approximation. Furthermore, the observed clustering coefficients are on average smaller than the approximation. This is reasonable since ξ can become arbitrary large but the clustering coefficient is bounded from above by 1. As a consequence the minimum of 1 and $2\theta\xi$ restricts the clustering coefficients observed on average in the simple case with $\beta = 1$ and $\varepsilon = 0$.

Of course, this simple scenario is insofar unrealistic as no experimental network should contain only wrong interactions and if it did it would be useless. However, as we will see later, the effect of false positive interactions on the clustering coefficient depends strongly

on ξ also for $\beta < 1$ and $\varepsilon > 0$. In addition, the degree of randomness in clustering is also an important factor.

3.3.2 Simulation results

To illustrate the effect of our model, corresponding simulations were performed for six different types of starting networks: (Poisson) random graphs (ER) (Erdős and Rényi, 1959), exponential networks with random (EX) and high (EH) clustering coefficients, power-law networks with random clustering coefficients (PL) and networks generated by the growth model described in section 2.4.2 which simulates the evolution of protein interaction networks and creates highly skewed networks similar to power-law networks. In the latter case, networks were generated with low (GL) and high (GH) clustering coefficients. High clustering coefficients in the GH networks were obtained by connecting duplicated notes with a probability p between 0.2 and 0.3. For the GL networks, duplicated nodes are not connected. As a consequence, these networks are clustered less than randomly since interactions between neighbors of a node which were created by duplication are avoided.

Exponential networks (EX and EH) and power-law (PL) networks were generated using the method by Chung and Lu (2002) described in section 2.3.2. Degree sequences for exponential networks were drawn from an exponential distribution $P(k) = \lambda e^{-\lambda k}$ for a constant λ. For PL networks, degree sequences were drawn from a power-law with exponential cut-off (see section 2.2.1) as described by Newman *et al.* (2001) using a combination of the transformation and rejection methods.

High clustering coefficients in exponential networks (EH) were created by introducing triangles into the network. Triangles were created by iteratively choosing a random node and connecting two random neighbors of this node. This is repeated until approximately the same clustering coefficients are obtained as for the GH networks. For small degree values, this leads to a deviation from the exponential distribution, but for the clustering coefficients considered, the deviation was only minor.

To simulate the effect of the Y2H methodology on the yeast interaction network, we generated networks for the described topologies, each containing 6,000 nodes (the approximate number of protein-encoding genes in yeast (Goffeau *et al.*, 1996)) and average degree values of 5, 10 and 20. For each combination of network topology and average degree, 50 networks were generated and simulation results were averaged over these 50 networks.

Analysis of simulated clustering coefficients

The observed clustering coefficients for the generated networks (Figure 3.4) vary greatly between network topologies and average degree values. With the exception of the EH and GH networks which have been created specifically to show high clustering, two dependencies can be observed. The clustering coefficients are highly correlated with the average degree of the networks but also with the asymmetry of the degree distribution. Since hubs have a large number of interactions, there is a higher probability that two interacting proteins interact with the same hub protein even if connections a randomly formed and as a

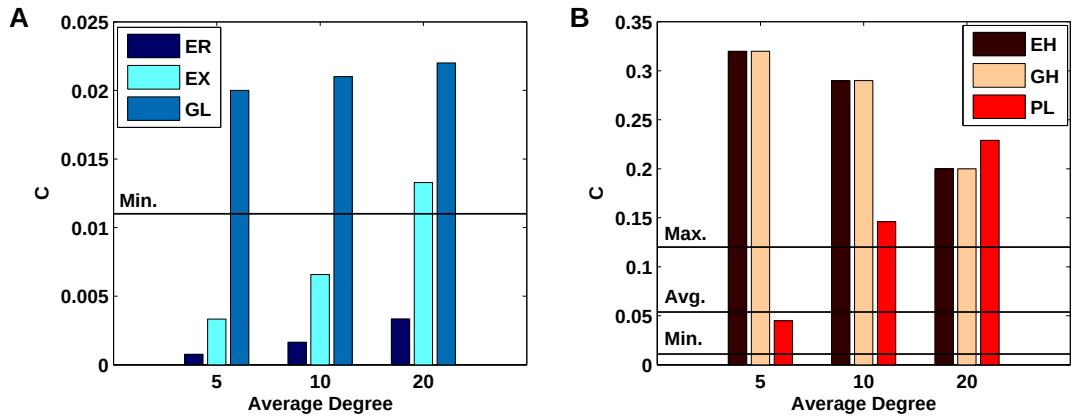


Figure 3.4: For all six network topologies and average degree values, 50 networks were generated and clustering coefficients averaged over those 50 networks. Clustering coefficients of the ER, EX and GL (**A**, from left to right) and the EH, GH and PL (**B**) networks are compared against the minimum, average and maximum clustering coefficient (horizontal lines) observed in experimental Y2H networks. For the EH and GH networks parameters were set such that both networks have approximately the same clustering coefficients.

consequence clustering coefficients are increased in highly skewed networks.

Since random ER graphs follow a Poisson distribution and thus are little skewed, they exhibit the lowest degree of clustering of any of the topologies. Compared to that, exponential networks have an increased tendency for high and low degree nodes. As a consequence, they tend to be higher clustered than the ER networks. Despite the fact, that the GL networks have lower clustering coefficients than expected randomly for the degree distribution, they still show higher clustering coefficients than the exponential networks due to their high degree of skewness. Even if the clustering coefficients of the GL networks are randomized by edge rewiring, they are still lower than the clustering coefficients of the highly skewed power-law networks (PL).

When comparing the clustering coefficients of the large-scale Y2H networks against the simulated topologies, we observe that all of the PPI networks show higher clustering than the random ER graphs. In general, the experimental networks have higher clustering coefficients than the exponential networks and even the GL networks. Only the human interaction network by Stelzl *et al.* (2005) is less clustered than all GL networks and exponential networks with high average degree values. Accordingly, only the EH and GH networks and the PL networks with high average degrees exhibit clustering coefficients which exceed those of all experimental Y2H protein-protein interaction networks.

These results as such do not exclude any of the topologies. However, when considering the effect of the different types of measurement errors on clustering coefficients, one should always keep in mind the original clustering coefficients we are starting from. In the following, the different effects of false negative and false positive interactions are again

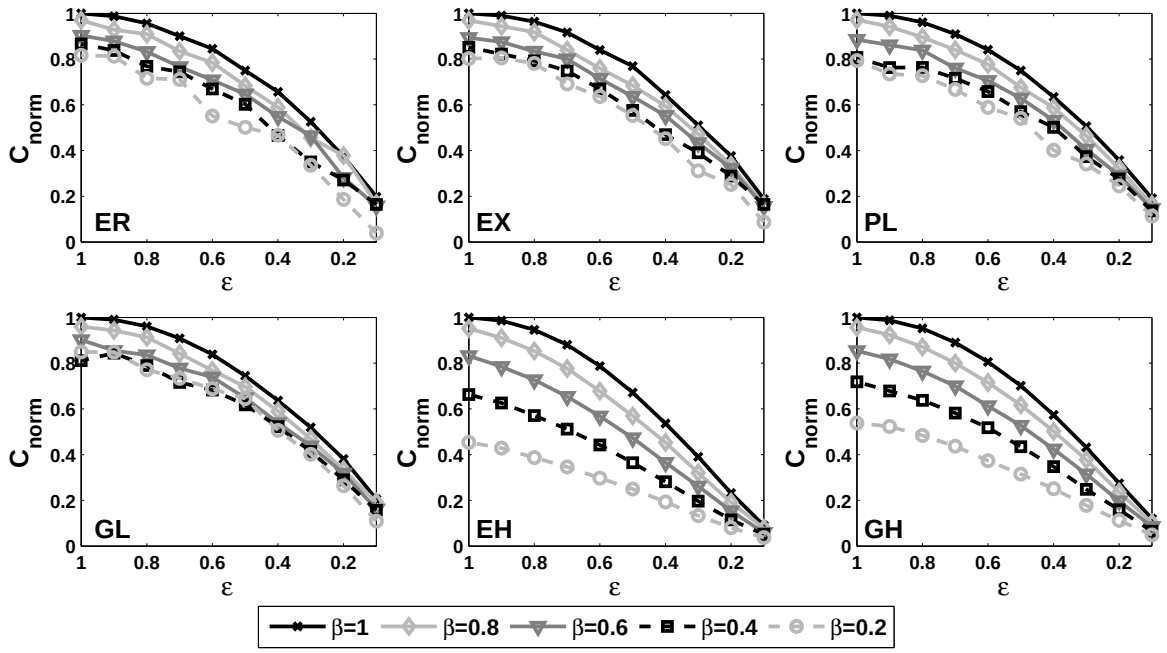


Figure 3.5: Impact of limited sampling on the average clustering coefficient for coverage rates below one. Results are shown for networks with average degree of 10, but are similar for all average degree values. The top row shows results for the randomly clustered ER, EX and PL networks and the bottom row for the less than randomly clustered GL networks and the highly clustered EH and GH networks. Highly clustered networks are affected to a much greater degree by low bait coverage rates than the randomly clustered or less than randomly clustered networks. Clustering coefficients were normalized by dividing by the original clustering coefficients.

considered separately from each other.

Missing interactions

Our theoretical results predict that both limited bait coverage and limited edge coverage lower clustering coefficients significantly regardless of network topology and average degree. Our simulations (see Figure 3.5) show that this prediction indeed holds. To illustrate the dramatic decrease in clustering due to false negatives, clustering coefficients of the sampled networks were normalized by dividing by the original clustering coefficients.

This shows that for all topology models the clustering coefficients of the sampled network are significantly lower than the clustering coefficients of the original networks for any value of bait or edge coverage. Furthermore, we can draw conclusions about similarities between the randomly or less than randomly clustered ER, EX, PL and GL networks on the one hand and the more than randomly clustered EH and GH networks on the other

hand. For all network topologies the effect of limited bait coverage is less severe than the effect of limited edge coverage. Yet, whereas limited bait coverage affects clustering in the ER, EX, PL and GL networks only to a minor degree, the effect on the highly clustered EH and PH networks is substantial (see Figure 3.5). Even at $\varepsilon = 1$, clustering coefficients in the EH and GH networks are significantly smaller for small values of β than in the other networks. For instance, at $\beta = 0.2$ they are only about half as high as in the original networks.

This observation is surprising since in our analytical derivations no such difference was observed. Nevertheless, it can be easily explained. In our derivations clustering coefficients were treated as continuous variables, whereas effectively they behave in a discrete manner since an edge can either exist or not. If we only consider nodes for which clustering coefficients before and after the simulation are greater than 0, no differences between highly and randomly clustered networks are observed. The differences observed are due to nodes for which $C_v > 0$ and $C_v^1 = 0$ and nodes for which $C_v = 0$ and $C_v^1 = 0$. In the first case, clustering coefficients decrease dramatically and stronger than expected, in the second case they do not decrease at all. In highly clustered networks the first type of nodes is much more common than in randomly clustered networks, whereas the second type of nodes is rarer. Accordingly, while for randomly clustered networks the effects on the two types of nodes probably cancel out each other to a large degree, there is an excess of the first type of nodes in highly clustered networks. This leads to the stronger reduction in clustering coefficients observed.

Spurious interactions

We have seen previously, that for $\beta = 1$ and $\varepsilon = 0$ the average clustering coefficient is expected to increase linearly with θ which is also confirmed in part by our simulations (Figure 3.6). However, for high values of θ, a deviation from the linear behavior can be observed which leads to a slower increase. As mentioned before, this is due to the violation of the independence assumption with increasing skewness. Thus, the higher the skewness in the network, the higher the deviation from the linear behavior. In Figure 3.6, topology models are sorted according to skewness. Accordingly, we observe that the more skewed a topology is, the smaller are the values of θ at which the observed clustering coefficients start to deviate from the linear behavior. This effect is most pronounced for the power-law networks, for which ξ predicts the highest increase in clustering due to false positive interactions. The effective increase turns out to be significantly less than predicted but is still much higher than for the other topologies, in particular for small values of θ.

So far, edge coverage was restricted to 0. Figure 3.7 illustrates the effect of different values of ε (but constant $\beta = 1$) and increasing θ on the clustering coefficient. For $\varepsilon > 0$, the effect on the clustering coefficient depends strongly not on the topology but on the degree of randomness in clustering. For two of the randomly clustered networks (ER and EX), the clustering coefficients increase linearly with θ for any ε. Indeed, if C is the original clustering coefficient, the resulting clustering coefficient C^2 can be approximated by $2\xi\theta + \varepsilon(2-\varepsilon)C$. As before, the PL networks deviate from this behavior and clustering

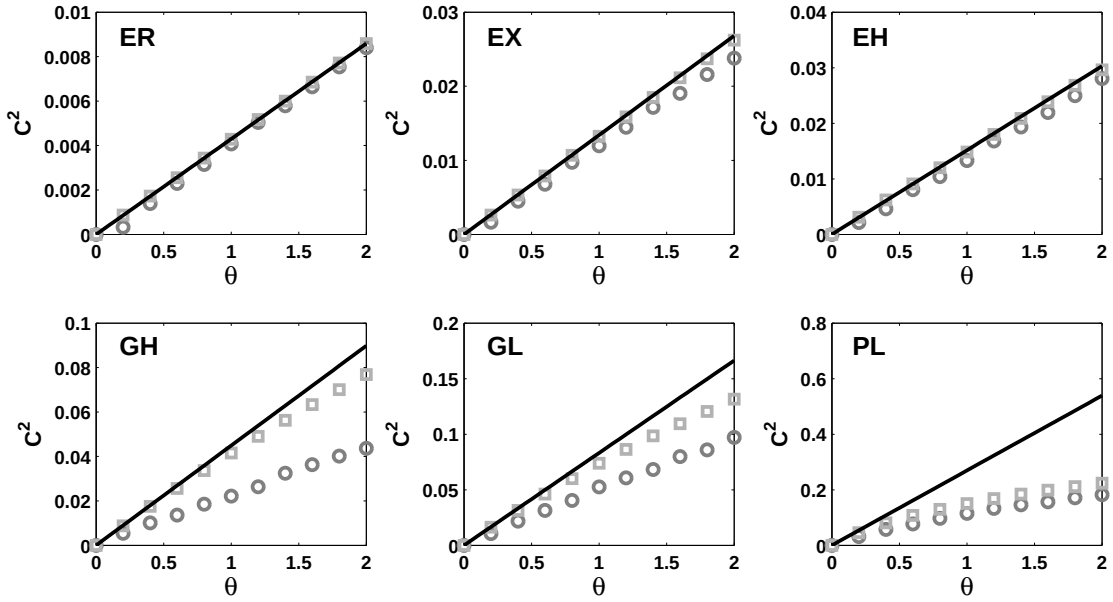


Figure 3.6: Observed and approximated effect of false positive interactions on the clustering coefficients. The average (dark grey circles) and maximum (light grey rectangles) clustering coefficients obtained in 50 simulations of false positive interactions with $\beta = 1$ and $\varepsilon = 0$ are compared against the clustering coefficients predicted by the approximation $2\theta\xi$ (black line) for networks of average degree 10. Topology models are sorted according to increasing skewness from top left to bottom right to illustrate the increasing deviation from the approximation with network skewness.

coefficients for higher values of ε increase more slowly than predicted. Furthermore, the higher ε, the lower is the rate of increase. As a consequence, the curves for $\varepsilon = 0$ and $\varepsilon = 1$ move towards each other for increasing θ.

For networks clustered less than randomly (GL), the average clustering coefficients for higher values of ε increase stronger than linearly at the beginning, until random clustering is reached in the network. From this point on a similar behavior is observed as for the randomly clustered networks. The contrary effect is found for highly clustered networks (EH and GH). In this case the clustering coefficients are reduced significantly by preferential attachment of false positives. Only when random clustering is reached in the network, clustering coefficients increase again depending on the value of ξ and thus on the asymmetry in the network. Nevertheless, the decrease in clustering due to missing interactions and false positives in these highly clustered networks can only be compensated for by very high error rates.

For $\beta < 1$, the effect of erroneous interactions on the average clustering coefficient is similar to the case in which all proteins are selected as baits (see Figure 3.8). Clustering

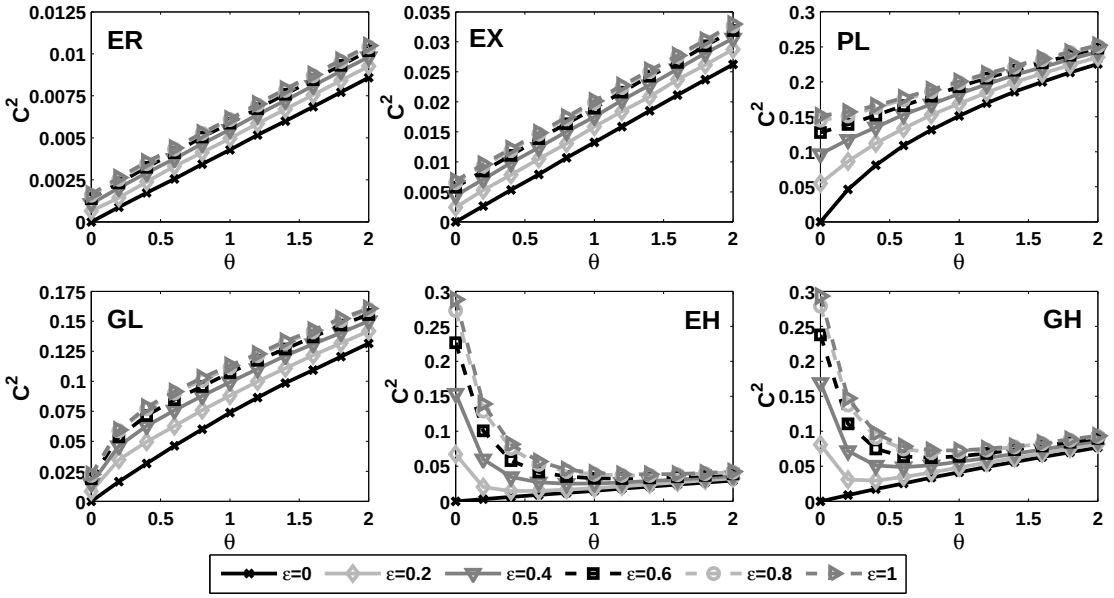


Figure 3.7: Spurious interactions can influence average clustering coefficients in two ways depending on the degree of clustering in the network. In randomly (ER, EX, PL) or less than randomly (GL) clustered networks, clustering coefficients can be increased by attaching false positive interactions. In highly clustered networks (EH and GH), clustering coefficients are – at least for reasonable error rates – decreased. Results are shown for average degree values of 10 and bait coverage $\beta = 1$.

coefficients can be increased as well for randomly clustered networks, but the increase turns out to be slightly less than before. A possible explanation for this observation might be that wrong interactions are only ever added between baits and preys (or other baits) but never between preys. Thus, for small values of β, baits are often connected to two preys which by definition of the model can never be connected. This results in smaller clustering in the network.

To illustrate the combined effect of different parameter values for the model, simulations were performed in which for each value of β and ε, θ was chosen such that the same fixed false positive rate of 50% was obtained (see Figure 3.9). In these simulations several observations could be made. First, of course, clustering coefficients tend to be highest for high values of edge coverage and decrease with edge coverage. Second, for the ER, EX and GL networks the clustering coefficients obtained are higher than the clustering coefficients in the original simulated networks even for small edge coverage rates, whereas for the PL networks this requires higher edge coverage. On the contrary, in the EH and GH networks the resulting clustering coefficients are always significantly smaller than the original clustering coefficients for the given false positive rate. Here, only extremely high values of θ

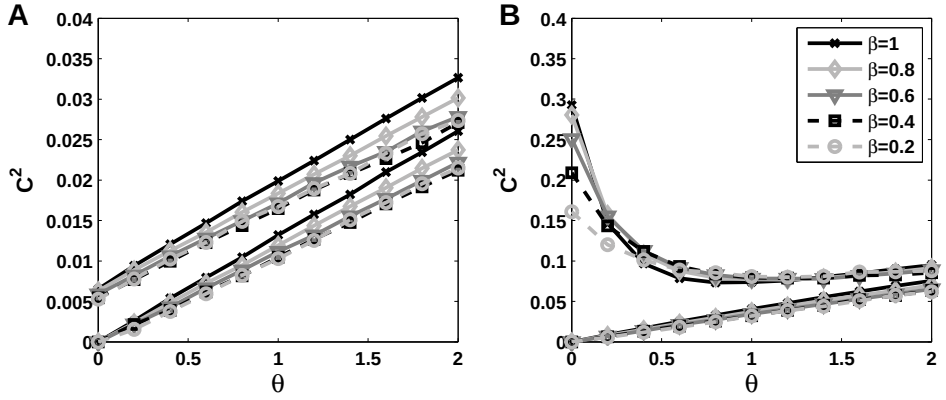


Figure 3.8: Impact of false positive interactions on clustering coefficients at different values of bait coverage. Results observed for values of bait coverage smaller than 1 and for edge coverage rates of 1 (upper curves) and 0 (lower curves). Results are shown for the EX (**A**) and GH (**B**) networks of average degree 10 but are similar for all topology models and average degree values. We observe that clustering coefficients for smaller values of β are, generally, slightly smaller then for $\beta = 1$, but the differences are minor.

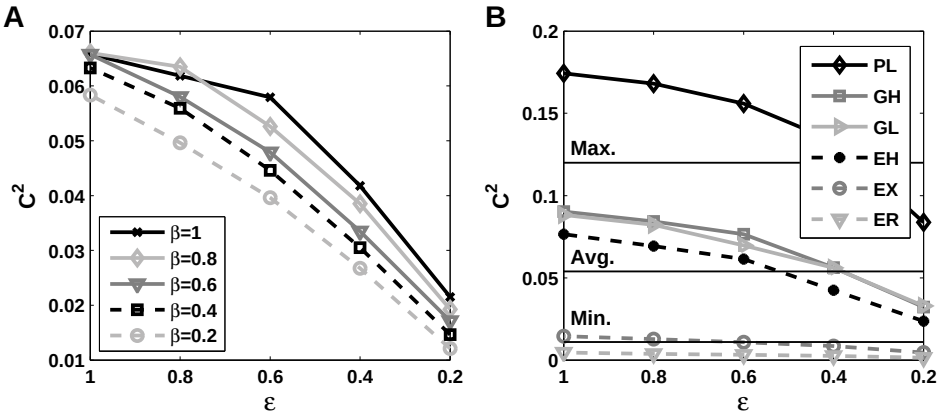


Figure 3.9: Resulting clustering coefficients at a fixed false positive rate. The combined effect of the different error mechanisms were analyzed by setting the false positive rate at a fixed value of 50% for networks of average degree 10. For each combination of β and ε, θ was then chosen accordingly. **A**) Resulting clustering coefficients at different bait coverage rates for the EH networks. Results for other topologies are similar but differences are less pronounced between different values of β. **B**) Maximum average clustering coefficient obtained for the different values of β considered for each topology. Minimum, average and maximum clustering coefficients observed in the real Y2H experiments are indicated by horizontal lines.

and, consequently, high false positive rates could increase clustering coefficients beyond the original value. In both cases, this is due to the different effects of false positive interactions on randomly and highly clustered networks. Furthermore, clustering coefficients tend to be similar for different values of β. The ER and EX networks show only minor differences, whereas stronger differences can be observed for the other network types. In this case, the differences are most pronounced for the highly clustered EH networks.

In order to compare the effects on clustering at 50% false positive rate between topology models, we computed for each topology the maximum over the averages for different values of β (see Figure 3.9 **B**). As can be seen, even by introducing false positive interactions clustering coefficients in ER and EX networks cannot be increased sufficiently to explain at least most of the observed Y2H networks by such a topology. The only topologies for which realistic clustering coefficients are observed are thus highly clustered exponential networks, the growth models and the power-law networks. Although EH networks were created with approximately the same clustering coefficients as the GH networks, the final clustering coefficients observed for these networks are nevertheless smaller than for the GH

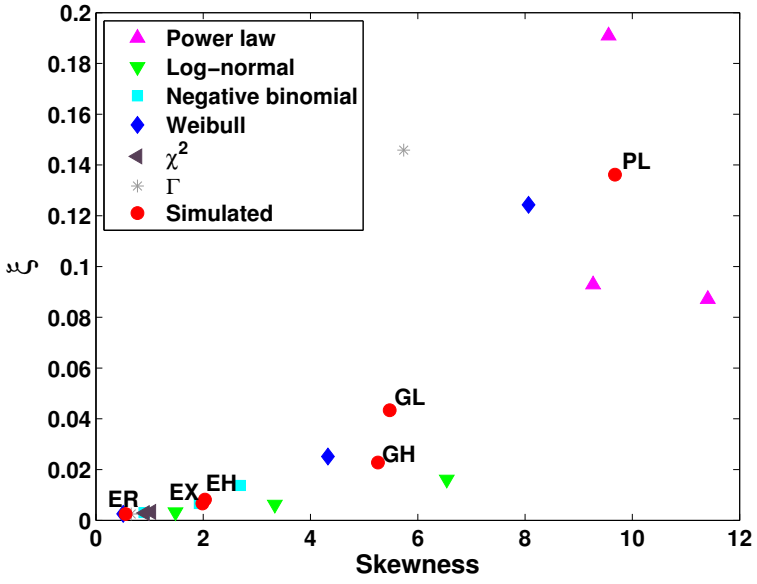


Figure 3.10: Correlation between the effect of false positive interactions and skewness. Values of ξ were computed for a several topology models additional to the ones for which full simulations of our model were performed and plotted against the skewness of the corresponding networks. Topology models considered include power-law, log-normal, negative binomial, Weibull, χ^2 and Γ distributions with varying parameters. Parameters were tuned such that average degree values of 10 were obtained and results were averaged over 50 networks generated for each topology. The topology models for which complete simulations were performed are indicated in red.

network. This can be explained by the fact that the increase in clustering for high θ as well as the lowest level up to which clustering coefficients decrease for smaller θ depend strongly on the skewness of the network topology which is higher for the GH networks than the EH networks.

Although we considered several possible topology models, there is an infinite number of possible topologies for which we did not perform simulations of our model. Nevertheless, the results presented above can be transfered to other topologies by taking into account the skewness of these models. If networks are clustered randomly, the clustering coefficients observed depend on the skewness of the corresponding degree distribution. Thus, highly skewed networks have high random clustering coefficients whereas slightly skewed or symmetric distributions exhibit very small clustering coefficients which are, in particular, smaller than clustering coefficients observed in real Y2H interaction networks. We have shown that missing interactions decrease these clustering coefficients even further. Only false positive interactions can increase clustering again in randomly clustered networks depending on the network topology. In networks clustered higher than randomly, clustering coefficients are decreased even by false positive interactions.

Our simulation suggest that ξ, although not a perfect approximation at least restricts from above the clustering coefficients observed in randomly clustered networks under the influence of false positive interactions. The higher ξ, the higher the increase in clustering due to false positive interactions in randomly clustered networks, although returns diminish with increasing ξ. Accordingly, we computed the values of ξ for a range of additional topologies and plotted them against the skewness of the corresponding networks (see Figure 3.10). For each topology model we generated 50 networks and averaged over the corresponding ξ and skewness values. As can be seen, ξ is highly correlated to the skewness of network models. Accordingly, false positive interactions can only increase the clustering coefficients of networks sufficiently which follow a highly skewed degree distribution.

3.4 Discussion

In their article, Han *et al.* (2005) raised the possibility that the apparent scale-free topology of experimental Y2H interaction networks is due to distorting effects of limited sampling in large-scale experiments and that by examining the degree distribution alone, the topology of the experimental interaction networks cannot be safely extrapolated to the complete interactome. In this context, our results indicate that based on additional topological characteristics such as the clustering coefficient, the range of possible topologies can be narrowed. Thus, although current large-scale PPI networks represent only a fraction of the interactomes, they can nevertheless be used to draw some inferences to the topological characteristics of the complete interactomes.

We have shown both analytically and in simulations that sampling with limited bait and edge coverage lowers the clustering coefficient tremendously for any of the examined network topologies. This result has several implications concerning the topology of the complete interactomes. In this setting, the clustering coefficients observed in protein-

protein interaction maps derived with high-throughput methods provide a lower bound on the clustering coefficients observed in complete interactomes. This furthermore suggests that the interactomes are highly clustered, much more than the simple random graph (ER), exponential (EX) or growth networks (GL). Accordingly, such topologies can be ruled out if the effect of spurious interactions is ignored. These findings do not eliminate the possibility that the original networks show a highly clustered topology different from a power-law topology.

Notwithstanding these considerations, we can use the relationship between clustering coefficients and bait and edge coverage to estimate the amount of error involved if we know both the original and resulting clustering coefficient and vice versa assess the original clustering coefficient based on the error rate and the observed clustering coefficient. In our simulations we found that in order to increase skewness in a network by limited sampling and thus to change the original distribution towards a power-law topology, bait coverage rates have to be lowered considerably. The degree to which they have to be lowered depends on the difference of the original topology to a power-law topology. Lowering edge coverage rates, on the other hand, does not have a sufficiently distorting effect. However, we have seen above that limited bait coverage leads to a significant reduction in clustering coefficients in highly clustered networks such as EH and GH. Thus, high original clustering coefficients would have to be assumed for the interactome if the observed interaction networks were sampled from a highly clustered distribution which is significantly different from a power-law distribution (e.g. an exponential distribution). If such a high degree of clustering appears unreasonable, the obvious conclusion is that the original interactome does in fact exhibit a power-law or a similar highly skewed topology.

We extended the sampling procedure described by Han *et al.* to cover the influence of false positive interactions on the topology of sampled networks. This leads to a more realistic model of Y2H experiments since spurious interactions are observed regularly in large-scale experiments. Without considering false positives the effect of sampling on the topology and the clustering coefficient might be underestimated or misinterpreted. In our model, interactions are introduced by a preferential attachment scenario in which the probability of obtaining wrong interactions depends on the degree of the nodes participating in an interaction. Furthermore, baits are more likely to acquire interactions than preys. This introduces a possible source of degree asymmetry in the model which is a consequence of the experimental set-up and not the topology of the network.

Based on the extended model, the conclusions drawn from the simple sampling model can be generalized. Preferential attachment of false positive interactions increases the clustering coefficient of networks which are clustered randomly (ER, EX and PL) or less than randomly (GL) but decreases the clustering coefficient for networks which are clustered higher than randomly (EH and GH) except for extremely high error rates. As the rate of increase in randomly and less than randomly clustered networks is only high for highly skewed networks, random graph and randomly clustered exponential networks still can be excluded confidently since unreasonably high error rates would have to be assumed to explain the clustering coefficients observed. Contrary to that, clustering coefficients of GL and even more so of PL networks can be increased sufficiently by introducing wrong inter-

actions to explain at least most of the observed clustering coefficients. Indeed at 50% false positive rate, similar clustering coefficients can be obtained for the GL networks as for the highly clustered exponential (EH) and growth networks (GH) whose clustering coefficients are decreased by wrong interactions.

Accordingly, our simulation results suggest that the interactome either follows a power-law or similarly skewed degree distribution or is highly clustered. Nevertheless, we can make the same argument as before, that changing e.g. an exponential towards a power-law topology requires small bait coverage rates and consequently high clustering coefficients in the original network.

For random and semi-preferential attachment, estimates for the expected increase in clustering for $\beta = 1$ and $\varepsilon = 0$ can be derived in the same way as for preferential attachment. However, the rate of increase is smaller for both random and semi-preferential attachment than for preferential attachment. Accordingly, at $\varepsilon > 0$, clustering coefficients can decrease even for randomly clustered networks. In the semi-preferential model, this is only the case for highly skewed networks such as the PL networks. In the random attachment scenario, this happens even for the slightly skewed exponential networks.

Simulations of false negative and positive interactions were only performed for networks with average degree values of 5, 10 and 20. Higher average degree values in the original networks lead to higher random clustering coefficients in the original networks and thus in the sampled networks. Hence, one might argue that the above conclusions are invalid if original average degrees only have to be increased sufficiently. However, such considerations are limited by what is actually observed in experimental networks. This can be illustrated by the following example. Suppose, a matrix Y2H screen ($\beta = 1$) results in a network with average degree $\bar{k}'$ of 5 and the false positive rate is estimated to be 50%. Then, edge coverage and original average degree $\bar{k}$ are related by the formula

$$\varepsilon(2 - \varepsilon) = \frac{\bar{k}'}{2\bar{k}}. \tag{3.15}$$

Thus, if $\bar{k} = 2.5$, ε is approximately 1. For $\bar{k} = 5$ it is 0.29, for $\bar{k} = 10$ it is 0.13, and so on. Accordingly, high average degree values can only be assumed if coverage rates are small. This on the other hand implies that although original clustering coefficients might be higher, the clustering coefficients resulting from the experiment are very small due to the low coverage rates.

The error mechanisms we proposed for our model are fairly simple and require few assumptions. Of course, many other error mechanisms are also possible (see e.g. Lin and Zhao, 2005) and we can never be sure that the way interactions are added describes the processes occurring in large-scale experiments accurately. As a consequence, the preferential attachment scenario was chosen to simulate the worst case in which false positive interactions also promote a scale-free topology in experimental networks regardless of the original topology. We showed that, even when assuming this worst case, conclusions can still be drawn to the topology of the interactome. Nevertheless, our results do not only apply to our model but can be generalized to a wider range of error mechanisms. Randomly removing edges from a network in general reduces clustering coefficients in this network.

On the other hand, adding edges to a network increases clustering only if the probability that triangles are created is at least as high as the probability that triangles exist in the original network. Random error processes, however, create most likely also random clustering coefficients. Accordingly, if the original networks are clustered higher than randomly, clustering coefficients are expected to decrease.

3.5 Conclusions

We conclude that measurement errors in large-scale experiments affect several aspects of the network topology apart from the degree distribution. The impact of the experimental set-up on these other characteristics may be used to infer the topology of the complete interactome. Here, we focused on the average clustering coefficient to evaluate the likelihood of different topological models for the interactome. Our analytical and simulation results indicate that some of the suggested topologies are highly unlikely and can be excluded with high confidence. Although only a selection of possible topology models was discussed, we have shown how the results can be transfered to other topologies as well. The most effective and most conclusive way to completely resolve the topology of the interactome would be to increase the coverage of the interactome by both many more experiments and by improving the false positive and false negative rates of large-scale methods. However, until this is achieved, we have shown that useful conclusions can still be drawn by modeling sampling effects.

Chapter 4

Degree correlations and network structure and stability

4.1 Introduction

In the previous chapter, we analyzed the effects of measurement errors on the degree distribution of protein-protein interaction (PPI) networks and showed that conclusions can be drawn from partial networks to the complete interactome despite many measurement errors. Although the degree distribution is the most commonly analyzed network characteristic, it does not characterize a network completely as the same number of connections can be formed in several ways without changing the degree distribution. For instance, low-degree nodes might associate preferentially either with other low-degree nodes or with hubs. Thus, two networks can have the same degree distribution and still differ in other aspects of network structure and react differently to perturbations.

Correlations between degree values of neighboring nodes were analyzed by Maslov and Sneppen (2002a,b) for the yeast interaction network determined by Ito *et al.* (2001). Maslov and Sneppen found that interactions between hubs were significantly suppressed in this network relative to a "null model" and that hubs were preferentially associated with low-degree nodes instead. Contrary to that, more recent studies (Coulomb *et al.*, 2005; Batada *et al.*, 2006) showed no such negative correlation between node degrees in yeast for high-confidence interaction sets. Aloy and Russell (2002) suggested that these contrasting results may be explained by a bias in the yeast-two hybrid system which artificially increases negative degree correlations.

To resolve the question whether interactions between hubs are suppressed or not, we performed a systematic analysis of experimentally derived PPI networks (Friedel and Zimmer, 2007). For this purpose, we compared these networks against reference networks showing no degree correlations (the "null model" of Maslov and Sneppen). Additionally, a model was developed to create reference networks with positive and negative correlations between the degrees of neighboring nodes, respectively. Based on simulations, we then evaluated how positive or negative degree correlations affect network structure and

tolerance of the networks to targeted deletions.

We found that negative degree correlations lead to less fragmentation of the original networks into connected components compared to positively correlated networks. On the other hand, such negative correlations increase the vulnerability of the corresponding networks to targeted deletions of hubs. This results in a higher rate of interaction loss and an increased disintegration rate in response to targeted deletions in negatively correlated networks. Thus, for any degree distribution, vulnerability to targeted attack on hubs can be increased by introducing negative correlations. On the other hand, positive degree correlations can decrease this vulnerability. Thus, the low attack tolerance of power-law networks (see section 2.2.5) can be both decreased and increased by modifying degree correlations in these networks.

Our results show that experimentally derived PPI networks, in particular Y2H networks, tend to be most similar to the "null model" networks. This suggests that there is no systematic bias towards negative degree correlations in Y2H experiments. Furthermore, although significant modifications in network structure are possible without changing the degree distribution, only a very small range of modifications is actually realized in PPI networks. Our results suggest that the mostly uncorrelated network structure of PPI networks might be a consequence of different selective disadvantages of both negatively and positively correlated networks.

4.2 Methods

4.2.1 Reference networks

Experimental networks were compared against three types of reference networks which were created by rearranging the connections in the network such that each node has the same degree as before. Uncorrelated reference networks (the "null model") were generated using the rewiring method by Maslov and Sneppen (2002b) (see section 2.3.1). These networks exhibit only random degree correlations given the degree distribution. Furthermore, we developed an algorithm to generate positively and negatively correlated reference networks from a starting network.

For both types of correlated reference networks, we start with a network containing only the nodes but not the edges of the original network. First, each node is assigned its degree value in the original network. We then choose iteratively the node v with the highest assigned degree whose current degree is lower than this assigned degree. Edges are added to this node until its degree value matches its original degree value. To add an edge a random node u is chosen from the remaining nodes with probability $P(u) \sim k_u^\tau$ for a fixed parameter τ. If the current degree of u is less than its assigned degree k_u, the interaction between v and u is added to the network.

To create negatively correlated reference networks τ was set to 0. As a consequence, each node is chosen with equal probability. Since low-degree nodes are most abundant in PPI networks, hubs will then be connected preferentially to low-degree nodes.

For positively correlated references τ was set to 3. As a consequence, high-degree nodes are chosen with higher probability and connections between hubs are increased.

4.2.2 Evaluation of degree correlations

To quantify degree correlations in a network we use the Pearson correlation coefficient r between degree values of connected nodes calculated over all edges in the network. Here, undirected edges are treated as two directed edges. Let $(e_1, \dots e_m)$ be the vector of all edges in the set of undirected edges and k_v the degree of a node v. Then we set x and y as two vectors of length $2m$ with $x_{2i-1} = k_u$, $x_{2i} = k_v$, $y_{2i-1} = k_v$ and $y_{2i} = k_u$ for $e_i = (u, v)$. The correlation coefficient r is then defined as

$$r = \frac{\sum_{i=1}^{2m}(x_i - \bar{x})(y_i - \bar{y})}{(2m-1)s_x s_y} \tag{4.1}$$

with $\bar{x}$ and $\bar{y}$ the sample means and s_x and s_y the sample standard deviations of x and y. Positive values of r indicate a positive correlation and negative values a negative correlation between the degrees of associated nodes. If the degree distribution is positively skewed, such as e.g. in power-law networks, we observe negative correlation coefficients even for the uncorrelated reference networks.

4.2.3 Simulation of measurement errors

To simulate the effects of measurement errors on degree correlations and network stability, 10% of the edges were removed randomly and replaced by the same number of edges. Here, four strategies were applied. First, 10% of the edges were rewired as described for the "null model". Second and third, edges were added using the method for creating either negatively or positively correlated networks after removing only 10% and not all edges from the starting network. And fourth, the preferential attachment method described in section 3.2.3 was used to add false positive interactions to the network after deleting 10% of edges. Here, the error rate θ is tuned such that approximately 10% wrong interactions are added.

4.2.4 Targeted deletion of nodes

Targeted deletion of hubs is simulated by iteratively deleting the node with the currently highest degree from the network (see section 2.2.5). For our purposes, nodes were not deleted with decreasing order of their degree in the original network as described by Albert *et al.* (2000). Instead, we recalculate the node degrees at each step and then delete the node with the currently highest degree. This different approach is used to avoid an artificial advantage for positively correlated networks. Since hubs are preferentially connected to other hubs in these networks, the deletion of some of the hubs will preferentially decrease the degree of other hubs. Consequently, some of the nodes which were hubs in the original

network might no longer be hubs after a few node deletion steps. Therefore, it is more appropriate to delete the nodes which then have the highest degree.

4.2.5 Analysis of network properties

For all networks considered and the corresponding reference networks, we analyzed the number of connected components and characteristic path length L between all proteins and between hubs only. A connected component is a maximal connected subgraph of the network. This means that a path exists between each pair of nodes in the same component and no path exists between any pair of nodes in different components.

Characteristic path length L for the complete networks was calculated as described in section 2.2.3. Furthermore, we specifically calculated characteristic path length between hubs as the average shortest paths between hubs. These shortest paths can also pass non-hub nodes. For this purpose, hubs are defined as the top 10% of nodes with the highest degree.

To analyze structural stability under targeted deletion of nodes, we recorded the development of network characteristics with progressive hub removal. The network characteristics considered are characteristic path length, efficiency (Latora and Marchiori, 2001), diameter (length of the longest path in the network), fraction of protein pairs connected by a path, fraction of edges remaining in the network after deletion (FER), size of the largest connected component, average component size and number of connected components. Since most of these network characteristics behave similarly as either characteristic path length or the fraction of edges remaining (FER) in the network, we focused on these two characteristics. Furthermore, we analyzed the number of connected components remaining (NCR) after deletion which shows a unique behavior.

4.3 Results

4.3.1 Protein-protein interaction networks

The following protein-protein interaction networks were analyzed: (i) networks from large-scale yeast two-hybrid (Y2H) experiments, (ii) networks extracted from the Database of Interacting Proteins (DIP) (Xenarios *et al.*, 2002) (version of April 2nd, 2006) and (iii) the yeast high-confidence interaction set compiled by Batada *et al.* (2006). In the first case, we used results from the large-scale Y2H studies of Ito *et al.* (2001) and Uetz *et al.* (2000) for yeast, Giot *et al.* (2003) for *Drosophila*, Li *et al.* (2004) for *C. elegans*, Rual *et al.* (2005) and Stelzl *et al.* (2005) for human and LaCount *et al.* (2005) for *P. falciparum*. From DIP we used the species-specific data sets for yeast, *Drosophila*, human and *E. coli* (see Table 4.1 for network characteristics).

For the Li and Giot networks, only high-confidence interactions were considered. For the Ito networks, both the high-confidence interaction set and the complete interaction set were analyzed separately to compare our results against the ones of Maslov and Sneppen.

Network	$\|V\|$	$\bar{k}$	L	# conn. comps.	r	FER
Yeast						
Ito *et al.* (2001) (complete)	3279	2.68	4.88	195	-0.176	0.22
Ito *et al.* (2001) (core)	797	1.89	6.14	143	-0.112	0.35
Uetz *et al.* (2000)	1005	1.80	7.49	177	-0.088	0.40
Batada *et al.* (2006)	2998	6.18	4.90	101	-0.047	0.37
DIP	4959	6.95	4.15	31	-0.133	0.26
Human						
Rual *et al.* (2005)	1549	3.37	4.36	118	-0.198	0.18
Stelzl *et al.* (2005)	1705	3.70	4.85	44	-0.191	0.16
DIP	1085	2.48	6.77	126	-0.004	0.36
Drosophila						
Giot *et al.* (2003) (core)	4651	2.01	9.43	591	0.023	0.46
DIP	7451	6.08	4.39	62	-0.081	0.24
Other species						
Li *et al.* (2004) (*C. elegans*, core)	1415	2.94	4.91	70	-0.176	0.19
E. coli (DIP)	1840	6.44	3.80	346	-0.086	0.12
LaCount *et al.* (2005) (*P. falciparum*)	1308	4.20	4.26	23	-0.025	0.31

Table 4.1: Network characteristics for experimental protein-protein interaction networks. Shown are the number of nodes $|V|$, the average degree $\bar{k}$, the characteristic path length L, the number of connected components, the correlation coefficient r and the fraction of edges remaining (FER) after deleting the 10% nodes with the highest degree.

Interactions from different experiments for the same organism were analyzed separately to compare them against each other. For each PPI network, 100 uncorrelated, positively and negatively correlated reference networks were generated, respectively, and results were averaged over the 100 individual simulation runs.

4.3.2 Degree correlations in PPI networks

The original PPI networks were compared against the three reference networks for the correlation coefficient r between the degrees of connected nodes. This comparison showed that the original networks tend to have correlation coefficients similar to or slightly smaller than the uncorrelated "null model" networks (see Figure 4.1). Higher correlation coeffi-

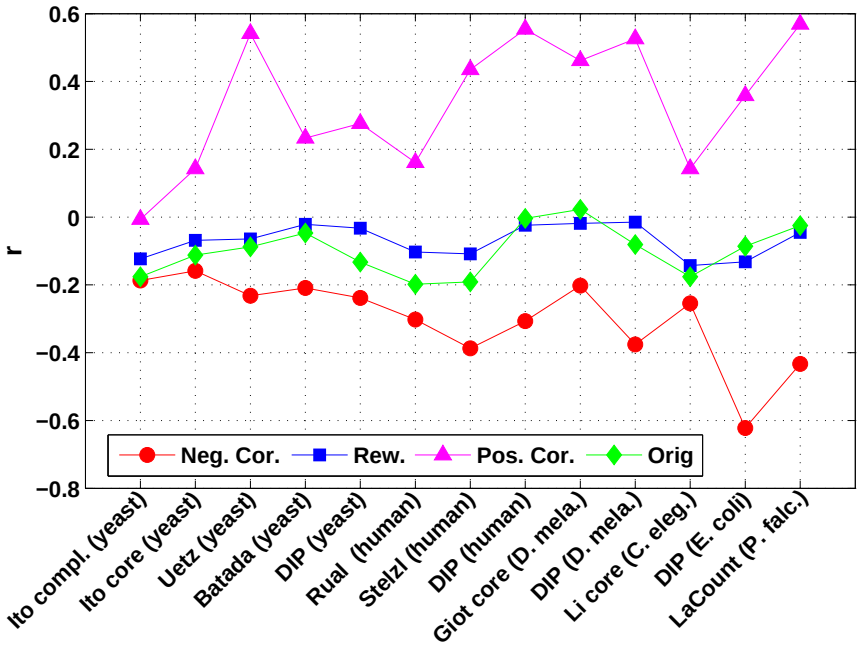


Figure 4.1: Correlation coefficients (r) observed in the original PPI networks and the corresponding negatively correlated, rewired and positively correlated reference networks.

cients than in the rewired networks are only observed for the *E. coli* and human interaction maps from DIP and the *P. falciparum* and *Drosophila* Y2H networks. Interestingly, the PPI networks with the highest similarity to the negatively correlated reference networks are the complete yeast interaction set from Ito *et al.* and to a lesser degree also the Ito core set. This is consistent with previously reported results (Maslov and Sneppen, 2002a,b). Contrary to that, the second large-scale yeast interaction set from Uetz *et al.* and the high-confidence network compiled by Batada *et al.* do not show a suppression of connections between hubs.

In general, the correlation coefficients of the rewired networks are also negative. This is a consequence of the positive skew in the degree distributions of the PPI networks (see section 2.2.1 for the definition of skewness) which leads to only few high degree nodes. As a consequence, hubs tend to be connected to low-degree nodes since those are most abundant even if connections between hubs are not suppressed. Differences in the degree distribution between the PPI networks might also explain why the correlation coefficients in the positively and negatively correlated networks are in some cases close to the correlation coefficients of the rewired networks and in some cases far apart.

Since large-scale experiments are very error-prone, we simulated the effects of measurement errors on the networks by randomly removing 10% of the interactions and adding another 10% in four different ways (see section 4.2.3). The four different strategies for simulating measurement errors resulted in slight variations in the correlation coefficients of the PPI networks. Nevertheless, the resulting correlation coefficients were still in general more similar to the rewired networks than to any of the reference networks.

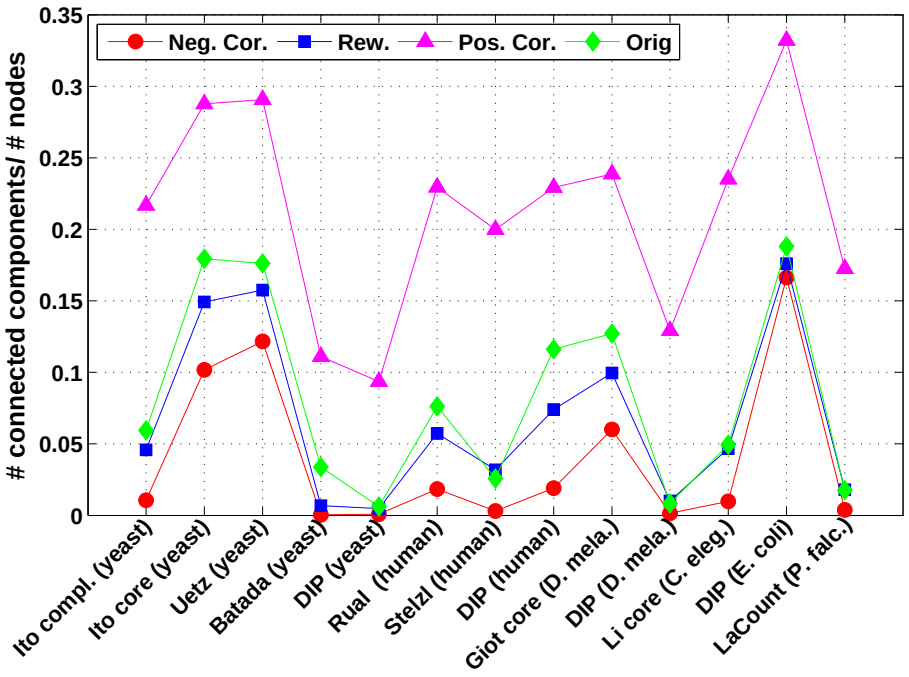


Figure 4.2: Number of connected components for the PPI networks and the negatively correlated, rewired and positively correlated reference networks. Since networks sizes and consequently also the number of connected components vary greatly between networks, numbers were scaled by dividing by the original network size. The highest number of connected components are always observed in the positively correlated networks and the lowest number in the negatively correlated ones.

4.3.3 Structural properties influenced by degree correlations

We calculated for the original and reference networks the number of connected components and the characteristic path length between all nodes and between hubs only. For the number of connected components significant differences can be observed between the different types of degree correlations in networks (see Figure 4.2). The number of connected components is highest in the positively correlated networks and lowest in the negatively correlated ones. Thus, positive correlations lead to an increased fragmentation of the network into separated clusters. Despite this trend, the PPI networks tend to consist of slightly more connected components than the rewired networks, even if they are characterized by smaller correlation coefficients than the latter.

For characteristic path length no consistent tendency can be observed (see Figure 4.3). In some cases positively correlated references have higher characteristic path lengths than negatively correlated references (e.g. for the human network by Rual *et al.*). In some cases it is the other way around (e.g. for the yeast network by Uetz *et al.*). If we restrict the calculation of the characteristic path length to paths between hubs (top 10% highest degree nodes), we observe the behavior expected from the degree correlations. The

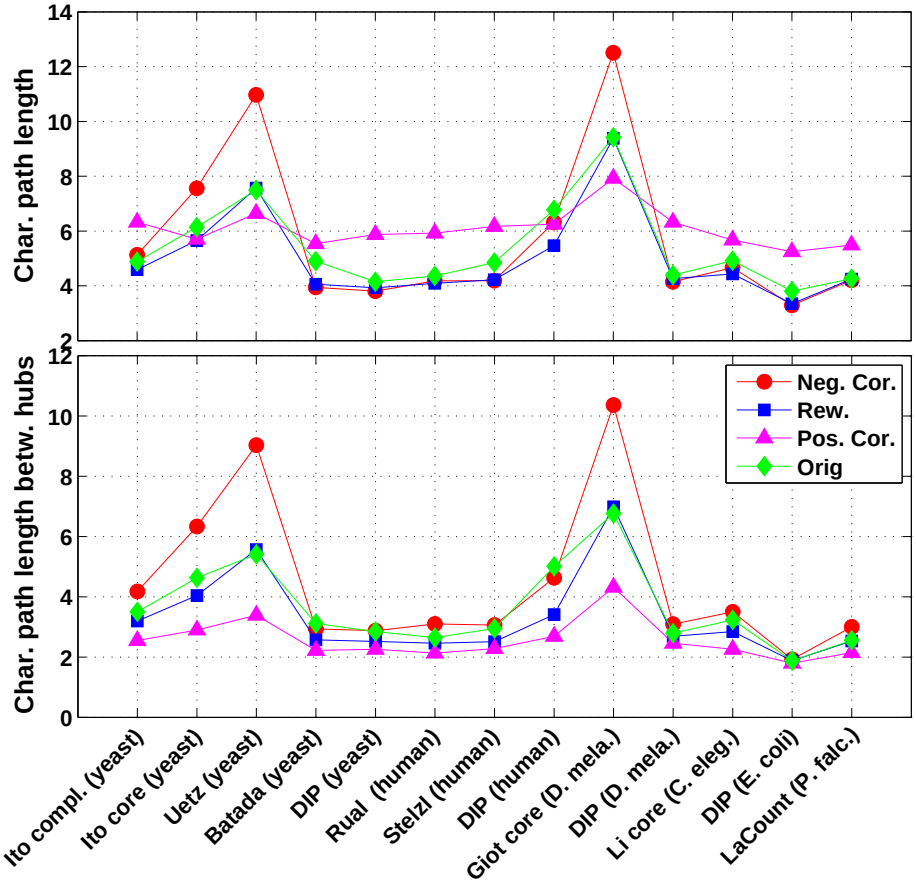


Figure 4.3: Characteristic path length is shown for all nodes (top) and for hubs only (bottom) for experimental PPI networks and corresponding reference models. No consistent tendency is observed for overall characteristic path length. If we restrict the calculation to hubs only, we observe that, as expected, the characteristic path length is longest in negatively correlated networks and shortest in positively correlated networks.

average distance between hubs is generally lowest for the positively correlated networks and highest for the negatively correlated networks (see Figure 4.3). Again, the protein-protein interaction networks tend to show characteristic path lengths between hubs similar to or slightly larger than the rewired networks. Despite the significant negative degree correlations observed in the complete and high-confidence interaction networks from Ito *et al.*, these networks actually have significantly shorter path lengths between hubs than the negatively correlated networks.

In all original PPI networks and all reference networks, hubs tend to lie in the same largest component (also called the giant component). Accordingly, the number of hubs connected by a path is in general significantly higher than would be expected for a random selection of 10% of nodes from the network. On average, paths between hubs make up between 0.99% and 6.7% of paths between all nodes. These values are highest for the

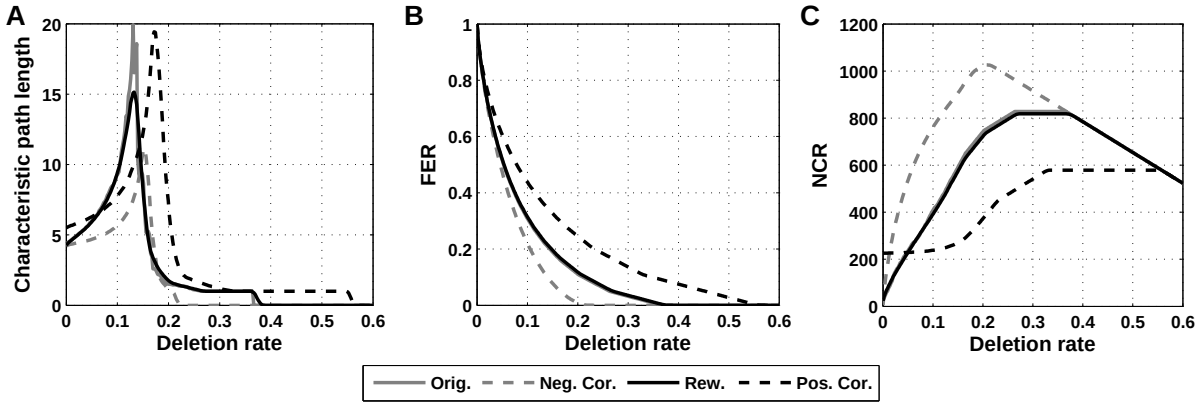


Figure 4.4: Development of network characteristics with increasing targeted deletion rate for the *P. falciparum* network (LaCount *et al.*, 2005) and the corresponding reference networks. Results are shown for the characteristic path length L (**A**), the fraction of edges remaining in the network (FER) (**B**) and the number of connected components remaining (NCR) (**C**). In this case, the original networks show almost the same behavior as the rewired networks.

positively correlated networks which consist of more components than the other network types.

4.3.4 Tolerance to targeted deletion

Apart from network properties, we compared PPI networks and reference networks for structural stability under targeted deletion of nodes by iteratively deleting the node with the currently highest degree from the networks (see section 4.2.4). We then analyzed the development of characteristic path length, the fraction of edges remaining in the network (FER) and the number of connected components remaining (NCR) with increasing deletion rate.

Development of network characteristics

Figure 4.4 shows the development of network characteristics with increasing deletion rate for the *P. falciparum* network from LaCount *et al.* (2005). In this case, the original network has almost the same correlation coefficient as the rewired networks, and accordingly, the development of network characteristics is most similar to the rewired networks. This Figure illustrates the typical behavior observed for the network properties considered. With increasing deletion rate, characteristic path length increases at first up to a point after which it decreases rapidly again as the network breaks apart into isolated components (see Figure 4.4 **A**). The rate of increase, the deletion rate at which the peak is observed and the rate of decrease afterwards can differ significantly between different network types.

In general, the characteristic path length of positively correlated networks increases

only slowly with increasing deletion rate and has a later and higher peak than observed for the rewired and even more so the negatively correlated references. This makes it difficult to compare the structural stability of networks by evaluating characteristic path length at one value for the deletion rate only. A small characteristic path length at one specific deletion rate does not necessarily imply tolerance to deletions. If it is due to a fragmentation of the network into isolated clusters, it actually suggests a lower tolerance.

On the other hand, the fraction of edges remaining in the network decreases continuously (see Figure 4.4 **B**) until all edges are deleted. In the same way, the number of connected components increases continuously until the network consists of isolated nodes only (see Figure 4.4 **C**). From this point on it decreases again. As a consequence, we can compare the structural stability easily by comparing the fraction of edges remaining (FER) in the network and the number of connected components (NCR) at a fixed deletion rate. In the following we consider a deletion rate of 10% of nodes.

Comparison of network stability

Figure 4.5 illustrates for each PPI network and the corresponding reference networks the fraction of edges remaining in the network and the number of connected components after 10% of the highest connected nodes were deleted.

The fraction of edges remaining after targeted deletion is highest for the positively correlated networks and lowest for the negatively correlated networks. Thus, negative correlations in a network lead to a higher rate of interaction loss and reduce the tolerance of the network to such deletions. Again, the PPI networks have a similar or slightly smaller fraction of edges preserved than the uncorrelated networks. The correlation coefficient of a network and the fraction of edges remaining after deletion are significantly correlated (Pearson correlation coefficient: 0.72) for the protein-protein interaction networks. Nevertheless, higher correlation coefficients do not necessarily lead to a higher fraction of edges retained. For instance, although the yeast network by Uetz *et al.* and the *E. coli* network have similar correlation coefficients, the *E. coli* network and all of the corresponding reference networks have a significantly smaller fraction of edges preserved than the Uetz *et al.* network.

The analysis of the number of connected components leads to similar conclusions about the reduced deletion tolerance of negatively correlated networks. Although positively correlated networks consist of more connected components to begin with, the deletion of hubs results in less connected components than in the negatively correlated networks and also the rewired networks. Thus, the positively correlated networks appear to break up into isolated clusters at a much lower rate. Again we observe that the PPI networks are most similar to the uncorrelated reference networks and not to the negatively correlated ones.

By changing the parameter τ (see section 4.2.1) in the creation of the positively and negatively correlated reference networks, correlation coefficients can be varied significantly. We performed simulations with different values for τ to show that our results do not only apply to the two values chosen (see Figure 4.6). Furthermore, simulations were also performed for different theoretical network models (random graph (ER), exponential (EX)

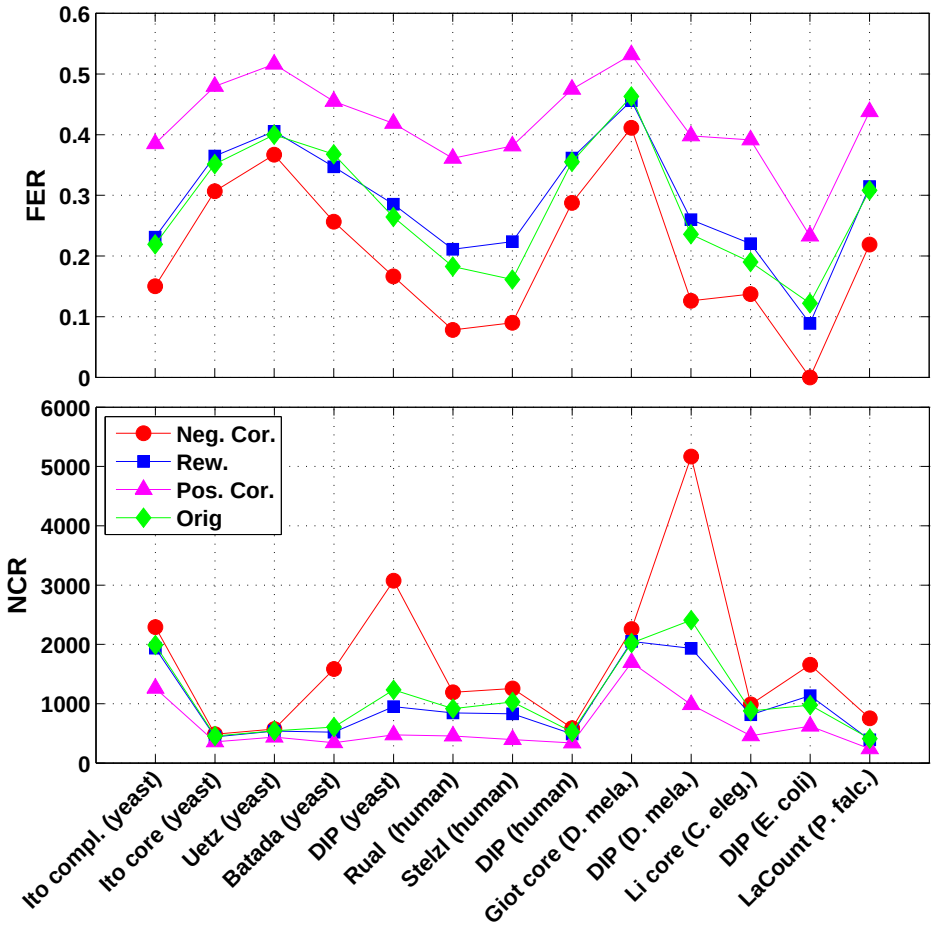


Figure 4.5: Changes in network structure after deletion of 10% of hubs. Fraction of edges remaining (FER) after deleting the 10% most highly connected nodes (top) and resulting number of connected components (NCR) (bottom) are shown for all PPI networks examined and the corresponding reference networks. Both network characteristics show that negatively correlated networks are most vulnerable to targeted deletion. Furthermore, the original PPI networks behave most similar to the uncorrelated references. Differences in NCR between different PPI networks are due to differences in network size.

and power-law networks (PL), see section 3.3.2) with different network sizes (200, 1000, 5000) and average degree values (2, 4, 6) (see Figure 4.7). These additional simulations confirm the results presented above. Thus, we can conclude that positive degree distributions increase the fragmentation of the original network but also its tolerance to targeted deletions.

4.4 Discussion

In this thesis, we investigated degree correlations in protein-protein interaction networks, the associated effects on network structure and the structural stability upon selective dele-

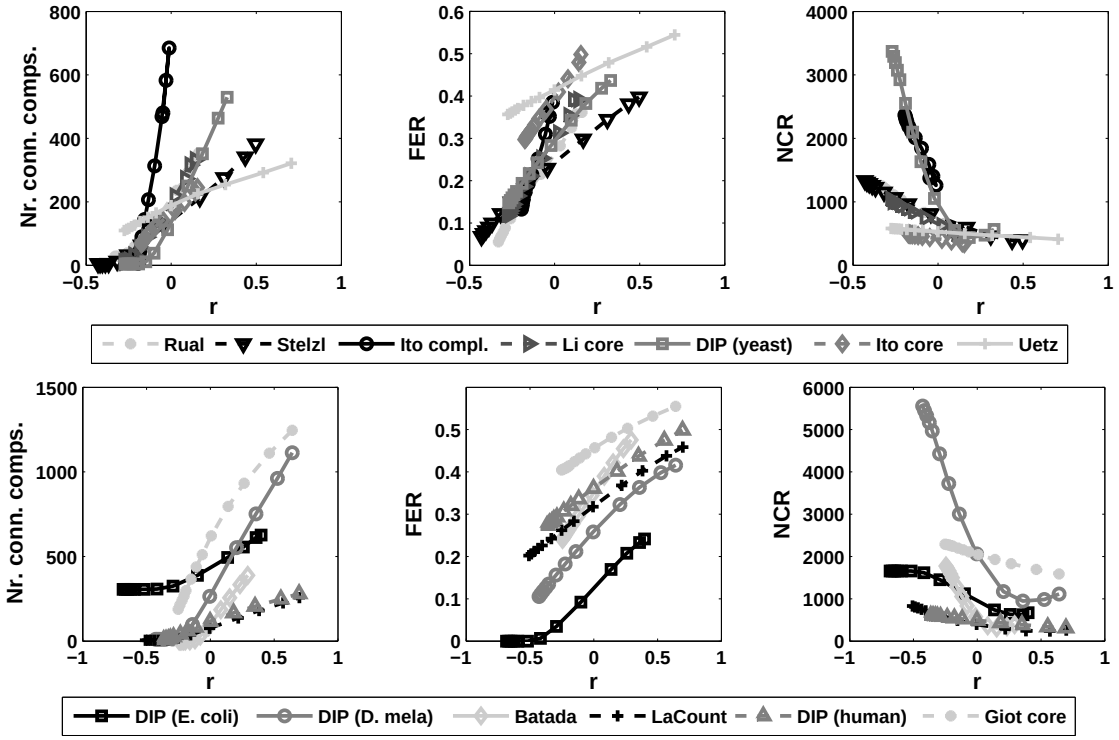


Figure 4.6: Results for reference networks with different values of degree correlations. Simulations were performed with reference networks created with different values of the parameter τ (13 different values between -0.3 and 5 were chosen). For each reference network, number of connected components in the original reference network, the fraction of edges remaining at a deletion rate of 10% (FER) and the number of connected components at this deletion rate (NCR) are plotted against the correlation coefficient r. The results for different values of τ for each PPI network are connected by lines to show the overall tendency. The number of connected components in the original reference networks as well as the fraction of edges remaining after 10% deletion rate increase continously with the correlation coefficient. The number of connected components at 10% deletion rate on the other hand decreases with the correlation coefficient. Although minor increases are again observed for high correlation coefficients, the number of connected components are still smaller than for the negatively correlated networks.

tion of hubs. For this purpose, we developed a model to simulate different types of degree correlations in networks. We compared several protein-protein interaction networks against negatively and positively correlated reference networks created with our model as well as the randomly correlated "null model" of Maslov and Sneppen.

Our results show that PPI networks are in general most similar to uncorrelated networks with regard to degree correlations and all other network properties considered. In this respect, they show a fairly consistent tendency across organisms and experiments. Only in a few cases did we observe considerable negative correlations such as in the com-

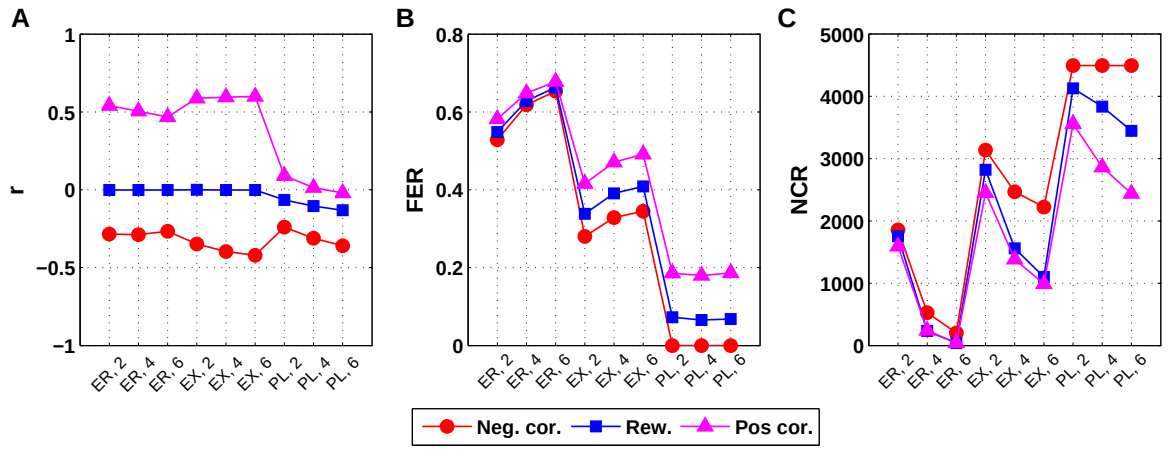


Figure 4.7: Network stability of the theoretical network models: Erdős and Rényi (ER) random graphs, exponential (EX) and power-law (PL) networks with 5000 nodes and average degree values $\bar{k}$ of 2, 4, and 6. Results are shown for (**A**) the correlation coefficients, (**B**) the fraction of edges remaining in the network at 10% deletion rate (FER) and (**C**) the number of connected components at this deletion rate (NCR).

plete yeast interaction network of Ito *et al.* (2001). It has been argued that the yeast-two hybrid system may artificially amplify negative degree correlations in protein-protein interaction networks (Aloy and Russell, 2002). Since we analyzed all large-scale Y2H interaction networks available for eukaryotes at the time of this study, our results suggest that Y2H experiments are not systematically biased towards negative degree correlations in the resulting networks. If they were, we would expect a much more pronounced tendency towards negative correlations in all networks and not only for one experiment.

The differences observed in structure and stability between positively and negatively correlated networks might explain why both seem to be avoided in PPI networks. Positive degree correlations lead to a fragmentation of the network into more connected components than in the negatively correlated or uncorrelated networks. However, hubs still tend to be located in the largest component. Thus, networks with different degree correlations do not differ in the location of hubs, but in the location of small-degree nodes. In negatively correlated networks, connections between hubs and low degree nodes include these low degree nodes into the largest component. Contrary to that, in positively correlated networks the low degree nodes cluster together to form smaller connected components.

Although modularity is desired in networks to prevent unwanted cross-talk, different modules still have to interact with each other to work in a coordinated fashion. This suggests that positive degree correlations are not realized in real protein-protein interaction networks because of the consequential increased fragmentation of the network. Furthermore, positively correlated networks are characterized by short distances between hubs. As suggested by Maslov and Sneppen (2002b), such short distances between hubs might make the network more vulnerable to random perturbations in protein concentrations. Therefore, a separation of hubs as observed for negatively correlated networks might be

favored.

On the other hand, positive degree correlations show a higher structural stability when hubs are deleted selectively. Although they consist of more components to begin with, they fall apart at a much lower rate than the other reference networks and less interactions are lost in the deletion process. In this case, the networks most affected by targeted deletion are the negatively correlated references. These differences in deletion tolerance may be explained as follows.

The deletion of a hub also removes interactions of its neighbors. As a consequence, in negatively correlated networks interactions of low degree nodes are removed, whereas in positively correlated networks interactions of other hubs are affected. However, these other hubs are most likely to be deleted in one of the next steps which would lead to the loss of the corresponding interactions anyway. Furthermore, in negatively correlated networks connections of low-degree nodes to a component are preferentially realized via hubs. Thus, the deletion of these hubs disconnects the low-degree nodes from the component. In the positively correlated networks, connections are more often via other low-degree nodes which are only deleted at a later stage.

If all interactions are more or less equally vital, the more interactions are lost the more damage is done to the cell. Furthermore, an increased disintegration of the network will prevent communication between modules and thus affect cellular processes. As a consequence, the differences in structural stability suggest a higher vulnerability of negatively correlated networks to a possible selective attack on hubs.

A biological interpretation of our results might explain why protein-protein interaction networks only show random degree correlations. Although negatively correlated networks may be more resilient to perturbations, they are also more vulnerable to a targeted attack at hubs. On the other hand, the high fragmentation of positively correlated networks might make them unfavorable as well. This suggests that both types of correlated network structures are selected against. Alternatively, network evolution processes might create more easily uncorrelated structures. In this case, positive or negative degree correlations may not be beneficial enough to lead to a deviation from these processes.

4.5 Conclusions

We showed that apart from the degree distribution, degree correlations can have a significant effect on network structure and the stability of the network under selective deletion of hubs. We observed that positive degree correlations lead to an increased fragmentation of the network into isolated components. Negative correlations, on the other hand, decrease the tolerance of the network to a selective deletion of hubs. Interestingly, we found for the PPI networks that they deviate only marginally from the uncorrelated "null model" both with respect to degree correlations and tolerance to targeted deletions. Thus, although large variations are possible, they are not realized at all in biological interaction networks. This may be explained by selective disadvantages associated with both types of degree correlations under different conditions.

Chapter 5

Analysis of herpesviral interaction networks

5.1 Introduction

In the previous chapters, we focused on the analysis of large intracellular interaction networks containing a thousand or more proteins and interactions. In this chapter, we focus on the analysis of small viral interaction networks which contain only a few hundred proteins and interactions at most. For viruses, or pathogens as such, two types of interaction networks can be distinguished: intraviral networks between viral proteins and virus-host interaction networks between virus and host proteins.

Contrary to large eukaryotic interactomes which so far could not be screened exhaustively, systematic screens of all possible interactions are feasible in viruses due to the small number of proteins. Nevertheless, apart from small RNA viruses (Flajolet *et al.*, 2000; Guo *et al.*, 2001; von Brunn *et al.*, 2007), only few virus interactomes have been studied in Y2H screens such as the bacteriophage T7 (Bartel *et al.*, 1996) and Vaccinia virus (McCraith *et al.*, 2000) interactomes. Recently, genome-scale Y2H studies on intraviral interactions have been published for the herpesvirus species Varicella zoster virus (VZV), Kaposi's sarcoma-associated herpesvirus (KSHV) (Uetz *et al.*, 2006) and Epstein Barr Virus (EBV) (Calderwood *et al.*, 2007). Additionally, interactomes for two other herpesvirus species, Herpes Simplex Virus 1 (HSV-1) and murine Cytomegalovirus (mCMV) as well as a second and independent interactome for EBV were identified with the Y2H system by the group of Jürgen Haas at the Max von Pettenkofer-Institut of the LMU München who had also conducted the screens on VZV and KSHV. In collaboration with the Haas group, we performed the combined computational analysis of the five herpesviral networks (Fossum *et al.*, 2008).

Herpesviruses form a family of large DNA viruses and are divided into three taxonomic subfamilies (α, β and γ) (McGeoch *et al.*, 2006). The α-herpesvirus lineage split off earliest from the common root about 400 million years ago (McGeoch *et al.*, 2006) while the β and γ subfamilies are more closely related (see Figure 5.1 **A**). Although all herpesviruses

have a similar virion architecture, they differ significantly in genome size, content and organization. Genome size ranges from $\sim$ 120 kbp, e.g. for VZV from the α-herpesviruses (Davison and Scott, 1986), to $\sim$ 240 kbp, e.g. for human Cytomegalovirus (hCMV) (Dolan *et al.*, 2004), a member of the β-herpesviruses. Gene content correlates with genome size and ranges between $\sim$ 70 (VZV) and $\sim$ 170 (hCMV) genes. Despite the large evolutionary distance, a common set of 41 core orthologs is conserved in all herpesvirus species (Davison, 2004; McGeoch *et al.*, 2006) (see Figure 5.1 **B**) which are generally involved in fundamental processes of the virus life cycle, such as DNA replication, processing and packaging, and which are to a large degree essential for viral replication in vitro (Dunn *et al.*, 2003; Yu *et al.*, 2003; Song *et al.*, 2005).

After primary infection, herpesviruses persist latently in host cells for the lifetime of the host (Hudnall and Stanberry, 2006) and, consequently, are some of the most common viruses. For instance, EBV prevalence in the human population has been estimated at up to 90% based on serological studies (Herbert *et al.*, 1995; Cohen, 2000; Hudnall *et al.*, 2008). Lytic reactivation of the herpesviruses has been associated with several diseases such as shingles (Gilden *et al.*, 2000) and a number of herpesviruses have been implicated in the development of cancer (Cohen, 2000; Smith *et al.*, 2002; Ganem, 2006). By comparing the interactomes of these important human pathogens we aim to identify common interactions important for the herpesviral life cycle and infection which may reveal potential strategies for medical treatment of herpesvirus infections.

The comprehensive analysis of the five herpesvirus interactomes showed that genome-wide screens of viral interactions are sufficiently sensitive for between-species comparisons to identify the basic structure of the interaction networks and a common core of interactions (Fossum *et al.*, 2008). We found that interactions are to a large degree conserved between orthologs in herpesviruses. Accordingly, the low coverage of individual Y2H measurements can be increased by the comparison of interactomes from several herpesviruses. In this way, biologically relevant interactions can be identified which may not be apparent from the interactome of each individual species alone.

In addition to intraviral interactions, we analyzed virus-host interactions for VZV and KSHV determined by the Haas group (Dong *et al.*, 2008) and the EBV virus-host interactions published by Calderwood *et al.* (2007). Bioinformatical analysis was performed together with Yu-An Dong. Recently, host-pathogen interactions from public databases were analyzed by Dyer *et al.* (2008) for 190 pathogen strains who found that pathogens preferentially target hubs and so-called bottlenecks, i.e. proteins which are important for many paths in the networks. Furthermore, they identified several biological processes enriched for pathogen targets. Apart from the large-scale EBV interaction networks, these results are mostly based on interactions from small-scale experiments and involve only relatively few interactions per species. In our comparison of results for the genome-scale studies in VZV, KSHV and EBV, we found that targeting of hubs and protein complexes appears to be a common mechanism of host interactions in herpesviruses. Furthermore, contrary to viral target proteins determined in small-scale experiments, cellular targets identified in the large-scale Y2H screen are not significantly enriched for important functional categories and biological processes.

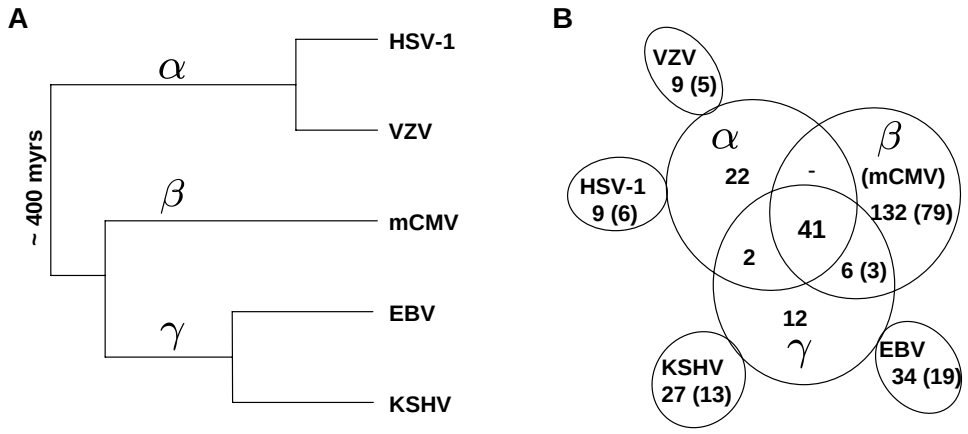


Figure 5.1: **A)** Phylogeny of the five investigated herpesviruses and their classification into the α, β and γ subfamilies. Edge length indicates the approximate evolutionary distance. **B)** Sunflower structure induced by the protein sets of the five viruses and their intersections. For each overlapping area the total number of shared proteins and the number of shared proteins with interactions in at least one species (in brackets) are indicated.

5.2 Methods

5.2.1 Analysis of intraviral interactions

From the five individual intraviral networks determined with the Y2H system by the Haas group for HSV-1, VZV, mCMV, EBV and KSHV, an overlay network was created by merging orthologous proteins into orthology groups and interactions between orthologous proteins. Orthology relationships were assigned based on Davison (2004). Each orthology group consists of proteins from at least two of the five herpesviruses. The 41 core orthology groups are a subset of all orthology groups and contain proteins from each of the five herpesviruses. The overlay network was then used to predict interactions between core proteins across species and to analyze network characteristics.

Furthermore, a phylogenetic tree was constructed for the core network and the complete network based on the Jaccard distance. The Jaccard distance between two species i and j is calculated as

$$d_{ij} = 1 - \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{5.1}$$

where C_i and C_j are the set of interactions between core and shared proteins in each species, respectively. The neighbor-joining algorithm of the PHYLIP package was then used for the tree construction (Felsenstein, 1989).

5.2.2 Analysis of virus-host interactions

Interactions between viral proteins from KSHV and VZV (from the Y2H study of the Haas group (Dong *et al.*, 2008)) and EBV (from the study of Calderwood *et al.* (2007)) to human proteins were connected to a network of human protein-protein interactions taken from the Human Protein Reference Database (HPRD) (Peri *et al.*, 2004) and the Biological General Repository for Interaction Datasets (BioGRID) database (Breitkreutz *et al.*, 2008). We then compared the distribution of degree and betweenness centrality (see section 2.2.6) for the viral targets against all other proteins in the human networks with the Kolmogorov-Smirnov test in R (R Development Core Team, 2007).

To identify functional groups which are significantly over-represented among human interaction partners of viral proteins, we performed a Gene Ontology (GO) over-representation analysis. The Gene Ontology provides a controlled and hierarchical structured vocabulary for describing the biological function of a gene or protein (Ashburner *et al.*, 2000). It is divided into three separate ontologies for describing the biological process, molecular function and cellular components for a gene. Annotations of GO terms to human genes and gene products were taken from the GO website (`http://www.geneontology.org/`).

GO terms over-represented among the viral targets are identified by calculating the p-value for finding at least l genes annotated with a specific GO term among the n viral targets. This p-value is calculated with the hypergeometric distribution.:

$$f(k|N,m,n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{5.2}$$

where k is the number of proteins targeted which are annotated with a specific GO term, m the total number of proteins annotated with this GO term, N the total number of proteins and n the total number of targeted proteins. The p-value that l or more targeted proteins belong to the specific functional category is then calculated as

$$p(k \geq l) = \sum_{k=l}^{\min(n,m)} f(k|N,m,n). \tag{5.3}$$

This corresponds to a one-sided Fisher's exact test (Fisher, 1935).

A GO term is annotated to a gene if either the term itself or one of its subterms in the ontology are annotated to that gene. GO over-representation analysis was performed with the Ontologizer package (Bauer *et al.*, 2008). P-values were corrected for multiple testing with the method of Benjamini and Yekutieli (2001), a more conservative version of the method of Benjamini and Hochberg (1995), which controls the false discovery rate (FDR) and does not require the tests to be independent. A significance threshold of 0.05 was used to extract significantly enriched GO categories.

5.2.3 Text mining

Literature interactions were identified using a combination of text mining and manual curation. A set of $\sim 87,000$ PubMed abstracts on herpesviruses was screened using the

text mining program ProMiner (Hanisch *et al.*, 2005) for occurrences of human proteins and proteins of any of the five viruses considered. Furthermore, human Cytomegalovirus (hCMV) proteins were included in the search as well. For identifying known viral interactions, 565 (for intraviral interactions) and 661 (for virus-host interactions) abstracts were selected which were found to contain at least two different proteins of the same virus or at least one viral and one human protein and some variation of the word 'interact'. Interactions were then extracted manually from the corresponding articles. In both cases, only direct, physical interactions were considered.

Furthermore, we evaluated whether publications on herpesviruses are enriched for targeted cellular genes. For this purpose, the fraction of herpesviral human target proteins found in herpesviral abstracts was compared against the same fraction for all human proteins ($\sim 30,000$) contained in the gene name dictionary used for text mining (Fundel and Zimmer, 2006). Again p-values were calculated with the hypergeometric distribution.

5.3 Results

5.3.1 Intraviral interactions of five herpesvirus species

The five Y2H screens of intraviral interactions performed by the Haas group (Fossum *et al.*, 2008) identified 111 interactions for HSV-1, 406 for mCMV and 218 for EBV (see Table 5.1). In combination with the previously published interactomes for VZV (173 interactions) and KSHV (123 interactions), altogether 1,031 intraviral interactions were obtained in the five herpesviral species. In the following we describe the analysis of these interaction networks (Fossum *et al.*, 2008).

Coverage of the five interactomes was evaluated against 120 previously published herpesviral interactions identified in our literature search. Of these 120 interactions, 17 (14.2%) could be detected in at least one virus. Of the 43 EBV interactions determined in the Y2H screen by Calderwood *et al.* (2007), only 6 (14.0%) could be confirmed in our screen. Similarly, only 5 out of 40 (12.5%) interactions determined in a Y2H screen of HSV-1 structural proteins (Lee *et al.*, 2008) were recovered. As Y2H studies generally suffer from low coverage (Huang *et al.*, 2007), these low conformation rates are not surprising. For instance, in the previous study of human interactions by Rual *et al.* (2005) only 2.3-8.4% of known interactions were recovered. On the other hand, this means that more than 97 % of interactions determined in these five screens have not been identified in other studies so far.

Network topology was analyzed for all five herpesviral interactomes. Due to the small number of nodes, reliable estimation of the degree distribution is difficult and not always meaningful (see Table 5.1 and Figure 5.2 **A** for the degree distribution of mCMV). As for eukaryotic networks, interactions are not uniformly or normally distributed among proteins and the degree distribution is highly asymmetric with most proteins having only few and a few proteins having many interactions. We first fitted the degree distributions to a power-law distribution using a simple linear fit in the logarithmic scale. This yielded estimates

Network properties	HSV-1	VZV	mCMV	EBV	KSHV
$\|V\|$	48	57	111	61	50
$\|E\|$	111	173	406	218	123
$\bar{k}$	4.63	6.07	7.32	7.15	4.92
γ (linear fit)	0.99	0.79	0.94	0.74	0.82
γ (ML)	1.87	1.71	1.63	1.64	1.76
C	0.249	0.393	0.244	0.400	0.146
C/C_{rand}	2.53	3.62	3.67	3.36	1.45
C/C_{rew}	0.80	1.00	1.24	1.14	0.69
L	2.79	2.34	2.84	2.44	2.84

Table 5.1: Network properties of the five herpesviral interaction networks. Results are shown for the number of interacting proteins $|V|$, the number of edges $|E|$, the average degree $\bar{k}$, the power-law coefficient γ estimated both with a linear fit and the maximum likelihood method (Newman, 2005), the clustering coefficient C, the ratio of the observed clustering coefficient to the average clustering coefficient in random graphs of the same size (C_{rand}) and in rewired networks (C_{rew}) and the characteristic path length L. Here, averages were obtained from 10,000 randomized networks each.

of the power coefficient < 1 for all five interactomes. As power laws with $\gamma < 1$ cannot be normalized, they are very uncommon (Newman, 2005). For eukaryotic interaction networks, values of γ between ~ 1.6 and 2.8 (Han *et al.*, 2005) are observed.

Using a maximum likelihood method (Newman, 2005), γ was estimated at 1.6-1.9 for the herpesviral networks. Although this is closer to the values observed for eukaryotic interaction networks, the fit to the observed degree distribution is not better than the linear estimator. Furthermore, the degree distribution can be fit equally well to an exponential degree distribution. Thus, due to the small size and observed variation, we cannot decide conclusively by which function the degree distribution is described best. We can only conclude that it is asymmetric with an excess of highly connected hubs compared to random graphs. The difference in degree between hubs and the remaining nodes with few connections is significantly less pronounced than in eukaryotic networks due to the smaller number of proteins in the network.

Absolute values of clustering coefficients in these networks are very high (see Table 5.1). However, the comparison to reference networks showed only little enrichment compared to random graphs of the same size and no enrichment compared to rewired networks with the same degree distributions. Characteristic path length is very short (< 3) as expected for such small networks and the maximum distance between two proteins (diameter) is only 5-7 interactions. Interestingly, tolerance to a targeted deletion of hubs is increased compared to eukaryotic interaction networks (see Figure 5.2 **B**) and a larger fraction of nodes can be removed before the network breaks down into isolated components. This can

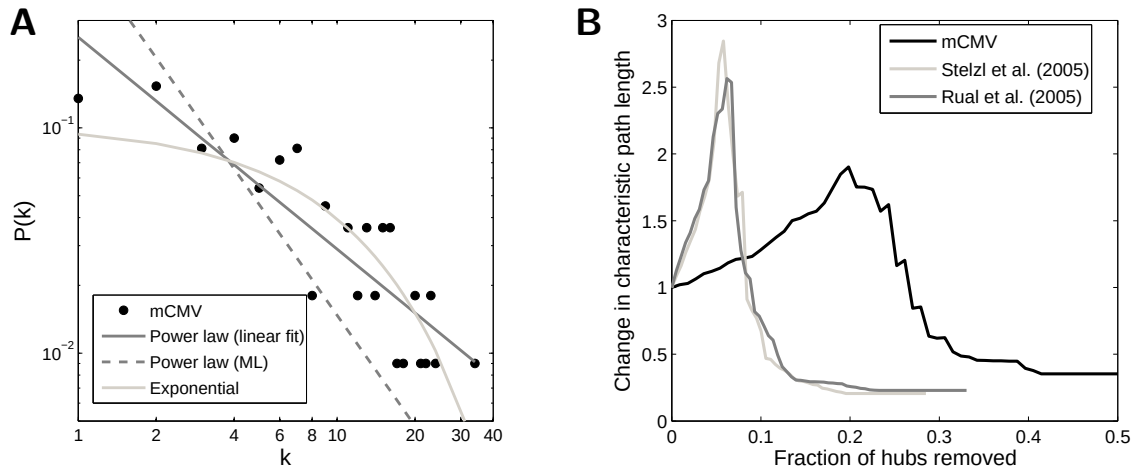


Figure 5.2: **A**) Degree distribution of the mCMV intraviral network and fitted power-law (both linear fit in the logarithmic scale and maximum likelihood (ML) estimate) and exponential distribution. **B**) Simulations of deliberate attack on the mCMV network in comparison to the human Y2H networks identified by Rual *et al.* (2005) and Stelzl *et al.* (2005) (see section 2.2.5). Characteristic path length is plotted as a multiple or fraction of the original values against the fraction of hubs removed. The herpesviral networks consistently exhibited a higher tolerance to targeted deletion, as the increase in path length is considerably smaller and the networks breaks down into isolated clusters at a much higher deletion rate.

be explained by the smaller size of the herpesviral networks and the consequently smaller differences between degrees of hub and non-hub proteins.

Herpesviral proteins can be divided into five groups based on their function in virus infection or egress or their localization within the virus capsid: DNA packaging, DNA replication, capsid, glycoproteins and tegument proteins. For each herpesviral interactome, we analyzed whether associations were significantly enriched within functional groups or between certain functional groups using the hypergeometric test but no significant association was observed. Furthermore, we compared for each species the number of tegument protein interactions against all other proteins using the Wilcoxon rank sum test since tegument proteins have a tendency to form spurious interactions. For the herpesviral networks, this did not appear to be a problem as no significant difference was observed between tegument and non-tegument proteins.

Based on the orthology assignments by Davison (2004) derived from sequence similarity and gene order, herpesviral interactomes were compared on the level of the individual interactions. In this case, more closely related species are characterized by higher sequence similarity between orthologous genes but also share more orthologous genes (see Figure 5.1). For some of these orthologous genes, proteins have diverged so much that the orthologous relationship can no longer be reliably identified from sequence similarity but only from

gene order and function. Contrary to previous inter-species comparisons which found only few interactions that were shared between species (Gandhi *et al.*, 2006), the five herpesviral interactomes analyzed here were derived with the same experimental protocols. Nevertheless, even in this case overlaps between the networks of the five herpesviral were quite small. Of 488 (409 non-redundant, i.e. conserved interactions are counted only once) interactions between proteins conserved in more than one species, only 140 (61 non-redundant) (28.7% or 14.9% non-redundant, respectively) interactions were identified in at least two screens.

Although overlaps between each pair of herpesvirus interactomes were relatively small, they were nevertheless significantly higher than observed for randomized orthology assignments. The same was true for the number of interactions between core orthologs conserved in 2, 3, and 4 species (see Figure 5.3 **A**). Randomized orthology assignments for a set of herpesviruses were obtained by first selecting the subnetwork of conserved proteins between the two or more species, and then randomizing the orthology assignments for these subnetworks. These results indicate that interactions are strongly conserved between species and that the low overlaps are largely a consequence of the low coverage of the Y2H system.

A common core of herpesviral interactions

To analyze the conservation of interactions between the core orthologs conserved in all three subfamilies, we generated an overlay of all protein interactions between these proteins detected in at least one of the five screens (core network, see Figure 5.3 **B**). As the core orthologs represent about half of the genes in HSV-1, EBV and KSHV respectively, the core network constitutes a significant part of their interactomes (see Figure 5.3 **C** for the non-core network). For mCMV, the core orthologs are less then 25% of all genes and the non-core network is much larger than the core network. Of a total of 218 non-redundant (283 in total) core protein interactions detected, 48 (113 in total) were found in at least 2, 8 in 3 and 5 in 4 species. No interaction was observed in all five interactomes.

We did not find a significant correlation between sequence identity and the number of conserved interactions detected (see Figure 5.4 **A**). For example, the interaction between the two tegument proteins UL11 and UL16 in HSV-1 was also detected in mCMV and EBV, i.e. in all three subfamilies, although sequence identity of UL11 is so low across subfamilies that it cannot be detected in all-against-all pairwise BLAST searches. In addition, interactions were not found to be preferentially conserved between closely related species and overlaps between the interactions in the core network did not correspond to the true phylogeny of herpesviruses (Figure 5.4 **B-C**). Indeed, the highest overlaps were observed between HSV-1 (α subfamily) and mCMV (β) as well as VZV (α) and EBV (γ) which belong to lineages separated earliest in herpesvirus evolution (McGeoch *et al.*, 2006). These results suggest that a common core of protein interactions has been conserved across all herpesvirus species despite long evolutionary distance.

In the overlay of all five herpesviral networks (Figure 5.3 **C**, sunflower structure), the core network shows up as the central subnetwork common to all herpesviruses. Subfamily- and species-specific networks are attached as leaves to this core. Only few connections

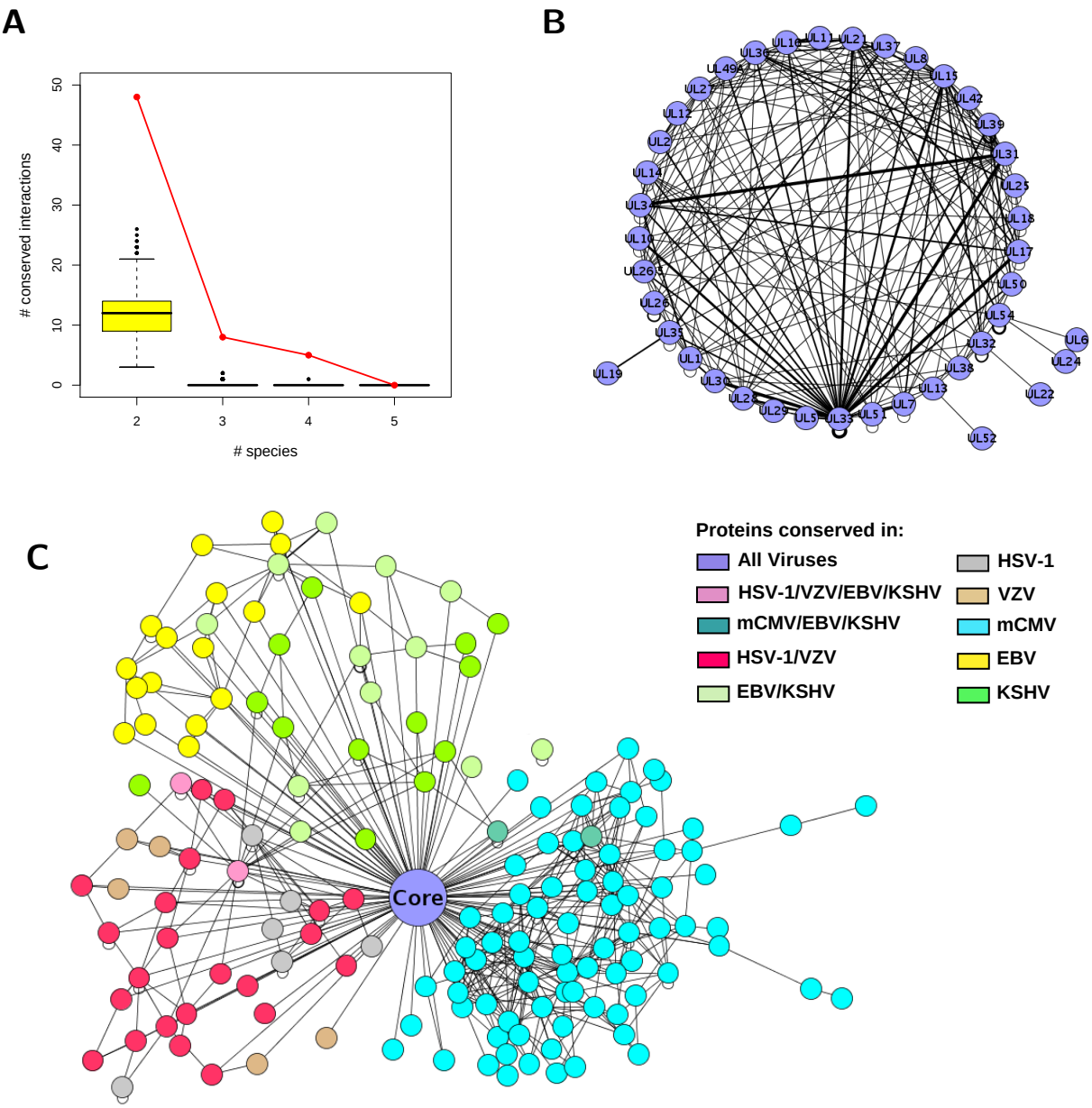


Figure 5.3: **A)** Comparison of the number of interactions conserved in 2, 3, 4 and 5 species for 1000 random orthology assignments (box plots) to the true number of interactions conserved in that many viruses (red line). **B)** Core interaction network between the 41 orthologous core proteins. The width of the edges indicates the number of species in which the interaction was detected. Nodes are labeled with the HSV-1 protein names. **C)** Interaction network between the non-core proteins of the five herpesviruses. All core proteins and the interactions between them are reduced to one central node (the core) surrounded by the leaves formed by the subfamily- and species-specific interaction networks.

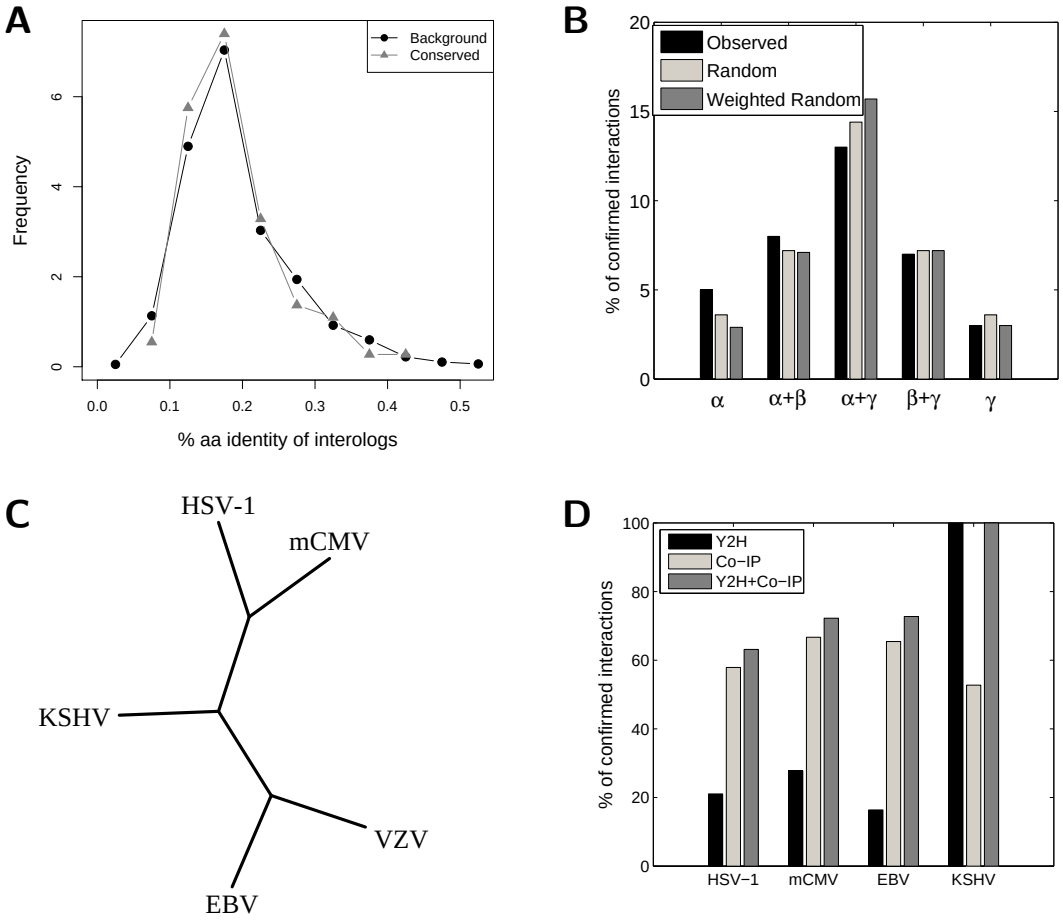


Figure 5.4: **A)** Distribution of amino acid sequence identity of the interacting proteins as compared to the background distribution of sequence identities for protein pairs not interacting. **B)** Distribution of core interactions conserved in two species within or across subfamilies compared against the random expectation if all possible combinations are equally likely or weighted based on the number of interactions in the core of each species. Interactions are not conserved preferentially between closely related species and no significant difference to the random expectation can be observed. **C)** Phylogenetic tree calculated from the herpesviral core and complete interaction networks. **D)** Histogram indicating the percentage of interactions predicted from KSHV to HSV-1, mCMV and EBV and confirmed either by Y2H, Co-IP or by Y2H and/or Co-IP.

exist between the subfamily-specific networks due to few shared proteins outside of the core. Consequently, the phylogeny calculated from the complete interaction network is the same as observed for the core. Our data provides evidence that the viral core network is extremely dense while the non-core network appears relatively sparse. However, since non-core interactions were tested in at most two species and not in five as the core interactions,

the non-core network may be equally dense. Indeed, no consistent difference was observed between the number of intraviral core and non-core interactions when considering each network separately.

To confirm that interactions between orthologous proteins are indeed conserved to a large degree, 92 interactions were predicted from 55 interactions detected in KSHV for the corresponding orthologs in HSV-1, mCMV and EBV and then tested with co-immunoprecipitation (Co-IP) by the Haas group. Predictions for VZV could not be verified due to experimental problems. 11/19 (58 %) of the predicted interactions could be confirmed in HSV-1, 12/18 (67 %) in mCMV and 36/55 (65%) in EBV, in comparison to 29/55 (53 %) in KSHV itself (see Figure 5.4 **D**). As negative controls, six interactions which were not detected in any of the initial Y2H screens were also tested in these four viruses (23 interactions in total, one could not be tested due to experimental problems). Of these interactions, two were tested positive in two viruses and one in only one virus. Taken together, this means that 5 out of 23 (21.7%) interactions were positive upon retesting. Although the confirmation rate of these negative controls is relatively high it is still much lower than observed for interactions determined in at least 1 screen. Furthermore, due to the low coverage of the Y2H system, many interactions, in particular weak interactions, were most likely missed by all of the Y2H screens. The positively tested controls may be examples of such interactions and not necessarily false positives.

Since the confirmation rate by Co-IP for the KSHV Y2H interactions is not higher than for the interactions predicted in HSV-1, mCMV and EBV, a high percentage of interactions appears to be conserved between core orthologs despite low sequence identity of some of the orthologs across subfamilies. To assess how complete the core network is, we evaluated the average number of new interactions which are added to the core network with each new Y2H screen. If core interactions indeed are conserved, coverage for the core network should increase with each new herpesviral interactome. Although the number of newly discovered interactions within the core steadily decreased with each new screen, saturation does not seem to be reached yet. On average, about 26% of interactions in the core network are identified for each individual species and about 13% of the interactions found for any of the other 4 species in the core. If we use this as coverage estimates for each screen, we expect that between 23-50% of the interactions in the core network have not been identified yet in any of the 5 screens. Thus, although coverage for the core network could be increased, a significant fraction of interactions is still missed.

Identification of a central viral protein

Most core proteins are essential and a majority can be found in herpesvirus virions, the infective particles in which the virus is released from the cell. The virion is composed of an icosahedral capsid, an amorphous tegument layer and a lipid bilayer membrane with embedded glycoproteins (McGeoch *et al.*, 2006). Using the high-coverage core network, a map of conserved protein interactions in the herpesviral virions was generated (see Figure 5.5 **A**). One outstanding example for a highly connected protein in this virion map is the mCMV M51 ortholog (HSV-1 UL33, VZV Orf25, EBV BFRF4 and KSHV Orf67.5). 11

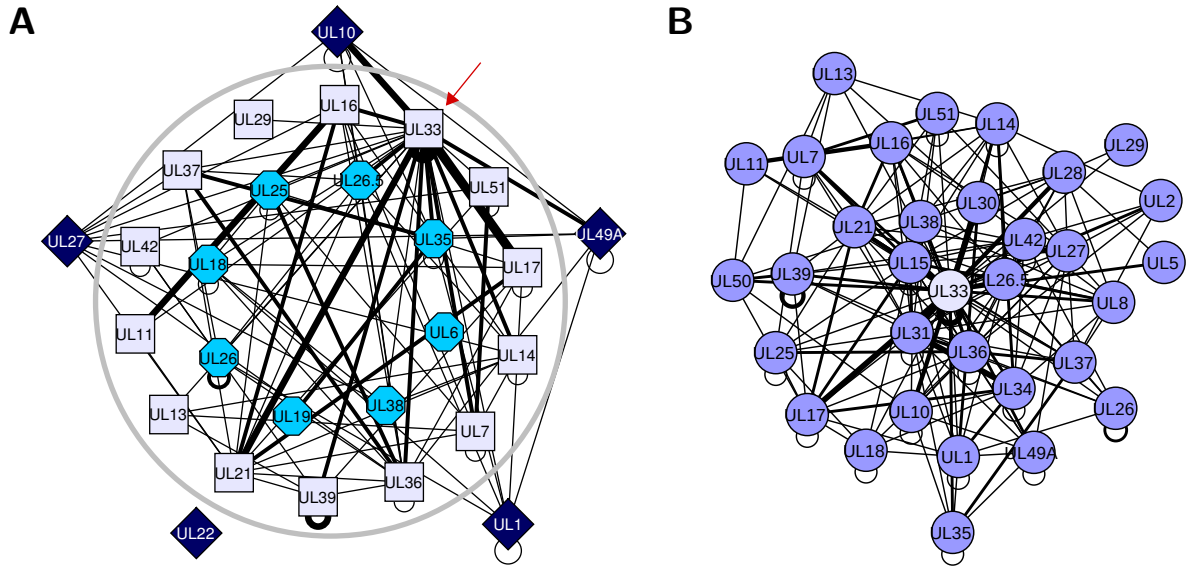


Figure 5.5: **A)** Schematic map of protein interactions between core orthologs within herpesvirus virions. The viral proteins indicated (or their orthologs) have been shown to be present in virion particles by proteomic analyses (Kattenhorn *et al.*, 2004; Varnum *et al.*, 2004). Capsid proteins (octagons) are shown in the inner layer, tegument proteins (rectangles) in the middle layer and glycoproteins (rotated squares) in the outer layer. The lipid bilayer envelope is indicated by a grey circle. Nodes are annotated with HSV-1 protein names and the width of the interactions indicates the number of species the interactions were detected in. UL33/M51 is indicated by an arrow. **B)** Interaction network of core orthologs interacting with UL33/M51.

out of 13 (86%) interactions in mCMV of M51 to other core proteins and 21 out of 33 (63%) interactions in all species taken together were found to be conserved in more than 1 species (see Figure 5.5 **B** for a map of the subnetwork of UL33/M51 and its interaction partners in the herpesviral core network).

Furthermore, 4 out of the 5 interactions observed in all but one of the 5 Y2H screens involved the UL33/M51 protein and most Y2H interactions of UL33/M51 in HSV-1 were confirmed by the Haas group even under high concentrations of an inhibitor which suppresses non-specific Y2H interactions. In an additional Co-IP screen of 22 M51 interactions in mCMV, 17 (77%) could be confirmed. This indicates that these interactions are very strong and that UL33/M51 is a major hub in the herpesviral interactome. Although UL33/M51 is among the most highly connected proteins in all five herpesviral networks, its prominent role only becomes apparent from the comparison and overlay of the five networks.

In all species apart from HSV-1, mCMV M51 orthologs interacted with M53 orthologs (VZV 27, EBV BFLF2, KSHV 69). Furthermore, in a later retesting of UL33 interactions in HSV-1, Jürgen Haas and co-workers found that the corresponding interaction with UL31,

the M53 ortholog in HSV-1, was clearly positive. While not much is known about M51 apart from its role in DNA packaging, M53 has been shown to be involved in nuclear egress in several species through its binding to M50 (HSV-1 UL34, VZV 24, EBV BFRF1, KSHV 67) (Reynolds *et al.*, 2001; Muranyi *et al.*, 2002; Fuchs *et al.*, 2002) which was confirmed in 4 of the 5 screens in this study. Both M50 in mCMV and UL34 in HSV-1 recruit protein kinase C to the nuclear membrane. The protein kinase C subsequently phosphorylates lamins to dissolve the nuclear lamina which allows the capsids to reach the inner nuclear envelope (Muranyi *et al.*, 2002; Park and Baines, 2006). In the Y2H screens, an interaction between M51 and M50 was observed in VZV and EBV and upon retesting also in HSV-1. Moreover, the Haas group showed by fluorescent labeling that M51 is targeted to the nuclear membrane by M50 and co-localizes with both M50 and M53. These results suggest that M51 forms a complex with M50 and/or M51 and is involved in nuclear egress via this complex.

Previous studies have indicated that UL33/M51 is involved in packaging of DNA in HSV-1 (al Kobaisi *et al.*, 1991), and that it interacts with the subunits of the terminase complex (HSV-1 UL28 and UL15) (Beard *et al.*, 2002). In the five Y2H screens, UL33 was observed to interact with UL15 and UL28 in three different species. Since many of its interaction partners are present in the virion tegument, this suggests that it plays a role in tegument formation and represents a possible connection between DNA packaging, nuclear egress and tegumentation. Studies in HSV have indicated that UL33 is associated with the external surface of capsids (Wills *et al.*, 2006), which would make such a dual role reasonable. While it is not known exactly how UL33 associates with the capsid, the interaction observed between M51 and the smallest capsid protein in mCMV and EBV (mCMV M48.2, EBV BFRF3) suggests a possible manner of association.

5.3.2 Herpesvirus-host interactions

In addition to intraviral interactions, virus-host interactions were screened by Y2H in KSHV and VZV by the Haas group (Dong *et al.*, 2008) and in EBV by Calderwood *et al.* (2007). For KSHV, a matrix approach was used to screen each pairwise combination of KSHV ORFs and 5,400 distinct human cDNAs which were also used by Stelzl *et al.* (2005) to screen human interactions. Contrary to that, a library approach was used to screen VZV ORFs against two human cDNA libraries which taken together cover approximately the complete human proteome. A library approach was also used by Calderwood *et al.* (2007) to screen interactions between EBV and human proteins. In total, 252 interactions were identified between 49 KSHV proteins and 131 human proteins and 828 interactions between 60 VZV and 753 human proteins (complete set). Furthermore, 173 interactions were found between 40 EBV proteins and 112 human proteins. If more stringent criteria were used to select for high-confidence (hifi) interactions in the VZV library screens, the number of interactions was reduced to 112 interactions between 33 VZV and 107 human proteins. Computational analysis of virus-host interactions was performed in collaboration with Yu-An Dong (Dong *et al.*, 2008) and is described in the following.

Only 2 interactions were found to be conserved between VZV and KSHV or EBV, re-

spectively, and only 1 between KSHV and EBV, all of them involving core orthologous proteins. Interestingly, the two proteins interacting with orthologous proteins in VZV and KSHV, HTATIP and COPS6, are also known cellular targets of HIV-1 proteins (Kamine *et al.*, 1996; Mahalingam *et al.*, 1998). A comparison of all cellular targets of VZV, KSHV and EBV revealed an additional 11 proteins which are targeted by both VZV and KSHV by non-orthologous proteins, 12 for VZV and EBV and 6 for EBV and KSHV. No human protein was found to interact with viral proteins from all three herpesviruses. Although these overlaps are small, they are nevertheless significantly higher than expected at random (hypergeometric test, FDR p-values < 0.05) for EBV, KSHV and the VZV high-confidence targets. Thus, these small overlaps may also be a consequence of low coverage and differences in the experimental set-ups.

In a comprehensive text mining analysis of literature, we identified over 300 published interactions between herpesviral and cellular proteins, including 3 for VZV, 102 for KSHV and 117 for EBV. Of these, only 1 interaction was recovered in KSHV and 4 in EBV but none in VZV which is to be expected as few interactions have previously been reported. As some of the human interaction partners of herpesviruses identified in these Y2H screens may have been linked with herpesviral infection in literature even though no actual interaction has been described or identified, we also analyzed the enrichment of viral targets within PubMed abstracts on herpesviruses. We found that targets identified in these Y2H screens were more than twice as likely to occur in herpesviral abstracts than expected from the overall frequency of human proteins in these abstracts (hypergeometric test, FDR adjusted p-values < 0.001). These results suggest that many of the observed targets have previously been mentioned in connection with various aspects of herpesvirus morphogenesis.

Connectivity of virus proteins and cellular targets

When comparing the number of intraviral interactions and interactions to host proteins for each viral protein, we found no significant correlation for KSHV. Contrary to that, intraviral and virus-host degrees were significantly correlated in VZV and EBV (Spearman's rank correlation, p-values < 0.05). In this case, both the intraviral interaction network of EBV described above and the network identified by Calderwood *et al.* (2007) were analyzed. While we did not observe a significant difference in the average number of intraviral interactions between core and non-core proteins for VZV and KSHV, core proteins were found to have significantly more interactions to human proteins (Fisher's exact test, p-values < 0.01). For EBV a significant difference was found for the intraviral degree between core and non-core proteins but not in virus-host degree. Contrary to the intraviral interactomes, no single protein stands out as a major hub in all three virus-host interaction networks and only the EBV BVRF1 protein (HSV-1 UL25, VZV Orf 34, mCMV M77, KSHV ORF 19) is highly connected in all of them.

In a recent study (Dyer *et al.*, 2008), it was found that viral proteins preferentially target highly connected proteins with high betweenness centrality. This study was based mostly on virus-host interactions determined in small-scale experiments, but did also include the large-scale results on EBV-human interactions. To investigate whether these

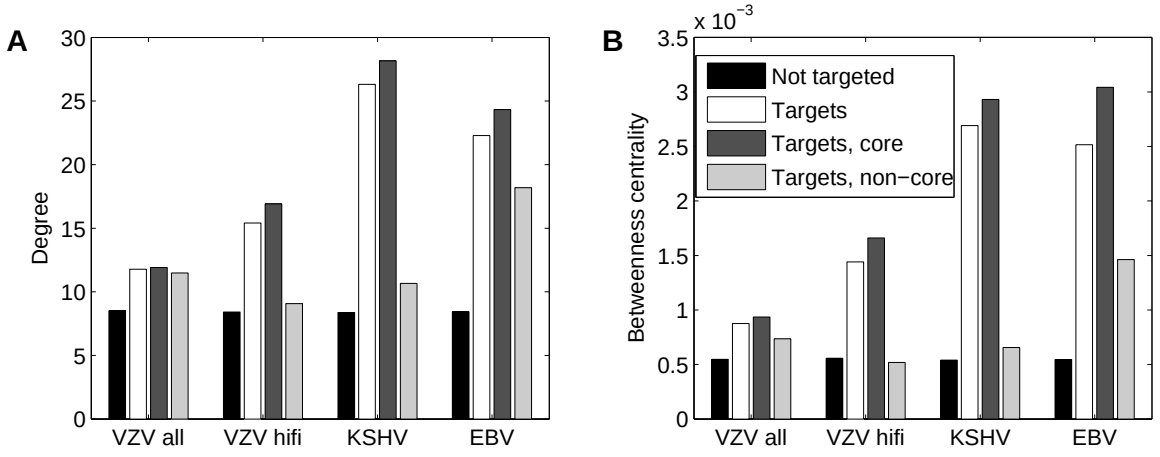


Figure 5.6: Average degree **(A)** and betweenness centrality **(B)** of viral targets in the large human network obtained from the HPRD (Peri *et al.*, 2004) and BioGRID (Breitkreutz *et al.*, 2008) databases. Results are shown for cellular proteins not targeted by the corresponding virus (black), all targets (white) of a virus, targets of core proteins (dark grey) and targets of non-core proteins (light grey). The high-confidence VZV (hifi) interactions were also analyzed separately.

results hold also for the VZV and KSHV viral-host interactions, we compiled a large set of human protein-protein interactions from the HPRD (Peri *et al.*, 2004) and BioGRID (Breitkreutz *et al.*, 2008) databases. This large data sets contains $\sim 45,000$ interactions between $\sim 10,000$ human proteins, i.e. approximately a third of the human proteome. Figure 5.6 shows average degree and betweenness centrality for the viral targets. Both average degree and betweenness centrality were significantly increased for cellular viral targets compared to the remaining proteins (Kolmogorov-Smirnov test, FDR p-values $< 10^{-3}$, see methods). Interestingly, this difference in connectivity and centrality was most pronounced for human proteins targeted by core proteins conserved in all five herpesviruses, whereas only little or no significant difference was observed for human proteins targeted by non-core proteins.

Since herpesviral proteins preferentially target hubs, we investigated whether the observed overlaps between viral-host interactomes can simply be explained by a preference for highly-connected proteins. For this purpose, we randomly sampled cellular target proteins from the complete network with the same degree distribution as the observed target proteins. For each pair of species, 1000 random target sets were sampled for each individual species. The p-value for the overlap between these species was then calculated as the fraction of samples with an overlap greater or equal to the actual observed overlap. For the comparison of EBV to KSHV and VZV overlaps were found to be significantly higher than observed for random target sets (FDR p-values < 0.02). The comparison of KSHV and VZV, however, showed no significant increase in the number of common interaction partners given the preference of both viruses to interact with highly connected cellular proteins.

Functional analysis of herpesviral interaction partners

Dyer *et al.* (2008) also analyzed the enrichment of GO functional categories among human proteins interacting with different groups of bacterial and viral pathogens. They identified several important biological processes which were enriched among cellular interaction partners of pathogens. For herpesviruses, enriched functional categories included viral reproduction, cell cycle and immune response as well as regulatory and signal transduction processes. EBV interactions were studied separately from the other herpesviruses and were derived mostly from the large-scale study of Calderwood *et al.* (2007). Interestingly, results for EBV differed significantly from the other herpesviruses with only few functional categories being enriched (protein binding, extracellular matrix and integrin binding).

We performed a GO over-representation analysis for the viral targets in the VZV and KSHV screens. If KSHV and high-confidence VZV targets were analyzed separately, no functional category was significantly over-represented. In a combined analysis of these cellular proteins, only the "protein binding" function was enriched among the targets. Only few additional categories are over-represented in the complete set of VZV interacting proteins and involve mostly mitochondrial proteins and protein complexes. Contrary to that, an analysis of cellular interactors of herpesviruses recovered from literature identifies many interesting functional categories, such as regulation of transcription and apoptosis, cell cycle and viral reproduction which is consistent with the results of Dyer *et al.* (2008). Since this enrichment is only observed in results from small-scale experiments, this may be either a consequence of the lower quality of large-scale screens or more likely due to a bias in the selection of interaction partners studied in the small-scale experiments. Furthermore, as interaction results from the small-scale experiments have also been used to annotate the function of the interacting cellular proteins, it is not surprising that many of the enriched functional categories focus on pathogen host-interactions (e.g. "entry into host cell" [FDR p-value 1.75×10^{-5}], "interaction with host" [1.78×10^{-5}], "response to virus" [0.0026]).

Although no interesting functional category was found to be over-represented among large-scale herpesvirus targets, human proteins involved in protein complexes taken from the CORUM database (Ruepp *et al.*, 2008) were enriched among viral targets (FDR corrected p-value < 0.005). In more than 90% of the cases protein complexes are targeted at only one or at most two positions. The only complex for which subunits were found to be interacting with proteins of all 3 viruses is the proteasome which is targeted at 4 different proteins (using the VZV high-confidence set). When including the complete VZV set, the ribosome and spliceosome are also found to be targeted by all 3 viruses.

5.4 Discussion

In a collaboration with the group of Jürgen Haas at the Max von Pettenkofer-Institut, we analyzed genome-scale Y2H screens of intraviral and virus-host interactions for several herpesviral species. Here, intraviral interactomes were studied in five species from all three herpesviral subfamilies. Although we observed little overlap between the five viral

networks in the Y2H screens, we could show that this is largely due to low coverage of the Y2H system and that interactions between proteins conserved in all species (core orthologs) are also conserved to a large degree. By combining interactions from all five screens, the low coverage of Y2H could be increased and important interactions and proteins could be identified.

Due to the small size of the intraviral networks, statistical analysis of the individual networks is difficult. The degree distribution cannot be reliably estimated and can be fitted equally well to a power-law and exponential distribution. Clustering coefficients correspond approximately to the values expected for the observed degree distributions and are only slightly higher than observed for random graphs of this size. As a consequence and in combination with the fact that characteristic path length is actually slightly higher than observed for random graphs of this size, these network would not be considered as "small world" networks in standard analyses, although they are small by default. Furthermore, functional groups and the number of interactions between them are too small to obtain significant results on preferential associations between them.

Thus, many interesting insights can only be obtained by a combined analysis of several related interactomes. Nevertheless, even then the small size of the networks is a limiting factor. For instance, the identification of conserved modules in networks based on direct and indirect interactions via a second protein (Sharan *et al.*, 2005), is hindered by the large fraction of proteins in the viral interactomes which are separated by no more than two interactions. On the other hand, the small size of the networks allows a more detailed analysis of the individual interactions than possible in large eukaryotic networks. This proved to be an advantage for instance for the analysis of the SARS interactome in which we also collaborated (von Brunn *et al.*, 2007).

The combined analysis of all five herpesvirus interactomes showed only few interactions conserved between herpesviruses. Nevertheless, conserved interactions were significantly enriched compared to the background expectation. From the five individual interactomes, we derived a core network containing all interactions observed between core orthologous proteins. The analysis of the core network showed that interaction overlaps between herpesviral species did not correspond to the known phylogeny derived from protein sequences. Indeed, highest overlaps were observed for species most distantly related. The same was true if we analyzed overlaps of the complete networks. These results suggest that interactions between orthologous proteins and in particular core orthologs are highly conserved and that the small overlaps between species are due to low coverage of the individual screens. To test this hypothesis, we predicted interactions from KSHV to HSV-1, mCMV and EBV and found that confirmation rates of the predicted interactions by Co-IP were comparable to confirmation rates of the original interactions in KSHV.

In a previous study, Yu *et al.* (2004) showed that interactions can be confidently transferred from one species to another if the joint sequence identity of orthologs is $> 80\%$. However, none of the herpesviral core orthologs show such a high degree of sequence conservation across species even within subfamilies. Furthermore, no correlation was observed between the number of species an interaction was identified in and the sequence identity of the interaction partners. Thus, our results indicate that sequence identity alone is not an

appropriate criterion for predicting herpesviral interactions from one species to another.

Several examples have been published for protein interactions conserved between different herpesviral subfamilies, e.g. the interaction between HSV-1 UL31 and UL34 (Reynolds *et al.*, 2001; Muranyi *et al.*, 2002; Lake and Hutt-Fletcher, 2004), and much of current knowledge about herpesvirus biology has been inferred for other species based on studies of HSV. Our study indicates that it is effectively possible to transfer intraviral interactions between orthologous proteins among herpesviruses. Thus, by generating an overlay network from several genome-wide Y2H screens in related species, the high number of false negative interactions within each individual analysis can be addressed and a more complete picture of the core interaction network can be obtained.

In the core network derived from the overlap of all five herpesvirus species, the UL33/M51 core ortholog shows up as an intraviral hub with a number of conserved interactions. Based on these highly conserved interactions, we can propose an additional functional role of UL33/M51 in tegumentation and nuclear egress which has so far not been characterized. Moreover, due to its central role for all five herpesviral interactomes, UL33/M51 should be a promising target for antiviral therapy (Loregian and Palú, 2005).

While intraviral interactions appear to be highly conserved in herpesvirus interactomes, only few conserved interactions and cellular target proteins could be identified in the large-scale screens of virus-host interactions on VZV, KSHV and EBV. Although these small overlaps may be a consequence of low coverage, a literature search also identified only few conserved virus-host interactions in herpesviruses. Of 136 previously published interactions between proteins conserved in at least two of the herpesviruses, only 7 (5.1%) have been described in more than one of the five herpesviral species discussed here. For intraviral interactions, this applies to 21 of the 87 (24.1%) interactions between conserved proteins recovered in our literature search. These results suggest that interactions of the herpesviruses with cellular proteins are less strongly conserved than intraviral interactions. This may be a consequence of the fact that viruses have to adapt to evolutionary changes in the viral host and, in particular, specific target proteins.

In a recent study on published pathogen-host interactions, Dyer *et al.* (2008) found that cellular interactors of pathogen proteins are generally highly connected and central to the host interactome. Our results show that this is also the case for human proteins targeted by herpesviral proteins. In particular, this applies to targets of core proteins which are conserved in all herpesviruses while targets of non-core proteins hardly differ from the remaining proteins in degree and betweenness centrality. As betweenness centrality is highly correlated with degree in the human network (Spearman correlation coefficient: 0.89, p-value $< 10^{-16}$), the high betweenness centrality of target proteins appears to be a consequence of their high degree values and not a distinguishing characteristic on its own. Indeed, previous studies have shown for protein-protein interaction networks that, contrary to regulatory networks, betweenness centrality is not a better predictor of essentiality than degree (Yu *et al.*, 2007).

In their study, Dyer *et al.* (2008) identified several biological processes which are preferentially targeted by pathogen proteins, such as cell cycle, apoptosis and immune response. While we could confirm these results for an extensive list of published herpesviral targets,

no significant preferential association was observed by the targets identified in the large-scale screens apart from a tendency of targets to be protein binding, which is not surprising. These conflicting results are not necessarily due to high error rates in the Y2H systems but very likely a consequence of a bias in the choice of interaction partners tested in small-scale experiments. In this case, promising candidates may have been chosen preferentially from important cellular pathways likely to be involved in virus infection and morphogenesis. Thus, large-scale experiments can provide a more unbiased view on the diverse cellular processes targeted by herpesviruses.

5.5 Conclusions

A comprehensive analysis of intraviral and virus-host interactions determined in large-scale Y2H experiments provided evidence that intraviral interactions within herpesviruses are highly conserved whereas virus-host interactions may evolve faster. Based on a comparison of five herpesviral interactomes, a common core of intraviral interactions was predicted and the low coverage of the Y2H system could be addressed. From the core network, UL33/M51 was identified as an important herpesviral hub which links DNA packaging, tegumentation and egress. Furthermore, we found that core orthologs are to a large degree responsible for the targeting of cellular hubs observed in the virus-host interaction networks. Contrary to previously published herpesviral interactors, targets identified in the large-scale screens are not enriched significantly for a specific biological function. As published results from small-scale experiments likely show an artificial bias towards important functional categories, large-scale screens may provide a more realistic picture of virus-host interactions.

Part II

Protein complexes

Chapter 6

Prediction of protein complexes

6.1 Introduction

In the first part of this thesis, we focused on the analysis of direct, physical protein-protein interactions most of which were identified with the Y2H system. Many of these interactions are formed within larger associations of proteins in complexes and diverse cellular processes are shaped by these protein complexes. Protein complexes can be predicted from networks of physical interactions but can also be directly identified with affinity purification techniques such as tandem affinity purification (TAP)(Rigaut *et al.*, 1999) (see section 2.1). So far, the only genome-scale studies on complexes have been performed with the TAP technique in the yeast *Saccharomyces cerevisiae* (Gavin *et al.*, 2006; Krogan *et al.*, 2006).

As described in section 2.1, in the TAP system affinity columns are used to purify a tagged bait protein and prey proteins interacting directly or indirectly with the bait from the cell. In the experiment of Gavin *et al.*, 1,993 distinct TAP-tagged proteins (baits) were purified successfully and 2,760 distinct proteins (preys) were identified in these purifications. In the experiment of Krogan *et al.*, 2,357 baits were purified and 4,087 preys identified. Ideally, the purification of a bait would result in the co-purification of all proteins contained in the same complex as the bait. However, like the Y2H system and all high-throughput methods, the TAP system produces measurement errors. Proteins binding unspecifically to the bait may be co-purified (false positives) while proteins from the same protein complex may fail to bind tightly enough (false negatives).

Because of these experimental errors and the large size of these data sets, sophisticated methods were developed in both studies to infer individual protein complexes from the raw purification data. However, the resulting complexes showed surprisingly little overlap (see results). After the publication of the original results, improved prediction methods were developed by Hart *et al.* (2007) and Pu *et al.* (2007). Here, the method by Pu *et al.* (2007) is based on the scoring method of Collins *et al.* (2007). Both methods used the data from the combined Gavin and Krogan purification experiments. Recently, two new methods were proposed by Rungsarityotin *et al.* (2007) and Zhang *et al.* (2008) but only

applied to the Gavin data.

A comparison of the different approaches outlines the important steps in complex determination (see Figure 6.1 **A**). First, purification experiments have to be combined for higher prediction quality. Second, scoring weights for individual protein interactions have to be determined from the purifications. Third, these scores should be converted to confidence values (between 0 and 1) assessing the likelihood of each protein interaction. Confidence values have a clear-defined minimum and maximum and, as a consequence, are more easily interpretable and comparable between experiments than arbitrary scores between 0 (or $-\infty$) and ∞.

A

	GA	KR	CP	HA	RU	ZH	BT
Experiment combination	-	-	S	U	-	-	U
Interaction scores	U	S	S	U	-	U	U
Confidence values	-	S	S	-	-	-	U
Network cutoff	U	S	S	S	-	S	U
Clustering	S	S	S	S	U	U	U
Shared Proteins	S	-	S	-	-	U	U

B

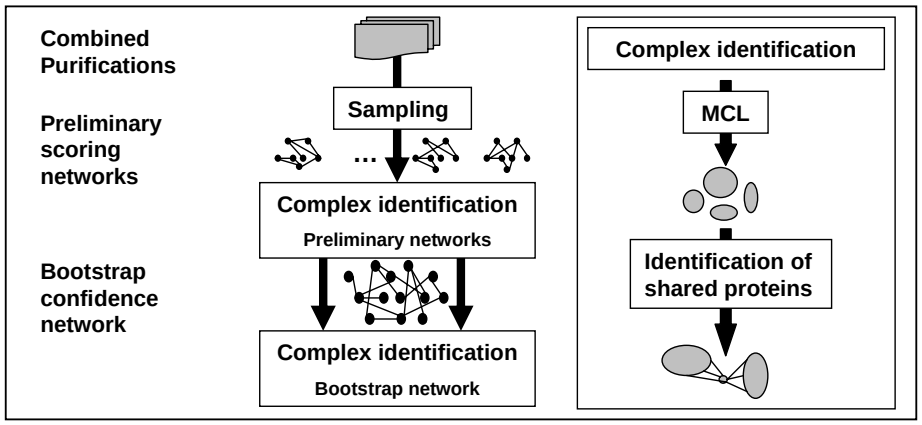


Figure 6.1: Outline of complex prediction methods. Figure **A** lists the major steps involved in complex identification and whether they are realized by the approaches of Gavin *et al.* (2006) (GA), Krogan *et al.* (2006) (KR), Collins *et al.* (2007) and Pu *et al.* (2007) (CP), Hart *et al.* (2007) (HA), Rungsarityotin *et al.* (2007) (RU), Zhang *et al.* (2008) (ZH) and the bootstrap approach described here (BT). Furthermore, we indicate whether a supervised approach based on additional training data is used (S, light grey) or an unsupervised approach (U, white). Dark grey indicates that the step is omitted. **B**) Outline of our bootstrap approach (see methods).

In a fourth step, a cutoff has to be applied on the network to select the most confident interactions. The restricted network then has to be clustered to obtain individual complexes. If the corresponding clustering method only produces disjoint complexes, a final step has to be included to identify proteins shared between complexes. This is necessary because proteins can be involved in more than one complex.

Apart from the combined method by Collins *et al.* (2007) and Pu *et al.* (2007), all described approaches leave out at least one of these steps. Since the method of Rungsarityotin *et al.* (2007) focuses on predicting the complexes directly from the experiments without calculating interaction scores first, it implements only the clustering step. Figure 6.1 **A** gives an overview on which methods implement which steps. With the exception of the method by Rungsarityotin *et al.* (2007), all approaches described above are supervised as they rely more or less heavily on the availability of additional training data in the form of known complexes for at least one step.

For yeast, manually curated complex data can be taken from the MIPS database (Mewes *et al.*, 2004) and the study of Aloy *et al.* (2004). Furthermore, complexes can be automatically extracted from Gene Ontology (GO) annotations (Ashburner *et al.*, 2000). Unfortunately, the resulting complexes are of lower quality than the manually curated ones (see results). Recently, manually curated protein complex data sets have become available for human and other mammals which cover about 12% of the protein-coding genes in human (Ruepp *et al.*, 2008). For other organisms, such information on complexes is not available which limits the applicability of the supervised approaches significantly.

Even if reference sets are available as for yeast, a large fraction of them have to be set aside as independent test sets to evaluate the quality of predicted complexes. Although some of the above mentioned approaches distinguish between test and training set by choosing one of the yeast reference sets for training and a different one for testing, these sets overlap to a large degree and, thus, are not sufficiently disjoint to guarantee a reliable performance estimate.

To deal with these problems, we developed an unsupervised algorithm for the identification of protein complexes from the purification data alone which implements all steps described above (Friedel *et al.*, 2008a). Since no additional information on protein complexes is required, our approach is not limited to yeast and other organisms for which many protein complexes have already been identified but can be applied easily to large-scale purification experiments for any species.

We show that our approach is equivalent to the best supervised methods both with regard to functional and localization similarity in the resulting complexes as well as predictive performance with regard to known yeast complexes. We find that our predictions and the Pu and Hart predictions on the combined data sets are much more consistent than the original Gavin and Krogan complexes. However, although the latest predictions are of similar overall quality, significant differences between the predicted complexes are still observed. This clearly illustrates the need for further investigations of the individual protein complexes.

6.2 Methods

The algorithm we propose for the unsupervised identification of protein complexes implements all six important steps (see Figure 6.1 **A**). Purification experiments are combined by pooling them. Interaction confidences are determined by first identifying preliminary complexes for bootstrap samples from the set of purifications (Efron, 1979; Efron and Tibshirani, 1994). The resulting confidence network is then clustered with the MCL algorithm (van Dongen, 2000) and proteins shared between complexes are identified in a post-processing step. Here, parameters are tuned based on intrinsic measures calculated from the networks and complexes alone. The last two steps are also used to determine the preliminary complexes for the bootstrap samples (see Figure 6.1 **B**). Details of the algorithm are described in the following.

6.2.1 Combination of experiments

While Gavin *et al.* (2006) used only one mass spectrometry method to identify proteins co-purified with a bait, Krogan *et al.* (2006) used two separate methods (MALDI-TOF and LC-MS/MS). Since mass spectrometry is one potential source of false positives and false negatives in the experiments, this approach increases the coverage and accuracy of the method. Proteins identified with both methods are highly confident, however the coverage is very low if only these proteins are considered (Krogan *et al.*, 2006). On the other hand, if all proteins identified by at least one method are used, the quality of the results is decreased.

As a trade-off between coverage and accuracy, we use the following approach. Prey proteins identified with both mass spectrometry methods are counted twice in a purification while proteins only identified with one method are counted only once. Thus, interactions between proteins identified with two methods will get a higher weight in the calculation of interaction scores described in the next section. Since Gavin *et al.* used only one mass spectrometry method, preys are only counted once for purifications from this experiment in the pooled set of both experiments.

6.2.2 Bootstrap sampling

To determine reliable confidence scores, the bootstrap technique (Efron, 1979; Efron and Tibshirani, 1994) is used to estimate how stable interactions are under perturbations of the data. A similar approach is utilized successfully for assigning confidence to phylogenetic trees (Felsenstein, 1985). In the following, let $\Phi = (\phi_1, \ldots, \phi_n)$ be the list of purifications and V the set of proteins. Each purification ϕ_i consists of one bait protein $b_i \in V$ and the preys $p_{i,1}, \ldots, p_{i,m} \in V$ identified for this bait in the purification ϕ_i:

$$\phi_i =< b_i, [p_{i,1}, \ldots, p_{i,m}] > . \tag{6.1}$$

From the list of purifications Φ, l bootstrap samples are created (in our case $l = 1,000$). Each bootstrap sample $S_j(\Phi)$ is created by drawing n purifications with replacement from

Φ. This means that the bootstrap sample $S_j(\Phi)$ contains the same number of purifications as Φ and each purification ϕ_i can be contained in $S_j(\Phi)$ once, multiple times or not at all. Multiple copies of the same purification are treated as separate purifications.

For each bootstrap sample $S_j(\Phi)$, we calculate preliminary interaction scores for each pair of proteins co-purified at least once. For this purpose, we use the socio-affinity scores described by Gavin *et al.* (2006) which compare the number of co-occurrences of two proteins against the random expectation using a combination of spoke and matrix model. No additional training data is required to compute them from a set of purifications.

The preliminary scoring network for each sample $S_j(\Phi)$ is then defined as $G_j = (V, E_j)$ with $E_j = \{(u, v) | u, v \in V \wedge w_{S_j(\Phi)}(u, v) \geq \tau_P\}$. Here, $w_{S_j(\Phi)}(u, v)$ is the socio-affinity score for the interaction between u and v in the bootstrap sample $S_j(\Phi)$ and τ_p a pre-defined cutoff which is applied to this network to filter for the most confident interactions and to allow for fast computation. From these preliminary networks, we then determine preliminary complex predictions for each individual bootstrap sample $S_j(\Phi)$ with the algorithm described in the following section.

6.2.3 Identification of protein complexes

The algorithm for the prediction of complexes from a network consists of two steps: clustering of the network and subsequent identification of shared proteins.

Clustering:

Networks are clustered using the Markov clustering algorithm (MCL) developed by van Dongen (2000). In a recent study (Brohee and van Helden, 2006), this algorithm was found to be superior to several other graph clustering algorithms for the identification of protein complexes. As a consequence, many approaches to complex identification from purification data use this method. The running-time of the MCL procedure is in $O(Nk^2)$ for a network with N nodes and a maximum degree of k. Thus, it is very fast for sparse networks. Its most important parameter is the *inflation parameter* which influences the granularity of the identified clusters, i.e. their size and number. The higher the inflation parameter, the smaller are the resulting clusters and the more clusters are identified.

All previous approaches based on MCL used additional training data in the form of known complexes to choose the optimal inflation parameter. We suggest to use an intrinsic measure which compares the resulting clusters against the original network from which the clusters were obtained. For this purpose, we use a performance measure for graph clustering proposed by van Dongen, the so-called *efficiency* (for details refer to van Dongen (2000)). Basically, a clustering is highly efficient if proteins in the same cluster are connected by edges with high weights and proteins in different clusters have no or only low weight connections.

To determine the optimal inflation parameter, we cluster the preliminary networks G_j for each sample $S_j(\Phi)$ with gradually increasing inflation parameters. For each inflation parameter we calculate the average efficiency over all samples. We found that for our

networks, the efficiency always reaches a maximum for a certain inflation parameter and decreases on either side of this value. Accordingly, the optimal inflation parameter is chosen as the one with the highest average efficiency across all samples. This inflation parameter is then used to cluster the preliminary networks for the bootstrap samples.

Identification of shared proteins:

The MCL algorithm, as most clustering methods, identifies only disjoint clusters. In real biological systems, however, proteins can be contained in more than one complex. If a protein has such strong associations with two complexes, the MCL procedure will either cluster those two complexes together or, if further associations between the complexes are missing, cluster this protein with only one of these complexes. We address this problem in a similar way as Pu *et al.* (2007) by post-processing the clusters obtained from the MCL algorithm. Contrary to them, we do not optimize this step based on proteins shared between known yeast complexes, but again use an intrinsic measure based on the scoring network.

The following criteria are used for adding shared proteins. First, a protein is only added to another complex if it has sufficiently strong interactions to this complex. Second, the interaction strength of the protein to the complex which is required to add the protein depends on the strength of interactions within the complex. Third, for large complexes strong interactions are only required to some of the complex proteins or, alternatively, weaker interactions to most of them. As a consequence, a protein p_i can be added to a complex C if

$$s(p_i, C) \geq \alpha \cdot s(C) \cdot \left(|C|^{-\gamma} / \, 2^{-\gamma} \right) \tag{6.2}$$

where $s(p_i, C)$ is the average interaction score of p_i to proteins of C and $s(C)$ the average interaction score within the complex. Interactions not contained in the network are given a weight of zero.

This threshold definition has two parameters, α and γ. Here, α defines how much weaker than $s(C)$ the connections to the complex are allowed to be and γ controls to which fraction of the complex the protein p_i has to have sufficiently strong interactions. The threshold decreases both with complex size and with complex confidence. To control the influence of complex size, a power function was chosen since it decreases steeply at first but then levels off for larger values. The power function is normalized to yield 1 for complexes of size 2. In this case the threshold depends only on the strength of the interaction between the two proteins.

Both parameters α and γ are set such that the weighted average score over all complexes after the post-processing is at least as high as a fraction λ of the original average score. For this purpose, α is set to λ and γ to the largest possible value such that this requirement is still met. This value of γ is identified with a binary search. As we only want to add proteins to complexes to which they are clearly associated, we set λ very high at 0.95. Proteins are added to complexes in parallel. Accordingly, the complex memberships and the average complex score are not updated until all proteins have been processed and the result does not depend on the order of the proteins.

6.2.4 Calculation of confidence scores and final complexes

The final confidence scores are then determined by calculating the so-called bootstrap network G_{BT} from the complexes identified for each bootstrap sample. In the bootstrap network, two proteins are connected by an edge if they are clustered together in at least one sample. The fraction of samples for which they are contained in the same complex provides the weight for the corresponding edge and the confidence for this association (between 0 and 1). Thus, the confidence score c_{BT} for an interaction (u, v) is defined as

$$c_{BT}(u,v) = \frac{1}{l}\sum_{j=1}^{l} \delta_{S_j(\Phi)}(u,v) \tag{6.3}$$

where $\delta_{S_j(\Phi)}(u,v)$ is 1 if u and v are in the same complex for bootstrap sample $S_j(\Phi)$ and 0 otherwise.

Final complexes are then obtained by applying the complex identification algorithm on this bootstrap network. For this purpose, the optimal inflation parameter determined in the previous step is chosen. No threshold is applied to the network before complex identification but the size of the network is limited by choosing a stringent cut-off τ_P for the preliminary socio-affinity networks.

More confident predictions can be obtained from the original complexes in the following way. First, all edges are removed from the network with weight lower than a threshold τ_C and then connected components are calculated for each complex separately in this restricted network. As a consequence, complexes can either shrink, be subdivided or be removed completely. This approach is more efficient than the alternative approach of first restricting the network and then repeating complex identification but yields almost identical complexes (see results).

6.2.5 Criteria for the evaluation of complex quality

Functional similarity within complexes:

Since protein complexes are formed to carry out a specific function, the function of proteins in the same complex should be relatively homogeneous. We evaluate the functional similarity between proteins predicted to be in the same complex by using the protein annotations of the Gene Ontology (GO) (Ashburner *et al.*, 2000) (see also section 5.2.2). The functional similarity of two proteins is quantified in terms of the *semantic similarity* of GO terms annotated to these proteins. Several variations of semantic similarity have been described (Resnik, 1999; Lin, 1998; Lord *et al.*, 2003; Schlicker *et al.*, 2006). Here, we use the relevance similarity proposed recently by Schlicker *et al.* (2006). This measure is based both on the closeness of two GO terms to their common ancestors as well as the level of detail of these ancestors.

The GO co-annotation score of a complex is the average relevance similarity of all protein pairs in this complex. The GO co-annotation score for a set of complexes is the

weighted mean over all complex scores and determined separately for the "biological process" and "molecular function" ontologies. The final co-annotation score is then calculated as the geometric mean of the "biological process" and "molecular function" GO scores.

Co-localization within complexes:

Since complexes can only be formed if the corresponding proteins are actually located together in the cell, a second quality measure is based on the similarity of protein localizations within a complex. For this purpose, we used the localization assignments and categories determined experimentally in yeast by Huh *et al.* (2003).

The *co-localization* score for a complex is defined as the maximum fraction of proteins in this complex which are found at the same localization. The average co-localization score is calculated as the weighted average over all complexes and is defined as

$$L = \frac{\sum_j \max_i l_{i,j}}{\sum_j |C_j|} \tag{6.4}$$

Here, $l_{i,j}$ is the number of proteins of complex C_j assigned to the localization group i and $|C_j|$ is the number of proteins in the complex C_j with localization assignments.

We do not calculate the co-localization score of a complex C_j as $\max_i l_{i,j} / \sum_i l_{i,j}$ as suggested by Pu *et al.* (2007). This method would give a score of 0.5 to a complex of two proteins where each protein is assigned to the same two co-localization categories. However, such a complex is perfectly co-localized and, accordingly, is given a value of 1 by our co-localization score.

Sensitivity and positive predictive value:

To evaluate the accuracy of the predictions, *sensitivity* (Sn) and *positive predictive value* (PPV) were calculated with regard to the following reference sets: (a) manually curated complexes from the MIPS database (Mewes *et al.*, 2004) (214 complexes after removing redundant complexes) and the study of Aloy *et al.* (2004) (101 complexes) as well as (b) complexes extracted from the SGD database (Dwight *et al.*, 2002) based on GO annotations (189 complexes). To compile the SGD set, GO-slim complex annotations to all yeast genes were taken from the SGD ftp site (`ftp://genome-ftp.stanford.edu/`).

We used the definition of sensitivity and PPV for protein complexes by Brohee and van Helden (2006). Both measures are calculated from the number $T_{i,j}$ of proteins shared between a reference complex R_i and a predicted complex C_j:

$$Sn = \frac{\sum_i \max_j T_{i,j}}{\sum_i |R_i|} \quad \text{and} \quad PPV = \frac{\sum_j \max_i T_{i,j}}{\sum_j \sum_i T_{i,j}}. \tag{6.5}$$

Thus, sensitivity evaluates which fraction of proteins in known complexes are recovered and positive predictive value determines how good the predicted complexes match to the known reference complexes. For the calculation of sensitivity only reference complexes

are considered for which $\max_j T_{i,j} \geq 1$. All described evaluation methods as well as the socio-affinity scores used for the preliminary networks were implemented by Jan Krumsiek as part of a student project.

6.3 Results

Bootstrap confidence values were calculated from the combined Krogan and Gavin purification experiments. This combined set contains 6498 purifications with 2995 distinct baits, more than half of which (1617) were purified more than once. On average, separate purifications of the same bait agree in about 28% of the retrieved preys between the two experiments. This is comparable to the agreement between purifications of the same bait within the Krogan data set (23%), but significantly smaller than within the Gavin set (49%).

From these purifications the final bootstrap network was calculated. We used a cut-off $\tau_P = 8$ on the socio-affinity scores to derive preliminary networks for each bootstrap sample. We chose a more stringent threshold than the one recommended by Gavin *et al.* (2006) for two reasons. First, the preliminary networks are much denser for the combined data than for the Gavin data alone and, as a consequence, the runtime of the MCL algorithm is increased considerably. Second, the final bootstrap network contains many more interactions than each individual preliminary network (in our case 20 times more). Thus, the more stringent threshold τ_P both reduces runtime of the bootstrapping step and at the same time limits the size of the resulting bootstrap network.

The final bootstrap network contains 62,876 interactions between 5195 distinct proteins. Because of this relatively small network size, no additional cut-off is necessary before predicting protein complexes with our approach. From the bootstrap network we predicted 893 complexes (denoted as BT-893) which contain 5187 distinct proteins (397 of those shared between complexes).

To compare our results against the smaller Pu and Hart predictions, more confident complexes were extracted from the original set with a threshold $\tau_C = 0.32$. This set contains 409 complexes with 1692 distinct proteins (101 shared between complexes) and will be referred to as BT-409 in the following. It is comparable in size with the Pu predictions of 400 complexes with 1914 distinct proteins (141 shared) and the Hart predictions of 390 complexes with 1689 distinct proteins (none shared). We also extracted a second set of 217 complexes (BT-217, 940 distinct proteins, 54 shared) at $\tau_C = 0.69$ which has a similar size as the MIPS complexes (214 complexes, 1190 distinct proteins, 119 shared). We compared our selection approach against the less efficient method of first restricting the network and clustering afterwards and found that the differences observed were negligible with sensitivities in both directions of about 0.97.

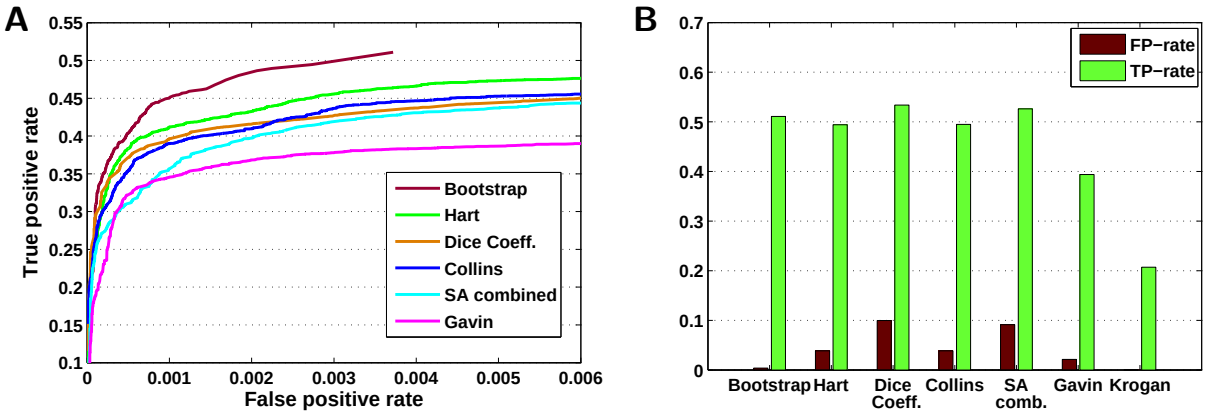


Figure 6.2: Accuracy of the interactions predicted with the bootstrap approach. **A**) ROC curves for the bootstrap, Hart, Dice coefficient and Collins scores, the socio-affinity scores for the combined data (SA combined) and the Gavin purifications alone (from top to bottom). Krogan scores are omitted since they performed significantly worse than any of the other scoring methods. True positive rates on the y coordinate are plotted against false positive rates on the x coordinate for gradually decreasing thresholds to predict an interaction from the scoring networks. **B**) Maximum false positive and true positive rate for each scoring network.

6.3.1 Evaluation of interaction networks

The quality of the bootstrap network in comparison to other suggested interaction scores was evaluated using a *receiver operating characteristic* (ROC) curve (Fawcett, 2006). In a ROC curve, true positive rates are plotted against false positive rates for gradually decreasing thresholds for predicting an interaction. True positive interactions were defined as interactions between proteins in the same MIPS complex. The large and small ribosomal subunits were excluded since they would otherwise make up 44% of the true positive interactions. True negative interactions were defined as interactions between proteins assigned to different MIPS complexes and cellular localizations by Huh *et al.* (2003).

Figure 6.2 **A** shows the resulting ROC curves for the Gavin, Krogan, Collins, Hart and bootstrap scores as well as socio-affinity scores and Dice coefficients (Zhang *et al.*, 2008) calculated from the combined experiments with the ProCope software package (Krumsiek *et al.*, 2008) (see section 6.3.7). We see that for all networks calculated from the combined data, the curve is generally steeper and reaches a higher level than for the scores calculated from each experiment alone. Furthermore, the bootstrap scores calculated with our method performed best at separating true interactions from noise. Among the scoring methods proposed after the publication of the original purification studies, the Collins scores and Dice coefficients performed worst. Nevertheless, they still performed slightly better for the given range than the socio-affinity scores computed from the combined experiments.

Figure 6.2 **B** illustrates the maximum possible true positive rate and corresponding false positive rate if all interactions in the scoring networks are predicted. Here, we see

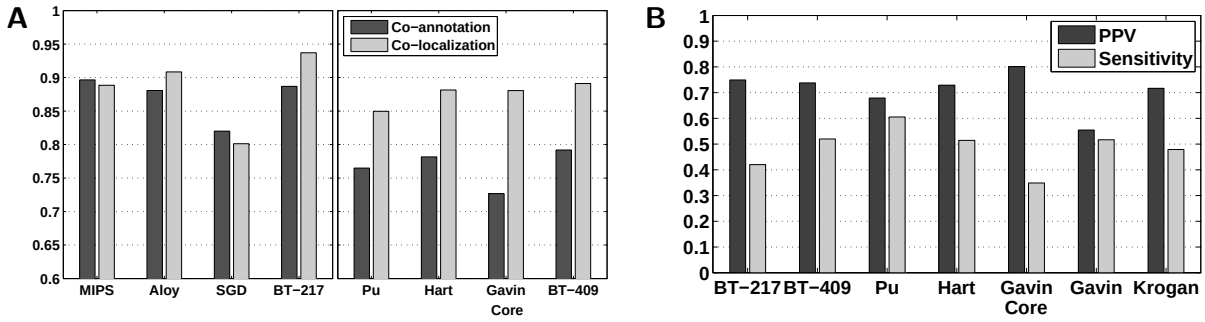


Figure 6.3: Functional and localization similarity and predictive accuracy of complexes. **A**) Co-annotation (dark grey) and co-localization (light grey) scores for the complexes taken from the MIPS and SGD databases and the publication by Aloy *et al.* (2004) as well as the highly confident BT-217 complexes on the left hand side and the Pu, Hart, Gavin core and BT-409 predictions on the right hand side. **B**) PPV (dark grey) and sensitivity (light grey) for the BT-217, BT-409, Pu, Hart, Gavin core, Gavin and Krogan complexes.

that only the socio-affinity scores and Dice coefficients can obtain a slightly higher true positive rate than the bootstrap confidence scores. However, this results in false positive rates 25 times as high as for the bootstrap confidence scores and can only be obtained by including interactions with very low scores. For all other scoring networks, the true positive rate of the bootstrap network cannot be reached, not even for high false positive rates.

6.3.2 Functional and localization similarity within complexes

To assess the quality of the predicted complexes, we calculated the co-annotation and co-localization scores for the MIPS, Aloy and SGD complexes as well as for the Pu, Hart, Gavin and Krogan predictions and the BT-409 and BT-217 complexes (see Figure 6.3 **A**). Furthermore, the Gavin core set was evaluated which contains only the core components defined by Gavin *et al.* (2006). In this analysis, we focused only on the predictions made on the combined set of purifications as well as the predictions from the original publications.

The lowest functional and localization similarity is observed for the Gavin and Krogan complexes (data not shown). By restricting to the more confident core components in the Gavin predictions, both co-annotation and co-localization can be increased significantly by 17% and 25%, respectively. Among all previously published prediction approaches, the highest co-annotation scores are obtained by the Pu and Hart predictions and the highest co-localization scores by the Hart predictions and the Gavin core set. In the MIPS and Aloy reference complexes functional and localization similarity is significantly higher. Among the reference complexes, the SGD complexes perform worst. While co-annotation is still higher than in the best predictions, co-localization is significantly lower.

When evaluating the complexes identified by our approach, we find that the BT-409 complexes perform significantly better than the Pu and Gavin core complexes with re-

gard to functional and localization similarity and slightly better than the Hart complexes. Moreover, the highly confident BT-217 complexes show similar co-annotation and higher co-localization scores than the manually curated MIPS and Aloy complexes. It should be noted though that a large fraction of the BT-217 complexes is already well-known as 64% of these complexes share at least two proteins with one of the reference complexes. In the BT-409 set, this applies only to 43% of the complexes.

6.3.3 Validation on reference complexes

By comparing the predicted complexes against the current knowledge on protein complexes in the form of the MIPS, Aloy and SGD reference complexes, the sensitivity and the PPV of the corresponding methods can be estimated. One should keep in mind, though, that these estimates may be unprecise due to the incompleteness of current knowledge.

Results for the comparison against the MIPS complexes are shown in Figure 6.3 **B**. Similar trends are observed for the comparison against all reference sets. The worst results are obtained by the original Gavin complexes, followed by the Krogan complexes. Here, the Gavin complexes are generally more sensitive but less accurate in their predictions than the Krogan complexes. By restricting the Gavin complexes to the core components, the PPV can be increased beyond that of any other prediction. However, this improvement comes at the cost of a very low sensitivity.

When comparing the performance of the BT-409, Pu and Hart complexes, we observe that none of the predictions is clearly superior to the other two. Although the sensitivity of the Pu complexes is slightly higher than for the other two approaches, the corresponding PPV is lower in return. Thus, it appears that all predictions cover the reference complexes with similar quality. The PPV of the BT-409 complexes can be increased slightly by restricting to the more confident BT-217 complexes, however the loss of sensitivity again is considerable.

6.3.4 Assessing predictions from the Gavin data alone

Since the recently published methods by Rungsarityotin *et al.* (2007) and Zhang *et al.* (2008) have only been applied to the Gavin data, we also applied our bootstrapping algorithm to this data set alone to compare it against these two prediction approaches. Due to the smaller size of the Gavin data set, we used a slightly lower τ_P of 7 on the preliminary networks to obtain 381 complexes with 1970 distinct proteins.

We compared our results against the original Gavin predictions, the Markov random field (MRF) predictions by Rungsarityotin *et al.* (2007) both from the matrix and spoke model and the predictions by Zhang *et al.* (2008). From the MRF matrix prediction we also selected the 381 complex predictions with highest quality score and for the Zhang *et al.* predictions we distinguished between complete and core predictions.

We found that functional similarity and co-localization (Figure 6.4) in the bootstrap complexes is higher than both in the complete and high-confidence MRF and Zhang predictions. PPV is slightly higher in the MRF and Zhang core predictions, but on the other

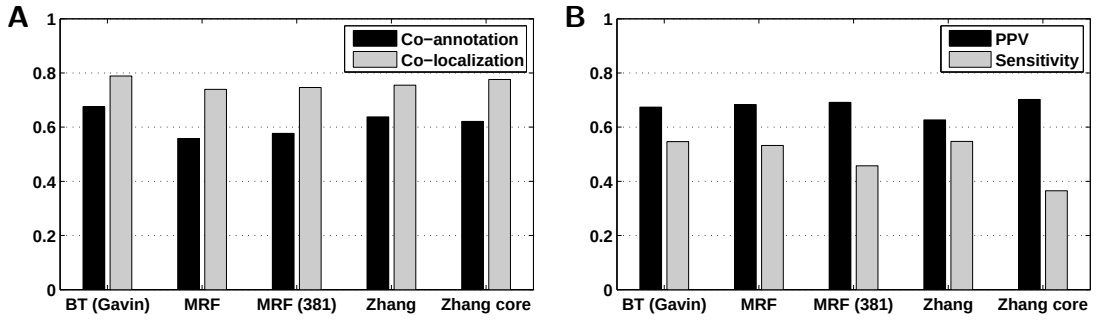


Figure 6.4: Functional and localization similarity (**A**) and positive predictive value and sensitivity (**B**) for the complex predictions from the Gavin data alone. Results are shown for the bootstrap (BT) predictions on the Gavin set, the Markov random field (MRF) predictions of Rungsarityotin *et al.* (2007) (for the matrix model, results for the spoke model are similar), the 381 best MRF matrix predictions and the predictions by Zhang *et al.* (2008) both for the complete and core set.

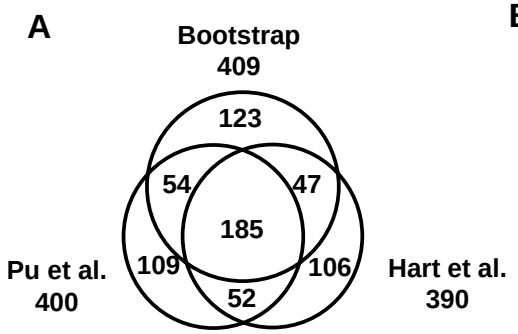
hand sensitivity is lower than in the bootstrap predictions. This shows that the complexes identified with our approach are of slightly higher overall quality than complexes predicted with these more recently published methods.

The bootstrap predictions on the Gavin data have significantly higher functional and localization similarity and predictive performance than the complete Gavin complexes. To compare our predictions against the high-confidence Gavin core predictions, we extracted a more confident set of 210 complexes with 1127 distinct proteins using a τ_C of 0.32 from our complete predictions. Although this complex set contains fewer complexes than the Gavin core predictions (423 complexes with 1128 proteins), the number of proteins is almost identical. We found that localization similarity is very similar between both predictions ($\sim$ 0.88), whereas functional similarity is significantly higher in the bootstrap complexes (bootstrap: 0.81, Gavin core: 0.73). Furthermore, the Gavin core predictions obtain a higher PPV (bootstrap: 0.7, Gavin core: 0.8) than the bootstrap complexes but only by having a significantly lower sensitivity (bootstrap: 0.46, Gavin core: 0.35).

6.3.5 Towards a consensus of complex predictions

Although functional and localization similarity within complexes is slightly higher for the BT-409 predictions than for the Pu and Hart predictions, the comparison against the reference complexes yielded very similar results for all three sets. In order to appreciate how much these predictions agree or diverge, we compared them at the level of the individual complexes.

First, we calculated PPV and sensitivity values in both directions for each possible pairwise combination of prediction methods. Here, we observed an average PPV of 0.85 and sensitivity of 0.72. This suggests that the agreement between each pair of these new predictions is much higher than between the Gavin and Krogan complexes for which we



B

	Bootstrap	Pu et al.	Hart et al.
Bootstrap		71.0%	77.8%
Pu et al.	85.9%		65.4%
Hart et al.	82.5%	81.1%	

Figure 6.5: Consensus between best supervised predictions and the bootstrap approach. (**A**) Venn diagram of the number of complexes which can be assigned consistently between the BT-409, Pu and Hart complexes. **B**) Fraction of one-to-many mappings among inconsistent predictions for each pairwise comparison (above the diagonal) and the average protein overlap between inconsistent protein complexes (below the diagonal).

observe an average PPV of 0.26 and sensitivity of 0.29.

In a second step, we calculated for each pair of prediction methods the number of complexes with (a) no significant overlap (at least 2 proteins) to the other set, (b) a significant overlap to exactly one complex in the other set which again has no other overlaps (consistent complexes) and (c) significant overlaps but without a one-to-one correspondence as in (b) (inconsistent complexes). In the second case, we also distinguished between complexes with an exactly matching counterpart and complexes which contain additional proteins in at least one of the predictions. The same analysis was also performed for all three sets together.

This analysis showed that about 25% of complexes in the pairwise comparisons of the BT-409, Pu and Hart complexes and 16% in the comparison of all three predictions are supported by only one method. This is much lower than observed between the Gavin and Krogan complexes, where 42% and 64% of the complexes, respectively, have no significant overlap to the other set

For the consistent complexes, results are shown in Figure 6.5 **A**. For more than half of complexes in this group the predictions agree exactly. In the remaining cases, 28-34% additional proteins are added by each method to the common core identified by all two or three predictions (see Figure 6.8 **B** for an example). Here, the consensus of each pair of predictions is much higher than for all predictions taken together. Nevertheless, even in the latter case the consensus is still larger with 185 complexes (46%) than between the Gavin and Krogan complexes where only 45 complexes ($< 10\%$) have a clear one-to-one correspondence between the predictions.

For the inconsistent complexes, we observe a one-to-many relationship in pairwise comparisons in $\sim$ 71% of cases. This means that a complex predicted by one method is subdivided into several complexes by the respective other method. Furthermore, inconsistent complexes agree in about 83% of the proteins between complex predictions. This

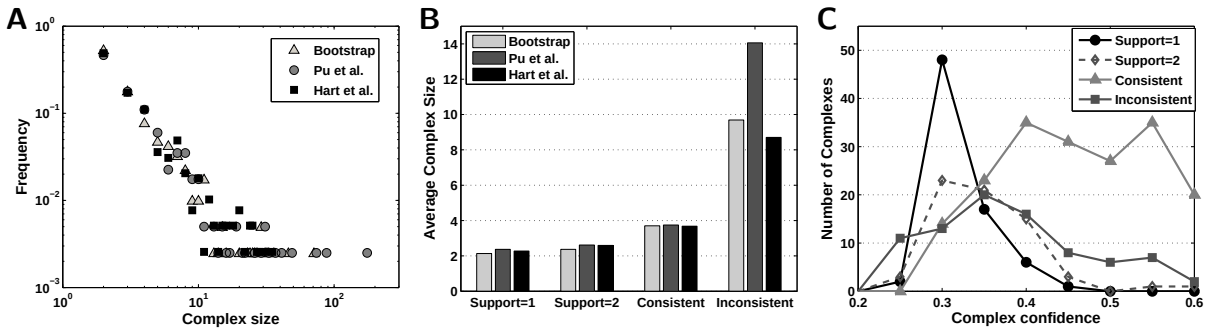


Figure 6.6: Size and confidence of predicted protein complexes. **A**) Distribution of complex size for the bootstrap, Pu and Hart predictions. **B**) Average complex size for all methods. **C**) Distribution of complex confidence scores for the bootstrap complexes (results for the Pu and Hart predictions show similar tendencies). Here, we distinguished between protein complexes which are supported by no other method (support = 1), by only one other method (support = 2) or can be mapped consistently or inconsistently between the bootstrap, Pu and Hart predictions.

indicates that although predictions identify approximately the same sets of proteins, they disagree with regard to the partitioning (see Figure 6.8 **C** for an example).

When we analyze the distribution of complex sizes for each prediction method (see Figure 6.6 **A**), we observe that it follows approximately a power-law distribution with $\sim 50\%$ of the complexes consisting of only two proteins and only 6-7% of complexes having a size larger than 10 proteins. Thus, the majority of protein complexes are heterodimers and only in few cases do we observe very large multi-subunit structures.

We also evaluated average complex size separately for complexes supported by only one or two methods or identified consistently or inconsistently by all methods (see Figure 6.6 **B**). Our results showed that complexes supported by only one method have very small sizes with 75-87% of complexes consisting only of an interaction of two proteins. Complexes supported by two methods show only slightly larger sizes. Medium complex sizes are observed for consistent complexes, where more than 50% contain more than 2 proteins. The inconsistent complexes generally are largest and 58-70% of those involve at least 5 proteins. We also found that average complex scores (see Figure 6.6 **C**) were smallest for complexes supported by only 1 or 2 methods and largest in complexes identified consistently by all complexes. For the inconsistent complexes only medium complex scores are observed.

For each group of complexes (support 1 or 2, consistent or inconsistent in a comparison of all predictions), we calculated functional similarity and co-localization separately (see Figure 6.7). These results show that complexes identified consistently by all methods have the highest functional similarity and co-localization, while complexes supported by only one method generally have the worst. Apart from these overall tendencies, striking differences can be observed between the methods. Complexes predicted only by the Hart method perform significantly better than complexes predicted only by either the bootstrap or the

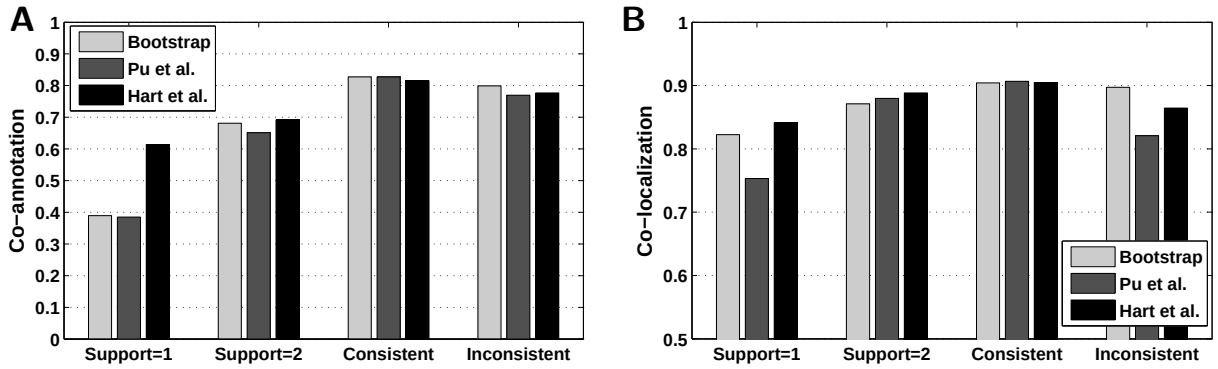


Figure 6.7: Comparison of complexes identified consistently and inconsistently by different methods. Average co-annotation (**A**) and co-localization scores (**B**) are shown for the four groups described in Figure 6.6.

Pu approach. On the other hand, the Hart predictions for the inconsistent complexes are of lower overall quality than the bootstrap predictions and the consistent complexes show lower functional similarities than both bootstrap and Pu predictions. On average, the Pu complexes perform worst across all groups.

6.3.6 Comparison of example complexes

We compared the BT-409, Pu and Hart predictions on example complexes which are supported by only two methods or identified consistently or inconsistently by all three methods. Figure 6.8 **A** shows a complex which is only found in the BT-409 and Hart predictions. Both predictions cluster the alpha and beta subunits of phosphofructokinase together. Furthermore, each approach predicts one additional protein of unrelated function. These additional proteins are probably false positives and, accordingly, the interactions to the additional proteins are scored significantly lower by both methods than the interaction between the two phosphofructokinase subunits.

Results for the SET3 histone deacetylase complex identified consistently by all approaches are shown in Figure 6.8 **B**. Here, the Hart approach predicts five additional proteins of unrelated or unknown function, but fails to identify other components of the histone deacetylase complex: CPR1 and HST1 which are both identified by the bootstrap and Pu approach. However, HST1 is only found in the BT-893 but not the BT-409 set. It has been shown that HST1 is only a non-essential subunit of this complex but an essential subunit of another complex (Pijnappel *et al.*, 2001). This may explain why the Hart approach does not identify this protein and why interaction scores of HST1 to this complex are quite weak both in the bootstrap and Pu predictions.

Figure 6.8 **C** shows an example for an inconsistently identified complex. Here, the bootstrap approach predicts three complexes: the PAF1 complex plus one protein of the FACT complex, the casein kinase CK2 complex with associated proteins as well as a smaller

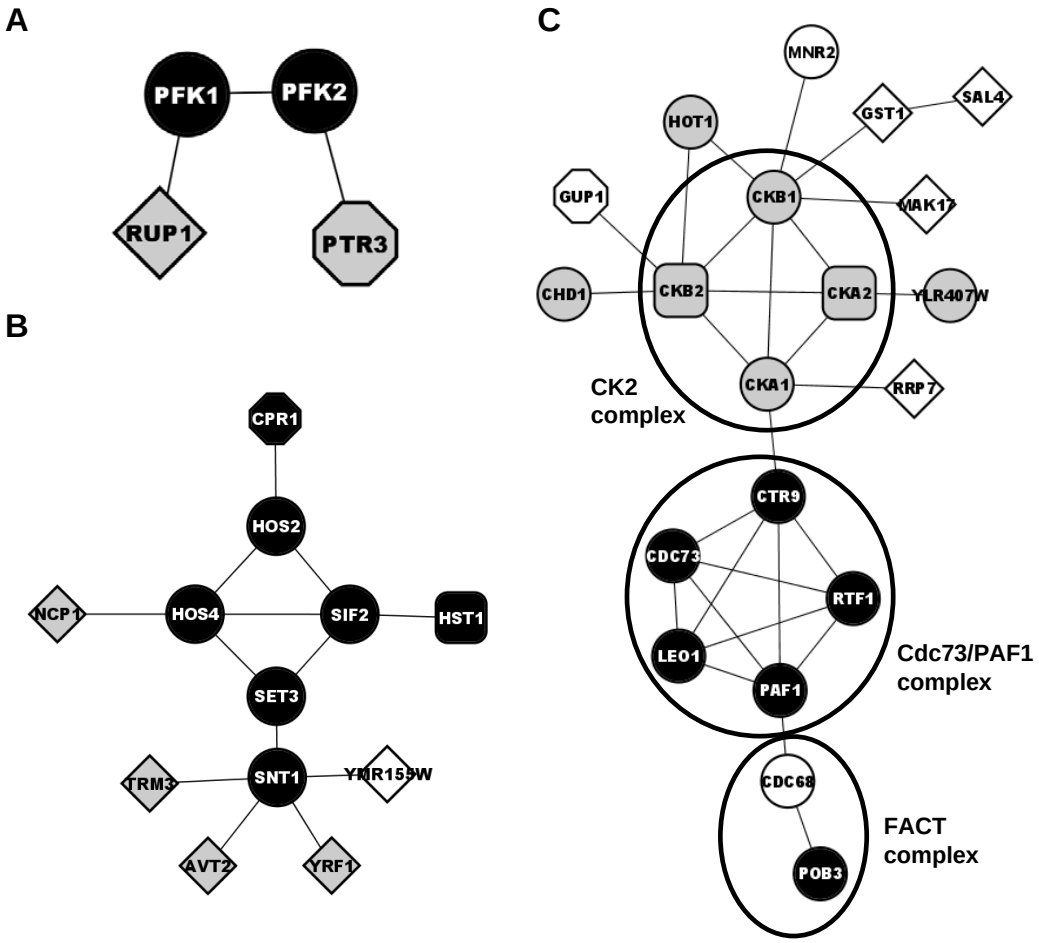


Figure 6.8: Comparison of BT-409, Pu and Hart predictions for example complexes. Figure **A** shows the phosphofructokinase complex (black) predicted only by the bootstrap and Hart approach. Predictions unique to the bootstrap and Hart complexes are indicated by octagons and rotated squares, respectively. Grey and white denote unrelated and unknown functions, respectively. Figure **B** shows the SET3 histone deacetylase complex (black) predicted by all three approaches. The following symbols are used: circles for proteins predicted by all approaches, octagons for predictions common to the bootstrap and Pu predictions, rounded rectangles and rotated squares for unique predictions by Pu *et al.* and Hart *et al.*, respectively. Colors are as in **A**. Figure **C** shows a comparison between the bootstrap, Pu and Hart predictions for the protein kinase CK2 complex and the PAF1 complex. Here, three complexes are predicted by the bootstrap approach (grey circles and rounded rectangles, black circles and grey rounded rectangles) while both Pu *et al.* and Hart *et al.* predict only one complex. Proteins unique to the bootstrap, Pu and Hart predictions are indicated by white octagons, circles and rotated squares, respectively. Important interactions were extracted with the maximum spanning tree approach described in chapter 7. Actual biological complexes found in the MIPS data set or in the literature are circled.

subcomponent of the CK2 complex consisting of one alpha and one beta subunit. The Pu and Hart methods group these complexes together into one complex. It has been shown that the CK2 complex associates with the PAF1 and FACT complexes during transcription (Krogan *et al.*, 2002). However, since the CK2 complex is involved in many more biological processes (Ahmed *et al.*, 2002), this association is not permanent which is reflected in the bootstrap predictions. The additional proteins added to the CK2 complex might be some of the many targets of the CK2 protein kinase (Ahmed *et al.*, 2002).

6.3.7 ProCope

The bootstrap algorithm as well as other methods for complex predictions are provided in the ProCope software package (Krumsiek *et al.* 2008, `http://www.bio.ifi.lmu.de/Complexes/ProCope/`). ProCope was developed together with Jan Krumsiek who implemented it as a student project. It provides implementations for socio-affinity scores (Gavin *et al.*, 2006), the purification enrichment scores of Collins *et al.* (2007), the scores developed by Hart *et al.* (2007) based on the hypergeometrical distribution, Dice coefficients (Zhang *et al.*, 2008) and our bootstrap confidence scores. For the clustering of scoring networks, ProCope provides access to the Markov Clustering algorithm (van Dongen, 2000) and implements several variants of hierarchical agglomerative clustering (Murthag, 1984). Proteins involved in more than one complex can be identified with our approach (see section 6.2.3) or the alternative method of Pu *et al.* (2007).

Furthermore, the evaluation methods described in section 6.2.5 are provided. The accuracy of scoring networks in predicting interactions within reference complexes can be assessed using receiver operating characteristic (ROC) curves. Predicted complexes can be evaluated by calculating sensitivity and positive predictive value compared to reference complexes (Brohee and van Helden, 2006) or functional similarity and co-localization within protein complexes. For evaluation of functional similarity, several alternative implementations of semantic similarity are included (Schlicker *et al.*, 2006). For evaluation of co-localization, ProCope implements our co-localization score as well as the alternative definition of Pu *et al.* (2007).

All methods have been tested extensively against the results of the original publications. Furthermore, ProCope has been designed to be easily extensible as well as highly efficient to allow for higher-order algorithms which require many repeated calculations such as the bootstrap algorithm.

The ProCope package provides a convenient graphical user interface (GUI) which can also be used as a Cytoscape plugin (Shannon *et al.*, 2003), command line tools suitable for batch job processing and a Java application programming interface (API). In the GUI, the user can quickly load purification data, calculate and cluster scoring networks, load evaluation data, compare complex sets, apply cut-offs on score networks or complex sets and much more. A Java Webstart version of the GUI is also available to start the program directly from within the webbrowser without the need for installing the package. All functionalities of ProCope can be accessed via a well-documented Java application programming interface (API) and integrated easily into new software programs. Furthermore,

the user interfaces of ProCope offer plugin functionalities to extend them by custom score calculation and clustering methods.

The potentials of ProCope can be illustrated with two examples. First, biologists can rapidly analyze data from new affinity purification experiments using our bootstrap approach or other previously published methods or a combination of those. Thus, time-consuming and error-prone reimplementations are avoided. Second, the existing functionalities of ProCope can be easily extended using the Java API. Researchers developing new prediction methods can rely on the infrastructure and evaluation methods provided by ProCope. New interaction score definitions or complex predictions methods can be implemented quickly by extending appropriate classes and the performance of the methods can be assessed without delay. In this way, the bootstrap algorithm can be modified by using other scoring methods than the socio-affinity scores or other clustering algorithms than the MCL algorithm. Moreover, new powerful approaches can be made available to the research community immediately using the plugin option of the user interfaces. Because of the easy access to its methods and the efficiency and extensibility of its algorithms, the ProCope package will be a valueable tool for researchers seeking to apply existing algorithms to new data or developing new and innovative prediction methods.

6.4 Discussion

In this chapter, we presented an algorithm for the prediction of protein complexes from purification experiments. It implements all necessary steps from the combination of different experiments up to the identification of shared proteins in an unsupervised manner. Accordingly, it does not depend on the availability of additional information on protein complexes and interactions and is not limited to organisms for which such an extensive knowledge exists. Therefore, our method can be applied to large-scale TAP experiments for any organism.

Intuitive and accurate confidence scores for protein interactions were obtained by application of the bootstrap technique. For this purpose, our complex identification method was applied to preliminary networks calculated from bootstrap samples to estimate the stability of interactions. The resulting confidence scores distinguished better between correct and wrong interactions than all other published scoring methods, in particular also better than the socio-affinity scores used for the preliminary networks.

The same complex identification method was then applied to the complete bootstrap network to yield a large set of complex predictions. From this large set, we extracted approximately the same number of high-confidence complexes (BT-409) as the so far best methods by Pu *et al.* (2007) and Hart *et al.* (2007). The comparison of functional and localization similarity within complexes showed slightly better results for the BT-409 complexes compared to the Pu and Hart predictions. Furthermore, the predictive performance with regard to known reference complexes proved to be equivalent. This suggests that meaningful complexes can be derived from the purification experiments without additional training data.

When analyzing the individual BT-409, Pu and Hart complexes, we found that about 60% of the complexes have a one-to-one correspondence in pairwise comparisons. Here, each prediction shows approximately the same agreement with each of the other predictions. When combining all three predictions, the fraction of complexes identified consistently drops to 46%. This shows, that the consensus between each pair of predictions is larger than between all three of them. Nevertheless, the degree of agreement is still significantly higher than observed between the original Gavin and Krogan predictions.

In general, complexes in the consensus set are assigned higher confidence values by each method than complexes not supported by all methods or identified inconsistently. This suggests that evidence in the experimental data for these consistently identified complexes is relatively clear while evidence for other complexes is more ambiguous. Such ambiguous information may be observed due to measurement errors but also if complexes are not permanently associated or connected by weak interactions.

Since low confidence scores indicate more unreliable complexes, the confidence of complexes should be taken into account for any analysis based on these protein complexes. However, since the more confident complexes tend to be already covered to a large degree by existing biological knowledge, new information may be found preferentially in the less confident ones. Thus, these should not be discarded per se but validated in additional experiments. Here, the original large set of complexes (BT-893) identified in this study can be a valuable resource for biological hypothesis generation and testing.

Since predictions of the bootstrap, Pu and Hart approaches differ significantly even though the overall quality is very similar, we analyzed in detail the quality of predictions for different categories of complexes. We found that the Hart predictions supported only by this method are of significantly higher quality than predictions supported only by the bootstrap or Pu algorithms. On the other hand, for complexes which are identified by all approaches but in slightly different compositions or partitions, the bootstrap algorithm performs better in identifying subunits of those complexes or the correct subdivison. Our results show that a comparison of predictions of different algorithms may lead to insights on essential or non-essential subunits of complexes and into interactions between complexes.

6.5 Conclusions

In this study, we have shown that highly accurate protein complexes and confidence scores can be obtained from tandem affinity purification results alone. As a consequence, the limiting factor in identifying protein complexes on a large scale for many species are experimental problems in transferring the TAP system from yeast to higher, multi-cellular eukaryotes. So far, only pilot studies have been published (Bürckstümmer *et al.*, 2006; Gloeckner *et al.*, 2007; Gregan *et al.*, 2007) which showed that mammalian complexes can be purified using variants of the TAP system. These results indicate that eventually large-scale TAP screens will be performed for a range of higher eukaryotes. Here, the bootstrap algorithm and the evaluation methods implemented in the ProCope library will prove to be highly valuable and will allow a reliable and rapid analysis of these new experiments.

Chapter 7

Identifying the topology of protein complexes

7.1 Introduction

In the previous chapter, we presented the bootstrap algorithm for predicting protein complexes from tandem affinity assays. As for most complex prediction approaches, complexes were predicted as sets of associated proteins and the substructure of the complexes or the physical interactions within the complexes was not considered. To analyze the individual complexes in more detail, we developed an approach to extract the physical interactions from the scoring networks for each complex and to identify the subcomponent structure within the protein complexes.

So far, few computational methods have been developed for analyzing the substructure of protein complexes. Aloy *et al.* (2004) used homology modeling and electron microscopy to at least partially resolve interactions between subunits of 54 experimentally derived complexes. The method of Hollunder *et al.* (2005, 2007a,b) identifies subsets of proteins which occur more frequently in different complexes than expected at random. As a consequence, this approach can only identify subcomplexes which occur in more than one complex. Gavin *et al.* (2006) distinguished between core elements and modules or attachments in their protein complex predictions but did not predict direct interactions.

Scholtens *et al.* (2005) and Bernard *et al.* (2007) developed approaches to model the physical topology of protein complexes from affinity purification results as well as physical interactions from yeast two-hybrid (Y2H) experiments. However, Scholtens *et al.* used this only as an intermediate step in predicting protein complexes and did not evaluate the actual interactions they predicted. Bernard *et al.* showed that accurate predictions can be obtained by applying their approach to combined affinity purification and Y2H results but did not evaluate to what degree their results depend on the Y2H interactions used additionally.

Here, we investigated whether the topology of protein complexes can be predicted from the affinity purification results alone (Friedel and Zimmer, 2008). The topology of a protein

complex describes both the direct, physical interactions within the complex (the complex scaffold) and its hierarchical substructure, i.e. the subdivision of the complex into smaller components. Since our bootstrap algorithm and most other methods for predicting protein complexes from affinity purification results calculate interaction scores as an intermediate step, we developed a method to extract the complex scaffolds from these densely connected scoring networks.

Our algorithm calculates the union of all maximum spanning trees from the interaction scores for each complex. The maximum spanning trees are then extended by interactions which are not accounted for by indirect interactions via other proteins. We applied our method to confidence scores and protein complexes calculated from the yeast affinity purification experiments of Gavin *et al.* (2006) and Krogan *et al.* (2006) with the bootstrap method described in chapter 6. We show that the interactions predicted by our approach are enriched for direct, physical interactions. Furthermore, the distance in the resulting network reflects the similarity of the subcomponent localization of the corresponding proteins and the substructure of the protein complexes can be resolved in a straightforward way.

7.2 Methods

In the following, let $C = \{C_1, \ldots, C_n\}$ be a set of protein complexes with C_i a set of proteins and $G = (V, E)$ a weighted network of interaction scores. Here, V is the set of all proteins and E the set of all interactions between them. We assume that all scores are confidence values in the range of 0 to 1. The function $w : E \rightarrow [0, 1]$ defines the weight, i.e. the confidence score, of each edge. Interactions not contained in the network are given a weight of 0. If the scoring method calculates general scores from $-\infty$ (or 0) to ∞, edge weights are scaled to $[0, 1]$.

We assume that each complex is connected in the network of actual physical interactions. This means that each protein can be reached from every other protein in the same complex by an indirect path of direct, physical interactions. This network of direct interactions is denoted as the scaffold of the complex and predictions are performed separately for each complex.

7.2.1 Maximum spanning trees

Our algorithm takes as input the network of interaction scores for all protein pairs in each complex. From these interaction networks, we predict interactions for the complex scaffolds by calculating maximum spanning trees. A spanning tree is a tree which connects all vertices in the network. The maximum spanning tree (MST) is the spanning tree which maximizes the sum of its edge weights. It can be calculated efficiently using the Kruskal or Prim algorithm (Kruskal, 1956; Prim, 1957; Cormen *et al.*, 2000). As several different MSTs can exist in a network, we determine the union of all possible MSTs to predict the set of direct interactions in the complex scaffold.

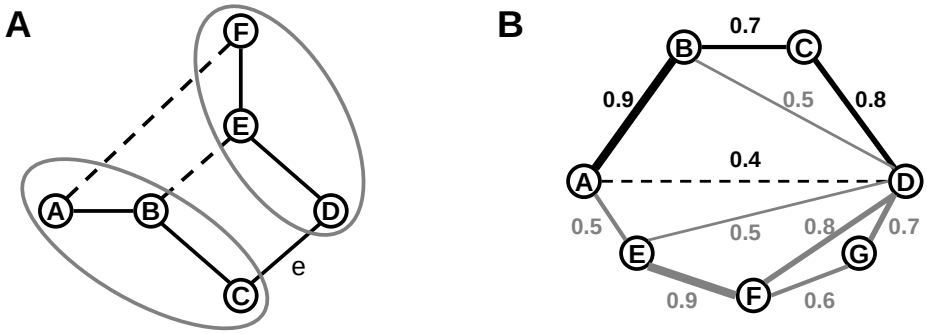


Figure 7.1: Methods for predicting physical interactions in protein complexes. Figure **A** outlines how interactions contained in at least one MST are identified. The solid lines show the MST T calculated first with the Kruskal or Prim algorithm. By removing edge e, a cut into the two sets $\{A, B, C\}$ and $\{D, E, F\}$ is created (grey ellipses). Dashed lines indicate edges crossing that cut with the same weight as e. By replacing e with any of these edges another MST T' is created. Thus, all of these edges are contained in at least one MST and added to the predicted scaffold. Figure **B** illustrates how MSTs are extended (for $\alpha = 1$). The current scaffold is indicated by solid lines. To determine if the interaction between A and D (dashed line) should be added to the network, we first find the optimal shortest path between the two nodes. In this case, this is $A \rightarrow B \rightarrow C \rightarrow D$ (black) which has a weight of $0.9 \cdot 0.7 \cdot 0.8 = 0.504$. Since the weight of edge (A, D) is smaller than 0.504, the edge is discarded. If the weight of (A, D) were larger, it would be added to the scaffold network.

To calculate all interactions contained in at least one MST, we do not have to compute all possible MSTs, but can find the relevant interactions starting from one arbitrary MST (see Figure 7.1 **A**). Deleting each edge e in turn from this MST yields a cut $Cut(T, e)$ – a partition into two sets – of the proteins in the complex. All edges crossing that cut, i.e. connecting proteins not in the same set, with the same weight as e are contained in at least one MST. Thus, the algorithm for predicting the MST scaffold consists of two steps. First, an MST is calculated either with the Kruskal or Prim algorithm. Second, all other edges contained in an MST are identified with the approach described above. With this algorithm, all edges in the union of all MSTs can be identified.

This can be shown in the following way. Let f be an edge in an arbitrary MST T' and T be the MST we start from. We identify f with our algorithm if there exists at least one edge $e \in T$ with $w(e) = w(f)$ such that f crosses the cut $Cut(T, e)$.

Adding f to T leads to a unique cycle $Cycle(T, f)$ and f crosses every cut induced by deleting any other edge e on this cycle from T. Thus, we only need to show that there exists at least one other edge e on that cycle with the same weight as f.

Lemma 7.1. $\forall e \in Cycle(T, f) \setminus \{f\}$, $w(e) \geq w(f)$.

Proof. By contradiction. Assume $\exists e \in Cycle(T, f) \setminus \{f\}$ with $w(e) < w(f)$. $T^* = T \setminus \{e\} \cup \{f\}$ is also a spanning tree since it is connected and contains no cycles. Then

$w(T^*) = w(T) - w(e) + w(f) > w(T)$. This is a contradiction to the assumption that T is an MST. □

Lemma 7.2. *$\exists e \in Cycle(T, f) \setminus \{f\}$ such that $w(e) = w(f)$.*

Proof. By contradiction. Assume that no such edge exists. From lemma 7.1 we then know that $\forall e \in Cycle(T, f) \setminus \{f\}$, $w(e) > w(f)$. Let $Cut(T', f)$ be the cut created by deleting f from T'. There exists an edge $e \in Cycle(T, f) \setminus \{f\}$ which crosses that cut. Thus, $T' \setminus \{f\} \cup \{e\}$ is a spanning tree with weight $w(T') - w(f) + w(e) > w(T')$. This is a contradiction to the assumption that T' is a maximum spanning tree. □

7.2.2 Extending the maximum spanning trees

Although the combination of all MSTs is no longer necessarily a tree, the resulting networks are extremely sparse and many physical protein interactions are missed. As a consequence, we extend this network by interactions which cannot be explained by an indirect interaction via other proteins in the MST scaffold. For this purpose, we compare an interaction (u, v) in the original network to the best indirect interaction between u and v in the current scaffold network. If the edge weight is at least as high as a factor α times the weight of the best indirect interaction, the interaction is added to the MST network. The resulting network is denoted as $eMST_\alpha$ and the parameter α tunes the density of the resulting scaffold network. Generally, α is set to 1.

For calculating the best indirect interaction between u and v we use the fact that all edge weights are confidence values in $[0, 1]$. As a consequence, the weight of an edge is interpreted as the probability that this edge is a physical interaction. The probability of a path P is calculated as the product of the edge probabilities on this path. Here, we assume independence between the edge probabilities. The weight of an indirect interaction between two proteins u and v is then defined as the maximum probability of any path

Algorithm: MST extension

- Calculate MST scaffold S
- Sort remaining edges
- $\forall\, e = (u, v) \notin S$ in non-increasing order
  - Find optimal path P in S from u to v
  - If $w(e) \geq \alpha w(P)$
    $\Rightarrow$ Add e to scaffold

Figure 7.2: Algorithm for calculating the extended MST networks ($eMST_\alpha$).

between them in the current scaffold (without the edge (u, v)). By taking the absolute values of the logarithms of the edges weights, the path with maximum probability can be efficiently calculated as the path with the smallest sum of transformed edge weights. This optimal path between a pair of nodes can then be efficiently calculated using Dijkstra's algorithm for shortest paths (Cormen *et al.*, 2000).

To identify interactions which cannot be explained by a path of sufficiently strong indirect interactions, we process candidate interactions in the order of non-increasing edge weights (see Figure 7.1 **B**, Figure 7.2). For each interaction e, we calculate the optimal alternative path P with weight $w(P)$ between the corresponding proteins in the current scaffold. The interaction e is added to the scaffold if $w(e) \geq \alpha w(P)$ and the scaffold is updated whenever a new interaction is identified. Since we never remove any interaction and consequently any path from the scaffold, there is a better alternative indirect interaction in the final scaffold for every interaction not contained in the scaffold. Furthermore, we can show that there is no better alternative indirect interaction for all interactions contained in the scaffold if $\alpha \leq 1$ (see below). This proves the correctness of the algorithm.

Lemma 7.3. *If $\alpha \leq 1$, we have for all edges $e = (u, v)$ in the final scaffold network S that $w(e) \geq \alpha w(P)$ for all alternatives paths P between u and v in S.*

Proof. By contradiction: Assume, there exists a path P for an edge e such that $w(e) < \alpha w(P)$. Since the weight of each edge is ≤ 1 and edge weights on a path are multiplied to get the path weight, we have for each edge $f \in P$ that $w(P) \leq w(f)$. Thus, $w(e) < \alpha w(f) \forall f \in P$ and $w(e) < w(f) \forall f \in P$ if $\alpha \leq 1$. As a consequence, all edges on this path have been processed before e and this path was already contained in the network at the time e is added. Thus, we have for the best alternative path P_{opt} between u and v at this time that $w(P_{opt}) \geq w(P)$. As a consequence, $\alpha w(P_{opt}) \geq \alpha w(P) > e$. This is a contradiction to the construction of the scaffold network. □

7.2.3 Baseline prediction algorithms

We compare our algorithm against two baseline predictors. The complete approach predicts all interactions within the complex as direct, physical interactions. The connected approach calculates the network G_τ for each complex where $\forall e \in E_\tau : w(e) \geq \tau$ and τ the largest value such that G_τ is connected.

7.3 Results

The MST and extended MST approaches were applied to interaction scores and complex predictions calculated with our unsupervised bootstrap approach (see chapter 6) from the combined results of the genome-scale TAP experiments of Gavin *et al.* (2006) and Krogan *et al.* (2006) in yeast. As we have shown in the previous chapter, these confidence scores are more accurate than any other scoring method and the medium (BT-409) and high-confidence (BT-217) bootstrap complexes are of the same quality as the best supervised

predictions and manually curated protein complexes, respectively. Of 62,876 interactions predicted by the bootstrap algorithm, 9,918 interactions (15.8% of the original set) are within the BT-409 complexes (complete approach). From this network we extracted 1,658 interactions with the MST approach and 3,085 interactions with the extended MST approach. Compared to that, the connected baseline approach yields 5,404 interactions which is more than 3.2 and 1.75 as many, respectively.

7.3.1 Reference interactions

To compile a reference set of physical interactions we extracted all yeast protein-protein interactions from the BioGRID database (Breitkreutz *et al.*, 2008) (release 2.0.45, Oct. 1st 2008) determined with the yeast two-hybrid (Y2H) method. Furthermore, yeast Y2H interactions from the recently published study by Yu *et al.* (2008) were included as they were not yet contained in the current BioGRID release. From the complete set of Y2H interactions, we extracted a second set of interactions from small-scale experiments in which ≤ 100 interactions were determined. We used only Y2H interactions for the reference sets, since other experimental methods, such as Co-IP or pull-down assays, do not only detect physical but also indirect interactions via other proteins.

Additionally, we predicted physical interactions from large-scale Y2H studies for other species and known domain-domain interactions extracted from three-dimensional structures of protein complexes. Y2H interactions were predicted for yeast from large-scale studies for other species (Giot *et al.*, 2003; Li *et al.*, 2004; Stelzl *et al.*, 2005; Rual *et al.*, 2005) using orthology assignments from the Inparanoid database (Berglund *et al.*, 2008). Interactions were predicted if both interaction partners had orthologs in yeast. Domain-domain interactions were taken from the iPfam (Finn *et al.*, 2005) and 3DID (Stein *et al.*, 2005) databases and mapped to the yeast proteome. Only interactions between different

Complex set	**Interaction network**			
	Y2H all	**Y2H small-scale**	**Y2H pred.**	**DD**
MIPS	56.9	93.6	19.5	17.6
	[0.04]	[0.16]	[0.06]	[0.02]
BT-409	80.8	138.9	60.0	13.8]
	[0.07]	[0.18]	[0.09]	[0.01]
BT-217	76.4	174.7	56.6	18.7
	[0.05]	[0.13]	[0.06]	[0.01]

Table 7.1: Enrichment (see equation 7.1) of Y2H and domain-domain interactions (DD) within the MIPS, BT-409 and BT-217 complexes. The second row for each combination of network and complex set specifies the fraction of Y2H interactions within protein complexes.

protein chains in the crystal structure were considered.

Table 7.1 shows a comparison of the reference networks against the BT-409 and BT-217 complexes and manually curated complexes from the MIPS database (Mewes *et al.*, 2004). The first row for each combination indicates the enrichment of Y2H interactions within complexes. Enrichment is calculated as $p_C/p_{\overline{C}}$ where

$$p_C = \frac{|E_C| \cap |E_{Y2H}|}{|E_C|} \quad \text{and} \quad p_{\overline{C}} = \frac{|E_{\overline{C}}| \cap |E_{Y2H}|}{|E_{\overline{C}}|}. \tag{7.1}$$

Here, E_C is the set of interactions within complexes, $E_{\overline{C}}$ the set of interactions between proteins in different complexes and E_{Y2H} the set of Y2H interactions. The second row of Table 7.1 specifies the fraction of Y2H interactions contained within complexes.

As can be seen, Y2H and domain-domain interactions are significantly enriched within protein complexes and the enrichment values appear to reflect the confidence of the corresponding Y2H set. The complete interaction set has much lower enrichment values than the interactions taken only from small-scale experiments. The same is true for the predicted Y2H and domain-domain interactions compared to the yeast Y2H interactions. Interestingly, the enrichment is generally higher in the bootstrap complexes than in the MIPS complexes. Even so, the fraction of interactions in the Y2H and domain-domain networks which connect proteins in the same complex is very small.

7.3.2 Evaluation of predictive accuracy

The predictive accuracy of the presented methods was again evaluated using *receiver operating characteristic* (ROC) curves (see chapter 6, page 86). In this case, true positive rate is the fraction of reference interactions within the BT-409 complexes recovered by the prediction methods. False positive rate is the fraction of interactions within the BT-409 complexes predicted to be in the scaffold but not contained in the reference network.

Our approach was compared against the complete and connected baseline classifiers. We could not evaluate the predictions of the approaches by Scholtens *et al.* (2005) and Bernard *et al.* (2007), which also model the topology of complexes, for the following reasons. The predicted interactions of the Scholtens *et al.* method cannot be obtained from the R implementation of the algorithm since they are only an intermediate step in predicting the complexes. For the predictions of Bernard *et al.*, on the other hand, no unbiased performance estimate can be obtained from the Y2H reference set as they used Y2H interactions as training data.

Figure 7.3 **A** shows the ROC curve for the MST and extended MST predictions and the baseline classifiers compared against the complete set of yeast Y2H interactions. Similar results can be observed for all reference sets. As can be clearly seen, significant improvements in predictive accuracy can be obtained with the MST approach. At a maximum true positive rate of 49.1%, only 13.6% false positives are predicted. At the same true positive rate, about 22.2% false positives are predicted by the baseline classifiers. The higher specificity of the MST approach results in a significantly lower sensitivity which can be

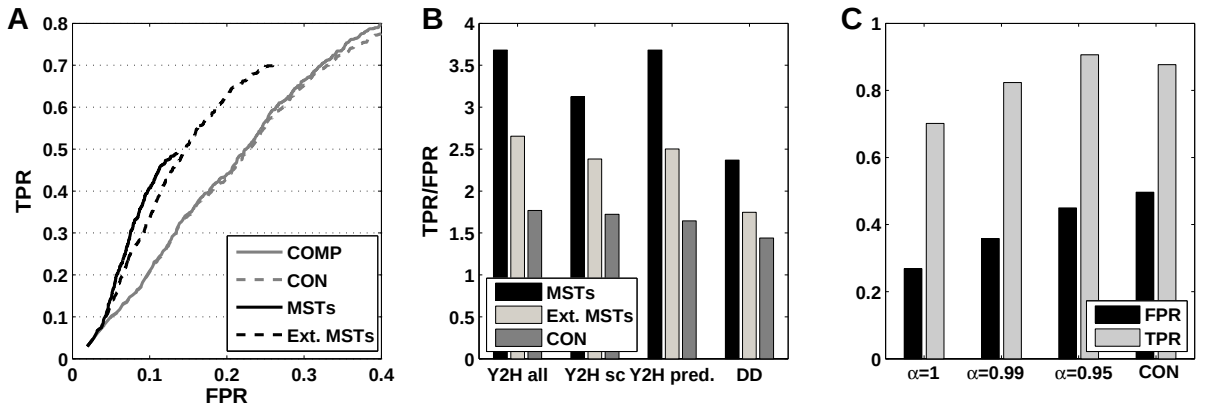


Figure 7.3: Accuracy of predicted physical interactions in the complex scaffolds. **A**) ROC curve for the interactions predicted by the complete (COMP), connected (CON), MST and extended MST (for $\alpha = 1$) approach compared to all yeast Y2H interactions within the BT-409 complexes. False positive rate (FPR) is plotted on the x-axis and true positive rate (TPR) on the y-axis. The curves for the complete and connected approach are almost identical in this range as almost all of the top scoring interactions of the complete network are also contained in the connected network. The networks differ mostly in the low scoring interactions contained additionally in the complete network. **B**) Ratio of true positive rate to false positive rate for the MST, extended MST and connected approach for the complete Y2H network (Y2H all), the small-scale Y2H interactions (Y2H sc), the predicted interactions from Y2H experiments for other species (Y2H pred.) and the domain-domain interactions from the 3DID and iPfam databases (DD). **C**) True positive and false positive rates for decreasing values of α and the connected networks in the complete yeast Y2H network.

increased by extending the MSTs. Although the false positive rate consequently increases as well, the overall performance of the extended MSTs is nevertheless significantly better than observed for the baseline predictions.

By comparing the ratios of true positive rate to false positive rate (TPR/FPR) (see Figure 7.3 **B**) we find that physical interactions are significantly enriched in the MST networks for all reference sets with TPR/FPR ratios almost twice as high than for the connected networks. For the extended MST networks, which improve the coverage of the MST method, the TPR/FPR ratios are lower but still increased compared to the connected approach.

Figure 7.3 **C** illustrates the true and false positive rates in the yeast Y2H network for decreasing values of α used for extending the MSTs. The more conditions are relaxed for extending the networks, the more interactions are added. As a consequence, more true interactions are recovered but also more wrong predictions are made. Nevertheless, more true positives can be recovered with the extended MST approach at a lower false positive rate than for the connected networks.

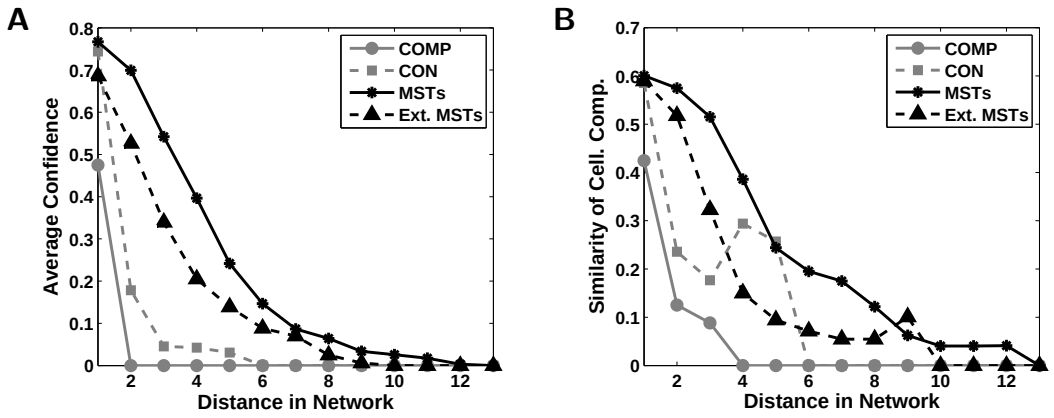


Figure 7.4: **A**) Correlation of the distance between a protein pair in the complete (COMP), connected (CON), MST and extended MST networks to the confidence of this pairwise interaction in the complete network. Averages are taken over all protein pairs with the same distance. **B**) Correlation of the distance to the fraction of GO cellular compartment annotations the two protein have in common.

7.3.3 Separation of substructures within complexes

In the network of physical interactions, proteins which are closely associated and contained in the same subcomponents of the complex are separated by only few physical interactions. Proteins in different subcomponents, on the other hand, are separated by many interactions. To evaluate this for the predicted interaction networks, we compared the distance of two proteins, i.e. the number of interactions on the shortest (unweighted) path between them, to the similarity of the subcomponents they are part of.

The distance of two proteins is significantly correlated to the confidence of the corresponding interaction in the original bootstrap network (Pearson correlation coefficient: -0.46 (complete network), -0.59 (connected approach), -0.91 (MST), -0.86 (extended MST)). Thus, the higher the confidence of an interaction between two proteins, the smaller is the distance in the resulting scaffold network (see Figure 7.4 **A**). Due to the short distances in the complete and connected networks where most proteins are directly connected, the correlation is significantly weaker than in the MST and extended MST networks.

To evaluate whether the distance of two proteins in the network reflects the similarity of their subcomponent localizations, it was compared against the fraction of Gene Ontology (GO) (Ashburner *et al.*, 2000) cellular component annotations they have in common (Figure 7.4 **B**). We used this simple similarity measure as more sophisticated measures such as the semantic similarity measure we used in chapter 6 are generally very high within proteins complexes. Here, the simple overlap measure allows for a more fine-grained analysis of the subcomponents within complexes.

As with interaction confidences, the similarity of the cellular component assignments generally decreases with increasing distance between the corresponding proteins. Further-

more, similarity decreases less rapidly for the MST and extended MST networks since the networks are more sparse and, as a consequence, the distances are larger. Thus, proteins involved in different subcomponents of a complex are separated from each other by many interactions in the predicted scaffolds, whereas proteins involved in the same subcomponents are close to each other. As the small distances in the baseline predictions make it difficult or even impossible in many cases to identify a substructure in the networks, resolution of the subcomponent structure is significantly improved by the MST and extended MST approaches.

Surprisingly, co-localization scores increase again at a distance of 4 for the connected network and at a distance of 9 for the extended MST network. This is due to the small number of protein pairs with this distance in the corresponding networks. Thus, outliers affect the average localization similarity more strongly.

7.3.4 Density of complex scaffolds

An analysis of the 195 complexes with size > 2 showed a negative correlation between the density of the complex scaffolds predicted by the extended MST approach and the size of the complexes (Spearman correlation coefficient: -0.34, p-value: 1.3×10^{-6}). 83 (42.6%) of these complexes are fully connected but they have an average size of only 3.8. Thus, for small complexes a globular structure is generally predicted with most proteins interacting physically. For the remaining complexes (average size 9.5) only 28.4% of interactions not contained in the MST networks are predicted additionally for the extended MST networks. Thus more sparse networks and as a consequence, more complex structures are predicted for larger complexes for which it is physically impossible that all proteins are directly connected.

We analyzed 20 complexes with size ≥ 10 and maximum distance in the network ≥ 4 in the extended MST network (Figures can be found at `http://www.bio.ifi.lmu.de/Complexes/Substructures/`). Based on the analysis of the extended MST networks for these complexes, we identified 5 cases for which partially overlapping complexes were clustered together due to many shared proteins (RNA polymerase; SWI/SNF and RSC chromatin remodelers; SAGA and TFIID complexes; INO80, SWR1 and NuA4 complexes; Rpd3 complexes). In these cases, interactions were only predicted between proteins in the same complex and only interactions of the shared proteins connect the individual complexes.

For 4 complexes, the spliceosome, the vacuolar ATPase, and the 90S preribosome (2 overlapping complexes), the known subcomponent structure is clearly visible in the network. For an additional 9 complexes with subcomplex annotations we found that most interactions are between proteins in the same subcomplex. Furthermore, in several cases core complex members which are central to the network can be distinguished while proteins interacting with the core members are at the network periphery. For the remaining 2 complexes, a clear spatial arrangement was observed. In the following, we illustrate with the example of the RNA polymerase how a detailed analysis of the complex scaffold can lead to a better understanding of the complex structure.

7.3.5 Analysis of the DNA-directed RNA polymerase

The DNA-directed RNA polymerase complex is one of the largest predicted complexes in the BT-409 set and contains 46 proteins. It effectively consists of three separate RNA polymerase complexes (RNA polymerase I, II and III) which have been clustered into one complex since they have many proteins in common. Such complexes which overlap to a large degree are a general problem for complex prediction algorithms and the approaches by Pu *et al.* (2007) and Hart *et al.* (2007) also cluster the three polymerases together. The crystal structure of polymerase II is known, whereas only little structural information is available for polymerases I and III (Cramer *et al.*, 2008).

Figure 7.5 shows the complex connections for the RNA polymerase in the connected and extended MST network. The complex was visualized using the organic layout function of Cytoscape (Shannon *et al.*, 2003) which clusters closely connected proteins together. In the complete bootstrap network, no substructure can be observed but all proteins form a tight cluster. In the connected network (see Figure 7.5 **A**), we observe at least a separation between the RNA polymerase III complex and the remaining proteins but polymerase I and II are too tightly connected to identify the substructure. It is only when proteins are colored by their cellular components that we detect that proteins from the same subcomponents are clustered together.

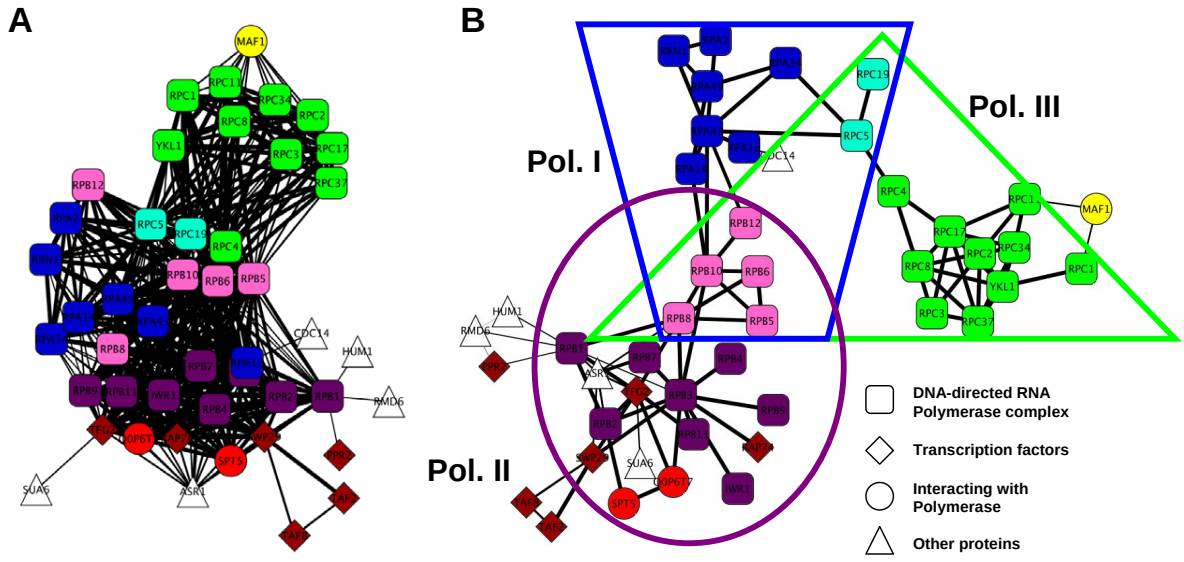


Figure 7.5: Predicted subnetworks for the DNA-directed RNA complex with the connected (**A**) and the extended MST approach (**B**). Colors indicate the three subcomponents: Polymerase complexes I (purple), II (dark blue) and III (green). Proteins contained in all 3 polymerases or the polymerases I and II are colored in pink and cyan, respectively. Rounded rectangles denote the actual polymerase proteins, rotated squares transcription factors and circles proteins interacting with the polymerases.

In the extended MST network (see Figure 7.5 **B**) the subdivision of the complex into polymerase complexes I, II and III can be clearly observed. The polymerase III complex is connected to the polymerase I complex by the two proteins which are part of both complexes (RPC19, RPC5). The latter one is connected to the polymerase II complex by a group of five proteins (RPB5, RPB6, RPB8, RPB10 and RPB12) contained in all three RNA polymerase complexes. Accordingly, physical interactions are only predicted between proteins contained in the same polymerase and no interactions are observed between different polymerases.

In the MST and extended MST network, the five proteins contained in all 3 polymerases are not directly connected to the other polymerase III proteins although they are subunits of this complex. If we relax the criterion for extending the MSTs ($\alpha = 0.99$), the interaction between RPB10 and RPC5, which was reported previously (Flores *et al.*, 1999), is added to the scaffold (see `http://www.bio.ifi.lmu.de/Complexes/Substructures/`). At first glance, this suggests that the interactions of the shared proteins to polymerase III are mediated via this interaction. However, if we look at the crystal structure of polymerase II and the model for polymerase III (Cramer *et al.*, 2008), we find that none of the common proteins are actually in physical contact in the complexes (possibly apart from RPB10 and RPB12).

Going back to the original purification experiments, we find that of the 7 interactions predicted between the common proteins, 6 interactions are bait-prey interactions which have been found to be very reliable (Bader and Hogue, 2002; Bader *et al.*, 2004) and 3 of those are identified in both directions (bait-bait interactions). This indicates that the association between these proteins is very strong. Since they do not appear to physically interact, this is probably a consequence of the fact that they are contained together in three different complexes. This close association of the five proteins can be identified reliably from the extended MST network.

Another example which illustrates the problems of affinity purification in distinguishing the physical interactions from indirect interactions is the RPB3 protein. The extended MST approach predicts 6 interactions to proteins contained in the polymerase II crystal structure. In 5 of those cases, the distance of these proteins in the crystal structure supports a physical interaction. For the interaction between RPB3 and RPB9, however, we find that the two proteins are located at different ends of the polymerase II in the 3D structure. Despite this fact, this interaction has a very high confidence score as RPB3 and RPB9 co-purified each other whenever one of these proteins was used as bait (6 times for RPB3 and 5 times for RPB9). In the crystal structure, RPB3 and RPB9 are linked by indirect interactions via RPB1 or RPB2. Since at least one of these proteins was always co-purified with both RPB3 and RPB9 as bait, this may explain the tight association between these proteins.

7.4 Discussion

Here, we presented an approach for predicting the topology of protein complexes described by the scaffold of direct interactions which spans the complex. First, our method calculates the union of all maximum spanning trees (MSTs) in the interaction score network for a protein complex. In a subsequent step, this network is iteratively extended by interactions which cannot be explained by a path of alternative indirect interactions. The MST approach is applicable to all weighted interaction networks and in particular to interaction scores calculated from affinity purification assays with any of the recently published scoring methods. Confidence scores which are required for extending the MSTs in our algorithm, can be obtained by scaling any type of scores to $[0, 1]$ or using our bootstrap approach implemented in the ProCope package (see section 6.3.7) to calculate confidence scores from affinity purification experiments.

Predictive performance of subnetworks calculated from bootstrap confidence scores was evaluated on experimentally determined direct, physical interactions from Y2H experiments and domain-domain interactions. Our results show that predictive accuracy can be increased significantly with our approach compared to baseline predictions. When comparing the individual protein complexes to the Y2H network, we observed that only 50-70% of predicted and manually curated complexes contain at least one Y2H interaction, and only 7-17% of the complexes are actually non-trivially connected (i.e. they are connected and contain more than two proteins) in the Y2H network. This suggests that many of the direct interactions within complexes have not been identified yet. In this case, the interactions predicted by our approach but not found in the Y2H network are promising starting points for experimental verification. Furthermore, due to false positives in the Y2H system, not all of the reference interactions may be true physical interactions. Here, the predicted scaffold may be used to cross-check the Y2H interaction networks for false positives.

Protein complexes are not simply disordered clumps of proteins but they have an internal substructure and a well-defined spatial arrangement in which not all proteins interact physically. In the network of physical interactions, proteins in the same subcomponents are closely connected whereas proteins in different subcomponents are separated by long paths of physical interactions. A comparison of the distance between protein pairs in our predicted networks and their similarity regarding the subcomponents they are part of, showed that similarity is negatively correlated to the distance. Thus, the networks reflect the modular substructure of the corresponding complexes. Although the same negative correlation was also observed for the baseline prediction approaches, distances in these networks are very short and, thus, do not allow for a reasonable resolution of the complex structure.

The extended MST approach predicts a globular structure for the majority of small complexes while sparser networks with more intricate structures are predicted for larger complexes. A detailed analysis of 20 of the largest complexes showed that the visual inspection of the predicted scaffold networks can identify subcomponents or overlapping complexes combined into one complex due to many shared subunits. Furthermore, core

members of the complex and proteins more loosely associated to the complex can be distinguished. A further analysis of those complexes for which a spatial arrangement was observed but no structure has been described yet may lead to new insights into these complexes beyond the individual protein subunits.

We illustrated this approach on the complex of DNA-directed RNA polymerases. While the substructure of the complex with three different RNA polymerases can only be partly observed in the baseline predictions, it is clearly evident in the network predicted with our approach. By relaxing the conditions for extending the MSTs slightly, the substructure of the complex can be further emphasized and important interactions can be identified. Furthermore, a comparison of the predictions and the original purification experiments to the 3D structure showed that the TAP system has limitations in distinguishing between indirect interactions and the actual physical ones. As a consequence, proteins interacting indirectly with each other can be co-purified repeatedly and with high specificity if they are contained together in several different complexes or are linked by indirect interactions through several proteins.

The approach we developed can be easily extended to include additional information in the form of known physical interactions or information from crystal structures on which proteins are too far removed to be physically interacting. In this way, interactions may be either enforced or forbidden when extending the MSTs and alternative indirect paths can be either created or removed.

7.5 Conclusions

In this chapter, we have presented an approach for post-processing protein complex predictions to allow for a more detailed analysis and comparison of complexes predicted from affinity purification results. Based on maximum spanning trees we infer physical interactions to identify and visualize the substructure of protein complexes in an intuitive way. We have shown that physical interactions are enriched within the predicted networks and that proteins in the same subcomplex are in close proximity in the complex scaffold while proteins from different subcomplexes are separated by many physical interactions. Thus, the complex topology can be inferred from purification results despite the experimental limitations of purification assays in distinguishing the actual physical interactions. Accordingly, the algorithm presented here supports the in-depth analysis of the predicted protein complexes beyond just the individual complex subunits.

Part III

RNA half-life

Chapter 8

Calculating RNA half-life from de novo transcription

8.1 Introduction

So far, we analyzed protein interactions and larger associations of proteins in complexes. Although high-throughput methods can now identify these interactions and complexes on a genome scale, they can only determine which interactions take place in the cell at some point but not when these associations are formed and how they are regulated. In particular for protein complexes it is important that all proteins are present in the cell at the same time in appropriate concentrations. As high-throughput methods for quantification of protein abundance (Ong and Mann, 2006; Bantscheff *et al.*, 2007; Hober and Uhlén, 2008; Colzani *et al.*, 2008) are not as advanced yet as methods for mRNA quantification, analysis of the regulation of biological processes is mostly performed at the mRNA level. Since genes are first transcribed to mRNA which is then used as a template for protein synthesis ("central dogma of biology", Crick, 1970), changes in mRNA abundance are assumed to be correlated to changes in protein abundance.

State-of-the-art gene expression profiling methods allow measurements of transcript abundance with microarrays on a genome scale, not only for entire genes but also individual exons, introns and the large fraction of non-coding genomic regions (Bertone *et al.*, 2004; Gardina *et al.*, 2006; Robinson and Speed, 2007). Standard approaches are focused on identifying differentially expressed genes whose transcript abundance is either up- or down-regulated between different conditions. As RNA levels in a cell are determined by the relative rates of RNA synthesis by polymerases and degradation by nucleases, constant transcript levels are maintained by an equilibrium of RNA synthesis and decay. As a consequence, changes in transcript levels may be caused by alterations in both synthesis or decay (Ross, 1995) which cannot be distinguished by standard gene expression profiling methods.

RNA decay rates have previously been determined by blocking transcription and then monitoring RNA decay over time (Bernstein *et al.*, 2002; Gutierrez *et al.*, 2002; Ragha-

van *et al.*, 2002; Selinger *et al.*, 2003; Yang *et al.*, 2003; Grigull *et al.*, 2004; Andersson *et al.*, 2006; Narsai *et al.*, 2007). Based on the assumption that RNA decay continues at the same rate after inhibition of transcription, decay rates for thousands of transcripts were obtained using microarray technology. However, transcriptional arrest induces a major stress response in the cell which influences regulatory mechanisms of RNA decay and can lead to stabilization of individual transcripts (Gorospe *et al.*, 1998; Blattner *et al.*, 2000; Brennan and Steitz, 2001; Bhattacharyya *et al.*, 2006). Additionally, it has been noted that the commonly used transcriptional inhibitor actinomycin-D (act-D) may stabilize some transcripts (Shyu *et al.*, 1989). Thus, this approach is inherently cell-invasive and, furthermore, cannot be combined with simultaneous measurements of RNA de novo transcription.

De novo transcription can be measured in a nondisruptive way by labeling newly transcribed RNA with 4-thiouridine (4sU) utilizing the nucleoside salvage pathways (Melvin *et al.*, 1978) followed by thiol-mediated isolation of the labeled newly transcribed RNA from total RNA (Melvin *et al.*, 1978; Woodford *et al.*, 1988; Ussuf *et al.*, 1995; Kenzelmann *et al.*, 2007; Dölken *et al.*, 2008). By combining this technique with standard microarray techniques, newly transcribed RNA can be directly measured for thousands of genes at the same time (Kenzelmann *et al.*, 2007; Dölken *et al.*, 2008). Furthermore, the integrated approach recently developed by Lars Dölken, our collaborator from the Max von Pettenkofer-Institut, allows the separation of total cellular RNA into both newly transcribed, labeled RNA and pre-existing, unlabeled RNA with high specificity (Dölken *et al.*, 2008).

This approach has the advantage that RNA decay rates can be monitored without transcriptional arrest. Changes in RNA synthesis and decay as well as their impact on total RNA levels can be measured simultaneously in a single experimental setting and RNA half-lives can be determined. Quantification of RNA synthesis increases sensitivity in detecting differentially expressed genes as significant changes can be observed in newly transcribed RNA before they have any noticeable effect on total RNA. Furthermore, analysis of transcript half-lives can lead to new insights into regulatory processes and provide a more dynamic picture of interactions in protein complexes (see the following chapter).

In the collaboration with Lars Dölken, we developed novel computational methods for the analysis of this new type of expresssion data and the calculation of transcript half-life (Dölken *et al.*, 2008). Based on the separation of total RNA into newly transcribed and pre-existing RNA and the analysis of all three RNA subsets, array data of the different RNA fractions can be normalized in an intuitive way using results for thousands of genes from single time point measurements within the same experiment. This approach also provides novel access to microarray data and probe set quality control, e.g. to determine the reliability of individual probe sets and select the most reliable ones if a single gene is represented by multiple probe sets on a microarray chip. We demonstrate that RNA half-life measurements based on decay, e.g. after blocking transcription, are inherently imprecise for medium- to long-lived transcripts. In contrast, RNA half-lives determined from newly transcribed/total RNA ratios are highly precise independent of transcript half-life. Moreover, significance of changes in newly transcribed and total RNA, e.g. after

interferon treatment, can be assessed in a straightforward way utilizing RNA half-lives to reliably detect differentially expressed genes.

8.2 Methods

8.2.1 Experimental data

Newly transcribed RNA in murine fibroblasts was labeled by Dölken *et al.* (2008) by culturing cells in the presence of 4-thiouridine (4sU) for 1 h. Total cellular RNA was isolated, thiol-specifically biotinylated and separated into labeled newly transcribed RNA and unlabeled pre-existing RNA using streptavidin coated magnetic beads. Three biological replicates each for newly transcribed, pre-existing and total RNA were analyzed using Affymetrix MG 430 2.0 arrays. In an additional experiment, newly transcribed and total RNA was measured after 1 h labeling for cells treated with interferon α (IFNα) (3 replicates) or IFNγ (3 replicates) or left untreated for comparison (3 replicates).

To compare the effect of different labeling times, newly transcribed and total RNA was also measured after 30 min of labeling (9 replicates) and newly transcribed RNA after 15 min labeling (3 replicates). In the latter case, total RNA measurements from 30 min labeling were used to calculate RNA half-lives. Furthermore, blocking of transcription with actinomycin-D (act-D) was evaluated by measuring total RNA after 1 h, 2 h and 3 h act-D treatment and without treatment. Microarrays were normalized using the GCRMA (Wu and Irizarry, 2004) algorithm of the BioConductor project (Gentleman *et al.*, 2004) in R (R Development Core Team, 2007).

On Affymetrix microarrays, as well as other oligonucleotide microarrays, genes are represented by short oligonucleotide sequences (probes). Each probe corresponds to a part of the sequence of a single known or predicted gene and RNA samples for this gene should hybridize to this probe. Several probes for the same gene are combined to a so-called probe set. Affymetrix arrays contain for each probe a so-called "mismatch" probe with a single nucleotide mismatch in the center of the probe sequence. The mismatch probe is used to quantify non-specific hybridization and to evaluate the reliability of probe sets and classify them either as "present", "marginal" or "absent" (see Affymetrix Inc. (2001) for details). For our analyses, only probe sets were considered with present calls for all replicates and RNA fractions considered.

8.2.2 Normalization by linear regression analysis

A crucial step in the calculation of RNA half-life is the normalization of RNA concentrations between the three different RNA fractions (newly transcribed RNA, pre-existing RNA and total RNA). As standard normalization approaches assume equal overall intensities for all arrays, a second normalization step is required to compensate for the different amounts of template mRNA present in newly transcribed RNA, pre-existing RNA and total RNA samples. In previous studies based on blocking transcription (e.g. using actinomycin-D)

this was done by either picking a few reference genes for which RNA half-life was determined independently (Yang *et al.*, 2003) or by fitting an exponential decay model to time series measurements (Narsai *et al.*, 2007). Thus, these methods either rely on very few genes or require many repeated measurements at different time intervals.

Here, we show that normalization can be achieved based on thousands of genes using only one time point when all three RNA fractions derived from a single sample are analyzed in parallel. RNA half-lives are calculated from the ratio of newly transcribed, labeled RNA (nt-RNA, L) and pre-existing, unlabeled RNA (p-RNA, U) to total RNA (t-RNA, N). After application of standard normalization algorithms (e.g. GCRMA), measured values for each fraction (L*, U* and N*) are assumed to be proportional to the true RNA concentrations with the same proportionality factor for all probe sets on the array. Before RNA half-life can be estimated, the proportionality factors have to be determined. However, since only the nt-RNA/t-RNA and p-RNA/t-RNA ratios are required for half-life calculation, we only have to identify two correction factors c_u and c_l with

$$\frac{U^*(t)}{N^*(t)}c_u = \frac{U(t)}{N(t)} \quad \text{and} \quad \frac{L^*(t)}{N^*(t)}c_l = \frac{L(t)}{N(t)}. \tag{8.1}$$

Since newly transcribed and pre-existing RNA should add up to the total RNA concentration, the correction factors c_l and c_u can be determined from the following equation:

$$\frac{L(t)}{N(t)} + \frac{U(t)}{N(t)} = 1 \Leftrightarrow \frac{L^*(t)}{N^*(t)}c_l + \frac{U^*(t)}{N^*(t)}c_u = 1 \Leftrightarrow \frac{U^*(t)}{N^*(t)} = \frac{1}{c_u} - \frac{L^*(t)}{N^*(t)}\frac{c_l}{c_u} \tag{8.2}$$

As a consequence, correction factors can be calculated with linear regression from microarray measurements for thousands of genes. If only either newly transcribed or pre-existing RNA has been measured in addition to total RNA, correction factors can be estimated based on the average half-life of RNA if this is known or estimated from some other source (see section 8.2.4).

8.2.3 Calculation of RNA half-life

With the beginning of 4sU treatment, newly transcribed RNA is labeled and the original pre-existing RNA decays at its normal rate. RNA decay has been shown to follow first-order kinetics (Lam *et al.*, 2001) with

$$\frac{dU}{dt} = -\lambda U \tag{8.3}$$

where λ is the decay rate for a given transcript. At time 0, the amount of pre-existing RNA corresponds to total RNA, i.e. $U(0) = N(0)$. At time t, the amount of pre-existing RNA is

$$U(t) = N(0)e^{-\lambda t}. \tag{8.4}$$

More intuitively, RNA decay is often described by the time required until half of the original amount of RNA is decayed. This is called the half-life of the mRNA and denoted

by $t_{1/2}$. The half-life $t_{1/2}$ can be calculated from the decay rate λ as $t_{1/2} = \ln 2/\lambda$ and as a consequence

$$U(t) = N(0)2^{-t/t_{1/2}}. \tag{8.5}$$

For our model, we assume that total RNA at time t is a multiple or fraction of the original concentration, i.e. $N(t) = \alpha(t)N(0)$ where $\alpha(t)$ is a function of time. This time-dependent factor is included to allow for variations in total RNA abundance, for instance due to cell growth or regulation of genes (see section 8.2.5). In the steady state case, $\alpha(t) = 1$.

As RNA de novo transcription compensates for the decay of RNA and the change in total RNA concentration, the amount of newly transcribed RNA $L(t)$ is composed of two parts: the amount of RNA necessary to compensate for decay ($L^c(t)$) and the amount of RNA necessary to change total RNA levels ($L^n(t)$). As $U(t) + L^c(t) = N(0)$, we have that

$$L^c(t) = N(0)\left(1 - 2^{-t/t_{1/2}}\right). \tag{8.6}$$

Furthermore,

$$L^n(t) = N(t) - N(0) = N(0)(\alpha(t) - 1). \tag{8.7}$$

If total RNA abundance is increased, $\alpha(t) > 1$ and $L^n(t) > 0$. If total RNA abundance is decreased and $\alpha(t) < 1$, $L^n(t)$ is negative and less RNA is transcribed than necessary to compensate for decay.

The total amount of newly transcribed RNA $L(t)$ is then calculated as the sum of equations 8.6 and 8.7:

$$L(t) = L^c(t) + L^n(t) = N(0)\left(\alpha(t) - 2^{-t/t_{1/2}}\right). \tag{8.8}$$

Since newly transcribed RNA is also subjected to RNA decay, $L(t)$ is not the total amount of RNA transcribed between time 0 and t but the net amount of newly transcribed RNA we observe at time t.

Using equations 8.5 and 8.8, the half-life $t_{1/2}$ is calculated as

$$t_{1/2} = -t \ln 2 / \ln\left(\frac{U(t)}{N(0)}\right) = -t \ln 2 / \ln\left(\alpha(t) - \frac{L(t)}{N(0)}\right). \tag{8.9}$$

Since $N(0) = N(t)/\alpha(t)$, we can rewrite the last equation as

$$t_{1/2} = -t \ln 2 / \ln\left(\frac{U(t)}{N(t)}\alpha(t)\right) = -t \ln 2 / \ln\left(\left(1 - \frac{L(t)}{N(t)}\right)\alpha(t)\right). \tag{8.10}$$

By inserting the normalized ratios measured for one probe set, the *probe set half-life* $t_{1/2p}$ can be estimated as

$$t_{1/2p} = -t \ln 2 / \ln\left(\frac{U^*(t)}{N^*(t)}c_u\alpha(t)\right) = -t \ln 2 / \ln\left(\left(1 - \frac{L^*(t)}{N^*(t)}c_l\right)\alpha(t)\right). \tag{8.11}$$

If a gene is represented by exactly one probe set, *gene half-life* $t_{1/2}$ is defined as the half-life measured for the corresponding probe set. On Affymetrix arrays, however, a large

number of genes are represented by several different probe sets. Due to quality differences between probe sets and experimental noise, half-life measurements of different probe sets for a single gene can result in dramatically different results. So far, it was not possible to identify the most reliable probe set in the present experimental setting and assess the level of noise. This can now be achieved based on the combined analysis of total, newly transcribed and pre-existing RNA derived from a single RNA sample. For this purpose a probe set quality score (PQS) is defined for each probe set based on the distance of the measurements from the linear regression line:

$$PQS = \exp\left(-\left|1 - \left(\frac{L^*(t)}{N^*(t)}c_l + \frac{U^*(t)}{N^*(t)}c_u\right)\right|\right). \tag{8.12}$$

Thus, the lower the distance, the higher is the PQS. For genes represented by several probe sets, gene half-life $t_{1/2}$ is consequently defined as the half-life for the probe set with maximum quality score.

To assess the overall rate of decay in a cell, median RNA half-life $t_{1/2m}$ is calculated which is defined as the median over all gene half-lives. To calculate $t_{1/2m}$, we first determine the median gene half-life separately from the pre-existing and newly transcribed RNA and then take the average of the two values. For this purpose, only normalized ratios < 1 are considered. As efficient biosynthetic labeling takes about 5 minutes to start after the addition of 4sU, the actual duration of labeling performed was reduced by 5 minutes to calculate mRNA half-lives (e.g. 55 minutes were used instead of 60, 25 instead of 30 and 10 instead of 15 minutes).

8.2.4 Normalization based on median half-life

If both newly transcribed and pre-existing RNA are measured, microarrays can be normalized in an intuitive way by estimating correction factors with linear regression analysis. If only newly transcribed or pre-existing RNA are measured, correction factors can still be determined if the median RNA half-life is known or estimated from other experiments. For this purpose, we use the definition of median half-life as the median over all gene half-lives. Here, the half-live for a gene i is denoted as $t_{1/2}(i)$ and the unnormalized nt-RNA/t-RNA and p-RNA/t-RNA ratios by $\frac{L_i^*(t)}{N_i^*(t)}$ and $\frac{U_i^*(t)}{N_i^*(t)}$, respectively. We then have that

$$\begin{aligned} t_{1/2m} &= \operatorname*{median}_i t_{1/2}(i) \\ &= -t \ln 2 / \ln\left(\operatorname*{median}_i \frac{U_i^*(t)}{N_i^*(t)} c_u \alpha(t)\right) \\ &= -t \ln 2 / \ln\left(\left(1 - \operatorname*{median}_i \frac{L_i^*(t)}{N_i^*(t)} c_l\right) \alpha(t)\right). \end{aligned} \tag{8.13}$$

Solving this equation for c_u and c_l, we obtain that

$$c_u = 2^{-t/t_{1/2m}} \alpha(t)^{-1} / \operatorname*{median}_i \frac{U_i^*(t)}{N_i^*(t)} \tag{8.14}$$

and

$$c_l = \left(1 - 2^{-t/t_{1/2m}} \alpha(t)^{-1}\right) / \operatorname*{median}_i \frac{L_i^*(t)}{N_i^*(t)}. \tag{8.15}$$

8.2.5 Modeling steady state, cell division and regulation

The factor $\alpha(t)$ included in our model describes the increase or decrease of total RNA over time and can be defined in different ways to model different scenarios. If $\alpha(t) = 1$, the amount of total RNA is assumed to remain approximately constant over time (steady state). Although RNA transcript levels in a standard cell culture are assumed to maintain a kind of steady-state level, half-life estimates based on newly transcribed/total RNA ratios have to consider that a small amount of RNA transcription can take place as cells grow and prepare for cell division. This can be modeled by defining $\alpha(t)$ as an exponential growth function such that the number of cells and, consequently, the amount of RNA for each gene have doubled after time T_2 ($\alpha(t) = 2^{t/T_2}$). As each cell divides after the completion of one cell cycle, T_2 is the cell cycle length (CCL) of the cell.

In the following, we show that half-life estimates can be easily converted between different models, even if the original p-RNA/t-RNA or nt-RNA/t-RNA ratios are unknown.

Lemma 8.1. *Two half-life estimates h_1 and h_2 calculated for different $\alpha_1(t)$ and $\alpha_2(t)$ are related by the following equation*

$$h_2 = h_1 \frac{-t \ln 2}{-t \ln 2 + h_1 \left(\ln \alpha_2(t) - \ln \alpha_1(t)\right)}. \tag{8.16}$$

Proof. Let r either be $\frac{U^*(t)}{N^*(t)} c_u$ or $1 - \frac{L^*(t)}{N^*(t)} c_l$ depending on which RNA fraction was used to calculate RNA half-life. According to the definition in equation 8.11, h_i is

$$h_i = \frac{-t \ln 2}{\ln\left(r \alpha_i(t)\right)}. \tag{8.17}$$

As a consequence $r = \frac{2^{-t/h_1}}{\alpha_1(t)}$. Inserting this in the definition of h_2, we get

$$\begin{aligned} h_2 &= \frac{-t \ln 2}{\ln(\frac{2^{-t/h_1}}{\alpha_1(t)} \alpha_2(t))} \\ &= \frac{-t \ln 2}{\frac{-t \ln 2}{h_1} + \ln\left(\alpha_2(t)\right) - \ln\left(\alpha_1(t)\right)} \\ &= h_1 \frac{-t \ln 2}{-t \ln 2 + h_1 \left(\ln \alpha_2(t) - \ln \alpha_1(t)\right)}. \end{aligned} \tag{8.18}$$

□

As gene half-life is calculated only from the one probe set with the highest quality score, calculation of individual half-lives does not have to be repeated to obtain results for

a different model $\alpha_2(t)$. Instead, new half-life estimates can be calculated from the results obtained from the original $\alpha_1(t)$ with equation 8.16.

Both the steady state and the cell division model can be extended to over- or under-expression of individual genes by an appropriate definition of $\alpha(t)$. This means that the total amount of RNA at time t is up- or down-regulated for a certain gene compared to the "normal" amount at time t. If β_N is the ratio of up- or down-regulation, $\alpha(t)$ is defined as $\beta_N \alpha'(t)$ where $\alpha'(t)$ is the function for the steady state or cell division model.

In theory, this would allow us to calculate half-life of a transcript if β_N is known. In practice however, estimates of up- or down-regulation are too noisy to get reliable half-life estimates. However, we can calculate the expected up- or down-regulation of total (β_N) or newly transcribed RNA (β_L) from each other.

Lemma 8.2. *The change in total RNA β_N and the change in de novo transcription β_L for a gene with half-life $t_{1/2}$ are related by the following formula:*

$$\beta_N = 1 + (\beta_L - 1)(1 - 2^{-t/t_{1/2}}\alpha(t)^{-1}). \tag{8.19}$$

Proof. Let $L(t)$ be the normal amount of newly transcribed RNA and $L'(t)$ the amount in a specific condition. The change in newly transcribed RNA can then be calculated as

$$\begin{aligned}
\beta_L = \frac{L'(t)}{L(t)} &= \frac{N(0)(\alpha(t) - 2^{-t/t_{1/2}}) + N(t)(\beta_N - 1)}{N(0)(\alpha(t) - 2^{-t/t_{1/2}})} \\
&= \frac{\alpha(t) - 2^{-t/t_{1/2}} + \alpha(t)(\beta_N - 1)}{\alpha(t) - 2^{-t/t_{1/2}}} \\
&= 1 + \frac{\beta_N - 1}{(1 - 2^{-t/t_{1/2}}\alpha(t)^{-1})}
\end{aligned} \tag{8.20}$$

This is equivalent to equation 8.19. □

This relationship is useful for assessing the significance of up- or down-regulation and identifying differentially expressed genes (see results). Due to the relationship of half-life estimates for different definitions of $\alpha(t)$ described in equation 8.16, the results for β_N and β_L do not depend on the choice of $\alpha(t)$. Any definition of $\alpha(t)$ can be used, in particular also $\alpha(t) = 1$.

8.3 Results

8.3.1 Median RNA half-life and probe set quality control

Median $t_{1/2m}$ of RNA in mouse fibroblasts was estimated with the linear regression method from the 1 h labeling experiment for which both newly transcribed (nt-RNA) and pre-existing RNA (p-RNA) were determined. It was performed separately for each replicate to obtain confidence intervals and the overall median $t_{1/2m}$ was estimated as the average

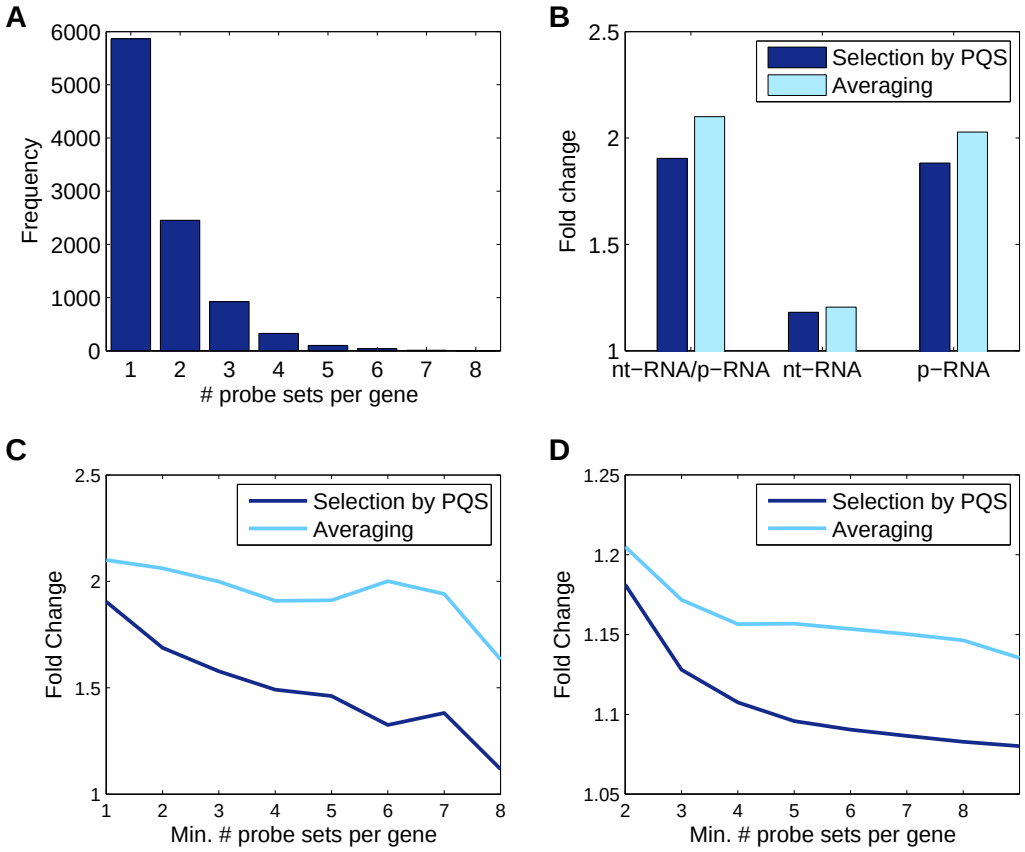


Figure 8.1: Evaluation of probe set quality control. **A)** Distribution of the number of probe sets per gene. Only probe sets are considered with present calls in all replicates for newly transcribed, pre-existing and total RNA. **B)** Average fold change of half-lives in the comparison of newly transcribed RNA against pre-existing RNA and the comparison of different replicates for newly transcribed RNA and pre-existing RNA, respectively. Results are shown for the probe set selection approach based on the probe set quality score (PQS) and the approach which calculates the average half-life of all probe sets for the same gene. **C-D)** Average fold change was calculated separately for all genes represented by at least 1, 2, 3, ..., 8 probe sets with the probe set selection and averaging methods and plotted against the minimum number of probe sets per gene (x-axis). Results are shown for the comparison between nt- and p-RNA **(C)** and different replicates for nt-RNA **(D)**. Fold changes are reduced significantly by selecting the most reliable probe set based on PQS.

of all replicates. Here and in the following, the steady state model was used ($\alpha(t) = 1$). Results for the cell division model are analyzed in section 8.3.4. Using all probe sets, median $t_{1/2m}$ was estimated at 295 minutes (~ 5 h). If only the highest quality probe sets are included for each gene (see section 8.2.3), $t_{1/2m}$ was slightly lower at 274 minutes with a 95% confidence interval between 225 and 323 minutes. Thus, median $t_{1/2m}$ in mouse cells is $\sim 4 - 5$ h.

About 40% of genes are represented by at least 2 probe sets on the microarrays used (see Figure 8.1 **A**). To compare our novel method of selecting the most reliable probe set for each gene against the standard approach of averaging probe set half-lives to obtain gene half-lives, we analyzed the difference of half-life estimates from nt-RNA and p-RNA and between different replicates (Figure 8.1 **B-D**). Difference in half-lives was evaluated in terms of the average fold change $\mathrm{fc} = \max\left(\frac{h_1}{h_2}, \frac{h_2}{h_1}\right)$ between two half-life estimates h_1 and h_2. In all comparisons, we found a clear improvement by the probe set selection method compared to the averaging approach.

To evaluate the performance of our method in the comparison of different replicates, we performed linear regression and half-life calculation for each replicate separately using only the probe set for each gene with the highest PQS for this replicate. Thus, probe set selection was performed independently for each replicate. Compared to the averaging approach, the differences in half-life estimates between different replicates were significantly reduced by the probe set selection method (Figure 8.1 **C-D**). Furthermore, as fold changes decrease with the number of probe sets per gene, the relative improvement increases.

A significant improvement by the probe set selection approach was also observed for a comparison of RNA half-life estimates from 30 min and 1 h labeling experiments in mouse (Wilcoxon rank test, p-value < 0.003). Here, probe sets were selected based on the results for the 1 h labeling experiments only. Furthermore, since the same probe set is selected independently for all three replicates significantly more often than expected at random, distinctive quality differences between probe sets are identified by our method. In particular, for the one gene represented by 8 probe sets on the mouse arrays with present calls for all replicates and RNA fractions, the same probe set is selected for all three replicates. The probability of this happening at random is < 0.016. This confirms that the deviation from the regression line is appropriate for assessing probe set reliability. In the following only the probe set with the highest PQS was used for genes represented by multiple probe sets on the microarray chips.

8.3.2 Accuracy of RNA half-life measurements

In the previous section, significant differences between estimates from nt-RNA and p-RNA were observed which are also apparent in the comparison of unnormalized nt-RNA/t-RNA (L^*/N^*) and p-RNA/t-RNA (U^*/N^*) ratios (see Figure 8.3.2 **A**). In particular for small nt-RNA/t-RNA ratios and large p-RNA/t-RNA ratios (i.e. long RNA half-lives), ratios can diverge significantly from the regression line. As variations seem to be equally pronounced in either direction and due to the large number of genes used for linear regression, median $t_{1/2m}$ and correction factors could nevertheless be estimated reliably. As a consequence, estimation of median half-life was stable even if only highly expressed genes were used.

Contrary to that, estimates for individual genes can vary significantly between pre-existing and newly transcribed RNA (Figure 8.3.2 **B**). Interestingly, differences between half-life estimates from nt-RNA for different replicates (fold change ~ 1.2, Figure 8.3.2 **C**) are significantly lower than for estimates from p-RNA (~ 1.9, Figure 8.3.2 **D**). While

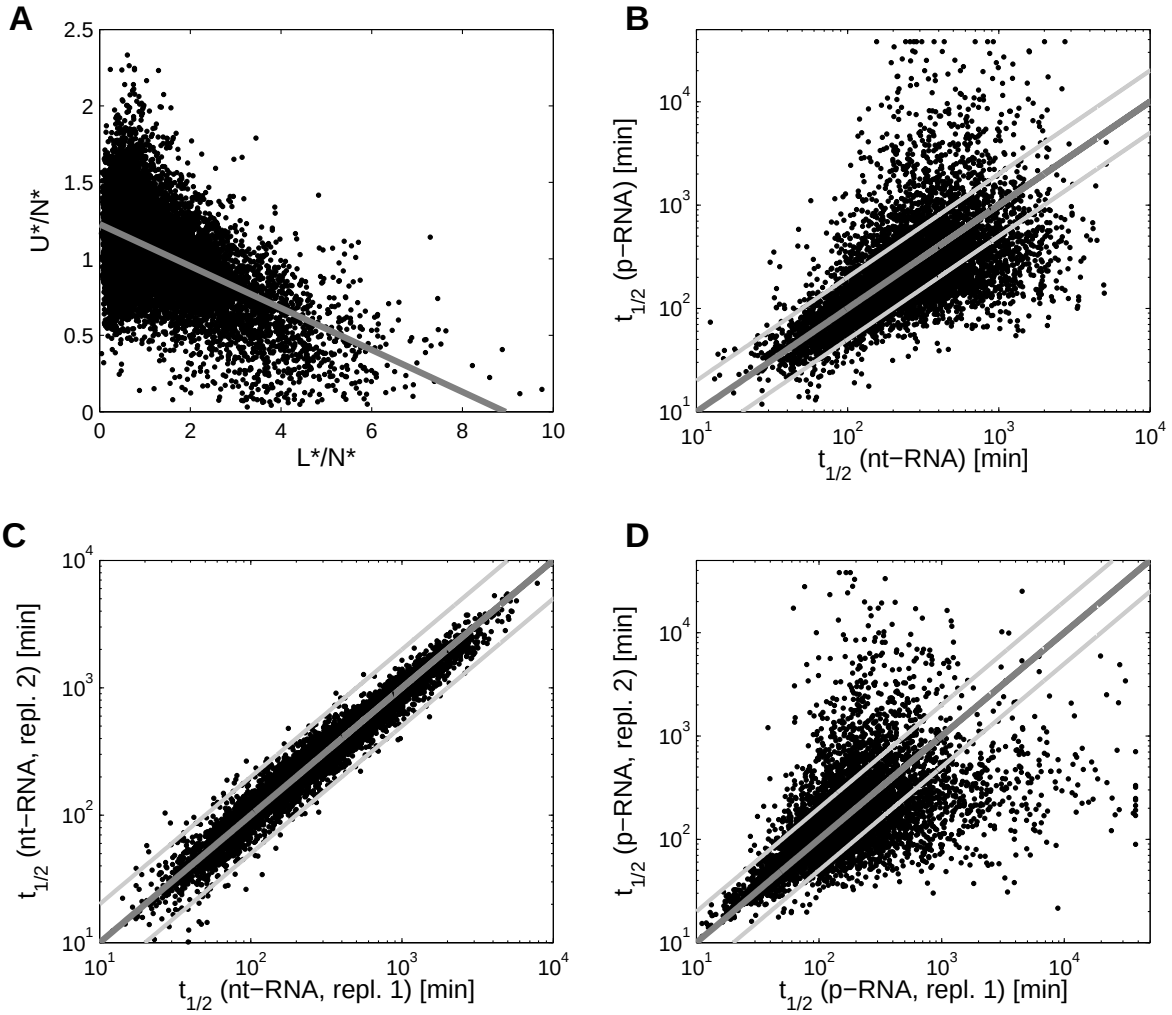


Figure 8.2: Comparison of RNA half-lives from nt- and p-RNA. **A)** Regression estimate (grey) for correction factors from unnormalized nt-RNA/t-RNA (L^*/N^*, x coordinate) and p-RNA/t-RNA ratios (U^*/N^*, y coordinate). **B-D)** Comparison of RNA half-life estimates from newly transcribed and pre-existing RNA (**B**) and between replicates 1 and 2 for newly transcribed (**C**) and pre-existing RNA (**D**). Equal estimates are indicated by the dark grey line, variation by a factor of 2 by light grey lines.

half-life estimates from nt-RNA are precise and highly reproducible for the complete range of half-lives, half-life estimates from p-RNA are only precise for very short half-lives but extremely variable for medium to long half-lives. Thus, with the exception of very short-lived transcripts, more reliable and precise half-life estimates can be obtained by measuring RNA de novo transcription and not decay.

The same observation was made for measurements of RNA decay after 1 h, 2 h and 3 h of act-D treatment (see Figure 8.3) which confirms that half-life estimates from RNA decay are only reliable for fast-decaying transcripts but not for the remaining genes. Fold

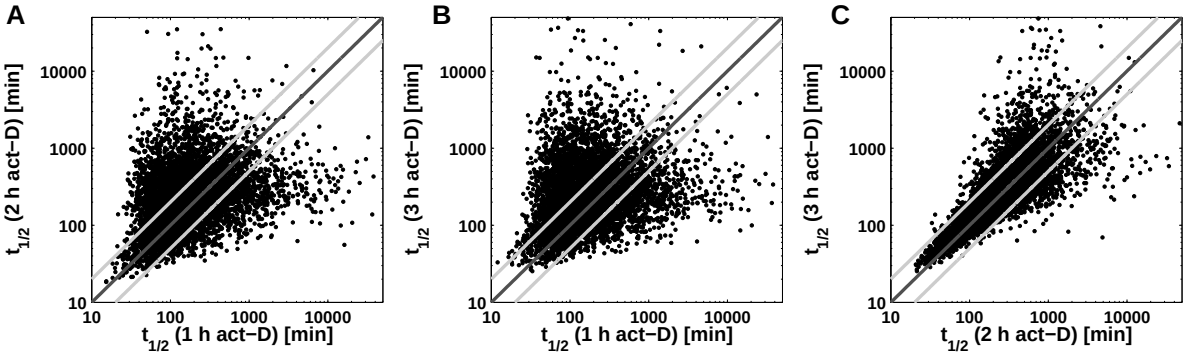


Figure 8.3: Comparison of RNA half-life estimates after actinomycin-D treatment. Half-life estimates from RNA decay after 1 h, 2 h and 3 h of act-D treatment are compared: **A**) 1 h against 2 h, **B**) 1 h against 3 h and **C**) 2 h against 3 h. Equal estimates are indicated by the dark grey line, a fold change of 2 by light grey lines.

changes between 2 h and 3 h measurements are significantly lower than between 1 h and 3 h measurements and in particular also between 1 h and 2 h measurements for which the absolute difference in act-D treatment length is the same. This indicates that estimates are more precise for larger t. However, since fold changes are still very large, this suggests that t would have to be increased by several hours to obtain accurate half-life estimates from RNA decay.

To further analyze these observations, we simulated the effect of measurement errors and noise. For this purpose, half-life estimates were obtained from nt-RNA and the expected nt-RNA/t-RNA and p-RNA/t-RNA ratios r at time t were calculated for each gene. We then simulated multiplicative measurement errors in measuring these ratios experimentally. For this purpose, we used that multiplicative errors in measuring each RNA fraction also result in a multiplicative error in the ratios.

Thus, two replicates were generated artificially for nt-RNA and p-RNA, respectively, by drawing the measured ratio r' for a gene from a normal distribution with mean r and standard deviation $r \cdot \sigma$ for a constant σ which describes the error rate. From these ratios, half-lives were then calculated. Simulation results for $\sigma = 0.2$ and $t = 1$ h and 10 h, respectively are shown in Figure 8.4. They show the same general trends as observed in the real experimental data. Precise results can be obtained from nt-RNA ratios for all genes after 1 h labeling but only for genes with short transcript half-lives from p-RNA. On the other hand, after 10 h labeling, accuracy of RNA half-lives from nt-RNA is lower than for half-lives from p-RNA.

These results can be explained in the following way. Suppose r_p is the p-RNA/t-RNA ratio and r_{nt} the nt-RNA/t-RNA ratio. At an error rate of 20%, we measure these ratios as $r_p(1 \pm 0.2)$ and $r_{nt}(1 \pm 0.2)$, respectively. To calculate RNA half-lives from nt-RNA, we calculate r'_p as $1 - r_{nt}$ which is within $1 - r_{nt}(1 \pm 0.2)$. Thus, if $r_p > r_{nt}$, i.e. if $r_p > 0.5$, r'_p has a lower error rate than r_p and, thus, provides more accurate half-life estimates.

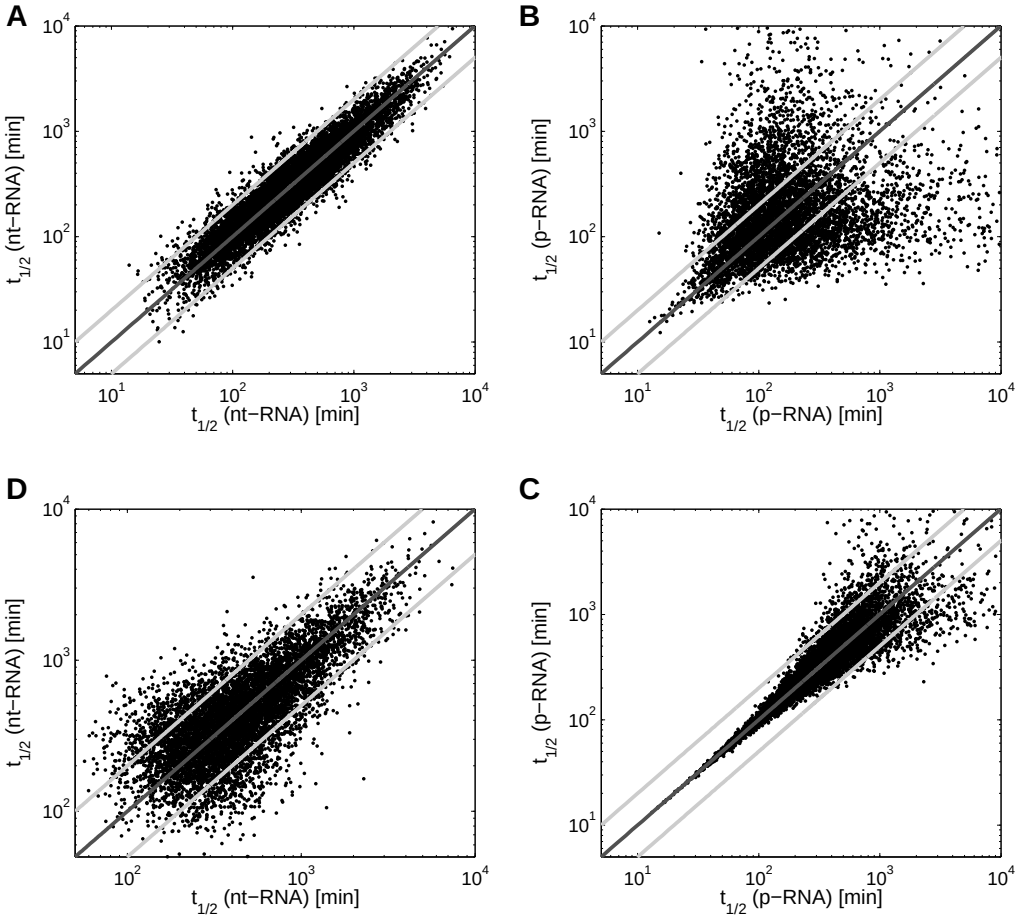


Figure 8.4: Simulation of measurement errors in RNA half-life determination. Measurement errors were simulated with $\sigma = 0.2$ and $t = 1$ h **(A-B)** and 10 h **(C-D)** for nt-RNA (**A, C**) and p-RNA (**B, D**) as described on p. 124.

For mouse fibroblasts, median $t_{1/2m}$ was estimated at ~ 5 h. This means that on average around 87% of pre-existing RNA is left after 1 h. Thus, at a 20% error rate r_p is measured within 0.87 ± 0.174 but r'_p within 0.87 ± 0.026. This indicates that estimates from nt-RNA are more reliable if $t < t_{1/2m}$ while otherwise estimates from p-RNA should be used.

8.3.3 Influence of labeling time

So far, we focused on the 1 h labeling experiments for which both newly transcribed RNA and pre-existing RNA were measured. Additional experiments have been performed for which only newly transcribed RNA and total RNA were extracted after 15 min, 30 min and 1 h of labeling (see section 8.2.1). In this case, microarray measurements were normalized based on the median $t_{1/2m}$ of 274 min as described in section 8.2.4. Since we

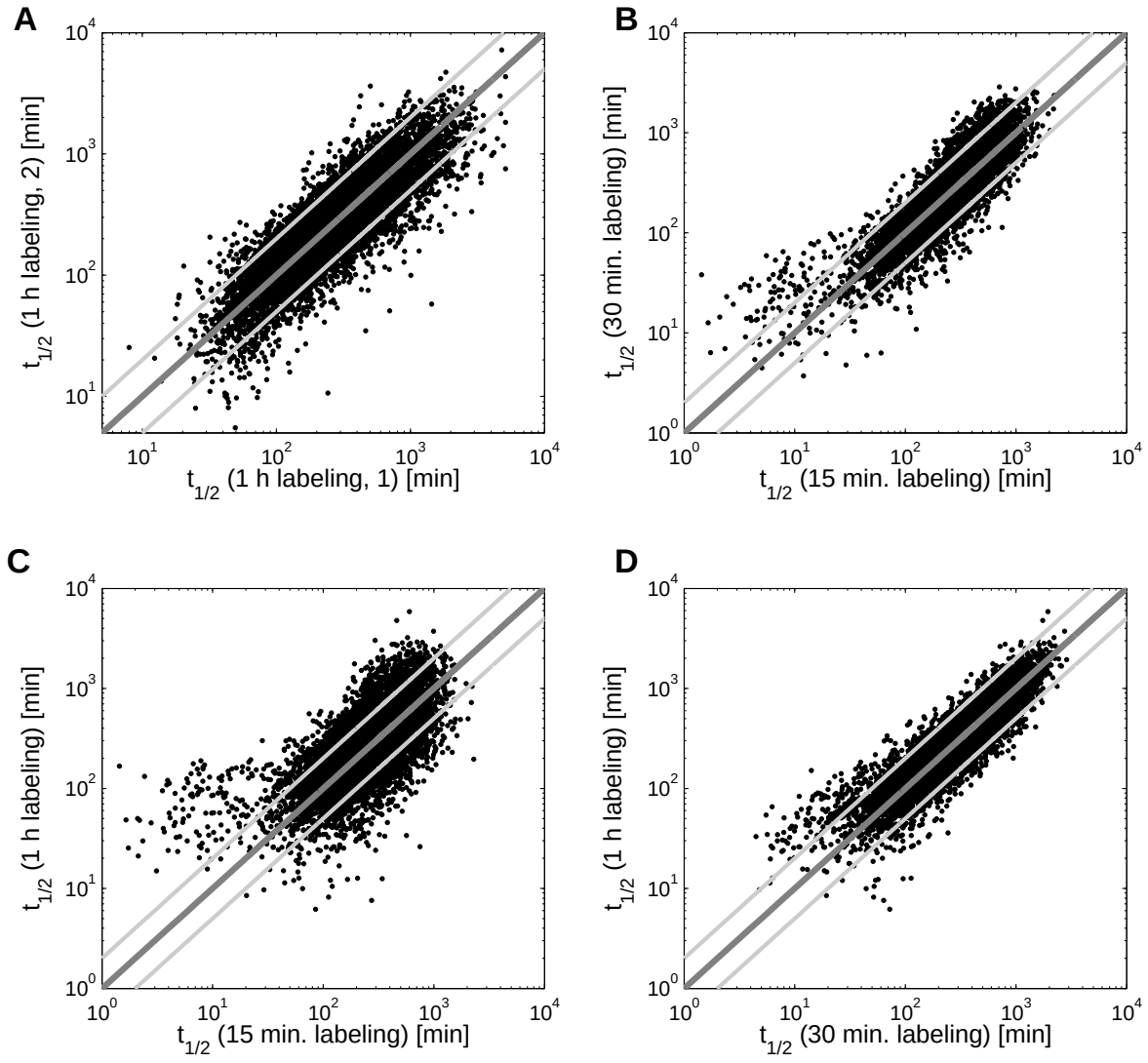


Figure 8.5: Comparison of RNA half-life estimates between **(A)** two different experiments with 1 h labeling, **(B)** 15 min and 30 min labeling, **(C)** 15 min and 1 h labeling and **(D)** 30 min and 1 h labeling. Equal estimates are indicated by the dark grey line. Light grey lines indicate a fold change of 2.

have already shown that more reliable estimates for RNA half-lives can be obtained from newly transcribed RNA, measurements of p-RNA are not necessary if the median $t_{1/2m}$ is established.

In a first analysis, we compared half-life estimates from the two independent experiments with 1 h labeling. (Figure 8.5 **A**). These results show a higher variability between results from different experiments than between different replicates in the same experiment, but half-life results are still very similar (fold change=1.49). For the comparison of different labeling times, the two 1 h labeling experiments were combined. For genes

represented by more than one probe set, the highest quality probe set was selected with present calls in both experiments. The same probe sets were chosen for the 15 and 30 min labeling results.

When comparing results from different labeling times, we generally observe a high correlation between the experiments. However, for short half-lives and to a lesser degree for long half-lives, half-lives from 15 min labeling are significantly lower than from 30 min and 1 h labeling. This indicates that nt-RNA/t-RNA ratios are increased after 15 min labeling compared to what would be expected from 30 min and 1 h labeling results. Since estimates from p-RNA are highly accurate for short half-lives, we also compared the estimates from p-RNA at 1 h labeling against the estimates from nt-RNA at 15 min labeling but observed the same situation as for half-lives from nt-RNA at 1h labeling.

As we could not reproduce this effect by simulating measurement errors for different labeling times, this indicates that decay of short-lived transcripts may not be accurately described by first-order kinetics. As newly transcribed RNA is also subjected to decay, more RNA is actually transcribed than the amount of pre-existing RNA that is decayed. Normally these additional transcripts are not measured as they are decayed before the end of labeling. However, if newly transcribed RNA is protected from decay for some time after transcription, e.g. the time it takes to transport mRNA from the nucleus to the cytoplasm where the ribosomes are located, more nt-RNA may be observed than the amount necessary to compensate for decay. This would distort half-life calculations and lead to shorter half-life estimates. For longer labeling, this effect is likely negligible whereas for short labeling of only a few minutes it may have a significant effect. This is consistent with the observation that differences between 30 min and 1 h labeling are less pronounced than between 15 min and 30 min labeling. It is not clear, however, why this effect should be most evident for extremely short-lived and long-lived transcripts.

8.3.4 Analysis of the cell division model

In the previous sections, steady state of total RNA concentrations was assumed. However, even if the relative abundance of each transcript is in steady state, absolute abundance can increase slightly as a small amount of RNA transcription takes place to allow for cell growth and cell division. In the steady state model, RNA concentrations are diluted as the cells grow since no additional RNA is transcribed. Contrary to that, the cell division model incorporates a small amount of additional RNA transcription with increasing cell volume. This model is described by $\alpha_{CD}(t) = 2^{t/T_2}$ where T_2 is the cell cycle length of the cell ($\sim$ 24 h for murine fibroblasts). For each probe set half-life estimates can be easily converted from the steady state model to the cell division model using equation 8.16. Thus,

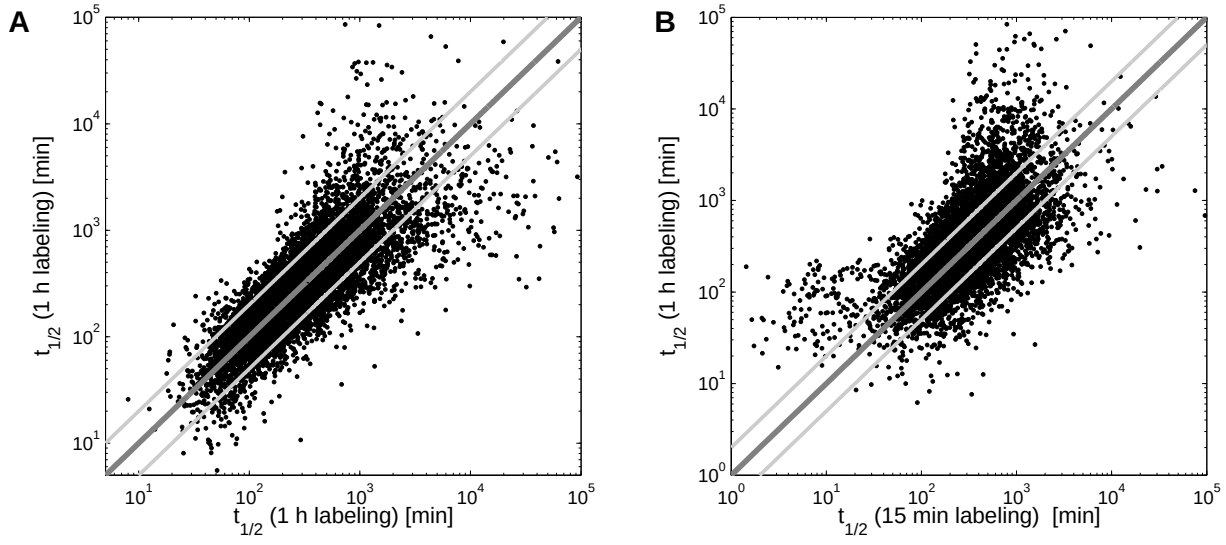


Figure 8.6: Comparison of RNA half-life estimates for the cell division model between **(A)** two different experiments with 1 h labeling and **(B)** 15 min and 1 h labeling. For short half-lives hardly any changes are observed compared to the steady state model, whereas for long half-lives variance of estimates is increased significantly.

the half-life $t_{1/2}^{CD}$ of a probe set in the cell division model is

$$\begin{aligned} t_{1/2}^{CD} &= t_{1/2}^{SS} \frac{-t \ln 2}{-t \ln 2 + t_{1/2}^{SS} \left(\ln \alpha_{CD}(t) - \ln \alpha_{SS}(t)\right)} \\ &= t_{1/2}^{SS} \frac{-t \ln 2}{-t \ln 2 + t_{1/2}^{SS} \left(\frac{t \ln 2}{T_2}\right)} \\ &= t_{1/2}^{SS} \frac{T_2}{T_2 - t_{1/2}^{SS}}. \end{aligned} \tag{8.21}$$

where $t_{1/2}^{SS}$ is the half-life in the steady state model. The difference between the steady state and cell division model depends strongly on transcript half-life. A steady-state half-life of 1 h is increased by only 3.4%, 5 h by 20% but 10 h already by 50%.

Since only the highest quality probe set was used for each gene to calculate median $t_{1/2m}$, it can be easily converted to the cell division model by converting individual gene estimates for each replicate and each RNA fraction separately and then recalculating mean and confidence interval. With this approach, median $t_{1/2m}^{CD}$ is calculated at 349 min (5.8 h) with a 95% confidence interval of 268-430 min.

Although the cell division model is more accurate in the biological sense, it leads to an amplification of measurement errors. As ratios are multiplied by $2^{t/T_2}$, error rates are also multiplied. An error of 10% for a half-life of 10 h in the steady state model multiplies to

an 18% error rate in the cell devision model and to over 30% for a half-life of 20 h. Figure 8.6 compares half-life estimates for the cell division model between the two different 1 h labeling experiments and 15 min and 1 h labeling. For short-lived transcripts, the results are effectively the same as for the steady state model. For long-lived transcripts, however, estimates from different experiments and labeling times vary much more than observed in the steady state model and considerable outliers are observed. Accordingly, variance between experiments is overestimated. Although long half-lives may be underestimated in the steady state model, results are more stable and reproducible between different replicates and, thus, the steady state model is more appropriate for comparisons between experiments or cell types.

8.3.5 Differential expression after IFN treatment

To evaluate the advantages of measuring de novo transcription and RNA half-lives for the detection of differentially expressed genes, effects of interferon (IFN) treatment in murine fibroblasts were analyzed. For this purpose, total abundance and de novo transcription after 1 h treatment with either IFNα or IFNγ was compared against results for untreated cells after 1 h labeling.

Using half-life estimates from the untreated reference experiments, expected changes in newly transcribed or total RNA can be calculated from the observed changes in total or newly transcribed RNA, respectively (see equation 8.19). The comparison of observed and expected changes (Figure 8.7) indicates that changes in total RNA abundance can be predicted accurately from the changes in RNA de novo transcription both if genes are differentially expressed and if they are not. For almost all genes ($> 99.9\%$), observed changes in total abundance and expected changes from de novo transcription vary by less than a factor of 2. Contrary to that, changes in de novo transcription can only be predicted from total RNA for genes for which total RNA is significantly changed. The reason for this is that for long-lived transcripts small variations in total RNA translate to large increases or decreases in newly transcribed RNA. Thus, small measurement errors can be misinterpreted as large changes in RNA transcription.

In standard microarray analysis, genes are generally classified as differentially expressed if transcript abundance is increased or decreased by at least a factor of 2 and/or based on the statistical significance of differences. As down-regulation of genes was only observed after IFNγ treatment for a few genes (3 genes in total RNA, 26 in newly transcribed RNA), we focused in our analysis on up-regulated genes. Using the relationship between fold changes in total or newly transcribed RNA, differentially expressed genes can be identified both based on observed and expected fold changes in two alternative ways. Coverage can be increased by classifying genes as differentially expressed if either observed or expected change is above the threshold. On the other hand, specificity can be increased by requiring that both observed and expected fold changes exceed the threshold.

To evaluate the advantages of such a combined analysis of RNA transcription and abundance for determing differentially expressed genes, we compared our results in murine fibroblasts against 97 IFNα-stimulated genes identified by Scherbik *et al.* (2007) using

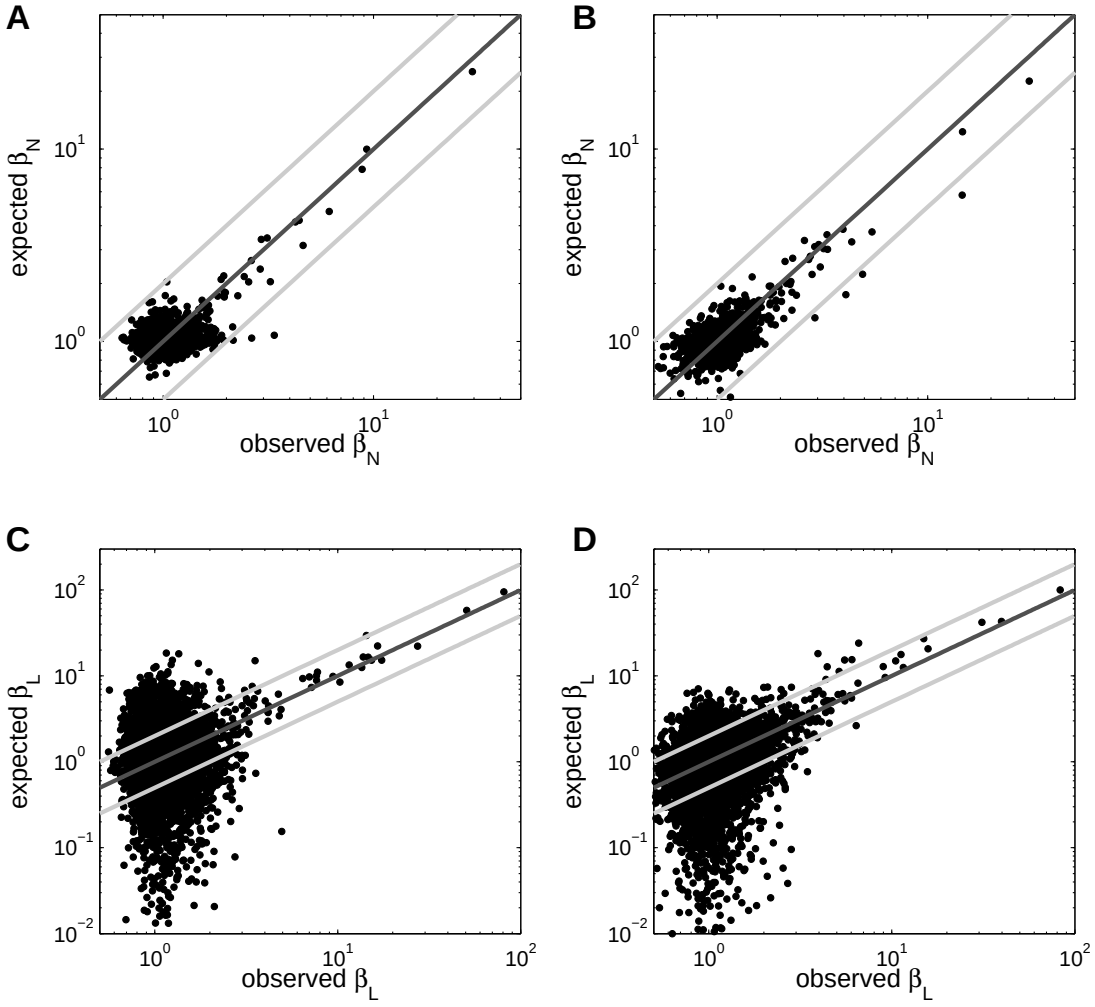


Figure 8.7: **A-B**) Observed relative change β_N in total RNA abundance and expected change from de novo transcription for IFNα (**A**) and IFNγ treated mouse fibroblasts (**B**). **C-D**) Observed relative change β_L in de novo transcription and expected change from total RNA abundance after IFNα (**C**) and IFNγ treatment (**D**). A two-fold variation is indicated by light grey lines, equality of observed and expected ratios by the dark grey line.

standard microarray analysis for mouse embryo fibroblasts. In this case, cell cultures where treated with recombinant Human IFNα for 1 h and then incubated with fresh medium for another 2 h hours before isolation of total RNA. To compare results, we classified genes from our experiments as differentially expressed from total and newly transcribed RNA, respectively, based on three approaches (observed $\beta_{N/L} \geq 2$, expected $\beta_{N/L} \geq 2$ or observed and expected $\beta_{N/L} \geq 2$) and then calculated the fraction of these genes also identified by Scherbik *et al.* (Figure 8.8).

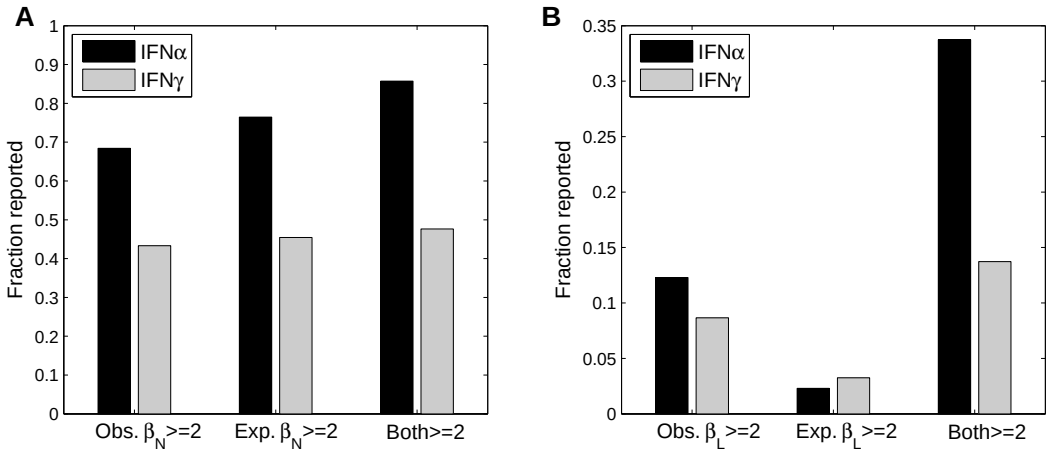


Figure 8.8: Comparison of differentially expressed genes after IFNα or IFNγ treatment to the IFNα induced genes identified by Scherbik *et al.* (2007). Genes were classified as differentially expressed in our data if either i) observed $\beta_{N/L} \geq 2$, ii) expected $\beta_{N/L} \geq 2$ or iii) both observed and expected $\beta_{N/L} \geq 2$. This was done separately for total (**A**) and newly transcribed RNA (**B**). For each set of genes classified as regulated, the fraction of genes is shown which were also identified by Scherbik *et al.* (2007). These results confirm that different genes are induced by IFNα and IFNγ since the overlap in the latter case to the regulated genes identified by Scherbik *et al.* is much lower.

As expected, the fraction of differentially expressed genes overlapping with the gene set identified by Scherbik *et al.* after IFNα treatment is higher for the IFNα treated cells than for the IFNγ treated cells. This confirms that although there is a significant overlap between IFNα and γ regulated genes, a large fraction of genes is regulated by only one of these cytokines (Sanda *et al.*, 2006). If we classify genes as regulated based on expected changes in total RNA, overlaps are actually higher than if predictions were made based on the actual observed changes. Using the consensus of both sets, overlaps could be further increased. On the other hand, changes in newly transcribed RNA calculated from total RNA are not appropriate for classification on their own. However, using a consensus of both observed and expected changes in nt-RNA, overlap to the predictions of Scherbik *et al.* could be increased significantly.

The number of differentially expressed genes which can be identified from changes in total RNA is relatively small (19 for IFNα treated cells, 30 for IFNγ). As median half-life in murine fibroblasts is ~ 5 h, de novo transcription would have to be increased by a factor > 8 for most genes to observe a two-fold change in total RNA after only 1 h of labeling. Genes for which transcription is less rapidly induced can only be detected from changes in newly transcribed RNA. Here, 244 and 277 genes show at least a two-fold increase of de novo transcription after 1 h of IFNα and IFNγ treatment, respectively. The consensus from observed and expected changes in de novo transcription yields 83 and 157 genes respectively. 35 of these genes are regulated by both IFNα and IFNγ.

The overlap of the consensus set to the Scherbik *et al.* predictions is significantly lower than for the differentially expressed genes identified from total RNA. Less than 35% of the IFNα regulated genes found in this analysis were also recovered in the Scherbik *et al.* study. These results indicate that many differentially expressed genes can be identified from changes in de novo transcription for which no significant change in total RNA abundance is visible after 1 h or even 3 h as in the Scherbik *et al.* study. Thus, we can identify differentially expressed genes from newly transcribed RNA which are only slowly induced and/or have long half-lives.

We evaluated 24 IFNα regulated genes with transcript half-life ≥ 10 h for which both expected and observed changes in newly transcribed RNA are ≥ 2, but total RNA is hardly changed at all ($\beta_N < 1.3$). Of these 24 genes, only 5 were also identified by Scherbik *et al.* (2007). As the mean increase in newly transcribed RNA was greater than 4 for these 5 genes, but only 2.5 for the remaining genes, total RNA abundance was probably increased sufficiently to detect the differential expressions for the 5 genes when Scherbik *et al.* extracted RNA 3 h after the beginning of IFN treatment. Genes with lower rates of RNA de novo transcription however, were still missed, such as the proteasome activator subunit 1 and 2 (PSME1, PSME2) which have been recently shown to be both IFNα and γ inducible (Shin *et al.*, 2007). Transcript half-life for these genes is > 18 h and 1 h after the beginning of IFNα treatment, de novo transcription was only increased by a factor of ~ 2.35 and total RNA only by a factor of ~ 1.12. Although longer IFN treatment would make it possible to detect the regulation of such long-lived transcripts from total RNA as well, fast and transiently regulated transcripts would likely be missed. Using newly transcribed RNA, both fast and transient and slow regulation effects can be detected even after short interferon treatment.

8.4 Discussion

With the newly developed method of Dölken *et al.* (2008) for labeling newly transcribed RNA with 4sU, the relative contributions of de novo transcription and decay of mRNA can be distinguished. For this purpose, we developed novel computational approaches for normalizing the three different RNA fractions (nt-RNA, p-RNA and t-RNA) obtained in the experiments and calculating RNA half-lives. As total RNA is the sum of newly transcribed and pre-existing RNA, microarray measurements can be normalized in an intuitive way using linear regression analysis.

Furthermore, a probe set quality score can be assigned to each probe set based on the distance of the observed nt-RNA/t-RNA and p-RNA/t-RNA ratios from the regression line. If genes are represented by several probe sets on the array, the quality score can be used to select the most reliable one. This probe set selection approach leads to more precise and reliable half-life estimates with a higher reproducibility than the standard approach of averaging estimates from different probe sets for the same gene. The more probe sets a gene is represented by, the higher is the gain in accuracy with the new selection approach. Moreover, as the same probe sets are selected independently for different replicates much

more often than expected randomly, significant quality differences between probe sets are identified.

Using these methods, we estimated median $t_{1/2m}$ in murine fibroblasts at $\sim 4.6 - 5.8$ h (steady state and cell division model, respectively). So far, median $t_{1/2m}$ has been determined in *E. coli* ($t_{1/2m} = 5$ min, cell cycle length (CCL)=20 min, Bernstein *et al.*, 2002), yeast ($t_{1/2m} = 20$ min, CCL=90 min, Wang *et al.*, 2002), *Arabidopsis* ($t_{1/2m} = 3.8$ h, CCL=19 h, Narsai *et al.*, 2007) and human HepG2 and Bud8 cells ($t_{1/2m} = 10$ h, CCL= 50 h, Yang *et al.*, 2003). These studies indicated that median half-life is correlated with the CCL in the corresponding cells with $t_{1/2m} \sim 20 - 25\%$ of CCL. For murine fibroblasts CCL is approximately 24 h and thus median $t_{1/2m}$ is approximately 19-24% of CCL. This confirms that $t_{1/2m}$ is coupled to CCL although the reason for this correlation between CCL and $t_{1/2m}$ is unclear.

A comparison of RNA half-life estimates from newly transcribed or pre-existing RNA showed that half-life estimates are more accurate if determined from newly transcribed RNA. For estimates from pre-existing RNA both after 4sU labeling and inhibition of transcription by act-D, tremendous deviations are observed between different replicates for medium to long half-lives. These estimates are only reliable for extremely short half-lives, whereas estimates from nt-RNA are reliable on the whole range of half-lives. Simulation of measurement errors showed that this effect is dependent on the labeling time. If length of labeling or transcription inhibition were increased beyond medium half-life, precision of half-lives from RNA decay should also increase.

We found that reliable RNA half-life estimates can be obtained after only 15 minutes of labeling. However, for fast-decaying transcripts more newly transcribed RNA is measured after 15 min than consistent with results from longer labeling. As this could not be explained by noise effects, it suggests that the short-lived transcripts may not follow simple first-order kinetics at least briefly after transcription. As differences between different labeling durations decrease with increasing labeling time, these effects can be neglected for longer labeling.

The above results were based on the assumption that total RNA abundance does not change over time. For a more realistic model of RNA decay, the increase in total RNA due to cell growth and division has to be taken into account. Half-life estimates for individual genes can be easily converted from the steady state model to the cell division model without having to repeat normalization. For short half-lives, estimates of the steady state and cell division model differ only to a minor degree. Contrary to that, long half-lives are underestimated in the steady state model whereas measurement errors are amplified by the cell division model. Thus, although the cell division model is more accurate in the biological sense, variations between experiments are overestimated. Even so, estimates are still more precise than estimates from RNA decay. For comparison of different conditions, cell lines or species (see next chapter), the steady state model is more appropriate than the cell division model as estimates are more stable and reproducible between experiments and replicates. Only for rapidly dividing cells such as yeast cells, the cell division model or estimates from p-RNA should be used.

As changes in RNA transcription are observable in newly transcribed RNA before they

have a significant effect on total RNA, differential expression after 1 h IFN treatment was analyzed both on total and newly transcribed RNA. From total RNA less than 20 IFNα induced genes could be identified, most of which had also been found in a previous microarray study of total RNA changes after IFNα treatment (Scherbik *et al.*, 2007). From newly transcribed RNA alone, more than 200 genes were identified for which a significant change in transcription was observed. Using transcript half-lives to compare observed and expected changes in newly transcribed RNA, this set could be restricted to a more confident set of 80-150 differentially expressed genes. Although these genes showed a significant enrichment compared to the results of Scherbik *et al.*, more than 60% of the IFNα induced genes were not identified in this previous study although a large fraction of them have already been characterized as interferon inducible in the literature.

One advantage of analyzing changes in newly transcribed RNA is that slowly induced genes with long half-lives can be identified. We analyzed 24 genes with extremely stable mRNAs whose transcription was up-regulated at least two-fold by IFNα. After 1 h of interferon treatment no significant change was observed in total RNA. Only for 5 of the genes with highest increase in transcription, significant changes were detected in total RNA by Scherbik *et al.* after 3 h. This shows that interferon induced differential expression can be detected from newly transcribed RNA as soon as 1 h after the beginning of interferon treatment which is not obvious in total RNA even several hours later. Thus, sensitivity of gene expression analysis is increased as both fast and transient as well as slow and long-term changes can be studied in one experiment.

8.5 Conclusions

In this chapter, we presented the approach for calculating RNA half-life from measurements of newly-transcribed, pre-existing and total RNA. Using our method, microarray experiments can be normalized in an intuitive way, median half-life of transcripts can be estimated and probe set quality can be assessed to select the most reliable probe set for each gene. Our results show that RNA half-lives calculated from newly transcribed/total RNA ratios are of superior accuracy compared to half-lives estimated from RNA decay. Assuming steady state more reliable estimates are obtained compared to a more biologically realistic cell division model. Based on newly transcribed RNA differential expression can be identified even shortly after changes in conditions and both short- and long-term effects can be quantified. Using half-life estimates, observed changes in total and newly transcribed RNA can be translated to expected changes in newly transcribed and total RNA, respectively, and regulated genes can be identified more confidently.

Chapter 9

A conserved role of RNA half-life

9.1 Introduction

In the previous chapter, we have shown how RNA half-lives can be derived from measurements of newly transcribed and total RNA. These half-lives are of superior accuracy than half-lives previously obtained from measurements of RNA decay after transcriptional arrest. As we found that RNA half-lives cover a large range of values from a few minutes to days, we investigated the function of such diverse half-lives and what can be learned about the regulation of biological process and protein complexes by studying transcript half-lives.

Previous studies on RNA half-lives from RNA decay after transcriptional arrest in yeast, *Arabidopsis* and human (Wang *et al.*, 2002; Gutierrez *et al.*, 2002; Yang *et al.*, 2003; Narsai *et al.*, 2007) found that gene function and transcript half-life seem to be closely related in eukaryotes. These observations and the fact that the majority of ortholog genes in *Arabidopsis* and human fall within the same region of RNA half-lives (Narsai *et al.*, 2007), suggest that RNA half-lives are conserved and play a functional role.

A possible explanation for the correlation between transcript half-life and gene function may be the linear relationship of transcript half-life and response time after a transcriptional stimulus (see Figure 9.1). De novo transcription of long-lived transcripts has to be increased tremendously relative to the normal amount of transcription to observe a significant change in total RNA concentration after a short time. Small changes in transcription can only lead to noticeable changes in total RNA after several hours. Contrary to that, for short-lived transcripts a small relative increase of transcription can lead to a significant and fast up-regulation of total RNA. Furthermore, as stable transcripts remain in the cell for a long time, transcription of the corresponding genes cannot be down-regulated quickly nor up-regulation be switched off in a timely manner. Thus, a short response time requires the fast decay of the corresponding transcripts and thus a high production rate at the steady state. As a consequence, rapid and flexible transcriptional regulation is only possible for short-lived transcripts.

To identify conserved patterns determining transcript half-life across species and cell

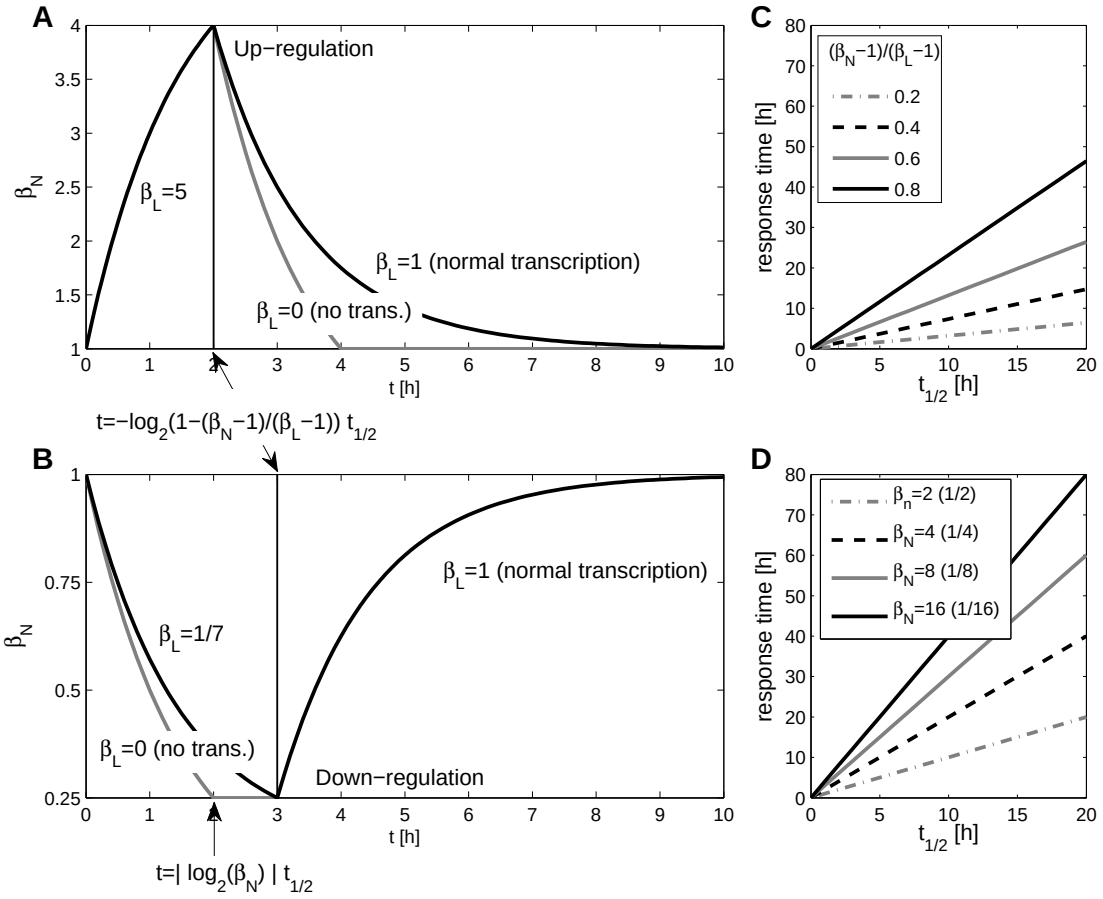


Figure 9.1: Simulation of response time for transcriptional regulation after a stimulus. **A**) Up-regulation of total RNA abundance by a factor $\beta_N = 4$ for $t_{1/2} = 1$ h. De novo transcription is increased by a factor $\beta_L = 5$, until $\beta_N = 4$. At this point, the stimulus is stopped and transcription returns to the normal rate ($\beta_L = 1$) or is switched off completely ($\beta_L = 0$, grey) until total RNA abundance has returned to normal levels. **B**) Down-regulation of RNA abundance ($\beta_N = \frac{1}{4}$, $t_{1/2} = 1$ h) by either down-regulation of transcription ($\beta_L = \frac{1}{7}$, black) or switching off transcription ($\beta_L = 0$, grey). As soon as $\beta_N = \frac{1}{4}$, de novo transcription is returned to the normal rate. **C**) The response time until a certain β_N is reached for a given change in de novo transcription β_L is $t = -\log_2(1 - \frac{\beta_N - 1}{\beta_L - 1})t_{1/2}$ where $t_{1/2}$ is the half-life of the transcript. Response time is illustrated for different values of $\frac{\beta_N - 1}{\beta_L - 1}$. **D**) If transcription is switched of, the response time until total RNA abundance has returned to the normal level after up-regulation or RNA abundance has been down-regulated by a factor β_N is $t = |\log_2 \beta_N| t_{1/2}$. Response time is plotted for several values of β_N. Values before parenthesis indicate the ratio of up-regulation β_N before transcription is switched off and values in parenthesis down-regulation of total RNA by a factor β_N.

types, we compared half-lives determined with the improved accuracy in human B-cells and murine fibroblasts from newly transcribed RNA for thousands of genes (Friedel *et al.*, 2008b). Here, microarray measurements of newly transcribed, pre-existing and total RNA were performed by Lars Dölken from the Max von Pettenkofer-Institut. This genome-scale comparison showed that RNA half-lives are highly conserved and are specifically correlated to gene function. Our results provide further evidence that fast and transient regulation of transcription and signal transduction is facilitated by rapid decay of the corresponding transcripts. Furthermore, transcripts of genes involved in energy or protein metabolism or encoding for protein complex subunits are highly stable.

Previous studies in yeast suggested that transcript decay rates for proteins involved in the same complex are coordinated (Wang *et al.*, 2002). As few annotated complexes were previously available for higher eukaryotes, this could not yet be confirmed for other species. Using the recently published compendium of mammalian protein complexes from the CORUM database (Ruepp *et al.*, 2008), we analyzed the coordination of transcript half-lives within complexes. Although RNA half-lives are highly correlated within protein complexes and also protein families, transcript half-life for individual genes can deviate significantly from other genes in the same complex or family to efficiently support the regulation of important biological processes, such as transcriptional regulation, energy metabolism and apoptosis. Therefore, careful analysis of these highly accurate data on RNA half-life can provide new insights into a broad spectrum of biological processes.

9.2 Methods

9.2.1 Experimental data

To investigate RNA turnover rates in mouse and human, we analyzed RNA half-lives in murine fibroblasts and human B-cells. The mouse fibroblast data set was described in the previous chapter (see section 8.2.1). For our purposes, we used estimates from 1 h 4sU labeling and the two separate experiments were combined. Newly transcribed, pre-existing and total RNA in human B-cells were measured by Lars Dölken and co-workers in the same way as described in the previous chapter. Cells were cultured in the presence of 4-thiouridine (4sU) for 1 h and total cellular RNA was then separated into labeled, newly transcribed RNA and unlabeled, pre-existing RNA. Human B-cells were chosen to compare the fibroblast data with a cell line of both a different species and cell type to identify conserved patterns for RNA half-life not specific to a certain cell type. Three biological replicates each for newly transcribed, pre-existing and total RNA were analyzed using HG U133 Plus 2.0 arrays.

All RNA fractions and replicates for the same species were normalized together with the GCRMA (Wu and Irizarry, 2004) algorithm of the BioConductor package (Gentleman *et al.*, 2004) in R (R Development Core Team, 2007). Normalized p-RNA/t-RNA and nt-RNA/t-RNA ratios were obtained with the linear regression method described in section 8.2.2. Only probe sets were considered with present calls in all replicates for all RNA

fractions. For each gene represented by several probe sets on the array, the highest quality probe set was selected as described in section 8.2.3.

9.2.2 Functional analysis of RNA half-lives

To identify functional groups which are significantly over-represented among short- or long-lived transcripts, we performed a Gene Ontology (GO) over-representation analysis (see section 5.2.2). Generally, GO over-representation analysis is performed by comparing the frequency of a GO term for a specific group of genes against the overall frequency or the frequency within another reference set. Since defining a cut-off to classify transcripts as either short-lived or long-lived would be rather arbitrary, we do not use this method but compare the distribution of half-lives for genes within a certain functional category against the overall distribution of half-lives. The significance of differences in the distributions is calculated with the Kolmogorov-Smirnov test (K-S test) in R. The same approach was used by Narsai *et al.* (2007) for the analysis of *Arabidopsis* half-lives.

P-values were corrected for multiple testing with the method of Benjamini and Yekutieli (2001), a more conservative version of the method of Benjamini and Hochberg (1995), which controls the false discovery rate (FDR) and does not require the tests to be independent. GO annotations for human and mouse were taken from the Gene ontology website (`http://www.geneontology.org/`). We only analyzed GO categories with at least 10 annotated genes. The root terms "biological process", "molecular function" and "cellular component" as well as direct subterms of the root terms were not considered as they are too unspecific. Correction of p-values was performed for all ontologies taken together and statistically significant results were determined at a significance level of 0.01.

9.2.3 Analysis of protein complexes and families

Protein complexes for human and mouse were taken from the CORUM database (Ruepp *et al.*, 2008) which provides a manually annotated collection of protein complexes in mammals. We extracted 1185 complexes for human and 285 complexes for mouse (CORUM version from June, 2008). Using orthology tables between human and mouse genes from the mouse genome database (MGD) (Eppig *et al.*, 2007), the human and mouse complexes were combined to a large set containing 1434 protein complexes. Since isoforms of protein complexes are described as individual entries in the CORUM database, some protein complexes are effectively represented several times in different variations. To avoid a bias due to the over-representation of some complexes, diversity of half-lives in complexes was calculated based on the complex co-membership network. In this network, proteins are connected if they are subunits of the same protein complex.

Protein family annotations for human and mouse were taken from the Pfam database (Finn *et al.*, 2008). We used the Pfam-A collection in which each family is described by a curated seed alignment for representative family members and additional family members have been identified by a search of sequence databases with the family profile Hidden Markov Model (HMM) built from the seed alignment. As families can overlap to a large

degree, we determined co-family networks for both species to analyze the similarity of half-lives in protein families.

The difference in half-life between two proteins p_1 and p_2 connected in a co-complex or co-family network was calculated as the fold change between the corresponding transcript half-lives h_1 and h_2:

$$fc(h_1, h_2) = \max\left(\frac{h_1}{h_2}, \frac{h_2}{h_1}\right). \tag{9.1}$$

Average fold change of half-lives in protein families and complexes was calculated as the geometric mean of all connected protein pairs and compared against results for random complexes or families and random half-lives (10,000 randomizations each). Protein complexes or families were randomized by repeatedly exchanging two random proteins from different complexes or families. Using this approach, the number of proteins for each complex or family and the number of complexes or families per protein remain constant. Half-lives were randomized by repeatedly exchanging the half-life of two random proteins. The resulting p-values were calculated as the fraction of randomizations for which the average fold change within complexes or families is lower or equal to the observed average.

9.3 Results

9.3.1 Median RNA half-life in murine fibroblasts and human B-cells

In the previous chapter, median half-life $t_{1/2m}$ in mouse fibroblasts was estimated with the linear regression method at 274 ± 49 min using the steady-steady model. Based on the 1 h labeling experiment, median $t_{1/2}$ in human B-cells was estimated at 315 minutes with the 95% confidence interval between 248 and 382 minutes. Taking into account the increase in total RNA due to cell growth and division, the estimates are slightly larger with 349 min in murine fibroblasts and 409 minutes in human B-cells (CCL of 24 h). Thus, in the steady state model median half-lives of human and mouse are quite similar ($\sim$ 5 h) but differ in the cell division model (5.8 h in mouse compared to 6.8 h in human). Nevertheless, even in this case the difference is not statistically significant due to the small number of replicates and large variation between estimates for individual replicates.

Half-life in human HepG2 cells was previously reported as $\sim$ 10 h (Yang *et al.*, 2003). Even if we take into account cell growth and division, median half-life $t_{1/2m}$ in human B-cells is at least 3 hours smaller. This suggests that median half-life can differ significantly between different cell types for the same organism. As mentioned previously, median half-life has been shown to correlate with the CCL of the cells. As the human HepG2 cells analyzed by Yang *et al.* had a CCL of $\sim$ 50 h which is about twice the CCL of the human B-cells, $t_{1/2m}$ in B-cells is expected to be about half of $t_{1/2m}$ in HepG2 cells. This corresponds approximately to the steady state estimate of $\sim$ 5 h.

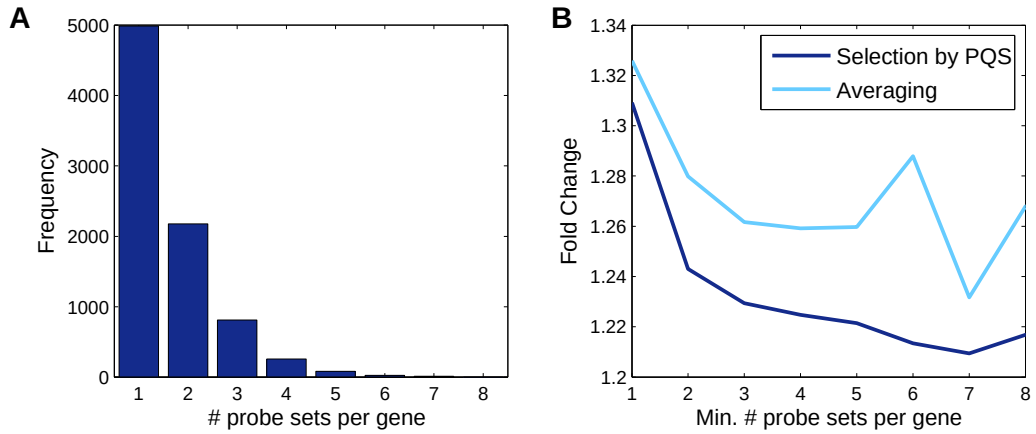


Figure 9.2: Probe set quality control in human B-cells. **A)** Distribution of the number of probe sets per gene for human. Only probe sets are considered with present calls in all replicates for newly transcribed, pre-existing and total RNA. **B)** Average fold change in half-life estimates for human B-cells between different replicates (y-axis) was calculated separately for all genes represented by at least 1, 2, 3, ..., 9 probe sets (x-axis). We compared the probe set selection method based on the probe set quality score (PQS) against the approach of averaging half-life estimates for different probe sets for the same gene. Fold changes are reduced significantly by the spot selection method and generally decrease with the minimum number of probe sets per gene.

9.3.2 Conservation of transcript half-life

Transcript half-lives were compared between human B-cells and mouse fibroblasts using the measurements from newly transcribed RNA since they are more precise than half-lives from measurements of pre-existing RNA (see section 8.3.2). Half-lives were calculated with the steady state model as results are more comparable and reproducible between experiments and replicates (see section 8.3.4). For each gene, the probe set with the highest quality score was selected. In this way we obtained half-life estimates for more than 8000 genes each for human and mouse.

For murine fibroblasts, we have previously shown that selection of probe sets based on the probe set quality score (PQS) increases accuracy and decreases variability of half-life estimates between replicates. The same analysis was also performed for the human B-cells (Figure 9.2). Again our results showed that probe set selection improves fold changes between replicates significantly compared to averaging of results. Furthermore, this confirms that quality differences can be identified between probe sets as the same probe set for a gene is chosen independently by all 3 replicates with significantly higher frequency than expected at random.

RNA half-lives are not normally distributed in both species but follow approximately a log-normal distribution with a long right tail (see Figure 9.3 **A-B**). Using orthology tables between human and mouse genes from the mouse genome database (MGD) (Eppig *et al.*, 2007), half-life estimates were compared for about 5000 genes with estimates for both

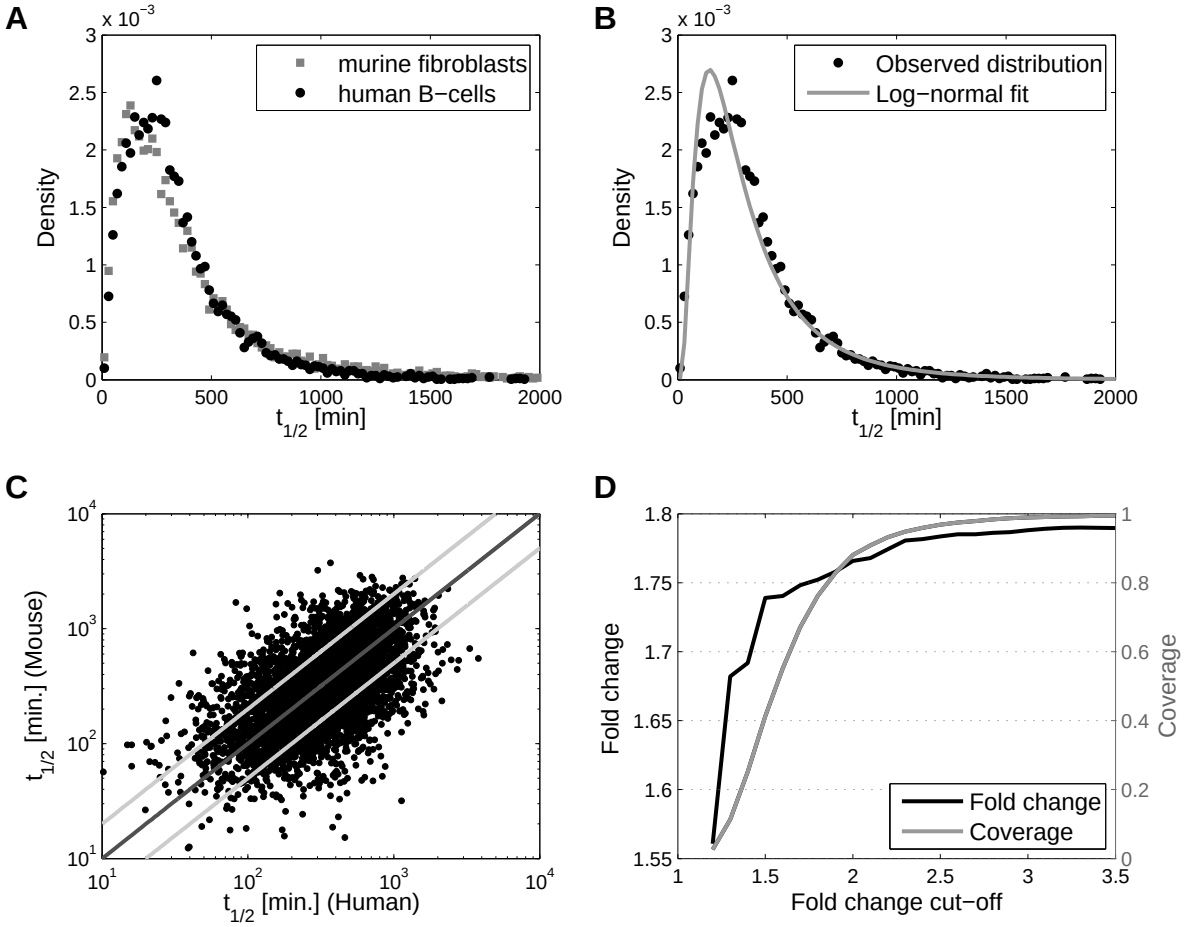


Figure 9.3: Conservation of RNA half-lives between murine fibroblasts and human B-cells. **A)** Distribution of half-lives in murine fibroblasts and human B-cells. **B)** Log-normal distribution fitted to distribution of half-lives in human B-cells. **C**) Comparison of half-lives between human and mouse. The dark grey diagonal indicates equal half-lives and the light grey lines a two-fold deviation. **D)** Average fold change between species was calculated selectively for those genes for which the fold change among different replicates in the same species is no higher than a specific cut-off (x-axis). For each cut-off, average fold change (black) between species for the selected genes and coverage (grey), i.e. fraction of genes included in the calculation of fold change, are shown. Fold changes decrease significantly if more selective cut-offs are chosen. Thus, the variation in the cross-species comparison is correlated to the variation in each individual species.

species (see Figure 9.3 **C**). The mean fold change between human and mouse was fc=1.8 and $\sim 67\%$ of genes were within the two-fold range. This deviation is significantly larger than observed for half-life estimates from newly transcribed RNA between the two separate experiments for mouse fibroblasts (fc=1.5, 84% of genes within the two-fold range) and in different replicates from the same experiment in mouse (fc=1.18, 98%) and human (fc=1.3,

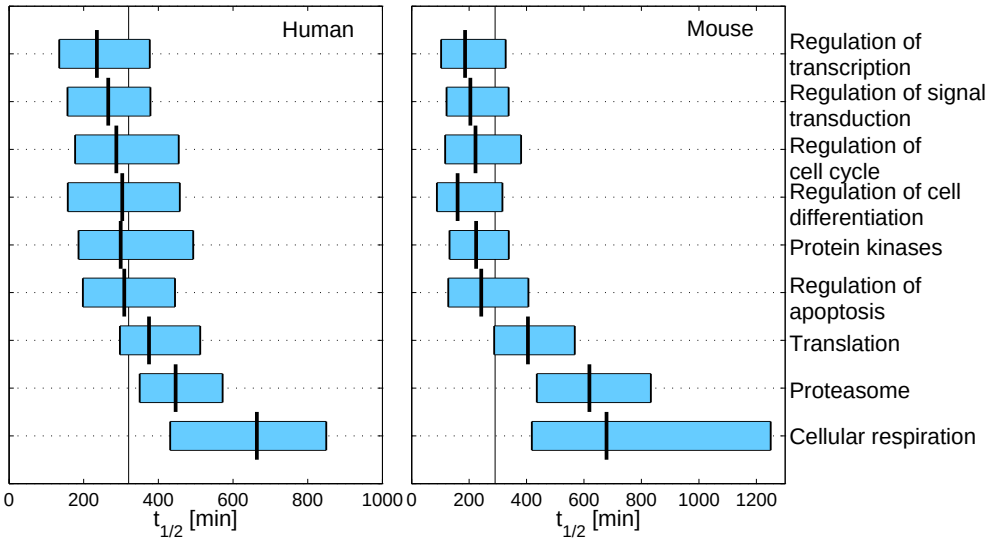


Figure 9.4: Characteristic RNA half-lives for different GO terms. Horizontal bars indicate the range of half-lives between the lower and upper quartile for each functional group. The median half-life for each GO term is indicated by short vertical lines and the overall median over all functional groups by the long vertical line.

92%). However, it is comparable to the variation between half-life estimates from newly transcribed and pre-existing RNA (fc=1.75 and fc=1.9 in human and mouse, respectively). These results show that RNA half-lives are much more conserved across species and cell types than a comparison from RNA decay would suggest (fc=2.44 between species for half-life estimates from pre-existing RNA).

Furthermore, a significant difference between the species was only observed for 18 ($< 1\%$) genes (t-test for unequal variance, FDR corrected p-values < 0.01). For three of these genes, the overall tendency is nevertheless the same, i.e. half-lives are either long or short in both species. This means that only for very few genes a qualitative difference is observed between human and murine cells. In addition, average fold changes in the genome comparisons can be decreased by filtering genes based on the variation between different replicates for the same species. Thus, the lower the variability within a species, the lower tends to be the variability between species (see Figure 9.3 **D**). These results suggest that a large fraction of the observed variability is due to noise and that RNA half-lives are highly conserved between species.

9.3.3 Regulation by fast transcript decay

Since previous studies have suggested a correlation between RNA half-lives and gene function, we identified GO categories significantly associated with either short or long half-lives separately for murine fibroblasts and human B-cells. We found that short-lived tran-

scripts were characteristic for genes involved in the regulation of transcription (FDR p-value $< 10^{-16}$) and signal transduction (p-value = 0.0066 (human) and 1.78×10^{-4} (mouse), see also Figures 9.4 and 9.5). For transcription, this has previously been confirmed both for *Arabidopsis* and human HepG2 and Bud8 cells (Gutierrez *et al.*, 2002; Yang *et al.*, 2003; Narsai *et al.*, 2007). Fast decay of signal transduction transcripts has so far only been reported for *Arabidopsis* (Narsai *et al.*, 2007).

Since short RNA half-lives allow a rapid and transient regulation on the transcriptional level, we investigated whether regulatory proteins in general show preferentially short RNA half-lives or whether this is limited to transcriptional and signal transduction regulation. For both species, we did not find a significant difference between the overall distribution of half-lives and the distribution for regulatory genes not involved in transcription and signal transduction. Contrary to that, a significant difference was observed between transcriptional and signal transduction regulatory proteins and the remaining regulatory proteins. Thus, specifically the regulation of transcription and signal transduction is supported on the transcriptional level by short half-lives. For other processes regulation may preferentially be performed post-transcriptionally, e.g. by protein modifications or protein decay by the proteasome after ubiquitination.

Furthermore, molecular functions and cellular components involved in transcription were also preferentially associated with short half-lives: transcription factor activity (p-value $< 10^{-10}$), DNA binding and zinc ion binding (p-value $< 10^{-16}$) and nucleus (p-value $< 10^{-8}$). Many zinc ion binding proteins are transcription factors such as kruppel-like factors, e.g. KLF10 ($t_{1/2} \sim 1$ h), or involved in signal transduction such as RASSF1 ($t_{1/2} \sim 25$ min). Median half-life of zinc finger proteins not associated with transcription, signal transduction or DNA binding is significantly higher than for all zinc binding proteins but still below the overall median. As the function of many of these genes is unknown, short RNA half-life may indicate so far unidentified transcription factors.

Median half-life of genes involved in the regulation of signal transduction is about 20-30 min higher than for transcriptional regulators but still about 1 h less than the overall median. Extremely short half-lives of < 1 h are observed for instance for the Cbp/p300-interacting transactivator (CITED2) (~ 25 min), NF-kappa-B inhibitor alpha (NFKBIA) and BCL6 (~ 45 min.), the G-protein-coupled receptor induced protein TRIB1 and TNF alpha induced protein 3 (TNFAIP3) (~ 1 h). CITED2, BCL6 and NFKBIA are also involved in the regulation of transcription (Bamforth *et al.*, 2001; Arima *et al.*, 2002; Kiernan *et al.*, 2003). TRIB1 belongs to a family which regulates the activity of MAP kinases and its mRNA has been previously shown to have short half-lives (Kiss-Toth *et al.*, 2004). TNFAIP3 is a highly regulated and rapidly induced anti-apoptotic gene (Liuwantara *et al.*, 2006).

Apart from transcription and signal transduction regulation and their super-categories, no other regulation category was significantly enriched for fast decay rates in human B-cells. In murine fibroblasts, we found that genes involved in regulation of cell differentiation and developmental processes (p-value=5.1×10^{-4}) and protein kinases (p-value=9.6×10^{-6}) were also significantly characterized by short half-lives (see Figure 9.4). Genes involved in development have also previously been found to be fast-decaying in human HepG2 cells

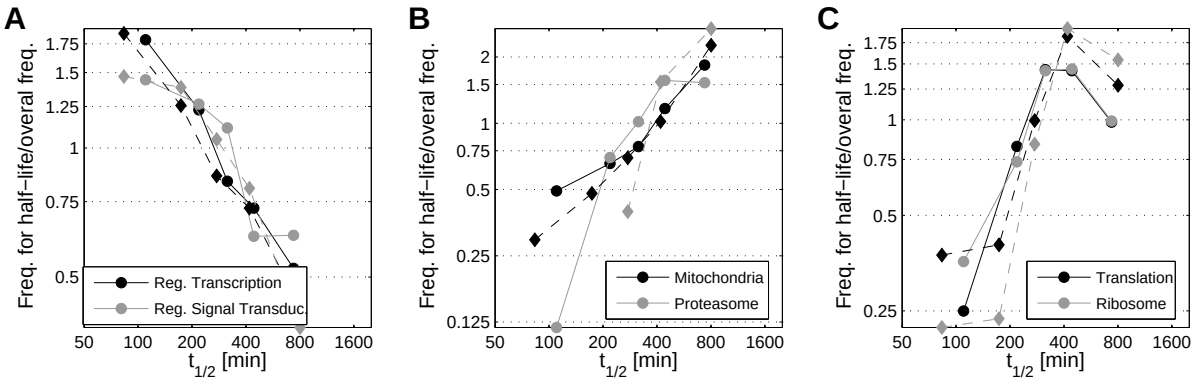


Figure 9.5: Correlation between transcript half-life and gene function. Frequency of GO categories for specific ranges of RNA half-life divided by the overall frequency of this GO category is plotted against RNA half-life. Results are shown for (**A**) genes involved in regulation of transcription (black) and signal transduction (grey), (**B**) mitochondrial (black) and proteasomal proteins (grey) and (**C**) translational (black) and ribosomal (grey) proteins. Results for human are indicated with solid lines, and results for mouse with dashed lines.

but not Bud8 cells (Yang *et al.*, 2003) and kinases in *Arapidopsis* (Narsai *et al.*, 2007). Furthermore, apoptosis and cell cycle transcripts were found to decay fast in human HepG2 cells. However, this observation could not be confirmed in the human B-cells and mouse fibroblasts. Although median half-life of transcripts involved in regulation of apoptosis and cell cycle is lower than the overall median this difference is not significant (see also Figure 9.4).

9.3.4 Slow transcript decay for energy and protein metabolism

Significant enrichment for very long RNA half-lives was found for genes involved in cellular respiration and energy metabolism. This has already been observed in yeast, human HepG2 and Bud8 cells and *Arabidopsis* (Wang *et al.*, 2002; Yang *et al.*, 2003; Narsai *et al.*, 2007). Our results show that this is true for all parts of cellular respiration. Enzymes and protein complexes involved in glycolysis, oxidative decarboxylation, the pentose phosphate pathway, the citric acid cycle and oxidative phosphorylation consistently have half-lives of more than 5 hours both in human and mouse (see Table 9.1). As most energy metabolism pathways take place in the mitochondria, mitochondrial proteins, in particular those located at the inner mitochondrial membrane, were also preferentially associated with long transcript half-lives.

Despite the overall long half-lives of cellular respiration protein, one interesting outlier was observed. While hexokinase I (HK-I) transcripts decay slowly ($t_{1/2} \sim 9$ h), transcripts of hexokinase II (HK-II) show short half-lives ($t_{1/2} \sim 1 - 4$ h) in both human and mouse. Phosphorylation of glucose by hexokinases is the first step of the glycolysis and hexokinase I is considered a "housekeeping gene" whose mRNA levels remain stable despite alterations

Enzyme	Gene(s)	$t_{1/2}$ [h] (human)	$t_{1/2}$ [h] (mouse)
Oxidative Phosphorylation			
NADH dehydrogenase	30	11.3	21.6
ATPases	16	11.6	14.3
Ubiquinol-cytochrome c reductase complex	6	10.0	17.4
Glycolysis and Pentose Phosphate Pathway			
Hexokinase 1	HK1	9.8	8.5
Hexokinase 2	HK2	3.6	0.9
Glucose-6-phosphate isomerase	GPI	18.4	14.3
6-Phosphofructokinases	PFKP, PFKM, PFKL	9.7	19.9
6-phosphogluconolactonase	PGLS	10.3	21.4
Fructose bisphosphate aldolases	ALDOA	6.8	9.2
Pyruvate kinase	PKM2	11.7	7.2
Pyruvate dehydrogenase	PDHA1, PDHB	14.5	6.5
Phosphogluconate dehydrogenase	PGD	11.4	6.6
Phosphopentose isomerase	RPIA	10.9	6.9
Transaldolase	TALDO1	9.9	28.6
Citric Acid Cycle			
Citrate synthase	CS	7.6	10.6
Aconitases	ACO1, ACO2	12.8	9.7
Isocitrate dehydrogenases	5	12.6	19.1
Oxoglutarate dehydrogenase	OGDH	5.1	13.0
Succinyl coenzyme A synthetases	SUCLG2, SUCLG1, SUCLA2	14.1	18.8
Succinate dehydrogenase	SDHC, SDHD, SDHA	6.8	21.1
Fumarase	FH	7.0	9.6
Malate dehydrogenases	MDH1, MDH2	10.7	18.6
Translation			
Ribosome (cytosolic)	55	6.2	7.5
Ribosome (mitochondrial)	55	6.8	9.4
Initiation Factors	29	5.4	4.1
Elongation factors	6	5.6	6.4
tRNA synthetases	20	6.3	6.2
Proteasome			
Proteasome	37	7.3	8.9

Table 9.1: Half-life estimates for genes involved in cellular respiration or protein synthesis and degradation. Gene names are shown for groups of no more than 3 proteins, otherwise the number of genes in that group are shown. Only genes were included for which half-life estimates were available in both human and mouse.

in glucose or insulin levels (Printz *et al.*, 1993) or feeding conditions (Soengas *et al.*, 2006). Contrary to that, expression of HK-II is induced by a variety of stimuli, such as glucose and insulin (Printz *et al.*, 1993; Osawa *et al.*, 1995; Jones and Dohm, 1997; Riddle *et al.*, 2000) which accelerates hexose catabolism. While HK-I is expressed in most tissues and is the most abundant hexokinase in tissues which do not depend on insulin to stimulate glucose uptake, HK-II is the predominant hexokinase isoenzyme in tissues responsive to insulin (Printz *et al.*, 1993). Increased expression of HK-II is also often associated with cancer cells (Mathupala *et al.*, 1995).

Thus, although these genes have the same function, different regulatory patterns have evolved to support non-redundant functional roles. These differences in gene regulation are reflected in the transcript half-lives. As HK-I is not regulated on the transcriptional level, stable protein concentrations are maintained by slow RNA turnover. Transcription of HK-II can be induced rapidly but can also be turned off promptly if the stimulus, e.g. insulin, is no longer present as the excess mRNA is quickly decayed.

Apart from energy metabolism, we found that translation in ribosomes and protein degradation by the proteasome, is characterized by long half-lives (see Table 9.1). However, while frequency of energy metabolism or proteasomal genes increases steadily with increasing RNA half-life and frequency of transcription and signal transduction regulatory genes decreases steadily, the frequency of translational genes, e.g. ribosomal proteins or translation initiation factors, peaks in the medium-to-long half-life range (see Figure 9.5). Nevertheless, the frequency of translational genes with extremely long-lived transcripts is still significantly higher than the frequency with very short-lived transcripts.

9.3.5 Coordination of transcript half-lives in protein complexes

The functional analysis showed that large protein complexes involved in energy and protein metabolism, i.e. the NADH dehydrogenase and ATPase complexes, the ribosome and the proteasome, were characterized by significantly longer transcript half-lives than the overall median. For smaller complexes we did not find a significant enrichment, however, the GO category "protein complex" as such was identified as significantly associated with slow decay rates. This was also true if all genes encoding for subunits of the above mentioned large complexes were not included (K-S test, p-value $< 10^{-7}$) and for subunits of the human and mouse complexes from the CORUM database (K-S test, p-value $< 3 \times 10^{-4}$).

For yeast complexes, decay rates of transcripts encoding subunits of the same protein complex were found to be highly coordinated (Wang *et al.*, 2002). To investigate whether this is also the case for human and mouse complexes, we analyzed average fold changes between protein subunits of the same complexes. Fold changes were compared against results for random complexes and half-lives, respectively (see section 9.2.3). In agreement with previous results, we found that transcript half-lives of subunits of the same complexes are indeed significantly more similar (p-value $< 10^{-4}$) than observed for random complexes or half-lives (see Figure 9.6). Here, the enrichment compared to random complexes was smaller than compared to random half-lives since transcript half-lives in protein complexes are generally high and, thus, much more similar by default.

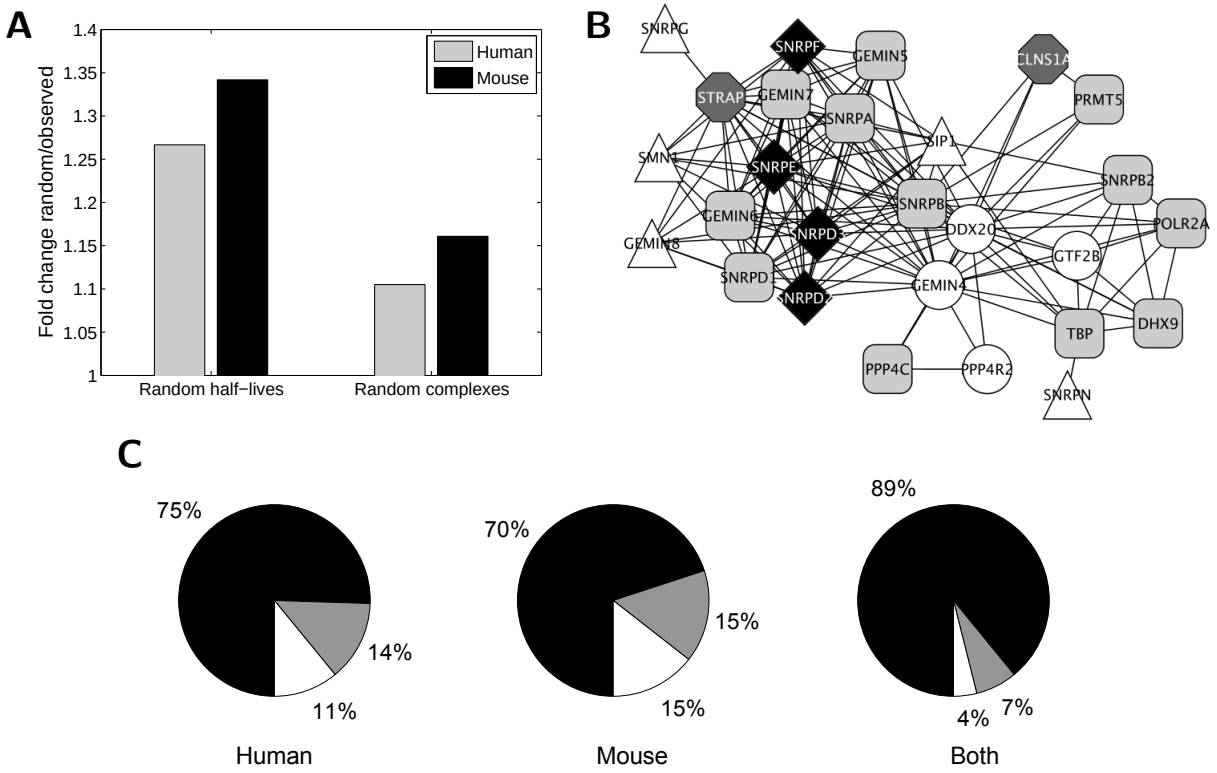


Figure 9.6: Similarity of RNA half-lives in protein complexes. **A**) Comparison of observed fold changes within complexes to fold-changes for random half-lives and random complexes. Here, ratios of fold changes for random half-lives and random complexes (average over 10,000 randomizations) to observed fold changes are shown. As all ratios are larger than 1, observed fold changes are significantly lower than random ones (p-values $< 10^{-4}$). **B**) Protein complexes containing the DDX20 protein are shown. Proteins are connected by an edge if they are subunits of the same complex. Proteins for which half-life estimates could not be obtained for at least one species are indicated by triangles. White circles indicate short, light grey rounded rectangles medium, dark grey octagons long and black rotated squares very long half-lives. **C**) Fraction of proteins for which transcript half-life is coordinated in all complexes they are part of (black), which deviate significantly for at least one complex but not all (grey) and which deviate significantly from all corresponding complexes (white). Proteins are classified as significantly deviating if the average fold change to other proteins in this complex is at least 50% higher than the fold change among the remaining proteins. For dimers, a fold change > 2 is considered as a significant deviation. In the comparison of human and mouse, deviation of a protein to a complex was taken as the minimum of the deviation in human and mouse, respectively.

Complexes with highly coordinated half-lives comprise the TRF1 telomere length regulation complex (fc $\sim$ 1.12), chaperonin CCT ring complex (fc $\sim$ 1.2), the proteasome (fc $\sim$ 1.5) and the NADH dehydrogenase (fc $\sim$ 1.52). The average fold change across

all complexes is 1.82 in human and 2.01 in mouse which is about 21-26% lower than for random half-lives and 9-14% lower than for random complexes. Despite this relatively high similarity of RNA half-lives in complexes, individual proteins can deviate significantly in transcript half-life from the remaining proteins in a complex.

One example is the SMN complex (see Figure 9.6) which is involved in many parts of RNA metabolism, such as snRNP assembly, pre-mRNA splicing and transcription (Pellizzoni *et al.*, 2001). Overall half-life in this complex is relatively high (> 6 h) with some subunits having half-lives > 9 h. However, two subunits, DDX20 and GEMIN4, have consistently short transcript half-lives in both human B-cells and murine fibroblasts. The DExD/H box RNA helicase DDX20 ($t_{1/2} \sim 1.5$ h) also interacts with several transcription factors and can act as a transcriptional repressor (Fuller-Pace, 2006). This suggests that RNA half-life of DDX20 is determined by its role in transcription regulation and not coordinated with most of the other SMN complex subunits. Furthermore, GEMIN4 ($t_{1/2} \sim 0.85 - 3.6$ h) forms an independent complex with DDX20 (Mourelatos *et al.*, 2002). Thus, within this smaller complex its half-life is coordinated.

We evaluated which fraction of proteins that deviate significantly in transcript half-life within a complex can be explained by half-life coordination with other complexes. For this purpose, proteins were classified as deviating from a complex if the fold change to the rest of the complex was increased by more than 50% compared to the fold change between the remaining complex subunits (fc > 2 for dimers). We evaluated the fraction of proteins whose transcript half-life was coordinated (i) with all complexes, (ii) at least one but not all and (iii) none of the complexes they are part of (Figure 9.6).

If we analyze proteins separately for human and mouse, we find that 25-30% of proteins deviate significantly from the remaining subunits of at least one of the complexes they associate with. In about 47% of these cases, the deviation cannot be explained by an association with another complex. In a combined analysis we find that a significantly smaller fraction of proteins deviates significantly both in human and mouse and only 51 proteins (4%) deviate significantly from all of the complexes they are contained in. Although this is a conservative estimate on the number of deviating complex subunits, these results show that half-life of most proteins is coordinated with other proteins in the same complex. Furthermore, only for a small fraction of deviating complex subunits can this deviation not be explained by half-life coordination with other complexes.

Examples for such outliers are given in Table 9.2. For instance in the CDT1-DDB1-CUL4A-RBX1 complex, the DNA replication factor CDT1 is periodically degraded by DDB1-CUL4A-RBX1 ligase after ubiquitination at the beginning of the S phase of the cell cycle. The CDT1 mRNA is down-regulated at cell cycle exit but is up-regulated again at the beginning of cell cycle (Xouri *et al.*, 2004). Up- and down-regulation of CDT1 abundance is supported by short transcript half-lives but stable protein and mRNA concentrations of the ligase subunits are maintained by slow RNA turnover. Thus, proteins which are regulated at the transcriptional level within a complex may be identified from their transcript half-lives.

In most of these cases the deviating complex subunits have short half-lives while overall half-life in the complex is long. Two examples indicate that complexes may be fast and ef-

Gene	$t_{1/2}$ **[h] (Protein)**	$t_{1/2}$ **[h] (Complex)**	**Complex name**
CDT1	1.2/1.36	9.0/7.5	Ubiquitin E3 ligase (CDT1, DDB1, CUL4A, RBX1)
AKAP1	2.7/1.85	10.8/9.3	AMY-1-S-AKAP84-RII-beta complex (signal transduction)
ASB6	1.0/0.66	4.5/4.6	Ubiquitin E3 ligase (ASB6, TCEB1, TCEB2, CUL5, RNF7)
BAX	38.8/12.14	2.1/3.3	BAX-BAK-IRE1alpha complex (apoptosis)
ST13	20.9/18.52	6.0/6.4	ASF1-histone containing complex
ARID2	2.3/1.71	8.6/5.7	PBAF complex (transcription, signaling)
SIAH1	0.9/1.52	4.4/4.9	Ubiquitin E3 ligase (SIAH1, SIP, SKP1A, TBL1X)
ZNF281	1.3/0.55	4.8/2.3	UTX-MLL2/3 complex (transcription, posttransl. mod.)
INOC1	1.6 /1.14	6.7/5.2	INO80 chromatin remodeling complex
MSL3L1	13.0 /18.21	3.8/1.6	MSL complex (DNA processing, protein modification)
BXDC2	1.5/1.63	6.1/6.5	Nop56p-associated pre-rRNA complex (ribosome biogenesis)
EPC1	1.1/0.69	5.1/2.8	NuA4/Tip60 HAT complex (transcription, cellular response to DNA damage, cell cycle control)
BCL6	1.0/0.58	4.8/4.5	Mi-2/NuRD-MTA2 complex (transcription repression, DNA processing, protein modification)
ZNHIT1	23.7/19.27	7.7/7.0	SRCAP-associated chromatin remodeling complex
TOE1	0.8/1.09	6.3/5.3	12S U11 snRNP (mRNA processing)

Table 9.2: Examples for complexes with diverging protein subunits with respect to RNA half-life. This table shows the 15 proteins with highest deviation of transcript half-life from all of the complexes they are contained in. For all of these proteins fold change to the remaining subunits in the complex is at least 90% higher than average fold change between these subunits. RNA half-lives in human/mouse are given as well as the complex to which deviation is lowest. With the exception of 4 complexes, half-lives of the deviating subunits are significantly lower than the overall half-life in the complex. This suggests that these subunits are regulated on the transcriptional level and that they are important for transcriptional regulation of the complexes. It is unclear, however, why in 4 cases individual subunits have significantly higher half-lives than the rest of the complex.

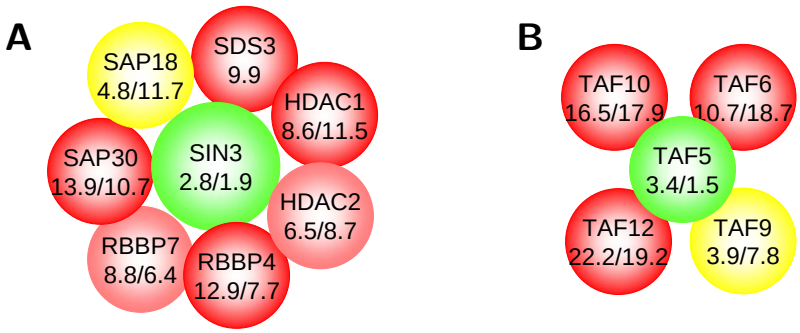


Figure 9.7: RNA half-lives for the SIN3 (**A**) and TFIID core (**B**) complexes. Short half-lives are indicated in green, median half-lives in yellow and long and very long half-lives in light red and red, respectively. Half-lives (in h) for human/mouse are also indicated. For SDS3, no half-lives could be obtained for murine fibroblasts. For TAF6, half-lives in murine fibroblasts were taken from the 30 min labeling experiments. For each complex, only one protein (center) has consistently short half-lives in both species.

ficiently regulated by regulating only the transcription of a few subunits: the SIN3/histone deacetylase (HDAC) complex and the TFIID complex. Even though these complexes are involved in the regulation of transcription which is generally characterized by short transcript half-lives, most of the complex subunits have long half-lives (see Figure 9.7).

The SIN3/HDAC complex consists of a highly conserved core of 8 subunits: SIN3, HDAC1, HDAC2, RBBP4, RBBP7, SAP30, SAP18 and SDS3 (Silverstein and Ekwall, 2005). Short transcript half-lives are only found for SIN3 which can interact with many other proteins with diverse functions and recruits them to the SIN3/HDAC core complex. In this way, it acts as a global transcriptional regulator for many cellular processes. Since SIN3 is an essential component of the SIN3/HDAC complex, this complex and, accordingly, several biological processes can be regulated efficiently by regulation of SIN3 expression.

The TFIID complex has five conserved subunits in common with the SAGA, PCAF, and TFTC complexes (Ogryzko *et al.*, 1998; Grant *et al.*, 1998; Wieczorek *et al.*, 1998; Albright and Tjian, 2000): the highly conserved TBP-associated factor (TAF) proteins TAF5, TAF6, TAF9, TAF10 and TAF12. These shared TAF proteins are necessary for the expression of 70% of genes (Lee *et al.*, 2000) but only TAF5 and the structurally similar TAF5L protein (Ogryzko *et al.*, 1998) have short RNA half-lives (1-3.5 h) both in human and mouse. Contrary to that, the other four proteins are characterized by medium to extremely large RNA half-lives ($>$ 16 h). Here, TAF6 half-life estimates were taken from 30 min labeling experiments in mouse as no estimates were available from 1h labeling. Our results suggest that a few complex subunits with short-half-lives are sufficient to regulate these complexes on the transcriptional level despite overall long RNA half-lives. This indicates that SIN3 and TAF5 play a central role in the regulation of the corresponding complexes. For TAF5 this fits well with results showing that it plays an important structural role in the formation of the complex (Leurent *et al.*, 2004) and is necessary for the DNA promoter binding activity (Boyer-Guittaut *et al.*, 2005).

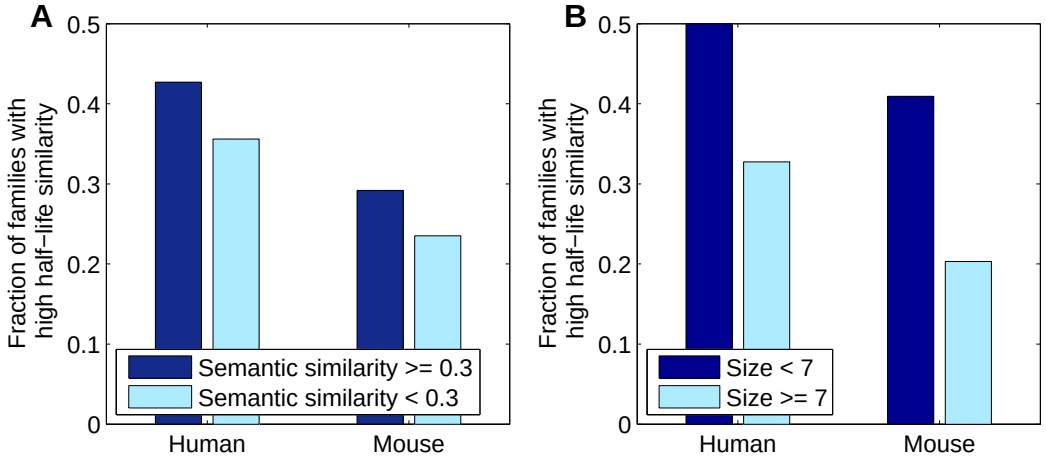


Figure 9.8: Correlation of RNA half-life similarity to functional similarity (**A**) and size (**B**) of protein families. Functional similarity was evaluated in terms of semantic similarity of GO annotations from the "molecular function" ontology (see section 6.2.5). Protein families were divided into two sets based on their semantic similarity (threshold 0.3, **A**) or size (threshold 7, **B**). For each set we calculated the fraction of families with high similarity in RNA half-life (average fold change < 1.8). If protein families were subdivided based on functional similarity, no significant difference was observed between the two sets (Fisher's exact test). If size was used as a criterion to divide families into two sets, the fractions differed significantly.

9.3.6 Similarity of transcript half-lives in protein families

In the previous sections, we have shown that transcript half-life is correlated to gene function. Since protein family members generally have similar functions, we investigated whether this functional similarity also correlates with a high similarity of transcript half-lives in protein families. Our results showed that transcript half-lives were more diverse within protein families than in protein complexes (fc ~ 2.2). Even so, fold changes were still statistically significantly smaller than for random families or random half-lives (p-value $< 10^{-4}$).

Fold change in families is significantly correlated with family size (Spearman correlation coefficient ~ 0.21, p-value $< 10^{-6}$). To evaluate if this is due to a higher functional similarity within small protein families, we calculated functional similarity within protein families using the semantic similarity measure on the "molecular function" ontology of the Gene Ontology (GO) (Schlicker *et al.*, 2006, see section 6.2.5). However, although size and functional similarity of protein families was negatively correlated (Spearman correlation coefficient ~ -0.11, p-value < 0.05), no significant correlation was observed between half-life similarity and functional similarity in families.

Furthermore, the fraction of protein families with high half-life similarity (fc < 1.8) was not significantly lower in families with low functional similarity (semantic similarity < 0.3) than in more functionally homogeneous families (see Figure 9.8). Contrary to that,

Gene	Symbol	$t_{1/2}$ [h] (human)	$t_{1/2}$ [h] (mouse)
Anti-apoptotic			
BCL-2	BCL2	3.80	3.74
BCL-XL	BCL2L1	1.96	1.11
BCL-W	BCL2L2	NA	1.75
A1	BCL2A1	3.72	NA
MCL1	MCL1	1.07	0.70
Pro-apoptotic			
BAX	BAX	**38.77**	**12.14**
BAK	BAK1	2.13	3.28
BOK/MTD	BOK	*4.64*	NA
BID	BID	**10.02**	*4.21*
BIM/BOD	BCL2L11	3.99	0.58
BAD	BAD	**11.46**	NA
HRK	HRK	*5.14*	NA
PUMA	BBC3	1.47	NA
BIK	BIK	*4.66*	NA
Uncategorized			
BCL-Rambo	BCL2L13	*4.09*	2.22

Table 9.3: RNA half-life estimates for genes in the BCL-2 family which contains both anti- and pro-apoptotic genes. Classification into anti- and pro-apoptotic genes was taken from Youle and Strasser (2008). Long transcript half-lives are indicated in bold and highlighted in grey and median transcript half-lives are shown in italics.

this fraction was increased significantly by more than 50% for small protein families (size < 7) compared to the other families (Fisher's exact test, p-value $< 10^{-4}$). This indicates that family size is a better indicator of transcript half-life similarity within protein families than functional similarity.

Indeed, we can find several examples of protein families with a highly specific function but diverging RNA half-lives. One such family is the hexokinase family for which we have previously seen that different regulatory patterns and non-redundant roles of the hexokinase I and II are supported by different RNA turnover rates. Another example is the BCL-2 protein family which contains both pro- and anti-apoptotic proteins (Youle and Strasser, 2008) for which both very short (e.g. MCL1, $t_{1/2} \sim 50$ min) and very long RNA half-lives (BAX, $t_{1/2} > 12$ h) are observed (see Table 9.3 and Figure 9.9). Long RNA half-lives are only observed for pro-apoptotic family members. As a consequence, an arrest in transcription following severe stress conditions could lead to a selective rapid decline of the short-lived transcripts of anti-apoptotic genes but not the long-lived transcripts of pro-apoptotic genes thereby promoting a pro-apoptotic state of the cell.

Interestingly, the pro-apoptotic BCL-2 family members BAX and BAK which share

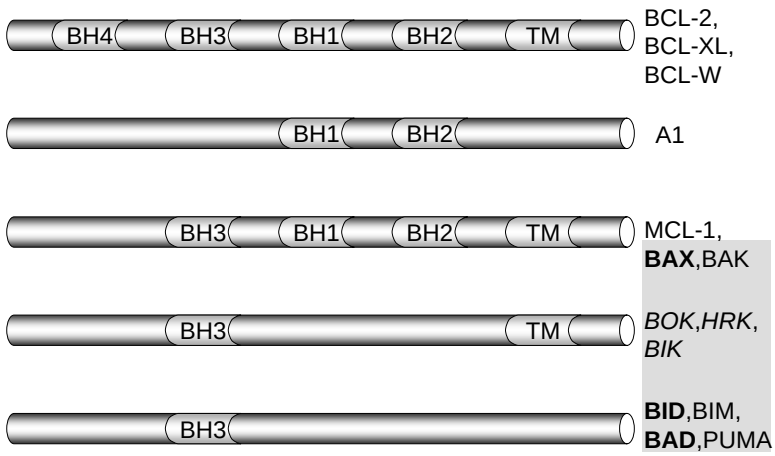


Figure 9.9: Schematic illustration of the domain structure of BCL-2 family genes. BCL-2 family proteins share between 1 and 5 conserved domains (BH1-4 and transmembrane domain (TM)) with BCL-2 (Youle and Strasser, 2008). Proteins with long transcript $t_{1/2}$ are indicated in bold and proteins with median transcript $t_{1/2}$ in italics. Pro-apoptotic genes are indicated by grey background.

a common domain structure (see Figure 9.9) and are generally assumed to substitute for each other (Wei *et al.*, 2001), exhibit tremendously different transcript half-lives ($>$ 12 h vs. 2-3 h). Increasing evidence suggests that BAX and BAK can play non-redundant roles and are regulated in different ways (Panaretakis *et al.*, 2002; Cartron *et al.*, 2003; Klee and Pimentel-Muinos, 2005; Samraj *et al.*, 2006). The observed differences in $t_{1/2}$ confirm such differences in the regulation of these proteins. Slow transcript turnover for BAX indicates that regulation is performed predominantly post-transcriptionally. Indeed, post-transcriptional regulation has been shown for BAX (Kim *et al.*, 2006). Contrary to that, short transcript half-life of BAK suggests that BAK is regulated to a large degree on the transcriptional level.

A comparison of half-life estimates in human and mouse showed that of 738 protein families with at least 2 members for which we obtained RNA half-lives in both human and mouse, 111 (15%) contain both family members with consistently fast- and slow decaying transcripts. As expected from the correlation between family size and half-life similarity, median size of these families is significantly higher than for the remaining families (22 vs. 5, Wilcoxon rank test p-value $< 10^{-16}$) and they include such large and functionally diverse families as the DEAD/DEAH box helicases and the Ankyrin repeat family. Even so, many of these families have a highly specialized function which is very similar between family members such as the hexokinase and BCL-2 families or the examples given in Table 9.4.

One of these examples is the CTP synthase. CTP synthase 1 (CTPS) has short RNA half-lives ($t_{1/2} \sim 3$ h), while turnover of CTPS2 RNA is quite slow ($t_{1/2} > 11$ h). CTP synthases are part of the pyrimidine biosynthesis pathway and catalyze the formation of

Gene(s)	$t_{1/2}$ **[h] (human)**	$t_{1/2}$ **[h] (mouse)**	**Gene(s)**	$t_{1/2}$ **[h] (human)**	$t_{1/2}$ **[h] (mouse)**	**Family**
PIAS4, PIAS1	2.6	1.3	PIAS2	9.2	7.4	protein inhibitor of activated STAT
TLE3, TLE4, TLE1	2.6	1.6	AES	13.6	18.6	groucho/TLE family
TCEB3	1.8	2.7	TCEA1	10.3	9.9	Transcription factor S-II
PARP8	2.6	3.5	PARP1	9.5	12.8	Poly(ADP-ribose) polymerase catalytic domain
NXT1, NXF1	1.4	2.3	NUTF2	14.6	8.4	Nuclear transport factor
HSPA5	2.7	2.8	HSPA4L	16.9	11.2	Hsp70 protein
CTPS	2.7	3.4	CTPS2	11.4	19.6	CTP synthase
GGA3	2.5	1.2	GGA2	7.2	14.2	ADP-ribosylation factor binding protein
BAP1	4.2	1.5	UCHL5	13.5	16.3	Ubiquitin carboxyl-terminal hydrolase
GNA13	3.7	1.9	GNAI3, GNAI2	12.4	12.8	G-protein alpha subunit
RGS2	0.6	0.9	RGS10	10.6	31.0	regulator of G protein signaling (RGS) protein

Table 9.4: Examples are shown for families with specialized functions but diverse transcript half-lives. Family members are indicated with either very short or very long transcript half-life in both murine fibroblasts and human B-cells. If more than one protein is contained in either subgroup, mean half-lives are shown, otherwise the half-life of the individual protein.

CTP. Since CTP is used in the synthesis of phospholipids and nucleic acids, CTP synthases are essential enzymes for all species (Stryer, 1995). Although regulation of protein activity by phosphorylation has been shown in human (Higgins *et al.*, 2007), little is known about transcriptional regulation of the CTP synthases. Our results make it possible to predict similar regulatory patterns as for the hexokinases. One isoform (CTPS) can be rapidly and transiently induced by transcriptional regulation while RNA concentrations of the other isoform (CTPS2) are stable and other regulatory mechanisms are likely predominant.

9.4 Discussion

Regulation of biological processes occurs at the transcriptional, translational and post-translational level. Although optimal control is only achieved by a supportive, coordinated regulation at all levels, important information on functional characteristics can already be obtained by analyzing a single level of regulation. While measurements of differential gene expression indicate which genes are regulated on the transcriptional level in a specific condition, analysis of RNA decay and turnover can provide insights on transcriptional regulation on a more general level. Here, short transcript half-lives indicate a strong transcriptional control of the corresponding genes while long half-lives suggest that post-transcriptional regulation is pre-dominant.

RNA decay has previously been studied in a wide range of species: *E. coli* (Bernstein *et al.*, 2002), yeast (Wang *et al.*, 2002), *Arabidopsis* (Gutierrez *et al.*, 2002; Narsai *et al.*, 2007) and human (Yang *et al.*, 2003). These studies indicated that transcript half-life is correlated to gene function. In particular, fast transcript decay of transcription related genes and slow decay of cellular metabolism transcripts were found to be important principles governing transcript half-life. These results were based on transcript half-lives from measurements of RNA decay after 1-3 h of transcriptional arrest. In the previous chapter, we have shown that these estimates, although highly specific for short half-lives, are unreliable for medium to long half-lives. Contrary to that, measurements of RNA de novo transcription provide more reliable and precise estimates on the whole range of RNA half-lives. Using this new approach, we determined precise RNA half-lives in human B-cells and murine fibroblasts for more than 8000 genes to investigate the conservation of RNA half-life between species and cell lines and identify specific patterns of RNA turnover.

A systematic comparison of half-lives between the two species for more than 5000 orthologous genes showed that transcript half-lives are highly conserved within species and cell types for the large majority of orthologous genes. For 67% of transcripts, half-lives did not diverge by more than a factor of two and only 18 genes showed a significant difference between species. Furthermore, variation between species was found to increase with the amount of variation observed within experimental replicates for each species. This suggests that to a large degree the observed variations are a consequence of natural variations and experimental noise within the individual experiments and do not present important differences between the species. As half-life estimates from RNA decay following transcriptional arrest are much more error-prone than half-lives from newly transcribed/total RNA ratios, a comparison based on results from decay rates would have considerably overestimated the variability between the species. Contrary to that, the analysis of RNA half-life estimates from newly transcribed RNA showed that transcript half-life is indeed highly conserved.

We performed a statistical analysis of GO functional categories to identify conserved relationships between transcript half-life and gene function. In consistency with previous results, a significant and conserved shift towards short half-lives was observed for genes involved in the regulation of transcription and signal transduction. A similar shift for genes involved in the regulation of cell cycle or apoptosis as proposed earlier (Yang *et al.*, 2003) could not be confirmed. Furthermore, although fast transcript decay allows a rapid

increase or decrease of steady-state RNA concentration due to transcriptional changes, regulatory genes as such do not exhibit a preference for fast mRNA decay unless they are involved in transcription or signal transduction. Nevertheless, as the majority of regulatory genes are involved in transcription or signal transduction, regulation of biological systems is supported in a powerful way by rapid turnover of mRNA.

Contrary to that, proteins responsible for energy metabolism and protein translation and degradation are characterized by most stable transcripts. This makes it possible to efficiently maintain stable mRNA concentrations. At the same time, regulation of these processes by transcriptional changes can only be very slow. More rapid regulation if necessary has to be mediated by post-transcriptional processes such as protein modifications or degradation. Interestingly, genes involved in translation have the highest frequency in the medium-to-long half-life range. Thus, mRNA stability is reduced compared to transcripts of energy metabolism or protein degradation genes. This could indicate that a greater degree of transcriptional control may be required for the constituents of the translational machinery.

A comprehensive analysis of more than 1000 known protein complexes in human and mouse showed that transcript half-lives within protein complexes are highly coordinated and biased towards long half-lives. Nevertheless, individual subunits can deviate significantly in transcript half-life from the remaining subunits. For instance, although the SIN3/HDAC complex and the TFIID complex are important transcriptional regulators, only a minor fraction of their subunits actually show short transcript half-lives while the majority of subunit transcripts are very stable. This may facilitate regulation of these complexes in an efficient and targeted way.

Since complex function depends on the availability of all essential components, down-regulation of complex activity can be easily accomplished by down-regulating the abundance of only one or few essential subunits. If down-regulation is only transient and lasts only for a short time, this is much faster and more efficient than first down-regulating the abundance of all proteins and then restoring abundance again to the original level for all of them. We identified a number of proteins which deviate significantly in RNA half-life from the other subunits of all complexes they are contained in. Based on these data, regulatory subunits can be predicted and an experimental analysis of these proteins may provide new insights on the regulation of large protein complexes.

Half-life within protein families was also found to be homogenous, although to a lesser degree than in protein complexes. We found that larger protein families generally are more variable in transcript half-lives than smaller ones. Interestingly, this does not appear to be correlated with the functional similarity within these protein families. Consequently, family size was found to be a much better predictor of variation in families than functional similarity.

In addition, we identified many examples in which closely related proteins with overlapping functions differ significantly in RNA half-life. These families cover such diverse and important biological processes as glycolysis, apoptosis and the pyrimidine biosynthesis pathway. Our results suggest that different regulatory patterns have been evolved by closely related family members to allow for non-redundant roles and differential and

fine-tuned responses to specific signals and stimuli. Thus, analysis of these differences in regulatory patterns based on mRNA half-life can lead to a better understanding of such non-redundant functional roles.

9.5 Conclusions

We performed a genome-scale comparison of transcript half-lives in human B-cells and murine fibroblasts to identify conserved mechanisms governing transcript half-life in mammals. Our results show that RNA half-life is strongly conserved and significantly correlated with gene function. The rapid turnover of transcripts involved in regulation of transcription and signal transduction allows for rapid and transient changes of total RNA abundance. In contrast, stable transcript concentrations are maintained by slow RNA decay for genes involved in cellular respiration, protein metabolism and protein complexes. Transcript half-lives are highly coordinated for protein complexes and families which indicates that protein subunits of the same complex and members of the same family are regulated in a similar way. Nevertheless, individual proteins can differ considerably from the remaining complex subunits and family members. These results suggest that efficient and targeted regulation of important complexes may be facilitated by rapid transcript decay and fast RNA turnover of only a few of their subunits. Furthermore, half-life differences in functionally similar protein families may have evolved to define differential regulatory patterns and non-redundant roles for these proteins. Therefore, the analysis of RNA half-life based on measurements of de novo transcription can lead to new insights into biological systems and the regulation and dynamics of protein complexes and families.

Part IV

Conclusions and outlook

Chapter 10

Conclusions and outlook

10.1 Contributions of this thesis

Advances in experimental techniques make it possible to study the system-level behavior of thousands of genes and proteins at a time. As proteins and genes can now be investigated in the context of their interaction and cooperation with other proteins and genes, biological systems can be explored at a so far unreached level of complexity and detail. Consequently, researchers in computational biology are challenged to develop new and efficient algorithms and analysis approaches to extract new insights on biological systems from the large data sets produced by high-throughput methods. In this thesis, we focused on the analysis of large-scale protein interactions maps and protein complexes purified from cell lysates and explored the dynamics and regulation of biological processes and protein complexes by investigating turnover of transcripts coding for the individual proteins.

Part I: Protein-protein interactions

The first part of this thesis focused on the analysis of physical protein-protein interactions determined in large-scale Y2H experiments. From the resulting interaction networks global properties of biological networks have been extrapolated for many species. However, as large-scale measurements are characterized by high numbers of missing or wrong interactions, it has been proposed recently that important topological properties of the network, such as the degree distribution, cannot be extrapolated from these partial networks to the complete interactomes. Contrary to that, we showed in this thesis that the topology of protein complexes can be inferred confidently despite high error rates if other network characteristics, such as the clustering coefficient, are taken into account as well.

For this purpose, we extended a basic model of measurement errors in Y2H experiments to incorporate false positive interactions in addition to false negative ones (chapter 3, Friedel and Zimmer (2006a,b)). Based on both analytical and simulation results, we showed that low coverage decreases clustering coefficients for any network while spurious interactions can increase them but only for randomly or less than randomly clustered networks. As the rate of increase and the overall clustering coefficient depend on the skewness

of the degree distribution, our results indicate that real interactomes are highly clustered and/or highly skewed. Thus, large-scale experiments can indeed provide valuable insights on the topology of interactomes even if interactions are missed or wrongly identified.

The analysis of the clustering coefficients showed that although the degree distribution is an important parameter of the network, other network characteristics can be modified without changing the number of interactions for each protein. Thus, connections in the network can be formed in different ways for the same overall degree distribution, for instance by promoting or suppressing interactions between the highly connected hubs. Suppression of interactions between hubs has previously been suggested as an important characteristic of protein-protein interaction networks as it may prevent propagation of perturbations in the network. As this observation was based on only one interaction network for yeast, we performed a systematic analysis of such correlations between degrees of interacting proteins within experimentally derived protein-protein interaction networks (chapter 4, Friedel and Zimmer (2007)). For this purpose, we developed a model to create networks with different types of degree correlations and analyzed structural and dynamic properties of the resulting networks.

Our results show that network structure and stability in networks are highly influenced by the way interactions are formed. Both suppression and promotion of interactions between hubs have favorable and unfavorable effects on network structure or stability. Furthermore, we showed that interactions between hubs are neither avoided nor preferred in biological protein interaction networks. Accordingly suppression of interactions between hubs is not a dominant feature and no bias is introduced by the high-throughput system. Although structural properties and stability of networks can be modified considerably by introducing different types of degree correlations, this does not seem to be the case in protein interaction networks.

Network analysis methods are generally focused on large networks of functionally diverse proteins. For smaller intra-pathogen or pathogen-host interaction networks new methods have to be developed. To this end, we collaborated with the Haas group at the Max von Pettenkofer-Institut of the LMU München which experimentally identified intraviral and virus-host interactions for several herpesvirus species in genome-scale Y2H experiments (chapter 5; Fossum *et al.* (2008); Dong *et al.* (2008)). The analysis of intraviral networks identified a common core of protein interactions conserved across all herpesvirus subfamilies despite an evolutionary distance of several hundred million years. By combining the interactomes, we could address the low coverage of the Y2H system and increase the number of interactions identified for each species significantly. From the combined interactome, interactions important for virus development and propagation can be identified which are not obvious from the individual interactomes.

The analysis of virus-host interactions, which we performed together with Yu-An Dong, showed that viral proteins target preferentially highly connected cellular proteins and protein complex subunits. Furthermore, interactions between the virus and host proteins, in particular the cellular hubs, are mediated more frequently by proteins conserved in all herpesviral species. As this indicates common mechanisms of virus infection, these experimental interactions may provide a valuable resource for the selection of drug targets.

Although the overlap to known viral-host interactions for herpesviruses was very low, a systematic text-mining screen showed that the identified virus targets were enriched within herpesvirus literature and, thus, have been previously linked to the herpesviral life cycle. Contrary to known targets of herpesviruses, the cellular interaction partners identified in the Y2H screens are not significantly enriched for specific biological processes or molecular functions. As this observation was made for three independent screens of herpesvirus-host interactions, this indicates a bias in the selection of potential interaction partners within small-scale experiments towards certain functional categories and is not necessarily a consequence of measurement errors in the large-scale Y2H system.

Part II: Protein complexes

In Y2H experiments only direct physical interactions are screened. These may be either transient as in signal transduction pathways or permanent or semi-permanent associations within protein complexes. Recently, experimental methods have been developed to purify protein complexes directly from the cell on a genome scale and have been applied to yeast by two separate groups. Like all high-throughput methods these methods are error-prone and, as a consequence, algorithms are necessary to predict the actual protein complexes from the raw purification data. Previous approaches generally relied more or less heavily on the availability of known protein complexes to train their algorithms which makes them only applicable to species for which protein complexes are well-documented. To address this problem, we developed an unsupervised algorithm which is based on the bootstrapping technique and independent of such additional training data (chapter 6, Friedel *et al.* (2008a)).

The confidence scores calculated by our bootstrap algorithm for individual protein interactions are more accurate than previously published scoring methods and the predicted protein complexes achieve the same quality as the best supervised approaches. An extensive comparison of complex predictions showed that the best previous prediction methods by Pu *et al.* (2007) and Hart *et al.* (2007) and our bootstrap algorithm (Friedel *et al.*, 2008a) agree to a much larger degree than the first prediction methods proposed by Gavin *et al.* (2006) and Krogan *et al.* (2006) and a common core of highly confident complexes is identified by all approaches. Our results show that highly accurate protein complexes can be determined with the bootstrap approach from purification results for any species even if its "complexome" has not been thoroughly annotated yet, i.e. no additional training data is available apart from the purifications. Furthermore, we developed the ProCope software package together with Jan Krumsiek (Krumsiek *et al.*, 2008) which provides implementations of our algorithm and previously published prediction approaches together with sophisticated evaluation methods. Using this software, researchers can quickly apply existing approaches and efficiently develop and evaluate new algorithms.

Protein complex prediction algorithms are generally focused on predicting the complexes as sets of proteins. They do not consider the internal structure of the complexes and do not distinguish between direct physical interactions and simple co-complex-membership. As most prediction algorithms calculate interaction scores as an intermediate step in com-

plex prediction, we developed an approach to identify the network of direct interactions within protein complexes from these scores (chapter 7, Friedel and Zimmer (2008)). Our algorithm is based on maximum spanning trees for each complex and predicts direct physical interactions with a superior quality compared to baseline approaches. As a consequence, the subcomponent structure of the complexes can be resolved in a straightforward way and important interactions can be identified. This is particularly important for the analysis of the predicted protein complexes. As they may contain false positives or consist of several interacting complexes, more in-depth analysis of these complexes is important to understand the actual complex structure.

Part III: RNA half-lives

When analyzing protein interactions or complexes, time-dependent dynamics are not considered as experimental methods can only identify that an interaction takes place at some point but not when and not under what conditions. To study the dynamics and regulation of biological systems, gene expression profiling methods are generally used as fast and reliable measurements of mRNA concentrations can be obtained with microarray technology. As mRNA is translated to proteins, mRNA abundance is assumed to correlate with protein abundance and changes in mRNA concentrations result in changes in protein concentration. Although recent advances in quantification of protein concentrations by mass spectrometry or protein microarrays (Ong and Mann, 2006; Bantscheff *et al.*, 2007; Hober and Uhlén, 2008; Colzani *et al.*, 2008) now also allow large-scale protein quantification, these methods are not as advanced yet as standard microarray technology.

Due to a constant turnover of mRNA in the cell, mRNA abundance is determined by a balance of RNA decay and transcription. The rate of RNA turnover determines how long RNA persists in the cell and how rapidly its abundance can be up- or down-regulated. Our collaborators from the Max von Pettenkofer-Institut, Lars Dölken *et al.*, recently developed an experimental method for measuring RNA de novo transcription and decay separately from total RNA abundance. This makes it possible to determine rates of RNA decay and transcription for thousands of genes at a time with standard microarray technology.

To support the analysis of this data, we developed methods for normalizing results from different RNA fractions in an intuitive way, for applying a quality control on probe sets and for calculating RNA half-lives (see chapter 8, Dölken *et al.* (2008)). First, our results show that highly accurate half-life estimates can be obtained from measurements of de novo transcription over the whole range of RNA half-lives. Contrary to that, estimates from decay are only accurate for short half-lives unless very long labeling times are used. Second, our probe set quality score allows to select the most reliable probe set for genes represented several times on the microarray and to improve accuracy of results compared to standard averaging approaches. Third, based on half-life estimates, changes in total and newly transcribed RNA can be correlated to each other and transcriptional regulation after a stimulus such as interferon treatment can be identified more reliably. In this way, both rapid and immediate as well as long-range effects can be identified in the same experiment by analyzing de novo transcription even if no effect is visible in total RNA abundance yet.

To better understand the mechanisms governing RNA decay and the functional role of the large variety of RNA decay rates, we analyzed RNA half-life in two mammalian species and different cell types in collaboration with Lars Dölken. In this way, we identified common patterns which are conserved across species and cell types in murine fibroblasts and human B-cells and found that RNA half-lives are highly conserved and significantly correlated with gene function. Our results showed that regulation of transcription and signal transduction is specifically supported by short half-lives. Contrary to that, genes involved in energy and protein metabolism are characterized by long transcript half-lives. In this way, stable mRNA concentrations are maintained for these central metabolic processes.

Furthermore, by analyzing transcript half-life, regulation of protein complexes could be investigated. We found that RNA half-lives were highly coordinated in protein complexes which indicates similar regulation patterns for complex subunits. A slightly lower but still significant half-life similarity was found in protein families. Despite this fact, individual proteins can differ significantly in half-life from the other proteins in the complex or family. Our results indicate that regulation of some protein complexes may be accomplished fast and efficiently by transcriptional regulation of only a few regulatory subunits. In protein families, diverse transcript half-lives may have evolved to create differential regulatory patterns and non-redundant functional roles of otherwise similar proteins or enzymes. Based on our observations, we could make several interesting predictions concerning the regulation of important protein complexes and biological processes such as transcription, apoptosis and the pyrimidine biosynthesis pathway.

10.2 Perspectives for future research

In this thesis, we described novel and sophisticated methods for studying cooperative associations of proteins and their regulation using high-throughput data. In the future, these methods will have to be applied to new experimental results and adapted and extended to deal with rapid advances in high-throughput techniques. Here, new developments are taking place in the identification of protein interactions and complexes and mRNA and protein quantification.

The yeast two-hybrid system is currently applied by the group of Jürgen Haas to systematically identify intra-pathogen and pathogen-host protein-protein interactions for a wider range of human pathogens. To analyze these interaction networks, methods developed for the analysis of herpesviral interactions will have to be compiled into a standard analysis workflow. As interactome screens are not performed for several related pathogen species for these pathogens, new methods will have to be developed to address the low coverage of the Y2H system and to identify important interactions using only the information from one species.

As yeast two-hybrid methods are inherently error-prone, significant efforts are currently being made to develop more reliable identification methods. Promising approaches include the protein-fragment complementation assay (PCA) (Michnick *et al.*, 2007) and the bioluminescence resonance energy transfer (BRET) system (Xu *et al.*, 1999). In the PCA

system, complementary fragments of a reporter protein are fused to two proteins which are tested for an interaction. If they interact, the fragments of the reporter proteins are brought in close proximity and fold back into the native and active structure of the reporter protein. This method makes it possible to study protein interactions in vivo in their native states and cellular locations and was recently applied on a genome scale to identify an in vivo interaction map in yeast (Tarassov *et al.*, 2008). In the BRET system, one candidate protein is fused to a bioluminescent luciferase and the other one to a fluorescent protein. If they interact, luciferase and flurescent protein are brought together close enough so that the color of the bioluminescent emission is changed by resonance energy transfer. In collaboration with the group of Prof. Ania Muntau (Dr. von Haunersches Kinderspital, LMU München) which is refining this method towards a high-throughput application, we are currently developing new methods for the analysis and evaluation of the large-scale BRET results.

While affinity purification methods have been successfully applied to yeast, large-scale applications in higher eukaryotes lag behind due to the higher complexity of these organisms. So far, no other genome-scale data sets have become available since the publication of the yeast studies. Nevertheless, recent developments in methological protocols (Bürckstümmer *et al.*, 2006; Gregan *et al.*, 2007; Gloeckner *et al.*, 2007) now make it likely that in the near future genome-scale screens of higher eukaryotes will be performed. As the unsupervised bootstrap algorithm we developed does not require known protein complexes for training, it can be immediately applied to such new studies. Here, the ProCope package can be used by researchers to quickly predict and evaluate protein complexes from these screens.

When we evaluated direct interactions predicted from purification assays on the RNA polymerase complexes, shortcomings of the TAP system in distinguishing between direct physical interactions and indirect interactions via proteins in the same complex became apparent. Here, it should be evaluated whether more complex models taking into account these limitations can increase prediction accuracy further. Possible extensions would be to explicitly model the probability of direct and indirect interactions in a second step after predicting the protein complexes. Furthermore, negative and positive information which excludes or enforces specific interactions may be included.

The new experimental approach to simultaneously measure RNA decay, transcription and abundance is currently applied by several groups, for instance to analyze the effect of virus miRNAs on transcript stability and abundance. To support these large-scale applications, we are planning to make the algorithms presented in this thesis available to the research community in the form of an efficient and easy-to-use software platform which can be integrated with other analysis tools. Furthermore, new approaches will have to be developed to identify transcripts with significant changes in transcript half-life between different conditions to compare for instance miRNA expressing cells to normal cells.

Finally, methods have to be adapted to new technological advances. The new generation of Affymetrix gene arrays now contain probe sets from the full length of each gene and not only from the 3' end. Each gene is represented by 26 probes and currently measurements from all probes are combined into one expression value for the whole gene. Using our probe

set quality score, high quality probes might be selected instead of using all 26 probes. On exon arrays, reliability of probes for individual exons but also the exons themselves can be assessed in the same way which can lead to new applications of this method for analysis of alternative splicing. Furthermore, advanced normalization methods might be developed based on the linear regression approach. In addition, next-generation sequencing technology (Genome sequencer by 454/Roche, Solexa sequencer by Illumina, the SOLiD system by Applied Biosystems) now makes it possible to quantify RNA abundance by rapid sequencing of transcripts isolated from cells (Wang *et al.*, 2008c). A comparison of measurements of RNA transcription and half-life obtained with different generations of microarrays or transcript sequencing could lead to a better understanding of the advantages and weaknesses of the different technologies.

Bibliography

Adams, M. D. *et al.* (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**(5013), 1651–1656.

Adams, M. D. *et al.* (1992). Sequence identification of 2,375 human brain genes. *Nature*, **355**(6361), 632–634.

Adams, M. D. *et al.* (2000). The genome sequence of Drosophila melanogaster. *Science*, **287**(5461), 2185–2195.

Affymetrix Inc. (2001). Statistical algorithms reference guide (`http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf`). Technical report.

Ahmed, K., Gerber, D. A., and Cochet, C. (2002). Joining the cell survival squad: an emerging role for protein kinase CK2. *Trends Cell Biol*, **12**(5), 226–230.

al Kobaisi, M. F., Rixon, F. J., McDougall, I., and Preston, V. G. (1991). The herpes simplex virus UL33 gene product is required for the assembly of full capsids. *Virology*, **180**(1), 380–388.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47.

Albert, R., Jeong, H., and Barabási, A. (2000). Error and attack tolerance of complex networks. *Nature*, **406**(6794), 378–82.

Albright, S. R. and Tjian, R. (2000). TAFs revisited: more data reveal new twists and confirm old ideas. *Gene*, **242**(1-2), 1–13.

Aloy, P. *et al.* (2004). Structure-based assembly of protein complexes in yeast. *Science*, **303**(5666), 2026–2029.

Aloy, P. and Russell, R. B. (2002). Potential artefacts in protein-interaction networks. *FEBS Lett*, **530**(1-3), 253–254.

Andersson, A. F. *et al.* (2006). Global analysis of mRNA stability in the archaeon sulfolobus. *Genome Biol*, **7**(10), R99.

Arima, M. *et al.* (2002). A putative silencer element in the IL-5 gene recognized by Bcl6. *J Immunol*, **169**(2), 829–836.

Ashburner, M. *et al.* (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1), 25–29.

Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, **20**, 991–7.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, **22**, 78–85.

Bairoch, A. and Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*, **19 Suppl**, 2247–2249.

Bamforth, S. D. *et al.* (2001). Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. *Nat Genet*, **29**(4), 469–474.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*, **389**(4), 1017–1031.

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–12.

Bartel, P. L., Roecklein, J. A., SenGupta, D., and Fields, S. (1996). A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet*, **12**(1), 72–77.

Batada, N. N. *et al.* (2006). Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol*, **4**(10), e317.

Bauer, S., Grossmann, S., Vingron, M., and Robinson, P. N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**(14), 1650–1651.

Beard, P. M., Taus, N. S., and Baines, J. D. (2002). DNA cleavage and packaging proteins encoded by genes U(L)28, U(L)15, and U(L)33 of herpes simplex virus type 1 form a complex in infected cells. *J Virol*, **76**(10), 4785–4791.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Bentley, D. R. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Benton, D. (1990). Recent changes in the GenBank on-line service. *Nucleic Acids Res*, **18**(6), 1517–1520.

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**(1), 51.

Berglund, A.-C., Sjölund, E., Ostlund, G., and Sonnhammer, E. L. L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, **36**(Database issue), D263–D266.

Bernard, A., Vaughn, D. S., and Hartemink, A. J. (2007). Reconstructing the topology of protein complexes. In *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2007, Oakland, CA, USA, April 21-25*, pages 32–46.

Bernstein, F. C. *et al.* (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, **112**(3), 535–542.

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N. (2002). Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci USA*, **99**(15), 9697–9702.

Bertone, P. *et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**(5705), 2242–2246.

Bhan, A., Galas, D. J., and Dewey, T. G. (2002). A duplication growth model of gene expression networks. *Bioinformatics*, **18**(11), 1486–93.

Bhattacharyya, S. N., Habermacher, R., Martine, U., Closs, E. I., and Filipowicz, W. (2006). Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, **125**(6), 1111–1124.

Blattner, C. *et al.* (2000). UV-Induced stabilization of c-fos and other short-lived mRNAs. *Mol Cell Biol*, **20**(10), 3616–3625.

Blattner, F. R. *et al.* (1997). The complete genome sequence of Escherichia coli K-12. *Science*, **277**(5331), 1453–1474.

Bollobás, B. (2001). *Random Graphs*. Cambridge University Press.

Boyer-Guittaut, M. *et al.* (2005). SUMO-1 modification of human transcription factor (TF) IID complex subunits: inhibition of TFIID promoter-binding activity through SUMO-1 modification of hsTAF5. *J Biol Chem*, **280**(11), 9937–9945.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J Math Sociol*, **25**, 163–177.

Breitkreutz, B.-J. *et al.* (2008). The BioGRID interaction database: 2008 update. *Nucleic Acids Res*, **36**(Database issue), D637–D640.

Brennan, C. M. and Steitz, J. A. (2001). HuR and mRNA stability. *Cell Mol Life Sci*, **58**(2), 266–277.

Brohee, S. and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**(1), 488.

Bürckstümmer, T. *et al.* (2006). An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat Methods*, **3**(12), 1013–1019.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**(5396), 2012–2018.

Calderwood, M. A. *et al.* (2007). Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA*, **104**(18), 7606–7611.

Cartron, P.-F. *et al.* (2003). Nonredundant role of Bax and Bak in Bid-mediated apoptosis. *Mol Cell Biol*, **23**(13), 4701–4712.

Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proc Natl Acad Sci USA*, **99**, 15879–82.

Chung, F., Lu, L., Dewey, T. G., and Galas, D. J. (2003). Duplication models for biological networks. *J Comput Biol*, **10**(5), 677–87.

Cohen, J. I. (2000). Epstein-Barr virus infection. *N Engl J Med*, **343**(7), 481–492.

Collins, S. R. *et al.* (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics*, **6**(3), 439–450.

Colzani, M., Schütz, F., Potts, A., Waridel, P., and Quadroni, M. (2008). Relative protein quantification by isobaric SILAC with immonium ion splitting (ISIS). *Mol Cell Proteomics*, **7**(5), 927–937.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2000). *Introduction to Algorithms, 2nd edition*. MIT Press, McGraw-Hill Book Company.

Coulomb, S., Bauer, M., Bernard, D., and Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci*, **272**, 1721–5.

Cramer, P. *et al.* (2008). Structure of eukaryotic RNA polymerases. *Annual Review of Biophysics*, **37**(1), 337–352.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563.

Davison, A. (2004). Compendium of human herpesvirus gene names. In *29th International Herpesvirus Workshop, Reno, USA.*

Davison, A. J. and Scott, J. E. (1986). The complete DNA sequence of varicella-zoster virus. *J Gen Virol*, **67 (Pt 9)**, 1759–1816.

Deane, C. M., Salwínski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, **1**, 349–56.

Dolan, A. *et al.* (2004). Genetic content of wild-type human cytomegalovirus. *J Gen Virol*, **85**(Pt 5), 1301–1312.

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, **14**(9), 1959–1972.

Dong, Y.-A., Koegl, M., Stelzl, U., Baiker, A., Friedel, C. C., Rose, D., Lutter, P., von Brunn, A., Schwarz, F., Kasmapour, B., Wanker, E., Koszinowski, U., Zimmer, R., Uetz, P., Fossum, E., and Haas, J. (2008). Herpesviral proteins preferentially target highly connected human host proteins. *Manuscript in preparation.*

Dorogovtsev, S. and Mendes, J. (2002). Evolution of networks. *Adv. Phys*, **51**, 1079–1187.

Dorogovtsev, S. N. and Mendes, J. F. (2001). Effect of the accelerating growth of communications networks on their structure. *Phys Rev E Stat Nonlin Soft Matter Phys*, **63**(2 Pt 2), 025101.

Dorogovtsev, S. N., Mendes, J. F., and Samukhin, A. N. (2000). Structure of growing networks with preferential linking. *Phys Rev Lett*, **85**(21), 4633–4636.

Dunn, W. *et al.* (2003). Functional profiling of a human cytomegalovirus genome. *Proc Natl Acad Sci USA*, **100**(24), 14223–14228.

Dwight, S. S. *et al.* (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Res*, **30**(1), 69–72.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog*, **4**(2), e32.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap.* Chapman & Hall, London.

Eisenberg, E. and Levanon, E. Y. (2003). Preferential attachment in the protein network evolution. *Phys Rev Lett*, **91**, 138701.

Eppig, J. T. *et al.* (2007). The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res*, **35**(Database issue), D630–D637.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8), 861–874.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein, J. (1989). PHYLIP - phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.

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

Finn, R. D. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res*, **36**(Database issue), D281–D288.

Finn, R. D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**(3), 410–412.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, **98**, 39–54.

Flajolet, M. *et al.* (2000). A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene*, **242**(1-2), 369–379.

Fleischmann, R. D. *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**(5223), 496–512.

Flores, A. *et al.* (1999). A protein-protein interaction map of yeast RNA polymerase III. *Proc Natl Acad Sci USA*, **96**(14), 7815–7820.

Fossum, E., Friedel, C. C., Rajagopala, S. V., Titz, B., Baiker, A., Schmidt, T., Kraus, T., Suthram, S., Bandyopadhyay, S., Rose, D., Uhlmann, M., Zeretzke, C., Dong, Y.-A., Boulet, H., Bailer, S. M., Koszinowski, U., Ideker, T., Uetz, P., Zimmer, R., and Haas, J. (2008). Comparative interactomics reveal evolutionary conserved herpesviral protein-protein interaction networks. *Submitted to PloS Pathogens*.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**(1), 35–41.

Friedel, C. C. and Zimmer, R. (2006a). Inferring topology from clustering coefficients in protein-protein interaction networks. *BMC Bioinformatics*, **7**, 519.

Friedel, C. C. and Zimmer, R. (2006b). Toward the complete interactome. *Nat Biotechnol*, **24**(6), 614–615.

Friedel, C. C. and Zimmer, R. (2007). Influence of degree correlations on network structure and stability in protein-protein interaction networks. *BMC Bioinformatics*, **8**, 297.

Friedel, C. C. and Zimmer, R. (2008). Identifying the topology of protein complexes from affinity purification assays. In *Proceedings of the German Conference on Bioinformatics, GCB 2008, Dresden, Germany, September 9-12*, pages 30–43.

Friedel, C. C., Krumsiek, J., and Zimmer, R. (2008a). Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. In *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2008, Singapore, March 30 - April 2*, pages 3–16.

Friedel, C. C., Dölken, L., Ruzsics, Z., Koszinowski, U., and Zimmer, R. (2008b). A conserved role of RNA half-life in mice and men. *Manuscript in preparation*.

Friedländer, M. R. *et al.* (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, **26**(4), 407–415.

Frishman, D. *et al.* (2001). Functional and structural genomics using PEDANT. *Bioinformatics*, **17**(1), 44–57.

Fromont-Racine, M., Rain, J. C., and Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet*, **16**(3), 277–282.

Fuchs, W., Klupp, B. G., Granzow, H., Osterrieder, N., and Mettenleiter, T. C. (2002). The interacting UL31 and UL34 gene products of pseudorabies virus are involved in egress from the host-cell nucleus and represent components of primary enveloped but not mature virions. *J Virol*, **76**(1), 364–378.

Fuller-Pace, F. V. (2006). DExD/H box RNA helicases: multifunctional proteins with important roles in transcriptional regulation. *Nucleic Acids Res*, **34**(15), 4206–4215.

Fundel, K. and Zimmer, R. (2006). Gene and protein nomenclature in public databases. *BMC Bioinformatics*, **7**, 372.

Gandhi, T. K. B. *et al.* (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, **38**(3), 285–293.

Ganem, D. (2006). KSHV infection and the pathogenesis of Kaposi's sarcoma. *Annu Rev Pathol*, **1**, 273–296.

Gardina, P. J. *et al.* (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.

Gavin, A.-C. *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–7.

Gavin, A.-C. *et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–6.

Gentleman, R. C. *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10), R80.

Gilden, D. H., Kleinschmidt-DeMasters, B. K., LaGuardia, J. J., Mahalingam, R., and Cohrs, R. J. (2000). Neurologic complications of the reactivation of varicella-zoster virus. *N Engl J Med*, **342**(9), 635–645.

Giot, L. *et al.* (2003). A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–36.

Gloeckner, C. J., Boldt, K., Schumacher, A., Roepman, R., and Ueffing, M. (2007). A novel tandem affinity purification strategy for the efficient isolation and characterisation of native protein complexes. *Proteomics*, **7**(23), 4228–4234.

Goffeau, A. *et al.* (1996). Life with 6000 genes. *Science*, **274**, 546, 563–7.

Gorospe, M., Wang, X., and Holbrook, N. J. (1998). p53-dependent elevation of p21Waf1 expression by UV light is mediated through mRNA stabilization and involves a vanadate-sensitive regulatory system. *Mol Cell Biol*, **18**(3), 1400–1407.

Grant, P. A. *et al.* (1998). A subset of TAF(II)s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. *Cell*, **94**(1), 45–53.

Gregan, J. *et al.* (2007). Tandem affinity purification of functional TAP-tagged proteins from human cells. *Nat Protoc*, **2**(5), 1145–1151.

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M. D., and Hughes, T. R. (2004). Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol*, **24**(12), 5534–5547.

Guo, D., Rajamki, M. L., Saarma, M., and Valkonen, J. P. (2001). Towards a protein interaction map of potyviruses: protein interaction matrixes of two potyviruses based on the yeast two-hybrid system. *J Gen Virol*, **82**(Pt 4), 935–939.

Gutierrez, R. A., Ewing, R. M., Cherry, J. M., and Green, P. J. (2002). Identification of unstable transcripts in Arabidopsis by cDNA microarray analysis: rapid decay is associated with a group of touch- and specific clock-controlled genes. *Proc Natl Acad Sci USA*, **99**(17), 11513–11518.

Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, **23**, 839–44.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., and Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, **6 Suppl 1**, S14.

Hart, G. T., Lee, I., and Marcotte, E. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, **8**, 236.

He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet*, **2**(6), e88.

Herbert, A. M., Bagg, J., Walker, D. M., Davies, K. J., and Westmoreland, D. (1995). Seroepidemiology of herpes virus infections among dental personnel. *J Dent*, **23**(6), 339–342.

Higgins, M. J., Graves, P. R., and Graves, L. M. (2007). Regulation of human cytidine triphosphate synthetase 1 by glycogen synthase kinase 3. *J Biol Chem*, **282**(40), 29493–29503.

Ho, Y. *et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–3.

Hober, S. and Uhlén, M. (2008). Human protein atlas and the use of microarray technologies. *Curr Opin Biotechnol*, **19**(1), 30–35.

Hollunder, J., Beyer, A., and Wilhelm, T. (2005). Identification and characterization of protein subcomplexes in yeast. *Proteomics*, **5**(8), 2082–2089.

Hollunder, J., Friedel, M., Beyer, A., Workman, C. T., and Wilhelm, T. (2007a). DASS: efficient discovery and p-value calculation of substructures in unordered data. *Bioinformatics*, **23**(1), 77–83.

Hollunder, J., Beyer, A., and Wilhelm, T. (2007b). Protein subcomplexes–molecular machines with highly specialized functions. *IEEE Trans Nanobioscience*, **6**(1), 86–93.

Huang, H., Jedynak, B. M., and Bader, J. S. (2007). Where have all the interactions gone? estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol*, **3**(11), e214.

Hubbard, T. *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res*, **30**(1), 38–41.

Hudnall, S. and Stanberry, L. (2006). *Tropical infectious diseases: principles, pathogens, and practice*, chapter Human herpesvirus infections, pages 590–620. Elsevier Churchill, Livingstone, Philadelphia (PA).

Hudnall, S. D., Chen, T., Allison, P., Tyring, S. K., and Heath, A. (2008). Herpesvirus prevalence and viral load in healthy blood donors by quantitative real-time polymerase chain reaction. *Transfusion*, **48**(6), 1180–1187.

Huh, W.-K. *et al.* (2003). Global analysis of protein localization in budding yeast. *Nature*, **425**(6959), 686–691.

Ispolatov, I., Krapivsky, P. L., and Yuryev, A. (2005). Duplication-divergence model of protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys*, **71**, 061911.

Ito, T. *et al.* (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, **98**, 4569–74.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**, 41–2.

Jones, J. P. and Dohm, G. L. (1997). Regulation of glucose transporter GLUT-4 and hexokinase II gene transcription by insulin and epinephrine. *Am J Physiol*, **273**, E682–E687.

Kamine, J., Elangovan, B., Subramanian, T., Coleman, D., and Chinnadurai, G. (1996). Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator. *Virology*, **216**(2), 357–366.

Kattenhorn, L. M. *et al.* (2004). Identification of proteins associated with murine cytomegalovirus virions. *J Virol*, **78**(20), 11187–11197.

Kenzelmann, M. *et al.* (2007). Microarray analysis of newly synthesized RNA in cells and animals. *Proc Natl Acad Sci USA*, **104**(15), 6164–6169.

Kiernan, R. *et al.* (2003). Post-activation turn-off of NF-kappa B-dependent transcription is regulated by acetylation of p65. *J Biol Chem*, **278**(4), 2758–2766.

Kim, B.-J., Ryu, S.-W., and Song, B.-J. (2006). JNK- and p38 kinase-mediated phosphorylation of Bax leads to its activation and mitochondrial translocation and to apoptosis of human hepatoma HepG2 cells. *J Biol Chem*, **281**(30), 21256–21265.

Kim, J., Krapivsky, P. L., Kahng, B., and Redner, S. (2002). Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys*, **66**, 055101.

Kiss-Toth, E. *et al.* (2004). Human tribbles, a protein family controlling mitogen-activated protein kinase cascades. *J Biol Chem*, **279**(41), 42703–42708.

Klee, M. and Pimentel-Muinos, F. X. (2005). Bcl-X(L) specifically activates Bak to induce swelling and restructuring of the endoplasmic reticulum. *J Cell Biol*, **168**(5), 723–734.

Krapivsky, P. L. and Redner, S. (2001). Organization of growing random networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, **63**(6 Pt 2), 066123.

Krapivsky, P. L., Redner, S., and Leyvraz, F. (2000). Connectivity of growing random networks. *Phys Rev Lett*, **85**(21), 4629–4632.

Krogan, N. J. *et al.* (2002). RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted proteomics approach. *Mol Cell Biol*, **22**(20), 6979–6992.

Krogan, N. J. *et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–43.

Krumsiek, J., Friedel, C. C., and Zimmer, R. (2008). ProCope–protein complex prediction and evaluation. *Bioinformatics*, **24**(18), 2115–2116.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, **7**, 48–50.

LaCount, D. J. *et al.* (2005). A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, **438**, 103–7.

Lake, C. M. and Hutt-Fletcher, L. M. (2004). The Epstein-Barr virus BFRF1 and BFLF2 proteins interact and coexpression alters their cellular localization. *Virology*, **320**(1), 99–106.

Lam, L. T. *et al.* (2001). Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol*, **2**(10), RESEARCH0041.

Lander, E. S. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Latora, V. and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.*, **87**(19), 198701.

Lee, J. H., Vittone, V., Diefenbach, E., Cunningham, A. L., and Diefenbach, R. J. (2008). Identification of structural protein-protein interactions of herpes simplex virus type 1. *Virology*, **378**(2), 347–354.

Lee, T. I. *et al.* (2000). Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature*, **405**(6787), 701–704.

Legrain, P. and Selig, L. (2000). Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett*, **480**, 32–6.

Leurent, C. *et al.* (2004). Mapping key functional sites within yeast TFIID. *EMBO J*, **23**(4), 719–727.

Levy, S. *et al.* (2007). The diploid genome sequence of an individual human. *PLoS Biol*, **5**(10), e254.

Li, S. *et al.* (2004). A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–3.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Lin, N. and Zhao, H. (2005). Are scale-free networks robust to measurement errors? *BMC Bioinformatics*, **6**, 119.

Liuwantara, D. *et al.* (2006). Nuclear factor-kappab regulates beta-cell death: a critical role for A20 in beta-cell protection. *Diabetes*, **55**(9), 2491–2501.

Lockhart, D. J. *et al.* (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, **14**(13), 1675–1680.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**(10), 1275–1283.

Loregian, A. and Palú, G. (2005). Disruption of protein-protein interactions: towards new targets for chemotherapy. *J Cell Physiol*, **204**(3), 750–762.

Mahalingam, S. *et al.* (1998). HIV-1 Vpr interacts with a human 34-kDa mov34 homologue, a cellular factor linked to the G2/M phase transition of the mammalian cell cycle. *Proc Natl Acad Sci USA*, **95**(7), 3419–3424.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat Methods*, **4**(8), 613–614.

Maslov, S. and Sneppen, K. (2002a). Protein interaction networks beyond artifacts. *FEBS Lett*, **530**(1-3), 255–256.

Maslov, S. and Sneppen, K. (2002b). Specificity and stability in topology of protein networks. *Science*, **296**, 910–3.

Mathupala, S. P., Rempel, A., and Pedersen, P. L. (1995). Glucose catabolism in cancer cells. Isolation, sequence, and activity of the promoter for type II hexokinase. *J Biol Chem*, **270**(28), 16918–16925.

McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci USA*, **97**(9), 4879–4884.

McGeoch, D. J., Rixon, F. J., and Davison, A. J. (2006). Topics in herpesvirus genomics and evolution. *Virus Res*, **117**(1), 90–104.

Melvin, W. T., Milne, H. B., Slater, A. A., Allen, H. J., and Keir, H. M. (1978). Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur J Biochem*, **92**(2), 373–379.

Mewes, H. W. *et al.* (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, **32**(Database issue), D41–D44.

Mewes, H. W., Albermann, K., Heumann, K., Liebl, S., and Pfeiffer, F. (1997). MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res*, **25**(1), 28–30.

Michnick, S. W., Ear, P. H., Manderson, E. N., Remy, I., and Stefan, E. (2007). Universal strategies in research and drug discovery based on protein-fragment complementation assays. *Nat Rev Drug Discov*, **6**(7), 569–582.

Middendorf, M., Ziv, E., and Wiggins, C. H. (2005). Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *Proc Natl Acad Sci USA*, **102**(9), 3192–7.

Mourelatos, Z. *et al.* (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev*, **16**(6), 720–728.

Muranyi, W., Haas, J., Wagner, M., Krohne, G., and Koszinowski, U. H. (2002). Cytomegalovirus recruitment of cellular kinases to dissolve the nuclear lamina. *Science*, **297**(5582), 854–857.

Murthag, F. (1984). Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly*, **1**, 101–113.

Narsai, R. *et al.* (2007). Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *Plant Cell*, **19**(11), 3418–3436.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, **45**, 167– 256.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, **46**, 323.

Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, **64**(2), 026118.

Ogryzko, V. V. *et al.* (1998). Histone-like TAFs within the PCAF histone acetylase complex. *Cell*, **94**(1), 35–44.

Ong, S.-E. and Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc*, **1**(6), 2650–2660.

Osawa, H., Printz, R. L., Whitesell, R. R., and Granner, D. K. (1995). Regulation of hexokinase II gene transcription and glucose phosphorylation by catecholamines, cyclic AMP, and insulin. *Diabetes*, **44**(12), 1426–1432.

Panaretakis, T., Pokrovskaja, K., Shoshan, M. C., and Grandér, D. (2002). Activation of Bak, Bax, and BH3-only proteins in the apoptotic response to doxorubicin. *J Biol Chem*, **277**(46), 44317–44326.

Park, R. and Baines, J. D. (2006). Herpes simplex virus type 1 infection induces activation and recruitment of protein kinase C to the nuclear membrane and increased phosphorylation of lamin B. *J Virol*, **80**(1), 494–504.

Pellizzoni, L., Baccon, J., Charroux, B., and Dreyfuss, G. (2001). The survival of motor neurons (SMN) protein interacts with the snoRNP proteins fibrillarin and GAR1. *Curr Biol*, **11**(14), 1079–1088.

Peri, S. *et al.* (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, **32**(Database issue), D497–D501.

Pijnappel, W. W. *et al.* (2001). The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes Dev*, **15**(22), 2991–3004.

Price, D. J. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.*, **27**, 292–306.

Prim, R. C. (1957). Shortest connection networks and some generalisations. *Bell System Technical Journal*, **36**, 1389–1401.

Printz, R. L. *et al.* (1993). Hexokinase II mRNA and gene structure, regulation by insulin, and evolution. *J Biol Chem*, **268**(7), 5209–5219.

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–15.

Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., and Wodak, S. J. (2007). Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics*, **7**(6), 944–960.

Qian, J., Luscombe, N. M., and Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol*, **313**(4), 673–81.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raghavan, A. *et al.* (2002). Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res*, **30**(24), 5529–5538.

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.

Reynolds, A. E., Ryckman, B. J., Baines, J. D., Zhou, Y., Liang, L., and Roller, R. J. (2001). U(L)31 and U(L)34 proteins of herpes simplex virus type 1 form a complex that accumulates at the nuclear rim and is required for envelopment of nucleocapsids. *J Virol*, **75**(18), 8803–8817.

Riddle, S. R. *et al.* (2000). Hypoxia induces hexokinase II gene expression in human lung cell line A549. *Am J Physiol Lung Cell Mol Physiol*, **278**(2), L407–L416.

Rigaut, G. *et al.* (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, **17**(10), 1030–1032.

Robertson, G. *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, **4**(8), 651–657.

Robinson, M. D. and Speed, T. P. (2007). A comparison of Affymetrix gene expression arrays. *BMC Bioinformatics*, **8**, 449.

Ross, J. (1995). mRNA stability in mammalian cells. *Microbiol Rev*, **59**(3), 423–450.

Rual, J.-F. *et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–8.

Ruepp, A. *et al.* (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, **36**, D646 – D650.

Rungsarityotin, W., Krause, R., Schodl, A., and Schliep, A. (2007). Identifying protein complexes directly from high-throughput TAP data with markov random fields. *BMC Bioinformatics*, **8**(1), 482.

Samraj, A. K., Stroh, C., Fischer, U., and Schulze-Osthoff, K. (2006). The tyrosine kinase Lck is a positive regulator of the mitochondrial apoptosis pathway by controlling Bak expression. *Oncogene*, **25**(2), 186–197.

Sanda, C. *et al.* (2006). Differential gene induction by type I and type II interferons and their combination. *J Interferon Cytokine Res*, **26**(7), 462–472.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235), 467–470.

Scherbik, S. V., Stockman, B. M., and Brinton, M. A. (2007). Differential expression of interferon (IFN) regulatory factors and IFN-stimulated genes at early times after West Nile virus infection of mouse embryo fibroblasts. *J Virol*, **81**(21), 12005–12018.

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.

Scholtens, D., Vidal, M., and Gentleman, R. (2005). Local modeling of global interactome networks. *Bioinformatics*, **21**(17), 3548–3557.

Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M., and Rosenow, C. (2003). Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res*, **13**(2), 216–223.

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, **6**(7), 639–645.

Shannon, P. *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**(11), 2498–2504.

Sharan, R. *et al.* (2005). Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA*, **102**(6), 1974–9.

Shin, E.-C. *et al.* (2007). Proteasome activator and antigen-processing aminopeptidases are regulated by virus-induced type I interferon in the hepatitis C virus-infected liver. *J Interferon Cytokine Res*, **27**(12), 985–990.

Shyu, A. B., Greenberg, M. E., and Belasco, J. G. (1989). The c-fos transcript is targeted for rapid decay by two distinct mRNA degradation pathways. *Genes Dev*, **3**(1), 60–72.

Silverstein, R. A. and Ekwall, K. (2005). Sin3: a flexible regulator of global gene expression and genome stability. *Curr Genet*, **47**(1), 1–17.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, **42**, 425–440.

Smith, J. S. *et al.* (2002). Herpes simplex virus-2 as a human papillomavirus cofactor in the etiology of invasive cervical cancer. *J Natl Cancer Inst*, **94**(21), 1604–1613.

Soengas, J. L., Polakof, S., Chen, X., Sangiao-Alvarellos, S., and Moon, T. W. (2006). Glucokinase and hexokinase expression and activities in rainbow trout tissues: changes with food deprivation and refeeding. *Am J Physiol Regul Integr Comp Physiol*, **291**(3), R810–R821.

Solé, R. V., Pastor-Satorras, R., Smith, E., and Kepler, T. B. (2002). A model of large-scale proteome evolution. *Adv. Complex Syst*, **5**, 43–54.

Song, M. J., Hwang, S., Wong, W. H., Wu, T.-T., Lee, S., Liao, H.-I., and Sun, R. (2005). Identification of viral genes essential for replication of murine gamma-herpesvirus 68 using signature-tagged mutagenesis. *Proc Natl Acad Sci USA*, **102**(10), 3805–3810.

Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, **33**(Database issue), D413–D417.

Stelzl, U. *et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–68.

Stryer, L. (1995). *Biochemistry (4th edition)*. W.H. Freeman & Company.

Stumpf, M. P. H., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA*, **102**, 4221–4.

Tanaka, R., Yi, T.-M., and Doyle, J. (2005). Some protein interaction data do not exhibit power law statistics. *FEBS Lett*, **579**, 5140–4.

Tarassov, K. *et al.* (2008). An in vivo map of the yeast protein interactome. *Science*, **320**(5882), 1465–1470.

Uetz, P. *et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–7.

Uetz, P. *et al.* (2006). Herpesviral protein networks and their interaction with the human proteome. *Science*, **311**(5758), 239–242.

Ussuf, K. K., Anikumar, G., and Nair, P. M. (1995). Newly synthesised mRNA as a probe for identification of wound responsive genes from potatoes. *Indian J Biochem Biophys*, **32**(2), 78–83.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Varnum, S. M. *et al.* (2004). Identification of proteins in human cytomegalovirus (HCMV) particles: the HCMV proteome. *J Virol*, **78**(20), 10960–10966.

Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Modeling of protein interaction networks. *ComPlexUs*, **1**, 38–44.

Velculescu, V. E. *et al.* (1997). Characterization of the yeast transcriptome. *Cell*, **88**(2), 243–251.

Venter, J. C. *et al.* (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

von Brunn, A., Teepe, C., Simpson, J. C., Pepperkok, R., Friedel, C. C., Zimmer, R., Roberts, R., Baric, R., and Haas, J. (2007). Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS ONE*, **2**(5), e459.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, **18**, 1283–92.

Wang, E. T. *et al.* (2008a). Alternative isoform regulation in human tissue transcriptomes. *Nature*.

Wang, J. *et al.* (2008b). The diploid genome sequence of an Asian individual. *Nature*, **456**(7218), 60–65.

Wang, Y. *et al.* (2002). Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA*, **99**(9), 5860–5865.

Wang, Z. *et al.* (2008c). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–2.

Wei, M. C. *et al.* (2001). Proapoptotic BAX and BAK: a requisite gateway to mitochondrial dysfunction and death. *Science*, **292**(5517), 727–730.

Wheeler, D. A. *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189), 872–876.

Wieczorek, E., Brand, M., Jacq, X., and Tora, L. (1998). Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II. *Nature*, **393**(6681), 187–191.

Wills, E., Scholtes, L., and Baines, J. D. (2006). Herpes simplex virus 1 DNA packaging proteins encoded by UL6, UL15, UL17, UL28, and UL33 are located on the external surface of the viral capsid. *J Virol*, **80**(21), 10894–10899.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol*, **15**(13), 1359–1367.

Woodford, T. A., Schlegel, R., and Pardee, A. B. (1988). Selective isolation of newly synthesized mammalian mRNA after in vivo labeling with 4-thiouridine or 6-thioguanosine. *Anal Biochem*, **171**(1), 166–172.

Wu, Z. and Irizarry, R. A. (2004). Preprocessing of oligonucleotide array data. *Nat Biotechnol*, **22**(6), 656–8.

Wuchty, S. (2004). Evolution and topology in the yeast protein interaction network. *Genome Res*, **14**, 1310–4.

Xenarios, I. *et al.* (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, **30**, 303–5.

Xouri, G. *et al.* (2004). Cdt1 and geminin are down-regulated upon cell cycle exit and are over-expressed in cancer-derived cell lines. *Eur J Biochem*, **271**(16), 3368–3378.

Xu, Y., Piston, D. W., and Johnson, C. H. (1999). A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *Proc Natl Acad Sci USA*, **96**(1), 151–156.

Yang, E. *et al.* (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res*, **13**(8), 1863–1872.

Yook, S.-H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, **4**, 928–42.

Youle, R. J. and Strasser, A. (2008). The BCL-2 protein family: opposing activities that mediate cell death. *Nat Rev Mol Cell Biol*, **9**(1), 47–59.

Yu, D., Silva, M. C., and Shenk, T. (2003). Functional map of human cytomegalovirus AD169 defined by global mutational analysis. *Proc Natl Acad Sci USA*, **100**(21), 12396–12401.

Yu, H. *et al.* (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, **14**(6), 1107–1118.

Yu, H. *et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**(5898), 104–110.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, **3**(4), e59.

Zhang, B., Park, B.-H., Karpinets, T., and Samatova, N. F. (2008). From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics*, **24**(7), 979–986.

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, **4**(8), e1000140.

Acknowledgements

I owe thanks to many people who supported and encouraged me during these last years while I was working on my thesis and I can only try to acknowledge them here.

First of all, I would like to thank my supervisor, Ralf Zimmer, for giving me the opportunity to write this thesis, for allowing me the freedom to pursue my own ideas and for helpful input and discussions.

I am very grateful to Franziska Schneider for being the heart and soul of our lab and always helping me out if the administration forms proved to be beyond me; my colleagues and labmates for many interesting discussions on and, in particular, off topic; my collaborators, Lars Dölken and Jürgen Haas for giving me so many interesting and challenging problems to work on; Florian Erhard and Jan Krumsiek, two talented students, whom I had the honor to supervise and work with during my PhD thesis.

Furthermore, I would like to thank Hans-Werner Mewes for reviewing this thesis and to Volker Heun and Hans-Peter Kriegel for being part of my dissertation committee.

I am indebted to my parents for always supporting me and being proud of me, my friends off- and online for reminding me that there is a life beyond the thesis and my cat, Jasmin, for forgiving me that I neglected her occasionally because of the thesis.

Last, I need to acknowledge Jorge Cham, who always managed to grasp the realities of grad student life with his PhD comics (`http://www.phdcomics.com/`) – sometimes to close for comfort – and made me realize that there are many other grad students out there just as intimidated as I was.

Lebenslauf

Name	Friedel	
Vornamen	Caroline Christina	
Geburtsdatum	6. Februar 1981	
Geburtsort	München	
Schulische Ausbildung	Max-Born-Gymnasium Germering, Abschluß: Abitur	*1991-2000*
Studium	Bioinformatik, Abschluß: B. Sc. Ludwig-Maximilians-Universität München & Technische Universität München	*2000-2003*
	Bioinformatik, Abschluß: M. Sc. Ludwig-Maximilians-Universität München & Technische Universität München	*2003-2005*
	Bioinformatik, Promotion Ludwig-Maximilians-Universität München	*2005-2008*
Stipendien	Stipendium nach dem Bayerischen Begabtenförderungsgesetz	*2000-2005*
Beruflicher Werdegang	Praktikum bei 4SC AG, Martinsried	*Sommer 2002*
	Studentische Hilfskraft LFE für Bioinformatik und Praktische Informatik, Ludwig-Maximilians-Universität München	*2002-2004*
	Wissenschaftliche Mitarbeiterin LFE für Bioinformatik und Praktische Informatik, Ludwig-Maximilians-Universität München	*seit 1.5.2005*

Publikationen

Caroline C. Friedel, Katharina H.V. Jahn, Selina Sommer, Stephen Rudd, Hans W. Mewes, Igor V. Tetko. *Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage.* Bioinfomatics 2005, 21:1383-1388. *2005*

Caroline C. Friedel, Ralf Zimmer. *Inferring topology from clustering coefficients in protein-protein interaction networks.* BMC Bioinformatics 2006, 7:519. *2006*

Caroline C. Friedel, Ulrich Rückert, Stefan Kramer. *Cost Curves for Abstaining Classifiers.* Proceedings of the third Workshop on ROC Analysis in Machine Learning, Pittsburgh, USA, 2006.

Caroline C. Friedel, Ralf Zimmer. *Toward the complete interactome.* Nature Biotechnology 2006, 24:614-615.

Chad A. Davis, Fabian Gerick, Volker Hintermair, **Caroline C. Friedel**, Katrin Fundel, Robert Küffner, and Ralf Zimmer. *Reliable gene signatures for microarray classification: assessment of stability and performance.* Bioinformatics 2006, 22:2356-2363.

Caroline C. Friedel, Ralf Zimmer. *Influence of degree correlations on network structure and stability in protein-protein interaction networks.* BMC Bioinformatics 2007, 8:297. *2007*

Albrecht von Brunn, Carola Teepe, Jeremy C. Simpson, Rainer Pepperkok, **Caroline C. Friedel**, Ralf Zimmer, Rhonda Roberts, Ralph Baric and Jürgen Haas. *Analysis of Intraviral Protein-Protein Interactions of the SARS Coronavirus ORFeome.* PLoS ONE 2007, 2:e459.

Florian Erhard, **Caroline C. Friedel**, Ralf Zimmer. *FERN - A Java Framework for Stochastic Simulation and Evaluation of Reaction Networks.* BMC Bioinformatics 2008, 9:356. *2008*

Jan Krumsiek, **Caroline C. Friedel**, Ralf Zimmer. *ProCope - Protein Complex Prediction and Evaluation.* Bioinformatics 2008, 24:2115-6.

Caroline C. Friedel, Ralf Zimmer. *Identifying the topology of protein complexes from affinity purification assays.* German Conference on Bioinformatics (GCB) 2008.

Lars Dölken, Zsolt Ruzsics, Bernd Rädle, **Caroline C. Friedel**, Ralf Zimmer, Jörg Mages, Reinhard Hoffmann, Paul Dickinson, Thorsten Forster, Peter Ghazal, Ulrich H. Koszinowski. *High resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis, abundance and decay.* RNA 2008, 14:1959-72.

Caroline C. Friedel, Jan Krumsiek and Ralf Zimmer. *Bootstrapping the interactome: unsupervised identification of protein complexes in yeast.* RECOMB 2008, LNBI 4955, pp. 3-16.

Caroline C. Friedel, Jan Krumsiek and Ralf Zimmer. *Bootstrapping the interactome: unsupervised identification of protein complexes in yeast.* Journal of Computational Biology, accepted. *2009*