Unraveling the Structure and Assessing the Quality of Protein Interaction Networks with Power Graph Analysis

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

von

Loic Royer
geboren am 19. Juni 1977 in Paris

Betreuender Hochschullehrer:	Prof. Dr.-Ing. Michael Schroeder Technische Unversität Dresden
Gutachter:	Prof. Dr. rer. nat. Ralf Zimmer LMU München

Tag der Einreichung:	06. 09. 2010
Tag der Verteidigung:	10. 11. 2010

Contents

Acknowledgements

I would like to thank all the people who made this work possible. First and foremost my parents Ligia and Alain Royer and my sister Maria-Pia Royer for their love and support – I have missed you dearly. I also want to thank Vanessa Carlos for her love and efforts proof-reading this work until 4 o'clock in the morning – *Trago-te no meu coração*

Michael Schroeder, my thesis advisor, deserves my sincere gratitude and admiration. Many doctoral students can only dream of the effective, demanding and friendly supervision that he provides. Michael's door is always open when you need counsel and his insights often unravel the tightest knots. Importantly, he knows how to lead a team and attain a goal – and he teaches those skills. Thank you Michael.

Christof Winter, who always brings with him a bright smile and a wide range of skills that prove many times invaluable. I have much esteem for his generous and caring attitude towards others and I have learned much from him. Thank you my friend.

Matthias Reimann, with whom I spent much time discussing power graphs. Without Matthias there would no fast power graph algorithm and no plugin for Cytoscape. But most importantly, it would not have been so fun.

Suzanne Mende, Martina Maisel, Andreas Hermann, Alexander Storch, and Maik Stiehler deserved much credit for providing the biological data for our applications. I also would like to thank Francis Stewart for his collaboration on the network compressibility project.

In Michael's group I have had the chance to work among friends. Thomas Wächter, Heiko Dietze, Andreas Doms – with whom I learned much during my first year developing GoPubMed. Rainer Winnenburg and Frank Dressel, with whom I had unforgettable discussions. I want to thank especially Rainer for his attention to details when proof reading my work, and Frank for our metaphysical discussions. Bill Andreopoulos who helped us develop the first version of the power graph algorithm. Conrad Plake and Jörg Hakenberg for their skills and hard work on the BioCreAtIvE challenge. Dimitra Alexopoulou, Anne Tuukkanen, Janine Roy, Annalisa Marsico, Simone Daminelli and Mathieu Clément-Ziza for their friendship and their efforts in proof-reading parts of this thesis.

I am grateful to our system administrators Alex Mestiashvili, Nick Dannenberg, and Gregor Friedrich for helping me solving technical problems and save my time for research.

I joined Michael Schroeder's group after being introduced by Liliana and Michael Alvers. I want to thank them for this as well as for their trust and friendship.

In Dresden the Biotec / MPI-CBG / MTZ community of colleagues and friends makes life as a PhD student a great experience. In particular I would like to thank Jakob Suckale, Ana García-Sáez, Gihan Dawelbait, Andreas Henschel, Julia Winter, and Lawrence Rajendran for their help and friendship. I also want to acknowledge my 'french connection' in Dresden who have been my family here for so long – Jens Grossmann, Lucie Lagard, Cyril Massimelli, Severine Massimelli, Agnes Baumes, Falk Hoffman, Francois Weissbuch, Rita Brauer, Juliette Cavaillier, Celine Caro, Sven Hörnich, Tiphaine Cattiau, and long before that Christopher Tuot and Richard Dominé.

Publications

Journal publications

Network Compression as a quality measure for protein interaction data
Loic Royer, A Francis Stewart, and Michael Schroeder
Submitted, 2010.
Contribution: Conceived and performed the experiments, analyzed the data.
Chapter 4 is based on this publication.

Tantalum coating induces earlier differentiation of mesenchymal stem cells compared with titanium surface.
Claudia Stiehler, Cody Bünger, Rupert Overall, **Loïc Royer**, Michael Schroeder, Morten Foss, Flemming Besenbacher, Mogens Kruhoffer, Moustapha Kassem, Klaus-Peter Gunther, Maik Stiehler
Submitted, 2010.
Contribution: Analyzed the gene expression data with power graph analysis and developed the oxidative stress hypothesis.
The 3^{rd} application in chapter 6 is based on this publication.

Expression profiling and network analysis reveal MELAS master regulators.
Susanne Mende, **Loic Royer**, Alexander Herr, Janet Schmiedel, Marcus Deschauer, Thomas Klopstock, Vladimir S. Kostic, Michael Schroeder, Heinz Reichmann, Alexander Storch
Submitted to Journal of Medical Genetics, 2010.
Contribution: Analyzed the gene expression data with power graph analysis.
The 2^{nd} application in chapter 6 is based on this publication.

Genome-wide expression profiling and functional network analysis upon neuroectodermal conversion of Human mesenchymal stem cells suggest HIF-1 and miR-124a as important regulators.
Martina Maisel, Hans-Jorg Habisch, **Loic Royer**, Alexander Herr, Javorina Milosevic, Stefan Liebau, Rolf Brenner, Johannes Schwarz, Michael Schroeder, and Alexander Storch.
Experimental Cell Research, 2010. **Impact factor 3.6.**
Contribution: Analyzed the gene expression data with power graph analysis.
The 1^{st} application in chapter 6 is based on this publication.

GoGene: gene annotation in the fast lane.
Conrad Plake, **Loic Royer**, Rainer Winnenburg, Jörg Hakenberg, and Michael Schroeder
Nucleic Acid Research, 2009. **Impact factor 7.4, cited by 7.**
Contribution: Designed bibliometric measures and ranking scheme.
The co-occurrence analysis in chapter 5 is based on this publication.

Unraveling protein networks with power graph analysis
Loic Royer, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder
PLoS Computational Biology, 2008. **According to the ISCB summer newsletter 2009 it was among the journal's top 3 most downloaded papers. Impact factor 6.2, cited by 12.**
Contribution: Conceived and performed the experiments, analyzed the data.
Chapter 3 is based on this publication.

Gene mention normalization and interaction extraction with context models and sentence motifs
Jörg Hakenberg, Conrad Plake, **Loic Royer**, Hendrik Strobelt, Ulf Leser, and Michael Schroeder
Genome Biology, 2008. **Impact factor 6.2, cited by 17.**
Contribution: Designed the recall component of the gene identifier.
The gene normalization in chapter 5 is based on this publication.

Conference publications

Identification of cancer and cell cycle genes with protein interactions and literature mining
Loic Royer, Conrad Plake and Michael Schroeder
Presented at the German Bioinformatics Conference, 2009.
Contribution: Conceived and performed the experiments, analyzed the data.
The cell cycle case study in chapter 5 is based on this publication.

Workshop publications

GoPubMed: Exploring PubMed with Ontological Background Knowledge.
Heiko Dietze, Dimitra Alexopoulou, Michael R. Alvers, Liliana Barrio-Alvers, Bill Andreopoulos, Andreas Doms, Jörg Hakenberg, Jan Mönnich, Conrad Plake, Andreas Reischuck, **Loic Royer**, Thomas Wächter, Matthias Zschunke, and Michael Schroeder
In Stephen A. Krawetz, editor, Bioinformatics for Systems Biology. Humana Press, 2008. **Cited by 8.**
Contribution: Contributed to the first versions of GoPubMed.

Me and my friends: gene mention normalization with background knowledge
Jörg Hakenberg, **Loic Royer**, Conrad Plake, Hendrik Strobelt, and Michael Schroeder
Proc 2nd BioCreative Challenge Evaluation Workshop, 2007. **Cited by 16.**
Contribution: Designed the recall component of the gene identifier.
The gene normalization in chapter 5 is based on this publication.

Improving Text Mining with Controlled Natural Language: A Case Study for Protein Interactions
Tobias Kuhn, **Loic Royer**, Norbert E. Fuchs and Michael Schroeder
Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences (DILS'06), 2006. **Cited by 13.**
Contribution: Directed the experiments.

Prova: Rule-based Java Scripting for Distributed Web Applications: A Case Study in Bioinformatics.
Alex Kozlenkov, Rafael Penaloza, Vivek, Nigam, **Loic Royer**, Gihan Dawelbait, and Michael Schroeder.
In Christopher Baker and Kei-Hoi Cheung, editors, Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. *Springer*, 2006. **Cited by 19.**
Contribution: Supervised student work.

Book chapters

Querying the semantic web: A case study.
Loic Royer, Benedikt Linse, Thomas Wächter, Francois Bry, and Michael Schroeder.
In Christopher Baker and Kei-Hoi Cheung, editors, Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. *Springer*, 2007.
Contribution: Conceived the case-study and Prova examples

Summary

Molecular biology has entered an era of systematic and automated experimentation. High-throughput techniques have moved biology from small-scale experiments focused on specific genes and proteins to genome and proteome-wide screens. One result of this endeavor is the compilation of complex networks of interacting proteins. Molecular biologists hope to understand life's complex molecular machines by studying these networks. This thesis addresses tree open problems centered upon their analysis and quality assessment.

First, we introduce power graph analysis as a novel approach to the representation and visualization of biological networks. Power graphs are a graph theoretic approach to lossless and compact representation of complex networks. It groups edges into cliques and bicliques, and nodes into a neighborhood hierarchy. We demonstrate power graph analysis on five examples, and show its advantages over traditional network representations. Moreover, we evaluate the algorithm performance on a benchmark, test the robustness of the algorithm to noise, and measure its empirical time complexity at $O\left(e^{1.71}\right)$ – sub-quadratic in the number of edges e.

Second, we tackle the difficult and controversial problem of data quality in protein interaction networks. We propose a novel measure for accuracy and completeness of genome-wide protein interaction networks based on network compressibility. We validate this new measure by *i*) verifying the detrimental effect of false positives and false negatives, *ii*) showing that gold standard networks are highly compressible, *iii*) showing that authors' choice of confidence thresholds is consistent with high network compressibility, *iv*) presenting evidence that compressibility is correlated with co-expression, co-localization and shared function, *v*) showing that complete and accurate networks of complex systems in other domains exhibit similar levels of compressibility than current high quality interactomes.

Third, we apply power graph analysis to networks derived from text-mining as well to gene expression microarray data. In particular, we present *i*) the network-based analysis of genome-wide expression profiles of the neuroectodermal conversion of mesenchymal stem cells. *ii*) the analysis of regulatory modules in a rare mitochondrial cytopathy: *Mitochondrial Encephalomyopathy, Lactic acidosis, and Stroke-like episodes* (MELAS), and *iii*) we investigate the biochemical causes behind the enhanced biocompatibility of tantalum compared with titanium.

Chapter 1

Introduction

1.1 Motivation

In the last decade, molecular biology has moved from small-scale experiments focused on specific genes and proteins to genome and proteome-wide screens. The hope is that life's complex molecular machines and processes can be reverse engineered by systematic experimentation. From self-assembling modular complexes that exert structural and catalytic activities to signal transduction pathways that process information in the cell, the challenge is to understand the emerging properties of the whole from its parts. How do interacting proteins form molecular complexes? How are internal and external stimuli acquired, processed, and acted upon by the cell? The representation of systems as *networks* of interacting units is the epitome of the transition from reductionism to holism, and a central tenet of Systems Biology (Gatherer 2010). Towards the goal of understanding the cell's molecular machines, an important first step is to unravel the structure and assess the quality of complex protein-protein interaction networks. This thesis tackles three open problems relevant to the analysis and assessment of these networks and their application to elucidate regulatory pathways. First, we address the problem of finding modules in complex networks. Second, we investigate the question of quality and coverage of protein interaction networks. Third, we find key master regulators in response to external stimuli with applications from stem cell research, disease and material biocompatibility.

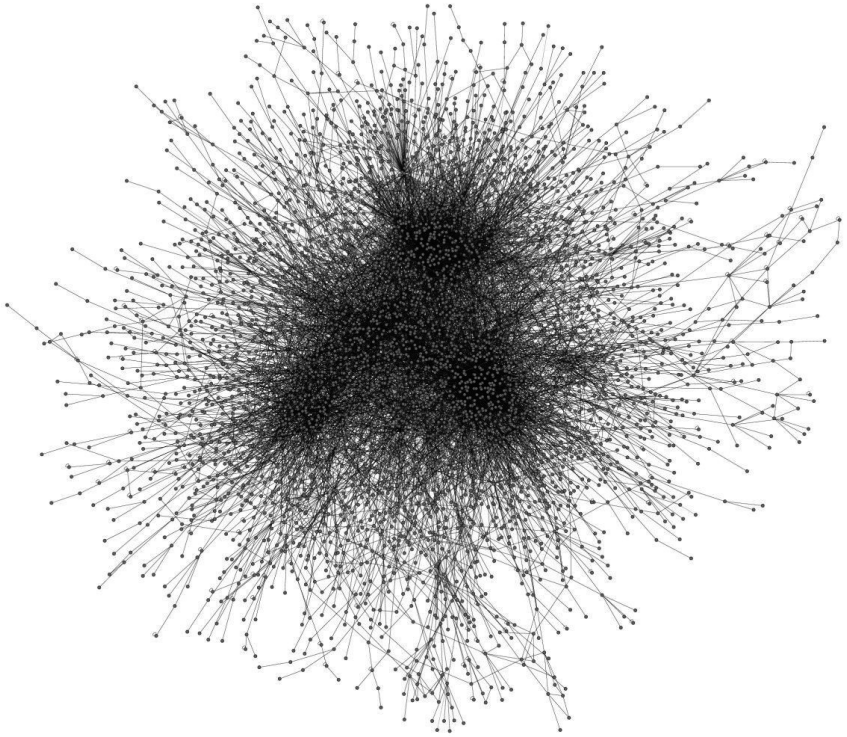


Fig. 1 **How to make sense of complex networks?** Human protein interaction network derived from experimental data by Rual et al. (2005) and Stelzl et al. (2005).

1.2 Definition of open problems

1.2.1 Open problem 1: Finding modules and preserving detail

Motivation Networks play a crucial role in biology and are often used to represent experimental results. Yet, their analysis and representation remains an open problem. Recent experimental and computational progress yields networks of increased size and complexity. There are, for example, small- and large-scale interaction networks, regulatory networks, genetic networks, protein-ligand interaction networks, and homology networks analyzed and published regularly. A common way to access the information in a network is through direct visualization, but this fails and just results in confusing “fur balls” (Fig. 1). On the other hand, clustering techniques manage to avoid the problems caused by the many nodes and edges by coarse-graining the networks and thus abstracting details. A fundamental question is: how to balance the necessity for abstraction together with the preservation of details?

Open problem How to find biologically relevant modules in protein interaction networks? In particular, how to convey without loss of information the subtle connection patterns within and between modules of proteins?

1.2.2 Open problem 2: Evaluating coverage and accuracy of protein interaction networks

Motivation In the last ten years, two experimental methods: affinity purification followed by mass spectrometry (AP/MS) and Yeast-two-hybrid (Y2H) – have emerged as popular genome-wide protein interaction mapping methodologies. Other approaches for reconstituting protein interaction networks range from computational and structural methods to manual curation and automated text-mining of large corpora of literature. Considerable obstacles have been encountered and the ways to assess data quality remain controversial. A comparison of the first genome-wide Yeast Y2H networks by Uetz et al. (2000) and Ito et al. (2001a), showed less than 20% overlap, which was slightly above random expectation and consequently raised serious challenges regarding the evaluation of data quality. Despite all of these efforts, the interaction space for most species is still sparsely explored and reliable gold standards are difficult to define (Yu et al. 2008). Consequently the problem of assessing the quality and coverage of protein interaction networks remains largely open.

Open problem How to computationally evaluate the quality of protein-protein interaction networks?

1.2.3 Open problem 3: Identification of master regulators and pathways from networks and gene expression data

Motivation Gene expression levels are controlled by a complex regulatory network involving transcription factors, microRNAs, and protein-mediated feed-back mechanisms. With the advent of gene expression micro array screens, it has become possible to measure gene expression levels genome-wide. This has allowed the investigation of gene regulatory mechanisms in the context of disease, cell differentiation, and signal transduction. Many theoretical frameworks, methodologies and tools exist to analyze gene expression datasets, but few exploit regulatory and protein-protein interaction networks to support the discovery of master regulators and pathways. Even less make use of novel network representations to determine which parts of a regulatory network are relevant and causative of changes in gene expression levels. Using a network representation that facilitates visual analytics it becomes feasible to directly analyze gene expression changes in their network context.

Open problem How to identify key master regulators and pathways with novel representations of regulatory and protein interaction networks? In particular, can we find key master regulators and pathways behind i) the neuroectodermal conversion of mesenchymal stem cells, ii) a rare mitochondrial cytopathy (MELAS), and iii) the enhanced biocompatibility of tantalum compared with titanium?

1.3 Thesis outline

The organization of this thesis is outlined in Fig. 2. After this introduction and the background section that reviews the relevant literature, the next four chapters provide answers to the three open problems. Chapter 3 introduces Power graph analysis as a novel approach for unraveling complex networks. In chapter 4 we apply power graph analysis to the evaluation of protein-protein interaction networks' quality and coverage. Chapter 5 reports contributions in text-mining and applications of text-mining to literature-derived networks. Chapter 6 presents results obtained by applying power graph analysis as well as text-mining methods to the identification of regulatory modules and pathways.

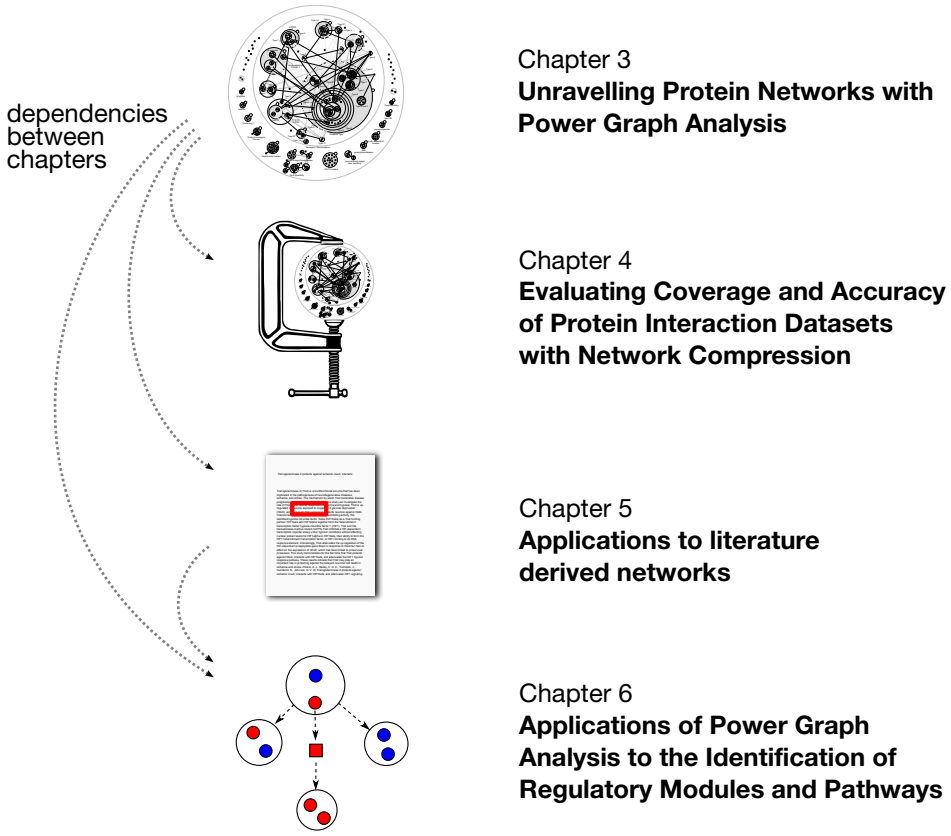


Fig. 2 **Thesis Outline.**

Chapter 2

Background

In this chapter, we review important biological facts about proteins and their interactions. In a first part, we give a detailed description of the main genome-wide interaction mapping techniques. Also, we examine the different ideas and models proposed for the topology of genome-wide protein interaction networks. In a second part, we review the problem of evaluating the quality of protein interaction networks and the ideas that have been applied to solve it. Finally, we review existing methods for protein network analysis and in particular visualization and clustering techniques.

2.1 Protein interaction networks

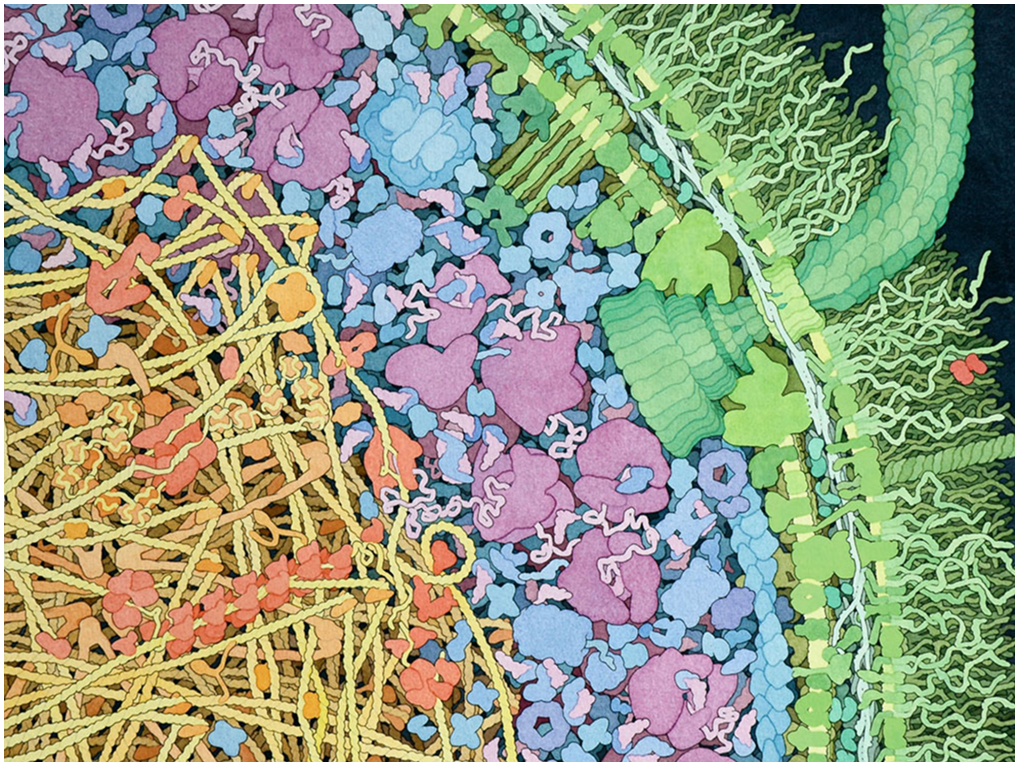


Fig. 3 **Cellular machines in the crowded interior of a bacterial cell.** This inspiring illustration by Goodsell (2009) shows the crowded interior of an Escheria coli bacterium. Proteins are the major biomolecular component of cells and amount to 70% of cellular mass. Proteins self-organize into complexes – modular cellular machines with mechanical, enzymatic and signaling functions. This image corresponds to a magnification of one million times ($\times 1,000,000$). Individual proteins can be discerned, but only large biomolecules are shown: proteins, nucleic acids, polysaccharides, and lipid membranes.

2.1.1 Proteins and their interactions

Proteins are one of the most important organic compounds for life and participate in almost every cellular process. They represent 15% of the cellular mass (Goodsell 2009), and live in a crowded space together with other biomolecules such as lipids, polysaccharides and nucleic acids (Fig. 3). Before the advent of high-throughput methods, proteins were studied following a reductionist approach by focusing on few proteins. Single genes and proteins can explain some diseases – e.g. sickle-cell disease caused by an Hemoglobin gene mutation. However, more complex diseases – cancer for example – arise from system's level disruptions. Therefore, it is necessary to examining the system of interacting proteins as a whole. In the following section we will review the relevant biology and highlight the transition from small-scale to large-scale experimentation, and how this has facilitated the study of emergent properties.

Protein interactions. Proteins interact and organize into molecular complexes – molecular machines – performing tasks essential to cellular life. Protein complexes have a structural or enzymatic activity, necessary for energy metabolism, DNA maintenance and replication. There exists a continuum of binding affinities between proteins: On the one hand we have stable protein binding with structural and mechanical functions. On the other hand we have weaker and typically transient protein interactions that are important for the transmission of information in the cell.

Modular interaction domains According to Pawson (2003) modular interaction domains form the basis of a molecular interaction code that coordinates assembly of protein complexes and networking between proteins. The assembly of multi-domain proteins from individual modular domains is a key mechanism of protein evolution and is at the core of the cell's proteomic regulation (Bornberg-Bauer et al. 2005; Pawson and Nash 2003). Multi-domain proteins contain both enzymatic domains that provide catalytic function and interaction domains that control enzymatic specificity and localization (Pawson and Nash 2003). There is a wide variety of domains – building blocks – available for mediating protein interactions as well as binding with DNA, RNA, and phospholipids. Fig. 4 shows a selection of interaction domains binding to both modified and unmodified peptide motifs, other domains, nucleic acids and phospholipids.

Domain-peptide and domain-domain interactions. Fig. 5 shows two types of binding mechanisms. On the one hand, we have domain-peptide interactions which are for example mediated by small peptide binding domains such as SH3, SH2, and PTB domains (Schlessinger 1994). These peptide binding domains confer specificity to multi-domain protein kinases and phosphatases which are responsible for most signal transduction in the cell (van der Geer and Pawson 1995; Pawson and Nash 2003). On the other hand, domain-domain interactions mediate the assembly of homodimeric and heterodimeric complexes with structural, catalytic, or regulatory functions (Fig. 5B).

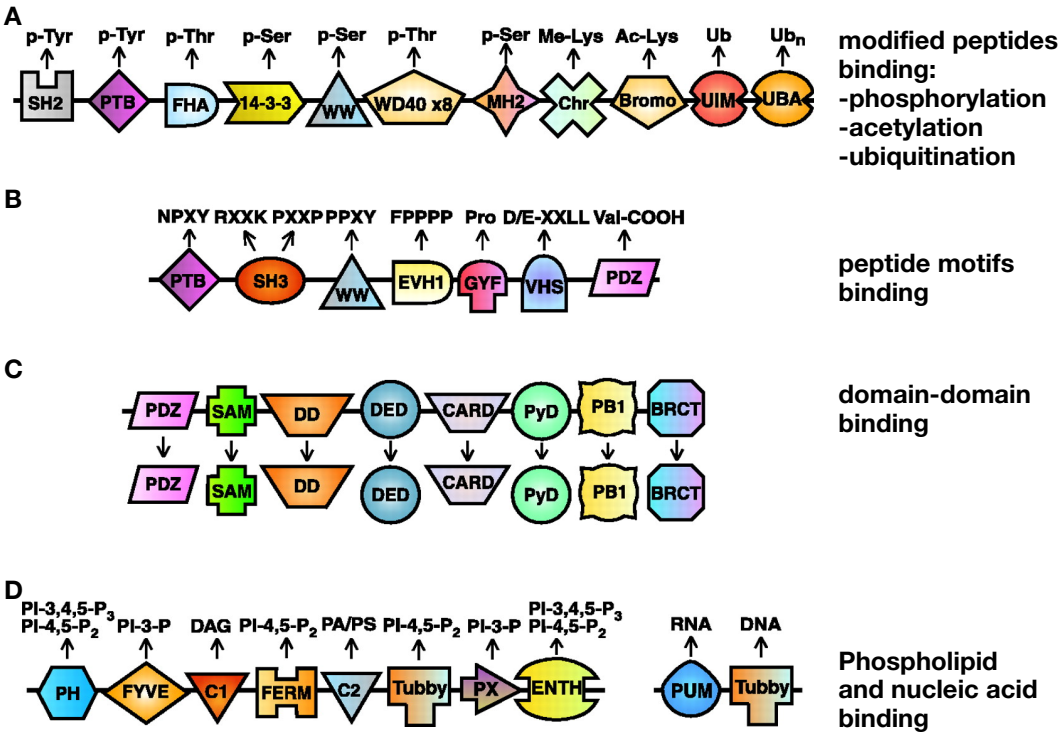


Fig. 4 **Interaction domains – building blocks for protein interactions.** A subset of interaction domains binding: proteins, nucleic acids and phospholipids is shown above. **(A,B)** Interaction domains may bind short peptide motifs. These bindings may be specific for certain post-translational modifications such as phosphorylation, acetylation and ubiquitination. **(C)** Homodimeric interaction domains. **(D)** Interaction domains may also bind to nucleic acids (DNA, RNA) and phospholipids (cell membrane) establishing a bridge between protein-protein interactions and the other molecular components of the cell. Figure adapted from Pawson and Nash (2003).

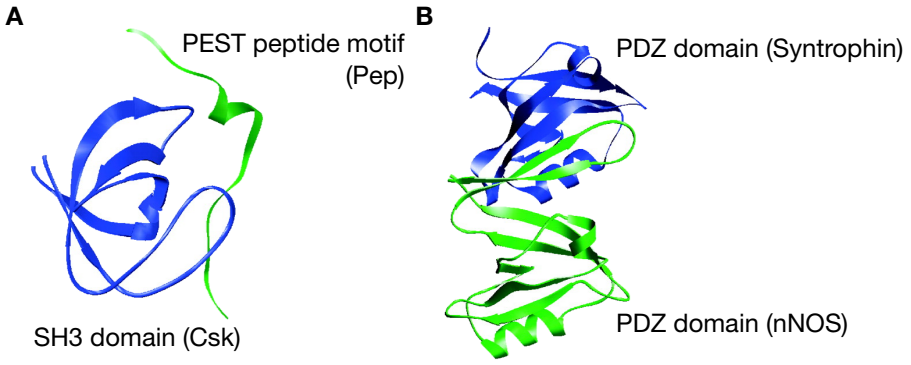


Fig. 5 **Two types of binding mechanisms. (A)** Domain-peptide binding. The SH3 domain of Csk (blue) is shown bound to the PEST peptide motif of the tyrosine phosphatase PEP (green). **(B)** Domain-domain interaction. A PDZ domain dimer of syntrophin (blue) and neuronal nitric oxide synthase (nNOS) (green). The beta-hairpin finger of nNOS is docked into the peptide binding groove of syntrophin. Figure adapted from Pawson and Nash (2003).

Cooperativity of protein interactions. The binding affinities between subunits are stronger within an assembled complex than between isolated subunit pairs (Fig. 6A).

This property called *cooperativity* is the mechanism underpinning the sequential self-assembly of protein complexes (Sorribas et al. 2007; Whitty 2008). The example illustrated on Fig. 6B shows the cooperative assembly of the IFN-beta 'enhanceosome' complex to DNA (Whitty 2008). Cooperativity shows that binary interactions are approximations of complex n-ary interactions.

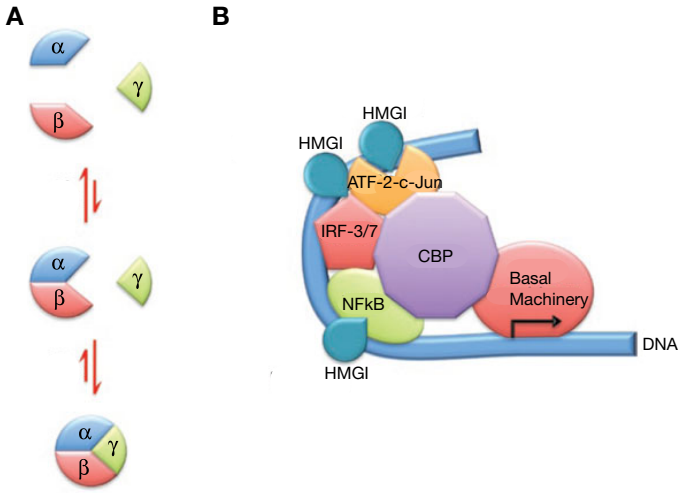


Fig. 6 **Cooperativity in protein interactions. (A)** Two proteins initially bind to form a complex $\alpha\beta$ that contains a high-affinity binding site for a third additional component γ. The resulting trimeric complex $\alpha\beta\gamma$ is stable even if all pairwise binary interactions are weak. **(B)** The IFN-beta 'enhanceosome' complex requires the cooperative assembly of multiple proteins to form a stable, functional complex on DNA. Figure and text adapted from Whitty (2008).

Experimental techniques for small-scale studies. Several experimental techniques have been developed to determine protein interactions (Shoemaker and Panchenko 2007a). The experimental gold standard is co-immunoprecipitation. Proteins are purified with a specific antibody and interaction partners are identified by western blotting. Other biochemical and biophysical approaches have been applied small-scale such as Yeast two-hybrid (Y2H), affinity purification followed by mass-spectrometry (AP/MS), protein complementation assay (PCA), chemical cross-linking, protein microarrays, fluorescence resonance energy transfer (FRET), and fluorescence correlation spectroscopy (FCS). Until about ten years ago, all studies of protein interactions were small-scale experiments conducted for few proteins. Comprehensive system-wide picture of proteins interactions required the development of novel experimental and computational resources and techniques: databases, *in silico* predictions, and high-throughput screens.

Towards large-scale – Literature curation and *in silico* predictions. Together with the systematic curation of protein interaction mentions from the biomedical literature (Reguly et al. 2006; Prasad et al. 2009), there has been intense efforts to consolidate all available data into standardized databases, including small-scale as well as large-scale experiments: IntAct (Hermjakob et al. 2004), MINT, BioGrid, DIP, HPRD (Prasad et al. 2009; Ceol et al. 2010; Breitkreutz et al. 2008; Salwinski et al.

2004; Aranda et al. 2010). However, as argued by Baumgartner et al. (2007) '*manual curation is not enough*'. This is particularly true for the curation of protein interactions from small-scale studies published in the biomedical literature. One approach is to apply text-mining methods to collect mentions of these interactions (Hoffmann et al. 2005). This remains an open problem with precision and recall well below 40% (Krallinger et al. 2008; Hakenberg et al. 2008). Alternatively, computational methods have also contributed with information on potential interactions on the base of shared evolutionary relationships (Valencia and Pazos 2002), atomic structures of proteins (Kim et al. 2006; Winter et al. 2006), and correlation with gene co-expression (Shoemaker and Panchenko 2007b).

High-throughput screens for large-scale protein interaction networks. Sanchez et al. (1999) defined the interactome as the whole set of molecular interactions in cells. The availability of comprehensive genome-wide networks comprising thousands of proteins and interactions has shifted the focus away from single interactions towards the study of proteome-wide networks. Recently, novel high-throughput approaches for Y2H (Uetz et al. 2000), AP/MS (Rigaut et al. 1999; Mann et al. 2001), and PCA assays (Tarassov et al. 2008) have been developed to characterize protein interactions at a larger scale, producing genome-wide networks of interacting proteins. As shown in Table 1, comprehensive protein interaction networks have been assembled for the following species: S. cerevisiae, C. elegans, D. melanogaster, H. pylori, H. sapiens, C. jejuni, T. pallidum, E. coli, Synechosystis, P. falciparum.

Table 1 **Genome-wide interactomes derived from large-scale Y2H, AP/MS and PCA screens.**

author	year	species	method	protocol	proteins	interactions	PubMed id
Uetz et al.	2000	Yeast	Y2H	library	806	644	10688190
Ito et al. (core)	2001	Yeast	Y2H	library	813	761	11283351
Ito et al. (full)	2001	Yeast	Y2H	library	3243	4367	11283351
Giot et al.	2003	D. melanogaster	Y2H	library	6988	20240	14605208
Stelzl et al.	2005	Human	Y2H	2-phase	1664	3083	16169070
Rual et al.	2005	Human	Y2H	library	1527	2529	16189514
Lacount et al.	2005	P. falciparum	Y2H	library	1272	2643	16267556
Sato et al.	2007	Synechosystis	Y2H	library	1915	3100	18000013
Parrish et al.	2007	C. jejuni	Y2H	2-phase	1326	11659	17615063
Titz et al.	2008	T. pallidum	Y2H	matrix	724	3627	18509523
Yu et al.	2008	Yeast	Y2H	library	2018	2705	18719252
Simonis et al.	2009	C. elegans	Y2H	library	1515	1748	19123269
Ho et al.	2002	Yeast	AP/MS	FLAG-tag	1693	8038	11805837
Butland et al.	2005	E. coli	AP/MS	TAP-SPA	1277	5324	15690043
Gavin et al.	2006	Yeast	AP/MS	TAP	1462	6942	16429126
Gavin et al.	2006	Yeast	AP/MS	TAP	1386	3244	16429126
Krogan et al.	2006	Yeast	AP/MS	TAP	2708	7123	16554755
Arifuzzaman et al.	2006	E. coli	AP/MS	His-tag	2457	8663	16606699
Collins et al.	2007	Yeast	AP/MS	TAP	1622	9070	17200106
Ewing et al.	2007	Human	AP/MS	FLAG-tag	2294	6449	17353931
Tarassov et al.	2008	Yeast	PCA	DHFR-based	1507	3030	18467557

In the following section we will give details and review the advantages and disadvantages of the three main experimental methods used for genome-wide interactome mapping – Y2H, AP/MS, and PCA.

2.1.2 Yeast two-hybrid (Y2H)

First introduced by Fields and Song (1989) the Yeast two hybrid system (Y2H) has become one of the most widely used techniques for discovering protein interactions. An interaction is detected when the binding of a transcription factor onto an upstream activating sequence (UAS) triggers the expression of its corresponding reporter gene. The reporter gene encodes for example the beta-galactosidase enzyme that causes bacteria to appear blue. As shown in Fig. 7A, the transcription factor consists of two domains: a binding domain that recognizes the UAS, and an activation domain that triggers the transcription. Two fusion proteins are prepared to test the interaction between the *bait* and *prey*. The bait is fused to the binding domain (BD), and the prey is fused to the activation domain (AD) – see Fig. 7B and C. In theory, each fusion protein alone is incapable of triggering the expression of the reporter gene. But in practice, the bait protein is sometimes capable of triggering transcription alone – leading to the problem of bait auto-activation (Ito et al. 2000). Even if both fusion proteins are present in the same cell, the two domains have to be close for triggering transcription. An interaction between the bait and the prey is defined by sufficient binding affinity which reconstitutes the transcription factor activity, and triggers reporter gene expression (Fig. 7D). In practice, bait and prey fusion proteins are brought together in one cell by mating a strain expressing the bait with a strain expressing the prey and selecting for diploids that carry both constructs. Activity of the reporter gene in these diploid cells is the read-out for the experiment.

Advantages. The main advantage of Y2H assays is their relative experimental simplicity and scalability. Interactions are detected *in vivo* without protein purification (Uetz et al. 2000; Ito et al. 2001a). A single mating operation between two genetically engineered Yeast strains is enough to test an interaction.

Limitations. There are many limitations to Y2H assays as discussed by Criekinge and Beyaert (1999) and recently by Koegl and Uetz (2007). First, the spurious activation of reporter genes by auto-activators and possibly other mechanisms such as spurious DNA binding of preys is a well known source of false positives (Rual et al. 2005). Also, indirect interactions bridged by endogenous proteins may lead to biologically irrelevant interactions, especially for proteins from higher organisms. Y2H assays are *in vivo*, but the interactions detected are initiated out of their real physiological context: the nucleus localization, the absence of co-factors, and native post-translational modifications may lead to both false positives and negatives. Also problematic is the potential toxicity of the fusion proteins when over-expressed in the Yeast nucleus. In general, bait and prey over-expression (cDNA) may lead to non-physiological binding affinities as these vary non-linearly with concentrations (Koudriavtsev et al. 2001).

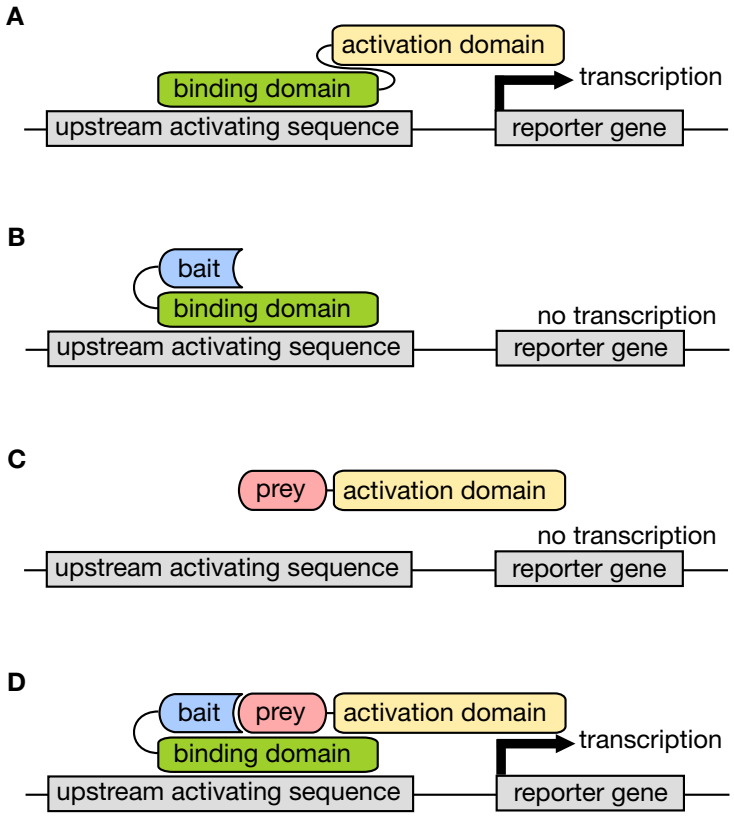


Fig. 7 **Yeast two-hybrid screening. (A)** A two domain transcription factor protein is responsible for the reporter gene expression. **(B,C)** Two fusion proteins are prepared, none of them is sufficient to independently trigger reporter gene expression. **(D)** The interaction of the bait with the prey brings the two domains in close proximity which is sufficient to reconstitute the transcription factor activity and thus express the reporter gene.

Large-scale Yeast two-hybrid. When applied large-scale for genome-wide interactome mapping Y2H screens raise several challenges. Each Y2H assay is designed to test the interaction of a pair of proteins. Testing all interactions genome-wide is unpractical. Assuming that Yeast has around 5,797 proteins, 18 million mating operations would be needed to test each pair for interaction. This approach – termed *matrix screen* – has become possible with the development of high-throughput array methodologies (Bartel et al. 1996; Uetz et al. 2000). However, these screens are necessarily restricted to species having small genomes such as T. pallidum which has an estimated 1,028 protein coding genes (Titz et al. 2008). The main challenge in genome-wide Y2H interactome mapping is to balance the scalability and qualitative aspects of the screens. Matrix screens are the most comprehensive and sensitive but their prohibitive scale is an obstacle. In the following we review the experimental strategies that have been devised to solve this problem while preserving comprehensiveness and quality.

Library based Y2H. In a library screen, one strain expressing a bait is mated with a pool of strains expressing different preys (Chien et al. 1991; Uetz et al. 2000). The diploids are selected for expression of the reporter genes and colonies of cells that

report interactions are picked (Fig. 8). At this point the identity of the preys interacting with the bait is unknown. In order to determine the identity of the preys, targeted sequencing is used. This approach involves fewer mating operations than the matrix approach but requires sequencing (Zhong et al. 2003). Moreover, since few colonies are picked for each bait, the number of interaction partners per bait is arbitrarily bounded. False positives may arise because of missing or under-represented strains due to prey protein toxicity (Zhong et al. 2003). While library screens solve the scalability of genome-wide Y2H screens, it does so at the expense of sensitivity.

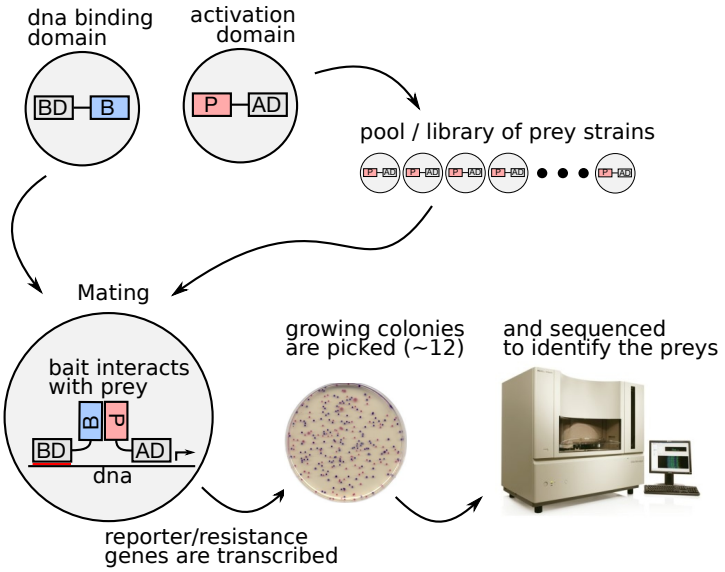


Fig. 8 **Library screens for Y2H interactome mapping.** Baits are mated to preys from a library (or pool) of prey strains. Only diploid cells that contain interacting prey and bait are selected and the resulting colonies are picked. While the identity of the bait is known for each colony, the prey needs to be identified by sequencing (Uetz et al. 2000).

Smart-pooling strategies – two-phase pooling. To mitigate the sensitivity issue of library screens, *smart pooling* strategies have been developed. One in particular – *two-phase pooling* – has been successfully applied to genome-wide interactome mapping (Stelzl et al. 2005; Parrish et al. 2007). As shown in Fig. 9, two-phase pooling is an algorithmic development: instead of relying on experimental techniques such as sequencing for prey identification, two-phase pooling implements a two phase search of the interaction space: first, it determines which baits are interacting and second, it screens preys against all baits that reported at least one interaction.

Upcoming smart-pooling strategies. Recently, even smarter pooling strategies for Y2H screening have been developed such as Shifted Transversal Design and Steiner-triple-system (Zhong et al. 2003; Jin et al. 2006; Xin et al. 2009). The common principle behind smart-pooling strategies is the redundant multiplexing of baits and preys for mating experiments followed by appropriate deconvolution algorithms for decoding the experimental results. Smart-pooling strategies can be as sensitive and specific as array-based matrix screens but at a fraction of the work-load (Xin et

al. 2009). Yet, as of 2010, no genome-wide interactome maps have been published using these novel strategies.

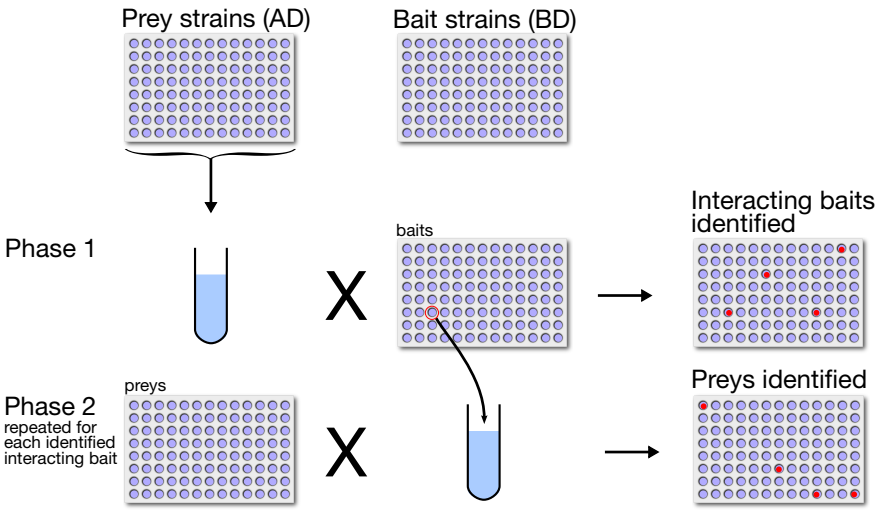


Fig. 9 **Two-phase pooling screens for Y2H interactome mapping.** Prey strains and bait strains are prepared independently and stored into several replicated arrays. In the first phase prey strains are pooled and mated to bait strains. Baits that interact with at least one of the pooled preys are detected. In the second phase interacting bait strains are individually mated to each prey strain to complete the identification of bait-prey pairs. Figure adapted from (Zhong et al. 2003).

2.1.3 Affinity purification followed by mass spectrometry (AP/MS)

Another approach to genome-wide interaction mapping is affinity purification followed by mass spectrometry (AP/MS). In this approach, a tagged bait protein is purified together with its binding proteins – the preys. As shown in Fig. 10, a fusion protein is prepared in which the bait protein is attached to a tag. Purification of bait and preys bound together is done across an affinity column, and the eluate is fed into a mass spectrometer for prey identification (Aebersold and Mann 2003). In this method, only the bait protein is engineered with a tag, while the prey proteins are in their native form. A single experiment may identify several preys and thus several interactions.

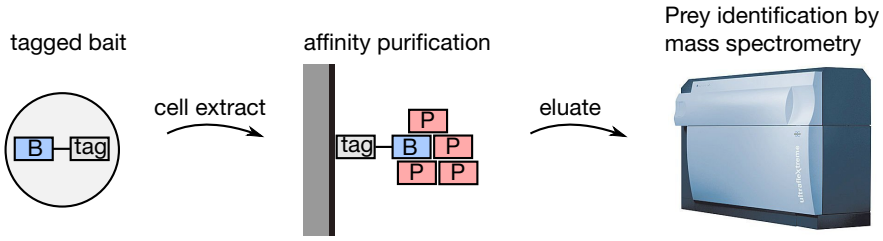


Fig. 10 **Affinity purification followed by mass spectrometry.** The three steps of AP/MS: tagging of bait proteins, affinity purification, prey identification by mass spectrometry. The expression mode may vary (overexpressed or physiological), as well as the tag (His, Flag, TAP), number of purification rounds, and mass spectrometry method used.

Different tags. The results of a AP/MS screen is strongly dependent on the ability of the tag to allow sensitive and specific purification. While His and FLAG tags have been used for interactome mapping (Ewing et al. 2007; Ho et al. 2002; Arifuzzaman et al. 2006), the state of the art is tandem affinity purification followed by mass spectrometry (TAP/MS) (Gavin et al. 2006; Krogan et al. 2006).

Tandem Affinity Purification (TAP) The TAP method was invented at the European Molecular Biology Laboratory by Puig et al. (2001). A fusion protein is prepared from the bait protein and a TAP tag. From the N-terminal the TAP tag consists of a calmodulin binding peptide (CBP), a tobacco etch virus protease cleavage site, and Protein A (Fig. 11A). Two rounds of purification are done (Fig. 11B and C) and the eluate is fed into a mass spectrometer for the identification of the preys (Fig. 11D). In the first purification contaminants are left on the column, whereas on the second purification contaminants are eluted. These two opposite rounds of purification achieve high complex purity and significantly reduce false positives when compared with single round purification protocols (Puig et al. 2001). Remarkably and despite the high purity, the TAP method is sensitive enough to detect interactions between proteins expressed at physiological levels (Puig et al. 2001).

Advantages. The main advantage of AP/MS is that the bait is engineered but not the preys which minimizes interference of the tag with the interaction interfaces. Moreover, the interactions occur *in vivo* and in context: complexes formed by the bait and the preys occur in their native context. In particular, the cellular localization and post-translational modifications are preserved. Detecting interactions in the context of other interaction partners is a strength of the approach because of the cooperativity of proteomic interactions (Whitty 2008; Sorribas et al. 2007). In TAP/MS the problems of contamination are mitigated by the highly effective purification process (Puig et al. 2001; Gavin et al. 2002).

Limitations. The main disadvantage of AP/MS screens is that the purification step may dissociate weakly interacting proteins (Gavin et al. 2002). Hence, in general AP/MS is inappropriate for detecting transient or weak interactions (Puig et al. 2001). However, Breitkreutz et al. (2010) showed recently that affinity purification techniques may be used to detect interactions between kinases/phosphatases and their regulatory subunits and substrates. This result casts doubts on the widely held belief that the sensitivity range of AP/MS does not encompass phosphorylation and signaling interactions (Yu et al. 2008).

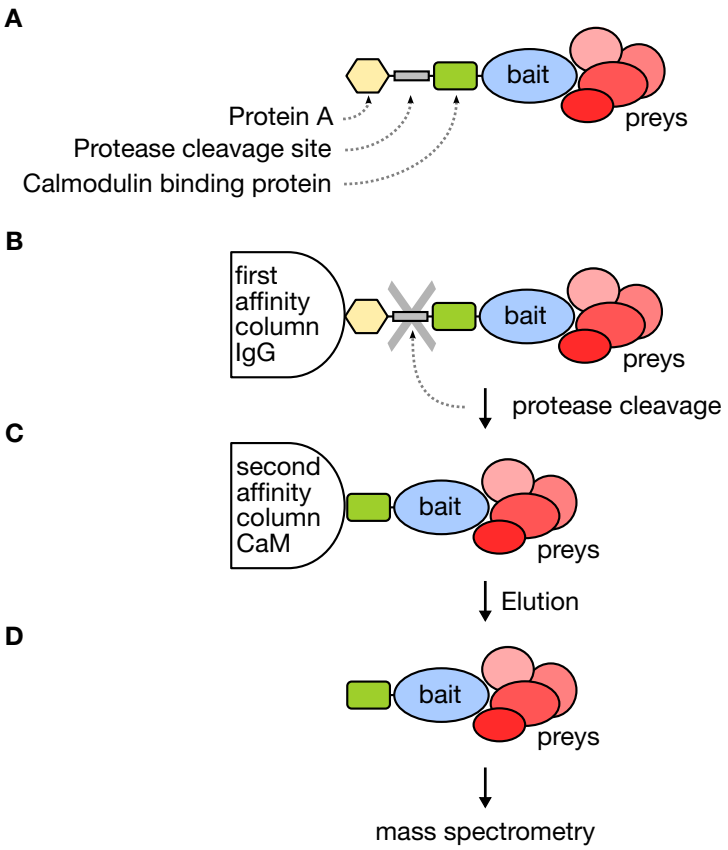


Fig. 11 **Tandem affinity purification followed by mass spectrometry. (A)** The fusion protein consists of a protein A, a protease cleavage site, a calmodulin binding protein, and the bait itself. The bait is bound to several preys which the method tries to identify. **(B,C)** The protein A binds to the affinity column IgG. Only the fusion protein and the preys remain bound to the column. A protease is then used to cleave the fusion protein. The calmodulin binding site, the bait and the preys are then released from the column. A second column with affinity to calmodulin is used for further purification. **(D)** After washing, the bait and preys are released with ethylene glycol tetraacetic acid (EGTA) and fed into a mass spectrometer for identification of the preys. Figure adapted from Puig et al. (2001).

Binary versus complex interactions Because AP/MS screens identify complexes, binary interactions can only be inferred indirectly from multiple purifications (Gavin et al. 2006; Krogan et al. 2006). The problem is further complicated by the ambiguity of the definition of interactions within complexes: Are two subunits of a complex not in direct physical contact, interacting? This point is crucial for the interpretation of AP/MS results and has led to two different interpretation models: spoke and matrix (Fig. 12). In the one hand, the spoke model underestimates interactions: the preys only interact with the bait but not with each other. On the other hand, the matrix model over-estimates interactions: the bait interacts with all preys and all preys interact with each another.

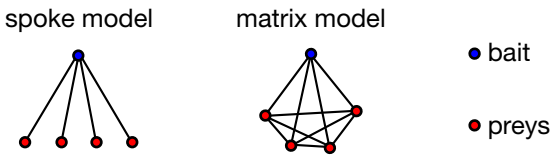


Fig. 12 **Matrix and spoke models for interpreting AP/MS results.** The spoke interpretation excludes the possibility of *indirect* interactions. The matrix model assumes all possible interactions between bait and preys. In reality, complexes resulting from purifications have both direct and indirect interactions, and not all subunits directly interact.

From complex purification to binary interactions To strike a balance between these extreme interpretations, other schemes based on the probability of occurrences of protein pairs in purifications (Hollunder et al. 2005), or based on the socio-affinity index and its variants have been developed (Gavin et al. 2006; Krogan et al. 2006). The socio-affinity index is the log-odds ratio of the number of times two proteins are observed together relative to what one would expect from their frequency in the dataset alone (Gavin et al. 2006). This provides a confidence value for each interaction. While these approaches give confidence scores for each potential interactions, other approaches aim at dissecting complexes in detail to determine the true interactions (Scholtens et al. 2005; Friedel and Zimmer 2009).

2.1.4 Protein-fragment complementation assay (PCA)

Recently, a novel approach for interaction detection – *protein-fragment complementation assay* (PCA) – was for the first time applied genome-wide in Yeast (Tarassov et al. 2008). Originally introduced for small-scale experiments by Remy and Michnick (2006), PCA proved to be a highly sensitive *in vivo* technique. It has been used in many applications, from drug discovery to protein design (Remy and Michnick 2007). Fig. 13 illustrates the method of Tarassov et al. (2008) for the first genome-wide PCA screen in Yeast. An enzyme (DHFR) consists of two complementary fragments whose activity is reconstituted upon complementation. Two fusion proteins are prepared. The first consists of the bait fused to the first fragment, and the second of the prey fused to the second fragment. If the bait does not bind or interact with the prey, the two fragments do not complement and the enzyme is inactive (Fig. 13A). Otherwise, if they do interact, the two fragments complement and the enzyme is active, reporting the interaction (Fig. 13B).

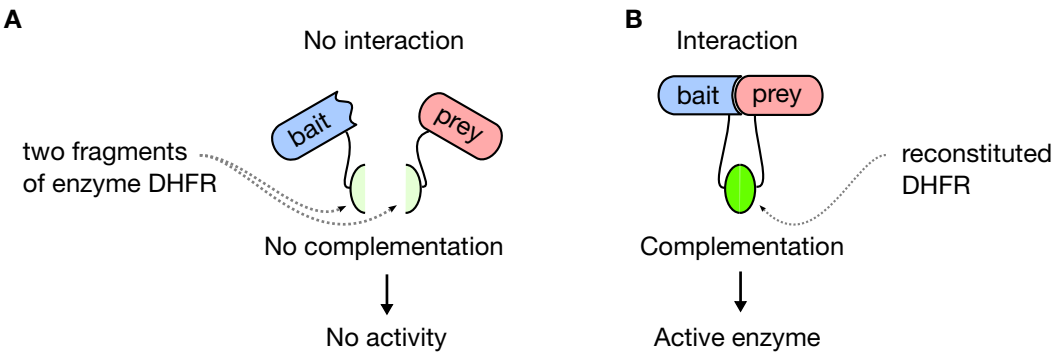


Fig. 13 **DHFR based protein-fragment complementation assay (Tarassov et al. 2008). (A)** If the bait does not interact with the prey, the two fragments of enzyme DHFR do not complement and the enzyme is inactive. **(B)** If the bait does interact with the prey, the two fragments are brought together and fold into their native structure – the enzyme is active and the interaction is reported.

Advantages. PCA screens overcome many limitations of both the Y2H and AP/MS screens because it initiates and detects interactions *in vivo* and in context. Tested proteins remain throughout the screen in their native biological state: correct post-translational modifications, correct localization, and availability of co-factors for co-operative interactions. In contrast to AP/MS, the interactions are *in vivo*, and in contrast to Y2H the interactions are functional (Tarassov et al. 2008). A definitive advantage over AP/MS is that proteins are tested for interaction in a pairwise fashion, resulting in binary interactions which avoids indirect interactions.

Limitations. The only aspect of PCA screens that may interfere with interaction detection are the reporter fragments themselves. Two problems may arise. First, the reporter fragments may drive the binding of the fusion proteins in the absence of a real interaction. The reversibility of fragment binding in DHFR-based PCA screens greatly mitigates this problem by preventing irreversible sequestration of complemented fusion proteins (Tarassov et al. 2008; Remy and Michnick 2007). Second, the fragments may disrupt binding depending on the structure of the fusion proteins and location of the interaction interfaces. This issue is common to all techniques – PCA, Y2H, and AP/MS – and is a fundamental limitation of interaction screens based on genetically engineered tagged proteins.

These three high-throughput and genome-wide interactome mapping techniques – Y2H, AP/MS, and PCA – have helped produce numerous interactome networks. In the following we review their topological characteristics.

2.1.5 Topological characteristics of interactome networks

What is the architecture and topology of current interactome networks? This question must be answered in the wider context of *complex networks* that represent entities and relationships of real-world systems, subject to dynamics and evolution (Gursoy et al. 2008). Resting on foundations from statistical mechanics, the field of *network*

science promotes the analysis of systems as graphs and searches for unifying principles (Strogatz 2001; Park and Newman 2004). In the following we review the topological properties of protein interaction networks and their biological interpretation.

Graphs. A graph $G = \left(V, E\right)$ is a set of nodes V and a set of edges $E \subseteq V \times V$ (Tutte 1998; Diestel 2005). For an edge between $u, v \in V$, we say that u is adjacent to v. Graphs can be undirected in which case $(u, v) \in E$ implies $(v, u) \in E$. Protein interaction networks can be abstracted as graphs: proteins are nodes and interactions are edges.

Clustering in protein interaction networks. The notion of clustering or edge-transitivity in networks was first introduced by Holland and Leinhardt (1971). Watts and Strogatz (1998) defined the network's clustering coefficient as the average local clustering coefficient defined for each node. The clustering coefficient $cc(u)$ of a node u is the proportion of interactions between the neighbors of u relative to the maximal number of potential interactions (see Fig. 14). Hence, this measures how close is the neighborhood of u to being totally connected.

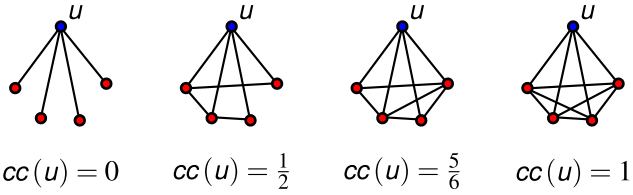


Fig. 14 **Clustering coefficient.** The clustering coefficient of a node u (blue) is the proportion of edges between its neighbors (red) relative to the total number of possible edges (here $\frac{4\times3}{2} = 6$). The clustering coefficient for the whole network is defined by taking the average for all nodes.

***Small-worlds* between order and randomness.** Early on Watts and Strogatz (1998) showed that the structure and organization of complex networks lies somewhere between order and randomness. They examined the characteristic path length of a graph G which is the average shortest path length between all node pairs. They observed that random networks have short characteristic path lengths but a low clustering coefficient, whereas ordered networks such as lattices are highly clustered but have long characteristic paths (Fig. 15). In contrast, complex real-world networks are both clustered and have short path lengths, a notion popularized as the *small-world* property (M. E. J. Newman 2003). This property was confirmed for protein interaction networks (Barabási and Oltvai 2004).

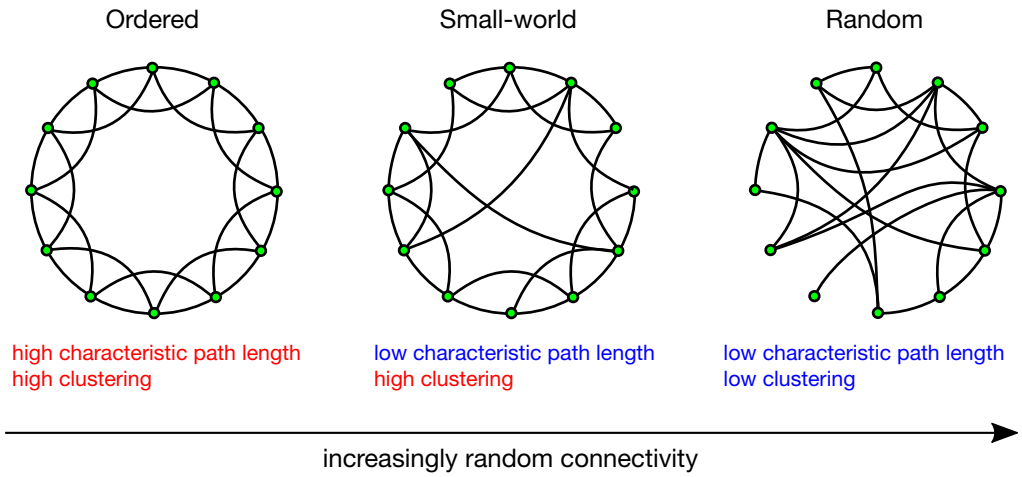


Fig. 15 **Small-world networks between order and randomness.** Ordered networks are characterized by long average paths whereas random networks are characterized by high clustering. Small-world networks have both properties. Figure adapted from Watts and Strogatz (1998).

Degree distribution and the scale-free property. The number of interaction partners of a protein in a network is a fundamental quantity also called the degree of the protein. Proteins may have few or many interaction partners and thus have a high or low degree. These differences can be quantified with the degree distribution. In their seminal work, Barabasi and Albert (1999) showed that the degree distribution of complex networks such as social communities, neural networks, and the World Wide Web, follow a power-law, implying that they have no characteristic *scale*. This result was initially accepted for protein interaction networks (Wagner 2001; Jeong et al. 2001; Rual et al. 2005) and has been interpreted as a signal of network evolution (Barabasi and Albert 1999) as well as conferring robustness to the underlying biological systems (Albert et al. 2000).

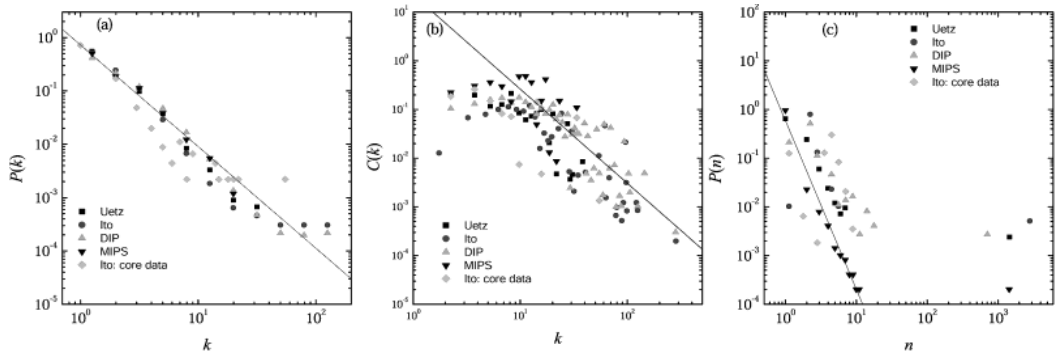


Fig. 16 **Scale-free property of Yeast protein interaction networks. (A)** Degree distribution of the four Yeast protein interaction networks. All datasets have a power law tail indicating that the underlying network has a scale-free topology. **(B,C)** Also exhibiting the scale-free property is the clustering coefficient distribution and cluster size distribution. Figure adapted from Yook et al. (2004).

However, the applicability of the power-law for protein interaction networks has been increasingly questioned. Lima-Mendez and van Helden (2009) showed that

the power law does not hold when appropriate statistical tests are applied. Other distributions have been reported to be a better fit, and the scale-free property may be a sampling artifact (Han et al. 2005; Thomas et al. 2003; Khanin and Wit 2006).

Hubs. Notwithstanding the controversy on the scale-free property of protein interaction networks, the existence of hubs – highly connected proteins – is a well established fact (Rual et al. 2005; Yook et al. 2004). Jeong et al. (2001) first observed in Yeast that hub proteins are often essential for survival. This result – termed the *centrality-lethality rule* – sparked interest in the biological explanations for hub proteins (Zotenko et al. 2008). Recently, Park and Kim (2009) revisited this question and also found a correlation between centrality and lethality. Ekman et al. (2006) showed that the many interaction partners of hubs may be explained by their enrichment in multiple and repeated domains accommodating many binding sites.

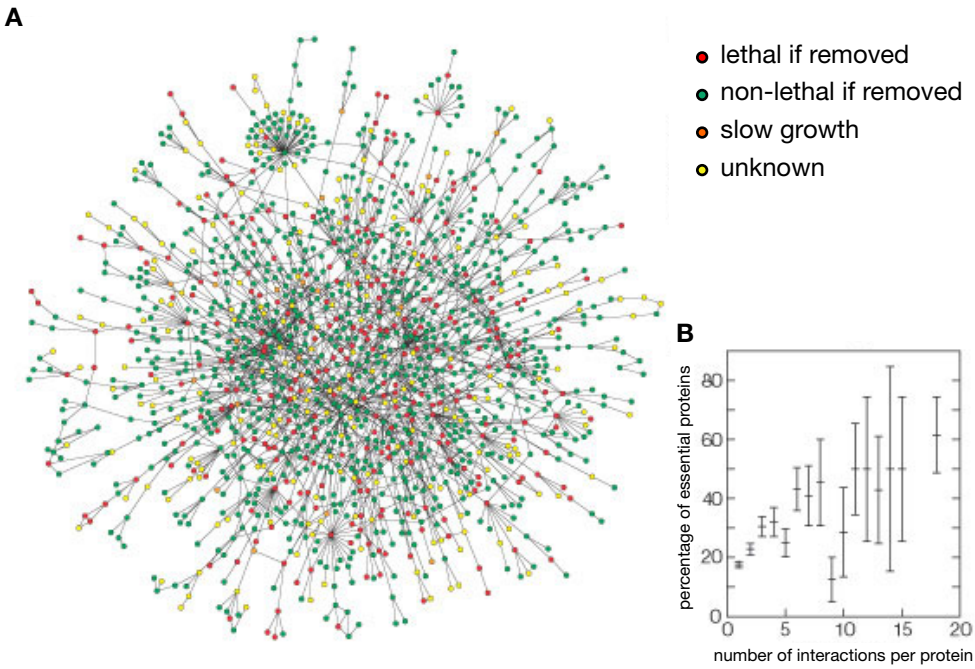


Fig. 17 **Centrality-Lethality rule. (A)** Largest connected component of the Yeast interactome obtained by Uetz et al. (2000). The phenotypic effect of removing a protein is indicated by its color in the network (see legend) **(B)** Percentage of essential proteins among proteins interacting with exactly k interaction partners. Observe that essential proteins are necessarily hubs but that the converse is not true. Figure adapted from Jeong et al. (2001).

Assortativity. Assortativity is the tendency for a network's nodes to be connected to others that are in some way similar or dissimilar (M. E J Newman 2002; M. E J Newman 2003; M. E. J. Newman 2003). As shown in Fig. 18, assortativity is defined as the proportion of homotypic interactions in a network. Homotypic interactions occur between proteins sharing some common property. In that sense, protein interaction networks have been shown to be assortative with respect to gene co-expression,

functional similarity, cellular localization, and phylogenetic profile similarity (von Mering et al. 2002; Jansen et al. 2002; Fraser et al. 2004).

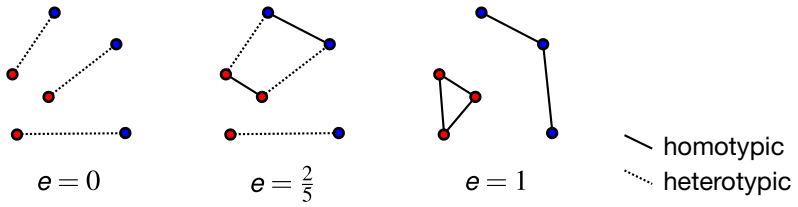


Fig. 18 **Quantifying assortativity.** A simple measure of assortativity in networks relying of the definition of homotypic interactions – adapted from M. E J Newman (2002). In the case of a two class label attached to each protein (for example lethality) homotypic interactions occur between proteins having the same class. When more than two classes are considered – for example in the case of protein cellular localization – homotypic interactions are defined disjunctively: if the two proteins are present together in at least one cellular compartment the interaction is deemed homotypic. The assortativity ratio *e* is then defined as the proportion of homotypic interactions.

Models for protein interaction network's structure and evolution. What evolutionary mechanisms and models explain the topological properties of protein interaction networks? Barabasi and Albert (1999) proposed *preferential-attachment* as a simple model explaining scaling in complex networks. In this model, the power-law degree distribution arises when newly introduced nodes are preferentially attached to already highly connected nodes. This model was given a biological interpretation with gene duplication and divergence models that implicitly follow the preferential attachment rule. In these models, newly introduced proteins are duplicates of preexisting proteins – sharing interactions but also diverging by loosing or gaining new partners (Rzhetsky and Gomez 2001; Pastor-Satorras et al. 2003; Middendorf et al. 2005; Ispolatov et al. 2005; Evlampiev and Isambert 2008). Supporting these models, Maslov et al. (2004) showed that distant paralogous proteins (around 20% sequence identity) have more similar interaction profiles than randomly selected protein pairs. Another study by Evlampiev and Isambert (2007) demonstrated that the scale-free topology of interactomes is the consequence of binding domain conservation. Another hypothesis is that simple properties of stickiness and promiscuity are enough to explain the collective organization of the networks (Deeds et al. 2006; Rachlin et al. 2006; Przulj and Higham 2006)

Size and topology of the true interactome. Results on the topology of interactomes are unreliable because of incomplete and noisy data. For example, estimates for the number of interactions in the Human interactome range from 130,000 to 650,000 (Venkatesan et al. 2009; Stumpf et al. 2008), and estimates on the reliability of high-throughput screens is typically well below 50

Many results on protein interaction networks ultimately rely on the quality of the underlying data. Defining, evaluating and comparing the quality of these networks from an experimental and computational point of view is controversial and an open problem. In the following we review the literature around this question.

2.2 Evaluation of interactome quality

2.2.1 Controversy on data quality

The question of data quality in protein interaction has encountered considerable obstacles and the ways to assess the quality of genome-wide interactome data remain controversial. Comparison of the first genome-wide Yeast Y2H networks by Uetz et al. (2000) and Ito et al. (2001b), showed less than 20% overlap, which was slightly above random expectation and consequently raised serious challenges regarding the quality of the data. By what criterion could Y2H data be confirmed? Shortly after, Gavin et al. (2002) performed one of the first large-scale screens using AP/MS (Shevchenko et al. 1999; Deshaies et al. 2002). Later screens by Gavin et al. (2006) and Krogan et al. (2006) were merged and filtered for false positives by Collins et al. (2007b). The quality of this interaction network was confirmed by an intensive study of a chromatin centered network by Shevchenko et al. (2008) termed *Chromatin Central*. Recently, Tarassov et al. (2008) completed the first genome-wide PCA screen in Yeast. Yet, no significant overlap has been found among the different methods: only 42 interactions between 287 proteins are found by Y2H, AP/MS and PCA (Fig. 19A and B).

Possible interpretations. Several interpretations can be given for the lack of agreement between experiments and across experimental methods. First, invoking the scientific principle of *reproducibility*, the lack of overlap between datasets such as Uetz et al. (2000) and Ito et al. (2001b) simply indicates poor data quality. Indeed, early reports gave false positive rates at 50-70% for large-scale Y2H screens (Ito et al. 2001b; Deane et al. 2002). Using a benchmark dataset, von Mering et al. (2002) reported that Y2H datasets had an estimated 1% coverage and 5% accuracy, whereas AP/MS methods had 35% coverage and 12% accuracy. Yet, recent estimates by Lemmens et al. (2010) put the accuracy of Y2H screens between 20 and 35%. Another interpretation is that experimental biases and differences in interaction search space might explain the lack of overlap (Venkatesan et al. 2009). Yu et al. (2008) argue that both Y2H and AP/MS methods have high specificity but explore distinct interaction subspaces (Fig. 19), with AP/MS favoring stable intra-complex interactions and Y2H transient inter-complex interactions.

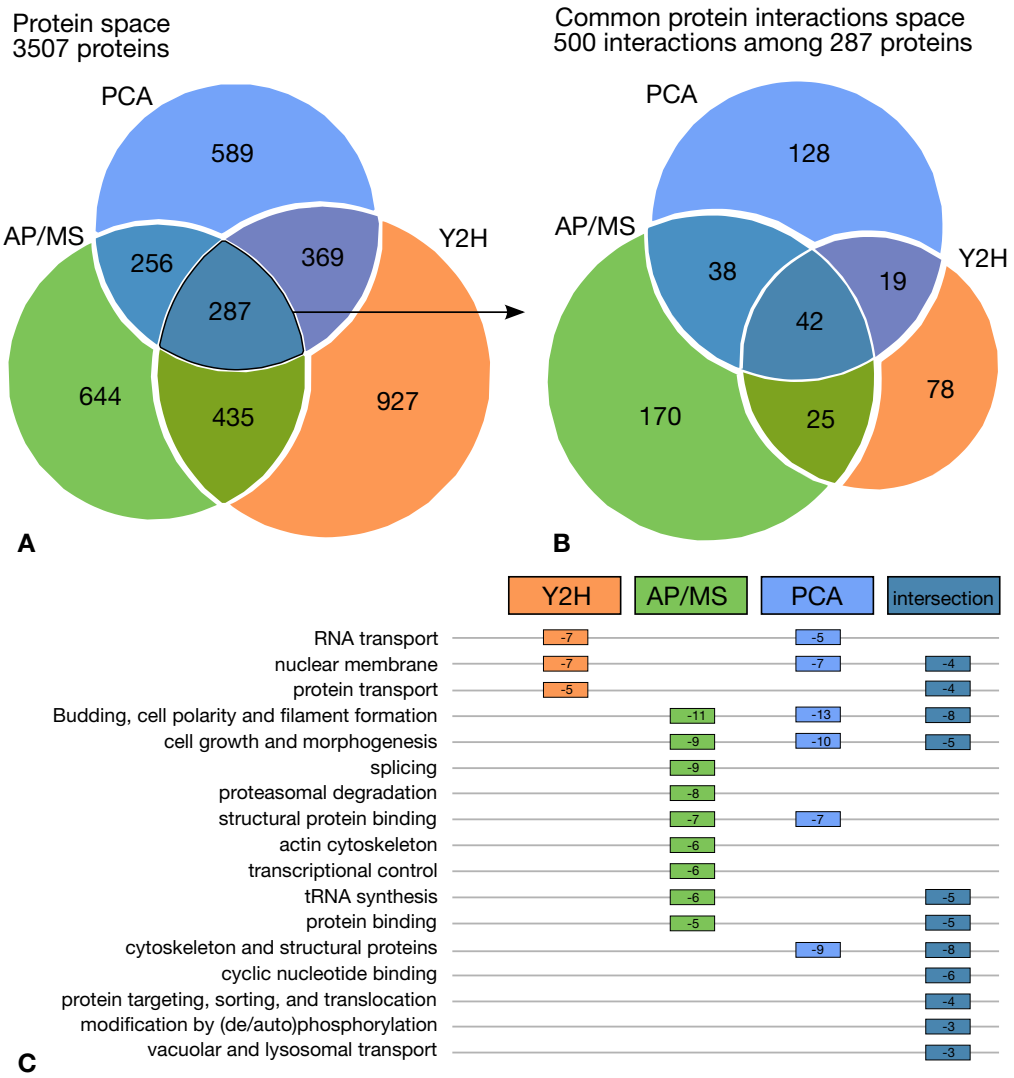


Fig. 19 **Low quality or distinct interaction spaces? (A)** Overlap between the subsets of Yeast proteins screened by Y2H, AP/MS, and PCA. In total, 3507 Yeast proteins were found to interact at least once by any of the three methods. Only 287 proteins were found by all methods to be part of an interaction. **(B)** Common protein interaction space. Between the 287 proteins explored by all methods, 500 interactions were reported in at least one experiment. Only 42 interactions were confirmed by all three methods. **(C)** Enrichment analysis of the common protein interactions space. Following the Venn diagram B, we show enriched MIPS annotations for proteins participating in interactions specific to each method (Y2H, AP/MS and PCA) and common to all (intersection). The number in each box is the p-value majoring exponent for the enrichment ($p < 10^x$). (Own analysis)

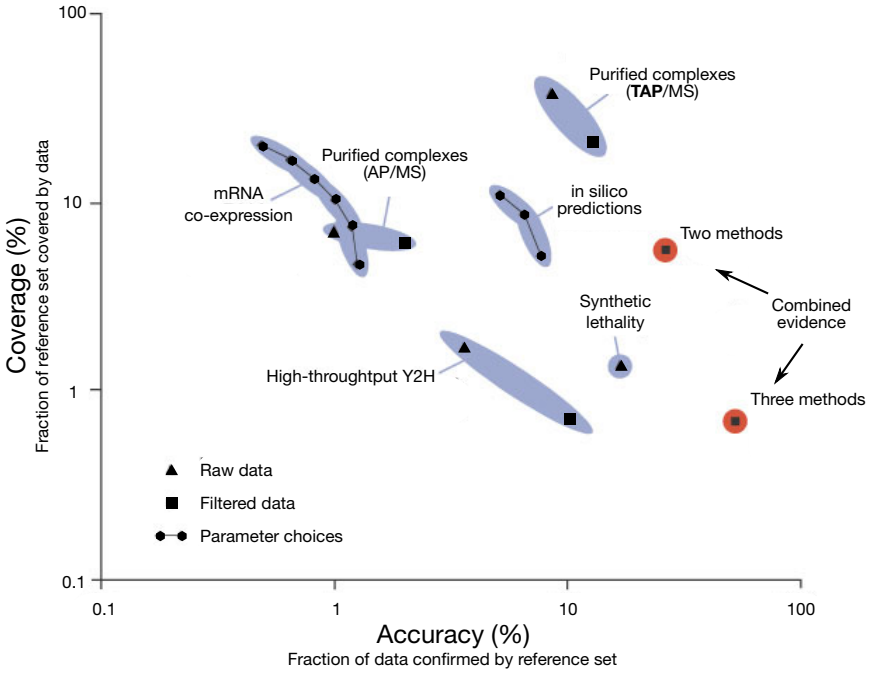


Fig. 20 **Yeast interactome benchmark based on reference set of 10,907 trusted interactions (von Mering et al. 2002).** Each dot in the graph represents an entire interaction dataset, and its position specifies coverage and accuracy. For most data sets, raw and filtered data are shown, demonstrating the trade-off between coverage and accuracy achieved by filtering. Figure and caption adapted from von Mering et al. (2002).

Network topology of the true and complete interactome. Comparison of AP/MS and Y2H methodologies is complicated by a fundamental difference: Y2H detects binary interactions whereas AP/MS detects complexes from which binary interactions are inferred. Furthermore, AP/MS datasets can be interpreted using the spoke model (only bait-prey interactions) or matrix model (all interacting with all). This complicates the comparison between Y2H and AP/MS networks: topological differences are difficult to disentangle from the experimental methodology. Yu et al. (2008) showed that networks derived from different experimental methodologies have different network topologies with AP/MS networks being more clustered than Y2H networks . Fig. 21 shows the differences in overall network organization between Y2H and AP/MS interactome networks. the combined Y2H network (Y2H-union) has a less clustered structure than the combined AP/MS network (combined-AP/MS) characterized by a lower clustering coefficient (Yu et al. 2008). Interestingly, the Yeast Y2H network has a markedly less clustered structure when compared to a high-confidence literature curation network. This point is further emphasized by recent data from the first large-scale in vivo protein-fragment complementation assay (PCA) in Yeast by Tarassov et al. (2008). The resulting network has a similar clustered topology as AP/MS screens, tilting the balance in favor of a clustered topology for interactomes (Tarassov et al. 2008). Furthermore, the analysis by Friedel and Zimmer (2006) showed that limited sampling in Y2H screens can significantly lower their clustering coefficients. The counter-argument explaining the low clustering of Y2H networks by transient signaling interactions is at odds with experimental

evidence. Breitkreutz et al. (2010) (see Fig. 1 A) found significant clustering among phosphatases, kinases and their substrates in Yeast.

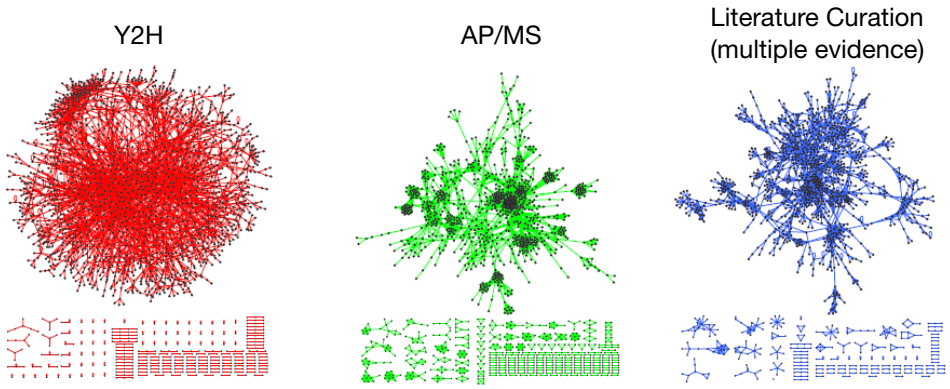


Fig. 21 **Different experimental methods produce markedly different network structures.** Yu et al. (2008) showed that the Yeast AP/MS interactome is more clustered than its Y2H counterpart. Figure and analysis from Yu et al. (2008).

Sensitivity and specificity. The quality of protein interaction networks is primarily measured by specificity – estimating the proportion of true interactions relative to the detected interactions (Ito et al. 2001b; Deane et al. 2002; von Mering et al. 2002). However there is increasing evidence that sensitivity and coverage are also important. Recently, Gerber et al. (2009) exhaustively screened interactions between 43 Streptococcus Pneumoniae proteins with 14,792 microfluidic affinity experiments (43^2 experiments repeated four times in both directions) and reached the conclusion that many interactions are missed by conventional screens. The recent work by Yu et al. (2008) shows that reaching saturation in Y2H screens requires multiple screens of the same interaction space. However, it remains that Y2H screens have a low sensitivity (see Fig. 22).

It is increasingly recognized that overall screen sensitivity is also a critical aspect of interactome quality. Venkatesan et al. (2009) analyzed the completeness, sensitivity, and specificity of interactome mapping methodologies and concluded that measures of interactome quality must take all possible sources of false positives and negatives into account. The challenge is to achieve both high sensitivity and high specificity.

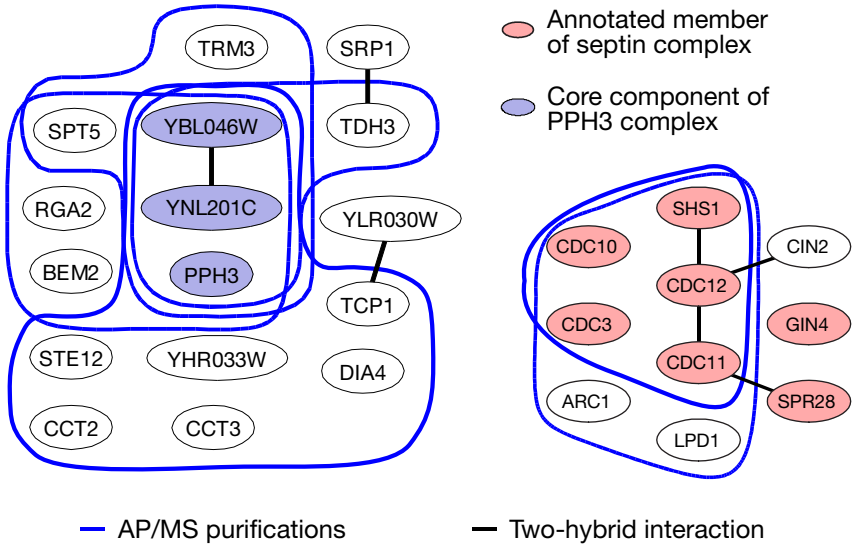


Fig. 22 **Overlap between binary Y2H interactions and AP/MS purifications.** The cores of the Septin and PHP3 complexes are identified by the overlap of AP/MS purifications. In contrast, few binary Y2H interactions are detected. Figure adapted from von Mering et al. (2002).

2.2.2 Approaches to assess the quality of protein interaction networks

Gold standards for evaluating interaction data quality. One approach to quality evaluation is to compare error-prone high-throughput data with interactions curated from literature on small-scale interaction studies. Indeed, manually curated interactions supported by multiple, independent pieces of evidence may be considered a gold standard (Prasad et al. 2009; Reguly et al. 2006). By recuration of a random sample of Human interactions mentioned in at least two publications, Cusick (Cusick et al. 2009) recently reported that 91.5% were correct. Several high-confidence datasets have been constructed by pooling information from literature curation and experimental data such as the 'binary-GS' dataset for Y2H by Yu et al. (2008) or the MIPS complex database for AP/MS by Mewes et al. (1999). Another approach is the compilation of interactions derived from 3D template structures – the Structural Interaction Network (Kim et al. 2006). While the coverage of known protein structures is still limited, this approach provides a network of high confidence interactions with verified binding interfaces. Despite all of these efforts, the interaction space for most species is sparsely explored and reliable gold standards are difficult to define. Consequently the problem of assessing the quality and coverage of protein interaction networks remains largely open.

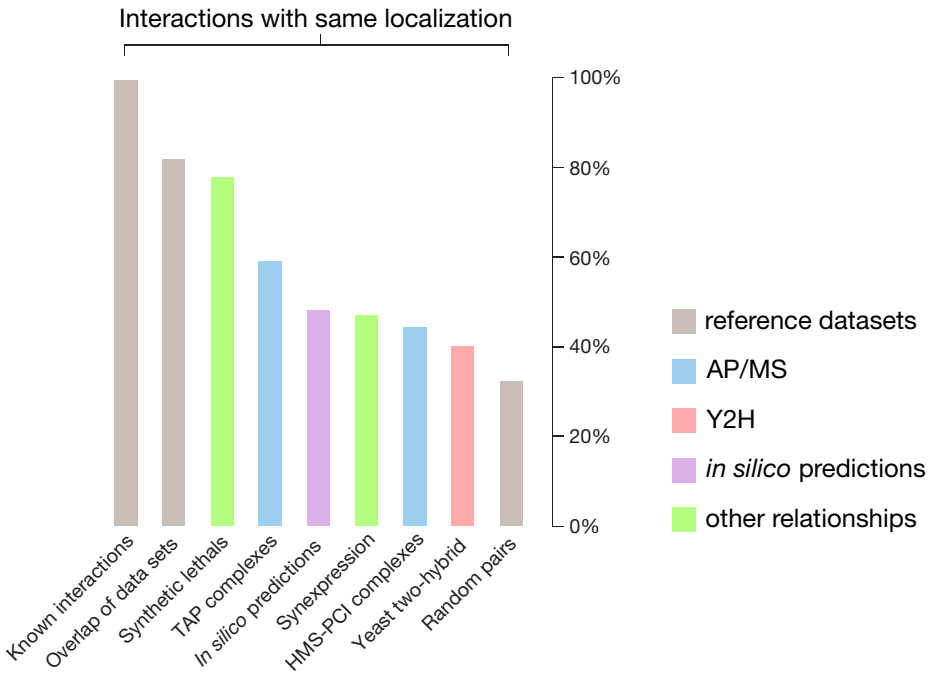


Fig. 23 **The fraction of interactions in which both partners have the same protein localization.** Combined Yeast networks derived from AP/MS, Y2H, and in silico predictions are compared to reference datasets, genetic and co-expression (synexpression) networks and a random baseline. Only proteins clearly assigned to a single category are considered. Figure and caption adapted from von Mering et al. (2002).

Validation by gene co-expression, functional similarity, cellular localization, and phylogenetic profile similarity. Protein interaction networks are assortative to gene co-expression, functional similarity, cellular localization, and phylogenetic profile similarity (von Mering et al. 2002; Jansen et al. 2002; Deane et al. 2002; Deng et al. 2003; Fraser et al. 2004; Yu et al. 2008). At the risk of circularity, this hypothesis can be used to evaluate interaction data quality. For example see Fig. 23 and Fig. 24 from von Mering et al. (2002) and Yu et al. (2008) which compare protein localization and gene co-expression for consolidated Y2H and AP/MS networks. This evidence favors AP/MS as the method producing the most assortative networks. However, the question remains whether assortativity is really a signal of network quality. Although widely accepted, this hypothesis remains unproven in the absence of truly reliable reference networks. Suthram et al. (2006) compared different approaches for assigning confidence values to Yeast protein interactions based on experiment type, protein function, localization, gene expression, and conservation information. Their conclusion is that the the best model – the work of Deng et al. (2003) – only uses two features: gene expression profile similarity and experiment type.

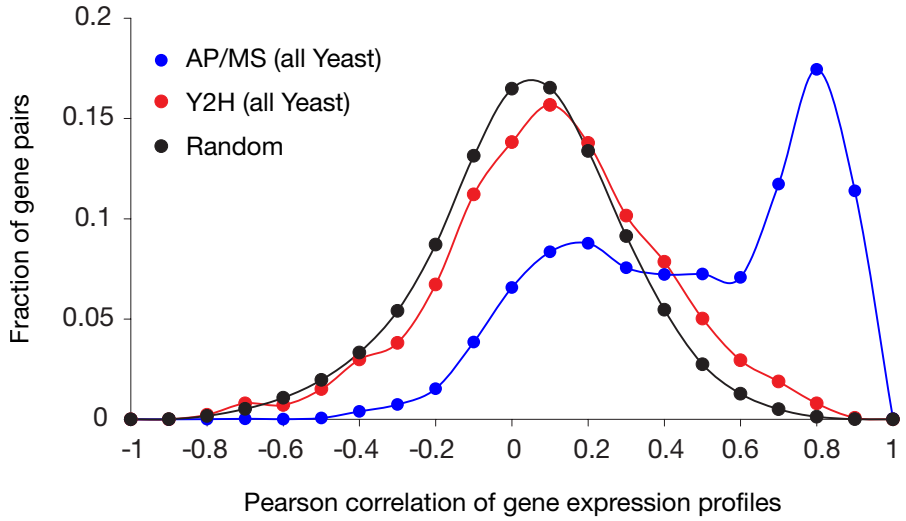


Fig. 24 **Gene coexpression correlation between interacting pairs.** Two consolidated Yeast protein interaction networks – Y2H-union and combined-AP/MS – are compared with respect to gene expression correlation. A random network is added for comparison. Note that Y2H gene expression correlation is slightly shifted in the direction of positive correlation (+0.1) when compared to random, whereas the combined AP/MS network distribution has a strong peak at around 0.8. Figure and caption adapted from Yu et al. (2008).

In the following we review the different approaches to the analysis of protein interaction networks.

2.3 Unraveling protein interaction networks

Previously we have discussed how topological descriptors such as clustering coefficient, degree distributions, assortativity as well as network motifs and patterns can help to characterize protein interaction networks. In the following we review different approaches for the analysis of protein interaction networks (Fig. 25). First, we discuss visual analytics applied on graphs and their adjacency matrices. Second, we review the literature for modules, motifs, and patterns in protein interaction networks. Third, we examine data-clustering approaches for networks and in particular agglomerative hierarchical clustering.

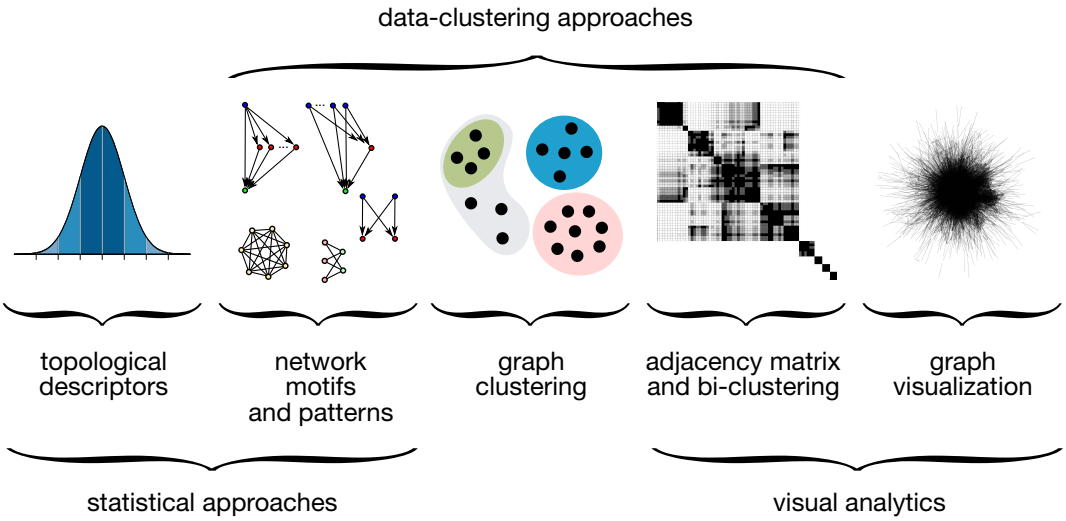


Fig. 25 **Approaches for unraveling protein interaction networks.** Different network analysis approaches share core concept from data-clustering, statistics and visual analytics. For example: topological descriptors such as the clustering coefficient are linked to network motif composition; network motifs can be found by clustering techniques; and clustering techniques such as bi-clustering are the basis of adjacency matrix visualization.

2.3.1 Visual analytics applied to protein interaction networks

Fur balls. A popular approach to the analysis of protein interaction networks is direct visualization. Fig. 26 shows four such examples published in high-impact publications (Rual et al. 2005; Bader et al. 2003; Stelzl et al. 2005; Kim et al. 2007). This approach results in pretty yet complicated *fur balls*. The problem stems from the many edges and edge crossings – caused in part by the small-world topology and the non-planarity of most graphs. Visual analytics on networks remains an open problem (Chen 2005) despite many advanced techniques for graph layout (Han and Ju 2003; Han and Byun 2004; Dwyer et al. 2006; Schreiber et al. 2009; Kojima et al. 2010) and interactive exploration (Shannon et al. 2003; Huttenhower et al. 2009; Hu et al. 2007).

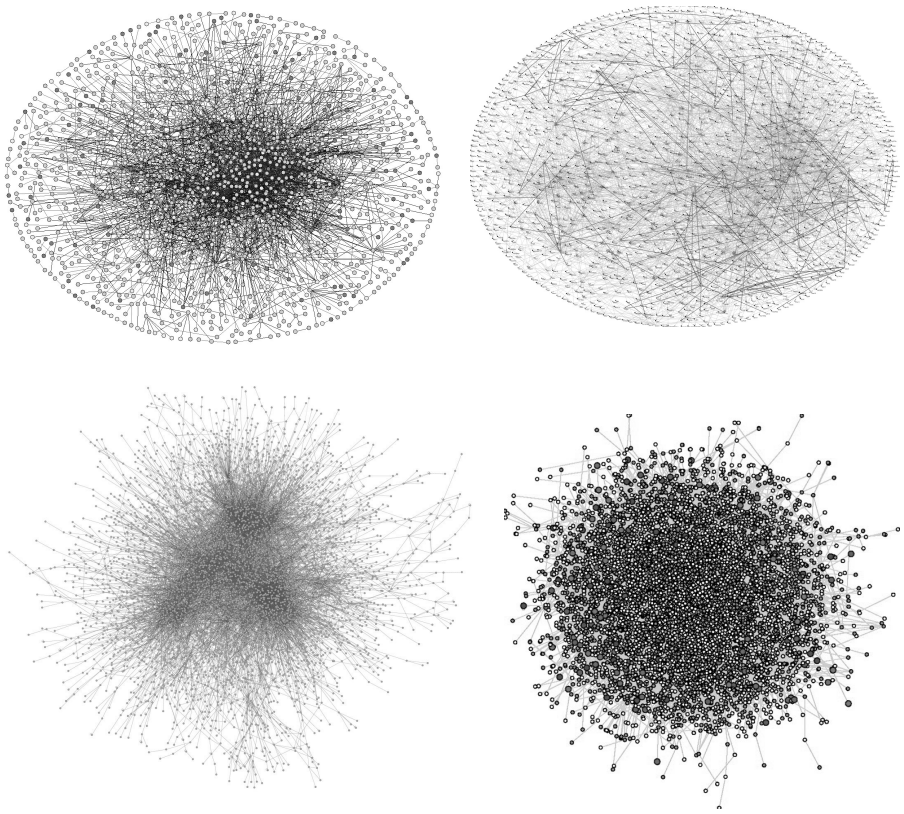


Fig. 26 **'Fur balls' published for protein interaction networks** Example networks by Rual et al. (2005); Bader et al. (2003); Stelzl et al. (2005); Kim et al. (2007).

Hypergraphs and other graph generalizations. Faced with the explosion in visual complexity caused by many edges and edge intersections, alternative graph formalisms have been proposed such as *hypergraphs*. Fig. 27 shows that instead of edges, hypergraphs have hyper-edges which are *sets* of nodes (Berge 1976). Ramadan et al. (2004) proposed hypergraphs as a convenient model for protein complexes. The main advantage of hypergraphs is that many edges can be abstracted as one hyper-edge (Klamt et al. 2009). However, metagraphs introduce many more layout problems (Klamt et al. 2009). Other alternative graph formalisms are *metagraphs* or *compound Graphs* in which the nodes are collapsed into *metanodes* (Hu et al. 2007).

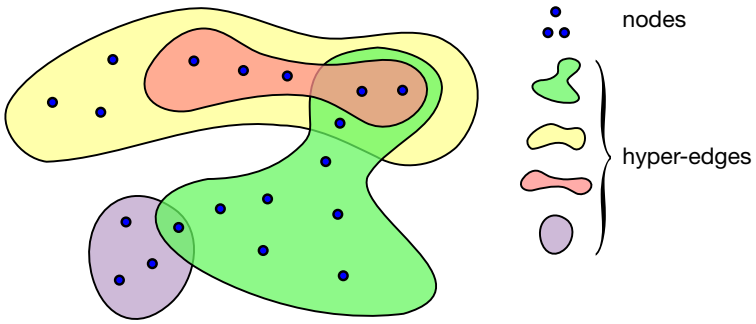


Fig. 27 **HyperGraphs.** In hypergraphs edges are not only a pair of nodes but instead an arbitrary large set of nodes.

Visualization tools and integrated frameworks. Several tools exist for visualization and visual analytics in biological networks. According to a survey by Suderman and Hallett (2007) more than 35 different network visualization tools exist in the biological domain. The most widely used are Cytoscape (Shannon et al. 2003), Pajek (Batagelj and Mrvar 1998), Osprey (Breitkreutz et al. 2003), Navigator (Motamed-Khorasani et al. 2007), VisANT (Hu et al. 2007), ProViz (Iragne et al. 2005), MOVE (Bosman et al. 2007) and GraphViz (Gansner and North 2000). Integrated frameworks and tools provide more than just network visualization but also storage and integration of other data types. Aragues et al. (2006) developed PIANA, a tool for the integration and analysis of protein interaction networks. Recently, Wu et al. (2009a) introduced a *Protein Interaction Network Analysis platform* (PINA) which integrates protein interaction networks from several databases and provides tools for network construction, filtering, analysis and visualization.

Matrix representations and bi-clustering techniques. Protein interaction networks can be indirectly visualized by their adjacency matrix. The adjacency matrix of a graph contains all of its connectivity information. The rows and columns of the matrix correspond to proteins in the network and each value in the matrix represents a confidence or probability. Since adjacency matrices are equivalent up to a permutation of columns and rows, the order of rows and columns is often chosen by bi-clustering (Cheng and Church 2000; Ding et al. 2006; Barkow et al. 2006). An example is the genome-wide PCA network by Tarassov et al. (2008) represented as a bi-clustered adjacency matrix (Fig. 28). On the on hand, the advantages of this representation is that the whole network can be represented without any overflow in visual complexity – no 'fur ball' effect. Also, it provides useful information about interaction profile similarity. On the other hand, the matrix representation is almost empty since for sparse protein interaction networks most of the information is located along the diagonal.

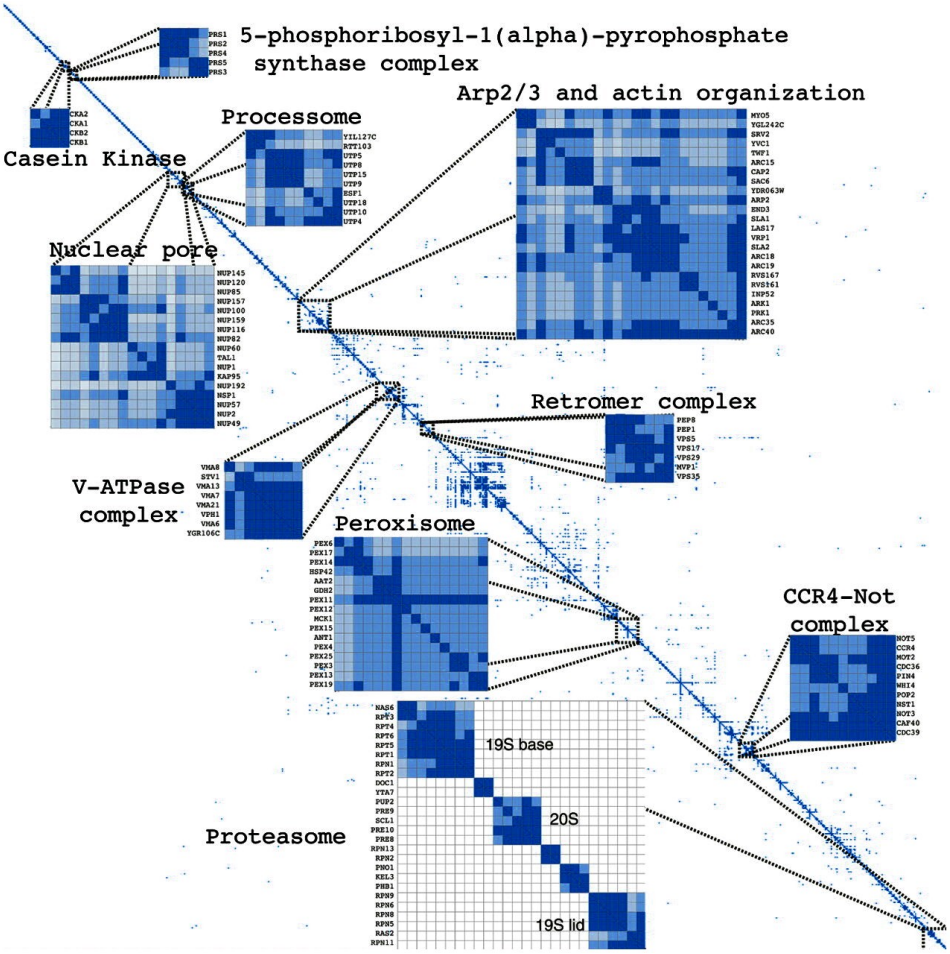


Fig. 28 **Bi-clustered matrix representation of a genome-wide PCA network.** Most of the connectivity information is close to the diagonal and only sparsely populates the non-diagonal regions of the matrix. Network and figure by Tarassov et al. (2008).

Bi-clustering (also called two-mode clustering or co-clustering) is a data-mining technique first introduced by Hartigan (1972) that simultaneously clusters the columns and rows of matrices. Several excellent reviews on the field exist e.g. Mechelen et al. (2004) and Tanay et al. (2005). It was first applied to psychology and social network analysis as a method for block-modeling (Breiger et al. 1975; Arabie et al. 1978). The term itself was first used by Mirkin (1996). Since then the method has been popular in bioinformatics for the analysis of microarray data (Cheng and Church 2000; Tanay et al. 2002; Kluger et al. 2003; Sheng et al. 2003; Koyuturk et al. 2004). Several software tools exist for performing biclustering such as BicAT (Barkow et al. 2006), cMonkey (Reiss et al. 2006), and for their visualization such as BiVoC (Grothaus et al. 2006), and BiVisu (Cheng et al. 2007). In particular, Ding et al. (2006) showed the relevance of bi-clustering to the analysis of protein-protein interaction data. Bein et al. (2008) showed the link between clustering rows and columns of adjacency matrices and the *Biclique partition problem*.

In the following we introduce the main patterns detected by bi-clustering, and review these network motifs and other inter-connection patterns in protein interaction networks.

2.3.2 Motifs and inter-connection patterns in protein interaction networks

Network motifs. First introduced by Alon (2007), network motifs are over-represented inter-connection patterns occurring in complex directed networks (Fig. 29). Initially discovered in E. coli gene regulation networks, network motifs have been found to be statistically over-represented in a wide range of networks including neuronal networks, food webs (Milo et al. 2002), phosphorylation networks (Breitkreutz et al. 2010) and protein interaction networks (Albert and Albert 2004; Wuchty et al. 2003). While most protein interaction networks are undirected, network motifs have been detected in directed phosphorylation networks. For each pair of interacting proteins the role of enzymes and substrates are known (Breitkreutz et al. 2010). Several tools based on efficient and scalable heuristics have been developed for the fast enumeration of network motifs (Wernicke and Rasche 2006; Koyutürk et al. 2004; Przulj et al. 2006).

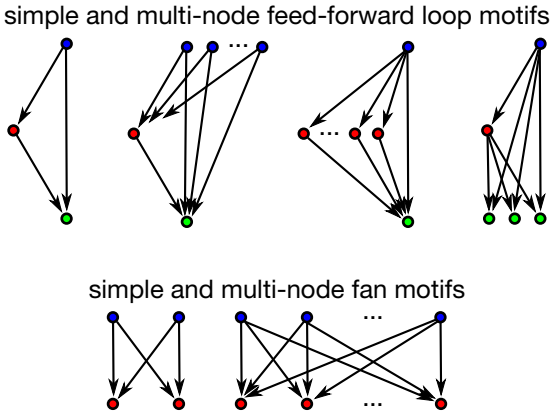


Fig. 29 **Simple and multi-node feed-forward and bifan motifs.** *Feed-forward loop motifs* act as persistence detector circuits or 'debounce' circuits that reject transitory signals. *Bifan* motifs perform combinatorial combination of input signals. Terminology by Kashtan et al. (2004); Alon (2007).

Bicliques, cliques and stars. Motifs in protein interaction networks can be decomposed into more primitive inter-connection patterns. Fig. 30 shows three types of sub-graphs: bicliques, cliques, and stars. Bicliques or *complete bipartite sub-graphs* are sub-graphs in which all nodes of a first set are adjacent to all nodes of a second set (Fig. 30A). A biclique between a set of m nodes and a set of n nodes is denoted $C(m,n)$. A clique or *complete graph* is a set of nodes which are all adjacent to each other; it is a special case of biclique for which the two sets are one and the same. A clique of size n nodes is denoted $C(n)$ (Fig. 30B). A special case of biclique arises when one of the two sets is a singleton node - then it is termed a star (Fig. 30C). A star between a node and a set of n nodes is denoted $C(1,n)$.

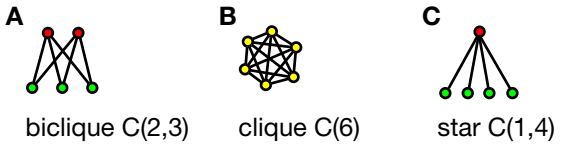


Fig. 30 **Biclique, clique and star. (A)** Biclique or *complete bipartite sub-graph*. **(B)** A clique or *complete graph*. **(C)** A star is a special case of biclique in which one of the two sets is a singleton node.

Abundance of cliques and bicliques in protein interaction networks. The abundance of cliques in protein interaction networks is the direct consequence of the existence of highly inter-connected cores within complexes (Gavin et al. 2006). The relative abundance of bicliques in protein interaction networks has been shown repeatedly – usually explained by domain mediated binding (Morrison et al. 2006; Thomas et al. 2003; Li et al. 2006b). This has in turn led to many approaches for predicting protein interactions from domain interactions (Kim et al. 2002; Deng et al. 2002; Ng et al. 2003; Nye et al. 2005; Liu et al. 2005; Rhodes et al. 2005; Patil and Nakamura 2005; Riley et al. 2005; Guimaraes et al. 2006; Jothi et al. 2006; Nye et al. 2006). Fig. 31 shows the example of a high-confidence protein interaction networks within the cytokinin signaling pathway in which bicliques are found abundantly.

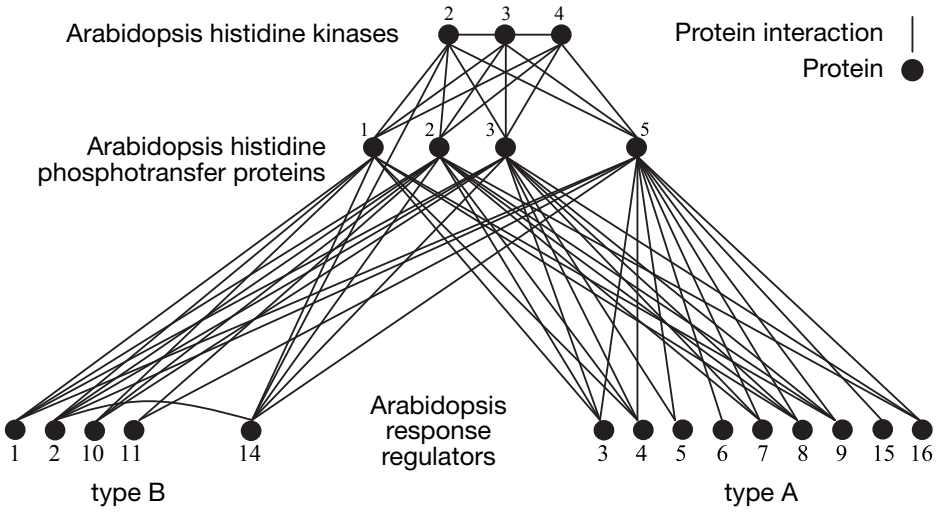


Fig. 31 **Example of bicliques in the cytokinin signaling protein interaction network of Arabidopsis thaliana.** In Arabidopsis thaliana three histidine kinases located on the membrane act as cytokinin hormone sensors. These sensors trigger a signal that is transmitted from the membrane to nucleus via phosphotransfer proteins and response regulators. The above high-quality network is the result of a small-scale Y2H screen 90% verified by *in vitro* co-affinity purification. Note the abundance of bicliques in the network. Figure and text adapted from Dortay et al. (2006).

Complexity of clique and biclique finding problems. Finding cliques or bicliques graph coverings and partitions is a challenging problem. For example finding the minimal partition of a graph into cliques is known to be NP-hard (Duh and Fürer 1997), and finding the minimal biclique partition is NP-complete (Kratzke et al. 1988). Peeters (2003) proved that the maximum edge biclique enumeration in bipartite

graphs is NP-complete. However, other problems such as enumerating all maximal bicliques of a graph can be done in polynomial time using a consensus method as shown by Alexe et al. (2004). It is also possible to determine if a graph has a biclique partition of at most k bicliques in polynomial time, as shown by Fleischner et al. (2009).

2.3.3 Network modules and graph clustering

Network modules

As defined by Hartwell et al. (1999): "a module is a discrete entity whose function is separable from those of other modules". The property of separability is best demonstrated when modules are *reused* in different complexes. One of the first evidences for modularity came from the analysis that accompanied the AP/MS interactome screen from Gavin et al. (2006). As shown in Fig. 32, Gavin et al. demonstrated in Yeast that protein complexes are hierarchically organized around cores with attachments and recurrent modules. These network modules were found by hierarchical clustering of a socio-affinity matrix (Gavin et al. 2006). Similar results were obtained by Krogan et al. (2006) and Collins et al. (2007b).

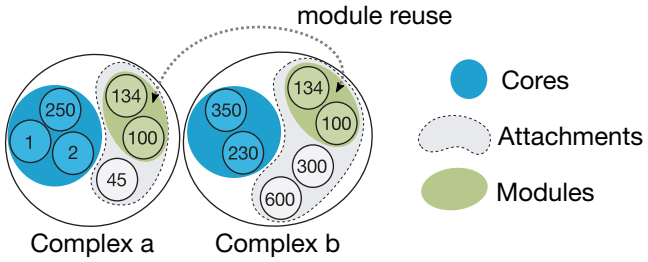


Fig. 32 **Cores, attachments and modules.** Gavin et al. proposed and verified the following framework for modeling the architecture of complexes. Complex cores are stable groups of proteins that are always found within a complex. Attachments are variable parts that may or not be present. Modules are attachments that are found in several different complexes. Figure adapted from Gavin et al. (2006).

Dense clusters for functional module and protein complex detection. A dense cluster in a protein interaction network can be identified as a group of highly interconnected proteins with few interactions to other groups – more intra- than interconnections (Bu et al. 2003; Georgii et al. 2009).

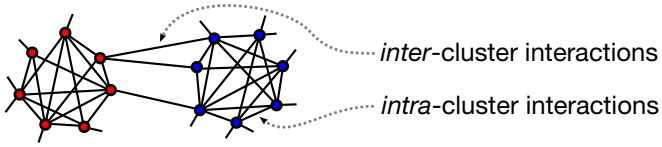


Fig. 33 **Dense clusters in networks.** Dense clusters are groups of cohesively interacting proteins with more *intra*-cluster interactions than *inter*-cluster interactions.

The notion of dense cluster in networks is an intuitive concept and has been intensively studied for the detection of complexes from protein interaction data (Spirin and

Mirny 2003; Palla et al. 2005; Hwang et al. 2006; Cui et al. 2008; Bu et al. 2003; Dunn et al. 2005; King et al. 2004; Pereira-Leal et al. 2004; Georgii et al. 2009). In particular, it is used in algorithms for complex detection such as socio-affinity score (Gavin et al. 2006), bootstrap confidence scores (Friedel et al. 2009), purification enrichment scores (Collins et al. 2007b), hypergeometrical distribution based scores (Hart et al. 2007), and dice coefficients (Zhang et al. 2008). These algorithms rely on popular graph clustering algorithms such as the Restricted Neighborhood Search Clustering (RNSC) algorithm (King et al. 2004), the MCODE algorithm (Bader and Hogue 2003), or MCL (Dongen 2000). Brohée and van Helden (2006) compared these algorithms and conclude that MCL is superior. Many of these algorithms have been made available in a framework developed by Krumsiek et al. (2008). The working hypothesis is that dense clusters correspond to functional modules which in turn correspond to protein complexes. This hypothesis is supported by the assortativity of protein interaction networks (as discussed previously).

Dense clusters are not necessarily network modules. There is an unfortunate confusion in the literature between the notion of dense cluster and network module. As argued by Wang and Zhang (2007) and Pinkert et al. (2010), the evidence provided by Gavin et al. (2006) shows that reused modules taking part in several complexes cannot be solely identified as groups of cohesively interacting proteins. Dense clusters are not necessarily modules. While dense clusters are useful for finding cores of complexes, they are not sufficient for finding modules – when defined as *reused* entities within the network. Another definition of *module* takes a higher-level approach and relies on the functional homogeneity of proteins that are hypothesized to function as an independent entity – usually a protein complex (Spirin and Mirny 2003; Cui et al. 2008). In that case *modularity* implicitly refers to the assumed modularity of molecular machines. In the following we show that dense clusters in graphs are not the only type of clusters that can be identified.

Two types of clusters in networks. Two types of clusters can be identified in networks: dense clusters and neighborhood-similar clusters (Fig. 34). This can be understood from the structure of the bi-clustered adjacency matrix: the edges characterizing dense clusters correspond to squares in diagonal region (Fig. 34A and C) while the edges characterizing neighborhood-similar clusters correspond to pairs of rectangles away from the diagonal (Fig. 34B and C). Another way to understand this point is to consider that neighborhood-similar clusters of nodes can be indirectly defined by clusters of edges or *link-communities* (Ahn et al. 2010). In Fig. 34B the cluster of nodes can be defined by a cluster of edges (biclique in blue) despite the absence of a direct connections between nodes. Currently, as observed by Ahn et al. (2010), most graph clustering algorithms are designed to detect clusters of nodes – typically by identifying groups of densely interconnected nodes.

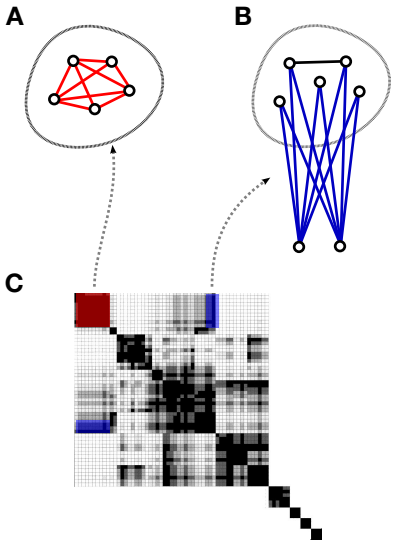


Fig. 34 **High density versus neighborhood similar clusters. (A)** High-density clusters are regions of the graph in which nodes are adjacent to each other. **(B)** In contrast, the nodes of highly neighborhood-similar clusters are not necessarily adjacent to each other but are instead adjacent to common neighbors. Note that high-density clusters are also neighborhood-similar clusters, but neighborhood-similar clusters are *not necessarily* high-density clusters. **(C)** Seen from the perspective of the bi-clustered adjacency matrix, both types of node clusters can be defined by clusters of edges – either squares along the diagonal for dense clusters and symmetric pairs of rectangles for neighborhood-similar clusters.

In the following we review existing graph clustering algorithms as essential tools for identifying modules in networks.

Graph clustering algorithms

Identification of modules used in complex detection algorithms rely on graph clustering algorithms. As shown in Table 2 numerous approaches have been developed. Graph clustering algorithms can be first classified in three main families depending on the kind of cluster returned: dense clusters, neighborhood similar clusters, or both. Most graph clustering algorithm aim at detecting dense clusters. The ideas used to tackle that problem come from diverse fields: statistical physics, graph theory, optimization, and algebraic graph theory. In the following we review these approaches.

Table 2 **Graph clustering algorithms.** The algorithms are listed in chronological order of publication. Several features can help characterize these algorithms. First the computational complexity, which is typically polynomial. Second, how much parameter tuning is required? This is important because it may strongly influence results as observed by Brohée and van Helden (2006). Third, can the algorithm make use of edge weights? Forth, the clustering procedure may be agglomerative, divisive, or return a fixed number of clusters. Fifth, the resulting clusters can be organized as a partition, as a hierarchy, or as an arbitrary covering of the nodes. Finally, which kind of clusters are identified: dense clusters or neighborhood-similar clusters? Most graph clustering algorithms return dense clusters (Ahn et al. 2010). For this table we relied on information provided by the respective authors as well as from the review by Andreopoulos et al. (2009) and the comparison by Brohée and van Helden (2006).

algorithm	name	complexity	tuning	weights	A/D/F	H/P/C	D/N/DN
Gallai (1967)	Modular decomposition	$O(n+e)$			A	H	D&N
Bron and Kerbosch (1973)	Bron-Kerbosch	$O\left(3^{\frac{n}{3}}\right)$			A	C	D
Blatt et al. (1996)	SPC	$O(ne)$	T	W	F	P	D
Matsuda et al. (1999)	Ncut-KL	$O(n+e)$	T	W	D	P	D
Edachery et al. (1999)	Kcliques	$O\left(n^3\right)$	T		A	P	D
Dongen (2000)	MCL	$O\left(n^3\right)$	T	W	D	P	D
Bolten et al. (2001)	SCC	$O(n+e)$		W	A	P	D
Pipenbacher et al. (2002)	ProClust	$O(n+e)$		W	A	P	D
Bader and Hogue (2003)	MCODE	$O\left(n^3e\right)$	T	W	A	P	D
King et al. (2004)	RNSC	$O\left(enc+c^2\right)$	T	W	F	P	D
White and Smyth (2005)	Spectral	$O(ne)$	T	W	F	P	D
Kim and Lee (2006)	BAG	$O(e)$			D	P	D
Ding et al. (2006)	Bi-clustering	$O(e)$	T	W	A	H	D&N
Andreopoulos et al. (2007)	MULIC	$O\left(n^2\right)$			A	H	D&N
Ahn et al. (2010)	Link Clustering	$O\left(n^2\right)$			A	C	D&N

$O(...n,e,c...)$	Big O notation with number of nodes(n), edges(e), and cliques(c) in the graph
T	Parameter tunning needed?
W	Edge weights considered?
A/D/F	Agglomerative / Divisive / Fixed number of clusters
H/P/C	Hierarchical / Partitive / Covering
D/N/D&N	Dense clusters / Neighborhood-similar clusters / both

Graph clustering algorithms for dense cluster identification:

• **Stochastic approaches.** An early trend in graph clustering came from statistical physics and is exemplified by *Super Paramagnetic Clustering* (SPC). Super Paramagnetic Clustering was first introduced by Blatt et al. (1996) and applied to gene co-expression networks by Getz et al. (2000), and to protein sequence homology networks by Tetko et al. (2005). Super Paramagnetic Clustering models the graph as an inhomogeneous ferromagnetic spin field. Nodes in the network are spins and edges are short range interactions between spins. Clusters are then identified as domains of spin-spin correlation. The *Markov CLuster* (MCL) algorithm introduced by Dongen (2000) considers the graph as a stochastic matrix. Inflation and expansion transformations are applied alternatively until a fixed point is reached. Clusters are the connected components in the resulting fixed point matrix.

• **Graph theoretic approaches.** The algorithm for enumerating all maximal cliques in an undirected graph is the typical example in this category. It has however a high computational cost with a time complexity of $O\left(3^{\frac{n}{3}}\right)$. Other algorithms rely on heuristics for identifying cliques and are polynomial in the number of nodes and edges.

For example, Matsuda et al. (1999) developed a clustering algorithm based on the identification of *p-quasi complete graph*. Similarly, Edachery et al. (1999) proposed a graph clustering algorithm that defines clusters as *distance-k cliques* in which the shortest path between two nodes is of at most length k. Recently, Kim and Lee (2006) proposed the BAG algorithm as a graph theoretic algorithm based on biconnected components and articulation points. In contrast, the *Molecular Complex Detection* (MCODE) algorithm by Bader and Hogue (2003) identifies densely connected regions in graphs with local neighborhood density vertex weighting. The measure of density is based on the local clustering coefficient to measure what the authors refer as the *cliquishness* around a node. It was applied to the detection of molecular complexes in large protein-protein interaction networks. For these reasons, empirical evaluations of time complexities are often more reliable.

• **Modular decomposition.** Another graph theoretic approach is modular decomposition – a well-characterized recursive partition of a graph into a hierarchy of modules (Gallai 1967). It has been discovered independently by researchers in game theory, graph theory, network theory (Bioch 2005). Modular decomposition has been applied for the design of efficient algorithms for graph pattern-matching (Habib et al. 2000), and graph drawing (Papadopoulos and Voglis 2006). In computational biology, it has been used to study the organization of sub-units within molecular complexes (Gagneur et al. 2004). Modular decomposition algorithms have been optimized to the point that the modular decomposition of a graph can be computed in linear time in the number of nodes n and edges e: $O(n+e)$ as shown by Tedder et al. (2008), McConnell and de Montgolfier (2005), and Cournier and Habib (1994).

• **Optimization-based approaches.** The *Restricted Neighborhood Search Clustering Algorithm* (RNSC) by King et al. (2004) is a clustering algorithm on graphs that relies on the optimization of a cost function defined on the set of possible clusterings. The authors search for optimal clustering by applying randomized local search techniques, and claim to obtain significantly lower-cost clusterings than other approaches.

• **Algebraic approaches.** Spectral clustering was first described by Ng et al. (2002). It may be applied to any data that can be represented as similarity graphs. It is based on the calculation of the eigenvectors and eigenvalues of the graph Laplacian matrix. It has been applied for the detection of communities in graphs (White and Smyth 2005).

Graph clustering algorithms for neighborhood-similar cluster detection

Other graph clustering algorithms can detect neighborhood-similar clusters as well as *edge clusters* or *link communities* as termed by Ahn et al. (2010). The following algorithms have been proposed: Bi-clustering (Hartigan 1972; Ding et al. 2006), Mulic (Andreopoulos et al. 2007), Pinkert et al. (2010) algorithm, and Link Clustering

(Ahn et al. 2010) (listed in Table 2). At the core of these algorithms is a measure of neighborhood similarity between nodes.

Neighborhood similarity. In the following u and v are two nodes in a graph G, and that $N(u)$ and $N(v)$ are the sets of neighbors of u and v, respectively. A measure of neighborhood similarity can be derived from the *Jaccard index* (Jaccard 1901). The *Jaccard neighborhood similarity* of two nodes u and v is:

$$J(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

By convention, if $|N(u) \cup N(v)| = 0$ then $J(u,v) = 0$. The differences between algorithms based on neighborhood-similarity come from different emphasis on the results. For example, the algorithms by (Pinkert et al. 2010; Ahn et al. 2010) put the emphasis on the notion of edge clusters termed for example *link communities* in Ahn et al. (2010). In the following we review agglomerative hierarchical clustering on which most neighborhood similarity based algorithms rely.

Comparison of the different graph clustering algorithms

Time complexity. As shown in Table 2 all graph clustering algorithms – except Bron and Kerbosch (1973) – are at least polynomial in the number of nodes, edges, or cliques. The complexities reported by the authors follow different conventions of what constitutes the worse case complexity. Moreover, the topology of the networks may have a very strong influence on the *effective average case complexity* – calling for a cautious comparison of the authors' reported complexities. In particular, complexities only parametrized by the number of nodes are difficult to compare because most primitive operations on graphs are defined on edges and not on nodes. An algorithm that is quadratic in n may be in fact cubic in e if for example Jaccard index calculations are considered atomic – $O(1)$ – instead of edge dependent – $O(e)$.

Parameter tuning. Given enough tunable parameters, a conveniently designed algorithm can be made to return any desired output. Therefore, it is highly desirable for algorithms to have as little parameters as possible. This spares the user the difficulty of finding the right combination of values – which may be a difficult computational problem. Hence, graph clustering algorithms should not need any parameter apart from the graph itself. This requirement is especially reasonable for algorithms that exhibit high parameter sensitivity. Among the algorithms that require parameter tunning (see Table 2) Brohée and van Helden (2006) noted that RNSC is more robust to parameter variation than MCL, MCODE, and SPC.

Hierarchies versus partition or coverings. One such parameter is the number of clusters returned. Some algorithms return a predefined number of clusters (SPC, RNSC, Spectral). In practice it is difficult to require that the algorithm discover the right number of clusters because it is usually application dependent (Tan et al. 2005). A solution to this problem is to return a hierarchy of clusters, providing a multi-scale

view into the graph. Hierarchies are returned by several algorithms such as modular decomposition (Gallai 1967), bi-clustering (Ding et al. 2006), and MULIC (Andreopoulos et al. 2007). An alternative to returning hierarchies and partitions is to return a collection of covering clusters. In this category we have the algorithm for enumerating all maximal cliques by Bron and Kerbosch (1973) and Link Clustering by Ahn et al. (2010). The advantage of overlapping clusters is application dependent. For example, in complex detection several different complexes may share modules.

Edge weights. Most graph clustering algorithms can make use of edge weights to obtain more accurate results. The hypothesis is that edges that have a very low weight have the least influence on the final result – removing lowest weight edges leaves the clusters invariant.

Dense clusters versus neighborhood similarity clusters. A more fundamental difference between graph clustering algorithms is the kind of clusters returned. We mentioned already that most graph clustering algorithms detect dense clusters. Only four can detect neighborhood similar clusters: modular decomposition, bi-clustering, MULIC, and Link Clustering. Modular decomposition often fails to return any cluster at all which explains its scarce use and poor implementation availability (Gagneur et al. 2004; Papadopoulos and Voglis 2006).

2.3.4 Agglomerative hierarchical clustering

Hierarchical clustering was first introduced by Sneath (2005). Its simplicity makes it probably the most elegant form of clustering. While this technique was known for a long time (Fitch and Margoliash 1967; Cavalli-Sforza and Edwards 1967), one of the first uses of hierarchical clustering in bioinformatics was for the multiple sequence alignment of sequences (Corpet 1988). Ten years later, Eisen's seminal paper on cluster analysis of gene expression data (Eisen et al. 1998) established the technique. In the context of network analysis it plays a role in neighborhood similarity based clustering algorithms as well as in bi-clustering schemes. In the following we describe the procedure:

Hierarchical clustering algorithm. Assume that we want to cluster n objects O_i. Further assume that we have a similarity measure $s\left(\{O_i,..,O_j\},\{O_k,...,O_l\}\right)$ defined on clusters of objects. This similarity measure captures how *similar* two clusters of objects are, and is often defined by a similarity measure between individual objects $s\left(O_i,O_j\right)$. The *Hierarchical Agglomerative Clustering Algorithm* proceeds as follows:

I Compute the similarity matrix containing the similarity between each pair of objects. Initially, treat each object as a singleton cluster.

II Find the most similar pair of clusters using the similarity matrix. Merge these two clusters into one cluster and update the similarity matrix by computing the similarity between the new merged cluster and existing clusters.

III Stop the algorithm if all objects are in one cluster. Otherwise, go to step II.

The result is a hierarchy of clusters.

Linkage methods. The definition of a similarity measure between clusters – or *linkage method* – is what distinguishes the different flavors of agglomerative hierarchical clustering. There are mainly three linkage methods mentioned throughout the literature: Single linkage method (*minimum*) (Johnson 1967), Complete linkage method (*maximum*) (Johnson 1967), Average linkage method (*unweighted pair-wise group mean average linkage*) (Sneath 2005).

Complete, single and average linkage methods. Given two clusters U and W, and a similarity measure $s(u,v)$ between nodes u and v, the three linkage methods are formally defined as follows:

- Maximum or complete linkage clustering:

$$s(U,W) = \max\{s(u,v) \,|\, u \in U \text{ and } v \in W\}$$

- Minimum or single-linkage clustering:

$$s(U,W) = \min\{s(u,v) \,|\, u \in U \text{ and } v \in W\}$$

- Mean or average linkage clustering:

$$s(U,W) = \frac{1}{|U||W|} \sum_{u \in U} \sum_{v \in W} s(u,v)$$

These linkage methods rely on the definition of the similarity measure $s(u,v)$ or equivalently dissimilarity $(1 - s(u,v))$ between objects (Fig. 35).

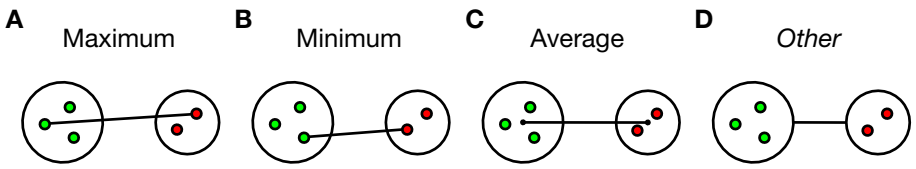


Fig. 35 **Maximum, minimum, average and *other* linkage methods. (A)** Maximum or complete linkage clustering. **(B)** Minimum or single-linkage clustering. **(C)** Mean or average linkage clustering (or unweighted pair group method with arithmetic mean – UPGMA). **(D)** Other linkage methods directly defined on clusters. Examples are Ward's minimal variance method (Ward Jr 1963) and the Hausdorf linkage method by Basalto et al. (2008).

Linkage methods characteristics. Complete (or maximal) linkage produces compact clusters (Jain et al. 1999), whereas single linkage produces elongated clusters (Nagy 1968). This is referred as the *chaining effect* of single linkage (Jain et al. 1999). Average linkage is a trade-off between single and complete linkage which main advantage is robustness to outliers.

Other linkage methods. Other possibilities exist for defining a similarity or dissimilarity between two clusters (Fig. 35D). For example, Ward's minimal variance method computes the variance of the union of the two sets (Ward Jr 1963). Another example is the Hausdorf linkage method that relies on the Hausdorf distance between point sets (Basalto et al. 2008).

Complexity of hierarchical clustering. In the absence of optimizations, the time complexity of average and complete linkage hierarchical clustering is $O\left(n^2 log(n)\right)$ computations of similarities between individual objects, where n is the number of objects to cluster. In contrast, single linkage clustering is easier with a complexity of $O\left(n^2\right)$ (Murtagh 1983).

2.4 Conclusion

In this chapter we reviewed the challenges facing genome-wide protein interaction mapping and discussed the advantages and limitations of high-throughput Y2H, AP/MS and PCA screens. Two open problems were highlighted. First, the important and controversial problem of data quality assessment. We showed that no consensus exists for experimental and computational methods for quality assessment. Second, assuming that the networks are reasonably represent the underlying molecular systems, how to best analyze these networks? We saw that visualization produces 'fur balls' and clustering algorithms abstract most details about individual interactions. Is there a way to represent networks that both identifies modules and preserves all information about the subtle connection patterns within and between modules?

In the following chapter 3 we will introduce power graph analysis as a novel representation for protein interaction networks. In chapter 4 we will show how this new representation can be interpreted as a compression algorithm for graphs and how it can be used for quality assessment. Finally in chapters 5 and 6 we will give a series of applications of power graph analysis to text-mined networks, stem cell research, disease, and material biocompatibility.

Chapter 3

Unraveling Protein Networks with Power Graph Analysis

3.1 Introduction

Networks play a crucial role in computational biology. Yet, their analysis and representation is still an open problem.

Power graph analysis is a lossless transformation of biological networks into a compact, less redundant representation, exploiting the abundance of cliques and bicliques as elementary topological motifs. Power graphs compress up to 85% of the edges in protein interaction networks and are applicable to all types of networks such as protein interactions, regulatory or homology networks. In this chapter we demonstrate with five examples the advantages of power graphs over traditional network representations. Investigating protein-protein interaction networks, we show how the catalytic subunits of the Casein Kinase II Complex are distinguishable from the regulatory subunits, how interaction profiles and sequence phylogeny of SH3 domains correlate, and how false positive interactions among high-throughput interactions are spotted. We apply Power Graph Analysis to large-scale protein interaction networks and show that they are significantly enriched in cliques and bicliques which are themselves enriched in Gene Ontology terms and InterPro domains and motifs. Additionally, we demonstrate the generality of power graph analysis by applying it to two other kinds of networks. We show how power graphs induce a clustering of both transcription factors and target genes in bipartite transcription networks, and how the erosion of a phosphatase domain in type 22 non-receptor tyrosine phosphatases is detected.

We show how to compute minimal power graphs by using neighborhood similarity clustering. We evaluate the algorithm on a benchmark of manually curated graphs and find that the algorithm produces the correct power graph in 86% of the cases. Moreover, we evaluate the algorithm's compression characteristics on two classes of random graphs: Erdös-Renyi-model (ER model) and Barabási-Albert-model (BA model). We show that in the general case – and independently of the model – the compression rate and the edge density are in affine relationship. In addition, we show that the algorithm's performance gracefully degrades as noise is added to protein-protein interaction graphs. We also empirically investigate the time complexity of the

algorithm and find that a tight lower-bound of the computation time follows a sub-quadratic power law in the number of edges.

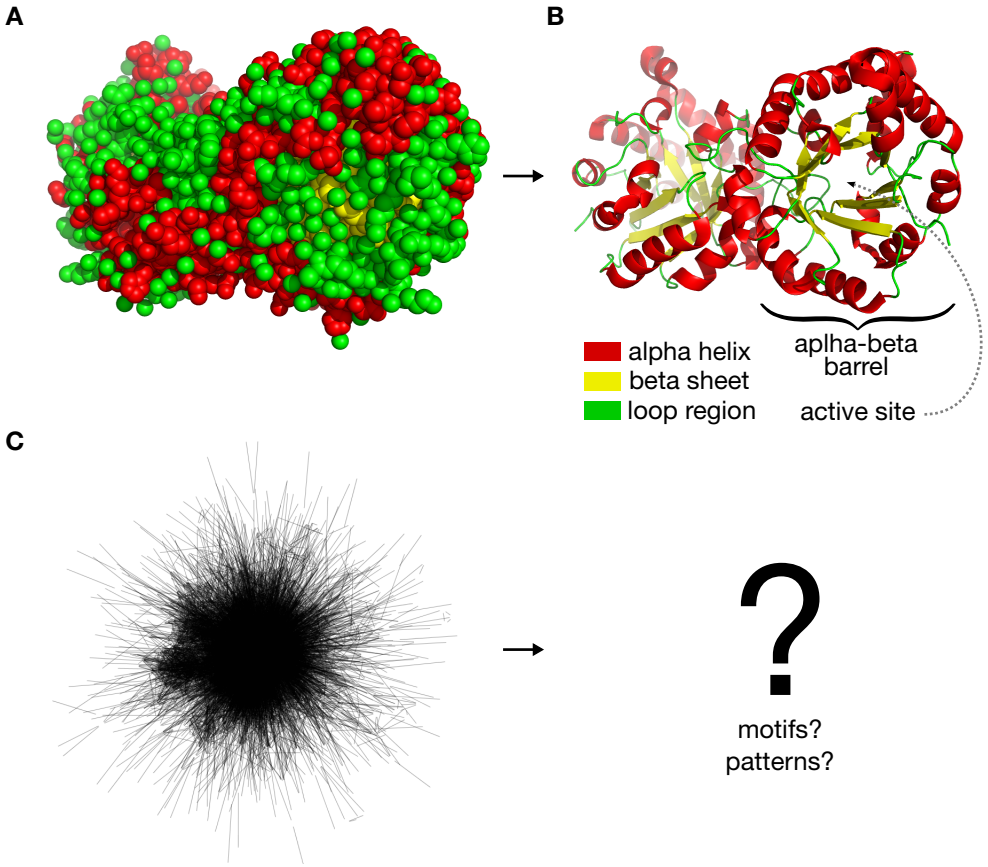


Fig. 36 ***'To comprehend is to compress!'*** **(A)** Atomic structure of an enzyme (triose-phosphate isomerase). Each atom is represented by a sphere (Kursula et al. 2004). Little insight onto the structure of the protein can be gathered from this representation. **(B)** Cartoon representation of the enzyme's structure highlighting alpha-helices (red) and beta sheets (green). The alpha-beta barrel where the active site is located is recognizable in the abstracted representation. **(C)** Yeast protein interaction network by Gavin et al. (2006) represented as a "fur ball". Which motifs and patterns can help to unravel protein interaction networks?

From motifs and patterns to understanding

Gregory Chaitin wrote: '*To comprehend is to compress!*' (Chaitin 2007). The basis for compressibility is often a bias in the statistics such as the over or under-representation of motifs and patterns (Salomon et al. 2007). We illustrate this principle in Fig. 36A with an example from structural biology. Knowing the location of each and every atom in a protein does not capture the structure's essence – as it resides at a higher level of abstraction. One of the key observations of Linus Pauling, Robert Corey, and Herman Branson in 1951 was the identification of two secondary structures in proteins: the alpha-helix and the beta-sheet (Eisenberg 2003). These two motifs are found in the backbone of most proteins together with loop regions. As shown in Fig. 36B the important features of protein structures such as active sites become apparent

when abstracting the molecule as the assembly of alpha-helices, beta-sheets, and connecting loops.

Motifs and patterns in complex networks. As shown in Fig. 36C, when visualized as graphs, protein interaction networks often appear as uninformative '*fur balls*'. Traditional graph representations show every edge and thus few insights can be garnered about the general architecture and organization of the network. Which recurrent motifs and patterns can help to better understand protein interaction networks and complex biological networks in general?

Bicliques, cliques and stars. How does the underlying biology manifest itself in the networks? Fig. 37 illustrates three recurrent motifs that have been reported in the literature.

Star motifs. The first motif is the star, representing a hub protein, frequently present in scale-free biological networks (Li et al. 2006a). The abundance of hub proteins may be explained by evolutionary models based on gene duplication, divergence (Taylor and Raes 2004) and preferential attachment (Barabasi and Albert 1999).

Clique motifs. The second motif is the clique, also referred to as complete graph: a set of completely interconnected proteins. In the core of molecular complexes, where the distinction between direct and indirect physical interactions is often blurred, protein interactions are observed to organize into cliques and bicliques (Gavin et al. 2006). Indeed, the completion of quasi-cliques and quasi-bicliques has been shown to successfully predict missing interactions between proteins (Bu et al. 2003). Cliques are a special case of reflexive bicliques.

Biclique motifs. The third motif is the biclique, also referred to as complete bipartite graph: all proteins in one group interact with all proteins in another group. Domain interactions have been reported to induce bicliques (Li et al. 2006b). Models of protein interaction networks based on interacting domains have been proposed in which complementary domains are shown to induce bipartite structures (Morrison et al. 2006; Thomas et al. 2003). Similarly, bicliques detected in protein interaction networks were used to discover motif pairs at interaction sites (Li et al. 2006b). In general, domain and protein interactions have been shown in many studies to be correlated. Domain bindings can be used to predict protein interactions, and conversely, protein interactions can be used to predict domain interactions (Kim et al. 2002; Deng et al. 2002; Ng et al. 2003; Nye et al. 2005; Liu et al. 2005; Rhodes et al. 2005; Patil and Nakamura 2005; Riley et al. 2005; Guimaraes et al. 2006; Jothi et al. 2006; Nye et al. 2006).

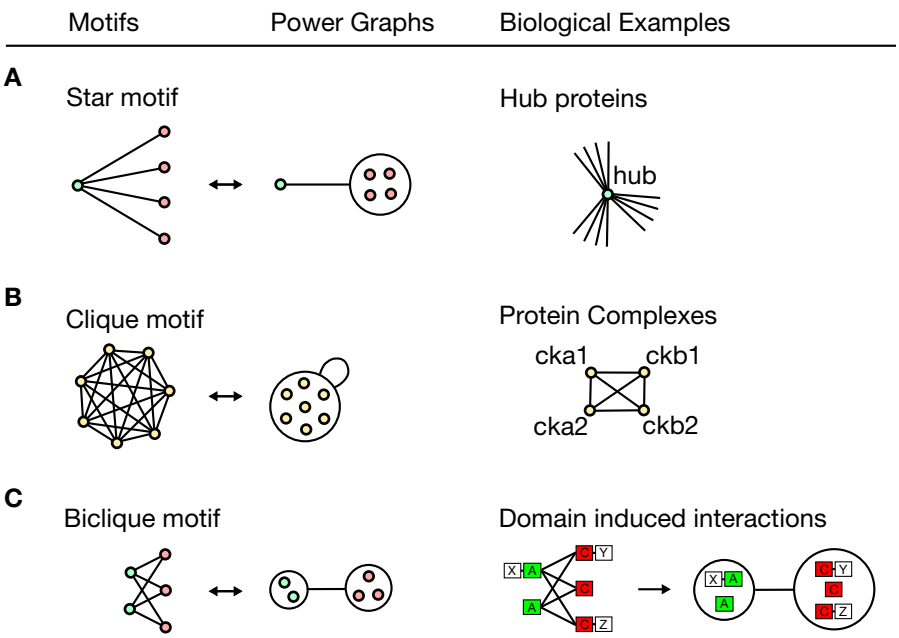


Fig. 37 **The three basic motifs in protein interaction networks: star, biclique, and clique.** Power nodes are sets of nodes, and power edges connect power nodes. A power edge between two power nodes signifies that all nodes of the first set are connected to all nodes of the second set. Note that nodes within a power node are not necessarily connected to each other. **(A)** Stars often occur because of hub proteins or when affinity purification complexes are interpreted using the spoke model (Gavin et al. 2006; Li et al. 2006a). **(B)** At the core of molecular complexes, protein interactions organize into cliques (Bu et al. 2003; Gavin et al. 2006). **(C)** Bicliques often arise because of domain-domain or domain-motif interactions inducing protein interactions (Morrison et al. 2006).

3.2 Unraveling Protein interaction networks

In the following we will show with three examples how power graph analysis groups proteins into biologically relevant modules. We first examine a small-scale network of SH3 domains and binding peptides that illustrates the modular binding of protein domains (Landgraf et al. 2004). The two following examples show how power graph analysis can assist in unraveling the internal structure of molecular complexes in two AP/MS networks from Gavin et al. (2006) and Krogan et al. (2006).

3.2.1 Example 1 – SH3 domain binding peptides

Before looking at examples of power graph analysis applied to large scale protein interaction networks, let us first examine a smaller network. Landgraf et al. (2004) used a combination of phage display and SPOT synthesis to discover peptides in the Yeast proteome binding to eight SH3 domains. We will show how power graph analysis reveals a relationship between neighborhood similarity in the network and sequence similarity of SH3 domains their binding motifs.

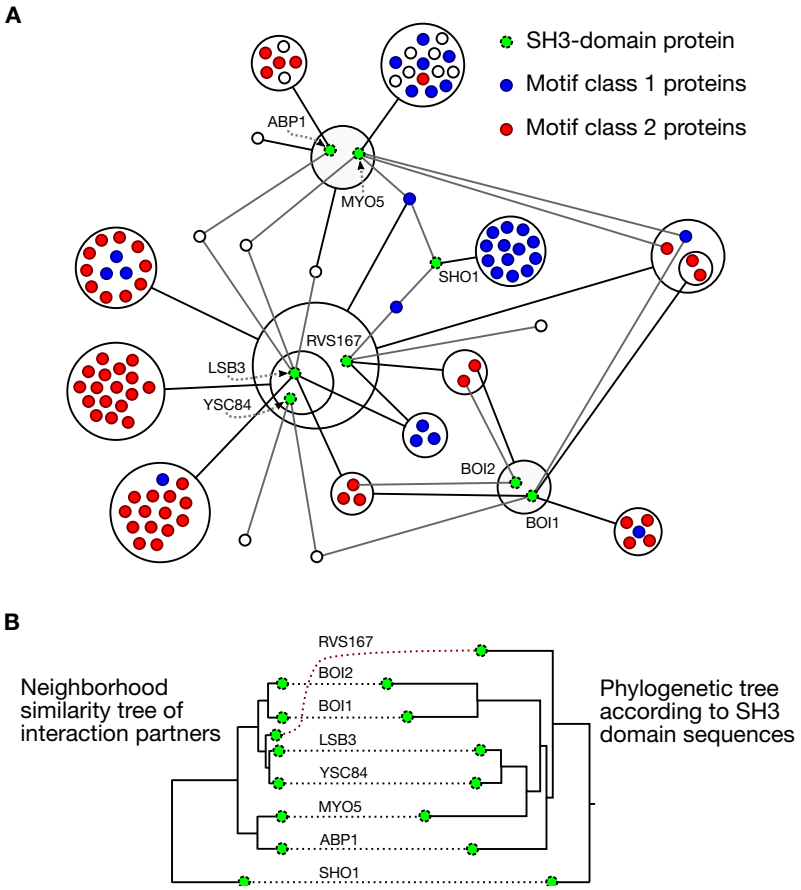


Fig. 38 **Interactions of SH3 carrying proteins with small peptides. (A)** Protein interaction network showing the 105 interaction partners of the SH3 domain carrying proteins: SHO1, ABP1, MYO5, BOI1, BOI2, RVS167, YHR016C and YFR024 (shown in green). The underlying network consists of 182 interactions represented here as 36 power edges – a reduction of 80%. Class 1 motif (RxxPxxP) proteins are shown in blue. Class 2 motif (PxxPxR) proteins are shown in red (Landgraf et al. 2004). Power graphs group proteins having similar binding motifs together. **(B)** Phylogeny and interaction profiles. Comparison of the phylogenetic tree of the SH3 domains sequences with the neighborhood similarity tree of interaction partners. The neighborhood similarity implied by the power graph reflects the sequence similarity of the SH3 domains.

SH3 domain to peptide network. Fig. 38A shows a power graph representation of the interaction network of SH3 domain carrying proteins (SHO1, ABP1, MYO5, BOI1, BOI2, RVS167, YHR016C and YFR024). The power graph representation achieves a reduction in complexity by diminishing the number of edges necessary for the representation by 80%. The proteins RVS167, YHR016C and YFR024 are in a power node together, showing the similarity of their neighborhoods. YHR016C and YFR024 are even more similar and have a power node of their own. Proteins that carry the SH3 domain are filled in gray. Power nodes of proteins bound by SH3 carrying proteins are enriched either in class 1 (RxxPxxP) or class 2 (PxxPxR) motifs (Landgraf et al. 2004).

Comparing sequence similarity and neighborhood similarity. We investigated how the interaction profiles of these eight SH3 carrying proteins relate to the domain

sequences. Fig. 38B shows a strong correlation between the phylogenetic tree of the SH3 domain sequences and the neighborhood similarity tree of interaction partners.

SH3 binding sequences correlates to peptide binding interaction profiles. The pair of SH3-carrying proteins YHR016C/YFR024, grouped in one power node in Fig. 38A, are also close in the neighborhood similarity tree. They are also close in the phylogenetic tree. The same holds for the pair BOI1/BOI2. However we also notice two discrepancies. Proteins ABP1 and MYO5 are grouped together in the neighborhood similarity tree – whereas they are not in the phylogenetic tree. Protein RVS167 has different placements in the two trees – RVS167 and YHR016C/YFR024 have similar interaction partners but dissimilar sequences.

3.2.2 Example 2 – Casein Kinase II Complex

A survey of the Yeast proteome by Gavin et al. (2006) showed the high modularity of Yeast molecular complexes. In the following we show how catalytic and regulatory subunits of the casein kinase II complex can be distinguished using power graph analysis. Fig. 39 shows the casein kinase II complex and its neighboring complexes.

Catalytic and regulatory subunits. The Casein kinase II has been implicated in cell cycle control, DNA repair, regulation of the circadian rhythm and other cellular processes. It is a tetramer of two catalytic alpha subunits, CKA1 & CKA2, and two regulatory beta subunits, CKB1 and CKB2. Remarkably, the power graph representation conveys immediately the difference between the alpha and beta pairs of subunits: the two alpha subunits are grouped together by one power node, and the beta subunits are grouped together by another power node. The reason for this is that the two alpha subunits have almost identical neighbors, which are in turn different from the neighbors shared by the beta subunits. The beta subunits are connected to the eIF3 sub-complex (NIP1, RPG1, PRT1) known to stimulate the binding of mRNA to ribosomes. The beta subunits are also connected – through the intermediary protein UTP22 – to a power node consisting of proteins ROK1, RRP7 and CMS1 that do not correspond to a known complex but that are all relevant to RNA processing, possibly a small complex. In contrast, the alpha subunits do not interact with these two groups, but instead with decarboxylase CAB3.

Neighboring complexes. Other complexes are visible in the power graph representation. For example the proteins POB3 and SPT16 are grouped together in one power node. They form a complex known as the heterodimeric FACT complex SPT16/POB3, involved in transcription elongation on chromatin templates. It is known that the FACT complex is activated by the Casein Kinase II Complex (Keller et al. 2001). Finally a group of two power nodes linked by a power edge, all of them interacting with the protein PAF1, form the PAF1 complex - a complex that associates with RNA polymerase II (Mason and Struhl 2003).

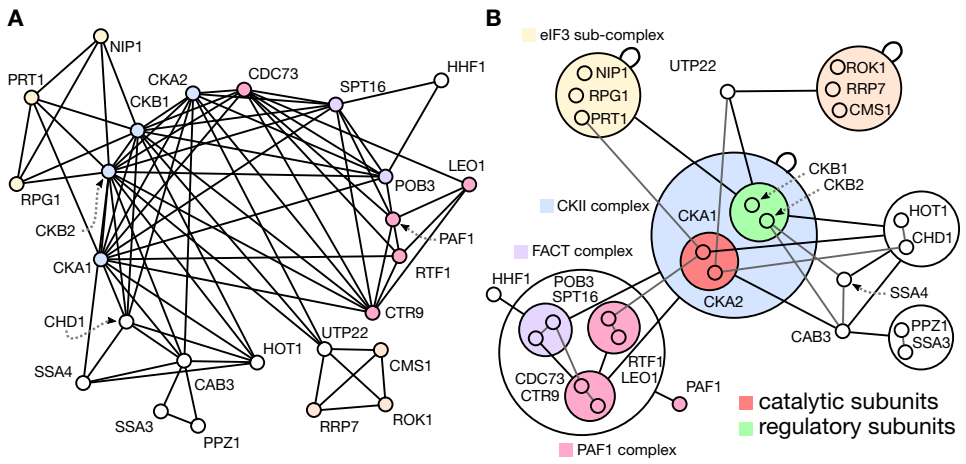


Fig. 39 **Casein Kinase II Complex and its neighboring complexes in Gavin et al. (2006). (A)** Network showing the four subunits – CKA1, CKA2, CKB1, and CKB2 – of the Casein Kinase II Complex and neighboring proteins from the FACT complex, sub-complex NIP1-RPG-PRT1 of eIF3, and PAF1 complex. **(B)** Corresponding power graph consisting of 30 power edges instead of 80 edges, thus an edge reduction of 62%. This simplification of the representation makes the separation of the regulatory subunits (CKB1, CKB2) from the catalytic subunits (CKA1, CKA2) immediately apparent without loss of information on individual interactions.

3.2.3 Example 3 – Untangling the nucleosome

Similarly to the survey of the Yeast proteome by Gavin et al. (2006), Krogan et al. (2006) investigated protein interactions using tandem affinity purification (TAP). Fig. 40A shows a subgraph of proteins surrounding the H1, H2A, H2B, H3 and H4 histone proteins. These proteins form the nucleosome, an octameric complex responsible for the packing of DNA into chromosomes.

Gene duplication – histone subunits subtypes. Interestingly, the subunits H2A, H2B, H3, and H4 come in pairs: HTA1/HTA2, HTB1/HTB2, HHT1/HHT2, and HHF1/-HHF2. This is an example of gene duplication (Taylor and Raes 2004), inducing a complete bipartite subgraph (biclique) of interactions between proteins expressing duplicated genes. In Yeast, HTA1, HTA2, HTB1, and HTB2 are nearly identical, with two and respectively four amino acids differing. HHF1 and HHF2 are identical proteins coded by different genes. Interacting with histones is the ORC Complex (Origin Recognition Complex), responsible for marking origin regions prior to DNA replication.

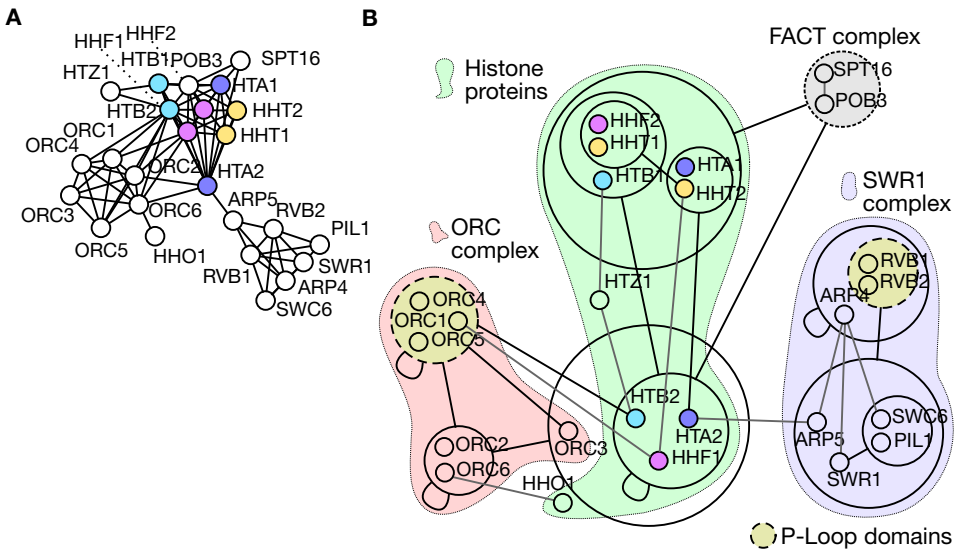


Fig. 40 **Histone protein interactions and neighboring proteins according to Krogan et al. (2006). (A)** Standard graph representation. **(B)** Power graph representation. The ORC complex is visible with a power node of proteins – ORC1/ORC4/ORC5 – carrying a nucleotide binding P-loop domain (SCOP:52540). Histones subtypes HTA1/2, HTB1/2, HHT1/2, and HHF1/2 share the same color. Histones HTA2, HTB2 and HHF1 are segregated from their twin subtypes HTA1, HTB1 and HHF2. The FACT complex SPT16/POB3 is again delineated.

Power graph of the nucleosome network. In Fig. 40B the corresponding power graph is shown. The ORC complex is a clique of six proteins which appears in the power graph representation as three power nodes linked by three power edges. One of these power nodes – ORC1/ORC4/ORC5 – interacts with HTB2 and is enriched in a specific domain: a nucleotide binding P-loop domain containing nucleotide triphosphate hydrolases. This same domain is found in the power node of proteins RVB1 and RVB2, which forms a biclique with ARP5, SWR1, PIL1, and SWC6, all related to the SWR1 complex.

Segregated histone subtypes. Surprisingly, histones HTA2, HTB2 and HHF1 are segregated from their twin subtypes HTA1, HTB1 and HHF2, as subunits ORC2 and ORC6 interact with HTA2, HTB2 and HHF1 and not with the HTA1, HTB1, and HHF2. This is in contrast to the identity/near identity of these pairs of histones. The power graph shows the separation between these two types of histones.

First hypothesis. Why have these mostly identical proteins different interaction partners? In the case of H2A histones, each subtype has been shown to be sufficient for cell viability, and no clear functional difference was reported apart from homozygous strains for *hta*1$^{-/-}$, exhibiting a slower growth (Kolodrubetz et al. 1982). Despite the near identity of these proteins, their interaction profiles are different which suggests that the interactions with ORC2 and ORC6 are false positives or false negatives - all or none of the histones interact with ORC2 and ORC6.

Second hypothesis. Yet, this hypothesis does not explain that co-regulated HTA2 and HTB2 are both seen interacting with ORC2 and ORC6, whereas the differently co-regulated HTA1 and HTB1 do not (Cherry et al. 1998). Moran et al. (1990) show that the promoter region of HTA2 and HTB2 is regulated by the amount of effective H2A+H2B expression. This mechanism is essential for ensuring a sufficient and balanced amount of histones during the S phase – when DNA replication takes place. An excess of H2A+H2B induces a 10 fold decrease in RNA production for HTA1 and HTB1. Thus, a possible explanation for not observing interactions between ORC2/ORC6 and HTA1/HTB1 is that under some circumstances – that may be triggered by the TAP methodology (the fusion of the TAP tag to the C-terminus) – the production of subtypes HTA1 is repressed. Moran et al. (1990) argue that the same feed-back regulation takes place for HTB1 as well as for all variants of HHT and HHF (Moran et al. 1990).

Power Graph Analysis helps to analyze high-throughput data by automatically highlighting the important information: in this case the separation of histones proteins into two differentially co-regulated groups, the P-loop domain containing subunits of the ORC complex and the FACT complex.

Overall, we see that power graphs give an insightful picture of the underlying biology. It should be stressed that these representations are obtained without the addition of biological background knowledge but instead by the network topology alone. Power graphs thus provide useful hints into the existence of complexes, their internal organization, and their relationships. Importantly, the power graph representation is a lossless representation, meaning that all and only interactions from the original network are represented faithfully. This is usually not the case for graph clustering methods.

In the following we apply power graph analysis to large-scale protein interaction networks.

3.3 Power Graph Analysis reveals hidden structures in protein interaction networks

In the following we show that protein interaction networks are significantly compressible when represented as poser graphs. This can be explained by the abundance of clique and biclique motifs. We validate the biological relevance of these motifs by finding a significant enrichment of power nodes in Gene Ontology terms and InterPro domains and motifs.

Edge reduction and conversion rate. As we have seen previously on specific examples, power graph analysis can help to disentangle complex protein interaction networks. A quantitative analysis requires the definition of measures. Here we introduce the edge reduction measure:

$$\textit{edge reduction} = \frac{\textit{edges} - \textit{power edges}}{\textit{edges}}$$

which is the proportion of edges collapsed in the power graph representation. Representing cliques and bicliques with power nodes and power edges allows to trade many edges for a hierarchy of power nodes. Power graphs have less power edges than edges in the original network as these get replaced by power nodes. To take into account the introduction of power nodes, we also compute the conversion rate of removed edges to power nodes:

$$conversion\ rate = \frac{edges - power\ edges}{non\ singleton\ power\ nodes}$$

From a visual complexity standpoint, trading edges for a hierarchy of sets of nodes is advantageous since the edges of a clique or biclique necessarily cross in two dimensions, whereas the circles delineating power nodes – by definition – do not.

Compressibility of protein interaction networks. Table 3 shows that up to 85% of the connectivity information in the networks is redundant. The conversion rate is correlated to both the average degree and edge reduction and thus adds little extra information.

Table 3 **Edge reduction and conversion rates for 13 protein interaction networks.**

Network	number of nodes	number of edges	average degree	edge reduction	conversion rate
Lim et al. (2006)	571	701	2.45	85%	12.1
Hazbun et al. (2003)	2243	3130	2.79	79%	13
Kim et al. (2006)	577	1090	3.78	67%	4.1
Gunsalus et al. (2005)	281	514	3.6	65%	4.6
Gavin et al. (2006)	1462	6942	9.4	64%	7.2
Ewing et al. (2007)	2294	6449	5.62	54%	6.6
Ito et al. (2001b)	3243	4367	2.69	53%	5.3
Rual et al. (2005)	1527	2529	3.31	50%	4.5
Krogan et al. (2006)	2708	7123	5.26	49%	4.5
Stanyon et al. (2004)	478	1778	7.43	48%	5.3
Butland et al. (2005)	1277	5324	8.33	43%	6.0
Arifuzzaman et al. (2006)	2457	8663	7.05	39%	5.4
LaCount et al. (2005)	1272	2643	4.16	38%	3.8

Protein interaction networks are significantly compressible. To evaluate the statistical significance of this result, we compare these levels of edge reduction to those of a network null model. We randomly rewire the networks while preserving the degree distribution of the network (Maslov and Sneppen 2002) and then recompute the corresponding power graphs. Fig. 41 shows the edge reduction for 13 protein interaction networks together with the box-plots for 1,000 randomly rewired networks. Computing the power graphs for 1,000 rewired networks per protein interaction network allows us to estimate the variance of the edge reduction and thus a z-score.

The z-scores obtained indicate that the original networks have significantly higher edge reductions than their rewired counterparts. At one extreme we have Gavin et. al. (2006) with a z-score of 242.

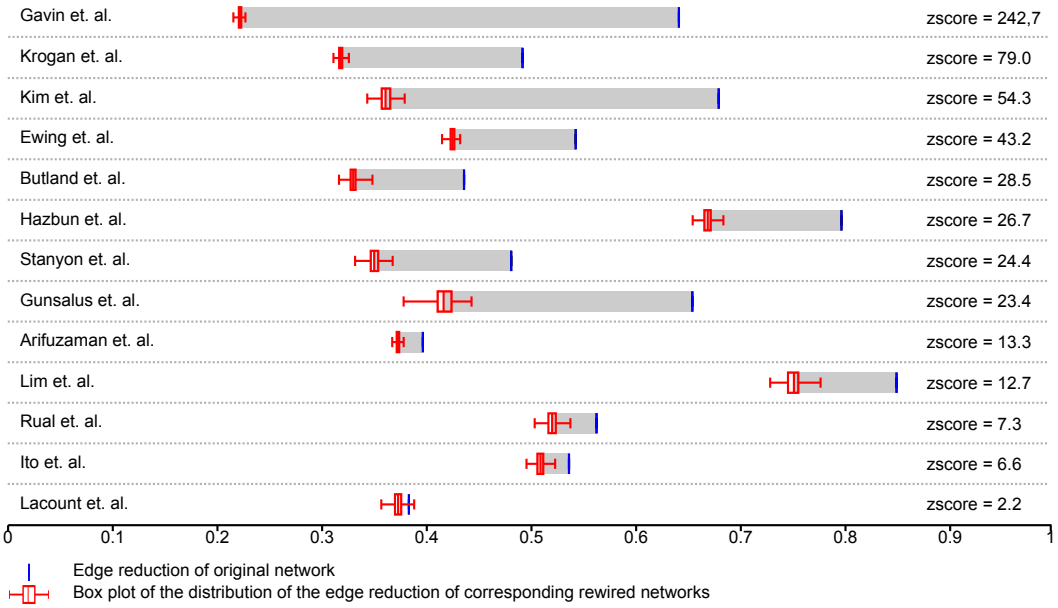


Fig. 41 **Comparison of 13 protein interaction networks to corresponding randomly rewired networks.** The distribution of edge reductions for rewired networks is represented as a box-plot: 50% of edge reduction values are inside the box and whiskers indicate min and max values. Most networks exhibit a significant deviation from the null model as indicated by high z-scores between 2.2 and 242.

Abundance of bicliques, cliques, and stars. Edge reduction and conversion rate are dependent on the abundance of stars, cliques and bicliques in the network – as these motifs require just one power edge to represent arbitrarily many edges. In particular, from the examples previously discussed (Casein Kinase II complex, and nucleosome), we expect cliques and bicliques to be the culprit. To ascertain that their abundance is indeed the explanation for the higher edge reductions, we examine the count of power edges having different sizes. Fig. 42 shows that power edges representing cliques and bicliques are abundant in the Gavin et. al. network, and absent for the corresponding rewired networks. Stars constitute most power edges found in the rewired networks at the exception of bicliques between groups of two nodes. Thus these protein interaction networks have significantly more cliques and bicliques than randomly rewired networks having the same number of nodes, and the same degree distribution.

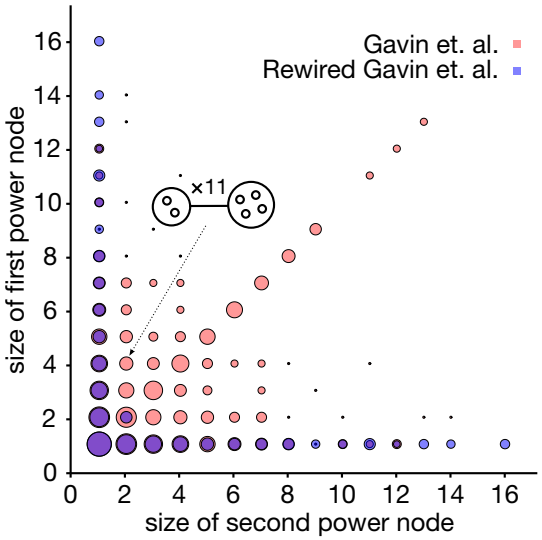


Fig. 42 **Distribution of bicliques, cliques and stars.** The area of each disc is proportional to the logarithm of the number of corresponding cliques and bicliques. Stars are found along the first column or row. For example, we find 11 bicliques between two nodes and 4 nodes. The diagram is symmetric along the diagonal. Protein interaction networks from Gavin et. al. (red) compared with corresponding rewired networks (blue). The high z-score (242) can be explained by significant abundance of cliques and bicliques compared to a random null-model obtained through rewiring. Note that the total count of cliques, bicliques, and stars, is not necessarily constant – even if number of edges is constant.

Having observed a significant abundance of cliques and bicliques, it remains the possibility that this is solely caused by experimental or methodological artifacts. However, we know of at least one case for which this cannot be the explanation: the Structural Interaction Network (SIN) by Kim et al. (2006). This network is a set of interactions carefully curated using structural information – all interactions reported are direct physical interactions explained by a known structural binding (Kim et al. 2006). This network exhibits a z-score of 54. Fig. 43 shows a close-up of a connected component of the SIN that illustrates its richness in structures – three cliques and two bicliques. The three cliques are enriched in Gene Ontology (Blake and Harris 2008) terms relevant to the spliceosome and to 35S primary transcript processing. Thus, proteins of this component are most likely part of the the ribosome and spliceosome machinery. Moreover, the examples previously given (Casein Kinase II complex, nucleosome, domain mediated interactions) in which power graphs give relevant insights on the structure of the networks, are often the rule and not an exception. For instance, the high z-score of Gavin's interaction network suggests that it is rich in structures with biological relevance.

Network motifs. These results corroborate studies that looked at network motifs identified as functional units in the context of biological networks (Alon 2007). Network motifs have been shown to admit generalizations composed of bicliques and stars (Kashtan et al. 2004). These patterns of interaction – characterized by a high connectivity – have been shown to be evolutionary conserved in the Yeast protein interaction network (Wuchty et al. 2003).

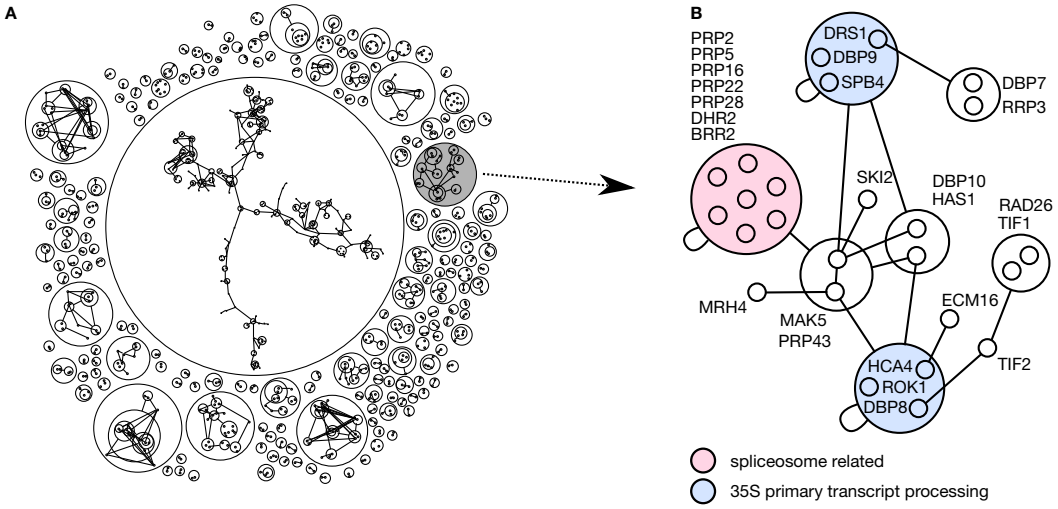


Fig. 43 **(A)** Close-up of a 25 node, 68 edges, connected component of the Structural Interaction Network (SIN) (Kim et al. 2006). **(B)** Power graph consisting of 17 power edges, thus an edge reduction of 73%. Three cliques enriched in GO terms related to 35S primary transcript processing and to the spliceosome become explicit in the representation.

Questioning the scale-free hypothesis. It has been argued recently that other distributions than the power-law are a better fit to the observed degree distributions of protein interaction networks (Khanin and Wit 2006; Thomas et al. 2003). It was also been shown that the scale-free property is not necessarily an intrinsic property of the networks, but could be an artifact caused by selection regularities in the sampling procedures (Stumpf et al. 2005; Han et al. 2005). Recently, Lima-Mendez and van Helden (2009) have argued that global properties of networks are averages that hide much detail. Other models for protein interaction networks, such as geometric random networks (Przulj et al. 2004) have been shown to be a better fit when looking at the motif composition of protein interaction networks. Our results show that the degree distribution does not characterize completely the idiosyncrasies of protein interaction networks: abundance of stars, cliques and bicliques is also an important signature.

3.3.1 Domain and gene ontology term enrichment of power nodes

We have previously argued that bicliques in protein interaction networks are a signal of shared protein domains. In the previous examples we showed that a power node of three proteins: ORC1, ORC4, and ORC5, have a P-loop domain. To further support the idea that power nodes are not artifacts of the networks topology but have a biological interpretation, we analyzed eleven networks for enrichment of power nodes in InterPro (Hunter et al. 2009), domains and motifs as well as for the enrichment in Gene Ontology annotations (GOA) (Blake and Harris 2008; Barrell et al. 2009). We found that for six networks more than half of the power nodes can be explained by a domain with a p-value of at least 10^{-3}. To a lesser extent, we also find that Gene Ontology terms also explain many power nodes – especially for networks derived from the Yeast interactome.

Method – hyper-geometric test. Our *null hypothesis* is that domain and Gene Ontology annotations are randomly distributed following an hyper-geometric distribution (Rivals et al. 2007). As shown in Fig. 44, we compute the *p*-value for each power node and annotation. We use Bonferroni's correction and compute a corrected *p*-value $p_c = mp$, where m is the number of tests performed for each network. In order to take into account missing domain annotations, we only consider power nodes with more than two thirds of their proteins annotated with at least one Gene Ontology term or at least one protein domain.

$$p(U,a) = 1 - \sum_{i=0}^{k-1} \frac{\binom{v}{i}\binom{n-v}{u-i}}{\binom{n}{u}}$$

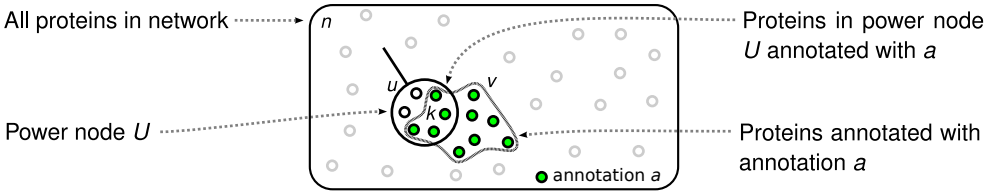


Fig. 44 **Evaluating the statistical significance of protein enrichment in Gene Ontology annotations and protein domains with the hyper-geometric test.** Power node U contains a number u of proteins. And a total of v proteins are annotated with annotation a. The number of proteins in U annotated with a is k. The p-value $p(U,a)$ is the probability that – by chance alone – k or more proteins in power node U are annotated with annotation a.

Results. Table 4 shows the distribution of *p*-values for the enrichment in Gene Ontology Annotations (GOA), with most *p*-values below 10^{-2}. Similarly, Table 5 shows that sufficiently annotated power nodes are significantly enriched in domains, with most *p*-values below 10^{-3}. A first observation is that protein domains are a better explanation for the grouping of proteins than GOA. However, the networks derived from the Yeast interactome show comparable enrichments in GOA and in protein domains. For other species such as Human, E. coli, and especially P. falciparum, the coverage and quality of annotations limits the ability of GOA to explain the power nodes. In contrast, domain annotations in InterPro are systematic and have high coverage. InterPro annotations are produced by the InterProScan program that searches for all occurrences of a domain or motif (Hunter et al. 2009).

Overall, these results show that the power nodes identified by power graph analysis have biological relevance and correspond to groups of proteins sharing protein domains and having a similar function.

Table 4 **Percentage of power nodes that are significantly enriched in Gene Ontology annotations.** Note that many protein are not annotated with Gene Ontology terms. Hence, only power nodes with an annotation coverage of more than two thirds are considered.

Protein interaction network	Species	$p < 0.001$	$p < 0.01$	% non-annotated
Ito et al. (2000)	Yeast	19%	32%	53%
Kim et al. (2006)	Yeast	86%	96%	90%
Gavin et al. (2006)	Yeast	75%	89%	44%
Krogan et al. (2006)	Yeast	79%	88%	48%
Ewing et al. (2007)	Human	13%	23%	8%
Rual et al. (2005)	Human	32%	32%	72%
Lim et al. (2006)	Human	8%	19%	10%
Arifuzzaman et al. (2006)	E. coli	4%	5%	44%
Butland et al. (2005)	E. coli	13%	19%	22%
LaCount et al. (2005)	P. falciparum	0%	31%	65%

Table 5 **Percentage of power nodes that are significantly enriched in protein domains and motifs.** Each protein is annotated with one or several protein domains or motifs from the InterPro database (Hunter et al. 2009). Most power nodes turn out to be enriched at a level of statistical significance of 1 per-thousand. Note that many proteins are not annotated with protein domains or motifs: only power nodes with an annotation coverage of more than two thirds are considered.

Protein interaction network	Species	$p < 0.001$	$p < 0.01$	% non-annotated
Ito et al. (2000)	Yeast	50%	100%	87%
Kim et al. (2006)	Yeast	90%	96%	0%
Gavin et al. (2006)	Yeast	70%	91%	3%
Krogan et al. (2006)	Yeast	78%	88%	6%
Ewing et al. (2007)	Human	53%	85%	9%
Rual et al. (2005)	Human	66%	66%	33%
Lim et al. (2006)	Human	39%	56%	10%
Arifuzzaman et al. (2006)	E. coli	46%	74%	0%
Butland et al. (2005)	E. coli	38%	72%	0%
LaCount et al. (2005)	P. falciparum	27%	53%	25%

3.4 Application to regulatory and sequence similarity networks

Other networks defined on proteins and genes also benefit from power graph analysis. An example are protein sequence similarity networks (Medini et al. 2006) in which nodes are proteins and edges represent BLAST E-values below a given threshold. These networks are geometric networks defined on the space of sequences with the BLAST *E*-value as a distance. Geometric networks are known to be saturated in cliques and bicliques (Przulj et al. 2004). Another example is the analysis of gene regulatory networks with power graph analysis. Gene duplication events and combinatorial sharing of transcription factor promoter regions create biclique motifs (Teichmann and Babu 2004; Alon 2007). Fig 45A illustrates a typical example in which bicliques arise from sharing regulatory motifs. For example,

in Yeast the genes for histone subunits HTA1 and HTB1 share the same promoter region, and are thus under the regulation of the same transcription factors. In the case of sequence similarity networks, cliques are often found for groups of highly similar proteins. Bicliques arise between otherwise more distant proteins that share similarity on a specific region, for example because of a shared domain (Fig 45B). A general principle by which cliques and bicliques occur in protein networks is now apparent: sequence regions such as domains and regulatory motifs are shared across different proteins. Their reuse as elementary building blocks and combinatorial arrangements causes cliques and bicliques in biological networks.

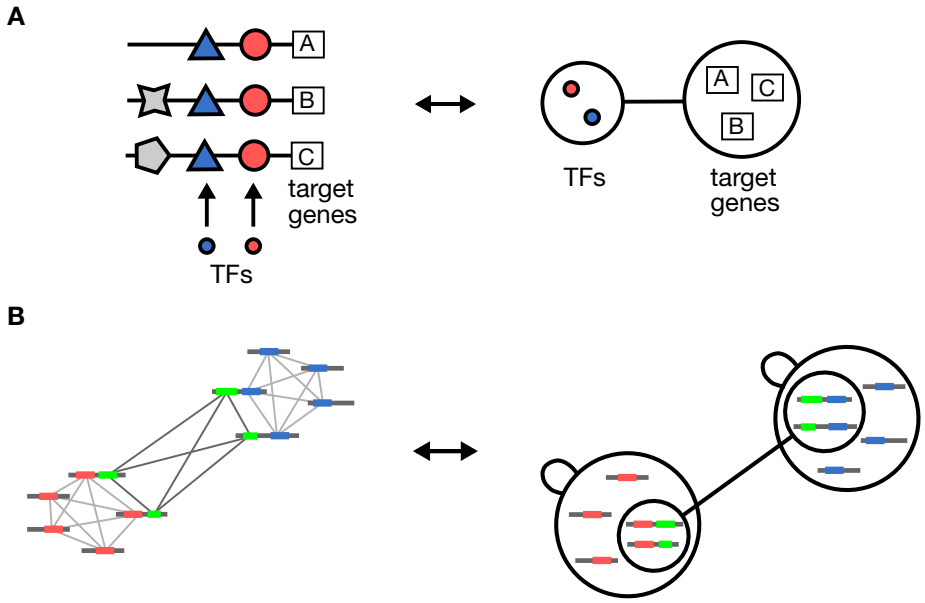


Fig. 45 **Examples of occurrences of bicliques in gene regulatory networks and sequence similarity networks. (A)** Bicliques may occur in regulatory networks owing to two reasons: transcription factors operate within complexes – combinatorial regulation – and regulatory motifs in promoter regions may be shared and repeated for different genes. **(B)** In sequence similarity networks, proteins sharing a sequence region of high similarity – i.e. a domain – induce cliques. Bicliques are induced between sub-groups of proteins owing to additional regions of sequence similarity.

3.4.1 Example 4 – Transcription factors to target genes network

Beyer et al. (2006) presented an integrative approach for assigning transcription factors to target genes in S. cerevisiae using data from chIP-chip experiments, known binding motifs, clusters of co-expression and other evidences. The result is a highly accurate bipartite network between transcription factors and target genes. The authors identified – among others – YAP1, YAP7 and MSN2 as part of a transcription factor module related to the stress response of S. cerevisiae. To investigate if a similar module could be identified with power graph analysis, we computed the power graph of the whole network and searched for the region containing YAP1, YAP7 and MSN2.

Regulatory modules. As shown in Fig 46, a group of transcription factors – SKN7, MSN2, MSN4, YAP1, YAP2(CAD1), and YAP7 – have similar target genes. Two sub groups are identified with different regulation profiles: SKN7/MSN2/MSN4 and YAP1/YAP2/YAP7. Also shown in Fig 46, target genes are grouped according to common transcription regulators. For example MSN2 and MSN4 both regulate 26 target genes predominantly involved in protein folding (p-value $< 10^{-5}$) and heat shock proteins (p-value $< 10^{-10}$). Interestingly, YAP1, YAP2 and YAP7 have in common 19 target genes involved in detoxification (p-value $< 10^{-6}$). The transcription factors MSN2, MSN4, and SKN7 are known to regulate the expression of genes in response to stresses, such as heat and osmotic shock, oxidative stress, low pH, glucose starvation, sorbic acid and high ethanol concentrations (Gasch et al. 2000).

Function of YAP7. YAP1, YAP2 and YAP7 are similar bZIP proteins of the YAP family characterized by unusual amino acid substitutions of their bZIP domains (Fernandes et al. 1997). It is known that YAP1 and YAP2 are involved in the transcriptional response to drugs, oxidative stress and metal detoxification (Gasch et al. 2000). YAP7 is nevertheless a poorly characterized transcription factor most similar – within the YAP family – to YAP6 whose over expression increases sodium and lithium tolerance (Mendizabal et al. 1998). The strong overlap of gene targets of YAP1, YAP2, and YAP7 and the common metal detoxification function of YAP1/YAP2 and YAP6, suggests that YAP7 also plays a role in metal detoxification.

Power Graph Analysis can decompose a bipartite network into an union of bicliques. This decomposition leads to a hierarchy of clusters of transcription factors linked to a hierarchy of clusters of target genes.

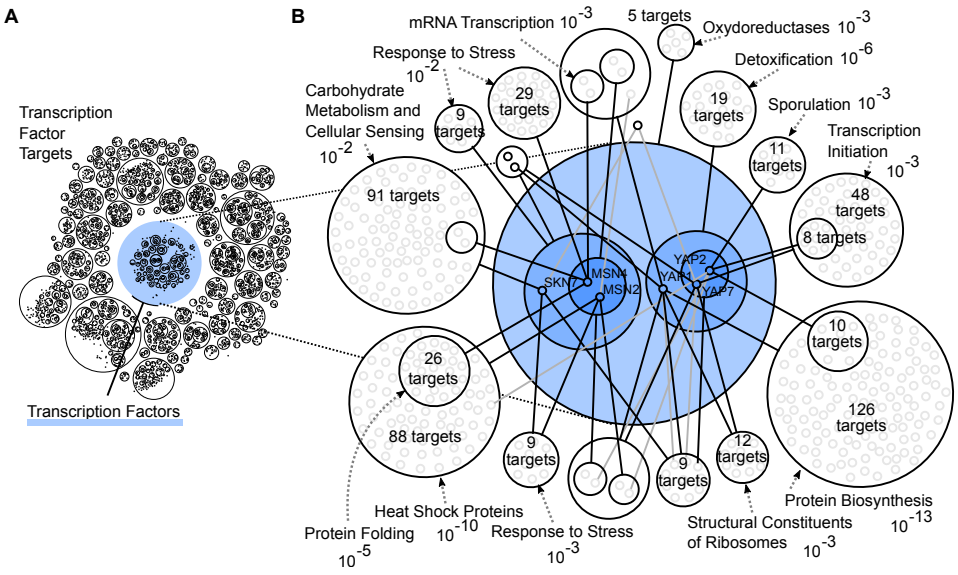


Fig. 46 **Power Graphs Analysis of a transcription regulation network. (A)** Power node hierarchy of the complete bipartite network between 158 transcription factors and 4217 target genes consisting of 13239 assignments. **(B)** Gene targets landscape of a group of transcription factors – SKN7, MSN2, MSN4, YAP1, YAP2(CAD1), and YAP7 – regulating the general stress response of S. cerevisiae. Target genes are grouped within power nodes and linked with power edges signifying the assignment of transcription factors to targets. Dominant GO categories in target gene sets are indicated with the p-value.

3.4.2 Example 5 – Human protein tyrosine phosphatase sequence similarity network

The protein tyrosine phosphatase (PTP) family (Pils and Schultz 2004) has a central role in signal transduction by controlling the phosphorylation state of tyrosine, serine and threonine residues. Tyrosine-specific protein phosphatases (EC:3.1.3.48) catalyse the removal of a phosphate group attached to a tyrosine residue.

Tyrosine phosphatase sequence similarity network. The network consists of 279 proteins and 4849 edges. Edges between two proteins corresponds to highly significant alignments of the sequences with a BLASTP *E*-value of at most 10^{-46}. The protein tyrosine phosphatase sequence similarity network is shown in Fig 47.

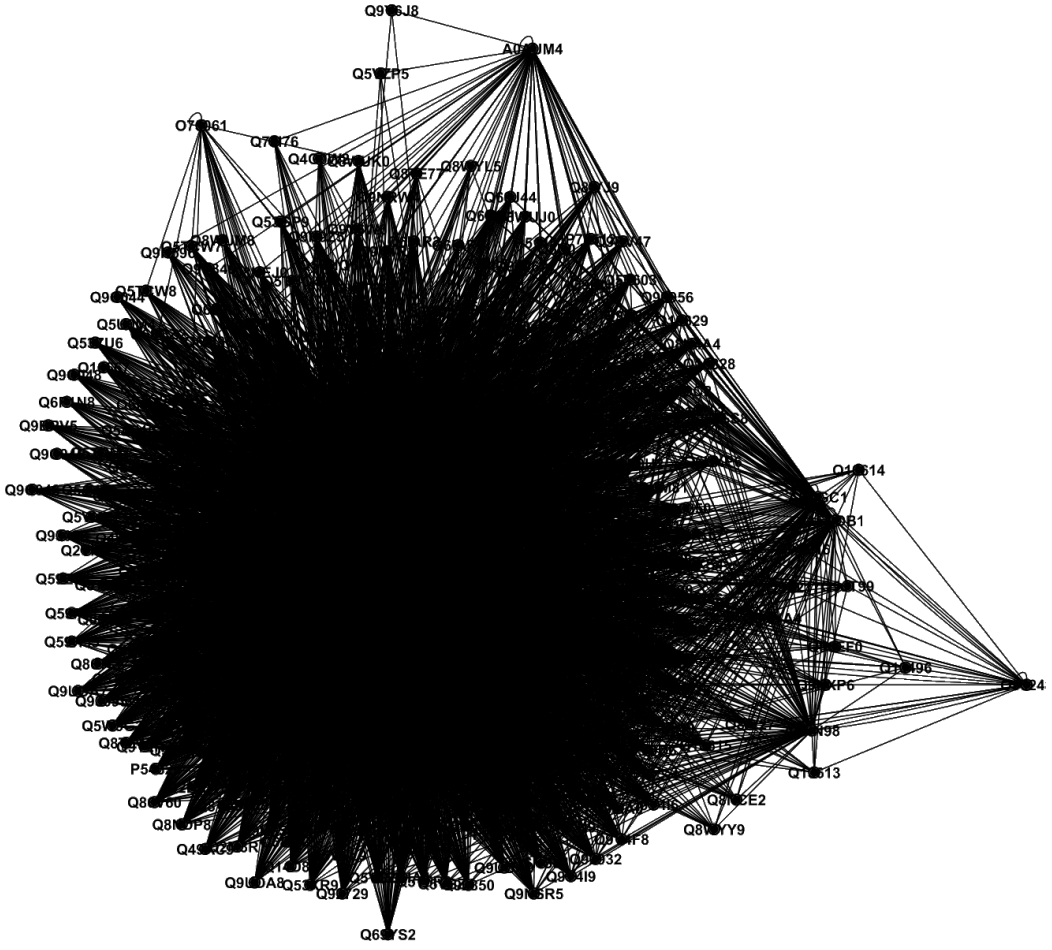


Fig. 47 **Sequence similarity network of all Human protein tyrosine phosphatases.** The underlying graph has 279 nodes and 4849 edges. Each node represents a Human protein tyrosine phosphatase, with an edge between two proteins corresponding to highly significant alignments with E-values of at most 10^{-46}. The network is obtained by an all against all BLASTP scan using the NCBI BLASTP tool (Ye et al. 2006). It is almost impossible to analyze the network in this form.

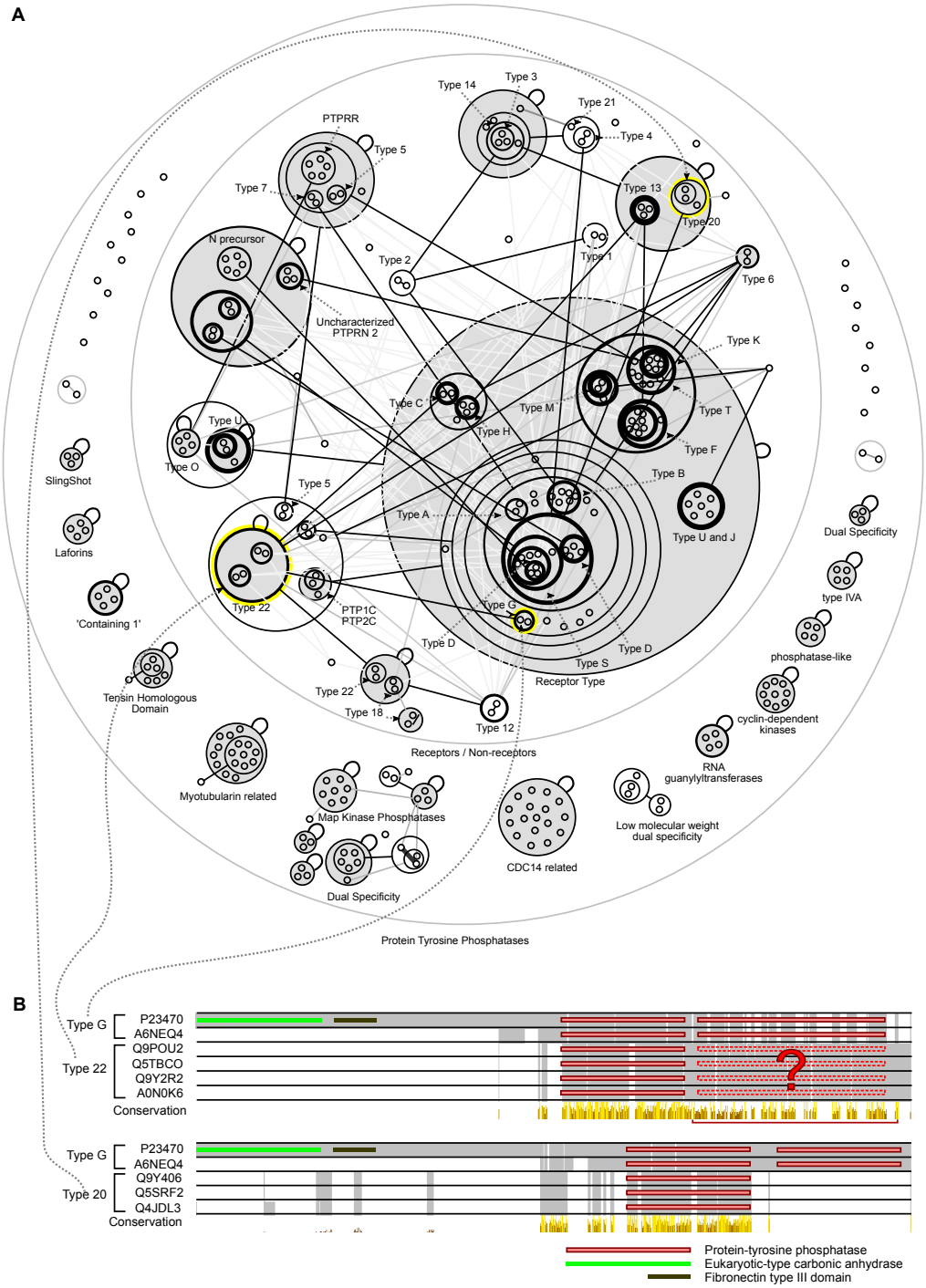


Fig. 48 **Power graph of the Human protein tyrosine phosphatase sequence similarity network. (A)** The power graph has 279 nodes, 95 non-singleton power nodes, and 209 power edges. Grayed power nodes correspond to totally connected sets of proteins. For example, all receptor type protein tyrosine phosphatases have an alignment *E*-value of at least 10^{-46}. Black power edges represent many edges of low E-values (lower than 10^{-46}), light-gray power edges abstract fewer edges and correspond to less significant sequence similarities. **(B)** Multiple sequence alignment for type G against type 22 and type G against type 20. The similarity observed in the power graph between type G and type 22 is explained by the homology between a region of type 22 non-receptors and the second copy of the tyrosine phosphatase domain of type G receptors. Negative control: type G and type 20 – which are not linked – do not share this similar region.

Tyrosine phosphatase sequence similarity power graph. The power graph of the PTP sequence similarity network is shown in Fig 48A. The network consists of 279 nodes, each one representing a protein. PTPs are usually classified into classical specific phosphatases, dual specificity phosphatases, and other minor classes, such as low molecular weight phosphatases and myotubularins. Classical specific phosphatases are further subdivided into receptor type and non receptor type. Unsurprisingly, because of their sequence similarities, the categories of receptor, non-receptor, and dual-specificity phosphatases are delineated by the power graph representation. For example the receptor type PTPs are grouped in one power node meaning they all are similar to one another with E-values below 10^{-46}. The same is observed for different classes of non-receptor type PTPs, such as myotubularins. Importantly, the different classes of receptor PTPs, such as types A, B, C, D, F, H, T are discriminated solely by shared similarity to non-receptor PTPs.

Optimal representation. The choice of a threshold for the *E*-value influences the power graph representation. We observe that for the value of 10^{-46} the power graph reveals the most details. For that value, the lossless reduction in complexity achieved by the power graph representation reaches 95% edge reduction – from 4849 edges to 209 with 95 power nodes. The clustering of proteins in the power graph corresponds to the known classification of PTPs: 82% of leaf power nodes (that do not contain power nodes) have all of their proteins belonging to exactly the same subfamily. While the previous results could have been obtained through the hierarchical clustering of the sequences, power graphs reveal additional details.

Similarity cross-links and domain erosion. Compared with traditional clustering methods, the cross-links between different regions of the hierarchy constitute a novel insight. For example, a group of 6 type B receptor PTPs are linked by a power edge to two type 2 non-receptor PTPs. Fig. 48B shows the multiple alignment of the corresponding sequences. While the common PTP domains are aligned for the six sequences, we also observe that the second copy of the tyrosine phosphatase domain of the two type G PTPs align to an unannotated region of about 370 amino acids with a sequence identity of 14% and a similarity of 39% (BLOSUM 62). This region corresponds with high probability (NorMD = 1.014) to a non-receptor phosphatase domain listed in ProDom – a database of automatically generated clusters of homologous sequence fragments (Bru et al. 2005). To verify if this region is responsible for the high similarity (E-value $< 10^{-46}$) between the type G receptor PTPs and type 22 non-receptor PTP, we compared the sequences of type G PTPs to a group of proteins to which they are not connected in the power graph: type 20 PTPs. As Fig. 48B shows, there is no region aligning with the second copy of the phosphatase domain. This result suggests that the second phosphatase domain of type 22 PTPs got eroded though the accumulation of mutations following a release in selective pressure.

Detection of similarity cross-links in the hierarchy is the contribution of Power Graph Analysis to the analysis of sequence similarity networks. These cross-links constitute a weak signal in networks and are difficult to detect. Evidence for this

domain erosion is carried by only eight similarity links between four and two proteins whereas the original network has 4849 edges. In the power graph representation it is one power edge among only 209.

3.5 Power graph algorithm

Before outlining the algorithm and its evaluation, we first formally define power graphs. We show how minimal power graphs grouping neighborhood similar nodes can be used to represent networks in a more succinct manner.

3.5.1 Power graphs

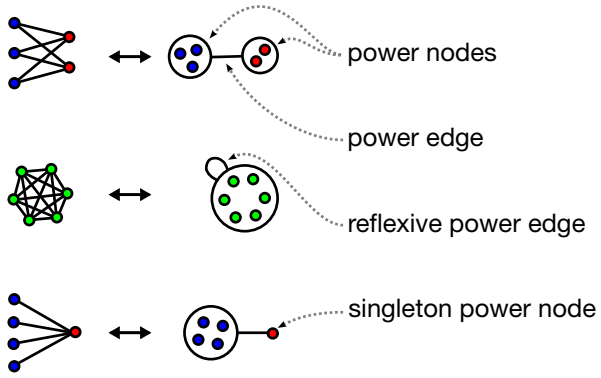


Fig. 49 **Power graphs.** Power nodes are sets of nodes. Power nodes of size one are called singleton power nodes. Power edges connect two power nodes. A power edge between two power nodes signifies that all the nodes of the first power node are adjacent to every node of the second power node. This leads to biclique motifs in the graphs, of which stars and cliques are special cases.

Definition. Given a graph $G = \left(V, E\right)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges, A *power graph* $\mathring{G} = \left(\mathring{V}, \mathring{E}\right)$ is a set of power nodes $\mathring{V} \subseteq \mathcal{P}(V)$ and a set of *power edges* $\mathring{E} \subseteq \mathring{V} \times \mathring{V}$. We say that two disjoint[1] power nodes $U, W \in \mathring{V}$ are adjacent if there is a power edge (U, W) in $\mathring{E}$. All power nodes in $\mathring{V}$ must participate in at least one power edge.

As illustrated in Fig. 49, a power graph $\mathring{G}$ represents graph G when the following holds: If and only if in $\mathring{G}$ two power nodes U and W are adjacent, then in G all nodes in U are adjacent to all W nodes[2]. Similarly, if and only if a power node is self-adjacent, then in G the nodes in U are all adjacent to each other[3]. There is one exception, we ignore self-adjacent nodes: $(u, u) \notin E$. It follows that power edges in $\mathring{G}$ represent bicliques, cliques and stars in G. Reciprocally, given a graph G, its bicliques, cliques and stars can be represented by power edges in $\mathring{G}$. In addition we further constrain the definition of power graphs by requiring the following two conditions:

Power node hierarchy condition: Any two power nodes are either disjoint, or one is included in the other. Therefore, power nodes form a hierarchy (Fig. 49A). This guarantees that the power node hierarchy can be represented in the plane which facilitates visualization.

---

1 such that $U \cap W = \emptyset$

2 $(U, W) \in \mathring{E}$ if and only if $\forall u \in U, \forall v \in W : (u, v) \in E$

3 $(U, U) \in \mathring{E}$ if and only if $\forall u, v \in U, u \neq v : (u, v) \in E$

Power edge partition condition: Each edge of the original graph is represented by one and only one power edge. In other terms, the power edges form a partition of the set of edges (Fig. 49B).

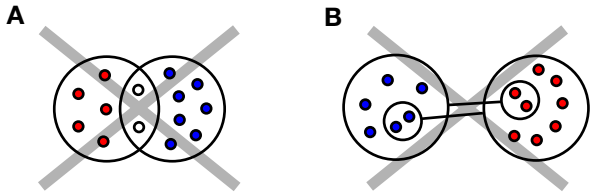


Fig. 50 **Definition of power graphs. (A)** Power node hierarchy condition – power nodes must form a hierarchy and thus no strict intersections are allowed. **(B)** Power edge partition condition – an edge must be represented by one and only one power edge.

Power graph equivalence. We can define the notion of equivalence between power graphs. Two power graphs $\mathring{G}_1$ and $\mathring{G}_2$ are equivalent if they represent the same graph G. Fig. 51 shows the example of the diamond graph that admits 20 equivalent - yet different - power graphs (including the trivial power graph which is the graph itself). Five of these have the minimum number of power edges and power nodes.

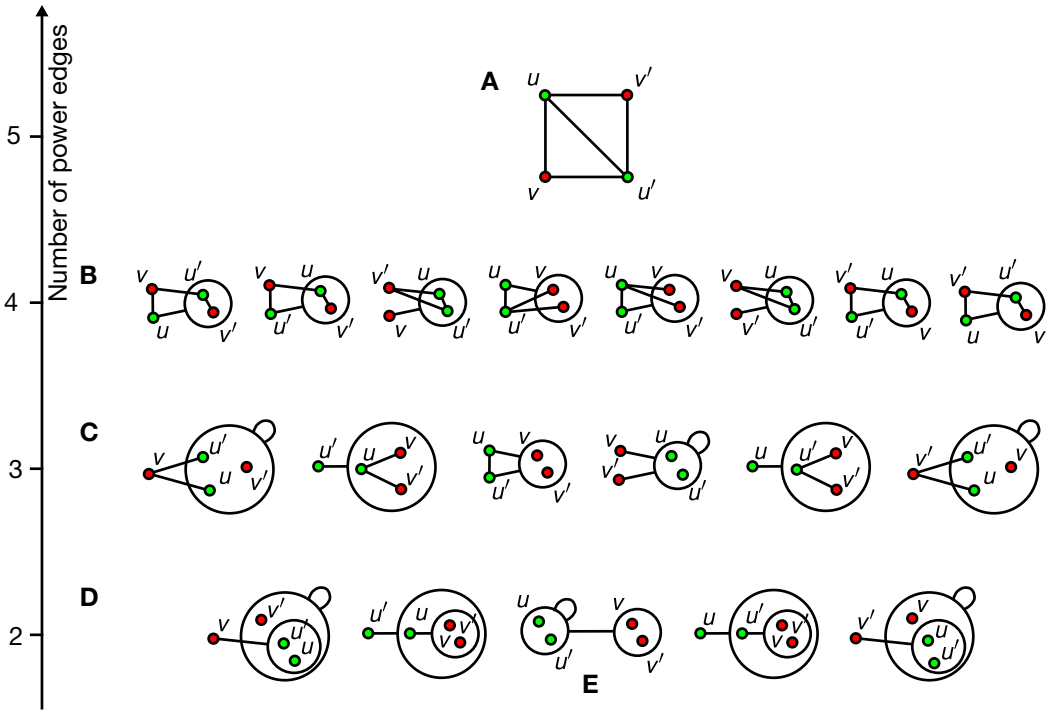


Fig. 51 **Equivalent power graphs for the diamond graph. (A)** A graph is a valid – yet trivial – power graph of itself. Note that node u has the same neighbors as node u', same holds for v and v'. **(B, C, D)** Power graphs of the diamond graph having four, three and two power edges. **(E)** This power graph is the optimal choice among the 5 minimal power graphs because it groups together nodes with similar neighborhoods (in this case identical).

Minimal power graphs. Among these five power graphs, one (indicated by **E**) is intuitively the most elegant representation of the diamond graph because it uses a minimum number of power edges and power nodes, and groups together nodes with similar neighborhoods. We define minimal power graphs as power graphs with the least number of power edges. For a given graph G, let's define an order on all its power graphs:

$$\mathring{G}_1 \leq \mathring{G}_2 \quad \text{iff} \quad \left|\mathring{E}_1\right| \leq \left|\mathring{E}_2\right| \text{ or } \left(\left|\mathring{E}_1\right| = \left|\mathring{E}_2\right| \text{ and } \left|\mathring{V}_1\right| \leq \left|\mathring{V}_2\right|\right)$$

Among other equivalent power graphs with the same number of power edges, minimal power graphs have the least power nodes. Minimal power graphs are not necessarily unique. As shown in Fig. 51 there are 5 minimal power graphs of the diamond graph.

Minimal power graphs based on neighborhood-similar power nodes. Among all minimal power graphs we choose those that group together nodes with similar neighborhoods (Fig. 51E). In the following we show how power nodes can be obtained by clustering the graph's nodes based on their neighborhood similarity.

3.5.2 Algorithm outline

Problem. *Let G be a graph. Find a minimal power graph $\mathring{G}$ representing G that groups together nodes with similar neighborhoods.*

We simplify the problem by first determining candidate power nodes and then searching for a minimal power graph built upon these power nodes:

- Candidate power nodes are found by neighborhood similarity hierarchical clustering of the graph's nodes. Candidate power nodes are sets of highly neighborhood similar nodes.
- A greedy search algorithm minimizes the number of power edges. We search for a minimal partition of the set of edges E into disjoint cliques and bicliques defined on a hierarchy of neighborhood-similar power nodes.

Simultaneous clustering of the nodes and edges. The power graph algorithm is the simultaneous clustering of the nodes and edges of a graph G. The pseudo-code for both steps of the algorithm is given in page 86. It consists in identifying clusters of nodes (line 1 to 13), and then clusters of edges – cliques and bicliques – as power edges between the candidate power nodes (line 14 to 41). As shown in Fig. 52, the first step of the algorithm searches for candidate power nodes by neighborhood similarity hierarchical clustering. The second step is the greedy search for a minimal power edge partition of G based on the candidate power nodes obtained after the first step.

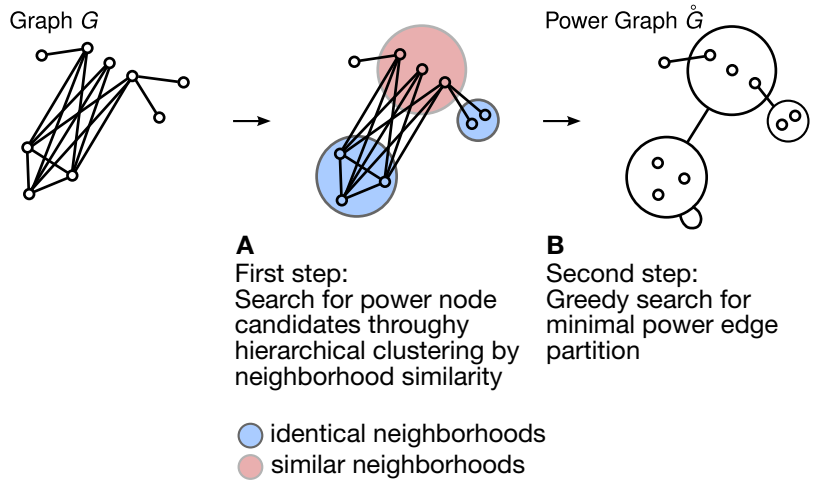


Fig. 52 **Outline of the power graph algorithm. (A)** First step: candidate power nodes are found using neighborhood similarity clustering of the nodes in *G*. **(B)** Second step: A greedy search for a minimal partition of edges in *G* into disjoint power edges (cliques or bicliques) is based on the candidate power nodes obtained in the first step.

3.5.3 First step – Search for candidate power nodes

The power graph algorithm is based on the observation that good candidate power nodes are node sets that have many common neighbors – their nodes have highly similar neighborhoods. As shown in Fig. 53, sets of nodes that have many common neighbors are more likely to be part of a clique or biclique – and are thus good candidates for power nodes.

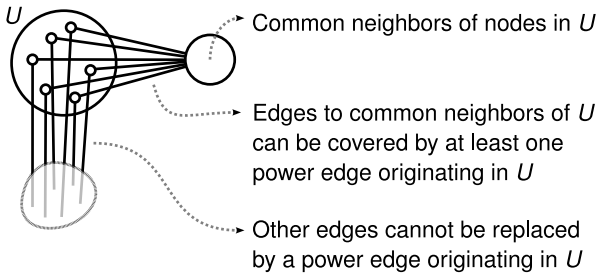


Fig. 53 **Candidate power nodes.** A *good* candidate power node *U* is characterized by its nodes having many common neighbors and few distinct neighbors – a property that can be quantified using neighborhood similarity.

Enumerating all potential power nodes is intractable since 2^n candidate power nodes must be considered for a graph on n nodes. Instead, the power graph algorithm finds candidates through an agglomerative hierarchical clustering based on neighborhood similarity (pseudo-code 1, lines 4 to 13).

Agglomerative hierarchical clustering. We use agglomerative hierarchical clustering based on the neighborhood similarity of nodes in the graph *G*. Partitive clustering schemes such as k-means are inappropriate because power nodes form a

hierarchy and not a partition on the set of nodes. We chose instead the simplest clustering scheme that returns hierarchical clusters. In the following we give details on the neighborhood similarity measure used.

Weighted sets. The Jaccard neighborhood similarity is defined for pairs of nodes. It can be generalized for pairs of *sets* of nodes using weighted neighborhood sets (Syropoulos 2001). Weighted sets are different from sets in that elements have a weight – each element is labeled with a positive real number. The cardinality of a weighted set is the sum of its weights. As shown in Fig. 54B, the intersection of two weighted sets is obtained by taking the minimum weight for each element. The union of two weighted sets is obtained by taking the maximum weight for each element. Note that the absence of an element in a weighted set corresponds to a weight of zero.

Generalized Jaccard neighborhood similarity. As shown in Fig. 54A, the weighted neighborhood set of U is the weighted set $N_w(U)$ of nodes adjacent to nodes u in U. The weight of each node u in $N_w(U)$ is the proportion of nodes in U that it is adjacent to. The generalized Jaccard neighborhood similarity is defined between two node sets U and V as:

$$J(U,V) = \begin{cases} \dfrac{|N_w(U) \cap N_w(V)| + \phi}{|N_w(U) \cup N_w(V)| - \phi} & \text{if } |N_w(U) \cup N_w(V)| \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Where ϕ is the correction term:

$$\phi = \frac{|N_w(U) \cap V| + |N_w(V) \cap U|}{2}$$

This correction is necessary to ensure that cliques and bicliques are treated equally. Note that in $N_w(U) \cap V$, the set V is taken as a weighted set in which all elements have weight one. When U and V are singleton sets $\{u\}$ and $\{v\}$, the correction ϕ is equal to one if u is adjacent to v, and to zero if not. Therefore, this measure of neighborhood similarity – defined between two sets of nodes U and V – is a generalization of the Jaccard neighborhood similarity for two nodes u and v since it gives the same result for two singleton sets: $J(\{u\},\{v\}) = J(u,v)$.

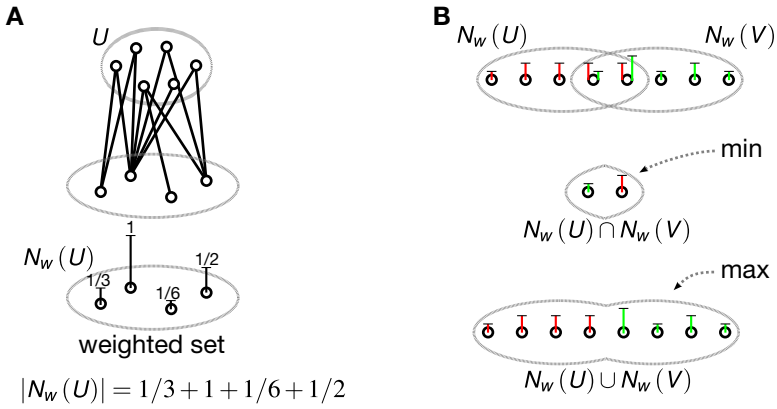


Fig. 54 **Generalized Jaccard neighborhood similarity. (A)** Weighted sets are different from sets in that elements have weights. The weighted neighborhood of node set U is the weighted set $N_w(U)$ – the weight of each node being the proportion of nodes in U to which it is adjacent. The cardinality of a weighted set is the sum of all weights. **(B)** The intersection of two weighted sets is the weighted set with the minimum weight for each node. The union of two weighted sets is the weighted set with the maximum weight for each element.

Correction term to treat cliques and bicliques equally. The correction term ϕ is a departure from the usual Jaccard index. As shown in Fig. 55, this correction guaranties the equal treatment of cliques and bicliques. This term comes from the requirement that nodes within an isolated clique or biclique must have *both* a neighborhood similarity of 1. If this special case is ignored then the pairwise similarities of nodes within a clique graph would not be 1 because nodes are not adjacent to themselves. Adding loop edges to all nodes would not remove the bias – but instead just displace it. Nodes within the biclique would then have a neighborhood similarity less than 1.

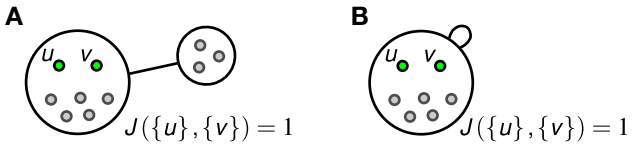


Fig. 55 **Correction term to treat cliques and bicliques equally.** The correction term ϕ in the neighborhood similarity is necessary for the equal treatment of cliques and bicliques. **(A)** The neighborhood similarity between two nodes u and v in a biclique is 1 – they have identical neighborhoods ($J(u,v) = 1$). **(B)** In a clique, the two nodes u and v would have identical neighborhoods only if loop edges would be allowed: (u,u) and (v,v). The correction term compensates the lack of loop edges so that in the case of cliques we also have: $J(u,v) = 1$. Note that choosing the convention that all nodes have loop-edges would only displace the problem: in that case nodes on either side of a biclique would not have a neighborhood similarity of 1.

Example of hierarchical clustering by neighborhood similarity. Fig. 56A shows a graph G of 11 edges on 10 nodes. Candidate power nodes can be found by neighborhood similarity hierarchical clustering. Table 6 lists the 9 cluster merging steps and the neighborhood similarities between the pairs of merged clusters and Fig. 56B

gives the corresponding dendrogram. As shown in Fig. 56C, among the 7 clusters, only two are power nodes of the minimal power graph $\mathring{G}$.

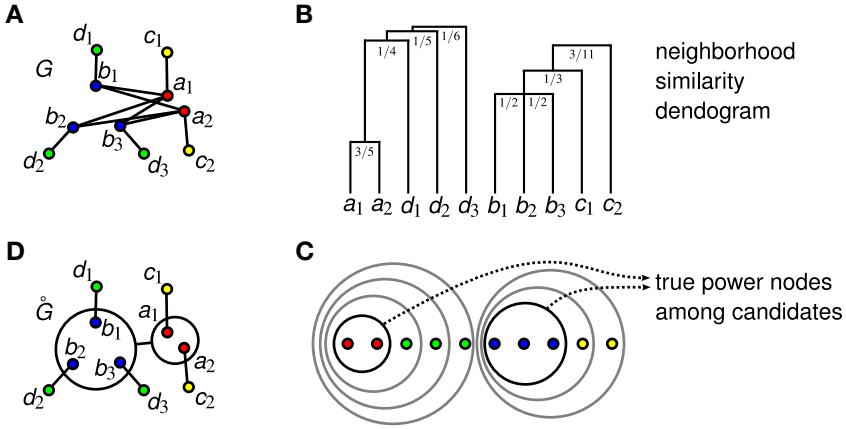


Fig. 56 **Example of neighborhood similarity hierarchical clustering. (A)** The example graph G is a biclique $C\left(\{a_1, a_2\}, \{b_1, b_2, b_3\}\right)$ with 5 additional edges – one for each of its 5 nodes. **(B)** Hierarchical clustering by neighborhood similarity returns 7 clusters. the two clusters $\{a_1, a_2\}$ as well as $\{b_1, b_2, b_3\}$ are clustered first as they have the most similar neighborhoods: $3/4$ and $1/2$, respectively. **(C)** These two clusters are also the only valid power nodes among the 7 candidates found by hierarchical clustering. **(D)** *The* minimal power graph $\mathring{G}_{min}$ of graph G.

Table 6 **Sequence of merging steps.** The hierarchical clustering by neighborhood similarity for graph G (Fig. 56A) is done in 9 successive cluster merging steps.

step	first set	second set	neighborhood similarity
1	$\{a_1\}$	$\{a_2\}$	0.6
2	$\{b_1\}$	$\{b_2\}$	0.5 [a]
3	$\{b_1, b_2\}$	$\{b_3\}$	0.5 [a]
4	$\{b_1, b_2, b_3\}$	$\{c_1\}$	$0.\overline{3}$
5	$\{b_1, b_2, b_3, c_1\}$	$\{c_2\}$	$0.\overline{27}$
6	$\{a_1, a_2\}$	$\{d_1\}$	0.25
7	$\{a_1, a_2, d_1\}$	$\{d_2\}$	0.2
8	$\{a_1, a_2, d_1, d_2\}$	$\{d_3\}$	$0.1\overline{6}$
9	$\{a_1, a_2, d_1, d_2, d_3\}$	$\{b_1, b_2, b_3, c_1, c_2\}$	0

a Step 2 and 3 may also be considered as a single merging step since the similarities are equal.

Completing the list of power node candidates Hierarchical clustering by neighborhood similarity provides an initial but not sufficient list of power node candidates. As shown in Fig. 57, overlapping cliques and bicliques can sometimes hide each other. The initial collection of candidates can be extended by adding for each candidate U the set corresponding to the maximal biclique originating in U. Let $N(U)$ be the set of common neighbors of nodes in U – it is the set of nodes adjacent to *all* nodes in U. For each cluster U, we add to the list of candidate power nodes its

neighborhood set $N(U)$ and second-order neighborhood set $N(N(U))$. There is no need to add $N^3(U)$ because $N^3(U) = N(U)$.

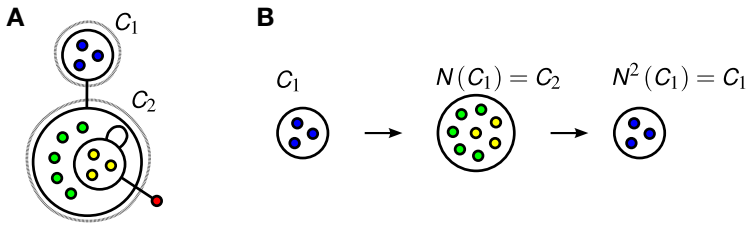


Fig. 57 **Completing the list of power node candidates (A)** Hierarchical clustering may not find all necessary candidates. While candidate power node C_1 is easily found by clustering, power node C_2 might not be found. In order to find the biclique $C(C_1, C_2)$ we need both C_1 and its set of common neighbors C_2. **(B)** For each cluster U we add the neighborhood set $N(U)$ and second order neighborhood set $N(N(U))$.

3.5.4 Second step – Search for power edges

In the second step, power edges are searched (pseudo-code 1, lines 14 to 41 and Fig. 58). Among all pairs of candidate power nodes (U, W) we retain all that induce a clique or biclique in the graph G (lines 14 to 22 and Fig. 58A) – these are the candidate power edges. Following the heuristic of making the locally optimum decision at each step, we perform a *greedy search*: we add the candidate power edges that cover most edges first with the hope of finding the global optimum (Cormen et al. 1990). As shown in Fig. 58B, the candidate power edges are put into a list sorted by decreasing size. The biggest candidate is removed from the list and considered first (Fig. 58C). Candidate power edges cannot be added to the power graph if they do not respect the hierarchy of power nodes (lines 27 to 30, and Fig. 58D) or if they cover an edge that is already covered by a power edge previously added to the power graph (lines 31 to 36, Fig. 58E). In these two cases we need to decompose the candidate (Fig. 58D and E) into smaller but compatible pieces that are put back into the sorted list. The search terminates when the list is empty and all candidates have been decomposed and eventually added to the power graph. If any edge (u, v) from the graph still needs to be covered in the power graph, they are added as singleton power edge: $\big(\{u\}, \{v\}\big)$ (lines 40 to 41).

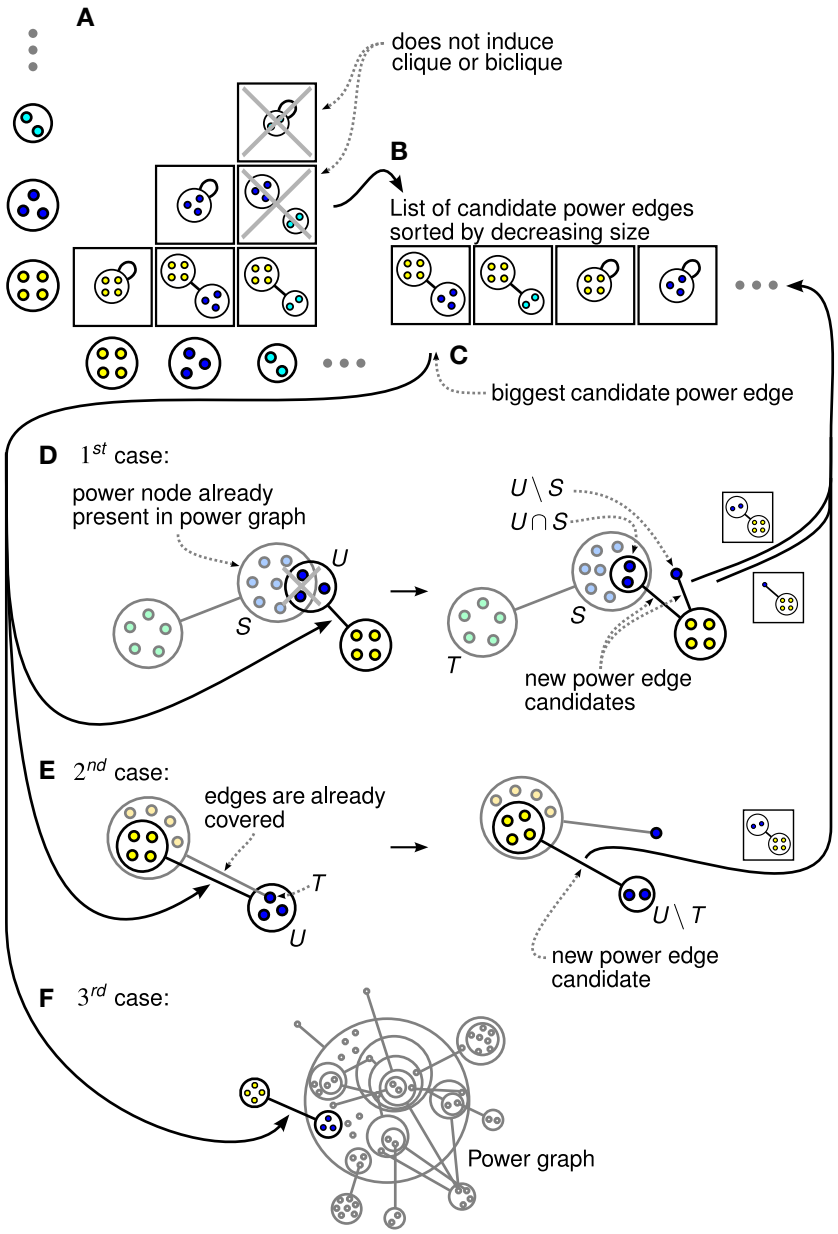


Fig. 58 **Greedy search for the minimal power graph. (A)** All candidate power edges are enumerated – only candidates inducing cliques (diagonal) or bicliques are considered (pseudo-code: line 16 to 22). Stars are found as a special case of biclique. **(B)** The candidates are added into a list sorted by decreasing size. The size of a power edge is the number of edges that it covers (line 24). **(C)** The biggest power edge is considered as first candidate (line 25). **(D)** 1^{st} case: one of the power nodes of the candidate power edge strictly intersects with a power node already present in the power graph. The candidate is decomposed and its pieces are put back into the sorted list (line 27 to 30). **(E)** 2^{nd} case: the candidate power edge covers edges also covered by a power edge already present in the power graph. The candidate is decomposed and if a piece remains it is put back into the sorted list (line 31 to 36). **(F)** 3^{rd} case: the candidate power edge can be directly added to the power graph. We then proceed with the next candidate in the list (line 39).

```
1:  Input: A graph G = (V, E)
2:  Output: A power graph Ġ = (V̊, Ė)
3:  Algorithm:
4:  ├ Initialize C and C′ to empty collections of node sets, and M to an empty matrix
5:  ├ For each node u in V, add to C and to C′ the singleton cluster {u}
6:  ├ Calculate for each pair (U, W) of clusters in C its neighborhood
       similarity s(U, W) and put it in the matrix M
7:  ├ While |C′| > 1:
8:  │ ├ Find one pair of clusters (U, W) with maximal similarity s_max from matrix M
9:  │ ├ Remove the two clusters U and W from C′
10: │ ├ Add the union of the two clusters U_new = U ∪ W to C and C′
11: │ └ Update neighborhood similarity matrix M: First, remove columns and rows
         of U and W. Second, calculate and add column and row for U_new
12: ├ For each cluster U in C: add to C the neighbor set N(U)
13: ├ Again, for each cluster U in C: add to C the neighbor set N(U)
14: ├ Initialize V̊ and Ė to empty sets, and L to an empty list
15: ├ Add for each node v in V a singleton set {v} to V̊
16: ├ For all unordered pairs (U, W) of node sets U and W in C:
17: │ ├ If U ∩ W = ∅ and if (U ∪ W, U × W) is a sub-graph in G:
18: │ │ ├ Add the power edge (U, W) to the list L
19: │ │ └ Compute for (U, W) its size: s(U, W) = |U||W|
20: │ └ If U = W and the U-induced graph in G is a clique:
21: │   ├ Add the power edge (U, W) to the list L
22: │   └ Compute for (U, W) its size: s(U, W) = ½|U|(|W| − 1)
23: ├ While list L is not empty:
24: │ ├ Sort the list L in descending order of power edge sizes s(U, W)
25: │ ├ Remove the first candidate power edge (U, W) from list L
26: │ ├ If the size of power edge (U, W) is two and if U = W then do nothing
27: │ ├ Else if there is a S in V̊ such that: U ∩ S ≠ ∅ but U ⊄ S and S ⊄ U:
28: │ │ └ Add to L the candidate power edges (U \ S, W) and (U ∩ S, W)
29: │ ├ Else if there is a S in V̊ such that: W ∩ S ≠ ∅ but W ⊄ S and S ⊄ W:
30: │ │ └ Add to L the candidate power edges (U, W \ S) and (U, W ∩ S)
31: │ ├ Else if there is a (S, T) in Ė such that: (U × W) ∩ (S × T) ≠ ∅:
32: │ │ └ If (S, T) covers not all edges of (U, W): ((U × W) ⊄ (S × T)):
33: │ │   ├ If U ⊂ S: Add to L the candidate power edge (U, W \ T)
34: │ │   ├ Else if U ⊂ T: Add to L the candidate power edge (U, W \ S)
35: │ │   ├ Else if W ⊂ S: Add to L the candidate power edge (U \ T, W)
36: │ │   └ Else if W ⊂ T: Add to L the candidate power edge (U \ S, W)
37: │ ├ Else if (U, W) is a clique (U = W):
38: │ │ └ Add power node U to V̊ and power edge (U, U) to Ė
39: │ └ Else: Add power nodes U and W to V̊ and power edge (U, W) to Ė
40: └ For each edge (u, v) not covered by any power edge in Ė
41:   └ add the singleton power edge ({u}, {v}) to Ė
```

Pseudo-code 1 **The power graph algorithm in detail.** The input is a graph $G = (V, E)$, and the output is a power graph $\dot{G} = (\mathring{V}, \dot{E})$. The first step (lines **4 to 13**) is the search for candidate power nodes. Hierarchical clustering on the set of nodes V is done using neighborhood similarity on node clusters. After line 13 the collection C contains these clusters, as well as for each cluster U its neighbors set and second-order neighbors set (added at lines **12 and 13**). The second step (lines **14 to 41**) is the greedy search for power edges. All cliques and bicliques induced by node sets in C are enumerated and their edge count is calculated (lines **14 to 22**). Power edges are then incrementally decomposed and eventually added to the power graph until all edges from G are covered by one and only one power edge.

3.6 Algorithm evaluation

We evaluate the power graph algorithm for minimality, scalability, robustness to noise, and time performance. First we evaluate the ability of the algorithm to reconstitute minimal power graphs from a manually designed benchmark of 15 minimal power graphs (page 89). Second, we evaluate the robustness of the algorithm to noise (page 91). Third, we examine the scalability of the algorithm when applied to dense networks (page 94). Forth, we give an empirical result on its time complexity (page 96).

3.6.1 Minimal power graph benchmark

We designed a benchmark of 15 power graphs that are minimal by design. First we explain the evaluation methodology and how we designed the power graphs of our benchmark. We then discuss the largest of these power graphs that was designed to test the robustness of the algorithm to deep and complex power node hierarchies. Finally we give the results and discuss the main result: the algorithm succeeds on 86% of the benchmark's power graphs.

Evaluation procedure. As shown in Fig. 59 we evaluate whether the power graph algorithm can reconstitute a power graph that has been designed to be minimal. First we unfold the power graph into its corresponding underlying graph, then this graph is given as input to the algorithm. We take a conservative approach and only consider two outcomes: it either succeeds or fails – the power graph is perfectly reconstituted or not. We did not find useful to quantify the discrepancies because in most cases the algorithm is able to perfectly recover the minimal power graph.

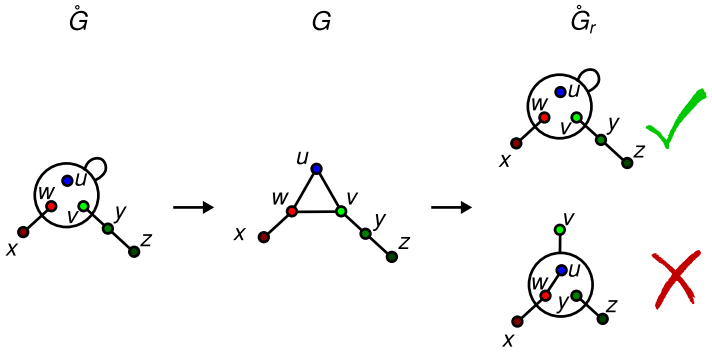


Fig. 59 **Evaluation procedure.** We unfold each power graph $\mathring{G}$ into its corresponding graph G. This graph is given to the algorithm as input and the resulting power graph $\mathring{G}_r$ is compared to the original power graph $\mathring{G}$. Two cases arise: if $\mathring{G}_r$ is identical to $\mathring{G}$ then we consider it a success, otherwise we consider that the algorithm fails for $\mathring{G}$.

Design of the benchmark set. We started from 14 small graphs that exhibit a variety of combinations of cliques, bicliques, stars, and single edges. For example, some of the chosen graphs had posed difficulties during the early phases of the algorithm's design. For each graph we manually searched for the minimal power

graph that best preserves the symmetries of the original graph and groups together nodes with similar neighborhoods (Fig. 60A to N). The first 14 power graphs are small (≤ 15) to make the manual verification of minimality tractable. In contrast, the last power graph (Fig. 60P) is large and was generated computationally to test the algorithm on deep power node hierarchies.

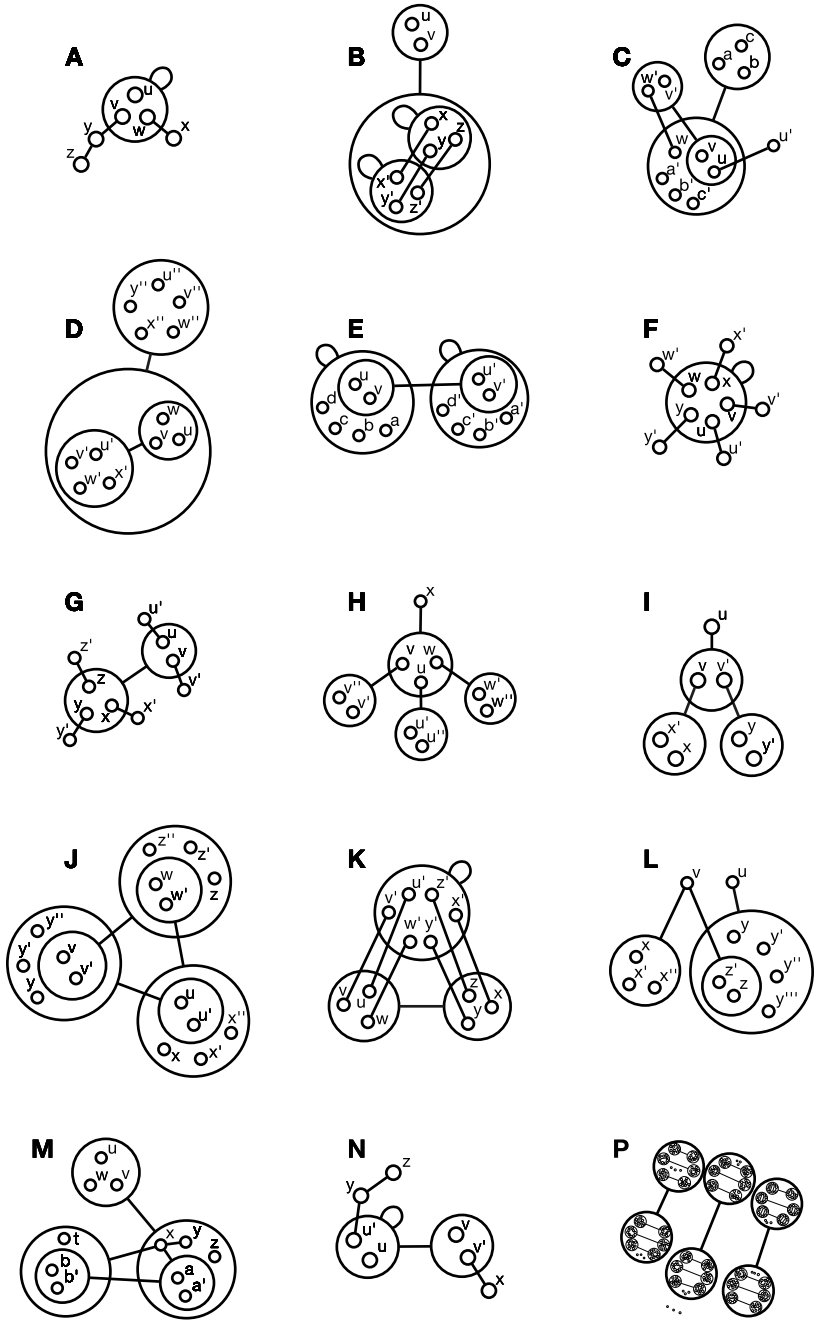


Fig. 60 **Benchmark set of minimal power graphs.** We test the power graph algorithm on 15 minimal power graphs. The algorithm succeeds in reconstituting 13 power graphs and fails on two: **B** and **I** (see Fig 62 for more details). **(P)** We also added to the benchmark a large power graph to test the robustness of the algorithm to deep power node hierarchies (see Fig 61 for more details).

Testing deep power node hierarchies. The last and the largest of the benchmark's power graphs (Fig. 60P) has a total of 387 power nodes and 981 power edges of which 129 are reflexive and 723 are single edges. It is recursively generated by adding power edges inside of power nodes (Fig. 61). The corresponding network has 1380 nodes and 182,505 edges. Remarkably the algorithm reconstitutes the power graph *exactly*. An important challenge is attaining minimality when generating such a large and complex power graph. While the first versions of the power graph algorithm could not reconstitute these power graphs, the current version can sometimes find more compact and elegant power graphs than a Human expert on smaller instances.

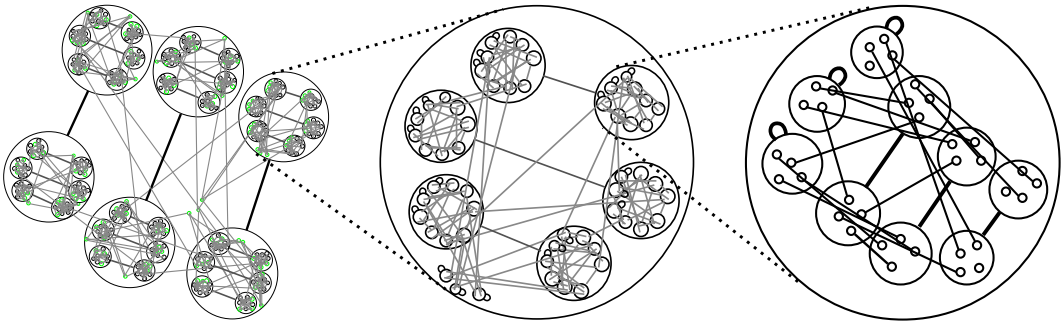


Fig. 61 **Testing the robustness of the algorithm to deep power node hierarchies** The last power graph in the benchmark set (Fig. 60P) is built by nesting three levels of power nodes. Each level is made of six distinct power nodes connected by three power edges, as well as three power nodes connected to themselves by a reflexive power edge. In addition, at each level single edges randomly link the cliques and bicliques with one another. Importantly, these edges never share a common node – this last point is important since it guarantees that the power graph is minimal.

Results. While the benchmark is limited in its scope – it is a small and biased selection – the results offer confidence that the power graph algorithm performs well in most cases: it perfectly reconstituted 13 out of 15 power graphs in the benchmark set (86%). As shown in Fig. 62, in two cases it fails. Failure typically happens for power edges between small power nodes (two to three nodes) that constitute a weak signal in the network.

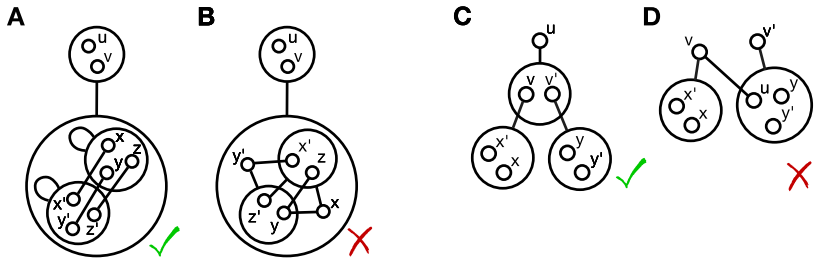


Fig. 62 **Two power graphs not recovered by the power graph algorithm.** The algorithm fails on two power graphs of the benchmark set (Fig. 60B and I). In both cases the errors can be traced to the breaking of some graph's symmetry. **(A)** The first power graph has a symmetric structure with two power nodes $\{x, y, z\}$ and $\{x', y', z'\}$ forming two bicliques in the underlying graph connected to another power node $\{u, v\}$. To complicate things, pairs of nodes in the cliques are linked by single edges: (x, x'), (y, y'), and (z, z'). **(B)** This last detail confuses the algorithm. The two cliques are not recognized and instead it finds a non-minimal solution: 7 instead of 6 power edges. **(C)** The second power graph is simply a balanced two level binary tree represented as three stars: $(u, \{v, v'\})$, $(v, \{x, x'\})$, $(v', \{y, y'\})$. **(D)** The algorithm fails to recognize the minimal power graph and instead groups u with y and y'.

3.6.2 Robustness to noise

Protein interaction graphs, false positives and false negatives. Because protein interaction networks suffer from false positives and negatives (Ito et al. 2001b; Deane et al. 2002; von Mering et al. 2002), we investigate the algorithm's robustness to noise. We compare power graphs of protein interaction networks before and after the addition, removal and rewiring of interactions.

Noise – false positive and false negative interactions. We chose a uniform noise model in which individual interactions are added, removed, or rewired (Erdős 1959). The noise level is the number of altered edges in proportion to the number of edges in the graph. A noise level of 0% means that no edges are added, removed or rewired. In the case of edge removal, a noise level of 100% means that all edges have been removed. In the case of edge addition, a noise level of 100% means that for each edge in the original graph there is an added edge. Finally, in the case of edge rewiring 100% means that every edge has been rewired.

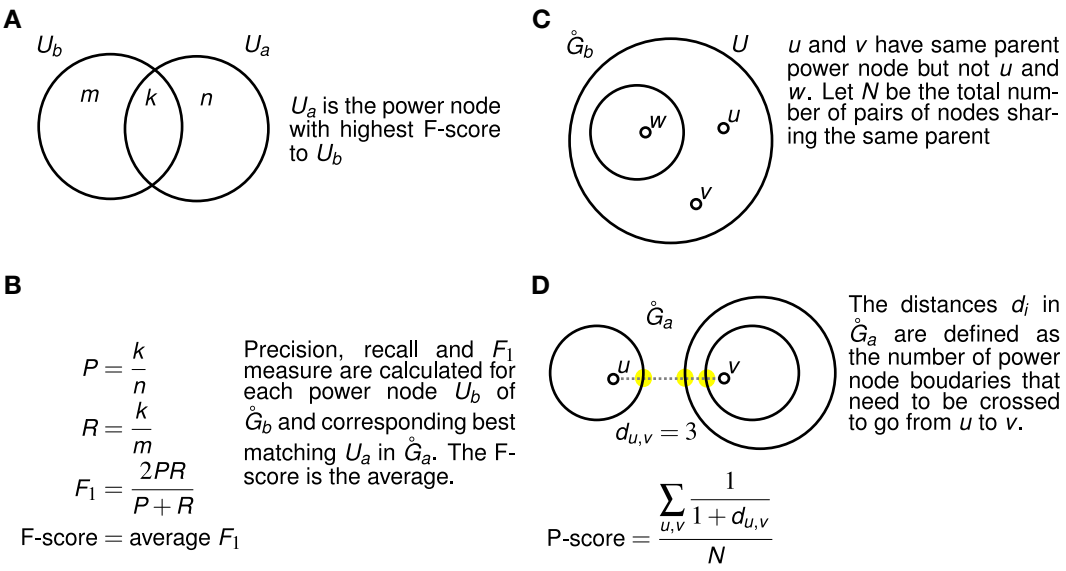


Fig. 63 **Definition of the F-score and P-score.** We use two different scores for comparing power graphs before and after the application of noise. **(A)** The F-score between the power graph before ($\mathring{G}_b$) and after ($\mathring{G}_a$) application of noise is calculated by matching each power node U_b of $\mathring{G}_b$ with the most similar power node U_a in $\mathring{G}_a$. **(B)** The F_1 measure compares the two sets U_b and U_a. The F-score is then defined as the average F_1 measure for all pairs (U_a, U_b). **(C)** The P-score measures how the distances between nodes in the power node hierarchy are affected by noise. In the following we *only* consider pairs of nodes (u, v) having same parent power nodes in $\mathring{G}_b$ – there distance is zero in the hierarchy. **(D)** In $\mathring{G}_a$, these nodes might not share the same parent anymore, we compute the distance $d_{u,v}$ as the number of power node boundaries that need to be crossed to go from u to v. From this distance we define the P-score (formula above) which is equal to 1 when all pairs of nodes still share the same parents ($d_{u,v} = 0$).

First evaluation method: F-score. We use two different methods to measure the effect of noise on the power node hierarchy. The first method – the F-score – is based on the F_1-measure. Let $\mathring{G}_b$ be the power graph before, and $\mathring{G}_a$ the power graph after the application of noise. For every power node U_b in $\mathring{G}_b$, we find the best matching power node U_a (Fig. 63A). The nodes in U_a are *predicting* the nodes in U_b. Precision and recall are calculated together with the F_1-measure – or harmonic mean between precision and recall (Fig. 63B).

Second evaluation method: P-score. The second method – the P-score – focuses on pairs of nodes and evaluates the extent to which nodes that have the same parent power node in $\mathring{G}_b$ remain close together in the power node hierarchy of $\mathring{G}_a$. Let two nodes u and v in $\mathring{G}_b$ have the same parent power node U (Fig. 63C). Let N be the total number of such pairs in $\mathring{G}_b$. Now consider their distance in the power node hierarchy of $\mathring{G}_a$. This distance $d_{u,v}$ is defined as the minimal number of power node boundaries that need to be crossed when going from u to v. This distance is converted into a similarity by application of the function $x \mapsto \frac{1}{1+x}$. The P-score between $\mathring{G}_b$ and $\mathring{G}_a$ is defined as the ratio between the sum of all $\frac{1}{1+d_{u,v}}$ and N (Fig. 63D).

Power graph analysis is robust to noise. Scatter plots of the F-scores and P-scores against the whole range of noise levels from 0% to 100% are shown in Fig. 64. The first observation is that – in a first approximation – there is a monotonous relationship between both scores and the noise level – indicating that power graphs degrade as more noise is applied. The scatter plots for the removal and rewiring of edges exhibit concavity (positive second derivative) – an indication that the effect of these two types of noise saturates from low to high levels of noise. For example, the graph by Gavin et al. (2006) has a strong F-score sensitivity to low edge rewiring levels – which is not the case of the SIN graph from Kim et al. (2006) (Fig. 64E). In contrast, for the addition of edges the plots are slightly convex, showing that power graphs are more resilient to low levels of edge addition than to removal or rewiring. For example, for the graph by Ito et al. (2001a) the P-score does not vary at low noise level (Fig. 64D). Overall the results show that there is no precipitous drop in the plots – power graph analysis is robust to unbiased noise models in which edges are removed, added and rewired.

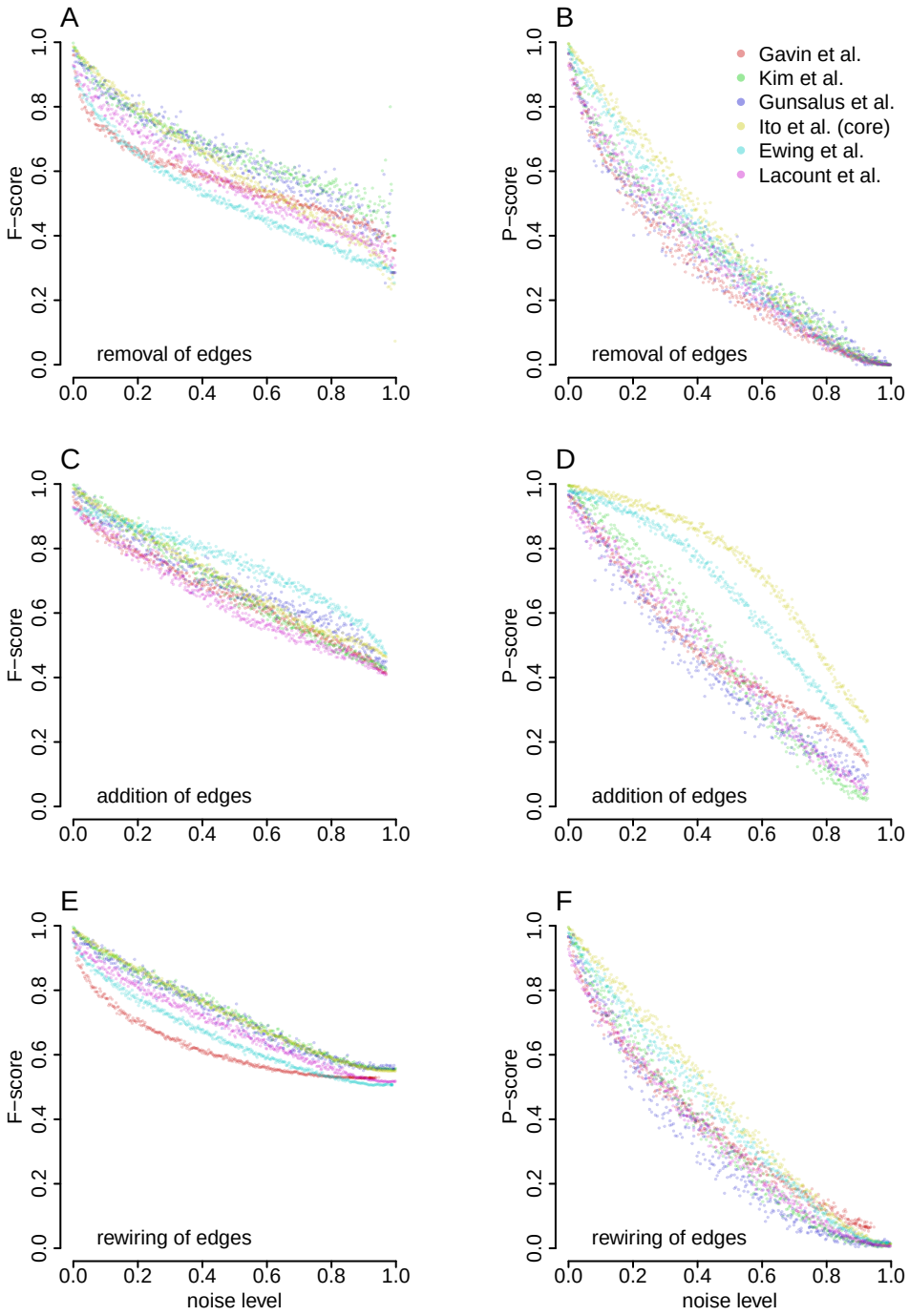


Fig. 64 **Robustness of power graph analysis to noise.** Comparison of power graphs of for six protein interaction graphs (Gavin et al. 2006; Kim et al. 2006; Gunsalus et al. 2005; Ito et al. 2001a; Ewing et al. 2007; LaCount et al. 2005) before and after application of noise. **(A, C, E)** Comparison based on the the F-score (Fig. 63A and B). **(B, D, F)** Comparison based on the the P-score (Fig. 63C and D). **(A, B)** Noise consists in the removal of edges, **(C, D)** addition of edges, **(E, F)** rewiring of edges.

3.6.3 Scalability

We conducted experiments to understand the behavior of the compression rate for high-density graphs of two important classes of synthetic random graphs: ER model

by Erdös and Rényi (1960) and synthetic scale-free graphs generated according to the preferential-attachment model of Barabasi and Albert (1999) (BA model).

Results. Fig. 65 shows how the compression rate behaves for the full range of edge densities and for three graph sizes: 150, 300, and 600 nodes. The edge density is the number of edges in the graph divided by the maximum number of edges: $\frac{n(n-1)}{2}$ where n is the number of nodes in the graph. A striking result is that in the general case and independently of the model, an affine relationship of the form $c = \frac{2}{3}e + \theta$ holds, where c is the compression rate, e is the edge density, and θ is a constant, dependent on both the model and number of nodes in the graph. For the same edge density, graphs generated according to the BA-model are in general more compressible by about 13% than graphs generated using the ER-model (a difference of about 0.13 in θ). For low edge densities this affine relationship does not hold anymore and the compression rate is then anti-correlated to the edge density. The compression rate reaches a minimum for an edge density between 0 and 0.2 and then steadily increases toward a compression rate of 1 for near-clique graphs of edge density close to 1. Increasing the number of nodes increases the affine model validity domain and shifts the curves down to lower compression rates.

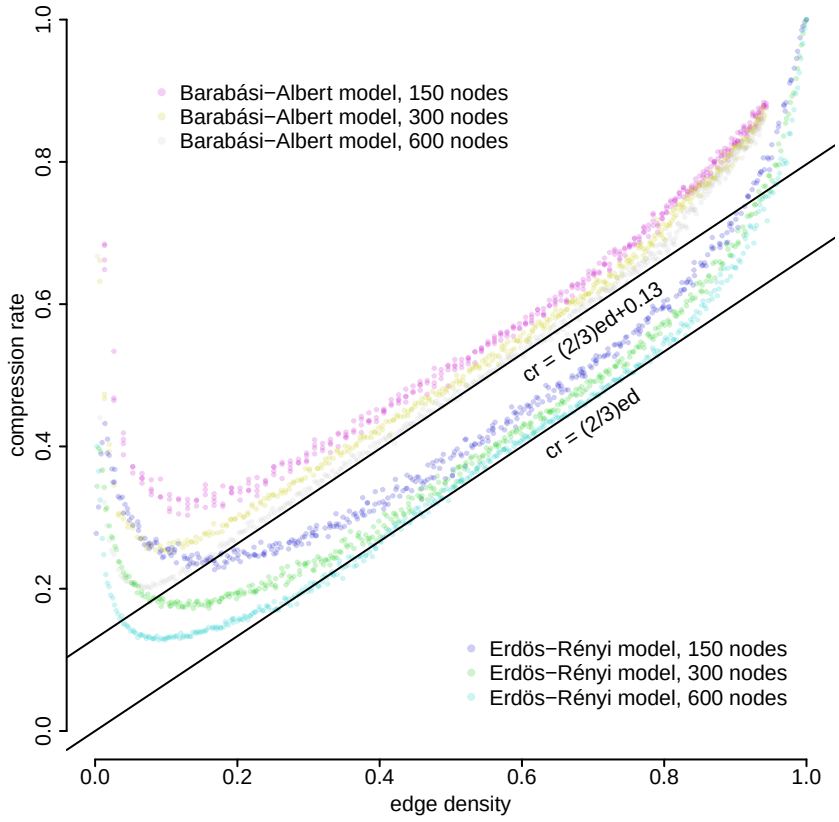


Fig. 65 **Scalability of power graph analysis, compression rate versus edge density.** Behavior of the compression rate for the full range of edge densities and for three graph sizes: 150, 300, and 600 nodes. The compression rate attains a minimum for an edge density between 0.1 and 0.2 and then increases linearly with a slope of $\frac{2}{3}$.

3.6.4 Time complexity of the power graph algorithm

In the following we address the question of how much time the algorithm needs to compute power graphs for graphs of different sizes? We will show that our implementation of the power graph algorithm admits a polynomial lower bound in the number of edges.

Empirical time complexity. The power graph algorithm is implemented in the Java language (Gosling and McGilton 1995). To evaluate the empiric time complexity of this implementation we collected graphs and ran the algorithm on a modern workstation (Quadricore Xeon 64 bit running at 2.67 GHz). A wide variety of graphs was used such as protein interaction graphs from the IntAct and BioGRID databases (Hermjakob et al. 2004; Stark et al. 2006), as well as random graphs generated according to the models by Erdös and Rényi (1960) or Barabasi and Albert (1999). Our first observation is that the duration of computation is mostly dependent on the number of edges and only marginally dependent on the number of nodes (Fig. 66). Fig. 67 shows that the relationship is almost quadratic: the time d in milliseconds needed to compute the power graph for a graph having e edges has the following tight lower-bound: $d \geq 0.00028\ e^{1.71} + 6$.

In both steps of the power graph algorithm the basic operations are defined on sparse sets, sparse vectors and sparse matrices. Operations such as intersections and unions of sparse neighborhood sets take an amount of time proportional to the number of neighbors – which in turn is dependent on the number of edges. This explains why the number of edges is the dominating factor.

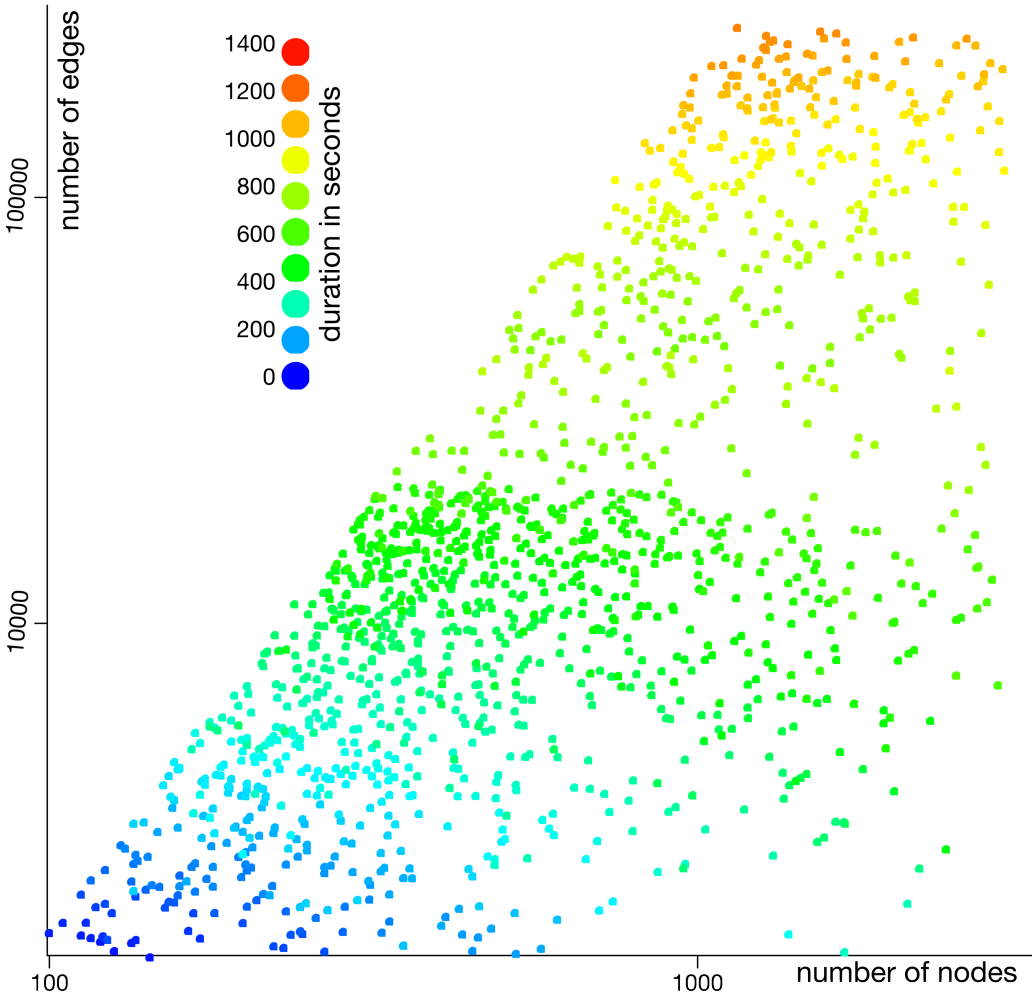


Fig. 66 **Time needed by the power graph algorithm on graphs with different number of nodes and edges.** The duration of computation is color coded – rainbow colors from blue, green to red represents values from 0 to 1,400 seconds. The upper-left half of the plot is empty because given a number of nodes n the number of edges is bounded by $\frac{n(n-1)}{2}$. An important observation is that the duration of computation is mostly dependent on the number of edges and only marginally on the number of nodes.

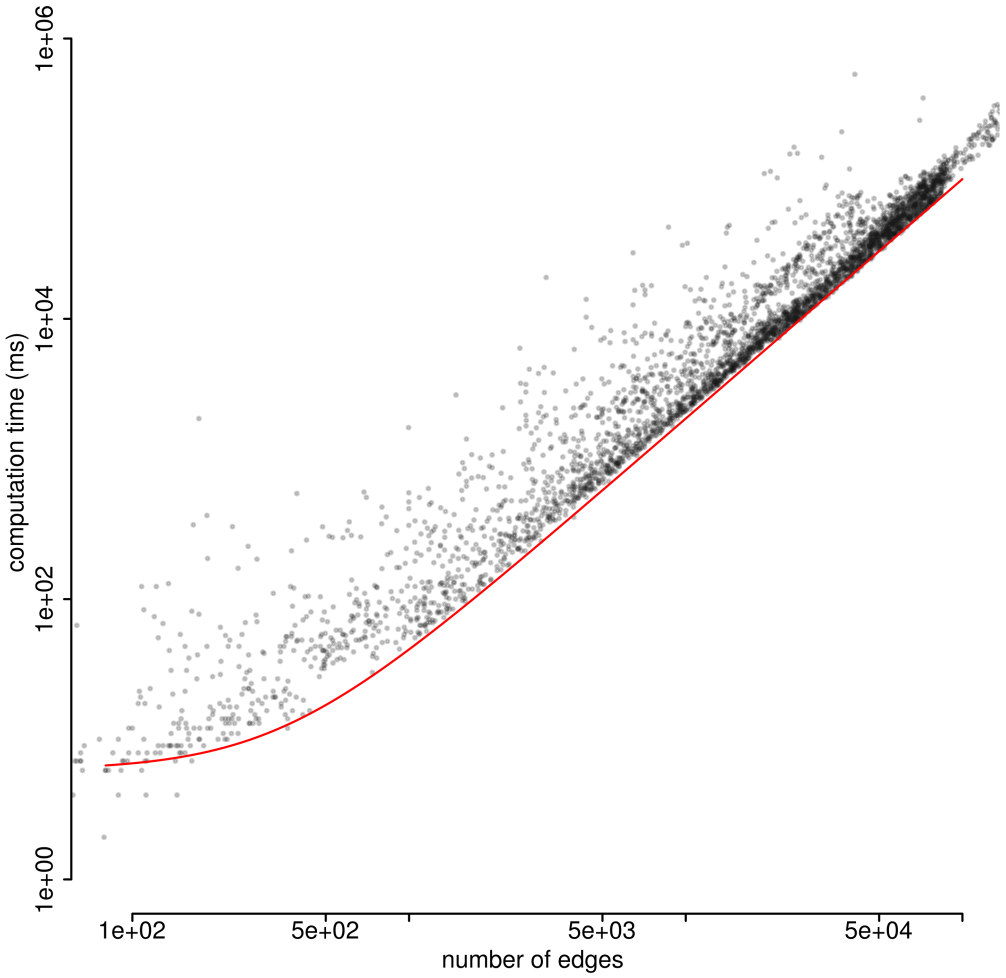


Fig. 67 **Computation time of power graph as a function of the number of edges.** Plot in log-log space of the number of edges and duration of computation for the same graphs as in Fig. 66. In red we plot the best fit for a tight lower-bound of the duration for a given number of edges. A fitted lower-bound model is: $d = 0.00028\ e^{1.71} + 6$ where e is the number of edges and d is the duration of computation in milliseconds. It is more reliable to look at the minimal duration because it attenuates the effect of occasional operating system interruptions as well as the residual dependence on the number of nodes.

3.7 Conclusion

"*Un bon croquis vaut mieux qu'un long discours*[4]."
Napoleon Bonaparte

Power graph analysis lies at the crossing point of clustering, network motif analysis, information compression, and visualization. With the previous examples and results we showed that power graph analysis reveals underlying biology when applied to protein interaction networks, regulatory and sequence similarity networks. It

4 A good sketch is better than a long speech

also leads to new insights and new hypotheses. In particular, we presented evidence that the similarity of interaction profiles for peptide-binding SH3 domains correlates with the sequence similarity of these domains (page 62). We also discussed how the difference of interaction profile – of otherwise near-identical histone subtypes – suggests that the TAP methodology interfered with the histone regulatory mechanisms, and led to low expression levels of histones subtypes HTA1 and HTB1 (page 63). Examining other types of networks, we showed that power graph analysis of a regulatory network by Beyer et al. (2006) led to the hypothesis that YAP7 is involved in metal detoxification (page 72). Finally, Power Graph Analysis, applied to a Human phosphatase sequence similarity network, reveals similarity cross-links in the hierarchy that are used to detect domain erosion in type 22 non-receptor protein phosphatases (page 74).

We have shown that the main reason behind the usefulness of power graph analysis is the observation that experimental protein interaction networks, bipartite regulatory networks, protein sequence similarity networks, and other biological networks have an abundance of cliques and bicliques. Cliques and bicliques have been previously noticed in biological networks (Morrison et al. 2006; Thomas et al. 2003; Li et al. 2006b; Pati et al. 2006). Here we argue that this abundance – which originates in the protein modularity and redundancy – constitutes a hallmark of their networks. The significant enrichment of power nodes in protein domains and GO terms further confirms that cliques and bicliques detected by power graph analysis are biologically relevant. In contrast to most graph clustering techniques, power graph analysis identifies these cliques and bicliques as carriers of a biological signal. Moreover, clustering algorithms on graphs often rely on the identification of highly connected regions. This approach works well for the detection of complexes and other regions of higher connectivity, but it fails for example in the case of bipartite regulatory networks: relevant clusters of transcription factors are not connected to each other but only to target genes. In protein interaction networks, relevant protein clusters are also defined by their neighboring proteins and not by their connectivity, as shown with the distinction between regulatory and enzymatic subunits of the casein kinase II complex.

We presented a fast and robust algorithm that can compute minimal power graphs by combining neighborhood similarity clustering for finding candidate power nodes with a greedy search for determining valid power edges. We have shown that the algorithm can reconstitute 86% of power graphs in our benchmark set of minimal power graphs (page 88). In addition we have shown that this algorithm scales well with networks of high density (page 94), and is robust to noise. We also investigated the time complexity of the algorithm and showed that it follows a sub-quadratic power law in the number of edges of exponent 1.71, and is marginally dependent on the number of nodes (page 96).

In the next chapter we will examine how power graph analysis can be used to evaluate the quality of protein interaction networks with the notion of *network compressibility*.

Chapter 4

Compressibility as a Novel Systemic Measure for Coverage and Accuracy of Protein Interaction Networks

4.1 Introduction

There is much debate about coverage and accuracy of genome-wide protein interaction networks. In the previous chapter we have shown that power graph analysis can be used to better understand the structure of protein interaction networks. Here we propose and validate network compressibility – computed with power graph analysis – as a novel measure for accuracy and completeness of genome-wide protein interaction networks. First, we verify the detrimental effect of false positives and false negatives. Second, we show that gold standard networks are highly compressible. Third, we show that authors' choice of confidence thresholds is consistent with high network compressibility. Forth, we present evidence that compressibility is correlated with co-expression, co-localization and shared function. Importantly, we also show that differences in network compressibility cannot be solely attributed to topological differences such as a lower average number of interaction partners or lower clustering coefficient. Examining the method specifics of affinity purification followed by mass-spectrometry and Yeast-two-hybrid screens we observe higher compressibility when using superior tagging methods, when maintaining physiological expression levels, and when employing smart-pooling strategies. Finally, we show that complete and accurate networks of complex systems in other domains exhibit similar levels of compressibility than current high quality interactomes.

Modularity, redundancy and cooperativity imply compressibility. We argue that the inherent cooperativity, modularity, and redundancy of molecular systems (Whitty 2008; Collins et al. 2007a) is reflected in their networks – leading to re-occurring patterns and motifs (Kashtan and Alon 2005). As explained in the previous chapter, protein interaction networks are compressible with power graph analysis because of the abundance of cliques and bicliques. As shown in Fig. 68, these patterns are caused by protein complex modularity and cooperativity, functional redundancy

and domain mediated interactions (Collins et al. 2007a; Whitty 2008; Breitkreutz et al. 2010). We expect these interaction motifs in high quality interactomes to produce a clear compressibility signal. Yet, network compressibility is reduced by false positives and false negatives.

We will show that network compressibility can be used to measure the network's content in patterns and motifs after subtracting compressibility that occurs by chance alone. This is reminiscent of the compressibility of genomic sequences due to the recurrence of similar sequences (Weiss et al. 2000; Herzel et al. 1994).

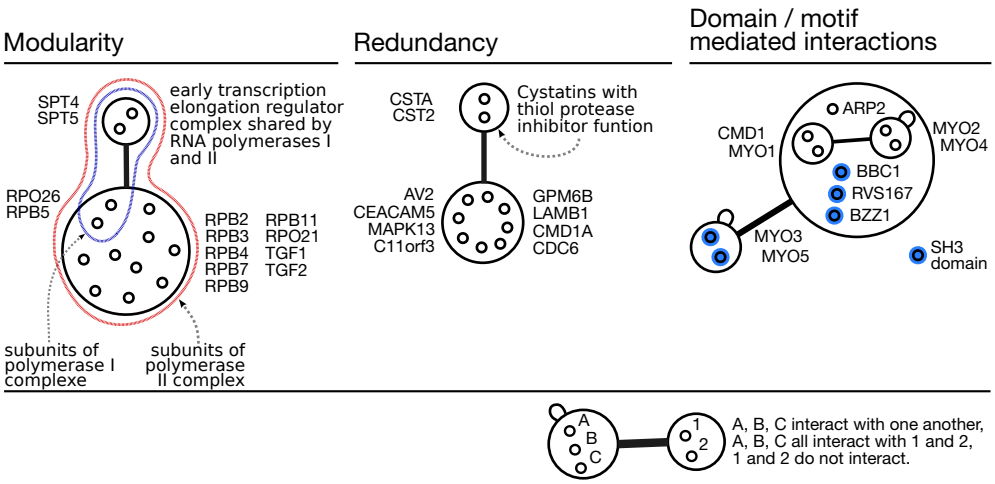


Fig. 68 **Modularity and redundancy in protein interaction networks.** Modularity is a hallmark of protein interaction networks (Gavin et al. 2006). In the network by Collins et al. (2007b) the proteins SPT4 and SPT5 have many common interaction partners. It forms the SPT4/SPT5 sub-complex – shared by both the polymerase I and II (Schneider et al. 2006) as well as complexes involved in mRNA capping and splicing (Lindstrom et al. 2003). Redundancy is seen, for example in the literature curated HPRD network (Prasad et al. 2009), as proteins of same function sharing interaction partners – here two thiol protease inhibitors. Domain and/or motif mediated interactions can overlap significantly as seen in the structural interaction network (SIN) in which two proteins, MYO3 and MYO5, have analogous interactions to 8 other proteins mediated by SH3 domains (Kim et al. 2006).

Entropy, compressibility, and Kolmogorov complexity. In computing, compression algorithms identify patterns in data and use these patterns to obtain compact representations, thus reducing data size. Lossless compression algorithms are reversible: the compressed representation is sufficient to recover the original data. In 1948, Shannon discovered a fundamental and unexceedable limit to lossless data compression based on the notion of entropy (Shannon 1948). Entropy is intrinsically dependent on the pattern statistics of the data. Following this first insight, Kolmogorov and Chaitin later generalized this notion and introduced program-size complexity as the length of the shortest program needed to specify data. As put forward by Chaitin: "to comprehend is to compress" (Chaitin 2007)

Chaitin's insight can be turned into an operational principle: compression algorithms can be used to analyze patterns and structure in data. For example, the information content of genomic sequences has been investigated in several studies (Weiss

et al. 2000; Herzel et al. 1994). It was applied to alignment-free sequence comparison by conditional Kolmogorov complexity (Li et al. 2001), and to protein sequence classification (Kocsor et al. 2006).

Similarly, there have been several attempts to quantify the information content of networks.

Estimating network entropy with compression algorithms. Early on, Rashewsky (1955) and Mowshowitz (1968) proposed to calculate the information content of graphs using Shannon's entropy formula. More recently, a definition of network entropy based on topology configuration was used to segregate random network models (Ji et al. 2008). Another definition based on local vertex functionals (Dehmer and Emmert-Streib 2008) was introduced with the goal of efficiently computing the entropy of large chemical graphs. Graph entropy has also been used to characterize the resilience and robustness of protein interaction networks (Demetrius and Manke 2005; Manke et al. 2006). All these definitions of network entropy rely on the simple idea that the more diverse the node neighborhoods, the higher the network entropy. For example, a network in which all nodes have nearly the same neighbors has a low entropy whereas a network for which all nodes have different neighborhoods will have a high entropy (Sun et al. 2008). If two nodes in a network have nearly the same neighbors then they are also nearly exchangeable – to recover the original network few interactions need to be rewired. This implies that the amount of information nescessary to encode both neighborhoods is less than the sum of that needed to encode each of them. This highlights the link between symmetry in networks and compressibility. The more a network has symmetries the more it is compressible. Recently, MacArthur et al. (2008) showed that 'real-world' complex networks are richly symmetric – much more than standard network models predict. This result suggests that compressibility can be used to characterize complex networks – a result that will be directly confirmed in this chapter. An indirect approach for measuring network entropy is to measure data size after applying a compression algorithm. The power graph algorithm can be considered a compression algorithm for graph and therefore an algorithm for measuring graph entropy. Similarly to power graph analysis, most approaches for graph compression exploit neighborhood similarity, non-uniform network motif statistics, and the scale-free property of complex networks (Lu 2002; Feder and Motwani 1995; Kao et al. 1998; Deo and Litow 1998; Randall et al. 2002; Boldi and Vigna 2004; Langville and Meyer 2004; Hannah et al. 2008).

Entropy, compressibility, and relative compressibility. Instead of measuring the entropy which varies according to the network's data size, we consider the entropic ratio. In network compression terms, the entropic ratio is the *absolute compression rate* of the network. For the sake of simplicity we define the absolute compression rate as the proportion of edges after compression compared to the number of edges before (see methods section for details and in depth discussion). For example, a compression rate of 70% means that among 100 edges in the original network, only 30 edges remain after compression. The compressibility of a network can also be measured relative to a random network model. We define the *relat-*

ive compression rate as the difference between the compression rate of a network and the compression rates of topologically equivalent random networks (see methods section for details). In the following *compressibility* will implicitly refer to *relative compression rate*.

In the following we give a four point validation of network compressibility as a quality measure for protein interaction networks.

4.2 Validation

First we validate the link between relative compression rate and network quality. We then compare the relative compressibility of all genome-wide interactomes and discuss how assay parameters such as protein expression level, tagging, and pooling strategies influence the networks' relative compressibility. Importantly, we show that relative compressibility is independent of the network topology such as average clustering coefficient and number of interaction partners. Finally, we verify that networks derived from completely and accurately known complex systems are compressible at levels similar to the best interactomes.

4.2.1 Validation 1 – False positives and false negatives decrease network relative compressibility

If relative compressibility measures the fidelity of the networks to the systems they represent, then the relative compression rate should deteriorate with the addition of noise to networks. Noise can be applied by randomly adding interactions – introducing false positives (FP) or by randomly removing interactions – introducing false negatives (FN). We consider two models for noise in protein interaction networks. In the Erdős–Rényi model (ER), the choice of interactions is independent of the network topology and all possible interactions are equally likely to be selected for addition or removal (Erdős 1959). In contrast, in the Barabási-Albert model (BA), the scale-free topology is preserved (Barabasi and Albert 1999). It is assumed that false positives are more likely for highly connected proteins ("the rich get richer") while false negatives are more likely for poorly connected proteins ("the poor get poorer"). This gives a total of four combinations: FN/ER, FP/ER, FN/BA, FP/BA which were applied to 12 Yeast networks (5 Y2H, 3 AP/MS, 1 PCA, 2 literature, 1 structure) adding and removing up to 60% of interactions. As shown in Fig. 69, we find that false positives and false negatives decrease the relative compression rates of networks – independently of the system from which the network is derived and independently of the model considered for false positives and false negatives. Thus, low sensitivity and low specificity implies low relative compression rate. Furthermore, relative compressibility decreases linearly with the increase of noise. For example, for the Collins network, each additional 2% of false positives or false negatives leads to a 1 percentage point decrease in relative compressibility.

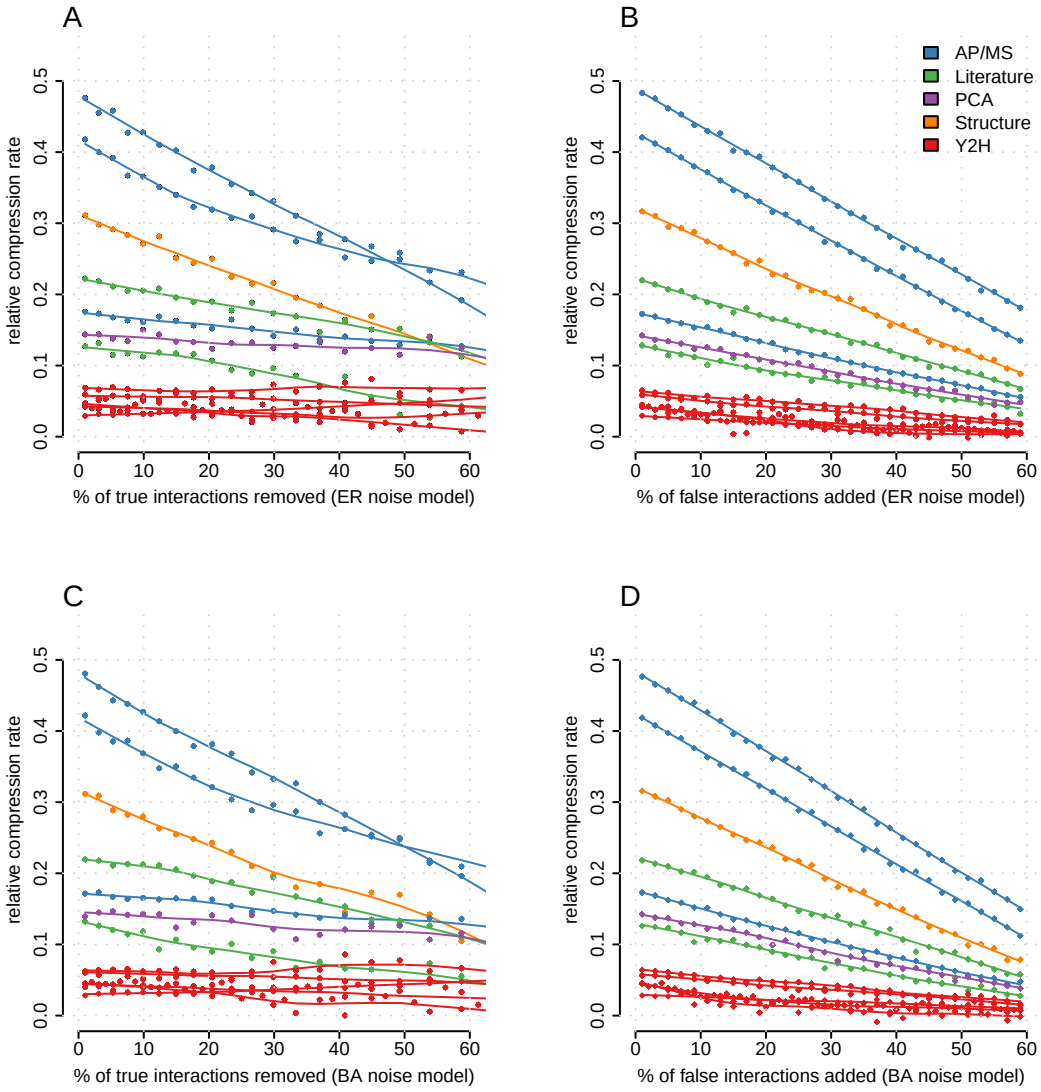


Fig. 69 **(A and B) Effect of removal/addition (ER model) of interactions on the relative compression rate in 12 Yeast networks.** In order to validate the relationship between network quality and relative compressibility, we investigate the effect of false positives and false negatives on the relative compression rate for up to 60% removed/added interactions. Independently of the experimental system or network topology, both false positives and false negatives consistently reduce the relative compression rate when the proportion of added or removed interactions is increased. **(C and D) Effect of removal/addition of interactions on the relative compression rate in** 12 **Yeast networks with the BA noise model.** Inspired by the Barabási-Albert preferential attachment model of network growth, we investigate the effect of false positives and false negatives biased towards highly connected proteins and lowly connected proteins, respectively. Therefore, the scale-free network topology is preserved and "interaction-rich proteins get richer and interaction-poor proteins get poorer". As for the random (ER) noise model, we observe that independently of the experimental system or network topology, both false positives and false negatives consistently reduce the relative compression rate. While both models give similar curves, the BA model decreases the relative compression rate by an additional 5% for high noise levels (60%).

4.2.2 Validation 2 – Relative compression rates correlate with published interaction confidences

Published interactomes are reported as binary interactions, i.e. either two proteins interact or not. Underlying these data are confidence scores – authors define a threshold and only report interactions above the threshold. Defining such a threshold is a difficult compromise since a conservative threshold may improve precision but lowers the coverage, while a generous threshold achieves the opposite effect. Thus, the threshold controls the amount of false positives and false negatives in the network and the question arises of how is this reflected in the compression rates. To answer this question we systematically analyzed the networks of Gavin (TAP/MS), Tarassov (PCA), and Parrish (Y2H) and computed the compression rates for networks defined by interactions above a minimum and below a maximum confidence score (see Fig. 70A). The results for all three networks is given in Fig. 70. First, we note that complete networks – lowest minimum and highest maximum – are not necessarily the most compressible. Second, with the exception of the network by Parrish, the most compressible sub-networks include the interactions of highest confidence. Moreover, including interactions of low confidence consistently decreases the relative compressibility of the corresponding sub-networks.

Gavin network (TAP/MS). Remarkably, for Gavin's network, the highest relative compression rate is found for a minimum confidence score (socio-affinity index) of 5 – a threshold recommended by the authors. We also observe the detrimental effect of both false negatives and false positives when imposing excessively high minimum or low maximal thresholds to the data: keeping only interactions with a score above 15 leads to similarly low relative compression rates as keeping only interactions with a score below 5.

Tarassov network (PCA). For Tarassov's network we find that the highest relative compressibility is found for a minimum score of 4 and a maximal score of 7. However, most sub-networks with high maximum thresholds have similar compressibilities (between 0.15 and 0.2) unless the minimum threshold is set too high (above 5). In agreement with this observation the authors choose to include most lower confidence interactions with a minimum threshold of 2.5. Interactions with a score above 5 form less network motifs and thus the sub-networks are lowly compressible. Yet, including these interactions together with interactions with a score above 4 gives more compressible sub-networks than without – indicating that these interactions belong to structures formed for slightly lower confidences.

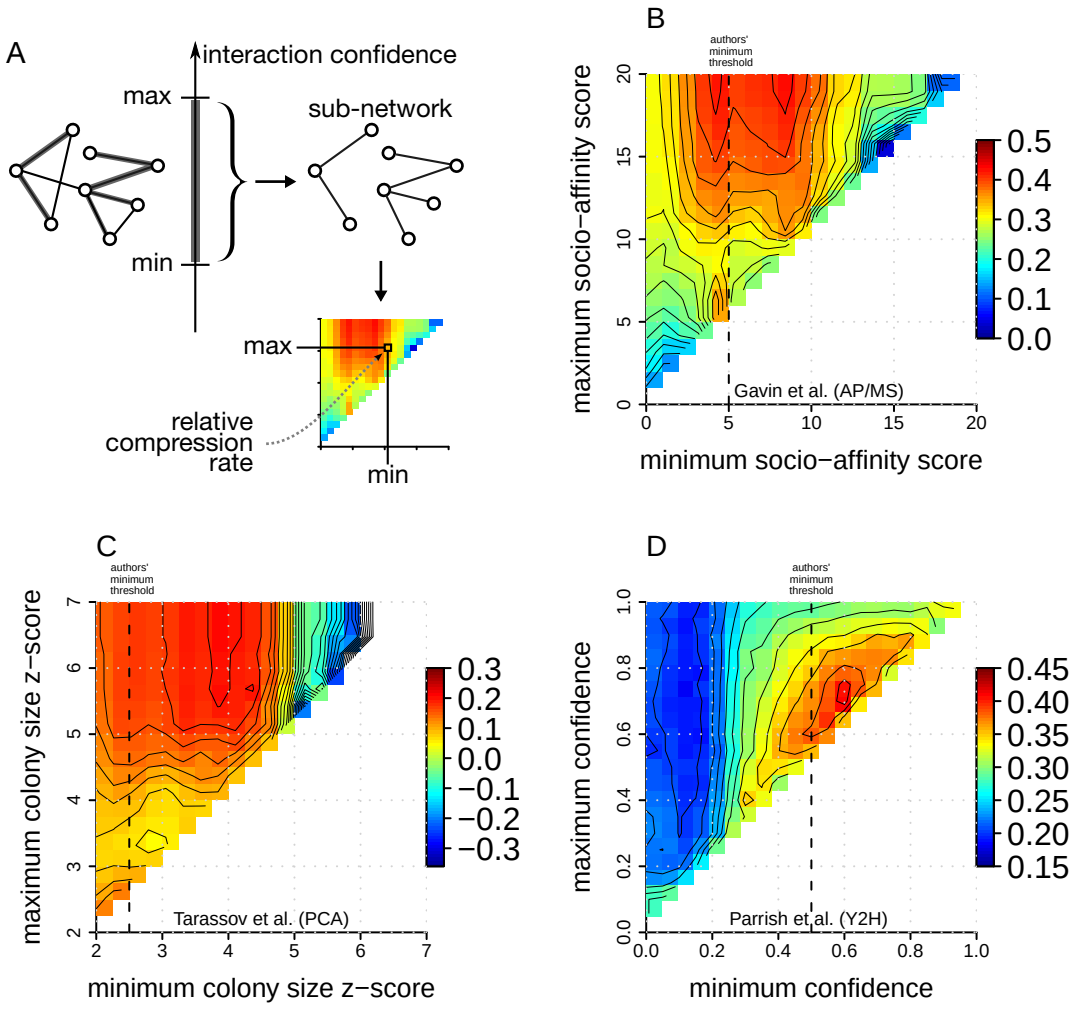


Fig. 70 **Correlating interaction confidence scores with relative compressibility. (A)** Measuring the relative compression rate of sub-networks obtained by slicing networks for different ranges of confidence scores. The color of each cell indicates the relative compression rate of each sub-network and the vertical dotted lines indicate the authors' choice of minimum confidence thresholds. **(B)** For Gavin's network we observe that the sub-network with the most interactions and the highest relative compression rate is found for a minimum socio-affinity score of 5 and a maximum of 20. This is in agreement with the minimum of 5 recommended by Gavin et al. – interactions with a lower score have reproducibility of less than 70% (Gavin et al. 2006). **(C)** For Tarassov's network we find that the highest relative compression rates are found for a minimum z-score of 4 and a maximal z-score of 7. However, lower confidence interactions do not significantly decrease the relative compressibility of the subnetworks – at most 2% relative compressibility points are lost when including lower confidence interactions (z-score from 2 to 7). This is in agreement with the relatively generous threshold of 2.5 used by the authors on the colony size z-score. **(D)** Parrish's network we observe low relative compressibility for sub-networks containing low confidence interactions (minimum < 0.3). In contrast to the Gavin and Tarassov networks, the highest relative compression rate is not found when including high confidence interactions. Instead, it is found for a sub-network with confidences between 0.6 and 0.7 which agrees with the author's threshold of 0.5 between high and low quality interactions.

Parrish network (Y2H). For Parrish's network we observe that interactions with confidence scores below 0.3 form sub-networks with low relative compression rates. In particular, we find that the sub-networks with lowest relative compression rates are found for a minimum of 0.10 and maximums below 0.75 – which indicates that interactions with a confidence around 0.1 are detrimental to relative compressibility. This is in agreement with the analysis by Parrish et al. (2007) which shows that interactions with a confidence of about 0.15 have the highest proportion of false positives. This is estimated from a training set of likely true positives and true negatives – see Fig. 2a in Parrish et al. (2007). Moreover, the peak in relative compression rate is found for a minimum threshold of 0.6, in agreement with the author's confidence threshold of 0.5 separating high from low confidence interactions. The value of 0.6 is in fact closer to the confidence score for which functional homogeneity between interacting proteins becomes significant – see Fig. 2c in Parrish et al. (2007). Surprisingly, and in contrast to the Tarassov and Gavin networks, interactions of *very* high confidence (above 0.7) are detrimental to the relative compressibility. These high confidence interactions do not fit together with the other high-confidence interactions (above 0.5).

4.2.3 Validation 3 – Author's gold standard datasets have highest relative compression rate

The network by Collins et al. (2007b) is a merge and re-analysis of the raw data from the Gavin and Krogan datasets aimed at improving coverage and reducing false positives. We observe that this dataset has a higher relative compression rate (48%) than both original datasets interpreted with the plain spoke model (Gavin 22% and Krogan 18%). This is in agreement with the author's assessment which showed that their consolidated dataset has a higher functional homogeneity than the Gavin or Krogan datasets – see Fig. 2 in Collins et al. (2007b).

Yu et al. (2008) compared their novel experimental dataset (CCSB-YI1) and their own merge of several datasets (Y2H-Union) to a gold standard of binary interactions derived from literature (CCSB-binaryGS). We find that this recent gold standard dataset has a higher relative compression rate (13%) than all Yeast Y2H datasets.

Ito et al. (2001a) discouraged the use of the Ito full dataset and instead recommended the use of only Ito core. We observe that the Ito core network has a slightly higher relative compression rate (of 2 percentage points). Since Ito full has the same if not a greater coverage than Ito core, we can assume that the difference in relative compression rate is attributable to false positives.

Similarly, false positive estimates by Lemmens et al. (2010) correlate with relative compressibility: the Stelzl dataset achieved the highest MAPPIT-retest success rate of 31% and also has a higher relative compression rate (20%) compared to the datasets from Rual (4%), Yu (CCSB-YI1, 6%), and Simonis (5%) – see Fig. 2 in Lemmens et al. (2010).

4.2.4 Validation 4 – Compressibility correlates with co-expression, co-localization and shared function

Assortativity in protein interaction networks refers to the preference of proteins to interact to other proteins that are similar or share certain properties (M. E. J. Newman 2003). It has been previously proposed as a means of evaluating network quality when applied to gene co-expression, functional similarity, cellular localization, and phylogenetic profile similarity (von Mering et al. 2002). Fig. 71A shows that the relative compression rate is highly correlated to the proportion of co-expressed genes pairs corresponding to interacting proteins (Spearman $\rho = 0.90$). There is a weaker correlation with function (Fig. 71B, $\rho = 0.65$) and with co-localization (Fig. 71B, $\rho = 0.67$), but only a weak correlation to phylogenetic profile similarity (Fig. 71D, $\rho = 0.43$). Several interesting observations can be made: First, gold-standard dataset CCSB-binaryGS (Yu et al. 2008) is consistently in the top 3 networks having higher relative assortativity ratios (Fig. 71A, B, C, and D). Second, Tarassov's dataset has the highest co-localization assortativity ratio – which is consistent with the fact that the PCA method is unique in that it detects *in-vivo* protein interactions within a 8 nanometer distance (Tarassov et al. 2008). Third, Ito full is the worst network for relative compressibility as well as for network assortativity. Forth, Collins network has consistently both higher assortativity and higher relative compressibility than the Gavin or Krogan datasets.

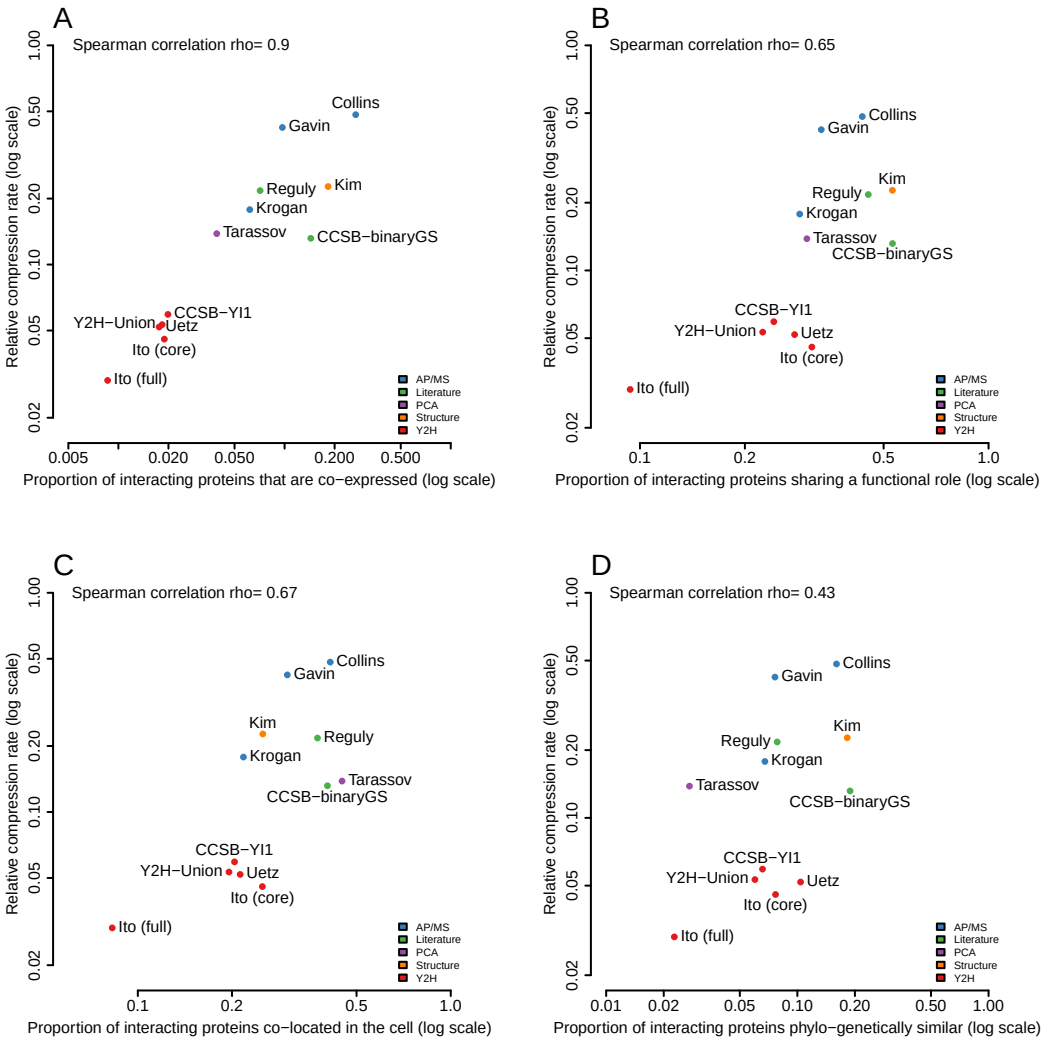


Fig. 71 **Correlation of the relative compression rate with gene co-expression, functional similarity, cellular localization, and phylogenetic profile similarity for 12 Yeast networks.** For all interacting pairs of proteins for which we have information about both, we compute the proportion – or assortativity ratio – of interacting proteins that are significantly co-expressed, share a cellular function, are found in at least one common cellular compartment, and have similar phylogenetic profiles. We normalize these ratios by subtracting the average proportion found for equivalent randomized networks similarly to the relative compression rate. **(A)** Relative compression rate versus relative proportion of interacting proteins that are co-expressed. The Spearman correlation ($\rho = 0.90$) is the highest of all four studied correlations. **(B)** Relative compression rate versus relative proportion of interacting proteins that share at least one functional role. **(C)** Relative compression rate versus relative proportion of interacting proteins that share at least one cellular localization. **(D)** Relative compression rate versus relative proportion of interacting proteins that have similar phylogenetic profiles. The low Spearman correlation ($\rho = 0.43$) indicates a poor correspondence between relative compression rate and shared evolution.

To summarize, the above four validation points substantiate our claim that higher network compressibility indicates corresponding higher network quality – understood as encompassing both coverage and accuracy. Next, we will discuss in detail how

the different experimental methods influence the relative compressibility of available genome-wide interactomes.

4.3 Analysis of all available interactomes

Relative compression rates of all genome-wide interactomes. Overall, we observe that Y2H networks are significantly less compressible. Table 7 lists the relative compression rates for *all* 21 genome-wide interactomes (13 Y2H, 7 AP/MS, 1 PCA), 5 entire databases (BioGRID, IntAct, DIP, MINT, HPRD), 2 literature curated networks and 1 structural interactome. AP/MS datasets are interpreted using the 'spoke' model thus preventing clustering effects (except Collins, see materials and methods section). To prevent a bias in the selection of datasets we defined a strict criteria for what constitutes a *large-scale genome-wide screen* (see methods section). Fig. 73 shows a plot of relative compression rates versus absolute compression rates for these networks. Absolute compression rates range from 30% to 70% and relative compression rates from 1% to 48%. Overall, we observe that Y2H networks are on average 6 times less compressible than all other networks. AP/MS networks have on average a relative compression rate of 21%, whereas it is 7% for Y2H networks. T-tests confirm that the relative compression rate of Y2H is significantly different from PCA, SIN, and literature curated networks ($p = 0.002$) and from AP/MS ($p = 0.01$). Fig. 72 shows that the maximal achieved relative compression rate has been increasing with time, indicating that progress in the methodologies is leading to networks with increasing richness in patterns and structure.

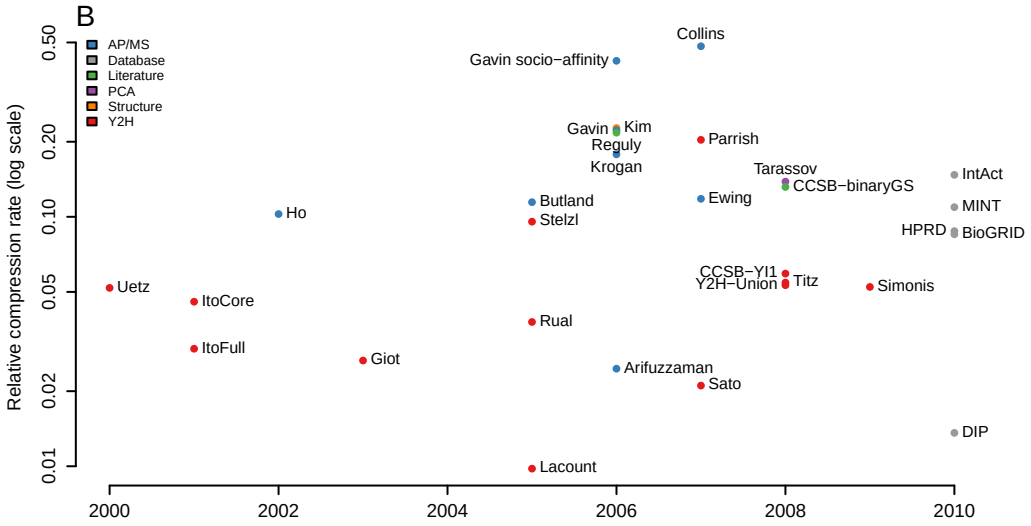


Fig. 72 **(B) Relative compression rates along time.** Progress has been made with higher relative compression rates achieved in recent years.

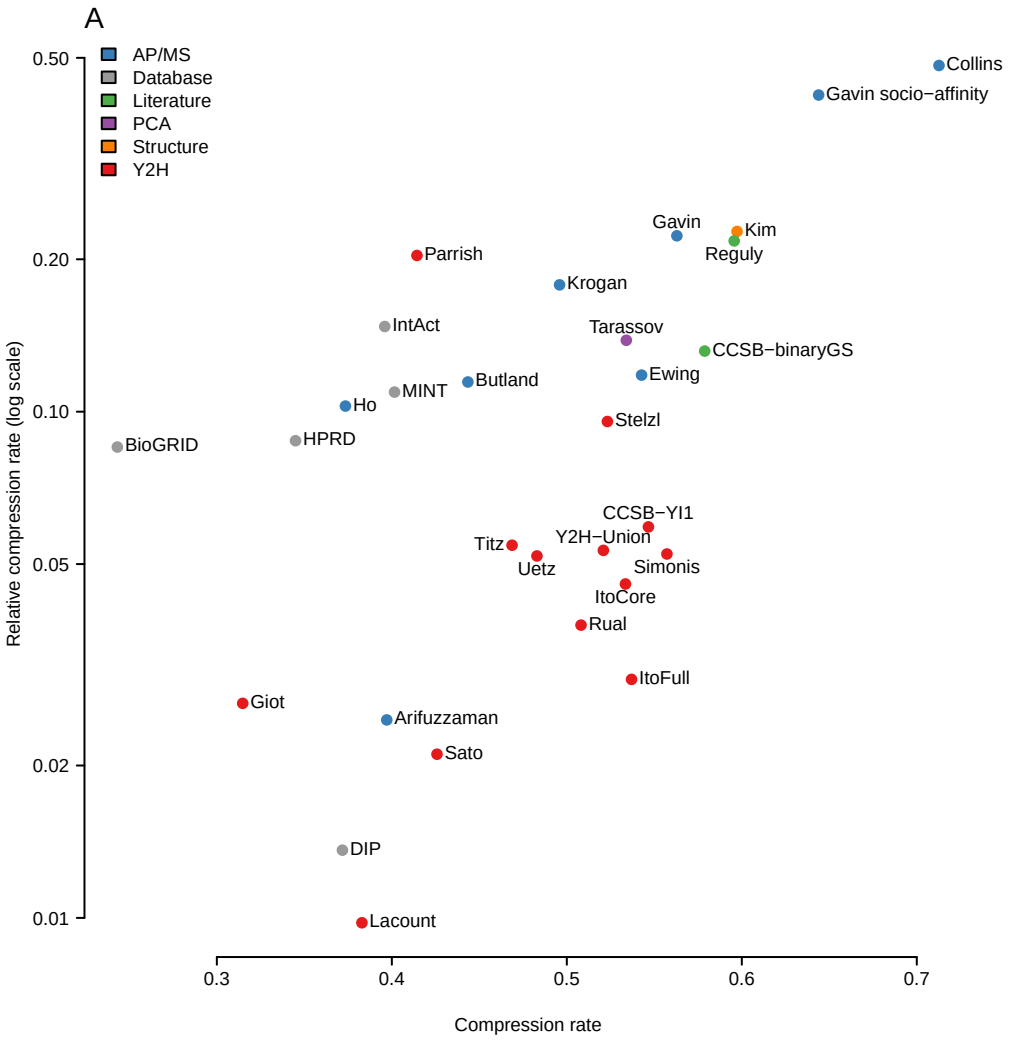


Fig. 73 **Compression rates and relative compression rates.** Relative compression rates plotted against compression rates for several types of large-scale networks: Y2H, AP/MS, PCA, and literature networks. Y2H networks have in general lower relative compression rates than AP/MS, Literature, Structure or PCA derived networks. More details are given in Table 7. *Important note*: by default all AP/MS datasets are interpreted using the spoke model. For the Gavin dataset we also add the network derived from socio-affinity scoring.

Average signal. To investigate the "average signal" of all available interactome data we computed the relative compression rate of all protein interaction data available in the multi-species databases: IntAct, MINT, BioGRID, and DIP. Most of these database averages cluster around a relative compressibility of 11% – with the exception of the DIP database which has a lower relative compressibility of 1.3%. One explanation is that DIP contains less large-scale genome-wide datasets (as defined in the methods section). The DIP database covers 8 large-scale datasets whereas MINT covers 11, IntAct covers 20, and BioGRID covers 18. Moreover, it covers fewer small-scale datasets (3,609 publications) than BioGRID (22,645 publications), IntAct (4,247 publications), and a similar number to MINT (2,942 publications).

Table 7 **Detailed information for figure 73.** The complete list of protein interaction networks analyzed is given together with the species, system, publication year, compression rate, relative compression rate, number of nodes and edges, average number of interaction partners (avg. num. of int. partners) and PubMed identifier of publication for referencing (Tarassov et al. 2008; Kim et al. 2006; Prasad et al. 2009; Reguly et al. 2006; Yu et al. 2008; Chatr-Aryamontri et al. 2008; Collins et al. 2007b; Gavin et al. 2006; Krogan et al. 2006; Ewing et al. 2007; Butland et al. 2005; Arifuzzaman et al. 2006; Parrish et al. 2007; Stelzl et al. 2005; Titz et al. 2008; Yu et al. 2008; Rual et al. 2005; Ito et al. 2001a; Sato et al. 2007; Uetz et al. 2000; LaCount et al. 2005; Ho et al. 2002; Simonis et al. 2009; Giot et al. 2003). Networks by Formstecher et al. (2005), Rain et al. (2001), and Li et al. (2004) are excluded because they are not comparable to the other networks – they are highly asymmetric (see methods section for detailed information on how the networks were compiled). Important note: by default all AP/MS datasets are interpreted using the spoke model, but for the Gavin dataset we also add the network derived from socio-affinity scoring.

author	species	system	year	compression rate	relative compression rate	number of proteins	number of interactions	average degree	PubMed Id
Collins et al.	Yeast	AP/MS	2007	0.71	0.48	1622	9070	11.18	17200106
Gavin et al. (socio-affinity)	Yeast	AP/MS	2006	0.64	0.42	1462	6942	9.50	16429126
Gavin et al. (spoke-model)	Yeast	AP/MS	2006	0.56	0.22	1386	3244	4.68	16429126
Krogan et al.	Yeast	AP/MS	2006	0.50	0.18	2708	7123	5.26	16554755
Ewing et al.	Human	AP/MS	2007	0.54	0.12	2294	6449	5.62	17353931
Butland et al.	E. coli	AP/MS	2005	0.44	0.11	1277	5324	8.34	15690043
Ho et al.	Yeast	AP/MS	2002	0.37	0.10	1693	8038	9.50	11805837
Arifuzzaman et al.	E. coli	AP/MS	2006	0.40	0.02	2457	8663	7.05	16606699
Aranda et al. (IntAct)	292 species	Database	2010	0.40	0.15	46011	162082	7.05	4,247 publ.
Ceol et al. (MINT)	332 species	Database	2010	0.40	0.11	29407	77954	5.30	2,942 publ.
Prasad et al. (HPRD)	Human	Database	2010	0.34	0.09	9463	35021	7.40	453,521 publ.
Breitkreutz et al. (BioGRID)	15 species	Database	2010	0.24	0.09	29499	229471	15.56	22,645 publ.
Salwinski et al. (DIP)	230 species	Database	2010	0.37	0.01	20685	58596	5.67	3,609 publ.
Reguly et al.	Yeast	Literature	2006	0.60	0.22	1536	2844	3.70	16762047
Yu et al. (CCSB-binaryGS)	Yeast	Literature	2008	0.58	0.13	1090	1263	2.32	18719252
Tarassov et al.	Yeast	PCA	2008	0.53	0.14	1507	3030	4.02	18467557
Kim et al. (SIN)	Yeast	Structure	2006	0.68	0.22	1178	2195	3.72	17185604
Parrish et al.	C. jejuni	Y2H	2007	0.41	0.20	1326	11659	17.59	17615063
Stelzl et al.	Human	Y2H	2005	0.52	0.10	1664	3083	3.71	16169070
Yu et al. (CCSB-YI1)	Yeast	Y2H	2008	0.55	0.06	1278	1641	2.57	18719252
Titz et al.	T. pallidum	Y2H	2008	0.47	0.05	724	3627	10.02	18509523
Yu et al. (Y2H-Union)	Yeast	Y2H	2008	0.52	0.05	2018	2705	2.68	18719252
Simonis et al.	C. elegans	Y2H	2009	0.56	0.05	1515	1748	2.31	19123269
Uetz et al.	Yeast	Y2H	2000	0.48	0.05	806	644	1.60	10688190
Ito et al. (core)	Yeast	Y2H	2001	0.53	0.05	813	761	1.87	11283351
Rual et al.	Human	Y2H	2005	0.51	0.04	1527	2529	3.31	16189514
Ito et al. (full)	Yeast	Y2H	2001	0.54	0.03	3243	4367	2.69	11283351
Giot et al.	D. melanogaster	Y2H	2003	0.31	0.03	6988	20240	5.79	14605208
Sato et al.	Synechocystis	Y2H	2007	0.43	0.02	1915	3100	3.24	18000013
Lacount et al.	P. falciparum	Y2H	2005	0.38	0.01	1272	2643	4.16	16267556

4.3.1 Y2H with two-phase pooling has best compression

First introduced by Fields and Song (1989), the Yeast two-hybrid system (Y2H) is a widely used technique for protein interaction testing. Applying Y2H for genome-wide interactome mapping raises scalability challenges which have been addressed with three approaches: library screens, matrix screens, and the recent smart-pooling screens such as two-phase pooling (Zhong et al. 2003).

Table 8 shows that the two most compressible networks – Stelzl and Parrish – were derived using two-phase pooling Y2H screens, the first having a lower screen-

ing completeness than the second. Parrish's network was derived from Campylobacter jejuni, a species with a small genome (1643 coding sequences), and 80% of all proteins were present as baits and preys. A screening completeness of 80% $\times$ 80% was achieved – 64% of all protein pairs where screened for interaction. In contrast, Stelzl et al. (2005) searched a sizable but smaller fraction (9%) of the much larger Human interactome search space (300 times larger). This observation suggests that already sensitive screens can deliver interactomes richer in patterns and motifs, if the quadratic size of proteomes can be overcome.

Table 8 **Strategies for Y2H screening.** There are three main strategies for large-scale genome-wide Y2H screens, briefly: i) matrix – all bait-prey pairs are tested, ii) library – preys are pooled and growing colonies are picked and then sequenced, and iii) two-phase pooling – preys are pooled in a first phase and in a second phase baits that reported interactions are pooled and screened against individual preys (see Zhong et al. (2003); Jin et al. (2006); Xin et al. (2009) for reviews). In the Parrish screen pools group 96 preys compared to 8 for Stelzl. The *screening completeness* is the proportion of the whole interactome search space that was accessible to the screen: $\frac{nb \times np}{n^2}$ where *nb* is the number of ORFs cloned for baits, *np* is the number of ORFs cloned for preys and *n* is the estimated number of protein coding genes. In practice, the assay and sampling sensitivity of Y2H screens greatly diminish the effective completeness (Venkatesan et al. 2009). For that same reason, screening completeness should not be misconstrued with assay sensitivity – for which the average number of interaction partners is a better indicator.

datatset	species	strategy	number of protein coding genes (ORFeome)	screening completeness	avg. num. of interaction partners	relative compression rate
Stelzl et al.	Human	two-phase pooling (8)	22,286	5%	3.7	10%
Parrish et al.	C. jejuni	two-phase pooling (96)	1,685	79%	17.5	20%
Titz et al.	T. pallidum	matrix	1,028	79%	10.0	5%
Rual et al.	Human	library	22,286	10%	3.3	4%
Simonis et al.	C. elegans	library	20,185	24%	2.3	5%
Giot et al.	D. melanogaster	library	14,144	60%	5.7	3%
Yu et al. (CCSB-YI1)	Yeast	library	5,797	81%	2.5	6%
Ito et al. (core)	Yeast	library	5,797	90%	1.8	5%
Uetz et al.	Yeast	library	5,797	85%	1.6	5%
Lacount et al.	P. falciparum	library	5,268	84%	4.1	1%
Sato et al.	Synechocystis	library	3,569	27%	3.2	2%

Lower sensitivity of library-based Y2H screens. The lower sensitivity of library based Y2H screens is also apparent if one examines the average number of interaction partners. Depending on the database – IntAct, BioGRID, Mint, HPRD or DIP – the average number of interaction partners per protein can be roughly estimated to be between 5 and 15. Published estimates similarly range around 5 and 8 (Grigoriev 2003). Interestingly, most library-based Y2H screens exhibit lower values than other strategies. For example, the Titz dataset was derived using the matrix approach for Y2H screening – all bait and prey pairs are tested individually – a potentially more sensitive strategy than library screens (Zhong et al. 2003). Similarly, two-phase pooling also seems to favor more interaction partners per proteins and thus can be deemed more sensitive.

Overall, Table 8 suggests that differences in relative compressibility between Y2H networks can be partly explained by the different screening strategies and their sens-

itivities. In contrast, screening completeness has a weaker influence on the relative compression rate than the overall effective sensitivity after taking assay and sampling sensitivity into account (Venkatesan et al. 2009).

4.3.2 AP/MS with knock-in and TAP-tagging has best compression

As Table 9 shows, one AP/MS network – Arifuzzaman et al. (2006) – has a low relative compression rate of 2% which is below the average for both Y2H and AP/MS datasets. It is also the only screen that uses both cDNA over-expression and the His-tag system instead of maintaining the physiological expression by knock-in tagging (von Mering et al. 2002), and achieving high purity by tandem affinity purification (TAP) (Gavin et al. 2002). We also observe the higher relative compression rate of Krogan or Gavin (knock-in) versus Ho (cDNA over-expression) in Yeast; and the higher relative compression rate of Butland (knock-in) versus Arifuzzaman (cDNA over-expression) in E. coli. More generally, the two expression modes can be distinguished by the relative compression rate of the corresponding networks (T-test with p-value below 5%).

Table 9 **Expression modes and tagging systems for AP/MS screening.** The Arifuzzaman dataset is an outlier when compared with other AP/MS datasets. A possible explanation is that it is the only screen that combined both non-physiological protein expression and His-tagging instead of the superior tandem purification procedure. Note: by default AP/MS datasets are interpreted using the spoke model. In addition we list the Gavin network derived by socio-affinity scoring (scores above 5). The Collins dataset relies on the same experimental data as the Krogan and Gavin datasets and is derived by a method similar to socio-affinity (Gavin et al. 2006).

datatset	species	expression modes	purification method	number of protein coding genes (ORFeome)	screening completeness	relative compression rate
Collins et al.	Yeast	physiological expression (knock-in)	TAP	5,797	80%	48%
Gavin et al. (socio-affinity)	Yeast	physiological expression (knock-in)	TAP	5,797	78%	42%
Gavin et al.	Yeast	physiological expression (knock-in)	TAP	5,797	78%	22%
Krogan et al.	Yeast	physiological expression (knock-in)	TAP	5,797	76%	18%
Butland et al.	E. coli	physiological expression (knock-in)	TAP/SPA	4,263	23%	11%
Ewing et al.	Human	over-expression (cDNA)	FLAG-tag	22,286	1%	12%
Ho et al.	Yeast	over-expression (cDNA)	FLAG-tag	5,797	10%	10%
Arifuzzaman et al.	E. coli	over-expression (cDNA)	His-tag	4,263	61%	2%

The above results show that experimental methods (AP/MS versus Y2H, pooling strategy, expression level, tagging) strongly influence relative compressibility. Next we show that organism complexity, network topology and under-sampling play a lesser role.

4.3.3 Relationship of relative compressibility with organism complexity, network topology and under-sampling

Organism complexity and relative compression rate. Table 8 and 9 show that differences between methods (two-phase pooling versus library, and physiological versus over-expression) have a stronger influence on the relative compression rate than differences in organism complexity as estimated by the ORFeome size. For example, for Y2H networks (Table 8), library screens have relative compression rates around 3% and differ in average by 2 percentage points from each other – independently of the species. In contrast, two-phase pooling screens have higher relative compression rates – above 10%. This shows that any species specific signal is probably hidden by a much stronger method specific signal.

Influence of the network topology on relative compressibility. On average Y2H networks have less interaction partners than AP/MS owing to the experimental method. Therefore, one reason for low relative compression rates in Y2H could be the low average number of interaction partners. However, Fig. 74A shows that the SIN (Kim), PCA (Tarassov), Stelzl, and literature curated networks have similarly low average number of interaction partners and yet have significantly higher relative compression rates. The same argument holds true for the clustering coefficient. Networks with low clustering coefficients but high relative compression rates exist (Ho, Ewing, Butland, Stelzl). We also observe that the clustering coefficient does not separate Y2H networks from other types of networks as well as does the relative compression rate (Fig. 74B). Indeed, lowly clustered networks can have high relative compression rates because the compression rate captures network motifs based on cliques *and bicliques*. Therefore, bipartite networks that do not contain a single clique – and thus have a clustering coefficient of zero – may still exhibit the whole range of compression rates. While both average number of interaction partners (average degree) and clustering coefficients are slightly correlated with the relative compression rate, these correlations do not constitute an explanation for the whole variability of the relative compression rate ($\rho = 0.31$ and 0.55 respectively). Moreover, Fig. 75 shows that the relative compressibility is largely independent of network size.

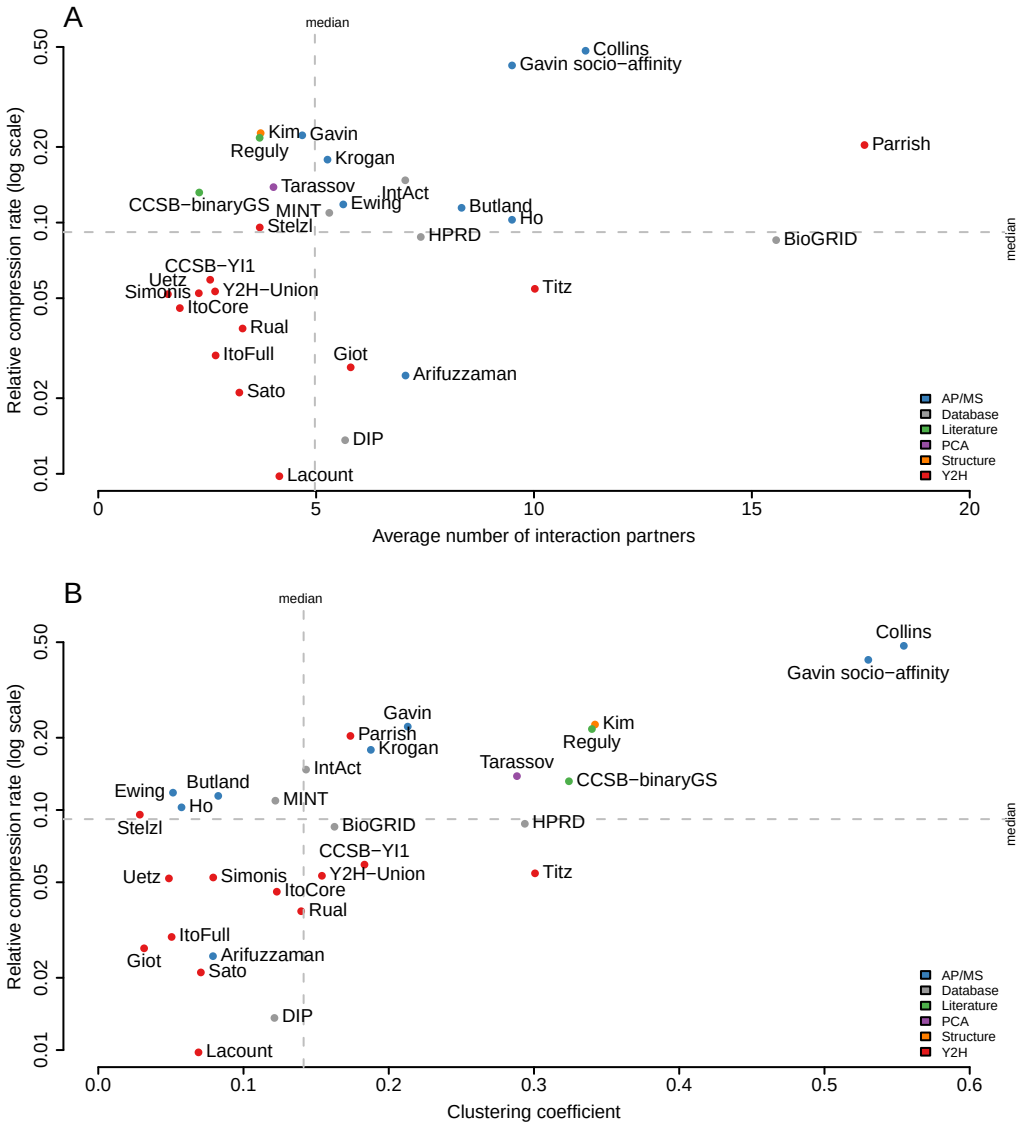


Fig. 74 **(A) Low average number of interaction partners is no reason for low relative compression rates.** While low relative compression rates imply low average number of interaction partners, low average number of interaction partners does not imply low relative compression rates. Note that the CCSB binary interaction gold standard (CCSB-binaryGS) has a similar average number of interaction partners as most Y2H networks and yet it has a higher relative compression rate. **(B) Relative compression rate versus clustering coefficient.** Similarly to the average number of interaction partners, we observe that a low clustering coefficient does not imply a low relative compression significance. For example, the Lacount dataset has a similar clustering coefficient (0.07) to the Butland dataset (0.08), and yet they differ in relative compression rates (11% difference). We also observe that the relative compression rate is better than the clustering coefficient at discriminating different screening methodologies.

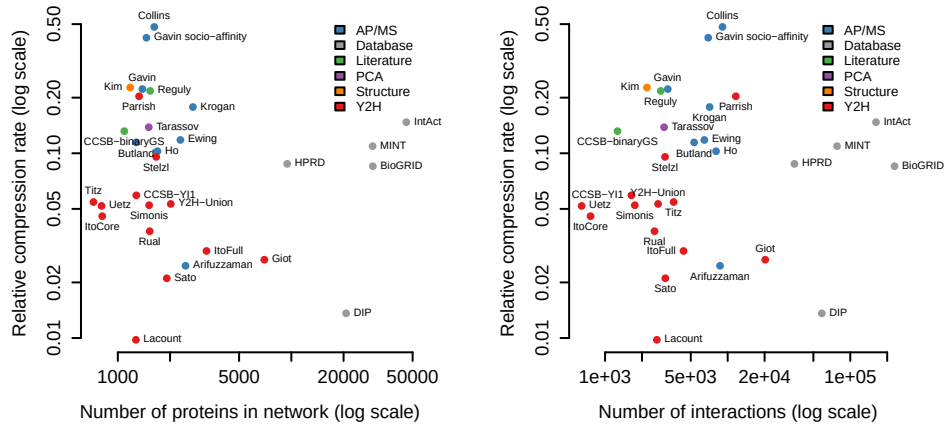


Fig. 75 **Relative compression rate versus the number of proteins and interactions in the networks.**

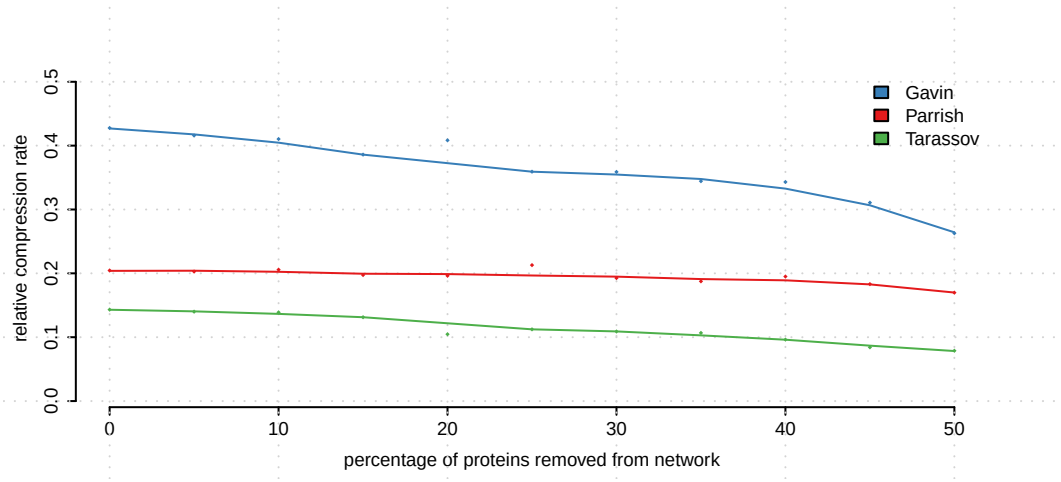


Fig. 76 **Influence of under-sampling on the relative compression rate.** The relative compression rate decreases slowly when proteins – and all their interactions – are removed from networks (compare to Fig. 69). For example, removing half of the nodes in the Parrish network decreases its relative compression rate by just 3 points. This shows that the effect of under-sampling is not as strong as the effect of false positives and negatives.

Effect of under-sampling on the relative compression rate Coverage in protein interaction networks is affected by false negatives but also by *under-sampling* – also termed *screening completeness*. Table 8 (Y2H) and Table 9 (AP/MS) show no clear correlation between screening completeness and relative compression rates. The strong link between compressibility and the experimental method (pooling strategy for Y2H and expression/tagging in AP/MS) hides any potential correlation. From a theoretical point of view, a strong effect is not expected since non-trivial cliques and bicliques are robust to random node removal. In practice, we observe the same behavior: we removed up to 50% of nodes from three networks – Gavin, Tarassov and Parrish – and observed that the relative compressibility marginally decreases even if

up to 50% of nodes are removed (Fig. 76). For Parrish's network it remains relatively constant. Comparing these results to those of Fig.69 leads us to the conclusion that under-sampling has less impact on the relative compressibility than false positives or false negatives.

4.3.4 How compressible are complete and accurate complex networks?

In the absence of at least one complete and accurate interactome map it is difficult to estimate the range of true relative compression rates. In particular, an important question is whether some of the high relative compression rates – above 30% – are a sign of an excess of repetitive patterns and motifs due to systematic errors in the data. To address this point, we compare the relative compressibility of current interactomes with that of accurate and complete networks derived from complex systems of interacting entities. Fig. 77 shows the same plot as Fig. 73A but overlaid with networks such as the C. Elegans neural network, Internet, network of North American airports, software module dependency in Java and CytoScape, and others (see methods for complete list and details). We observe that all complex systems networks have a relative compression rate of at least 15% and on average 25%. There is one exception: the north American power grid has a relative compression rate of just 5%. From manual inspection of the different networks, we reached the conclusion that a possible explanation is the network's planarity: it is the only one in which the entities and their interactions are strongly constrained in two dimensions. In the other networks the interacting entities are embedded in higher dimensional spaces and have more freedom to interact – a characteristic shared with protein interaction networks. Fig. 77 suggests that a relative compressibility between 15% and 50% is a signature of networks derived from complex systems whose structure is completely and accurately known. Similar levels of relative compressibility are expected for complete and accurate protein interaction data.

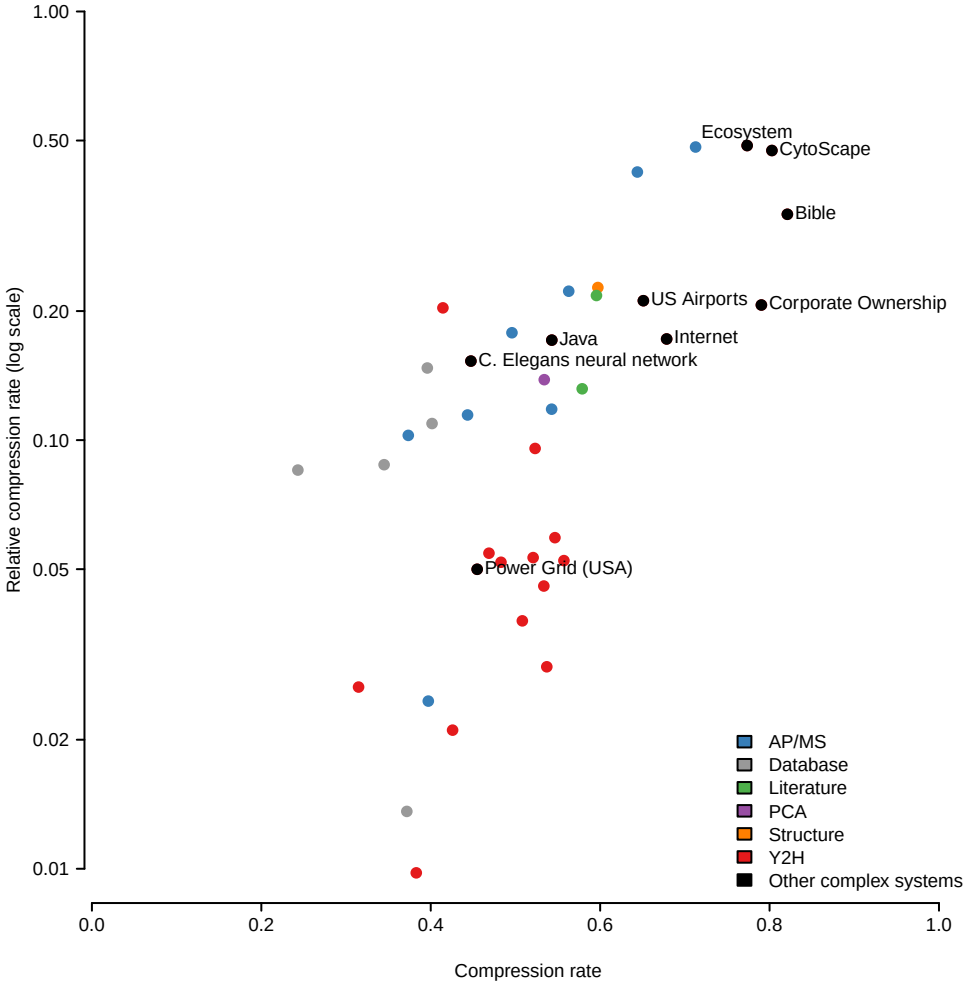


Fig. 77 **Comparing protein interaction networks with the accurate and complete networks of other complex systems.** In order to estimate the relative compression rate of true and complete interactome maps we computed the relative compression rates of a wide range of networks derived from complex systems from ecology, neuroanatomy, software engineering, and the Internet.

4.3.5 Example – zooming into chromatin remodeling complexes

As argued by Lima-Mendez and van Helden (2009), global properties of networks are an average that hides much detail. Therefore, let us consider the patterns underlying compressibility in more detail.

Richness in network motifs. Fig. 78A-D shows the size and number of motifs obtained from selected networks plotted as disc charts. The number and size of each disc represents the abundance of cliques and bicliques of different sizes. Networks with a high relative compression rate (AP/MS, SIN, PCA; Fig. 78A,B,D) are rich in cliques and bicliques involving many proteins, whereas networks of low relative compression rate (Y2H, Fig. 78C) are depleted.

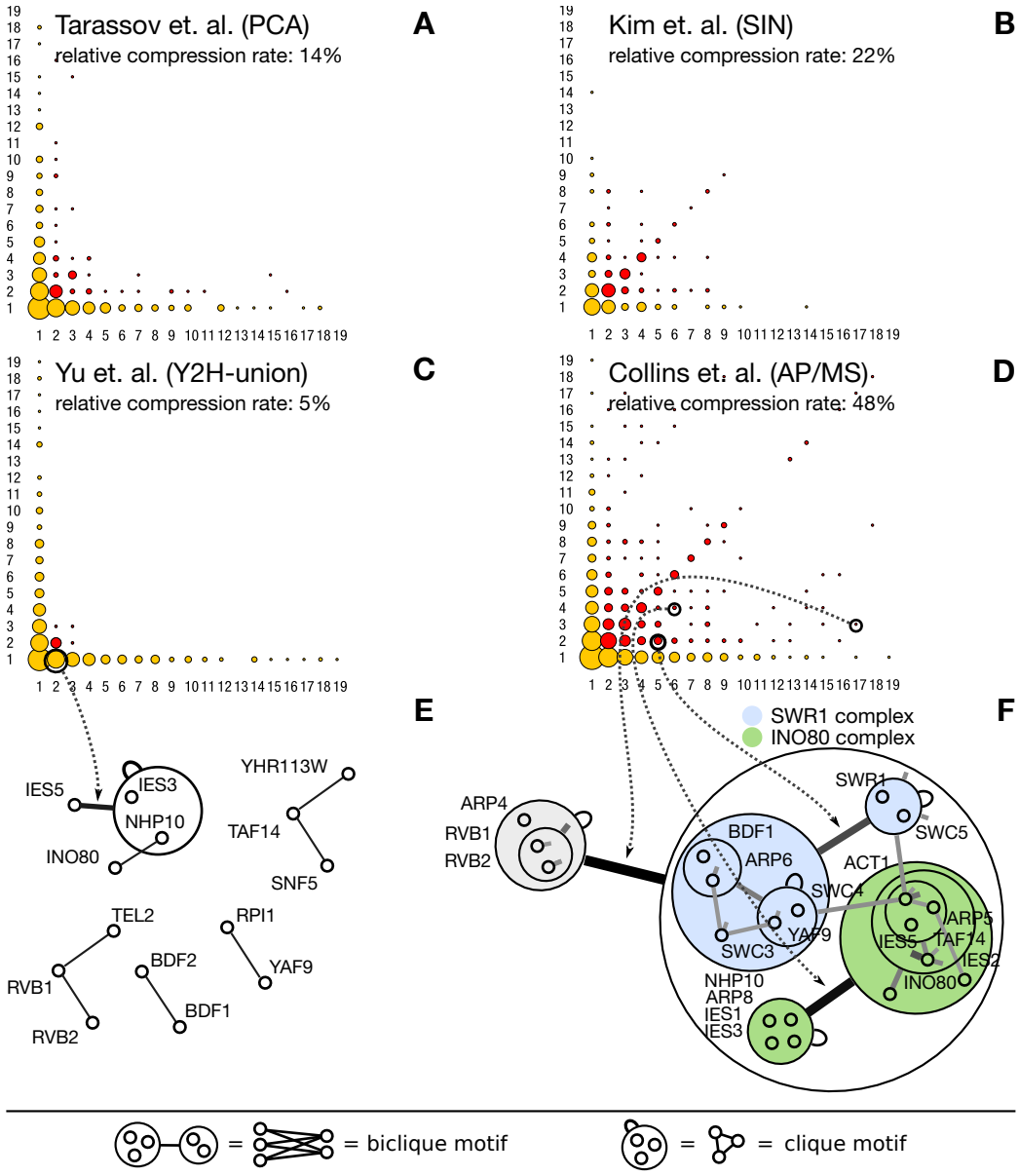


Fig. 78 **High relative compression rate explained by the richness in network motifs. (A-F)** The disc charts show the distribution of network motifs – bicliques, cliques and stars – found by power graph analysis. The radius of each disc at a point (m, n) represents – on a log scale – the number of motifs for which m proteins interact with n other proteins. High relative compression rate corresponds to denser disc charts and thus to many large cliques and bicliques. **(C)** The Y2H-union network from Yu - which has the highest relative compression rate of all Y2H networks (13 in Table 7) – has a depleted disc chart. **(D)** Collins' AP/MS network has one of the highest relative compression rates and also has one of the densest disc chart. **(E)** The same proteins as in **F** are looked at in the Y2H-union network – only the RVB1/RVB2 sub-complex is visible. **(F)** A modular sub-complex of three essential proteins: RVB1, RVB2, and ARP4 is seen participating in both the INO80 and SWR1 complexes.

Example – INO80 and SWR1C complexes. Fig. 78F shows an example from the Collins et al. (2007b) network , which has been confirmed by intense examination in Shevchenko et al. (2008). Here, three proteins – RVB1, RVB2, and ARP4 – interact with 17 other proteins in two chromatin remodeling and DNA repair complexes.

RVB1 and RVB2 are the subunits of a hetero-dodecameric DNA helicase (Torreira et al. 2008). ARP4 is an essential actin-related protein which binds to histone H2A (Harata et al. 1999). These three proteins are common subunits in two different complexes: INO80 (Shen et al. 2000) and SWR1C (Wu et al. 2009b). While RVB1 and RVB2 constitute an interaction unit as a helicase, they also form a module with ARP4 employed in these two chromatin remodeling complexes. The other 17 components of INO80 and SWR1C are found in the biclique motif. Overall, the modularity of these molecular complexes provides the biological basis for the network's significant compressibility. Some of the interactions between sub-units of the INO80 and SWR1C might be false positives, but these occur between proteins that are in the same complex or that indirectly interact. The effect of these false positives on the compressibility is thus negligible compared to that of true stochastic false positive occurring between otherwise unrelated proteins. In contrast, only the binary interaction between RVB1 and RVB2 is found in the Y2H-union dataset (Yu et al. 2008).

4.4 Materials and Methods

4.4.1 Network datasets

Exhaustive compilation of protein interaction networks. We collected all (21) large-scale genome-wide protein interaction networks derived from experimental data published between 2000 and 2009. The data files where obtained directly from the supplementary material of the publications. In the cases where the interaction data was not provided in the supplementary material or in the companion website, we obtained the data from one of the interactome databases – Biogrid, Intact, Mint, or DIP. Moreover, we did an automatic scan of these four databases and verified that we had collected all experimental datasets satisfying our strict inclusion criteria: we only consider experimental protein interaction networks that are genome-wide in intent and symmetric. We exclude dataset focused on proteins of a specific biological function.

Asymmetric networks. In symmetric networks the sets of baits and preys are largely overlapping. We exclude highly asymmetric datasets because they are not comparable to symmetric ones. For example, if the number of baits is small in comparison to the number of potential preys. Networks from Formstecher et al. (2005); Rain et al. (2001) map interactions around 102 and 261 baits respectively against several thousand preys. Another example is the network by Li et al. (2004) which is a highly asymmetric C. elegans protein interactome map between about 2,000 baits and 15,000 preys. This asymmetry introduces a bias in their relative compression rates and makes them incomparable to the other networks (9% and 18% for the Li and Formstecher datasets respectively).

Screening completeness. In the case of species with large proteomes such as D. melanogaster, C. elegans, and Human, the screening completeness of individual

datasets may be low. However, if the experiment has largely overlapping and symmetric sets of baits and preys – and is unbiased as well as genome-wide in intent – we included it (for example the Rual and Stelzl datasets).

Spoke versus matrix. In the case of AP/MS datasets we interpreted the data using the spoke model. For the Gavin dataset we also add the network derived from socio-affinity scoring (binary interactions with a socio-affinity score above 5) for comparison.

Reference networks. In addition to these experimental networks we added two literature curated datasets (Reguly et al. 2006; Yu et al. 2008), and a network derived from protein structures (Kim et al. 2006). To estimate the "average" signal of all the interactome data available we also considered the networks derived from the whole protein interaction data compiled in the BioGRID, Intact, Mint, DIP, and HPRD databases. The different species forming distinct and independent connected components of the network – hence giving a species-averaged signal.

Overlap between datasets. Some of the datasets overlap: the Ito full dataset contains the same interactions as the Ito core dataset with the addition of lower confidence interactions. The network by Collins et al. (2007b) is a computational reanalysis of the experimental data by Gavin et al. (2006); Krogan et al. (2006) with a similar method to Gavin's socio-affinity. The Y2H-Union dataset from Yu et al. (2008) is a merge of three high quality Y2H datasets: Ito-core, Uetz and the recent CCSB-YI1 (Yu et al. 2008; Ito et al. 2001a; Uetz et al. 2000).

4.4.2 Relative and absolute relative compression rates

Compression rate. Compression rates for protein interaction networks and rewired networks were calculated with the power graph algorithm (see chapter 3 page 57).

The compression rate of a network is calculated from a power graph by computing the edge reduction. If the original network has $|E|$ edge and the power graph $|E'|$ edges, then the compression rate is:

$$c = \frac{|E| - |E'|}{|E|}$$

The compression rate is between 0 and 1. If the power graph has the same number of edges as the original network, then the compression rate is 0. The maximal compression rate is achieved for a completely connected network, which reduces to one power edge.

A simple definition. Clique/biclique membership is not covered in the measure of compression rate because it only assesses the number of edges before and after compression. There are two reasons for our choice:

First, simplicity – our goal is to keep the measure as simple as possible. Combining reduction of nodes and edges into one measure leads directly to a number of subsequent questions: Are they of equal importance? Should they be weighted? How should they be combined?

Second, compression with and without nodes strongly correlate. Fig. 79 plots compression rate defined solely on edges *versus* compression rate defined on edges and nodes. The high correlation coefficient of $\rho = 0.94$ shows that the dominating factor in the compressibility of interactomes are edges and thus nodes can be ignored.

Measuring both clique and biclique content. An important point is that compressibility as measured by power graphs can capture network motifs based on cliques but also based on bicliques. Therefore, a bipartite network that does not contain a single clique can still exhibit the whole range of compression rates. Therefore networks with a clustering coefficient of zero may still have high compression rates – see *ecosystem* network in Table 10.

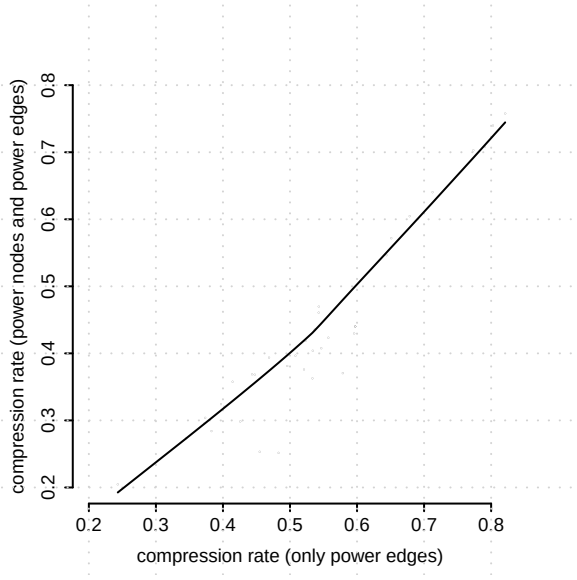


Fig. 79 **Edge reduction based on power edges compared to edge reduction based on power nodes and power edges.** We chose the simplest definition of compression rate: we compare the number of edges after and before compression. Counting power edges (after compression) is sufficient because power edges include the information about the two sets that are connected. As shown above, considering power nodes in addition to power edges does not significantly change the compression rate.

Relative compression rate. The relative compression rate measures an original network's compression rate in relation to an average random network of same topology. To compute the relative compression rate one generates 1000 random networks following the null model (see below) and computes the average compression rate. The relative compression rate measures by how much the original network's compression rate differs from the average random compression rate:

$$c_{rel} = c - \overline{c_{random}}$$

Where $\overline{c_{random}}$ is the mean of the compression rates for the random networks. For example, a relative compression rate of 0.1 means that the compression rate is 0.1 – 10% points – higher than the average compression rate of equivalent random networks. The relative compression rate is a more relevant measure than the compression rate because a certain level of compressibility is always expected, even from random networks. Fig. 80 shows the compression rates plotted against the average compression rates of topologically equivalent and randomly rewired networks.

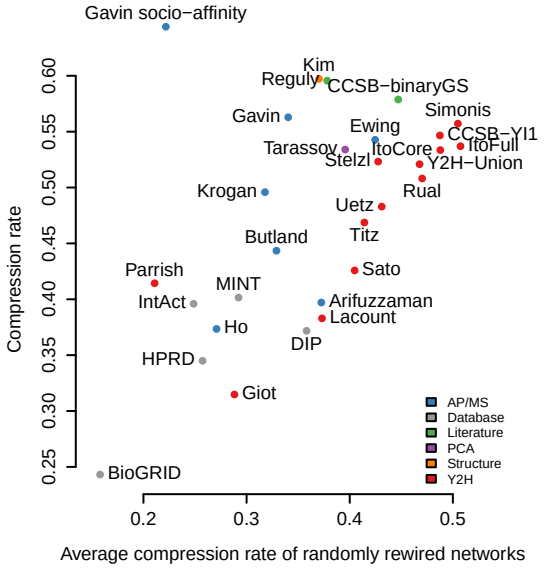


Fig. 80 **Compression rate versus the average compression rate of randomly rewired networks of same topology.** The relative compression rate is computed by taking the difference between the absolute compression rate and the average compression rate of randomly rewired networks with the same topology.

4.4.3 Random networks and network noise

Network null model – degree preserving random rewiring. Given a protein interaction network, we generate a large (1000) population of randomly rewired networks. These random networks have the same number of nodes and edges, as well as the same number of interaction partners per node and hence the same degree distribution as the original network. These networks are generated by randomly re-wiring the original network (Maslov and Sneppen 2002). Two randomly chosen interactions A-B and C-D are replaced by two new interactions A-C and B-D. This preserves the number of edges per node. This operation is repeated a number of times which is a multiple of the number of edges in the network – thus ensuring that almost all edges are rewired at least once. Moreover, each random network is generated from a previously rewired network and thus correlation with the original protein interaction network is unlikely.

Models for false negatives and false positives. For the results in Fig. 69 we used two models for false positives and false negatives. The first model – ER for Erdős–Rényi – consists in randomly adding or removing interactions. The interaction

partners are drawn from a uniform distribution over all proteins following the exponential model first described by Erdős and Rényi (Erdős 1959). The second model – BA for Barabási-Albert – consists in randomly removing interactions from poorly connected proteins and randomly adding interactions to highly connected proteins. Interaction-rich proteins get richer and interaction-poor proteins get poorer. The interaction partners are drawn from a distribution in which the probability for each protein is proportional (or inversely proportional) to the number of its interaction partners (Barabasi and Albert 1999). For both models we analyzed the influence of false positives (added interactions) and false negatives (removed interactions) separately, thus leading to four different models: ER false negatives, ER false positives, BA false negatives, BA false positives. *Important note:* since we consider symmetric genome-wide screens where the set of baits is largely overlapping to the set of nodes, we don't need to consider the bait or prey status of proteins in our noise models.

Analysis of false negatives and false positives' influence on the relative compressibility. We generated networks with simulated false positives and false negatives for 12 Yeast protein interaction networks. For each of the four models we considered 30 different levels of false positives and negatives from 1% to 60% – in total 1,440 networks. For each of these 1,440 networks we generated 1,000 topologically equivalent networks to measure the relative compressibility of the networks – more than 1.4 million compression rates were computed. The full calculation required 50,000 CPU-hours on a 2,500 CPUs supercomputer.

4.4.4 Correlations

Correlating interaction confidence scores with relative compressibility. We obtained the raw interaction confidence scores for the three datasets by Gavin et al., Parrish et al., and Tarassov et al. (provided in the supplementary material of the publications). As illustrated on Fig. 70A, we extracted sub-networks by selecting interactions with confidence scores within a given minimal and maximal value. For each pair (min, max) corresponds a sub-network for which we computed the compression rate. The relative compression rate was obtained as the difference between the compression rate of each sub-network and the compression rate of the whole network after randomization (see procedure described previously). In this context, the compressibility is measured relative to the random baseline compressibility of the whole network. This is required because otherwise sub-networks richer than the whole network in motifs and patterns would not be detected. Cells close to the diagonal represent small confidence intervals and thus correspond to small sub-networks. Unfortunately, few publications offer the raw unfiltered interaction data with confidence scores – we agree with Hart et al. (2006) that a wider availability of such raw data would greatly benefit new analysis on error rates.

Correlation of network compressibility with co-expression, co-localization, shared function, and phylogenetic similarity. We correlate interactions with gene co-expression, cellular function, cellular co-localization, and phylogenetic profile sim-

ilarity for 12 Yeast networks and for all interacting pairs of proteins for which we have complete information. We use the following assortativity ratio:

$$e = \frac{H}{H+E}$$

Where H is either the number of homotypic interactions for which the proteins are significantly co-expressed, share a cellular function, are found in at least one common cellular compartment, or have significantly similar phylogenetic profiles. $H+E$ are all the interactions – homotypic and heterotypic – for which we have complete information about *both* interacting proteins. We use data compiled by Lee et al. (2004)) for defining co-expression and phylogenetic similarity. We consider that two proteins are co-expressed if they have a log-likelihood score above 2, and phylogenetically similar if the log-likelihood score is above 1.5. Shared function was measured using the Gene Ontology (GO) molecular function (MF) and biological processes (BP) annotations as provided by the SGD database (Hong et al. 2008). For co-localization, we use the genome-wide protein localization data from Huh et al. (2003). Two proteins are co-localized if they share at least one cellular compartment, and two proteins share cellular function if they have at least one common GO term (BP or MF). As for the relative compression rate we normalize these assortativity ratios by subtracting the average proportion found for equivalent randomized networks. We thus compute the *relative* assortativity ratio:

$$r_{rel} = r - \overline{r_{random}}$$

Where $\overline{r_{random}}$ is the mean ratio obtained for topologically equivalent randomly rewired networks (see above for network null-model). In Fig. 71 the x-axis is r_{rel} (relative assortativity ratio) and the y-axis is c_{rel} (relative compression rate).

4.4.5 Networks of complex systems

We collected nine networks from the network science literature derived from complex systems of interacting entities (Table 10). These networks were chosen for their accuracy and completeness: the Internet network, software module dependencies in Java and Cytoscape, North American airport network, ownership relationships of American corporations, a food web in South Florida, co-appearance relationships between characters in the Bible, North American power grid network, and the neural network of C. elegans (the latter has been completely and accurately mapped because of its small size).

Table 10 **Networks of complex system's are compressible.** Network relative compressibility in the range $15\% - 50\%$ is typical of complete and accurate networks derived from complex systems. Note: the *South Florida Ecosystem* network has a clustering coefficient of zero because it is a *strict* bipartite network – the relative compressibility is not solely measuring clique content and clustering in networks.

network	source	year	number of nodes	number of edges	average degree	clustering coefficient	relative compression rate
South Florida Ecosystem	Heymans et al. (2002)	2000	381	2,137	11.2	zero	0.48
Cytoscape class dependencies	Cytoscape	2009	615	3,463	11.2	0.26	0.47
Bible co-appearance network	Knuth (1993)	1993	130	743	11.4	0.77	0.33
US Airports	Colizza et al. (2006)	2007	500	2,980	11.9	0.61	0.21
Corporate Ownership	Norlen et al. (2002)	2002	7,253	6,711	1.8	0.01	0.20
Java library class dependencies	Java	2006	1,538	7,817	10.1	0.39	0.17
Internet (autonomous systems)	Leskovec et al. (2005)	2006	22,963	48,436	4.2	0.23	0.17
C. elegans neural network	White et al. (1986)	1986	297	2,148	14.4	0.29	0.15
Power Grid (USA)	Watts and Strogatz (1998)	1998	4,941	6,594	2.6	0.08	0.04

4.5 Conclusion

"*I made this letter longer than usual,*
because I lack the time to make it short."
Blaise Pascal

Over the past years, numerous genome-wide protein interaction datasets have been published. They have been obtained by different experimental methodologies sparking a discussion on data quality and coverage. Since proteomic interactions are inherently co-operative, modular and redundant, interactomes are expected to contain re-occurring motifs and patterns which can be detected by measuring their relative compressibility. The relative compression rate compares the compression rate of a given network to that of random networks of the same topology. We propose the relative compression rate as a measure of the richness of interaction data in patterns and structure – richness which is affected by both false positives and false negatives. In this perspective, data quality has to be understood as encompassing both sensitivity and specificity because both high sensitivity at the expense of specificity, and high sensitivity at the expense of specificity, are detrimental to understanding the proteomes' complex molecular systems.

We underpin the relationship between relative compressibility and data quality as follows. First, by showing that adding noise (both FP and FN) negatively affects relative compressibility independently of the noise model and kind of network. Second, gold standard datasets and community-recognized higher quality datasets exhibit higher relative compressibility. Third, an assessment of confidence thresholds based solely on the relative compressibility agrees with the authors' own benchmarks and analyses. Forth, we show that relative compressibility correlates with co-expression, co-localization, and shared function. Finally, we show that well characterized complex systems from other domains also exhibit relative compressibility levels similar to that of many protein interaction networks – thus suggesting that accurate and complete interactomes are also significantly compressible.

We screened all 21 genome-wide interactome datasets available, 5 complete interaction databases, as well as three other networks. We found that networks derived from Y2H data show significantly less relative compressibility than networks derived from other experimental methods. To some extent this is attributable to the lower sensitivity of Y2H screens, which are biased towards binary, transient, and non-cooperative interactions. Possibly, the consistently low average number of interaction partners of Y2H networks indicates that the high selection stringency employed to achieve high specificity leads to more depleted networks (Uetz et al. 2000; Ito et al. 2001b). However, other types of networks with equally low average number of interaction partners, or similarly low clustering coefficients still report higher relative compressibility. Our results suggest that advances in Y2H screening strategies – in particular two-phase pooling – can bring the relative compressibility of Y2H networks to levels similar to AP/MS and PCA networks (above 10% for Stelzl and Parrish datasets). In fact smarter pooling strategies for Y2H screening are being developed and tested such as Shifted Transversal Design and Steiner-triple-system, thus paving the way for higher sensitivity and accuracy (Zhong et al. 2003; Jin et al. 2006; Xin et al. 2009). We also observed a similar effect of the experimental method on the relative compressibility of AP/MS screens. Networks derived from state-of-the-art purification procedures (Tandem affinity purification, TAP) and detecting interactions of baits expressed at physiological levels (knock-in versus cDNA over-expression) exhibit higher relative compressibility.

Chapter 5

Applications to Literature Derived Networks

5.1 Introduction

In this chapter we present an application of power graph analysis to the analysis of co-occurrence networks derived from literature. Automated extraction and analysis of biomedical knowledge from the literature has become a critical problem. More than 18,000,000 biomedical articles have been indexed in the PubMed database and several thousand new articles are added every day. Text-mining has been applied to this problem with mixed results[1]. One of the indispensable first steps necessary for text-mining of biomedical knowledge is the unambiguous identification of gene mentions in text – or *gene mention normalization*. We first present a context-based gene mention normalization algorithm that achieved the best result of over 81% F_1-measure at the BioCreAtIvE II text-mining challenge (Hakenberg et al. 2008). This allows us to precisely identify Human genes throughout the biomedical literature and associate to each gene mention an entry from gene databases. We use this resource to build a gene co-occurrence network derived from literature – two genes co-occur if they are mentioned together in significantly many abstracts. We show how to explore this large network with power graph analysis and how it can be used to gather insights into Human genes important for the cell cycle progression. In particular, we show how combining text-mining and power graph analysis can be used to suggest novel annotations of genes. In addition, we show that 25% of cell-cycle genes found in a high-throughput RNAi screen by Kittler et al. (2007) can be confirmed using protein interaction data.

5.2 Background

Before the advent of high-throughput experimental assays, it was feasible for researchers to put the results of their experiments into the wider context of evidences published in the biomedical literature. Today, gene expression microarray, RNAi screens, and other high-throughput assays produce long lists of genes that

1 Hirschman et al. 2005a; Krallinger et al. 2008.

need to be put into context. This problem is exacerbated by the insufficient manual curation in databases such as e.g. *Entrez Gene*[2], *UniProt*[3], or *GOA*[4]. Every day more than 2,000 new abstracts are added to PubMed which now totals more than 18,000,000 citations (Baumgartner et al. 2007) – it is impossible to manually curate and enter all this information in databases. Among these abstracts at least 2.74 million mention Human genes (Plake et al. 2009). The plot in Fig. 81 shows the number of first mentions of Human genes in PubMed per year. Text-mining of biomedical literature holds the key to unlock this textual knowledge.

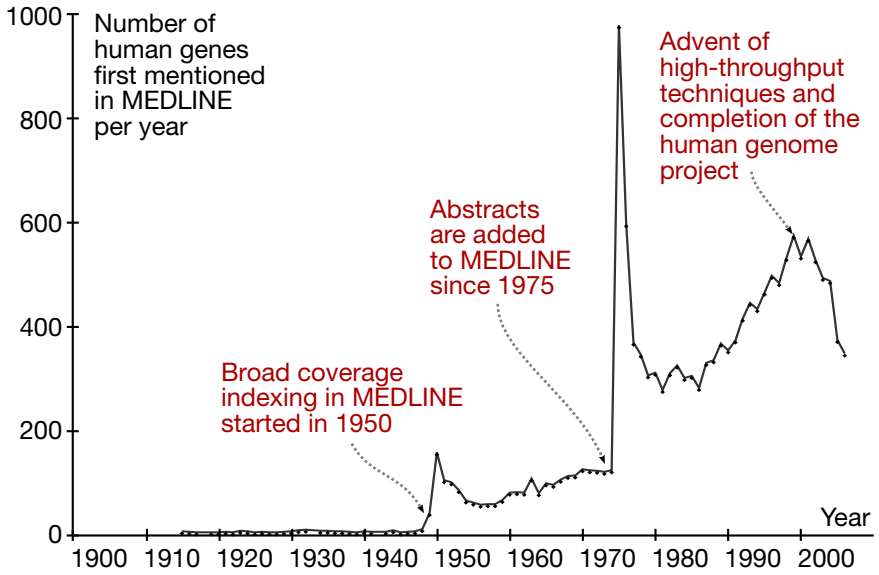


Fig. 81 **Number of Human genes first mentioned in PubMed abstracts per year.** Two artifacts of the indexing process are visible: in 1950 when broad coverage indexing started at NCBI and in 1975 when abstracts and not just titles started to be added to PubMed. Since the year 2000 the rate of introduction of novel gene mentions has slowed down.

BioCreAtIvE I and II. To address this question, the BioCreAtIvE[5] challenges were organized. BioCreAtIvE is a community-wide effort for the evaluation of text mining and information extraction systems applied to the biological domain. The first BioCreAtIvE challenge was conducted in 2004 (Blaschke et al. 2003) and the second in 2006 (Krallinger et al. 2008). Fifteen teams participated to the first challenge and more than twenty for the second challenge. The BioCreAtIvE tasks included:

Gene mention recognition: Find mentions of genes in biomedical text without necessarily linking the mention to a specific database entry.

Gene mention normalization: Find mentions of genes in text and identify them to specific database entries.

Gene function annotation: Annotation of gene products based on evidence found in biomedical text.

2 Maglott et al. 2007.
3 Consortium 2009.
4 Barrell et al. 2009.
5 Critical Assessment of Information Extraction systems in Biology

Protein-protein interaction: Find mentions of interacting proteins in biomedical text.

In the following we first focus on the problems of *gene recognition and normalization* and review the different approaches that have been developed.

5.2.1 Recognizing and Normalizing gene names

Gene mention recognition (GR)

Gene name nomenclatures are either non-existent or poorly respected (Hanisch et al. 2003). Moreover, genes are often described such as "p65 subunit of NF-kappaB" instead of being given short names or abbreviations. Therefore, it is a non-trivial problem to locate gene mentions in biomedical text since other terms may be confused for gene mentions. In Drosophila the problem is even worse because researchers have a tradition of giving humorous names to genes. For example, mutations in the genes *Ken and Barbie* result in no externally visible genitalia. Another difficulty is to precisely delimit gene mentions: among "p65", "p65 subunit" and "p65 subunit of NF-kappaB" which are correct gene mentions?

BioCreAtIvE I GR results. The results of the first BioCreAtIvE challenge showed that high F_1-measure of over 80% are achievable (Yeh et al. 2005). The benchmark consisted of 15,000 manually curated sentences from sentences from MEDLINE (Tanabe et al. 2005). Zhou et al. (2005) obtained the best performance with a F_1-measure of 82% using Support Vector Machine (SVM) and discriminative Hidden Markov Models (HMM). Hakenberg et al. (2005) systematically evaluated the features for gene name recognition and found that character sequence statistics were the most informative.

BioCreAtIvE II GR results. Better results were obtained with a maximal F_1-measure of 87.2% and even 90% when combining the prediction of all 21 teams together (Smith et al. 2008). the best performing system by Ando (2007) used a semi-supervised method which exploits unlabeled data in addition to the labeled training data.

Gene mention normalization (GN)

In the biomedical literature, the same gene or protein is often mentioned by different synonymic names and abbreviations. On the other hand, gene and protein names are also polysemous – different genes may have the same name or abbreviation. Therefore, a challenging task is to link gene mentions to specific gene or protein databases entries. Already before the first BioCreAtIvE challenge, Morgan et al. (2004) had proposed a gene recognition and normalization system for aiding the curation process of the FlyBase[6] database. Their system was based on Hidden Markov Model (HMM) approach and achieved a F_1-measure of 72% (precision 88% and recall 61%) on a benchmark.

6 Drysdale and Consortium 2008.

BioCreAtIvE I GN results. The results of the BioCreAtIvE I challenge showed that for Yeast a F_1-measure of 92% could be attained. But for Mouse and Fly, the results were not as good, due in part to more ambiguity in the gene naming conventions. The best F_1-measure for Fly was 82% and for Mouse 79% (Hirschman et al. 2005b). Among the best performing systems is ProMiner which recognizes gene mentions based on up-to-date dictionaries of genes and protein names (Hanisch et al. 2003). ProMiner reached a F_1-measure of approximately 80% for Mouse and Fly, and about 90% for Yeast (Hanisch et al. 2005). Later, it was extended to use conditional random fields for the recognition of variation in biomedical terms (Klinger et al. 2007). Similarly, Fundel et al. (2005) proposed a simple approach based on gene name synonym lists. They obtained an F_1-measure of 89.7% for Yeast and of 77.3% for Mouse.

BioCreAtIvE II GN results. For BioCreAtIvE II, instead of normalizing Yeast, Fly and Mouse gene mentions, the task was to normalize Human gene mentions. The organizers provided 281 expert-annotated abstracts containing 684 gene identifiers for training, as well as a blind test set of 262 documents containing 785 identifiers. Three systems achieved F_1-measures between 80% and 81%. By pooling the results from the 20 participating teams, the organizers attained a F_1-measure of over 90% – a performance comparable to Human experts (inter-annotator agreement) (Morgan et al. 2008).

5.2.2 Mining networks from literature

In the following we review network-based tools developed for navigating the biomedical literature as well integrative frameworks that combine other data such as gene co-expression.

Networks for navigating the literature. Early web-based tools such as MedMiner by Tanabe et al. (1999) and PubMatrix by (Becker et al. 2003) offered the analysis of cDNA microarrays data based on text-mined information from GeneCards and PubMed (Safran et al. 2010). A popular resource is iHOP by Hoffmann and Valencia (2004). Information Hyperlinked over Proteins (iHOP) is a navigable network of sentences from PubMed with proteins and genes as hyperlinks. It contains half a million sentences and 30,000 different genes from nine species including Human, Mouse and Yeast (Hoffmann et al. 2005). Other tools soon followed linking not just genes and proteins but also other biological terms such as diseases, drugs and cell types. Chen and Sharp (2004) presented Chilibot, a text-mining tool for building relationship networks among biological concepts, genes, proteins, and drugs. Plake et al. (2006) proposed AliBaba, an interactive tool for the extraction and visualization of associations between cells, diseases, drugs, proteins, species and tissues from PubMed abstracts.

Integrative approaches. von Mering et al. (2003) developed STRING, a database of predicted functional associations between proteins derived from genomic context,

high-throughput experiment data, conserved co-expression, and text-mined biomedical knowledge. This is probably the most comprehensive resource of its kind with 261,033 orthologous genes from 89 fully sequenced genomes. Simiarly, Köhler et al. (2006) developed ONDEX – a database integration tool featuring text mining based knowledge extraction and methods for graph-based analysis. Sivachenko and Yuryev (2007) developed the integrative pathway analysis platform – *Pathway studio* – for the analysis and navigation of molecular networks for drug discovery.

5.3 Human gene mention normalization

The starting point for mining gene and protein knowledge from literature is the unambiguous normalization of gene mentions in text: a gene mention must be linked to a specific gene or protein database entry. In this section we will present our context-based gene mention normalization algorithm that achieved the best result of over 81% at the BioCreAtIvE II text-mining challenge (Hakenberg et al. 2008). This resource is available as *GoGene*[7], a search engine for genes and proteins that sorts its results hierarchically according to the Gene Ontology and MESH (Plake et al. 2009).

Context model based normalization. Biologists reading articles usually don't have a problem identifying genes in text. They have a defined scope of interest and read articles relevant to their research field. Moreover, articles and gene mentions are not considered in isolation, but in the context of molecular functions, biological processes, diseases, etc. We solve the gene normalization problem by mimicking the Human approach of using background knowledge to define contexts. These contexts consist of biological function, cellular localization, diseases, species, mutation, and other biological information that may be associated to genes and that are mentioned in text. For each PubMed abstract and for each of its gene mentions we define a context. Gene mention normalization is thus the problem of finding for a gene mention and its textual context, the best matching gene context.

An example. As shown in an example in Fig. 82, our approach to gene normalization uses context models. In the absence of a context, the gene mention (highlighted in grey) is ambiguous since there exist five distinct Human genes named *p54*. The abstract contains terms such as *RNA helicase*, *Human*, *chromosome 11* that provide a context. The best match is the *p54* gene with the Entrez Gene identifier 1656. This particular gene has a similar context to the context of the gene mention: it is a *Human* gene located on *chromosome 11* and has been annotated with the Gene Ontology term *RNA helicase*. The other candidate genes are not RNA helicases, or are located in a different chromosome (chromosome 3 or 6).

7 www.gopubmed.org/gogene/

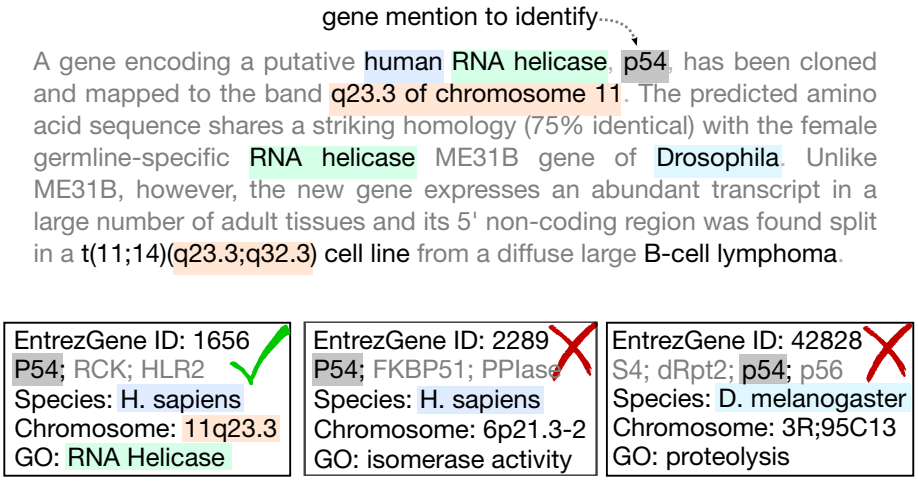


Fig. 82 **Example of gene mention normalization using context models.** Terminology relevant to function, location, disease, etc. is identified in text and defines the textual context, which is matched against the potential gene contexts. Among the three candidates identities for this gene mention p54, only one is a Human RNA helicase on chromosome 11.

In the following we explain in more detail how high recall and high precision can be achieved for gene mention normalization. Fig. 83 illustrates the complete workflow.

5.3.1 Recall – syntactic flexibility through regular expressions

The starting point of the normalization of genes in text is the compilation of gene names from gene and protein databases. For each name in our reference database we construct a regular expression which is flexible to syntactical variations. Gene names may be identifiers, abbreviations, or whole phrases (page 136). Hence, for distinct types of gene names we apply distinct methods to obtain the regular expressions that reflect differences in syntactic flexibility (page 137). We compile these regular expressions into a single finite state automaton that is used for finding gene mentions in text (page 138). A gene mention is often matched by several regular expressions and thus it is associated to several identifiers from the reference database.

Reference database for Human genes. The Entrez Gene database lists 23,438 Human genes which we use as our reference identifiers. To ensure high recall, we compile for each gene all its known synonyms from Entrez Gene (Maglott et al. 2007) and add relevant protein names and synonyms from UniProt (Consortium 2009).

Gene name classification. All synonyms from the lexicon are classified into either of four categories. We treat any instance of these four groups differently concerning the way we generate regular expressions:

- database identifiers,
- abbreviations and acronyms,
- compound names,
- unlikely gene names.

Database identifiers referred to identifiers and accessions codes from UniProt, HGNC, VEGA, KIAA, FLJ-DB, and several others (Consortium 2009; Prasad et al. 2009; Wilming et al. 2008; Kikuno et al. 2004; Ota et al. 2004). Abbreviations and acronyms in our scheme are names that have zero or one white space, a mix of upper and lower case characters, digits, and symbols, or only upper case letters. We treat words that start with an upper case letter but continue with lower cases as compound names. Compound names are names with multiple white spaces or that are not grouped as abbreviations. We filter out unlikely names ("AA", "ORF has no N–terminal 'Met',it may be non–functional", single letters, numbers) and concentrate on the other three groups instead. In the following we explain how we generate regular expressions for each of these synonym classes.

Database identifiers. Database identifiers often follow a strict syntactic format, we search for them using predefined regular expressions. A match triggers an immediate normalization of the referenced gene/protein. No further disambiguation is required if it is an explicit database entry.

Abbreviations and acronyms. To generate regular expressions for abbreviations and acronyms, we segment each name into components. This segmentation is triggered by strong and weak bonds within a name. White spaces and hyphens are strong bonds, weak bonds occur for every other change in the flow of characters (between upper and lower case letters; between letters and digits.) We also introduce weak bonds for the first and last letter in a sequence of letters. For each segment, we generate potential variations based on observations in the lexicon list and training data. In general, variations are allowed for changes in the surface pattern of:

- letter sequences such as MYD, Myd, myd, MyD,
- switches between Roman and Arabic numbering such as 2, ii, II
- single letters for Greek characters such as α, a, A, alpha, Alpha
- special single letters such as R, r, or receptor and L, LG, l, or ligand.

Possible variations for each segment are combined into a regular expression; all expressions for all segments define an expression for the whole abbreviation, with any kind of gap in between. Examples are:

- HER2 = { HER, HeR, Her, her } [-]? { 2, ii, II }
- IFN-gamma = { IFN, Ifn, IfN, ifn } [-]? { g, G, gamma }
- MYD88 = { MYD, MyD, Myd, myd } [-]? { 88 }
- CYP1A1 = { CYP, CyP, Cyp, cyp } [-]? { 1, i, I } [-]? { a, A, alpha } [-]? { 1, i, I }
- CD95R = { CD, Cd } [-] { 95 } [-]? { r, R, receptor }

In addition, abbreviations of Human gene names often feature an additional "h" at the beginning, so this is added as optional to every abbreviation.

Compound names. We segment compound names at white spaces. Every segment (or token in this case) is treated similarly to abbreviations. Tokens that resemble English words (initial upper or lower case, then all lower case letters) have less variations in their capitalization (all lower-case or initial upper-case). Some tokens in a

compound name are often left out in text, such as “protein”, “domain”, “antigen”, and “channel” – we thus encode all these as optional in the regular expression.

Regular expression filtering. We remove regular expressions that match any of 7,700 manually curated stop words. Hand–crafted rules also remove matches such as “or 45” and “and 1”, triggered by too loose regular expressions for the gene names “Or45” and “And–1”, respectively. For some names, such as “protein 1” or “antigen 2”, we do not generate regular expressions that allow for variability, but instead require exact matches.

Scanning text with a finite state automaton. All regular expressions for all gene names are compiled into a single finite state automaton. The end states of the automaton correspond to each and every potential match – accumulating all corresponding Entrez Gene identifiers.

Additional identifiers for ambiguous gene mention. We add gene identifiers found in similar abstracts when several gene identifiers are found for an ambiguous mention. Comparing the text at hand to texts collected from the Entrez Gene summaries and other textual sources linked to specific genes, we are able to add missing gene identifiers.

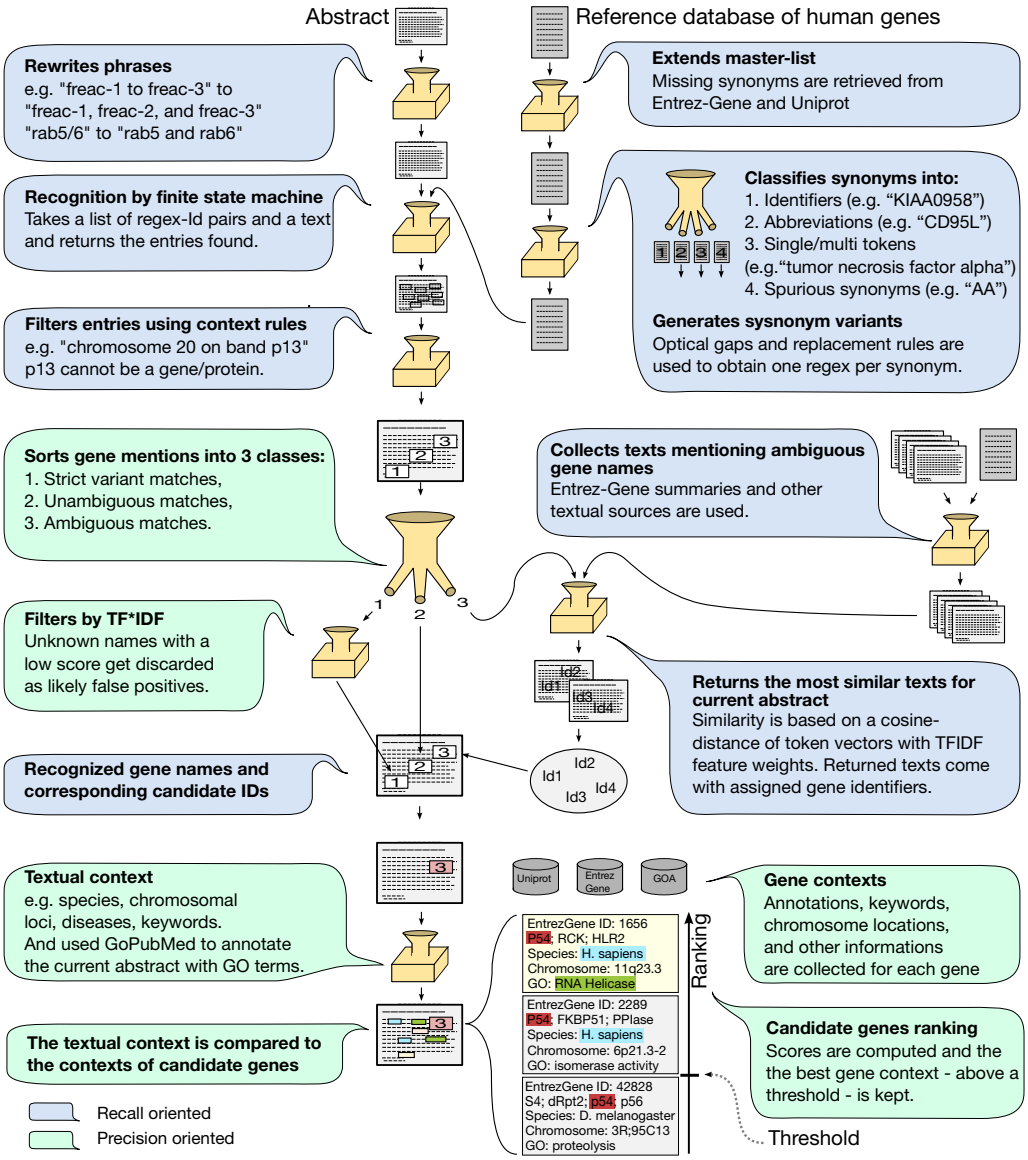


Fig. 83 **Gene mention normalization workflow.** Human gene mention normalization starts with a piece of text – generally a biomedical abstract from PubMed – and a reference database for Human gene names. After extending the reference list of gene names, a finite state machine locates gene mentions in the abstract. We filter gene mentions using the text immediately surrounding a gene. Ambiguous gene mentions are re-evaluated by looking at similar texts and adding relevant gene identifiers to improve recall. Finally the textual context of each gene mention is compared to the contexts of each candidate gene identifier. The candidate gene identifiers for each gene mention are then ranked and – if above a given threshold – the best is chosen as the correct gene identifier.

5.3.2 Precision – ranking candidates by context similarity

Gene mentions are identified in text and assigned to one or several candidate gene identifiers. The problem is to choose the correct identifier. Our algorithm builds and compares contexts for both the text in which a mention is found, and for the candidate genes. For each gene mention in a given context, the best matching gene context likely corresponds to the correct gene.

Gene contexts. Background knowledge from Entrez Gene, UniProt, and GOA is collected for each of the 23,438 genes. For each gene we define a context model as a collection of items such as ontology terms, keywords, and whole text fragments:

- Entrez Gene summary text,
- Gene Ontology terms from Entrez Gene, GOA and UniProt,
- *"Gene Reference Into Function"* from Entrez Gene (GeneRIFs),
- chromosomal location,
- names of protein interaction partners,
- associated diseases,
- keywords and functions from UniProt,
- mutations,
- protein domains found in the gene products (for protein coding genes),
- tissue specificity from UniProt,

Text fragments are for example whole paragraphs from the Entrez Gene summaries that describe a gene, or descriptions of a gene's implications in diseases from UniProt.

Gene mention filtering. Gene mentions are filtered out according to heuristics. These invoke the immediate context of a name, that may contain evidence that the name refers to a different species (not Human or mammal), that the name refers to a disease, that it is an unspecific mention (of a protein family), or that is a common English word.

Text contexts. The context model for abstracts is a bag-of-words from which we excluded about 80 stop words and other non–discriminative words such as "gene" and "protein". We also use Gene Ontology terms mined from text available from GoPubMed (Doms and Schroeder 2005).

Comparing contexts. Gene Ontology terms, keywords, and text fragments are compared separately. Each comparison yields likelihoods that measure the similarity of the current text with the textual knowledge available on each gene. We combine the likelihoods into a normalized confidence score between 0 and 1.

For Gene Ontology terms we use a similarity measure that takes into account the shortest path via the lowest common ancestors in the hierarchy, as well as the depth of this lowest common ancestor in the overall hierarchy (comparable to Schlicker et al. (2006)). Since potentially several terms are associated to each gene or found

in each text, we chose to keep the most similar pair of terms as representing the similarity between gene and text contexts GO terms.

For keywords (not part of the Gene Ontology), we calculate the fraction of terms occurring in the abstracts among all terms. For text fragments, we calculate the cosine distance of both bag–of–word representations and the normalized overlap (fraction of tokens from the disease description that also occur in the abstract). Correct matches for mutations and chromosomal locations trigger the maximal score. In most cases these annotations are enough to unambiguously identify a gene.

Ranking of candidate genes. Once a confidence score has been calculated for each candidate gene, we pick the gene identifier (Entrez Gene identifier) with the highest score. In some cases the score is too low and the gene mention is discarded.

5.3.3 Related work

A Similar system to ours is the work by Fundel and Zimmer (2007) that achieves an F_1-measure of 80.5 % on the BioCreAtIvE gene normalization task. This system uses a dictionary containing the Human gene names taken from Hugo, EntrezGene, and SwissProt. In addition, this system filters the names of cell lines and diseases that resemble gene names. Disambiguation is done by calculating the cosine similarity of the abstract with known candidate gene synonyms.

5.3.4 Results at the BioCreAtIvE II competition

Our gene normalization system was developed using a publicly available training set of PubMed abstracts with gene mentions identified to Entrez Gene identifiers. The evaluation was done on a test set provided by the organizers of the BioCreAtIvE II challenge. Our results for the gene normalization task are shown in Table 11. In addition to the performance on the training set, and official results on the test set, we show performance of the current system which has been further improved.

Table 11 **Gene normalization evaluated on the BioCreAtIvE II training and test set.** Results on the test set reflect the expected performance of the system and are the official results of the competition (independently evaluated). The last row shows performance improvement in the aftermath of BioCreAtIvE II.

Short description of the submitted run	Precision	Recall	F_1-measure
Training set	82.1	81.6	**81.8**
Training set, no filtering, no disambiguation	20.2	**92.7**	33.1
Test set	78.9	83.3	**81.0**
Test set, no disambiguation	49.6	**87.5**	63.3
Test set, current performance	90.7	82.4	**86.4**

Our best official F_1-measure on the test set is 81% – the current system achieves 86.4%. Our maximum recall values are between 87.5% and 92.7% (test and training set, respectively). Table 12 shows the influence of different context types on

the performance. Chromosomal locations have the greatest influence on precision (+64.5%), but unfortunately not all abstracts contain such information. To maintain high recall, Gene Ontology (GO) terms are needed (losing only 9.9% in recall, but gaining 36.6% in precision).

Table 12 **Breakdown of the Impact of different context types on Human gene mention normalization.** Starting from a baseline configuration (pure recognition of named entities, see text), each context type was evaluated *separately*. In addition, we present the influence of filtering by the immediate context in a sentence, for example by excluding wrong species. Table adapted from Hakenberg et al. (2008)

Context type	Precision (%)	Recall (%)	F_1-measure (%)
Baseline: no context	9.7	**91.1**	17.5
+ GeneRifs	50.8	78.3	**61.6**
+ GO terms	46.3	**81.2**	59.0
+ EntrezGene summaries	49.0	66.7	56.5
+ Diseases	22.7	43.9	29.9
+ Functions	50.8	72.5	59.7
+ Keywords	53.0	53.6	53.3
+ Chromosome locations	**74.2**	14.8	24.7
+ Tissues	39.4	29.1	33.4
+ Immediate contexts (heuristics)	23.5	89.8	37.2

In the following we will see ho we can use this resource to mine a gene-gene co-occurrence network from the whole biomedical literature and apply it to functional prediction of Human cell-cycle genes.

5.4 Human gene literature co-occurrence network

One way to analyze the large corpus of text-mined gene mentions is to construct a gene co-occurrence network. Two genes co-occur if they are mentioned together in at least one abstract. We use the hypergeometric test to filter-out the least significant pairs of genes.

Mining PubMed abstracts for gene co-occurrences. We use an offline version of PubMed (all abstracts until 2008) and the gene mention normalization algorithm previously described to find 2.74 million abstracts mentioning at least one Human gene. Two genes are said to co-occur if they are mentioned together in at least one abstract. We construct a network consisting 851,954 edges representing all co-occurrences between 9,774 Human genes.

Statistical filtering. Assuming a precision of 90.7% for identifying genes, the precision for identifying a pair of genes is 90% $\times$ 90% = 81%. Thus we expect at least about 20% of gene-gene co-occurrences to be incorrect. We remove from the network statistically insignificant co-occurrences using the hypergeometric distribution (King et al. 2004). Given two genes g_1 and g_2 mentioned both at least once together in the literature, we compute a p-value for the significance of their co-occurrence as follows:

$$p(g_1, g_2) = 1 - \sum_{i=0}^{k-1} \frac{\binom{v}{i}\binom{n-v}{u-i}}{\binom{n}{u}}$$

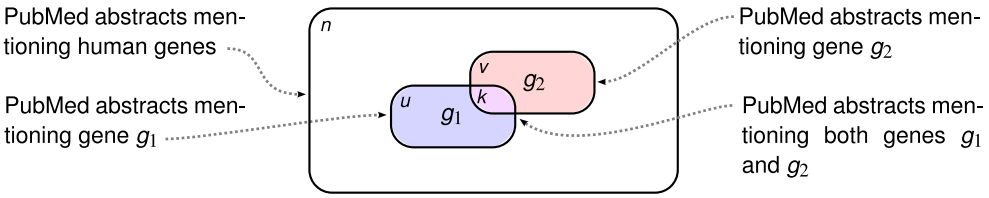


Fig. 84 **Evaluating the statistical significance of gene–gene co-occurrences with the hypergeometric test.** There is a total of n = 2.74 million PubMed abstracts mentioning Human genes. Gene g_1 is mentioned in u abstracts and gene g_2 is mentioned in v abstracts. The number of abstracts mentioning both g_1 and g_2 is k. The p-value $p(g_1, g_2)$ is the probability that g_1 and g_2 co-occur by chance alone in k or more abstracts given u, v, and n.

The p-value corresponds to the probability that the two genes g_1 and g_2 are mentioned together in k or more abstracts assuming that their occurrences are random and independent from each other. This is equivalent to the value obtained from a one-tailed version of Fisher's exact test (Mehta et al. 1984). We use Bonferroni's correction and compute a corrected p-value $p_c = mp$, where m is the number of co-occurrences tested for significance (Strassburger and Bretz 2008). We chose a threshold p-value at $p_c < 10^{-3}$ and remove all co-occurrences that have a higher p-value (higher p-value implies lower significance).

After this statistical filtering, the network consists of 9,774 genes and 42,049 edges representing statistically significant co-occurrences. Each gene is co-occurring with an average of 8.6 other genes (average degree).

Comparison to Human protein–protein interactions.
We compare this co-occurrence network with protein–protein interaction (PPI) data. For this we use both protein– protein interactions from the HPRD and BioGRID databases (Prasad et al. 2009; Stark et al. 2006) totaling 47,900 unique interactions between 9,460 Human proteins. The HPRD database provides high quality manually curated interactions from the literature.

Fig. 85, shows the little overlap between gene co-occurrences and protein interactions. Only 13% (5,741) of all co-occurrences correspond to known protein interactions. Careful inspection of the network reveals other types of relationships between genes than the interactions between their corresponding proteins. For example, GSTA1 and GSTT1 are both glutathione S-transferases and thus it makes sense that they are both mentioned together in 12 abstracts (p-value $< 10^{-11}$, GSTA1 and GSTT1 are mentioned a total of 614 times and 857 times, respectively.) Sometimes, the relationship described in the literature is not a physical interaction but instead a genetic interaction – for example the two transcription factors SP1 and USF1. There is more than 40 abstracts in PubMed mentioning both transcription factors together as regulating genes in concert, and at least one mentions their genetic interaction (Ge et al. 2003). In another example karyopherin beta (KPNB1) and RAN binding protein 1 (RANBP1) relationship is the stabilizing role of RANBP1 for the interaction of KPNB1 with RAN (Chi et al. 1996).

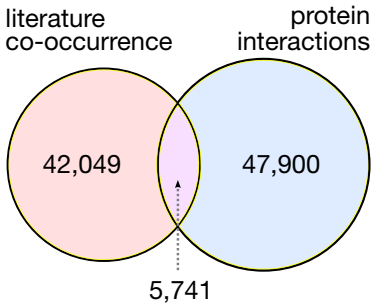


Fig. 85 **Comparing gene–gene co-occurrences with protein–protein interactions.** The protein interactions are compiled from the HPRD and BioGRID databases (Prasad et al. 2009; Stark et al. 2006) totaling 47,900 interactions between 9,460 Human proteins. Only 5,741 literature gene co-occurrences are interactions, and many interacting genes are not found co-occurring in literature. This highlights both the difficulty of mining information from literature and the existence of other relationships between proteins other than direct physical interactions.

The Human gene literature co-occurrence network provides the whole literature at a glance for genes and their relationships – whether they are functional, physical, genetic, or other.

Power Graph Analysis. We compute a power graph using the power graph algorithm (see page 78) and obtain a power graph consisting of the same 9,774 nodes

(genes), 3,776 power nodes (groups of genes), and 24,381 power edges. Among these power edges, 1,011 (2%) are reflexive and thus represent cliques, 5,941 (14%) correspond to stars and bicliques and 17,429 (41%) are unclustered and remain as single edges. The edge reduction is of 42%.

As shown in Fig. 86, the power nodes form a complex hierarchy from which a few preeminent features can be seen such as the hubs in the network. Table 13 lists the three most connected genes – or hubs – in the network: IFNG, AKT1, and TP53. Note that AKT1 is the most connected gene but is not the most mentioned.

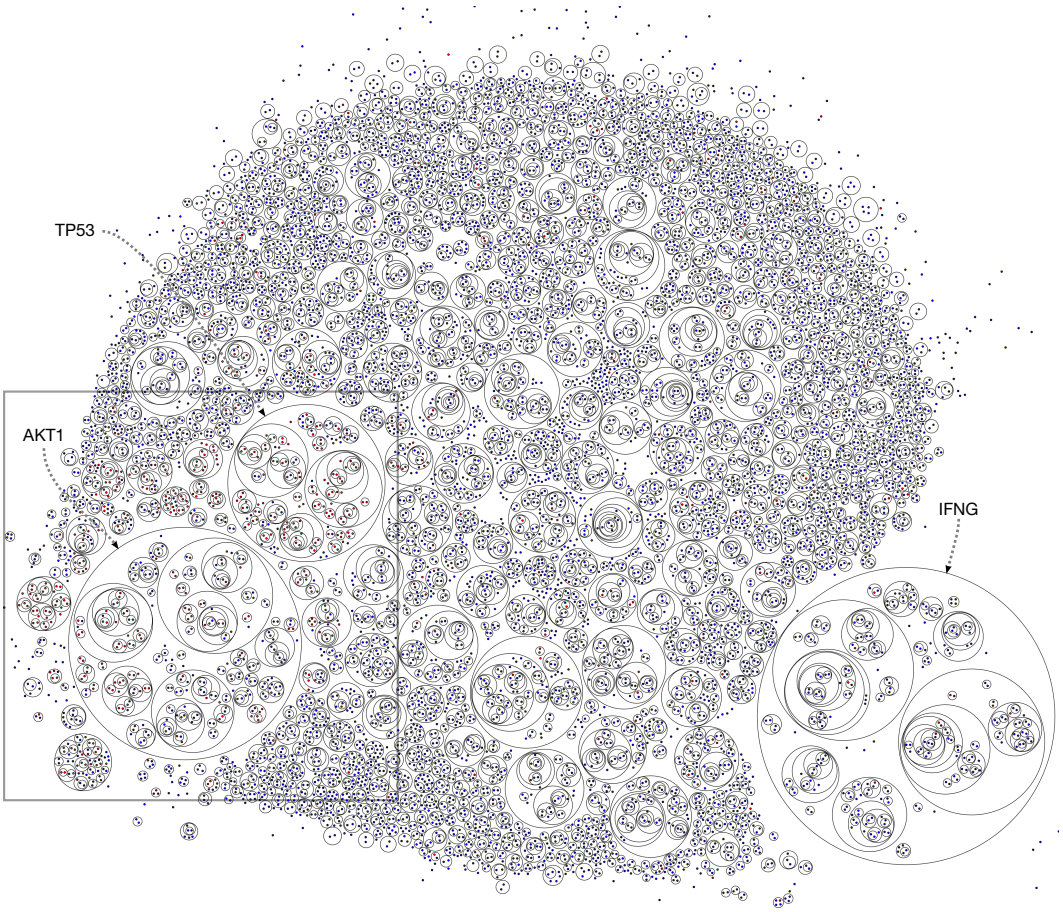


Fig. 86 **Power graph of the Human gene co-occurrence network.** The network is mined from 2.74 million PubMed abstracts and consists of 9,774 genes and 42,049 co-occurrences. The power graph has 3,776 power nodes additionally to the 9,774 singleton nodes (genes), and 24,381 power edges. Only power nodes and reflexive power edges (cliques) are shown. Power nodes that form cliques are drawn in green. The three biggest power nodes correspond to hubs in the network: *interferon gamma* (IFNG), *v-akt murine thymoma viral oncogene homolog 1* (AKT1), *tumor protein p53* (TP53). All three are often mentioned in the literature together with other gene because of their central role in cancer. The dashed box corresponds to a region enriched in cell cycle genes (see Fig. 87).

Table 13 **Top 3 most co-occurring Human genes in PubMed.** AKT1, IFNG and TP53 are most mentioned Human genes in the biomedical literature and are hubs in the Human gene co-occurrence network. The number of mentions in PubMed abstracts is given as well as the number of genes that they co-occur with.

gene name	mentions	co-occurring genes	description
AKT1	13127	270	protein kinase involved in apoptosis, signal transduction of growth factors, and development.
IFNG	39451	251	Interferon-gamma – cytokine critical for innate and adaptive immunity against viral and intracellular bacterial infections as well as for tumor control.
TP53	38619	191	Gene encodes tumor protein p53 which regulates cell cycle arrest, apoptosis, senescence, DNA repair, and changes in metabolism.

5.4.1 Case study – cell cycle genes

In addition to the text-mining of gene mentions in the biomedical literature, we can also mine gene annotations. By mining Gene Ontology (GO) terms relevant to cell-cycle, we can establish a comprehensive map of Human cell-cycle genes. In the following we compare known cell-cycle genes and genes found by text-mining and show that 40.6% of cell-cycle Gene Ontology Annotations (GOA) are confirmed by text-mining. However, some of the unconfirmed annotations are false negatives, many of the genes found by text-mining are indirectly relevant to the cell-cycle.

Literature co-occurrence of genes with cell cycle terms. GO has 352 terms under *cell cycle*. For example: *mitosis*, *G1 phase*, *M phase*, *chiasma formation*, or *spindle elongation*. These terms are identified literally in text using techniques from the GoPubMed search engine (Doms and Schroeder 2005; Dietze et al. 2008). We find 210.000 abstracts mentioning a gene together with a cell cycle term. Similarly as for gene-gene co-occurrence we filter out insignificant co-occurrences by using the hypergeometric test with Bonferroni correction ($p < 10^{-3}$).

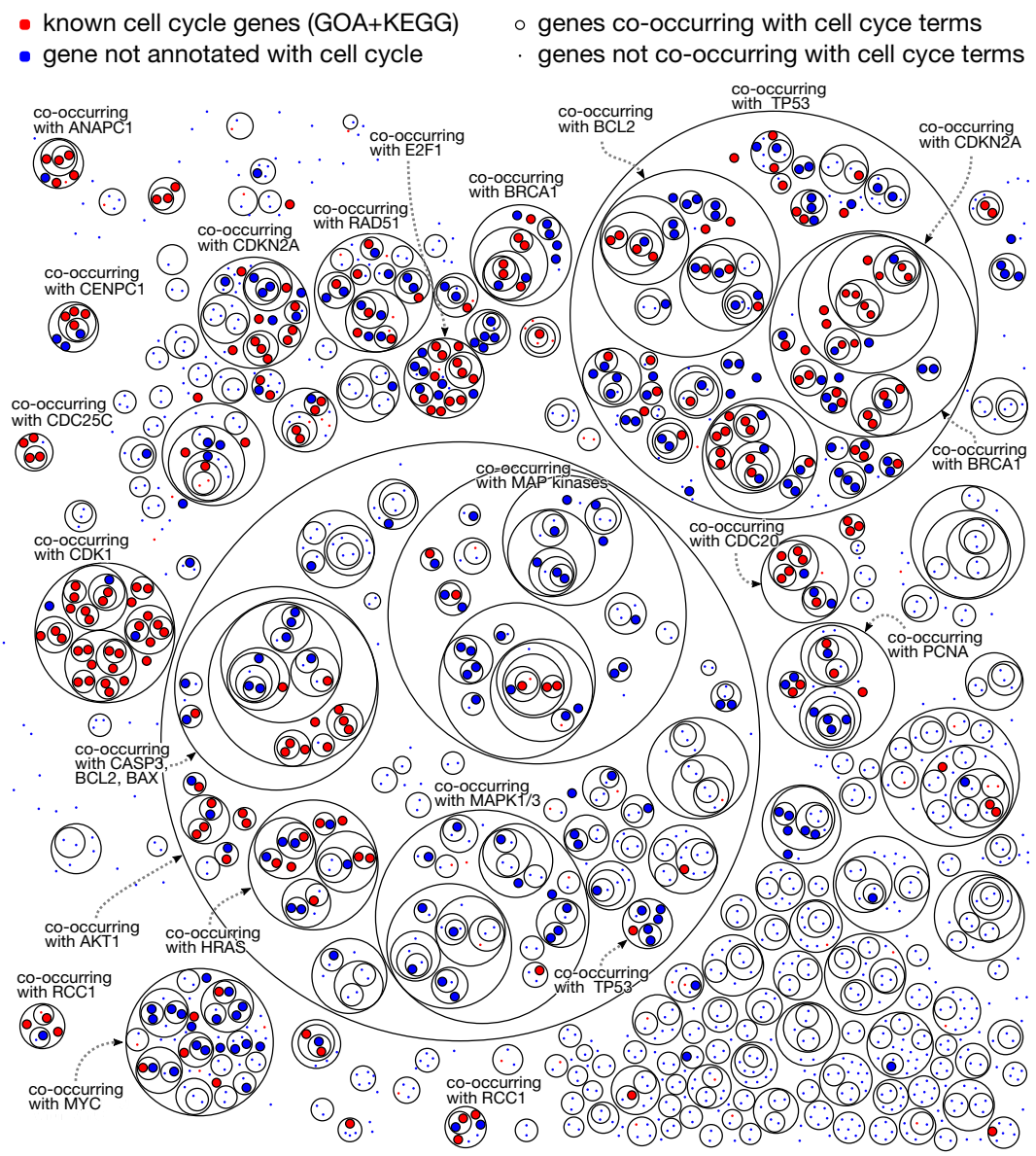


Fig. 87 **Region of the gene–gene co-occurrence power graph enriched with cell cycle genes.** Known cell cycle genes according to GOA and KEGG annotations are shown in red (blue if not known to be relevant to cell cycle). Genes that co-occur with cell cycle terms are shown with a bigger radius than those that do not. GOA and KEGG cell cycle annotations partially overlap with genes co-occurring with cell cycle terms in literature. Using the gene–gene co-occurrence network we see that cell cycle genes cluster together in the same power nodes. For example, genes co-occurring with CDK1, BCL2, TP53, and CDKN2A are annotated or/and co-occur in the literature with cell cycle terms.

Comparing GOA and KEGG with text-mining for cell cycle annotations. We evaluate how many genes annotated with cell cycle terms can be recovered using text-mining. Cell cycle annotations are compiled from GOA (Barrell et al. 2009) and KEGG (Okuda et al. 2008) databases. Table 14 summarizes the results: 40.6% of cell cycle genes are confirmed by text-mining of literature co-occurrences with cell cycle GO terms.

Table 14 **Confirming cell cycle genes using GO term co-occurrence.** We show the results confirming cell cycle genes annotated in the Gene Ontology Annotation (Barrell et al. 2009) and KEGG (Okuda et al. 2008) databases. The prediction is done with literature co-occurrence of genes with cell cycle terms. Predicting known cell cycle genes from GOA can be done at a maximal recall of 55%. When predicting GOA + KEGG cell cycle annotations the recall drops to 40.6%.

Prediction	Recall (%)	Precision (%)	F_1-measure(%)	p-value($<$)
GOA	**55.7**	32.2	40.4	10^{-186}
GOA+KEGG	40.6	**43.7**	**41.4**	10^{-206}

The discrepancy between GOA cell cycle annotations and co-occurrences mined from literature reveals the strengths and weaknesses of both resources. The GOA annotations (Barrell et al. 2009) are the result of manual curation of literature but also of the automatic conversion of annotations of UniProt/Swiss-Prot keywords (Consortium 2009), Enzyme nomenclature (EC numbers) classes, and sequence information such as protein families and domains (Hunter et al. 2009). KEGG annotations are related to biochemical pathways and thus are similar to those provided by the enzyme nomenclature and classification system (Bairoch 1994). In general, not all annotations present in GOA are reflected in the literature – which explains a fraction of the 60% of genes annotated with cell cycle terms but not found by text-mining (Table 14).

The precision is 43.3%, meaning that 56.7% of text-mined cell cycle genes are either wrong or possibly missing from GOA and/or KEGG. Since the manual curation of the literature for associations is a slow and unsystematic process, it is reasonable to postulate that many of the text-mined cell cycle co-occurrences are in fact correct annotations. In the following we support this hypothesis with several examples.

Missing annotations in GOA and KEGG. We give here three examples of genes that have not been formally annotated with GO cell cycle terms but for which we find evidence of their role in cell cycle. As shown in Fig. 88B the genes DDAH2, TERT, KRAS are the only three nodes contained in a power node totaling 15 genes which are not annotated with cell cycle in GOA or KEGG. These 15 genes all co-occur in the literature with TP53 and MYC, but also specific sub-groups co-occur with other important cell cycle genes such as CDK1, CDK2, CDKN1A, and CDKN1B. There is evidence that KRAS – a small GTPase – modulates the cell cycle by both positive and negative regulatory pathways (Fan and Bertino 1997). And the gene TERT encodes for a telomerase reverse transcriptase known to maintain cell cycle by preventing telomere shortening and thus apoptosis (Parkinson et al. 2008).

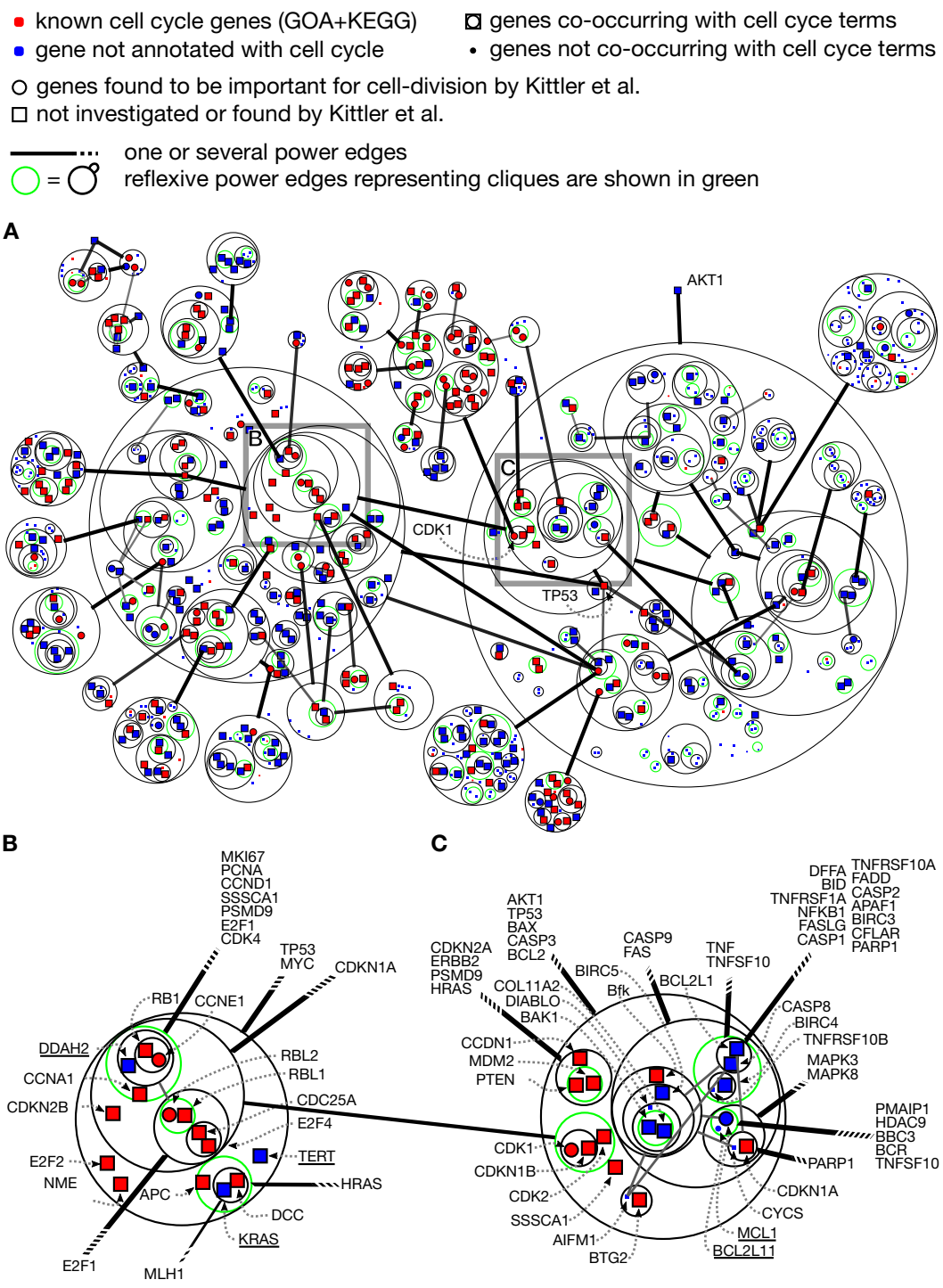


Fig. 88 **Cell cycle genes according to GOA and KEGG compared to genes co-occurring with cell cycle terms. (A)** Same region of the gene–gene co-occurrence power graph as in Fig.87 but with a different layout and a selection of the power edges connecting the biggest power nodes. Most genes are annotated or co-occurring with cell cycle terms. **(B)** Close-up for genes co-occurring with TP53 and MYC. Only three – DDAH2, TERT, and KRAS – are not annotated with cell cycle terms in GOA and KEGG (blue). However, evidence exists in the literature that at least TERT and KRAS play a role during cell cycle (Parkinson et al. 2008; Fan and Bertino 1997). **(C)** Close-up of for genes co-occurring with AKT1, TP53, BAX, CASP3, and BCL2. Two groups can be distinguished depending on whether the genes co-occur with CASP9 and FAS. If they do, most of them are not annotated with cell cycle terms but instead to a closely associated process – apoptosis. Two genes MCL and BCL2L11 forming a power node co-occurring with two MAP kinases MAPK3 and MAPK8 are found by Kittler et al. (2007) to be important for cell division but are not yet annotated as such in GOA or KEGG. And yet, we find MCL1 as often co-occurring with cell cycle terms.

Evidence from a recent cell cycle RNAi screen. Evidence for novel cell cycle genes also comes from a recent genome-wide high-throughput RNAi screen from Kittler et al. (2007) that identified 1351 genes important for cell division in HeLa cells. Among these genes, 243 were already known to be associated to cell cycle progression and 882 were previously associated with other functions. To illustrate how our network analysis corroborates the experimental evidence from Kittler et al. (2007) let us examine two examples: MCL and BCL2L11. The RNAi screen showed that knocking down both genes leads to a G phase arrest in HeLa cells. In Fig. 88C both genes appear in a power node dominated by cell-cycle and apoptosis genes. Both are specifically mentioned with two map kinases MAPK3 and MAPK which belong to a wider group of cancer genes – BCL2L1 and BCL2L11 – co-occurring together with apoptosis genes, CASP9 and FAS. According to Fujise et al. (2000), MCL1 function is to enhance cell survival by inhibiting apoptosis which in turn plays an important role in the regulation of cell cycle progression. Similarly, Morton et al. (2009) presents some indirect evidence for the link between cell cycle regulation and BCL2L11.

5.4.2 Text-mining and protein interactions

In the following, we investigate the possibility of confirming some of the genes from the Kittler et al. (2007) screen with independent evidence from the biomedical literature. In addition to text-mined associations, we test protein homology and protein interactions. In particular, we use HPRD by Prasad et al. (2009) as a source of high-quality manually curated Human protein-protein interactions.

Guilt-by-association. We transfer association to cell-cycle progression to all homologous proteins and direct interaction partners of genes significantly co-occurring with cell-cycle terms in the literature. We experiment with several decision functions but observe that simply transferring to all direct neighbors of a gene performs best.

Results. As shown in Table 15, text-mining alone does not lead to a significant overlap with Kittler genes ($p = 0.51$). The use of sequence homology marginally improves the result's significance ($p = 0.24$). Only by using protein interactions (HPRD) do we find a significant overlap with a p-value of 0.016. Of the 850 novel cell-cycle genes identified in Kittler et al. (2007), 24% can be confirmed by literature mining combined with high confidence protein interaction networks.

Table 15 **Corroborating cell cycle genes using GO term co-occurrence, protein sequence homology and protein interactions.** Confirming the novel cell-cycle genes of Kittler et al. (2007) is not possible using pure text-mining ($p = 0.51$), is only marginally improved using protein sequence homology ($p = 0.24$), but is possible using protein interactions from HPRD ($p = 0.016$).

Confirmation method	Recall (%)	p-value($<$)
Cell cycle term co-occurrence	3.4	0.51
+ protein sequence homology	6.7	0.24
+ protein interactions (HPRD)	**24.3**	**0.016**

Examples. Figure 89 shows a sub-graph of the HPRD network with genes confirmed by our method. For example, let's examine the gene IRF3 – an interferon regulatory factor. It forms a complex with CREBBP and interacts with it in the network (Yang et al. 2002). Moreover, CREBBP co-occurs with cell-cycle terms such as 'DNA replication checkpoint', 'centriole replication', 're-entry into mitotic cell-cycle'. These co-occurrences together with the interaction between IRF3 and CREBBP is our evidence for a link between IRF3 and cell-cycle. The importance of IRF3 for the cell-cycle can be further confirmed – target genes of IRF-3 are themselves involved in cell-cycle as shown by Andersen et al. (2007).

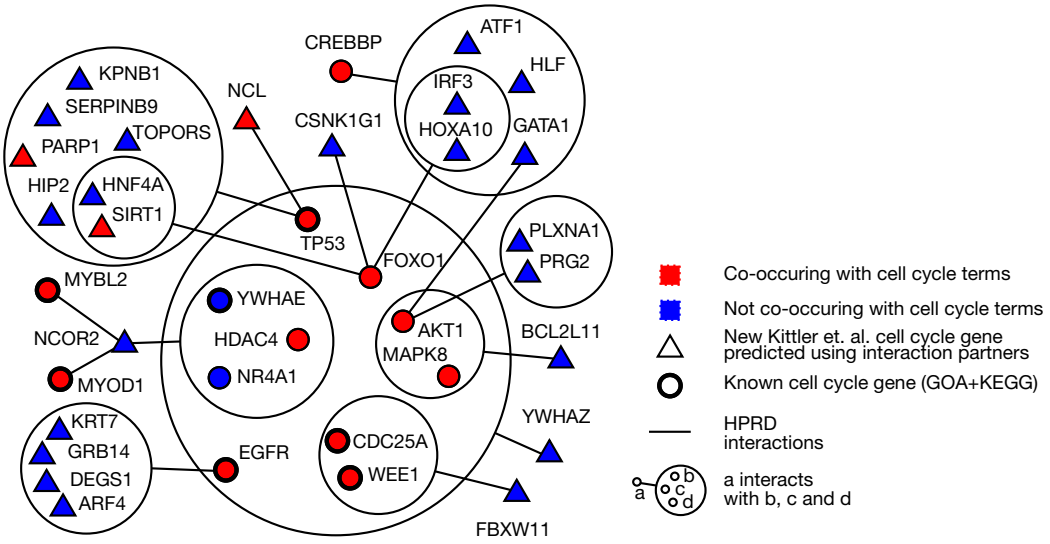


Fig. 89 **Cell cycle genes according to GOA and KEGG compared to genes co-occurring with cell cycle terms.** Example HPRD Sub-network of protein interactions for new cell cycle genes (Kittler et al. 2007) confirmed by guilt by association. Genes like IRF3 and NCOR2 (see upper right power node) can be confirmed using both gene–cell cycle term co-occurrence and high quality protein interactions from HPRD.

5.5 Conclusion

"*Information is not knowledge*"
Albert Einstein

We showed that text-mining of gene mentions has attained a level of confidence sufficient for large-scale mining of gene mentions from the literature. Our current system for gene mention normalization achieves a F_1-measure of 86.5% on the BioCreAtIvE II benchmark dataset. Using this resource, we can construct a large gene-gene co-occurrence network that partially overlaps with known gene and protein interactions. This network contains relevant relationships between genes and proteins that are for example direct or indirect interactions as well as genetic or regulatory interactions. Using power graph analysis we explored this large co-occurrence network and discussed its use as a functional map of genes. In particular, we showed that genes important for the cell-cycle can be identified as co-occurring with several key genes such as CDK1, BCL2, TP53, and CDKN2A. We showed that recovering GOA

annotations by text-mining is feasible with 56% recall and 32% precision. However we also showed that text-mining alone cannot confirm cell-cycle genes found in the RNAi screen from Kittler et al. (2007). Instead, known protein interactions from high quality databases (HPRD) are needed to confirm 25% of novel cell-cycle genes. Another solution to the problem of mining textual knowledge is to make it accessible from the beginning. We proposed already in 2006 to let the authors and editors summarize the main facts in a controlled natural language (Attempto Controlled English) (Kuhn et al. 2006).

Chapter 6

Applications of Power Graph Analysis to the Identification of Regulatory Modules and Pathways

6.1 Introduction

This chapter presents the application of power graph analysis to the identification of regulatory modules and pathways from gene expression microarray data.

First we present the network-based analysis of genome-wide expression profiles of the neuroectodermal conversion of mesenchymal stem cells. We found that HIF-1alpha and miR-124a are master regulators that tightly control a network of deregulated genes. Remarkably, the importance of HIF-1alpha was confirmed experimentally by immunoblotting.

Second, we present the analysis of regulatory modules in a rare mitochondrial cytopathy: *Mitochondrial Encephalomyopathy, Lactic acidosis, and Stroke-like episodes* (MELAS). We investigated the hypothesis that nuclear compensatory responses to mitochondrial mutations can be traced upstream to implicate transcription factors. We also present the putative discovery of a link between MELAS and another rare disease – Sjögren's Syndrome. Two of the transcription factors regulating MELAS genes – IRF-8 and NF-Y – are also known to play a role in Sjögren's Syndrome. Our results suggest that these two transcription factors are the most promising candidates for key regulators in both MELAS and Sjögren's Syndrome.

Third, we investigate the biochemical causes behind the enhanced biocompatibility of tantalum compared with titanium. Our hypothesis is that more reactive oxygen species are released on titanium than on tantalum. This may explain why Human mesenchymal stem cells (hMSCs) cultured on tantalum surfaces reach a steady-state in gene expression levels sooner than on titanium surfaces. In this context we find four key master regulators: P53, NF-Y, IRF-1 and NRF2 involved in sensing and responding to oxidative stress. Corroborating this finding we also detect a strong signal in the selenoaminoacid metabolism pathway – an essential component of the anti-oxidative arsenal of the cell.

6.2 HIF-1alpha and miR-124a as master regulators of mesenchymal stem cells neuroectodermal conversion

During animal embryogenesis, embryonic cells commit to one of three distinct germ layers: endoderm, mesoderm, or ectoderm. Mesenchymal stem cells (MSCs) are mesoderm-derived multipotent stem cells, able to differentiate *in vitro* or *in vivo* into a variety of cell types such as endothelial cells, adipocytes, myocytes, chondrocytes, or osteoblasts. Recently, bone-marrow-derived Human mesenchymal stem cells (hMSCs) were shown to break barriers of germ layer commitment and differentiate *in vitro* into cells with neuroectodermal properties. Hermann et al. (2006) reported on a protocol for the efficient conversion of hMSCs into a neural stem cell like population (Human marrow-derived NSC-like cells or hmNSC).

Here we present work done to investigate the transcriptome alterations during this conversion. The transcriptomes of hMSCs and hmNSCs were obtained by Affymetrix oligonucleotide microarray profiling and analyzed by power graph analysis applied on regulatory and protein interaction networks. The result of the analysis is the identification of regulatory molecules involved in the neuroectodermal conversion process. Two potential master regulators, HIF-1 and microRNA miR-124a, were found. The key role of HIF-1alpha was shown in a follow-up experiment: HIF-1alpha is more active in hmNSCs than in hMSCs.

6.2.1 Background and methods

Neuroectodermal conversion of Human mesenchymal stem cells. Human mesenchymal stem cells (hMSCs) were isolated from bone marrow collected after routine surgical procedures on four adult patients. The conversion of hMSCs into hmNSCs was triggered by culture in hypoxic conditions (3% O_2, 5% CO_2, and 92% N_2) with the addition of growth factors EGF (Epidermal Growth Factor) and FGF-2 (Basic Fibroblast Growth Factor) – see Fig. 90. The neuronal fate of the converted cells was shown by immunostaining of the neural stem cell marker nestin (Fig. 91), and electrophysiological experiments (Hermann et al. 2004).

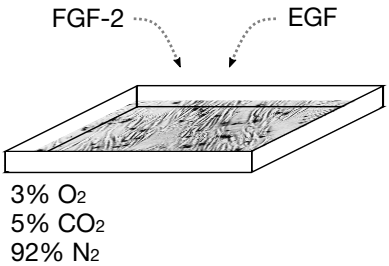


Fig. 90 **hMSCs to hmNSCs conversion protocol.** Human marrow-derived mesenchymal stem cells (hMSCs) are converted into Human marrow-derived NSC-like cells (hmNSC) under hypoxia (3% O_2) and growth factors EGF (Epidermal growth factor) and FGF-2 (Basic fibroblast growth factor) (Hermann et al. 2004).

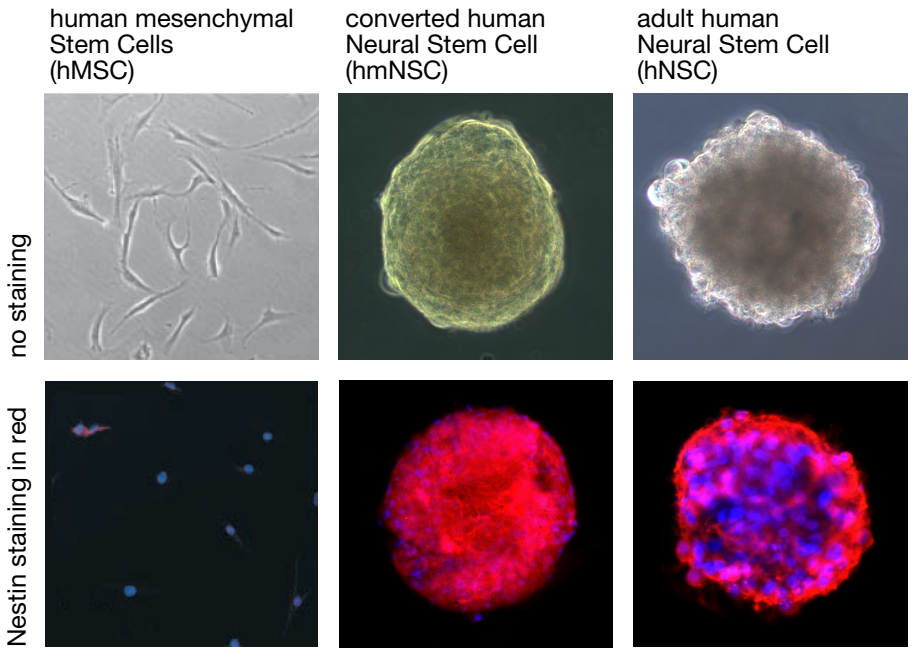


Fig. 91 **Validation of the conversion protocol.** Converted cells (hmNSC) present the neural stem cell marker nestin – in red – characteristic of adult Human neural stem cells (hNSC). Nuclei are counterstained in blue with DAPI. Adapted from Hermann et al. (2004)

Microarray analysis. Microarray analysis was done using Affymetrix U133A chips containing 22,215 probe sets representing at least 12,905 individual genes. Experimental data collected by the Storch research group and data from the NCBI GEO database (Barrett et al. 2009) were pooled and normalized using the RMA-algorithm (Irizarry et al. 2003). Data previously published by Maisel et al. (2007) was used for comparison with primary adult Human NSCs. *Deregulated genes* are defined as having a two-fold higher or lower expression in hmNSCs compared with hMSCs – resulting in 760 up-regulated and 1,001 down-regulated genes.

TRANSFAC and HPRD. The TRANSFAC database (release 11.4) contains data on transcription factors, their experimentally-proven binding sites, and target genes (Wingender 2008). It compiles binding sites and derives positional weight matrices that represent binding site motifs. These motifs are used to predict regulatory links between transcription factors and gene promoters that have not yet been studied in detail. To complement the regulatory information provided by TRANSFAC we used the database of manually curated Human protein-protein interactions of the Human Protein Reference Database (HPRD)[1].

Power graph analysis. The TRANSFAC database was used to build a network linking transcription factors to target genes. This network comprises 1,782 transcription factors and 7,085 target genes. We employed power graph analysis to explore the network. In particular, we investigated the possibility that groups of target genes

1 Prasad et al. 2009.

may be deregulated in concert – thus implicating shared upstream transcription factors. We use the high quality but low coverage TRANSFAC data as a stringent noise filtering on the microarray data.

6.2.2 Results

HIF-1alpha and miR-124a. Within the overall TRANSFAC transcription factor network (Fig. 92A), we identified two regions under shared regulatory control enriched in hMSCs to hmNSCs deregulated genes (Fig. 92B). The two transcription factors regulating these two groups are the HIF-1 transcription factor complex (HIF-1alpha/HIF-1beta also called ARNT) and microRNA miR-124a. In addition, transcription factors of the STAT[2] family – namely STAT1 and STAT3 – were also found to be relevant because of their role in neuronal survival (Dziennis and Alkayed 2008).

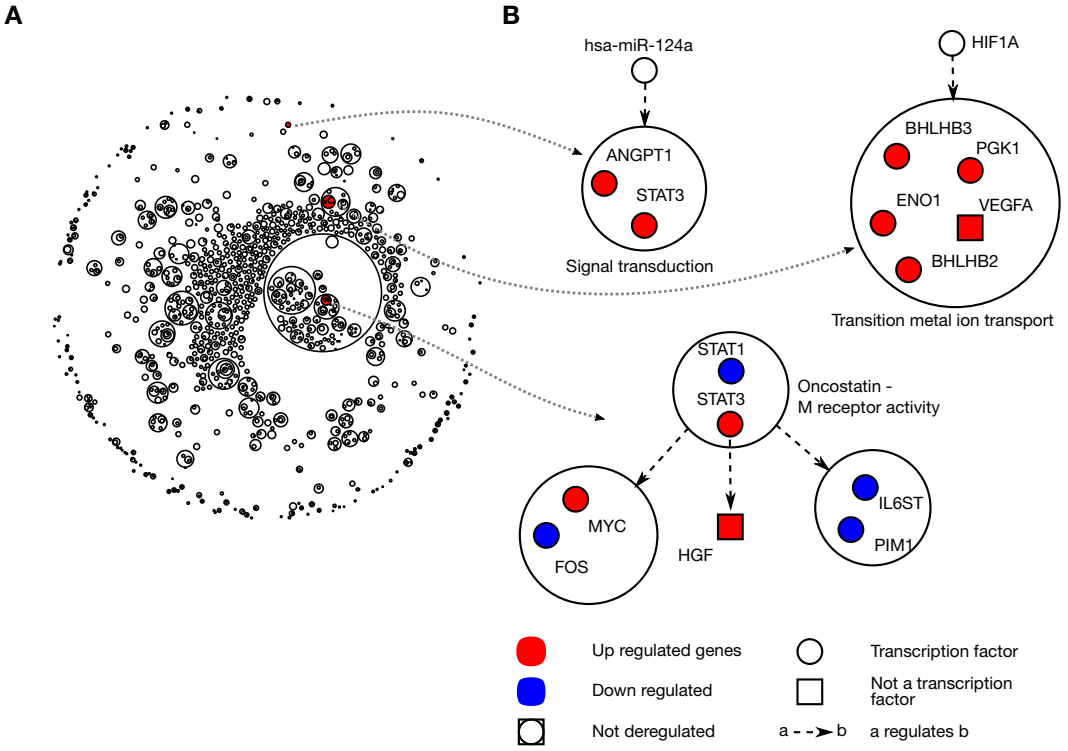


Fig. 92 **Identifying transcription factors upstream of deregulated genes. (A)** Overview of the TRANSFAC network power graph. **(B)** Transcription factors and miRNAs upstream of power nodes containing many deregulated genes. Labels indicate GO terms for which power nodes are significantly enriched ($p < 0.05$).

Integrating protein interactions. To this transcriptional sub-network we add manually curated protein interactions from the HPRD database[3], as well as regulated genes which interact with target genes or transcription factors (Fig. 93). This integrated view shows that HIF-1alpha and miR-124a have common targets such as STAT3 and c-Myc. In particular, c-Myc – which is highly up-regulated (4.7-fold) in

2 Signal transducer and activator of transcription.
3 Prasad et al. 2009.

hmNSCs compared with hMSCs – is a master regulator of the cell cycle, involved in stem cell maintenance (Singh and Dalton 2009).

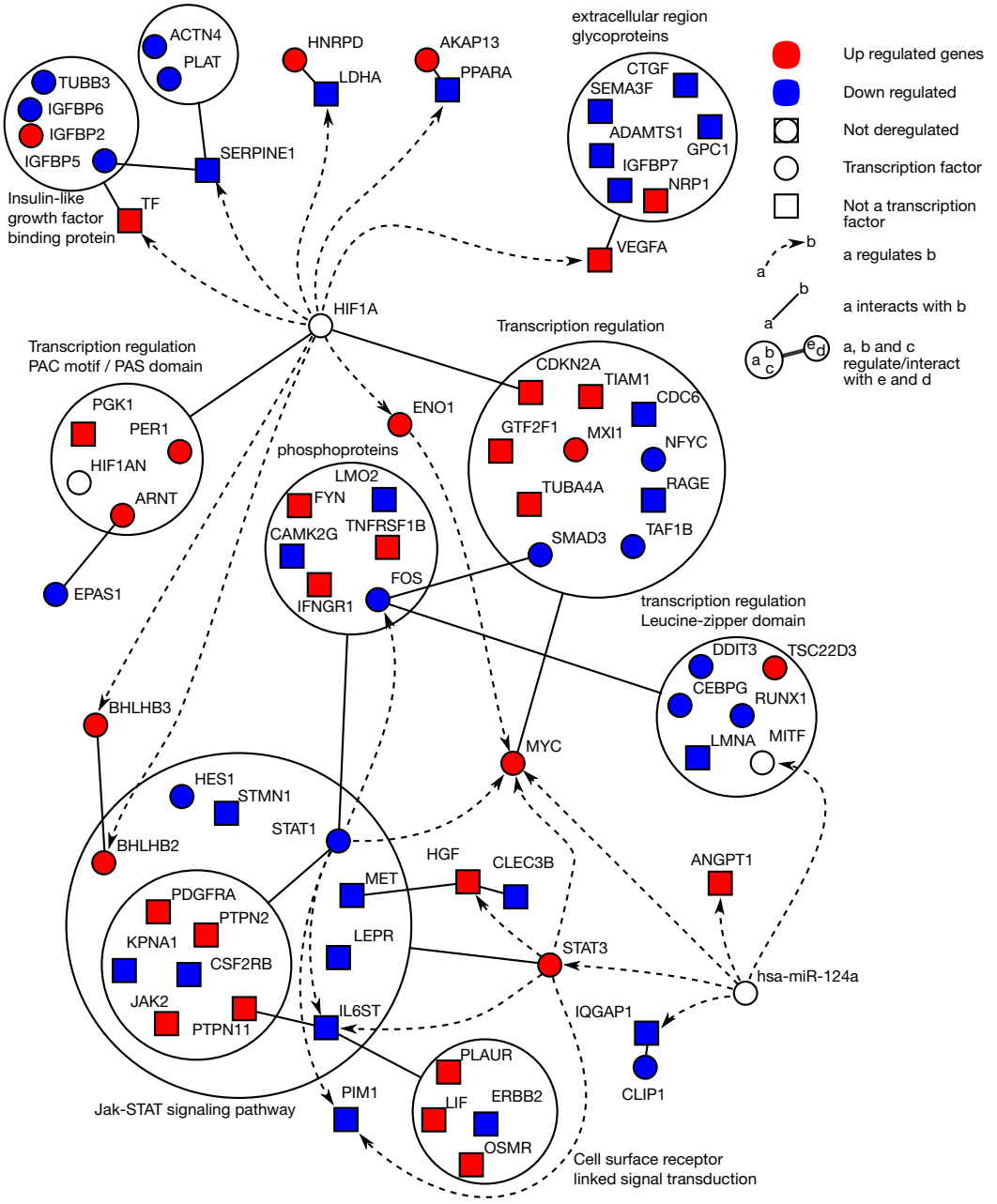


Fig. 93 **Master regulators – HIF-1alpha and miR-124a – and their target genes, complemented with high-confidence protein interactions from HPRD (Prasad et al. 2009).**

6.2.3 Validation

HIF-1alpha: oxygen level sensor. Activity of the HIF-1 complex is regulated by post-translational modifications of HIF-1alpha including its hydroxylation and subsequent degradation (Maxwell et al. 1999; Ivan et al. 2001). As shown in Fig. 94, the hydroxylation of HIF-1alpha is prevented during hypoxia, allowing HIF-1alpha to escape proteolysis, and activate transcription (Ratcliffe 2007). One of the key culture

conditions for hMSCs to hmNSCs conversion is hypoxia (3% O_2). Since the HIF-1 complex is post-translationally regulated, only HIF-1alpha protein levels – and not mRNA levels – are expected to increase. In the following we show immunoblotting results that confirm this hypothesis.

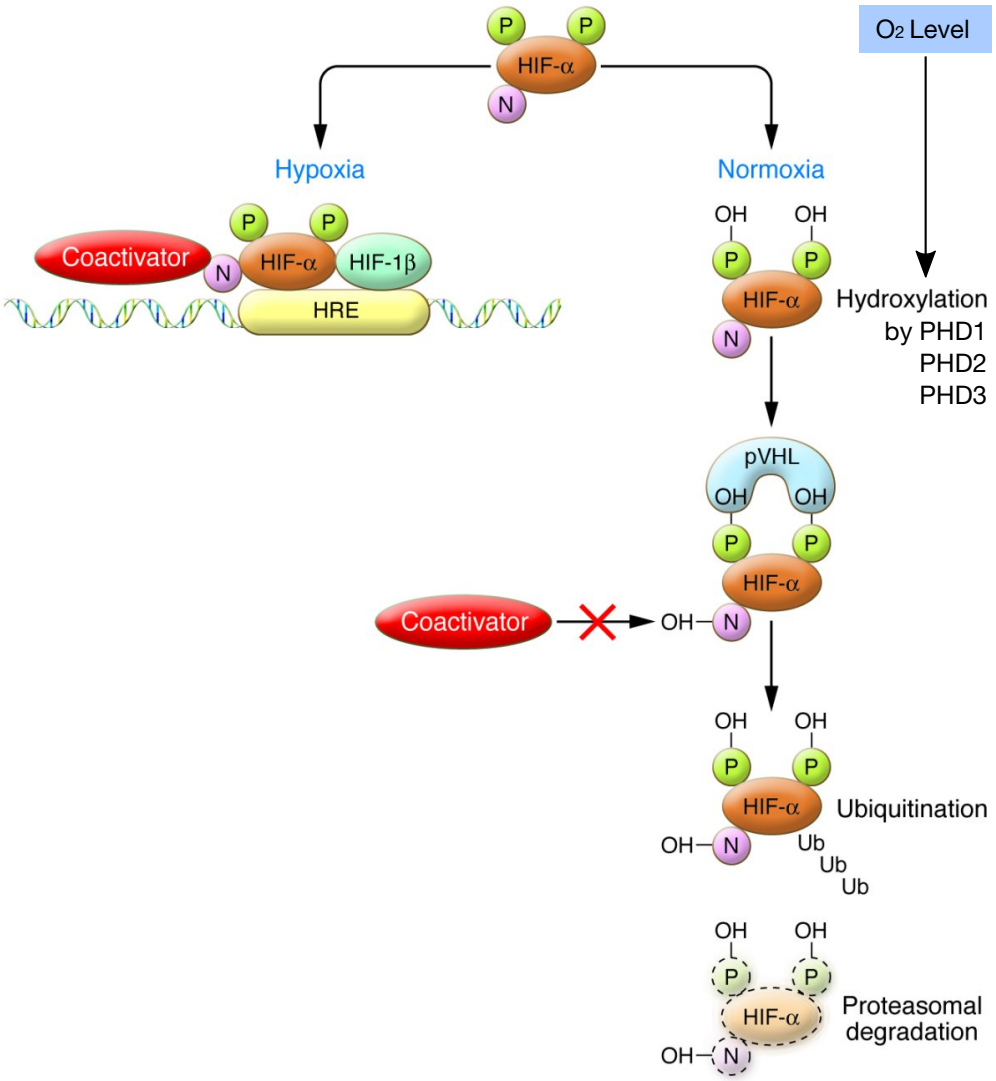


Fig. 94 **HIF activity under hypoxic and normoxic conditions.** In normoxia (normal oxygen level), hydroxylation of two proline residues promotes HIF-alpha binding to pVHL and HIF-alpha destruction via the ubiquitin/proteasome pathway, while the hydroxylation of an asparagine residue blocks association with coactivators. The HIF prolyl hydroxylases, PHD1, PHD2, and PHD3, are responsible for sensing the oxygen level and for consequently hydroxylating HIF-alpha. In hypoxia, these processes are suppressed, allowing HIF-beta subunits (both HIF-1alpha and HIF-2alpha) to escape proteolysis, dimerize with HIF-1beta, recruit coactivators, and activate transcription via hormone response element (HREs). Legend: N, asparagine; P, proline; OH, hydroxyl group; Ub, ubiquitin. Illustration from Ratcliffe (2007)

Experimental validation of HIF-1alpha activity. The stabilization of HIF-1alpha was confirmed by immunoblotting. Fig. 95 shows that level of HIF-1alpha protein is significantly increased in hmNSCs compared with hMSCs (2 fold increase in intens-

ity compared with beta-actin, p-value $< 5\%$). This variation can only be attributed to post-translational regulatory mechanisms – likely those on Fig. 94 – since HIF-1alpha mRNA levels were unchanged upon conversion (micro-array data). This result shows that gene expression data can miss important post-translational aspects of regulatory pathways. Using existing knowledge on regulatory networks, network analysis techniques can help filling the gaps in our understanding. A definitive confirmation of the role of HIF-1alpha would be its knock-down – using RNAi for example – and subsequent analysis of the conversion process. Interestingly, HIF-1alpha hydroxylases PHD1, PHD2 and PHD3 [4] are up-regulated upon conversion. One hypothesis is that there exists a feedback loop that senses the level of HIF-1alpha and adjusts the concentration of the hydroxylases accordingly. Upon conversion the level of HIF-1alpha is higher and therefore the concentration of hydroxylases is adjusted.

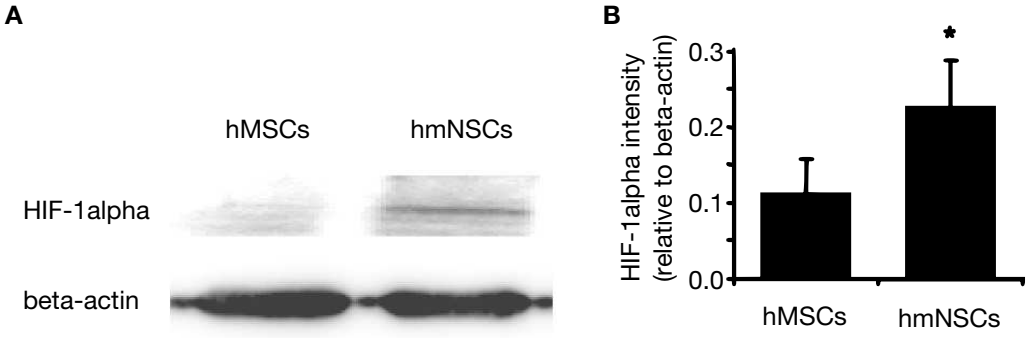


Fig. 95 **Importance of HIF-1alpha confirmed by immunoblotting.** Western blot analysis of HIF-1alpha in hmNSCs compared with hMSCs shows increased expression of HIF-1alpha protein upon neuroectodermal conversion (Maisel et al. 2010). The star (*) represents p-value < 0.05.

MicroRNA master regulator – miR-124a. The second master regulator besides HIF-1 shown in Fig. 93 is miR-124a. MicroRNAs (miRNAs) are short non-coding RNAs (about 22 nucleotides) that have been implicated in fine-tuning gene regulation (Filipowicz et al. 2008). How does miR-124a regulate its target genes? Ribonucleoprotein complexes (miRNPs) – which bring together proteins and miRNAs – have been proposed by Dostie et al. (2003) as a possible mechanism for miR-124a's regulatory role: miR-124a was found to associate with Gemin3, a putative DEAD-box RNA helicases which is a component of the survival of motor neurons (SMN) complex. Another mechanism proposed by Yoo et al. (2009) is the switching of chromatin-remodeling complexes by miR-124a.

Fig. 96 shows that in the last two years more than 15 publications mention miR-124a in the context of neurons (Source: GoPubMed statistics by Dietze et al. (2008)).

4 prolyl hydroxylases 1, 2, and 3

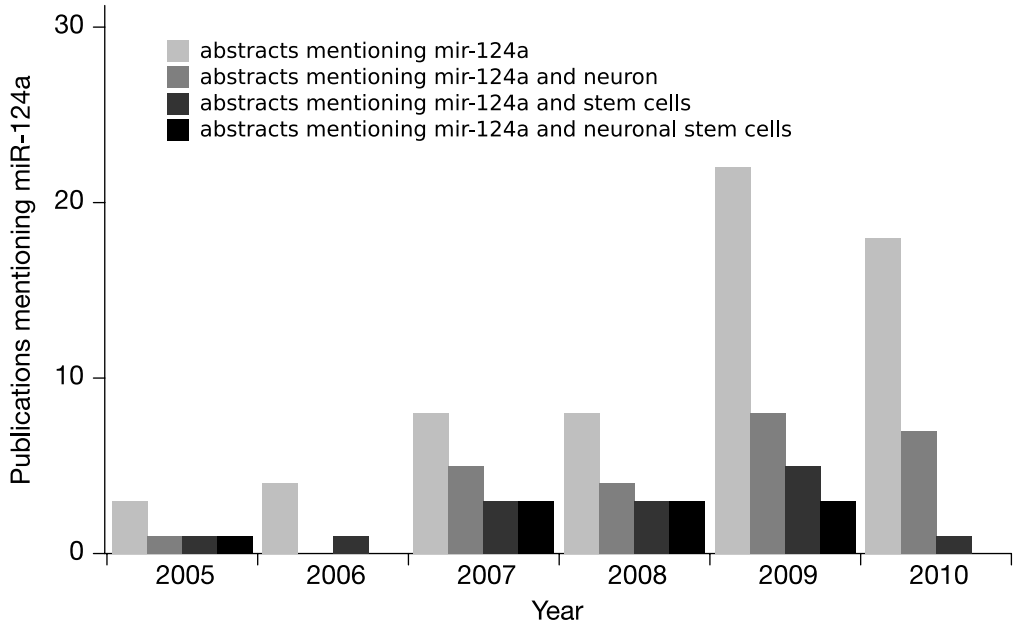


Fig. 96 **MicroRNA mir-124a in the literature.** Mentions of mir-124a in the context of neurons, stem cells, and neuronal stem cells. Source: GoPubMed statistics by Dietze et al. (2008).

Independent evidence for the role of miR-124a in neuronal stem cell differentiation. Several independent studies show that miR-124a has a specific role in neuronal tissue development and maintenance. It has been reported that miR-124a is mainly expressed in the brain, particularly in neurons from the developing and adult nervous system (Lim et al. 2005) and that it modulates embryonic stem cell-derived neurogenesis (Krichevsky et al. 2006). Indeed, miR-124a has been shown to maintain the neuronal phenotype, cell-specific characteristics of neurons, and to be one of the 36 miRNAs uniquely expressed in Human ES cells (Conaco et al. 2006; Suh et al. 2004). Lagos-Quintana et al. (2002) conducted a tissue specific identification of miRNAs from Mouse. Their study shows that miR-124a is a dominant miRNA in the cortex, cerebellum, and midbrain, accounting for 25% to 48% of all miRNAs identified by cloning. Moreover, miR-124a is conserved between invertebrates and vertebrates hinting at its importance. Recently, Cheng et al. (2009) further established the link between miR-124a and neurogenesis in mammalian stem-cell niches.

Role of miR-124a in neurogenesis and neuronal activity. Recently, there has been evidence supporting the role of miR-124a in neuronal tissue and activity. Fischbach and Carew (2009) identified miR-124a's critical role in synaptic plasticity and memory. Clark et al. (2010) showed that miR-124a controls gene expression in the sensory nervous system of Caenorhabditis elegans, and Arora et al. (2010) showed that miR-124a affects mRNA expression during mammalian retinal development.

Together with our own results, these reports confirm the role of miR-124a as a master regulator of gene expression in neuronal tissue.

6.3 MELAS – master regulators and link to Sjögren's Syndrome

Here we present the analysis of regulatory modules in a rare mitochondrial cytopathy (MELAS) caused by a mitochondrial DNA mutation. The goal of this project is to verify the existence of a nuclear compensatory responses to the mutation. We find three candidate master regulators: IRF-8, NF-Y, and HIF-1. We also present the putative discovery of a link between MELAS and another rare disease – Sjögren's Syndrome. Two of the transcription factors regulating MELAS genes – IRF-8 and NF-Y – are also known to play a role in Sjögren's Syndrome. We postulate that these two transcription factors are highly promising master regulator candidates for MELAS nuclear pathway.

6.3.1 Background and methods

MELAS A3243G. *Mitochondrial Encephalomyopathy, Lactic acidosis, and Stroke-like episodes*, or MELAS, is a mitochondrial cytopathy that was first characterized in 1984 by Pavlakis et al. (1984). Mitochondrial cytopathies are primarily caused by mitochondrial DNA (mtDNA) mutations. As illustrated in Fig. 97, an heteroplasmic A>G mtDNA base mutation at nucleotide 3243 in the tRNALeu(UUR) gene has been identified as the main cause for the MELAS syndrome in 80% of the cases (Sproule and Kaufmann 2008; Goto et al. 1990). Mitochondrial mutations are heteroplasmic, which means that wild-type and mutant mtDNA coexist in cells and tissues of patients.

Symptoms and molecular phenotype. MELAS symptoms affects most organs and tissues. In most cases, the first symptoms appear in childhood following a period of normal development. However, some patients suffer a relatively mild disease progression with first symptoms appearing late in life (Chinnery et al. 1997). MELAS patients present a broad spectrum of clinical phenotypes, but have primarily a buildup of lactic acid in their bodies, a condition called lactic acidosis. The subsequent increased acidity in the blood can lead to vomiting, abdominal pain, extreme tiredness, muscle weakness, and difficulty in breathing. Diabetes mellitus and hearing loss can also be part of the syndrome (Chinnery et al. 1997). As shown in Fig. 98, sometimes a microscopic accumulation of abnormal mitochondria can be seen as ragged-red fibers – a typical manifestation of mitochondrial diseases usually found in muscle tissue.

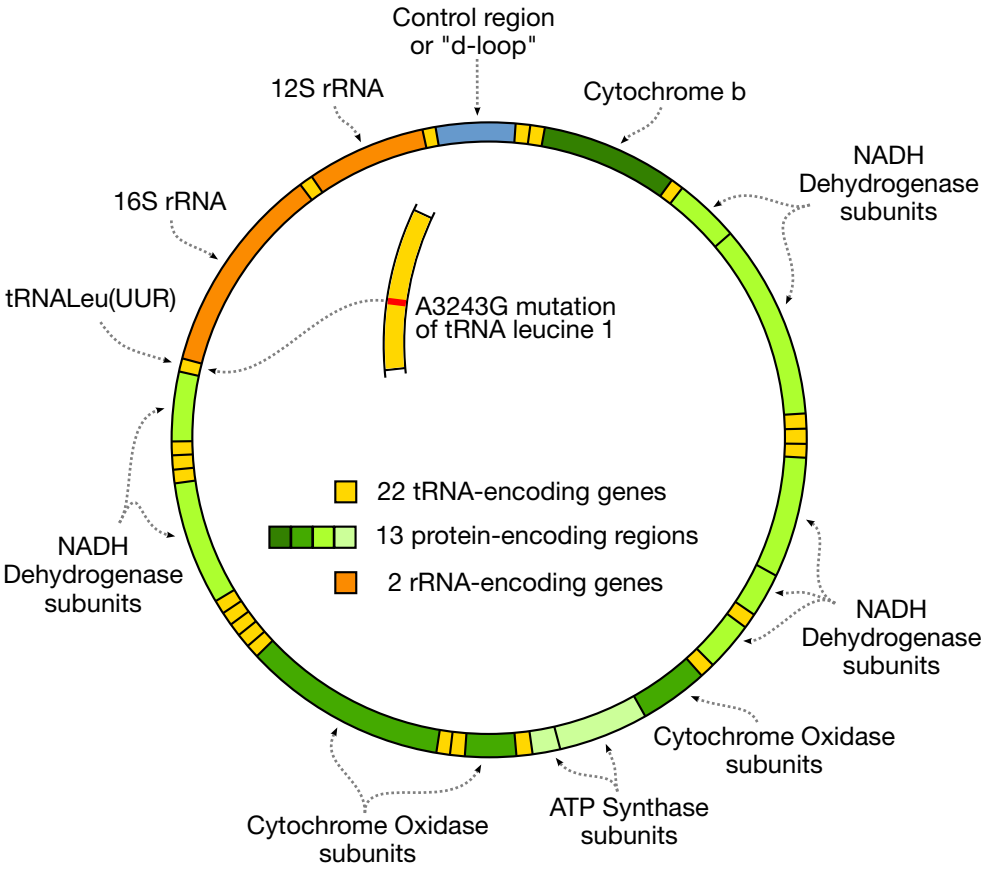


Fig. 97 **Map of the Human mitochondrial DNA (mtDNA).** In humans, 100 to 10,000 mtDNA molecules are present per cell. Mitochondrial DNA has 22 tRNA encoding genes, 2 rRNA encoding genes (16S and 12S), and 13 transport chain encoding genes (Complex I, III, IV, and ATP synthase). The MELAS syndrome is caused by a heteroplasmic base mutation, i.e. not all mtDNA molecules have the mutation. The mutation is the exchange of base G for base A at nucleotide position 3243 of the leucine tRNA (A3243G of tRNALeu) (Goto et al. 1990).

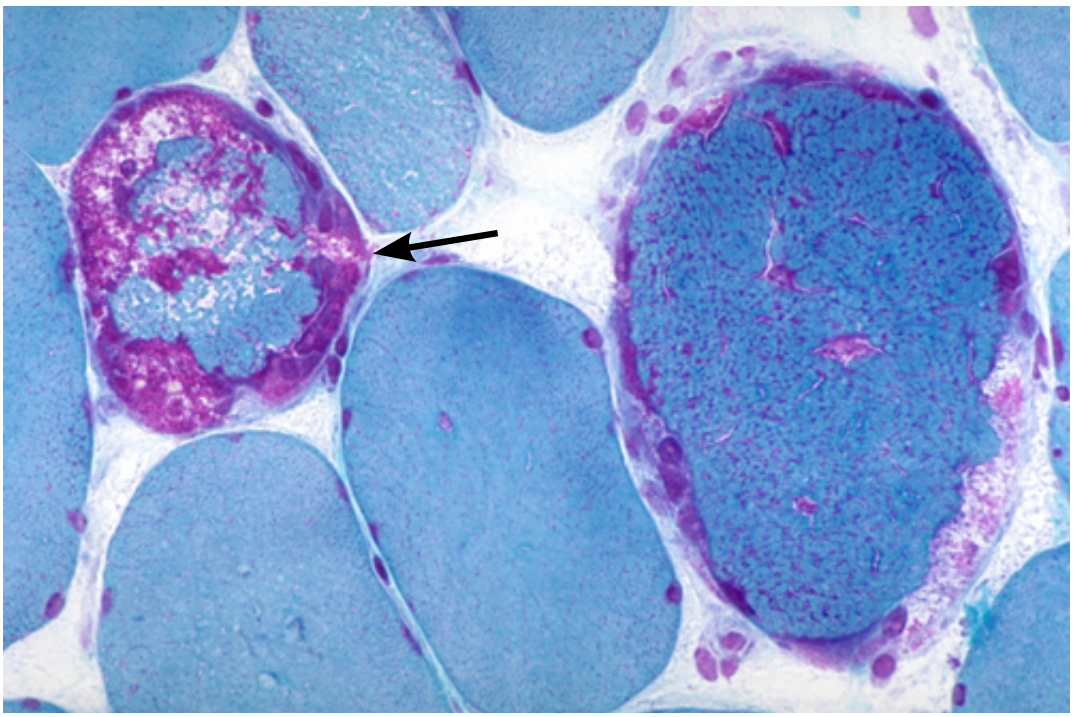


Fig. 98 **Ragged-red fibers hallmark of mitochondrial diseases.** Frozen sections of muscle stained with Gomori trichrome. Myofibers with red stained deposits are called ragged red fibers and are the hallmark of mitochondrial disorders. Photograph by Agamanolis (2009).

Nuclear compensatory response hypothesis. The heteroplasmic nature of mtDNA mutations is assumed to define the clinical outcome of the disease (Chinnery et al. 1997). So far, no association between the proportion of mutant mtDNA – or level of heteroplasmy – and the course or severity of the disease was observed (Karppa et al. 2005). An hypothesis is that secondary mutations or alterations in nuclear gene expression patterns induced by mitochondrion-to-nucleus signaling may play a significant role in the modulation of the primary mtDNA defect.

Genome-wide expression profiling. To test this hypothesis, genome-wide expression profiling experiments were conducted on peripheral blood samples from ten A3243G MELAS patients compared to twenty age- and sex-matched healthy controls. Human Genome U133A 2.0 Arrays (HG-U133A) were used to analyze the expression level of 14,500 well-characterized Human genes. The top 1,000 probesets were selected. The corresponding 789 Human genes showed the highest expression variance between patients and controls.

Hierarchical clustering. Hierarchical clustering of the microarray data showed that patient transcriptome patterns can be classified into three groups. This clustering did not correlate with age at onset, disease duration or disease severity[5], suggesting that these clinical parameters do not have a significant influence on overall

5 as measured with the *Newcastle Mitochondrial Disease Adult Rating Scale.* (NMDAS) see Schaefer et al. (2006)

gene expression patterning (Fig. 99). Furthermore, neither age at blood sampling nor sex can explain the clusters because the controls were matched for age and sex.

Heteroplasmy levels correlate with nuclear genes expression levels. The individual load of mutant mtDNA was measured by a qPCR approach originally described by Nomiyama et al. (2002). This methodology allows the parallel detection of total and mutant mtDNA. Different levels of heteroplasmy were found to correspond to different clusters (Fig. 99). The transcriptomes of the first cluster of patients were found to have mtDNA loads between 20 and 40%, the second cluster of patients had a high mutant mtDNA load, and the third cluster loads were below 20%. This result clearly demonstrates a correlation of the level of heteroplasmy with nuclear gene expression pattern.

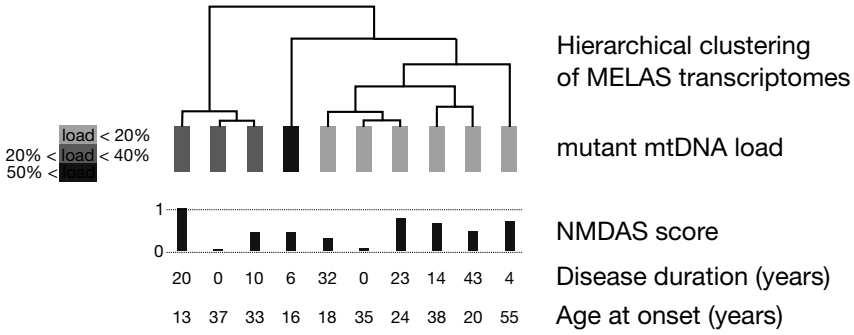


Fig. 99 **Hierarchical cluster analyses of gene expression pattern of MELAS patients.** Hierarchical clustering was done based on the gene expression microarray data of ten A3243G MELAS patients. Pearson correlation of gene expression was used to compare two patients and average linkage was chosen as agglomeration rule. Three clusters could be found to correlate the load in mutant mtDNA. In contrast, the NMDAS scores, disease duration and age at onset did not correlate with the clusters found (Mende et al. 2010).

Question – Which master regulators explain the link between mutant mtDNA and nuclear gene expression? The previous results lead to the hypothesis that the relative abundance of mutant A3243G mtDNA molecules is pivotal for the differences in nuclear gene expression responses. The question is which nuclear master regulators could explain the link between mutant mtDNA load and nuclear gene expression differences. To investigate this hypothesis, power graph analysis was done on an integrated network of transcription factors and protein interactions focused on the 789 genes found to be most deregulated in MELAS.

TRANSFAC, HPRD and power graph analysis. We followed the same methodology as in the previous section (page 155). We used the TRANSFAC-derived network of transcription factors to target genes, and analyzed its power graph with the intent of finding key transcription factors upstream of deregulated genes. In addition, we integrated Human protein-protein interactions from HPRD to complete the network.

State-of-the-art text-mining for gene and disease identification. We use our state-of-the-art text-mining techniques to establish statistically significant links between diseases. For this we rely on statistically significant gene-disease co-occurrences throughout the whole biomedical literature. See chapter 5 page 131 for the details.

6.3.2 Result 1 – MELAS master nuclear regulators

Identification of key transcription factors upstream of deregulated genes. The first observation is that among the 7,085 target genes, only 186 belong to 789 deregulated MELAS genes. Moreover, MELAS genes are scattered throughout the network – only few power nodes contain several deregulated genes. We hypothesize that concentration of deregulated target genes in specific power nodes indicates that the activity of the corresponding transcription factor causes the deregulation. As shown in Fig. 100, power nodes enriched in deregulated MELAS genes were identified manually. The transcription factors regulating these groups of genes are:

- *interferon regulatory factors* (IRFs): IRF-1, IRF-2, and IRF-8,
- *nuclear factor Y*: NF-Y,
- the *hypoxia inducible factor complex*: HIF-1[6],
- *Pituitary Octamer Unc domain class 2 transcription factor 1*: POU2F1[7].

From the proportion of deregulated genes under their control, the most promising transcription factors are IRF-8, NF-Y, and HIF-1.

Protein interactions. To complete the picture and verify whether the four regulatory modules could be integrated into a coherent network, we added manually curated protein interactions from the HPRD database – which is the state of the art database for protein interactions (Prasad et al. 2009). Interactions were added between proteins present in the four regulatory modules or between proteins of the regulatory modules (product of a target genes or transcription factors) as well as other deregulated proteins not initially found in the TRANSFAC network.

MELAS master nuclear regulators. As shown in Fig. 101, the target proteins of the four regulatory modules form a tightly linked protein network regulated by transcriptional regulators: *IRF-2*, and *IRF-8*, *HIF-1alpha/HIF-1beta(ARNT)*, *NF-Y*, and *cAMP responsive element-binding protein (CREB)-related transcription factor* (CREBBP). In addition, PGC1-alpha (PPARGC1A) – a known master regulator of mitochondrial biogenesis – is also part of the integrated network (Wu et al. 1999). This analysis suggests that the regulation of HIF-1alpha/HIF-1beta(ARNT) and NF-Y is altered by the MELAS mutation – both of which are not regulated at the RNA level but instead by post-translational mechanisms (Manni et al. 2008).

6 HIF-1alpha/HIF-1beta complex also called ARNT

7 also called Oct-1

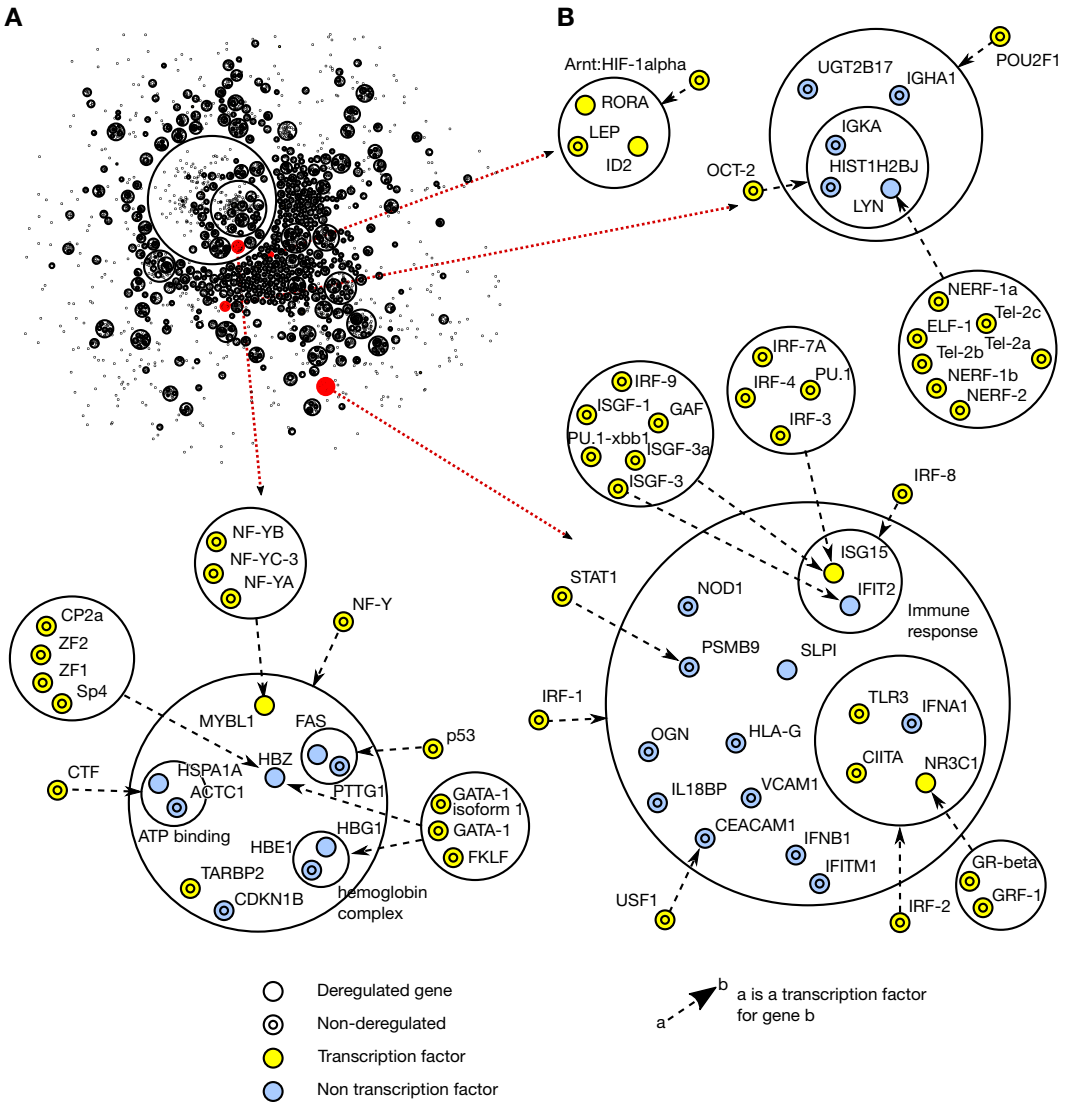


Fig. 100 **Key transcription factors from TRANSFAC upstream of MELAS deregulated genes. (A)** Power nodes of the TRANSFAC power graph. Power nodes enriched in MELAS deregulated genes are marked in red. **(B)** Power nodes enriched in deregulated genes are shown together with the corresponding regulating transcription factors.

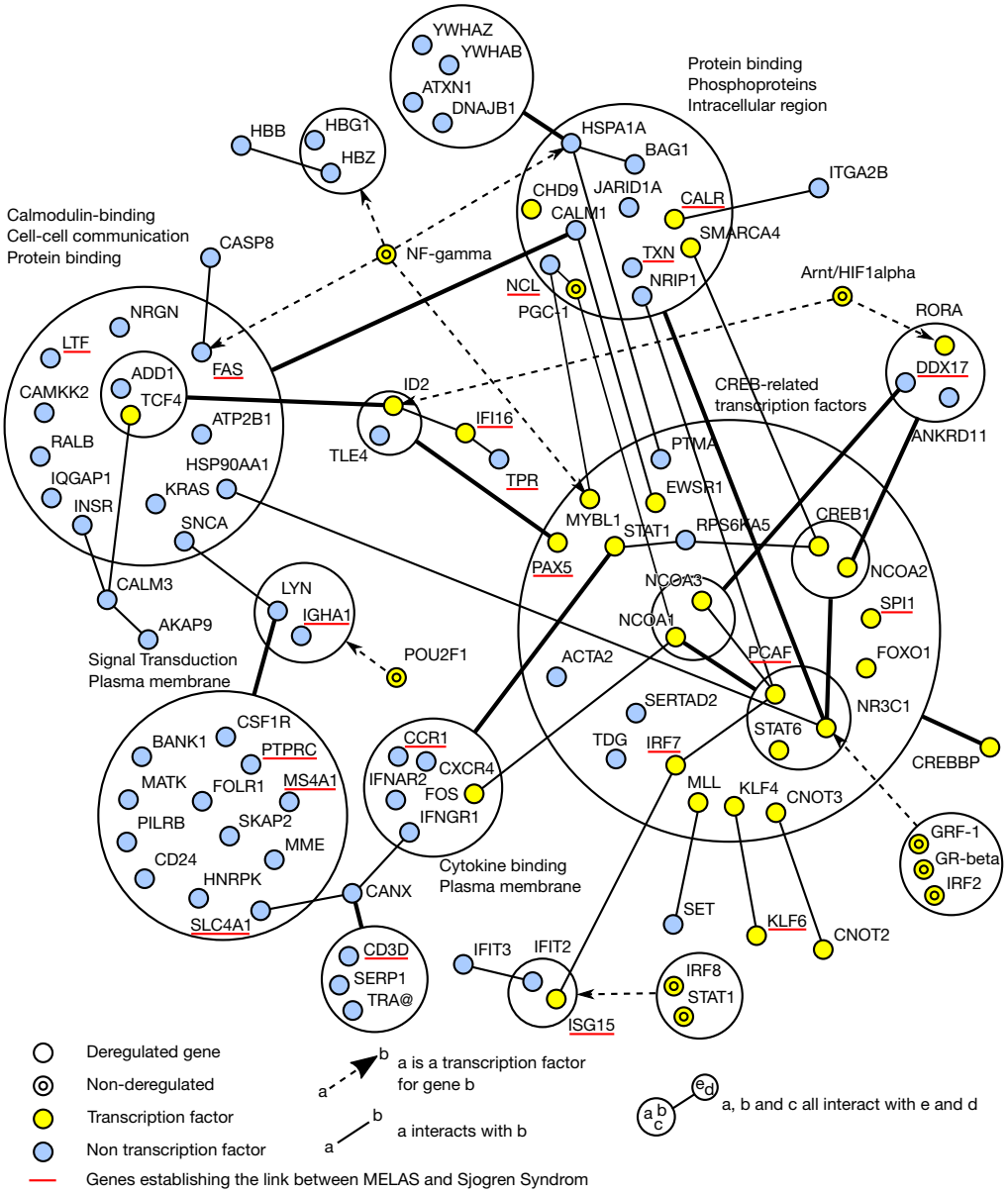


Fig. 101 **Key transcription factors upstream of MELAS deregulated genes in the context of their protein interactions.** Power graph of the network induced by transcription factors and target gene products complemented by protein interactions from HPRD. Note the two kinds of edges in the network: protein interactions and regulatory interactions. Genes establishing the link to Sjögren's Syndrome are underlined in red.

6.3.3 Result 2 – Link to Sjögren's Syndrome

Since MELAS is a rare (prevalence 0.27%) and under-recognized disease (Manwaring et al. 2007), finding other diseases with similar causative genes and pathways could provide complementary information. Here we discuss our discovery of a link between MELAS and Sjögren's Syndrome – a slightly more common (prevalence: 1.3%) autoimmune disease. Sjögren's Syndrome also known as *Mikulicz disease* or *Sicca syndrome* is an autoimmune disease, destroying the exocrine glands that produce tears and saliva (Delaleu et al. 2008).

Linking MELAS deregulated genes with Sjögren's Syndrome genes found by literature-mining. To establish this link we obtained data on statistically significant gene-disease co-occurrences throughout the whole biomedical literature. MELAS deregulated genes and disease terms are identified in biomedical abstracts. Diseases that are consistently mentioned together with MELAS deregulated genes are reported and a p-value is computed to evaluate the statistical significance[8].

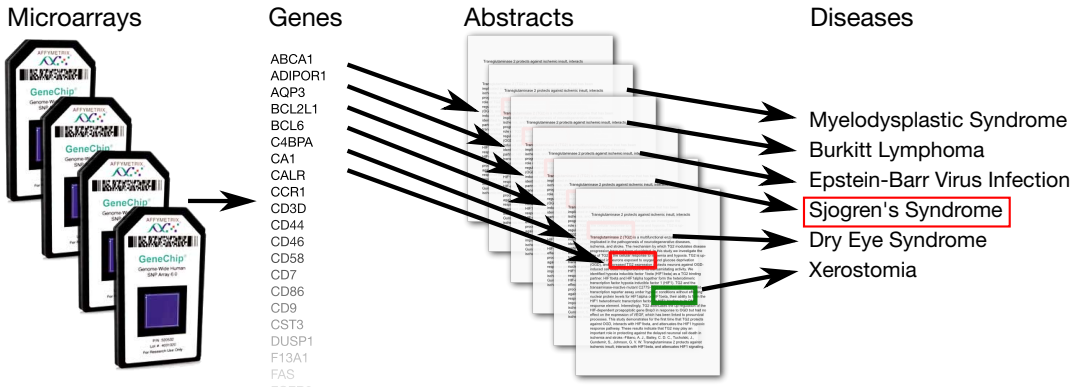


Fig. 102 **From deregulated genes in MELAS to Sjögren's Syndrome.** We identify the MELAS deregulated genes in abstracts from the biomedical literature and find diseases that are mentioned together with these genes.

The result is shown in Table 16. MELAS deregulated genes are enriched in two categories of diseases: i) exocrine gland related and ii) blood related. The occurrence of genes related to blood diseases could be explained by the peripheral blood origin of the samples and thus might be an experimental artifact. However, the first category of diseases – and in particular the link to Sjögren's Syndrome – is more intriguing. The genes explaining this connection are only partially overlapping with the blood related genes (40% see Table 17) which reduces the likelihood that it is an artifact of the collection procedure. About 8% of the MELAS deregulated genes establish a statistically significant link ($p = 1.44 \times 10^{-5}$) to Sjögren's Syndrome.

8 The p-value is computed using the hypergeometric distribution, see page 143 for more details.

Table 16 **MELAS genes enrichment for diseases.** Two categories of diseases – exocrine gland related and blood related – are often mentioned in the literature together with MELAS deregulated genes. Of particular interest is the link between MELAS and Sjögren's Syndrome. Xerostomia or dry mouth disease is a closely related disease to Sjögren's Syndrome.

Disease category	Disease	Coverage (%)	*p*-value
exocrine glands	Sjögren's Syndrome	7.76	1.44×10^{-5}
exocrine glands	Xerostomia	7.86	2.08×10^{-4}
blood	Myelodysplastic Syndromes	19.25	1.49×10^{-5}
blood	Bone Marrow Diseases	18.66	2.25×10^{-5}
blood	Leukemia	19.65	8.53×10^{-4}

Linking genes. An example of a gene establishing a link between MELAS and Sjögren's Syndrome is CD3D. Hjelmervik et al. (2005) investigated gene expression in salivary glands of Sjögren's Syndrome patients and controls. We cite: "CXCL13 and CD3D were expressed in $\geq$ 90% of primary Sjögren's Syndrome patients and in $\leq$ 10% of the controls." This excerpt shows that CD3D is also implicated in Sjögren's Syndrome. The same holds for all genes in Table 17.

Table 17 **List of 68 Human genes found to establish a statistically significant link between MELAS and Sjögren's Syndrome.** Only 40% of the genes establishing the link are blood related. In the table legend *blood* refers to genes that establish a link to blood related diseases; *network* refers to genes that are also found in the integrated network of Fig. 101 (underlined in red). Note: the 'at' in IGH@ is not a typo but refers to the *immunoglobulin heavy locus* gene.

gene name	id	blood	network	gene name	id	blood	network
ABCA1	19			IL6R	3570	✔	
ADIPOR1	51094	✔		IRF7	3665		✔
AQP3	360			KLF6	1316		✔
BCL2L1	598	✔		KRT1	3848		
BCL6	604			LCN2	3934		
C4BPA	722			LTF	4057	✔	✔
CA1	759			LYZ	4069	✔	
CALR	811		✔	MMP9	4318		
CCR1	1230	✔	✔	MS4A1	931	✔	✔
CD3D	915	✔	✔	MX1	4599		
CD44	960			NCL	4691	✔	✔
CD46	4179			NUP210	23225		
CD58	965	✔		ORM1	5004		
CD7	924	✔		PAX5	5079	✔	✔
CD86	942	✔		PPBP	5473	✔	
CD9	928	✔		PRDX2	7001	✔	
CST3	1471			PSMB1	5689		
DUSP1	1843			PTPRC	5788	✔	✔
F13A1	2162			S100A8	6279		
FAS	355		✔	SLC4A1	6521		✔
FCER2	2208			SLPI	6590		
FCGR3A	2214	✔		SNRPD3	6634		
FCGR3B	2215	✔		SP110	3431	✔	
GZMA	3001			SPI1	6688	✔	✔
H1F0	3005	✔		SPTBN1	6711		
HLA-DQA1	3117			SSB	6741		
HLA-DQB1	3119			STAT1	6772	✔	
HLA-DRB1	3123			TFRC	7037	✔	
HLA-DRB3	3125			TLR8	51311		
HLA-DRB4	3126			TMEM1	7109	✔	
IFI16	3428	✔	✔	TPP1	1200		
IGH@	3492	✔		TPR	7175		
IGHA1	3493		✔	TXN	7295		✔
IGHD	3495			XCL1	6375		

Independent evidence for a link between MELAS and Sjögren's Syndrome. Lindal et al. (1992) searched for mitochondrial alterations in 472 muscle biopsy specimens. For 49 patients they found abnormal accumulation of mitochondria. For several of these patients they could establish a genetic inheritance pattern: three had MELAS and seven had Sjögren's Syndrome. This shows that Sjögren's Syndrome is also characterized by mitochondrial dysfunction similar to MELAS: ragged-red fibers were found in at least two patients as reported by Torbergsen et al. (1991). This independent evidence confirms the plausibility of the link between MELAS and Sjögren's Syndrome, and offers promising avenues for understanding the molecular mechanisms of MELAS. Moreover, note that 30% (20) of Sjögren's Syndrome genes appear in the network on Fig. 101.

Common master regulators between MELAS and Sjögren's Syndrome? Two of the MELAS regulators also regulate two Sjögren's Syndrome genes: NF-Y regulates FAS and IRF8 regulates ISG15 (see Fig. 101 where Sjögren's Syndrome genes are underlined in red). While only indirect evidence implicates NF-Y to Sjögren's Syndrome (Cha et al. 2004), several reports implicate its target FAS[9]. One of the key proteins in Sjögren's Syndrome is Ro52 – a E3 ligase known to be a major target of autoantibodies (Rhodes et al. 2002). Ro52 is known to mediate ubiquitination of several members of the interferon regulatory factor (IRF) family (Espinosa et al. 2009). In particular, Kong et al. (2007) showed that IRF-8 interacts with Ro52. In contrast, little evidence could be found implicating the other two regulators HIF-1 and POU2F1 in Sjögren's Syndrome. We conclude that IRF-8 and NF-Y are the two most promising candidates for master regulators in MELAS and Sjögren's Syndrome. Both IRF-8 and NF-Y transcriptionally regulate MELAS genes and are implicated in Sjögren's Syndrome – a disease that we postulate to share key molecular mechanisms with MELAS.

6.3.4 Discussion

We found evidence that the underlying mtDNA mutation causing MELAS (A3243G of tRNALeu) is linked to a nuclear compensatory response altering the regulatory effect of transcription factors CREBBP, HIF-1alpha/HIF-1beta, IRF-8 and NF-Y. These factors are known to be involved in cellular energy homeostasis and responses to energy failure. Differences in these responses could explain the clinico-genetic diversity of MELAS patients. Moreover, we found a link between MELAS and another rare disease: Sjögren's Syndrome. This statistically significant link is supported by 68 genes that are deregulated in MELAS and consistently mentioned in the literature in the context of Sjögren's Syndrome. Remarkably, two of the transcription factors involved in the MELAS nuclear compensatory response – IRF-8 and NF-Y – are also known to be implicated in Sjögren's Syndrome.

9 Saito et al. 1999; Tsuzaka et al. 2007; Bolstad et al. 2000; Mullighan et al. 2004.

6.4 Superior Biocompatibility of Tantalum versus Titanium

Metallic bone implants are commonly applied in the fields of orthopedic surgery and dentistry. An implant's durable osseointegration requires that bone precursor cells attach, proliferate and differentiate on the implant surface. Stiehler et al. (2008) previously observed that tantalum (Ta) exhibits superior biocompatibility than titanium (Ti) when tested on Human mesenchymal stem cells (hMSCs). Since both surfaces tested are topographically smooth, this difference must be attributed to chemical reactions occurring at the interface between metal and cells.

The aim of this study is to understand the differences in biocompatibility between titanium and tantalum using gene expression time-series for hMSCs cultured on the respective surfaces. We focused on the top 1,000 most temporally varying genes and found that on tantalum the regulatory response reaches a steady-state sooner than on titanium. Moreover, we find several genes and pathways that exhibit a differential response to tantalum versus titanium. First, we find key transcription factors: NRF2, EGR1, IRF-1, NF-Y and P53 involved in cell response to oxidative stress. Second, we find a metabolic pathway at the heart of the selenoaminoacid metabolism which is an essential part of the anti-oxidative machinery of the cell. These results suggest that higher concentrations of reactive oxygen species at the titanium-MSCs interface cause oxidative stress that delay the attachment, proliferation, and differentiation of MSCs on titanium surfaces compared with tantalum.

6.4.1 Background and methods

Biocompatibility of titanium and tantalum. Metallic implant materials are widely used in the field of orthopedics, oral and maxillofacial surgery. Titanium (Ti) metal is well known for its biocompatibility with bone tissue and is the most widely used bone implant material (Tschernitschek et al. 2005). The cells attach and adhere to the metallic implant surface, proliferate, and differentiate into osteoblasts. The durable osseous fixation of the implant is promoted by extracellular matrix mineralization (Groessner-Schreiber and Tuan 1992).

The promises of tantalum. Tantalum (Ta) is another promising but less used biomaterial. Tantalum in metal form has been used for plates, suture wires, and radiographic bone markers in limited areas of orthopedic and craniofacial surgery with excellent results for more than 60 years (Matsuno et al. 2001). Studies have shown that it is comparable if not superior to titanium's chemical and mechanical properties, including high malleability, ductility, corrosion resistance, low solubility and low toxicity (Balla et al. 2010). Under normal conditions, metals such as aluminum, titanium, tantalum do not corrode because a thin and protective oxide layer forms spontaneously on the surface exposed to oxygen or humidity. Like titanium, tantalum's electrically non-conductive oxide layer prevents electron exchange, and thus any oxidoreduction (redox) reactions from occurring. It is known that a low redox reaction rate is import-

ant for the biocompatibility of materials – redox reactions can denature proteins and prevent osteointegration (Zitter and Plenk 1987).

Assessing biocompatibility with mesenchymal stem cells. One way to assess the biocompatibility of a material is to study its interaction with mesenchymal stem cells (MSCs). hMSCs play a crucial role in the process of bone regeneration and biomaterial fixation (Mistry and Mikos 2005; Bruder et al. 1998). Stiehler et al. (2008) conducted the first study comparing the interactions of Human hMSCs with smooth, clean and well-characterized titanium, tantalum, and chromium surfaces (Fig. 103). They observed that hMSCs adherent to smooth tantalum (Ta) surfaces demonstrate superior biocompatibility compared with titanium and chromium.

Controlling for surface differences between titanium and tantalum. Atomic force microscopy (AFM) is a scanning probe technique suitable for analyzing surface topography on the nanometer scale (Binnig et al. 1986). AFM brings a sharp tip at the end of a cantilever close to the surface, recording the interaction forces between the tip and the surface in the *pN* regime. Scanning the surface produces high-resolution topographical images. Stiehler et al. (2008) showed with AFM that both coatings have a root-mean-square (RMS) roughness of less than 1.5*nm*. Since both tantalum and titanium surfaces tested are smooth, the only differences are necessarily chemical in nature. Moreover, using X-ray photoelectron spectroscopy, Stiehler et al. (2008) also showed that the native oxide layers have a thickness of 2 to 5*nm*, with no contaminants apart from carbon atoms.

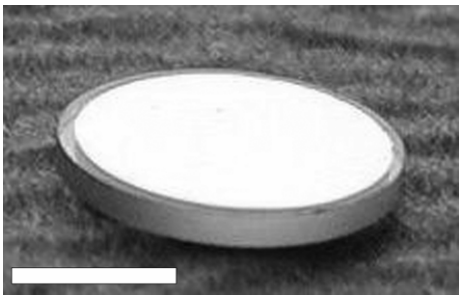


Fig. 103 **Macrograph of metal-coated float glass disc.** Bar = 20 mm. Figure adapted from Stiehler et al. (2008).

Evaluating biocompatibility with immortalized Human hMSCs. Human hMSCs were immortalized by transfection with a retroviral vector carrying the TERT gene encoding the catalytic subunit of the Human telomerase. Cells were cultivated on coated discs (6 replicates per coating type and time point) using minimal essential medium (MEM) supplemented with 1% fetal bovine serum to reduce concentration of interfering proteins. The cells were maintained in a humidified atmosphere of 37℃ and 5% CO2. After cell cultivation for 1, 2, 4, or 8 days, RNA was pooled – resulting in 2 pooled replicate RNA samples per time point.

Gene expression analysis. mRNA levels were measured using Human Genome U133 Plus 2.0 arrays (Affymetrix) containing more than 22,000 probesets. Normalization of data was done using the RMA method (Gentleman et al. 2004; Irizarry et al. 2003). Probes were mapped to 17,507 unique Entrez gene identifiers. Expression values from replicate arrays were averaged. The top 1,000 genes of highest temporal expression variance for Ta/Ti ratios were selected. In Table 18 we show the 20 most up-regulated genes on Ta after 4 days of cultivation. Many of those genes are involved in processes such as cell-adhesion, cell migration, and bone formation.

Table 18 **The 20 most up-regulated genes on Ta after 4 days of cultivation.** i) Genes associated with the processes of cell adhesion and cell migration. ii) Genes involved in the process of bone formation.

Gene name	Function	Fold	
parathymosin	DNA replication	3.09	
Rho GDP dissociation inhibitor (GDI) alpha⋆	cell motion, cell adhesion	2.55	
plectin 1	cytoskeleton-membrane attachment	2.44	i
chromosome 20 open reading frame 149	unknown	2.43	
myosin regulatory light chain	regulation of muscle contraction	2.28	
adaptor-related protein complex 2, alpha 1 subunit	endocytosis	2.28	
talin 1	cell motion, cell adhesion	2.25	i
tissue non-specific alkaline phosphatase	ossification	2.22	ii
cortexin 1	membrane	2.21	
zink finger protein, multitype 1	metal ion binding	2.2	
sorbin and SH3 domain containing 3	cell-substrate adhesion	2.17	i
interferon induced transmembrane protein 3 ($1-8U$)	response to biotic stimuli	2.15	
FOS-like antigen 1	chemotaxis	2.13	i
cysteine-rich protein 1 (intestinal)	metal ion binding	2.12	
zyxin	cell adhesion, metal ion binding	2.12	i
interferon induced transmembrane protein 1 ($9-27$)	negative regulation of cell proliferation	2.1	ii
collagen, type VI, alpha 1	extracellular matrix-receptor interaction	2.1	i, ii
ADAM metallopeptidase	metal ion binding, extracellular matrix synthesis	2.09	ii
basigin (Ok blood group)	cell surface receptor linked signal transduction	2.08	
SAC3 domain containing 1	cell division	2.07	

Table 19 **Known half-reactions of titanium, chromium and tantalum compounds (Bard et al. 1985).** Each reaction read from left to right is a reduction, from right to left an oxidation (oxidation is loss, and reduction is gain of electrons). E° is the standard electrode potential of reversible redox half-reactions. The bigger the E°, the easier the compound can be reduced and the more oxidizing it is to other compounds. In the context of titanium and tantalum oxide layers two reactions and the corresponding oxides are relevant: TiO_2 and Ta_2O_5. Note that the tantalum oxide is less susceptible to be an oxidizer than the titanium oxide because it has a lower E°. An example of poor biocompatible material is chromium which is known to be carcinogenic and to have highly oxidizing and soluble oxides (Cohen et al. 1993).

Half-reaction			E°(Volt)
$Ti^{2+} + 2e^-$	$\rightleftharpoons$	$Ti(s)$	-1.63
$TiO(s) + 2\,[\mathrm{U{+}200A}]\,H^+ + 2\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$Ti(s) + H_2O$	-1.31
$Ti_2O_3(s) + 2\,[\mathrm{U{+}200A}]\,H^+ + 2\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$2\,TiO(s) + H_2O$	-1.23
$Ti^{3+} + 3\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$Ti(s)$	-1.21
$TiO^{2+} + 2\,[\mathrm{U{+}200A}]\,H^+ + 4\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$Ti(s) + H_2O$	-0.86
$2\,[\mathrm{U{+}200A}]\,TiO_2(s) + 2\,[\mathrm{U{+}200A}]\,H^+ + 2\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$Ti_2O_3(s) + H_2O$	-0.56
$TiO^{2+} + 2\,[\mathrm{U{+}200A}]\,H^+ + e^-$	$\rightleftharpoons$	$Ti^{3+} + H_2O$	$+0.19$
$Ta_2O_5(s) + 10\,[\mathrm{U{+}200A}]\,H^+ + 10\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$2\,[\mathrm{U{+}200A}]\,Ta(s) + 5\,[\mathrm{U{+}200A}]\,H_2O$	-0.75
$Ta^{3+} + 3\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$Ta(s)$	-0.60
$Cr^{3+} + 3\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$Cr(s)$	-0.74
$Cr^{3+} + e^-$	$\rightleftharpoons$	Cr^{2+}	-0.42
$CrO_4^{2-} + 4H_2O + 3e$	$\rightleftharpoons$	$Cr(OH)^{3+}(s) + 5OH^-$	-0.13
$Cr_2O_7^{2-} + 14\,[\mathrm{U{+}200A}]\,H^+ + 6\,[\mathrm{U{+}200A}]\,e^-$	$\rightleftharpoons$	$2\,[\mathrm{U{+}200A}]\,Cr^{3+} + 7\,[\mathrm{U{+}200A}]\,H_2O$	$+1.33$

Hypothesis – The stability of tantalum's oxide layer explains its enhanced biocompatibility versus titanium. The main aspect that influence the likelihood of redox reactions on the titanium and tantalum surfaces is the oxide layer stability. Table 19 summarizes the known oxidoreduction half-reactions involving titanium and tantalum. In principle, tantalum oxide layer (Ta_2O_5) is more stable and less oxidizing than its titanium counterpart (TiO_2). More direct experimental evidence of the superior stability of the tantalum oxide layer compared with other metals including titanium is provided in Fig. 104. Zitter and Plenk (1987) tested the resistance to current flow of different metal-solutions interfaces in increasing potentials. The results show that tantalum surface oxide layer is more stable than titanium or nobium oxide layers.

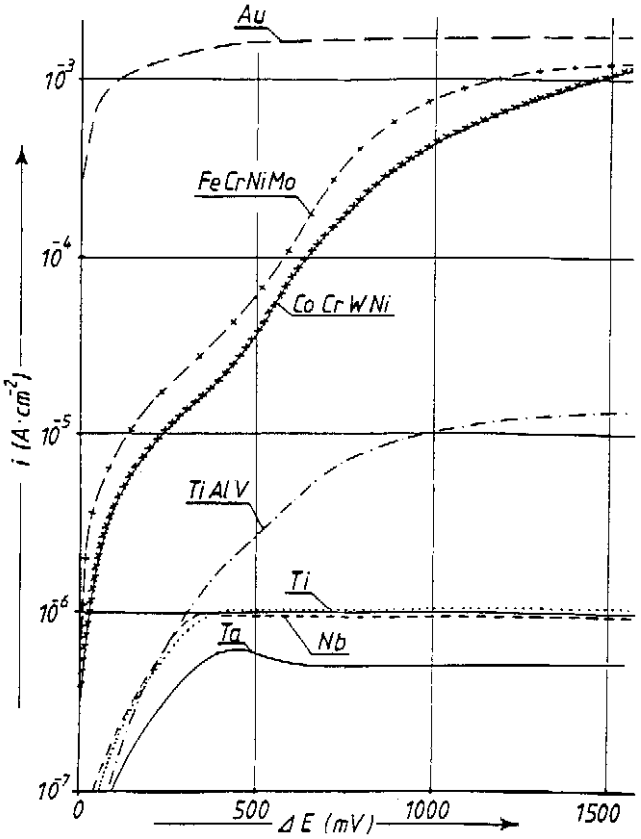


Fig. 104 **Current density as a function of the potential difference in a saline solution for different metals (Ti, Ta, Nb, Au) and alloys (FeCrNiMo, CoCrWNi, and TiAlV).** This curve reflects how resistant the different metal-solution interfaces are to current flow. Tantalum, titanium and nobium have the lowest currents because of their highly dense, stable, and dielectric oxide layers that prevent electron circulation and redox reactions from occuring. In contrast, gold (Au) exhibits the lowest resistance and is indeed known for its poor biocompatibility (Zitter and Plenk 1987). Importantly, this shows that tantalum is potentially the most biocompatible metal tested because it is the most inert from a redox point of view. Figure adapted from Zitter and Plenk (1987).

TRANSFAC, HPRD and power graph analysis. We followed the same methodology as in the previous sections (page 155). We used the TRANSFAC-derived network of transcription factors to target genes, and analyzed its power graph with the intent of finding key transcription factors upstream of differentially expressed genes.

Pathway analysis. Functional and pathway enrichment analysis was done using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al. 2003).

6.4.2 Results

hMSCs gene expression levels reach a steady state sooner on tantalum than on titanium. As shown in Fig.105 the top 1,000 genes with highest temporal variation in expression levels reach a steady state sooner on tantalum than on titanium. Upon culture in different conditions, the hMSCs react by activating or deactivating different pathways. The chemical stress induced by the oxide layers affects cells cultured both on titanium and on tantalum – explaining rapid changes in expression levels during the first 4 days. hMSCs cells undergo a transition from a proliferating state to a differentiating state. After 8 days the cells have adapted to the new conditions. Our hypothesis is that the chemical stress induced by the oxide layer on hMSCs is stronger on tantalum than on titanium, thus explaining the delay of hMSCs in reaching the steady-state.

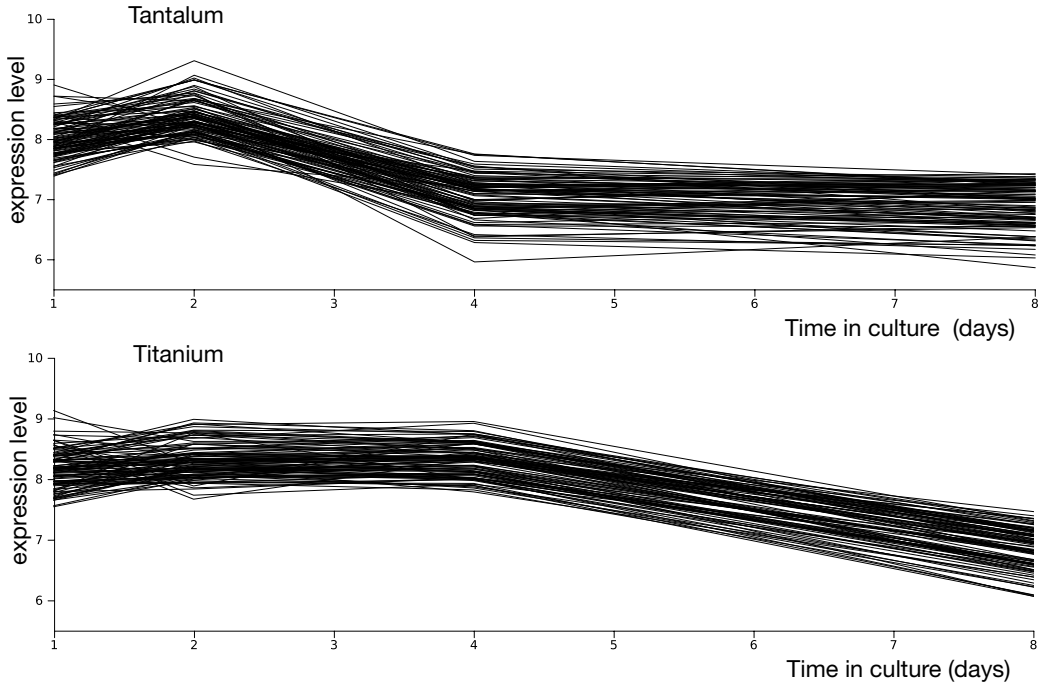


Fig. 105 **Gene expression levels of hMSCs reach a steady state sooner on tantalum than on titanium.** The gene expression levels were measured for day 1, day 2, day 4 and day 8 – other values are linearly interpolated. The gene expression levels exhibit a transient response that stabilizes later for titanium than for tantalum. Here we show the expression profiles for 97 genes that sustain high expression levels until day 4 on titanium but only until day 2 on tantalum. On day 8 the gene expression levels have reached a steady state for both metals.

Functional profile of the top 1,000 genes. Table 20 shows that most of the genes involved in the differential response of hMSCs to culture on titanium versus tantalum belong to signaling pathways, as indicated by a high enrichment in post-translational

modifications: phosphoproteins (50%), acetylation (21%), and transcription factors (coiled coil 17%l, zinc-finger 13%). Also, many genes are involved in cell cycle regulation and progression: cell division and mitosis (10%) and response to DNA damage (3%).

Table 20 **Functional enrichment of the top 1,000 genes with highest temporal variation.** For each term, the number and percentage of genes annotated with the term is given, the *p*-value (Fischer exact test) as well as the Bonferroni corrected *p*-value.

Term	Count	Percentage	*p*-value	Fold Enrichment	Bonferroni
phosphoprotein	532	53	$< 10^{-21}$	1.36	$< 10^{-18}$
coiled coil	178	17	$< 10^{-11}$	1.67	$< 10^{-9}$
acetylation	218	21	$< 10^{-11}$	1.54	$< 10^{-8}$
cell cycle	60	6	$< 10^{-9}$	2.43	$< 10^{-6}$
cell division	38	3	$< 10^{-7}$	2.68	$< 10^{-4}$
DNA damage	31	3	$< 10^{-6}$	2.83	$< 10^{-3}$
mitosis	29	3	$< 10^{-6}$	2.93	$< 10^{-3}$
DNA repair	27	3	$< 10^{-4}$	2.64	$< 10^{-2}$
zinc-finger	128	13	$< 10^{-4}$	1.43	$< 10^{-2}$
alternative splicing	456	46	$< 10^{-4}$	1.14	$< 10^{-2}$

Identification of master regulators. Power graph analysis of the transcription factor to target gene network (TransFac) revealed that the transcription factors NRF2, EGR1, IRF-1, and particularly NF-Y and P53 may play major roles in the differential gene expression response (Fig.106). It is known that P53 interacts with the heterotrimeric transcription factor NF-Y to form a P53/NF-Y, a complex that modulates the expression of key cell cycle genes in response to DNA damage (Peart and Prives 2006; Benatti et al. 2008). NF-Y binds to a CCAAT-box, thus regulating the expression of glutathione peroxidase 4 (GPX4) – one of the most potent anti-oxidant compounds produced by the cell (Huang et al. 1999). Glutathione peroxidases (GSH) catalyze the reduction of hydrogen peroxide and other oxidative compounds that induce cell damage. One particularly relevant target of P53/NF-Y is FAS. FAS is up-regulated on the first days of culture and its expression level decreases and stabilizes in the last two days – sooner on tantalum than on titanium. FAS is an apoptosis-inducing protein whose activation has been associated with reactive oxygen species (ROS) (Wang et al. 2008). Note that two other transcription factors found – IRF-1 and NRF2 – are also implicated in oxidative stress response, ROS level sensing and GSH expression control (Chan et al. 2001; Hickling et al. 2010).

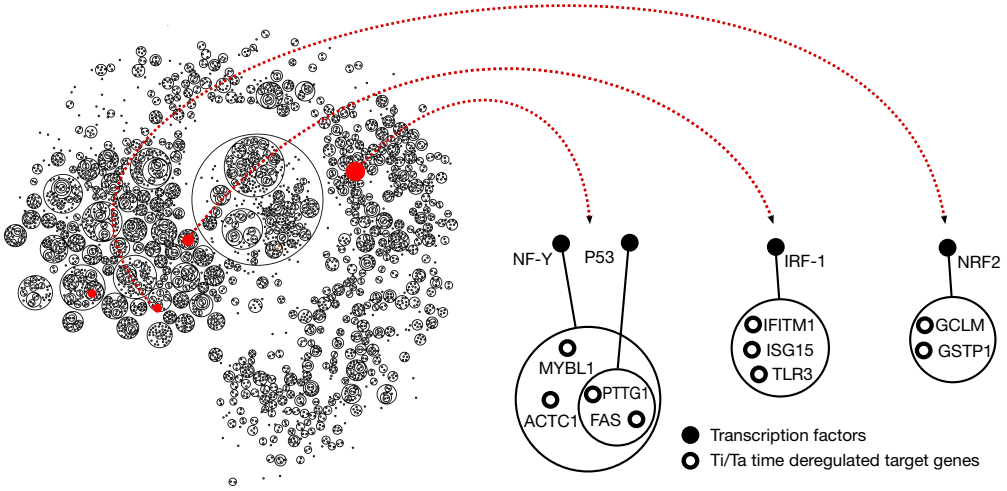


Fig. 106 **Master regulators involved in the different response hMSCs on tantalum versus titanium.**

Importance of Reactive Oxygen Species (ROS) for biocompatibility. Reactive oxygen species (ROS) are a broad class of small reactive molecules – such as oxygen ions and peroxides – that contain the oxygen atom . ROS are oxidative agents that are destructive to both DNA and proteins and are known to induce apoptosis (Cardaci et al. 2008). The previous results suggest that the release of ROS by titanium and tantalum oxide layers may trigger apoptosis related pathways dominated by P53, NF-Y, IRF-1 and NRF2. It is already known that titanium nanoparticles induce oxidative stress and apoptosis in cultured cells (Park et al. 2008). It has also been shown by Achanta and Huang (2004) that P53 has a direct role in sensing oxidative DNA damage. There is also evidence of a direct link between titanium and P53 DNA damage sensing (van Kooten et al. 2000). Note that ROS delay the cell cycle through DNA damage but also inhibit the differentiation of osteoblastic cells as shown by Mody et al. (2001).

In order to check this hypothesis we searched for KEGG pathways that explain the delayed stabilization of gene expression levels on titanium compared with tantalum. Table 21 shows the 5 most significant pathways associated with the 1,000 genes of highest temporal variance. The cell cycle pathway is the most significantly enriched. This could explain the delayed transition between proliferating hMSCs and differentiated cells. Also, we find a significant enrichment of the P53 signaling pathway, confirming P53 as master regulator.

Moreover, we found a strong signal in the selenoamino acid metabolism pathway (Fig. 107). We find that almost all enzymes (4 out of 5), in a directed enzymatic chain starting from selenomethionines and ending with selenocysteines, exhibit delayed gene expression stabilization on titanium. Selenoproteins such as glutathione peroxidase are potent antioxidant proteins that require a selenocysteine selenoamino-acid (Arnér 2010). This finding brings additional confidence to our claim that oxidative stress mediated by ROS released from the oxide layer is a key mechanism behind the differences in gene-expression levels between titanium and tantalum.

Table 21 **Pathway enrichment for the top 1,000 genes with highest temporal variation.** For each term, the number and percentage of genes annotated with the term is given, the *p*-value (Fischer exact test). We use the DAVID service and its EASE score for deciding if an enrichment is significant (Dennis et al. 2003). Note the high *p*-value of the *selenoamino acid metabolism* pathway. The *p*-value under-estimates the statistical significance of this finding because it is computed for the whole pathway and not for the specific contiguous metabolic path (as shown in red in Fig.107).

Pathway	Number of genes	*p*-value
Cell cycle	14	10^{-3}
RNA degradation	9	10^{-3}
mTOR signaling pathway	7	10^{-2}
P53 signaling pathway	8	2.63^{-2}
Ubiquitin mediated proteolysis	11	6.05^{-2}
Selenoamino acid metabolism	4	9.97^{-2}

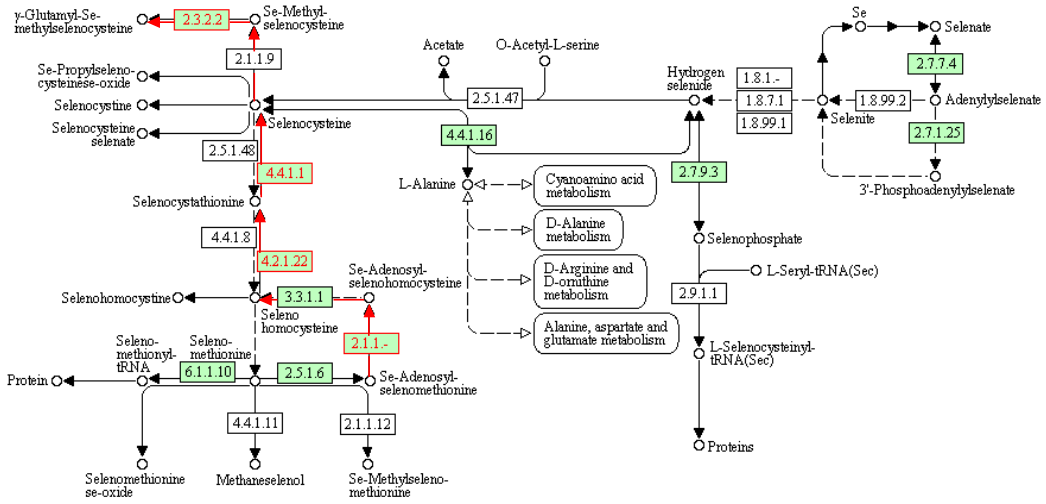


Fig. 107 **Selenoamino acid metabolism.** Selenoamino acids are amino acids in which selenium has been substituted for sulfur (Arnér 2010). Selenium is often found as selenomethionine because it is indistinguishable from methionine – and is thus a natural source of selenium. A more functional form in which selenium is present in cells is selenocysteine. In contrast to selenomethionine, no free reserves of selenocysteine exists because of its high reactivity. Selenocysteine is a key amino acid that is present in several reductase enzymes that have an anti-oxidant function such as glutathione peroxidases, thioredoxin reductases, and glycine reductases (Kryukov et al. 2003). Strikingly, we observe that the enzymatic chain converting selenomethionine to selenocysteine is reaching steady state sooner in tantalum compared with titanium – suggesting that the cells needed antioxidative selenoproteins longer when cultivated on titanium. Enzymes enclosed in red boxes correspond to genes that are reaching steady state sooner in Ta than in Ti. Only enzymes colored in green can be mapped to Human genes.

6.4.3 Discussion

The main difference in the gene expression profiles of hMSCs on tantalum versus titanium is that hMSCs cultivated on tantalum reach a steady state sooner than on

titanium. We analyzed the genes behind this difference with power graph analysis applied to the TRANSFAC network. We found four candidate master regulators P53, NF-Y, IRF-1 and NRF2 that share a functional role in reactive oxygen species (ROS) sensing, DNA damage sensing and cell-cycle control (Fig. 108). Because it is known that the titanium oxide layer is less stable and more prone to releasing ROS than the tantalum oxide layer, we propose that ROS are the external stimuli behind the differential response. This initial hypothesis is corroborated by the detection of a strong signal in the enzymatic chain converting selenomethionine to selenocysteine – an important tool in the anti-oxidative arsenal of cells.

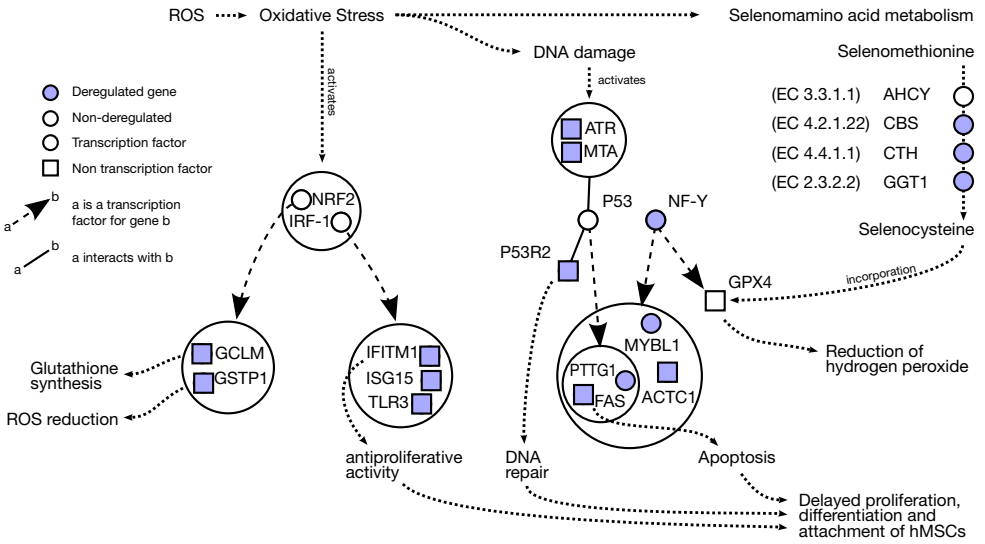


Fig. 108 **Summary.** Higher ROS levels on titanium lead to a longer transitory response to oxidative stress and a more sustained activation of the anti-oxidative machinery of hMSCs. This in turns delays proliferation, differentiation and attachment of the cells on titanium surfaces compared with tantalum surfaces.

Another possibility which could also explain the differences in biocompatibility is the different adsorption characteristics of the titanium and tantalum metals. Nagayasu et al. (2005) showed that titanium and tantalum oxide particles exhibit different adsorption characteristics, with tantalum being generally less adsorptive than titanium. Corroborating this hypothesis is the result by Sousa et al. (2008) who showed that osteoblast adhesion and morphology on titanium oxide depends on the competitive pre-adsorption of albumin and fibronectin. Differences in absorptivity could vary the thickness of a protein-based layer thus modulating the release of ROS from the underlying oxide layer.

6.5 Conclusion

"*If people do not believe that mathematics is simple,*
it is only because they do not realize how complicated life is."
John von Neumann

In this chapter, power graphs enabled the exploration of complex networks, the identification of regulatory modules, and the visualization of pathways. First, we showed that neuroectodermal conversion of mesenchymal stem cells is controlled by two master regulators: HIF-1alpha and miR-124a. Remarkably, the role of HIF-1alpha was confirmed by immunoblotting. Second, we discovered a striking connection between two rare diseases: MELAS and Sjögren's Syndrome, and identified two master transcription factors: IRF-8 and NF-Y. Third, we identified oxidative stress as a likely cause for the enhanced biocompatibility of tantalum compared with titanium.

Chapter 7

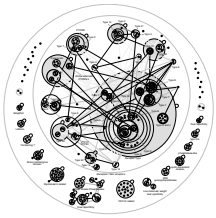
Conclusion and outlook

In the following we revisit the three open problems addressed in this thesis by summarizing its key contributions. We also discuss the limits and possible improvements.

7.1 Contributions of this Thesis

7.1.1 Revisiting open problem 1

How to find biologically relevant modules in protein interaction networks? In particular, how to convey without loss of information the subtle connection patterns within and between modules of proteins?



In chapter 3 we presented power graph analysis, a novel network analysis approach that combines in a unique way ideas from visualization, data clustering, network motif analysis, and information compression. We showed on several examples how power graph analysis reveals the biology underpinning protein interaction, regulatory and homology networks. In the following we review from different perspectives the unique advantages of power graph analysis, and in particular how it can convey without loss of information the subtle connection patterns within and between network modules.

Contributions to graph clustering. In the one hand, traditional graph clustering algorithms search for densely connected regions in the network and abstract these as clusters. On the other hand, network visualization techniques display all the graph's connectivity information but lack *structure*. Power graph analysis can do both. We showed that it is possible to obtain a graph-like representation that preserves all the connectivity information, while providing structure as clusters of nodes (power nodes) and cluster of edges (power edges). Moreover, table 2 shows that only four other graph clustering algorithms – Modular decomposition, bi-clustering, MULIC, and Link Clustering – can detect clusters of nodes defined by common neighborhoods. And only two algorithm have the ability to detect and represent edge clusters – Bi-clustering and Link Clustering. From a computational point of view the power

graph algorithm has a competitive sub-quadratic complexity of $O\left(e^{1.71}\right)$ in the number e of edges. This is better than Link Clustering or MULIC which have a complexity of $O\left(n^2\right)$ with n being the number of nodes, or worse $O\left(e^3\right)$ when considering the number of edges e (Andreopoulos et al. 2006; Ahn et al. 2010). The lower time complexity of the the power graph analysis algorithm is attributable to its optimized sparse data structures.

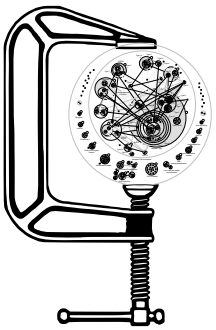
Contributions to network visualization. Aside from the algorithmic and graph clustering contributions, power graph analysis is also unique in that it is the first graph clustering algorithm that allows a graph-like visual representation of both the node and edge clusters. While bi-clustered adjacency matrix representations of graphs can also detect and represent nodes and edges clusters, power graphs are unique in representing these in a graph-like manner. From a visualization point of view, power graph analysis can drastically reduce the 'fur-ball' effect that plagues network visualization. This is possible because of the lossless transformation of node and edge clusters into symbolic representations: power nodes and power edges. A good example is on page 74 where we showed that the power graph representation can reduce by 95% the visual clutter – representing 4849 edges with only 209 power edges and 95 power nodes.

Availability. The power graph algorithm is widely accessible to the community. The algorithm and a specially tailored visualization engine has been implemented by Matthias Reimann as a plugin for Cytoscape (Shannon et al. 2003) .

Key contributions summarized:

- Graph clustering algorithm capable of finding both dense clusters and neighborhood similar clusters.
- Clusters of nodes but also clusters of edges are identified.
- The connectivity information between these clusters is preserved.
- Cliques and bicliques – which are abundant and relevant for biological networks – are explicitly represented.
- Power graphs are lossless and intuitive graph-like representations that can reduce visual clutter by up to 95%.
- Fast algorithm in $O\left(e^{1.71}\right)$ for networks with e edges, marginally dependent on the number of nodes.

7.1.2 Revisiting open problem 2



How to computationally evaluate the quality of protein-protein interaction networks?

In chapter 4 we presented network compressibility as a novel measure to evaluate the quality of protein interaction networks. We conducted a four point validation by *i*) testing the effect of false positives and negatives, *ii*) verifying that gold standard networks are highly

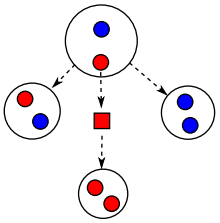
compressible, *iii*) checking that confidence thresholds are consistent with high network compressibility, and *iv*) verifying the correlation with co-expression, co-localization and shared function. Moreover, we showed that best-quality protein interaction networks exhibit compressibility levels similar to those of accurately known complex systems. Additionally, higher compressibility is observed when using superior Y2H and AP/MS protocols. Overall, these results establish that network compressibility is correlated to network quality. It is to date the most extensive study of the network quality and compressibility covering 21 genome-wide interactomes (13 Y2H, 7 AP/MS, 1 PCA), as well as literature, structure-derived, and consolidated datasets. Importantly, it is also the first quality measure for protein interaction networks that is defined solely on the network information – without the need of any additional biological data.

Key contributions:

- Extensive four point validation of network compressibility as quality measure for protein interaction networks.
- First quality measure solely based on the network's information.
- Most extensive survey – covering 21 genome-wide interactomes – of genome-wide protein interaction network quality and compressibility.

7.1.3 Revisiting open problem 3

How to identify key master regulators and pathways with novel representations of regulatory and protein interaction networks? In particular, can we find key master regulators and pathways behind i) the neuroectodermal conversion of mesenchymal stem cells, ii) a rare mitochondrial cytopathy (MELAS), and iii) the enhanced biocompatibility of tantalum compared with titanium?



In chapter 6 we demonstrated the use of power graph analysis for the analysis of three gene expression datasets produced by our collaborators in Alexander Storch's research group, on neurodegenerative diseases, and Maik Stiehler's research group, at the clinic for orthopedics (Dresden University of Technology). In all three cases, master regulators and pathways were identified and their relevance was confirmed by our collaborators. Remarkably, our prediction of the role of HIF-1alpha in the neuroectodermal conversion of mesenchymal stem cell was confirmed experimentally by immunoblotting.

Key contributions:

- The neuroectodermal conversion of mesenchymal stem cells is controlled by two master regulators: HIF-1alpha and miR-124a. The role of HIF-1alpha was confirmed experimentally by immunoblotting.
- Discovery of a hidden link between two rare diseases: MELAS and Sjögren's Syndrome.
- Identification of two master transcription factors in MELAS: IRF-8 and NF-Y.

- Oxidative stress is the likely cause for the enhanced biocompatibility of tantalum compared with titanium.

7.2 Limitations and possible improvements

In the following we review the limitations of our approach and possible improvements for future work.

7.2.1 Power graph analysis

Overlapping power nodes. In the definition of power graphs we require that power nodes form a hierarchy. This requirement facilitates layout of power graphs and is motivated by practical considerations. In general, arbitrarily overlapping power nodes cannot be drawn in the plane using discs since Venn diagrams are limited to 3 sets (Venn 1880). It is a sensible requirement from a visualization point of view, but it could be relaxed for graph clustering by allowing node and edge cluster overlap. This is reminiscent of the approach proposed by Ahn et al. (2010) in which link communities give rise to overlapping node clusters.

Networks with different types of edges. The power graph algorithm considers all edges in the graph as belonging to the same class. However, some applications may require several types of edges. For example, one may want to compute a power graph on a mixed network of protein interactions and gene regulation. In that case, a desirable outcome would be to guaranty unmixed power edges – edge clusters should respect the separation between edge classes. Otherwise, the meaning of a power edge in terms of individual edges is ambiguous and the transformation would not be lossless.

Edge weights, lossless transformation, and edge clustering. The current implementation of the power graph algorithm can use edge weights. This is implemented by adapting the Jaccard neighborhood similarity for edge weights normalized between 0 and 1. Does it always makes sense to group together power edges of low and high confidence? An alternative would be to cluster edges of *similar weights* together. This hints at a more general algorithm recast purely in terms of edge similarity. In that setting a similarity measure *between edges* is defined by connectivity but also by edge attribute and label similarity. While outside of the scope of this thesis, we should note that it is possible to define a similarity measure between edges that favors clique and biclique clusters.

Incomplete cliques and bicliques. In power graph analysis the requirement that power nodes and power edges represent *complete* cliques and bicliques might be seen as a limit to its applicability to noisy datasets. Instead, one could consider the notions of p-clique and p-biclique. A p-clique is a set of nodes V such that each node $v \in V$ is adjacent to at least $p|V|$ other nodes in V, where p is a proportion between

0 and 1. Similarly, a p-biclique is two node sets U and V such that all nodes $u \in U$ are adjacent to at least $p|V|$ nodes in V and all odes $v \in V$ are adjacent to at least $p|U|$ nodes in U. When $p = 1$ p-cliques and p-bicliques are complete cliques or bicliques. As p decreases towards 0 p-cliques and p-bicliques lose more and more edges but the distribution of edges remains balanced between nodes. Only the second step of the algorithm (*power edge search*) would need to be modified to detect p-cliques and p-bicliques. Currently, the greedy search prioritizes candidate power edges by their size which is the number of underlying edges. In the case of incomplete power edges a possible measure could be the product of p with the size of the candidate power edge. However a better approach would be to use a Pareto ranking approach to avoid extreme cases in which p is too low or in which the candidate power edge is too small.

It remains that cliques and bicliques are relatively robust to the removal or addition of few single edges. As shown in Fig. 109, removing an edge (u, v) from a clique or biclique has the effect of just shrinking it by one or two nodes respectively.

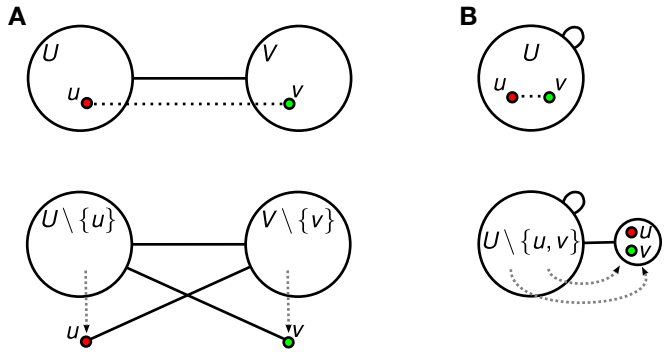


Fig. 109 Robustness of power edges to edge removal. Removing a single edge (u, v) from a clique or biclique underlying a power edge has the consequence of removing just u and v from it. **(A)** Removing an edge (u, v) from the biclique underlying a power edge (U, V) leads to its graceful degradation into a smaller power edge $\left(U \setminus \{u\}, V \setminus \{v\}\right)$. **(B)** Same holds for a reflexive power edge. Removing an edge (u, v) from the clique underlying a reflexive power edge (U, U) has the only consequence of just removing both nodes u and v.

Putting the emphasis on statistically significant network motifs. In chapter 3 and 4 we showed that protein interaction networks and complex networks in general are rich in cliques and bicliques (see Fig. 41, Fig. 78, Fig. 42, and Fig. 77). Moreover, large bicliques and cliques are less likely to occur in networks by chance alone and are thus more *significant* than smaller ones. In contrast, stars happen by chance in random networks. Currently, the power graph algorithm ignores these statistical aspects and concentrates on the size of the motifs (number of underlying edges). An alternative would be to give priority to highly significant network motifs in the greedy search by computing a p-value for each candidate power edge. This can be done using a generalized hypergeometric test to quantify the likelihood of overlap between neighborhood sets. This approach would have the advantage of guaranteeing the detection of significant cliques and bicliques. However, this would imply a much higher time complexity for the algorithm.

7.2.2 Network compressibility and quality

Noise in networks. We evaluate our quality measure using two noise models: ER and BA. The first model considers a uniform distribution for picking nodes for the added/removed interactions (ER), whereas the second considers a distribution inversely proportional to the degree distribution (BA). The strength of these two models is their simplicity. Yet, one could consider more sophisticated models closer to the experimental details and protocols of Y2H, AP/MS, and PCA screens. For example, in the case of AP/MS screens, one could devise a noise model taking into account the observed interaction propensity of proteins – possibly by estimating binding kinetics and simulating pull-down. In the case of Y2H library screens, it could be interesting to estimate how the uneven distributions of strains in the prey library could be used to model false negatives. However, it remains that stronger arguments and additional experimental evidence are needed to justify more complicated noise models. In a wider context, a framework for simulating experimental noise, sampling, and other experimental effects similar to that proposed by Venkatesan et al. (2009) would help clarify the relationship between experimental details and network relative compressibility.

Novel motifs for higher network compressibility. In chapter 4 we measure network compressibility with the power graph algorithm. However, other graph compression schemes exist. Therefore the question is whether our quality measure can be improved by a better compression algorithm. Compression algorithms can attain the *Shanon limit* to data compression only if they are able to completely capture the data's statistics. In our case we exploit the over-representation of cliques and bicliques in complex networks and in particular protein interaction networks. First, let us note that any graph compression algorithm solely based on the identification of dense clusters would be inferior to power graph analysis because of their inability to detect neighborhood similar clusters and in general edge clusters (SPC, Kcliques, MCL, SCC, ProClust, MCODE, RNSC, Spectral, and BAG). Similarly to power graph analysis, most graph compression algorithms rely on the neighborhood similarity between nodes in the network (Lu 2002; Feder and Motwani 1995; Kao et al. 1998; Deo and Litow 1998; Randall et al. 2002; Boldi and Vigna 2004; Langville and Meyer 2004; Hannah et al. 2008), or exploit symmetries and graph isomorphisms within the networks (Rashewsky 1955; Mowshowitz 1968; Dehmer and Emmert-Streib 2008). By examining these different approaches it appears that graph clustering algorithms rely on three fundamental ingredients:

- A choice of network motifs,
- an algorithm capable of identifying these motifs,
- a representation language in which these motifs can be combined to specify a unique graph.

What other motifs – aside from cliques and bicliques – could be relevant for complex networks? Alon (2007)'s generalized network's motifs are not good candidates since they are easily reduced to few stars and bicliques. A good hint comes from the networks that power graph analysis fails to compress. For example, consider a *chain* in which the nodes a, b, c, d, e are adjacent: $(a, b), (b, c), (c, d), (d, e)$ and such that

only a and e are adjacent to other nodes – hence b, c, d have only two neighbors each. This network motif could be replaced by a single *chain-edge* (a, b) to which the information about b, c, d and their order in the chain would be attached. Similarly, node cycles which are closed chains of nodes could also be abstracted and hence compressed. There is one kind of biological network for which these two motifs could greatly improve the measure of compressibility: metabolic networks, and in general enzyme-substrate networks.

Adjacency matrix compression algorithm. Another possible improvement of our compression algorithm would be to compress the binary image corresponding to the bi-clustered adjacency matrix. Consider the adjacency matrix as an image in which white pixels correspond to edges and black pixels corresponds to the absence of edges. The first step of the algorithm would simply permute the lines and columns in order to maximize the size of uniformly colored rectangles and squares in the image. Note that this step is the node clustering step of the current algorithm. The second step would compress this binary image by – for example – finding a minimum number of rectangles covering all and only white rectangles. In order to attain higher compression levels we would search for a minimal covering using white and black rectangles – or rectangles acting as XOR operators. This would further reduce the minimal number of rectangles needed to represent the graph. Even higher compression could be attained if regions of the image could be copied and pasted into other similar regions. If two regions in the image are similar enough, their common information plus the differences would weight less information than if separately encoded.

7.3 Outlook

Advances in molecular biology are moving the field away from purely descriptive models and closer to reverse engineering the cell. Recently, Gibson et al. (2010) demonstrated the creation of a bacterial cell controlled by a chemically synthesized genome. This milestone, together with genome sequencing entering an industrial phase, and efforts to decode the epigenome under way (Barski et al. 2007; Ji et al. 2010) show that the next frontier is likely the interactome. Can the same level of resolution be attained? For this, many challenges need to be tackled. First, increasing quality and coverage is needed through the development of superior experiments (Tarassov et al. 2008; Xin et al. 2009; Breitkreutz et al. 2010). Second, the analysis of the interactome will need to be extended to cover the time dimension, protein post-translational states, as well as DNA-protein and lipid-protein interactions (Aparicio et al. 2005; Dioum et al. 2009). At that point, richer systems biology models capturing enzymatic reactions, binding and transport kinetics will replace the network abstractions – revealing the true nature of the interactome.

References

Achanta G and Huang P (2004). Role of p53 in sensing oxidative DNA damage in response to reactive oxygen species-generating agents. *Cancer Res* **64**: 6233–6239

Aebersold R and Mann M (2003). Mass spectrometry-based proteomics. *Nature* **422**: 198–207

Agamanolis DP (2009). *Neuropathology*. Ed. by AAMC MedEdPortal. Northeastern Ohio Universities College of Medecine

Ahn YY, Bagrow JP and Lehmann S (2010). Link communities reveal multiscale complexity in networks. *Nature* **466**: 761–764

Albert, Jeong and Barabasi (2000). Error and attack tolerance of complex networks. *Nature* **406**: 378–382

Albert I and Albert R (2004). Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics* **20**: 3346–3352

Alexe G, Alexe S, Crama Y, Foldes S, Hammer PL and Simeone B (2004). Consensus algorithms for the generation of all maximal bicliques. *Discrete Applied Mathematics* **145**: 11 –21

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

Andersen J, Vanscoy S, Cheng TF, Gomez D and Reich NC (2007). IRF-3-dependent and augmented target genes during viral infection. *Genes Immun* **9**: 168–175

Ando R (2007). 'BioCreative II gene mention tagging system at IBM Watson'. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Citeseer, pp. 101–103

Andreopoulos B, An A and Wang X (2006). Bi-Level Clustering of Mixed Categorical and Numerical Biomedical Data. *International Journal of Data Mining and Bioinformatics* **1**: 19–56

Andreopoulos B, An A, Wang X, Faloutsos M and Schroeder M (2007). Clustering by Common Friends Finds Locally Significant Proteins Mediating Modules. *Bioinformatics* **23**: 1124

Andreopoulos B, An A, Wang X and Schroeder M (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform* **10**: 297–314

Aparicio O, Geisberg JV, Sekinger E, Yang A, Moqtaderi Z and Struhl K (2005). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Mol Biol* **Chapter 21**: Unit 21.3

Arabie P, Boorman S and Levitt P (1978). Constructing blockmodels: how and why. *Journal of mathematical psychology* **17**: 21–63

Aragues R, Jaeggi D and Oliva B (2006). PIANA: protein interactions and network analysis. *Bioinformatics* **22**: 1015–1017

Aranda B et al. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**: D525–D531

Arifuzzaman M et al. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res* **16**: 686–691

Arnér ESJ (2010). Selenoproteins-What unique properties can arise with selenocysteine in place of cysteine? *Exp Cell Res* **316**: 1296–1303

Arora A, Guduric-Fuchs J, Harwood L, Dellett M, Cogliati T and Simpson DA (2010). Prediction of microRNAs affecting mRNA expression during retinal development. *BMC Dev Biol* **10**: 1

Bader GD and Hogue CWV (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**: 2

Bader GD, Betel D and Hogue CWV (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**: 248–250

Bairoch A (1994). The ENZYME data bank. *Nucleic Acids Res* **22**: 3626–3627

Balla VK, Bodhak S, Bose S and Bandyopadhyay A (2010). Porous tantalum structures for bone implants: Fabrication, mechanical and in vitro biological properties. *Acta Biomater* **12**: 120

Barabasi and Albert (1999). Emergence of scaling in random networks. *Science* **286**: 509–512

Barabási AL and Oltvai ZN (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**: 101–113

Bard AJ, Parsons R and Jordan J (1985). *Standard Potentials in Aqueous Solutions*. Ed. by Marcel Dekker. New York: Marcel Dekker, New York

Barkow S, Bleuler S, Prelic A, Zimmermann P and Zitzler E (2006). BicAT: a biclustering analysis toolbox. *Bioinformatics* **22**: 1282–1283

Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C and Apweiler R (2009). The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* **37**: D396–D403

Barrett T et al. (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**: D885–D890

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I and Zhao K (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837

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

Basalto N, Bellotti R, Carlo FD, Facchi P, Pantaleo E and Pascazio S (2008). Hausdorff clustering. *Phys Rev E Stat Nonlin Soft Matter Phys* **78**: 046112

Batagelj V and Mrvar A (1998). 'Pajek - Program for Large Network Analysis'. In: vol. 21. International Network for Social Network Analysis, pp. 47–57

Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G and Hunter L (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* **23**: i41–i48

Becker KG, Hosack DA, Dennis G, Lempicki RA, Bright TJ, Cheadle C and Engel J (2003). PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* **4**: 61

Bein D, Morales L, Bein W, C.O. Shields J, Meng Z and Sudborough I (2008). Clustering and the Biclique Partition Problem. *Hawaii International Conference on System Sciences* **12**: 475

Benatti P, Basile V, Merico D, Fantoni LI, Tagliafico E and Imbriano C (2008). A balance between NF-Y and p53 governs the pro- and anti-apoptotic transcriptional response. *Nucleic Acids Res* **36**: 1415–1428

Berge C (1976). *Graphs and hypergraphs*. North-Holland Pub. Co.

Beyer A, Workman C, Hollunder J, Radke D, Moller U, Wilhelm T and Ideker T (2006).

Integrated assessment and prediction of transcription factor binding. *PLoS Comput Biol* **2**: e70

Binnig, Quate and Gerber (1986). Atomic force microscope. *Phys Rev Lett* **56**: 930–933

Bioch J (2005). The complexity of modular decomposition of boolean functions. *Discrete Applied Mathematics* **149**: 1–13

Blake JA and Harris MA (2008). The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics* **Chapter 7**: Unit 7.2

Blaschke C, Hirschman L, Yeh A and Valencia A (2003). Critical assessment of information extraction systems in biology. *Comp Funct Genomics* **4**: 674–677

Blatt, Wiseman and Domany (1996). Superparamagnetic clustering of data. *Phys Rev Lett* **76**: 3251–3254

Boldi P and Vigna S (2004). The webgraph framework I: compression techniques : 595–602

Bolstad AI, Wargelius A, Nakken B, Haga HJ and Jonsson R (2000). Fas and Fas ligand gene polymorphisms in primary Sjögren's syndrome. *J Rheumatol* **27**: 2397–2405

Bolten E, Schliep A, Schneckener S, Schomburg D and Schrader R (2001). Clustering protein sequences–structure prediction by transitive homology. *Bioinformatics* **17**: 935–941

Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA and Weiner J (2005). The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci* **62**: 435–445

Bosman D, Blom E, Ogao P, Kuipers O and Roerdink J (2007). MOVE: A Multi-level Ontology-based Visualization and Exploration framework for genomic networks. *In Silico Biol* **7**: 35–59

Breiger R, Boorman S and Arabie P (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology* **12**: 328–383

Breitkreutz A et al. (2010). A global protein kinase and phosphatase interaction network in yeast. *Science* **328**: 1043–1046

Breitkreutz BJ, Stark C and Tyers M (2003). Osprey: a network visualization system. *Genome Biol* **4**: R22

Breitkreutz BJ et al. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**: D637–D640

Brohée S and Van Helden J (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**: 488

Bron C and Kerbosch J (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* **16**: 575–577

Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S and Kahn D (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* **33**: D212–D215

Bruder SP, Jaiswal N, Ricalton NS, Mosca JD, Kraus KH and Kadiyala S (1998). Mesenchymal stem cells in osteobiology and applied bone regeneration. *Clin Orthop Relat Res* : S247–S256

Bu D et al. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res* **31**: 2443–2450

Butland G et al. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**: 531–537

Cardaci S, Filomeni G, Rotilio G and Ciriolo MR (2008). Reactive oxygen species mediate p53 activation and apoptosis induced by sodium nitroprusside in SH-SY5Y cells. *Mol Pharmacol* **74**: 1234–1245

Cavalli-Sforza L and Edwards A (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* **19**: 233

Ceol A, Aryamontri AC, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L and Cesareni G (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**: D532–D539

Cha S, Brayer J, Gao J, Brown V, Killedar S, Yasunari U and Peck AB (2004). A dual role for interferon-gamma in the pathogenesis of Sjogren's syndrome-like autoimmune exocrinopathy in the nonobese diabetic mouse. *Scand J Immunol* **60**: 552–565

Chaitin GJ (2007). *Meta Math! The Quest for Omega*. Pantheon Books

Chan K, Han XD and Kan YW (2001). An important function of Nrf2 in combating oxidative stress: detoxification of acetaminophen. *Proc Natl Acad Sci U S A* **98**: 4611–4616

Chatr-Aryamontri A, Zanzoni A, Ceol A and Cesareni G (2008). Searching the protein interaction space through the MINT database. *Methods Mol Biol* **484**: 305–317

Chen C (2005). Top 10 unsolved information visualization problems. *IEEE Comput Graph Appl* **25**: 12–16

Chen H and Sharp BM (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5**: 147

Cheng KO, Law NF, Siu WC and Lau TH (2007). BiVisu: software tool for bicluster detection and visualization. *Bioinformatics* **23**: 2342–2344

Cheng LC, Pastrana E, Tavazoie M and Doetsch F (2009). miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche. *Nat Neurosci* **12**: 399–408

Cheng Y and Church GM (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* **8**: 93–103

Cherry JM et al. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**: 73–79

Chi NC, Adam EJ, Visser GD and Adam SA (1996). RanBP1 stabilizes the interaction of Ran with p97 nuclear protein import. *J Cell Biol* **135**: 559–569

Chien CT, Bartel PL, Sternglanz R and Fields S (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A* **88**: 9578–9582

Chinnery PF, Howell N, Lightowlers RN and Turnbull DM (1997). Molecular pathology of MELAS and MERRF. The relationship between mutation load and clinical phenotypes. *Brain* **120**: 1713–1721

Clark AM, Goldstein LD, Tevlin M, Tavaré S, Shaham S and Miska EA (2010). The microRNA miR-124 controls gene expression in the sensory nervous system of Caenorhabditis elegans. *Nucleic Acids Res* **1**: 14

Cohen MD, Kargacin B, Klein CB and Costa M (1993). Mechanisms of chromium carcinogenicity and toxicity. *Crit Rev Toxicol* **23**: 255–281

Colizza V, Barrat A, Barthélemy M and Vespignani A (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci U S A* **103**: 2015–2020

Collins SR et al. (2007a). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**: 806–10

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS and Krogan NJ (2007b). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**: 439–450

Conaco C, Otto S, Han JJ and Mandel G (2006). Reciprocal actions of REST and a microRNA promote neuronal identity. *Proc Natl Acad Sci U S A* **103**: 2422–2427

Consortium U (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37**: D169–D174

Cormen TH, Leiserson CE and Rivest RL (1990). *Introduction to Algorithms*. Cambridge, MA: MIT Press

Corpet F (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* **16**: 10881

Cournier A and Habib M (1994). A new linear algorithm for modular decomposition. *Lecture Notes in Computer Science* **787**: 68–84

Criekinge WV and Beyaert R (1999). Yeast Two-Hybrid: State of the Art. *Biol Proced Online* **2**: 1–38

Cui G, Chen Y, Huang DS and Han K (2008). An algorithm for finding functional modules and protein complexes in protein-protein interaction networks. *J Biomed Biotechnol* **2008**: 860270

Cusick ME et al. (2009). Literature-curated protein interaction datasets. *Nat Methods* **6**: 39–46

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

Deeds EJ, Ashenberg O and Shakhnovich EI (2006). A simple physical model for scaling in protein-protein interaction networks. *Proc Natl Acad Sci U S A* **103**: 311–316

Dehmer M and Emmert-Streib F (2008). Structural information content of networks: graph entropy based on local vertex functionals. *Comput Biol Chem* **32**: 131–138

Delaleu N, Immervoll H, Cornelius J and Jonsson R (2008). Biomarker profiles in serum and saliva of experimental Sjögren's syndrome: associations with specific autoimmune manifestations. *Arthritis Res Ther* **10**: R22

Demetrius L and Manke T (2005). Robustness and network evolution–an entropic principle. *Physica A: Statistical Mechanics and its Applications* **346**: 682 –696

Deng M, Mehta S, Sun F and Chen T (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res* **12**: 1540–1548

Deng M, Sun F and Chen T (2003). Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput* : 140–151

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**: P3

Deo N and Litow B (1998). 'A structural approach to graph compression'. In: *Proc. of the MFCS Workshop on Communications*, pp. 91–101

Deshaies RJ, Seol JH, McDonald WH, Cope G, Lyapina S, Shevchenko A, Shevchenko A, Verma R and Yates JR (2002). Charting the protein complexome in yeast by mass spectrometry. *Mol Cell Proteomics* **1**: 3–10

Diestel R (2005). *Graph theory*. Springer

Dietze H et al. (2008). 'GoPubMed: Exploring Pubmed with Ontological Background Knowledge'. In: *Ontologies and Text Min-*

ing for Life Sciences : Current Status and Future Perspectives. Ed. by Michael Ashburner, Ulf Leser and Dietrich Rebholz-Schuhmann. Dagstuhl Seminar Proceedings 08131. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany

Ding C, Zhang Y, Li T and Holbrook S (2006). Biclustering Protein Complex Interactions with a Biclique Finding Algorithm : 178–187

Dioum EM, Wauson EM and Cobb MH (2009). MAP-ping unconventional protein-DNA interactions. *Cell* **139**: 462–463

Doms A and Schroeder M (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* **33**: W783–W786

Dongen S (2000). Graph clustering by flow simulation. *Centers for mathematics and computer science (CWI), University of Utrecht* : 49–57

Dortay H, Mehnert N, Bürkle L, Schmülling T and Heyl A (2006). Analysis of protein interactions within the cytokinin-signaling pathway of Arabidopsis thaliana. *FEBS J* **273**: 4631–4644

Dostie J, Mourelatos Z, Yang M, Sharma A and Dreyfuss G (2003). Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA* **9**: 180–186

Drysdale R and Consortium F (2008). FlyBase : a database for the Drosophila research community. *Methods Mol Biol* **420**: 45–59

Duh R and Fürer M (1997). 'Approximation of k-set cover by semi-local optimization'. In: *Proc. 29th Ann. ACM Symp. on Theory of Comp.* ACM, pp. 256–265

Dunn R, Dudbridge F and Sanderson CM (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* **6**: 39

Dwyer T, Koren Y and Marriott K (2006). Drawing directed graphs using quadratic programming. *IEEE Trans Vis Comput Graph* **12**: 536–548

Dziennis S and Alkayed NJ (2008). Role of signal transducer and activator of transcription 3 in neuronal survival and regeneration. *Rev Neurosci* **19**: 341–361

Edachery J, Sen A and Brandenburg F (1999). Graph clustering using distance-k cliques. *Lecture Notes in Computer Science* : 98–106

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863–14868

Eisenberg D (2003). The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A* **100**: 11207–11210

Ekman D, Light S, Björklund AK and Elofsson A (2006). What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol* **7**: R45

Erdös P and Rényi A (1960). Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**: 17–60

Erdős P. Rényi A (1959). On Random Graphs I. *Publ. Math. Debrecen* **6**: 290–297

Espinosa A et al. (2009). Loss of the lupus autoantigen Ro52/Trim21 induces tissue inflammation and systemic autoimmunity by disregulating the IL-23-Th17 pathway. *J Exp Med* **206**: 1661–1671

Evlampiev K and Isambert H (2007). Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst Biol* **1**: 49

Evlampiev K and Isambert H (2008). Conservation and topology of protein interaction networks under duplication-

divergence evolution. *Proc Natl Acad Sci U S A* **105**: 9863–9868

Ewing RM et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3**: 89

Fan J and Bertino JR (1997). K-ras modulates the cell cycle via both positive and negative regulatory pathways. *Oncogene* **14**: 2595–2607

Feder T and Motwani R (1995). Clique partitions, graph compression and speeding-up algorithms. *Journal of Computer and System Sciences* **51**: 261–272

Fernandes L, Rodrigues-Pousada C and Struhl K (1997). Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol Cell Biol* **17**: 6982–6993

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

Filipowicz W, Bhattacharyya SN and Sonenberg N (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9**: 102–114

Fischbach SJ and Carew TJ (2009). MicroRNAs in memory processing. *Neuron* **63**: 714–716

Fitch W and Margoliash E (1967). A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochemical Genetics* **1**: 65–71

Fleischner H, Mujuni E, Paulusma D and Szeider S (2009). Covering graphs with few complete bipartite subgraphs. *Theoretical Computer Science* **410**: 2045 –2053

Formstecher E et al. (2005). Protein interaction mapping: a Drosophila case study. *Genome Res* **15**: 376–384

Fraser HB, Hirsh AE, Wall DP and Eisen MB (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A* **101**: 9033–9038

Friedel CC and Zimmer R (2006). Toward the complete interactome. *Nat Biotechnol* **24**: 614–5

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

Friedel CC, Krumsiek J and Zimmer R (2009). Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J Comput Biol* **16**: 971–987

Fujise K, Zhang D, Liu J and Yeh ET (2000). Regulation of apoptosis and cell cycle progression by MCL1. Differential role of proliferating cell nuclear antigen. *J Biol Chem* **275**: 39458–39465

Fundel K and Zimmer R (2007). Human Gene Normalization by an Integrated Approach including Abbreviation Resolution and Disambiguation. *Second BioCreative Challenge Evaluation Workshop* :

Fundel K, Güttler D, Zimmer R and Apostolakis J (2005). A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics* **6 Suppl 1**: S15

Gagneur J, Krause R, Bouwmeester T and Casari G (2004). Modular decomposition of protein-protein interaction networks. *Genome Biol* **5**: R57

Gallai T (1967). Transitiv orientierbare Graphen. *Acta Mathematica Academiae Scientiarum Hungaricae* **18**: 25–66

Gansner ER and North SC (2000). An open graph visualization system and its applications to software engineering. *Software — Practice and Experience* **30**: 1203–1233

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D and Brown PO (2000). Genomic expres-

sion programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**: 4241–4257

Gatherer D (2010). So what do we really mean when we say that systems biology is holistic? *BMC Syst Biol* **4**: 22

Gavin AC et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141–147

Gavin AC et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631–636

Ge Y, Jensen TL, Matherly LH and Taub JW (2003). Physical and functional interactions between USF and Sp1 proteins regulate human deoxycytidine kinase promoter activity. *J Biol Chem* **278**: 49901–49910

Gentleman RC et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80

Georgii E, Dietmann S, Uno T, Pagel P and Tsuda K (2009). Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* **25**: 933–940

Gerber D, Maerkl SJ and Quake SR (2009). An in vitro microfluidic approach to generating protein-interaction networks. *Nat Methods* **6**: 71–74

Getz G, Levine E and Domany E (2000). Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A* **97**: 12079–12084

Gibson DG et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**: 52–56

Giot L et al. (2003). A protein interaction map of Drosophila melanogaster. *Science* **302**: 1727–1736

Goodsell D (2009). *Machinery of Life*. Copernicus Books

Gosling J and McGilton H (1995). The Java Language Environment: A White Paper. Sun Microsystems

Goto Y, Nonaka I and Horai S (1990). A mutation in the tRNA(Leu)(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature* **348**: 651–653

Grigoriev A (2003). On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res* **31**: 4157–4161

Groessner-Schreiber B and Tuan RS (1992). Enhanced extracellular matrix production and mineralization by osteoblasts cultured on titanium surfaces in vitro. *J Cell Sci* **101 (Pt 1)**: 209–217

Grothaus GA, Mufti A and Murali TM (2006). Automatic layout and visualization of biclusters. *Algorithms Mol Biol* **1**: 15

Guimaraes KS, Jothi R, Zotenko E and Przytycka TM (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biol* **7**: R104

Gunsalus KC et al. (2005). Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature* **436**: 861–865

Gursoy A, Keskin O and Nussinov R (2008). Topological properties of protein interaction networks from a structural perspective. *Biochem Soc Trans* **36**: 1398–1403

Habib M, McConnell R, Paul C and Viennot L (2000). Lex-BFS and partition refinement, with applications to transitive orientation, interval graph recognition and consecutive ones testing. *Theoretical Computer Science* **234**: 59–84

Hakenberg J, Royer L, Plake C, Strobelt H and Schroeder M (2007). 'Me and my friends: gene mention normalization with background knowledge'. In: *Proc 2nd BioCreative Challenge Evaluation Workshop*

Hakenberg J, Bickel S, Plake C, Brefeld U, Zahn H, Faulstich L, Leser U and Scheffer T (2005). Systematic feature evaluation for gene name recognition. *BMC Bioinformatics* **6 Suppl 1**: S9

Hakenberg J, Plake C, Royer L, Strobelt H, Leser U and Schroeder M (2008). Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol* **9 Suppl 2**: S14

Han JDJ, Dupuy D, Bertin N, Cusick ME and Vidal M (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* **23**: 839–844

Han K and Byun Y (2004). Three-dimensional visualization of protein interaction networks. *Comput Biol Med* **34**: 127–139

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

Hanisch D, Fluck J, Mevissen HT and Zimmer R (2003). Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput* : 403–414

Hanisch D, Fundel K, Mevissen HT, Zimmer R and Fluck J (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics* **6 Suppl 1**: S14

Hannah D, Macdonald C and Ounis I (2008). Analysis of Link Graph Compression Techniques. *Lecture Notes in Computer Science* **4956**: 596

Harata M, Oma Y, Mizuno S, Jiang YW, Stillman DJ and Wintersberger U (1999). The nuclear actin-related protein of Saccharomyces cerevisiae, Act3p/Arp4, interacts with core histones. *Mol Biol Cell* **10**: 2595–2605

Hart GT, Ramani AK and Marcotte EM (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**: 120

Hart GT, Lee I and Marcotte ER (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8**: 236

Hartigan J (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association* : 123–129

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999). From molecular to modular cell biology. *Nature* **402**: C47–C52

Hazbun TR et al. (2003). Assigning function to yeast proteins by integration of technologies. *Mol Cell* **12**: 1353–1365

Hermann A et al. (2004). Efficient generation of neural stem cell-like cells from adult human bone marrow stromal cells. *J Cell Sci* **117**: 4411–4422

Hermann A, Liebau S, Gastl R, Fickert S, Habisch HJ, Fiedler J, Schwarz J, Brenner R and Storch A (2006). Comparative analysis of neuroectodermal differentiation capacity of human bone marrow stromal cells using various conversion protocols. *J Neurosci Res* **83**: 1502–1514

Hermjakob H et al. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**: D452–D455

Herzel, Ebeling and Schmitt (1994). Entropies of biosequences: The role of repeats. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **50**: 5061–5071

Heymans JJ, Ulanowicz RE and Bondavalli C (2002). Network analysis of the South Florida Everglades graminoid marshes and comparison with nearby cypress ecosystems. *Ecological Modelling* **149**: 5 –23

Hickling KC, Hitchcock JM, Oreffo V, Mally A, Hammond TG, Evans JG and Chipman JK (2010). Evidence of oxidative stress and associated DNA damage, increased proliferative drive, and altered gene ex-

pression in rat liver produced by the cholangiocarcinogenic agent furan. *Toxicol Pathol* **38**: 230–243

Hirschman L, Yeh A, Blaschke C and Valencia A (2005a). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6 Suppl 1**: S1

Hirschman L, Colosimo M, Morgan A and Yeh A (2005b). Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics* **6 Suppl 1**: S11

Hjelmervik TOR, Petersen K, Jonassen I, Jonsson R and Bolstad AI (2005). Gene expression profiling of minor salivary glands clearly distinguishes primary Sjögren's syndrome patients from healthy control subjects. *Arthritis Rheum* **52**: 1534–1544

Ho Y et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**: 180–183

Hoffmann R and Valencia A (2004). A gene network for navigating the literature. *Nat Genet* **36**: 664

Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C and Valencia A (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE* **2005**: pe21

Holland PW and Leinhardt S (1971). Transitivity in structural models of small groups. *Comparative Groups Studies* : 107–124

Hollunder J, Beyer A and Wilhelm T (2005). Identification and characterization of protein subcomplexes in yeast. *Proteomics* **5**: 2082–2089

Hong EL et al. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res* **36**: D577–D581

Hu Z, Mellor J, Wu J, Kanehisa M, Stuart JM and DeLisi C (2007). Towards zoomable multidimensional maps of the cell. *Nat Biotechnol* **25**: 547–554

Huang HS, Chen CJ and Chang WC (1999). The CCAAT-box binding factor NF-Y is required for the expression of phospholipid hydroperoxide glutathione peroxidase in human epidermoid carcinoma A431 cells. *FEBS Lett* **455**: 111–116

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS and O'Shea EK (2003). Global analysis of protein localization in budding yeast. *Nature* **425**: 686–691

Hunter S et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**: D211–D215

Huttenhower C, Mehmood SO and Troyanskaya OG (2009). Graphle: Interactive exploration of large, dense graphs. *BMC Bioinformatics* **10**: 417

Hwang W, Cho YR, Zhang A and Ramanathan M (2006). A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol Biol* **1**: 24

Iragne F, Nikolski M, Mathieu B, Auber D and Sherman D (2005). ProViz: protein interaction visualization and exploration. *Bioinformatics* **21**: 272–274

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264

Ispolatov I, Krapivsky PL and Yuryev A (2005). Duplication-divergence model of protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys* **71**: 061911

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S and Sakaki Y (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combin-

ations between the yeast proteins. *Proc Natl Acad Sci U S A* **97**: 1143–1147

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y (2001a). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**: 4569–4574

Ito T, Tashiro K and Kuhara T (2001b). [Systematic analysis of Saccharomyces cerevisiae genome: gene network and protein-protein interaction network]. *Tanpakushitsu Kakusan Koso* **46**: 2407–2413

Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS and Kaelin WG (2001). HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. *Science* **292**: 464–468

Jaccard P (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* **37**: 241–272

Jain A, Murty M and Flynn P (1999). Data clustering: a review. *ACM computing surveys (CSUR)* **31**: 264–323

Jansen R, Greenbaum D and Gerstein M (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**: 37–46

Jeong H, Mason SP, Barabási AL and Oltvai ZN (2001). Lethality and centrality in protein networks. *Nature* **411**: 41–42

Ji H et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **13**: 12

Ji L, Bing-Hong W, Wen-Xu W and Tao Z (2008). Network Entropy Based on Topology Configuration and Its Computation to Random Networks. *Chinese Physics Letters* **25**: 4177–4180

Jin F, Hazbun T, Michaud GA, Salcius M, Predki PF, Fields S and Huang J (2006). A pooling-deconvolution strategy for biological network elucidation. *Nat Methods* **3**: 183–189

Johnson S (1967). Hierarchical clustering schemes. *Psychometrika* **32**: 241–254

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

Kao M, Occhiogrosso N and Teng S (1998). Simple and efficient graph compression schemes for dense and complement graphs. *Journal of Combinatorial Optimization* **2**: 351–359

Karppa M, Herva R, Moslemi AR, Oldfors A, Kakko S and Majamaa K (2005). Spectrum of myopathic findings in 50 patients with the 3243A>G mutation in mitochondrial DNA. *Brain* **128**: 1861–1869

Kashtan N, Itzkovitz S, Milo R and Alon U (2004). Topological generalizations of network motifs. *Phys Rev E Stat Nonlin Soft Matter Phys* **70**: 031909

Kashtan N and Alon U (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A* **102**: 13773–13778

Keller DM, Zeng X, Wang Y, Zhang QH, Kapoor M, Shu H, Goodman R, Lozano G, Zhao Y and Lu H (2001). A DNA damage-induced p53 serine 392 kinase complex contains CK2, hSpt16, and SSRP1. *Mol Cell* **7**: 283–292

Khanin R and Wit E (2006). How scale-free are biological networks. *J Comput Biol* **13**: 810–818

Kikuno R, Nagase T, Nakayama M, Koga H, Okazaki N, Nakajima D and Ohara O (2004). HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE. *Nucleic Acids Res* **32**: D502–D504

Kim PM, Lu LJ, Xia Y and Gerstein MB (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**: 1938–1941

Kim PM, Korbel JO and Gerstein MB (2007). Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**: 20274–20279

Kim S and Lee J (2006). BAG: a graph theoretic sequence clustering algorithm. *Int J Data Min Bioinform* **1**: 178–200

Kim WK, Park J and Suh JK (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform* **13**: 42–50

King AD, Przulj N and Jurisica I (2004). Protein complex prediction via cost-based clustering. *Bioinformatics* **20**: 3013–3020

Kittler R et al. (2007). Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nat Cell Biol* **9**: 1401–1412

Klamt S, Haus UU and Theis F (2009). Hypergraphs and cellular networks. *PLoS Comput Biol* **5**: e1000385

Klinger R, Friedrich CM, Mevissen HT, Fluck J, Hofmann-Apitius M, Furlong LI and Sanz F (2007). Identifying gene-specific variations in biomedical text. *J Bioinform Comput Biol* **5**: 1277–1296

Kluger Y, Basri R, Chang JT and Gerstein M (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* **13**: 703–716

Knuth DE (1993). 'The Stanford GraphBase. A Platform for Combinatorial Computing'

Kocsor A, Kertész-Farkas A, Kaján L and Pongor S (2006). Application of compression-based distance measures to protein sequence classification: a methodological study. *Bioinformatics* **22**: 407–412

Koegl M and Uetz P (2007). Improving yeast two-hybrid screening systems. *Brief Funct Genomic Proteomic* **6**: 302–312

Kojima K, Nagasaki M and Miyano S (2010). An efficient biological pathway layout algorithm combining grid-layout and spring embedder for complicated cellular location information. *BMC Bioinformatics* **11**: 335

Kolodrubetz D, Rykowski MC and Grunstein M (1982). Histone H2A subtypes associate interchangeably in vivo with histone H2B subtypes. *Proc Natl Acad Sci U S A* **79**: 7814–7818

Kong HJ et al. (2007). Cutting edge: autoantigen Ro52 is an interferon inducible E3 ligase that ubiquitinates IRF-8 and enhances cytokine expression in macrophages. *J Immunol* **179**: 26–30

Koudriavtsev A, Jameson R and Linert W (2001). *The law of mass action*. Springer, Berlin

Koyuturk M, Szpankowski W and Grama A (2004). Biclustering gene-feature matrices for statistically significant dense patterns : 480–484

Koyutürk M, Grama A and Szpankowski W (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics* **20 Suppl 1**: i200–i207

Kozlenkov A, Penaloza R, Nigam V, Royer L, Dawelbait G and Schroeder M (2006). Prova: Rule-Based Java Scripting for Distributed Web Applications: A Case Study in Bioinformatics**. 4254**: 899–908

Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, Hirschman L and Valencia A (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol* **9 Suppl 2**: S1

Kratzke T, Reznick B and West D (1988). Eigensharp graphs: Decomposition into

complete bipartite subgraphs. *Trans. Amer. Math. Soc.* **308(2)**: 637–653

Krichevsky AM, Sonntag KC, Isacson O and Kosik KS (2006). Specific microRNAs modulate embryonic stem cell-derived neurogenesis. *Stem Cells* **24**: 857–864

Krogan NJ et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**: 637–643

Krumsiek J, Friedel CC and Zimmer R (2008). ProCope–protein complex prediction and evaluation. *Bioinformatics* **24**: 2115–2116

Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigó R and Gladyshev VN (2003). Characterization of mammalian selenoproteomes. *Science* **300**: 1439–1443

Kuhn T, Royer L, Fuchs NE and Schroeder M (2006). Improving Text Mining with Controlled Natural Language: A Case Study for Protein Interations :

Kursula I, Salin M, Sun J, Norledge BV, Haapalainen AM, Sampson NS and Wierenga RK (2004). Understanding protein lids: structural analysis of active hinge mutants in triosephosphate isomerase. *Protein Eng Des Sel* **17**: 375–382

Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, Rawlings C, Verrier P and Philippi S (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* **22**: 1383–1390

LaCount DJ et al. (2005). A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* **438**: 103–107

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W and Tuschl T (2002). Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12**: 735–739

Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R and Cesareni G (2004). Protein interaction networks by proteome peptide scanning. *PLoS Biol* **2**: E14

Langville A and Meyer C (2004). Deeper inside pagerank. *Internet Mathematics* **1**: 335–380

Lee I, Date SV, Adai AT and Marcotte EM (2004). A probabilistic functional network of yeast genes. *Science* **306**: 1555–1558

Lemmens I, Lievens S and Tavernier J (2010). Strategies towards high-quality binary protein interactome maps. *J Proteomics* **73**: 1415–1420

Leskovec J, Kleinberg J and Faloutsos C (2005). Graphs over time: densification laws, shrinking diameters and possible explanations : 177–187

Li D, Li J, Ouyang S, Wang J, Wu S, Wan P, Zhu Y, Xu X and He F (2006a). Protein interaction networks of Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster: large-scale organization and robustness. *Proteomics* **6**: 456–461

Li H, Li J and Wong L (2006b). Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics* **22**: 989–996

Li M, Badger JH, Chen X, Kwong S, Kearney P and Zhang H (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* **17**: 149–154

Li S et al. (2004). A map of the interactome network of the metazoan C. elegans. *Science* **303**: 540–543

Lim J et al. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125**: 801–814

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS and Johnson JM (2005). Mi-

croarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**: 769–773

Lima-Mendez G and Van Helden J (2009). The powerful law of the power law and other myths in network biology. *Mol Biosyst* **5**: 1482–1493

Lindal S, Lund I, Torbergsen T, Aasly J, Mellgren SI, Borud O and Monstad P (1992). Mitochondrial diseases and myopathies: a series of muscle biopsy specimens with ultrastructural changes in the mitochondria. *Ultrastruct Pathol* **16**: 263–275

Lindstrom DL, Squazzo SL, Muster N, Burckin TA, Wachter KC, Emigh CA, McCleery JA, Yates JR and Hartzog GA (2003). Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. *Mol Cell Biol* **23**: 1368–1378

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

Lu H (2002). Linear-time compression of bounded-genus graphs into information-theoretically optimal number of bits : 223–224

MacArthur BD, Sánchez-García RJ and Anderson JW (2008). Symmetry in complex networks. *Discrete Applied Mathematics* **156**: 3525 –3531

Maglott D, Ostell J, Pruitt KD and Tatusova T (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **35**: D26–D31

Maisel M et al. (2007). Transcription profiling of adult and fetal human neuroprogenitors identifies divergent paths to maintain the neuroprogenitor cell state. *Stem Cells* **25**: 1231–1240

Maisel M et al. (2010). Genome-wide expression profiling and functional network analysis upon neuroectodermal conversion of human mesenchymal stem cells suggest HIF-1 and miR-124a as important regulators. *Exp Cell Res* **13**: 234

Manke T, Demetrius L and Vingron M (2006). An entropic characterization of protein interaction networks and cellular robustness. *J R Soc Interface* **3**: 843–850

Mann M, Hendrickson RC and Pandey A (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**: 437–473

Manni I, Caretti G, Artuso S, Gurtner A, Emiliozzi V, Sacchi A, Mantovani R and Piaggio G (2008). Posttranslational regulation of NF-YA modulates NF-Y transcriptional activity. *Mol Biol Cell* **19**: 5203–5213

Manwaring N, Jones MM, Wang JJ, Rochtchina E, Howard C, Mitchell P and Sue CM (2007). Population prevalence of the MELAS A3243G mutation. *Mitochondrion* **7**: 230–233

Maslov S and Sneppen K (2002). Specificity and stability in topology of protein networks. *Science* **296**: 910–913

Maslov S, Sneppen K, Eriksen KA and Yan KK (2004). Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* **4**: 9

Mason PB and Struhl K (2003). The FACT complex travels with elongating RNA polymerase II and is important for the fidelity of transcriptional initiation in vivo. *Mol Cell Biol* **23**: 8323–8333

Matsuda H, Ishihara T and Hashimoto A (1999). Classifying molecular sequences using a linkage graph with their pairwise similarities. *Theoretical Computer Science* **210**: 305–325

Matsuno H, Yokoyama A, Watari F, Uo M and Kawasaki T (2001). Biocompatibility and osteogenesis of refractory metal implants, titanium, hafnium, niobium, tan-

talum and rhenium. *Biomaterials* **22**: 1253–1262

Maxwell PH, Wiesener MS, Chang GW, Clifford SC, Vaux EC, Cockman ME, Wykoff CC, Pugh CW, Maher ER and Ratcliffe PJ (1999). The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature* **399**: 271–275

McConnell R and De Montgolfier F (2005). Linear-time modular decomposition of directed graphs. *Discrete Applied Mathematics* **145**: 198–209

Mechelen IV, Bock HH and Boeck PD (2004). Two-mode clustering methods: a structured overview. *Stat Methods Med Res* **13**: 363–394

Medini D, Covacci A and Donati C (2006). Protein Homology Network Families Reveal Step-Wise Diversification of Type III and Type IV Secretion Systems. *PLoS Comput Biol* **2**: e173

Mehta CR, Patel NR and Tsiatis AA (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **40**: 819–825

Mende S, Royer L, Herr A, Schmiedel J, Deschauer M, Klopstock T, Kostic VS, Schroeder M, Reichmann H and Storch A (2010). Expression profiling and network analysis reveal MELAS master regulators. *submitted* :

Mendizabal I, Rios G, Mulet JM, Serrano R and De Larrinoa IF (1998). Yeast putative transcription factors involved in salt tolerance. *FEBS Lett* **425**: 323–328

Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S and Frishman D (1999). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **27**: 44–48

Middendorf M, Ziv E and Wiggins CH (2005). Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *Proc Natl Acad Sci U S A* **102**: 3192–3197

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U (2002). Network motifs: simple building blocks of complex networks. *Science* **298**: 824–827

Mirkin B (1996). *Mathematical classification and clustering*. Springer

Mistry AS and Mikos AG (2005). Tissue engineering strategies for bone regeneration. *Adv Biochem Eng Biotechnol* **94**: 1–22

Mody N, Parhami F, Sarafian TA and Demer LL (2001). Oxidative stress modulates osteoblastic differentiation of vascular and bone cells. *Free Radic Biol Med* **31**: 509–519

Moran L, Norris D and Osley MA (1990). A yeast H2A-H2B promoter can be regulated by changes in histone gene copy number. *Genes Dev* **4**: 752–763

Morgan AA, Hirschman L, Colosimo M, Yeh AS and Colombe JB (2004). Gene name identification and normalization using a model organism database. *J Biomed Inform* **37**: 396–410

Morgan AA et al. (2008). Overview of BioCreative II gene normalization. *Genome Biol* **9 Suppl 2**: S3

Morrison JL, Breitling R, Higham DJ and Gilbert DR (2006). A lock-and-key model for protein-protein interactions. *Bioinformatics* **22**: 2012

Morton LM et al. (2009). Risk of non-Hodgkin lymphoma associated with germline variation in genes that regulate the cell cycle, apoptosis, and lymphocyte development. *Cancer Epidemiol Biomarkers Prev* **18**: 1259–1270

Motamed-Khorasani A, Jurisica I, Letarte M, Shaw PA, Parkes RK, Zhang X, Evangelou A, Rosen B, Murphy KJ and Brown TJ (2007). Differentially androgen-modulated genes in ovarian epithelial

cells from BRCA mutation carriers and control patients predict ovarian cancer survival and disease progression. *Oncogene* **26**: 198–214

Mowshowitz (1968). Entropy and the complexity of the graphs. i: An index of the relative complexity of a graph. *Bull. Math. Biophys.* **30**: 75–204

Mullighan CG, Heatley S, Lester S, Rischmueller M, Gordon TP and Bardy PG (2004). Fas gene promoter polymorphisms in primary Sjögren's syndrome. *Ann Rheum Dis* **63**: 98–101

Murtagh F (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal* **26**: 354

Nagayasu T, Imamura K and Nakanishi K (2005). Adsorption characteristics of various organic substances on the surfaces of tantalum, titanium, and zirconium. *J Colloid Interface Sci* **286**: 462–470

Nagy G (1968). State of the art in pattern recognition. *Proc. IEEE* **56**: 836–862

Newman MEJ (2002). Assortative mixing in networks. *Phys Rev Lett* **89**: 208701

Newman MEJ (2003). Mixing patterns in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **67**: 026126

Newman MEJ (2003). The structure and function of complex networks. *SIAM Review* **45**: 167

Ng A, Jordan M and Weiss Y (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **2**: 849–856

Ng SK, Zhang Z and Tan SH (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* **19**: 923–929

Nomiyama T, Tanaka Y, Hattori N, Nishimaki K, Nagasaka K, Kawamori R and Ohta S (2002). Accumulation of somatic mutation in mitochondrial DNA extracted from peripheral blood cells in diabetic patients. *Diabetologia* **45**: 1577–1583

Norlen K, Lucas G, Gebbie M and Chuang J (2002). 'Visualization and Analysis of the Telecommunications and Media Ownership Network.' In: *Proceedings of International Telecommunications Society 14th Biennial Conference*

Nye TMW, Berzuini C, Gilks WR, Babu MM and Teichmann SA (2005). Statistical analysis of domains in interacting protein pairs. *Bioinformatics* **21**: 993–1001

Nye TMW, Berzuini C, Gilks WR, Babu MM and Teichmann S (2006). Predicting the strongest domain-domain contact in interacting protein pairs. *Stat Appl Genet Mol Biol* **5**: 5

Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S and Kanehisa M (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* **36**: W423–W426

Ota T et al. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**: 40–45

Palla G, Derényi I, Farkas I and Vicsek T (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**: 814–818

Papadopoulos C and Voglis C (2006). Drawing graphs using modular decomposition. *Lecture Notes in Computer Science* **3843**: 343

Park EJ, Yi J, Chung KH, Ryu DY, Choi J and Park K (2008). Oxidative stress and apoptosis induced by titanium dioxide nanoparticles in cultured BEAS-2B cells. *Toxicol Lett* **180**: 222–229

Park J and Newman MEJ (2004). Statistical mechanics of networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **70**: 066117

Park K and Kim D (2009). Localized network centrality and essentiality in the yeast-

protein interaction network. *Proteomics* **9**: 5143–5154

Parkinson EK, Fitchett C and Cereser B (2008). Dissecting the non-canonical functions of telomerase. *Cytogenet Genome Res* **122**: 273–280

Parrish JR et al. (2007). A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biol* **8**: R130

Pastor-Satorras R, Smith E and Solé RV (2003). Evolving protein interaction networks through gene duplication. *J Theor Biol* **222**: 199–210

Pati A, Vasquez-Robinet C, Heath LS, Grene R and Murali TM (2006). XcisClique: analysis of regulatory bicliques. *BMC Bioinformatics* **7**: 218

Patil A and Nakamura H (2005). Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics* **6**: 100

Pavlakis SG, Phillips PC, DiMauro S, Vivo DCD and Rowland LP (1984). Mitochondrial myopathy, encephalopathy, lactic acidosis, and strokelike episodes: a distinctive clinical syndrome. *Ann Neurol* **16**: 481–488

Pawson T (2003). Organization of cell-regulatory systems through modular-protein-interaction domains. *Philos Transact A Math Phys Eng Sci* **361**: 1251–1262

Pawson T and Nash P (2003). Assembly of cell regulatory systems through protein interaction domains. *Science* **300**: 445–452

Peart MJ and Prives C (2006). Mutant p53 gain of function: the NF-Y connection. *Cancer Cell* **10**: 173–174

Peeters R (2003). The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics* **131**: 651 –654

Pereira-Leal JB, Enright AJ and Ouzounis CA (2004). Detection of functional modules from protein interaction networks. *Proteins* **54**: 49–57

Pils B and Schultz J (2004). Evolution of the multifunctional protein tyrosine phosphatase family. *Mol Biol Evol* **21**: 625–631

Pinkert S, Schultz J and Reichardt J (2010). Protein interaction networks–more than mere modules. *PLoS Comput Biol* **6**: e1000659

Pipenbacher P, Schliep A, Schneckener S, Schönhuth A, Schomburg D and Schrader R (2002). ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics* **18 Suppl 2**: S182–S191

Plake C, Schiemann T, Pankalla M, Hakenberg J and Leser U (2006). AliBaba: PubMed as a graph. *Bioinformatics* **22**: 2444–2445

Plake C, Royer L, Winnenburg R, Hakenberg J and Schroeder M (2009). GoGene: gene annotation in the fast lane. *Nucleic Acids Res* **37**: W300–W304

Prasad TSK et al. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res* **37**: D767–D772

Przulj N, Corneil DG and Jurisica I (2004). Modeling interactome: scale-free or geometric? *Bioinformatics* **20**: 3508–3515

Przulj N, Corneil DG and Jurisica I (2006). Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics* **22**: 974–980

Przulj N and Higham DJ (2006). Modelling protein-protein interaction networks via a stickiness index. *J R Soc Interface* **3**: 711–716

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M and Séraphin B (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**: 218–229

Rachlin J, Cohen DD, Cantor C and Kasif S (2006). Biological context networks: a mosaic view of the interactome. *Mol Syst Biol* **2**: 66

Rain JC et al. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature* **409**: 211–215

Ramadan E, Tarafdar A and Pothen A (2004). 'A Hypergraph Model for the Yeast Protein Complex Network'. In: *Proceedings of the Sixth IEEE International Workshop on High Performance Computational Biology*

Randall KH, Stata R, Wiener JL and Wickremesinghe RG (2002). The Link Database: Fast Access to Graphs of the Web. *Data Compression Conference* **0**: 0122

Rashewsky (1955). N. Rashewsky, Life, information theory, and topology. *Bull. Math. Biophys.* **17**: 229–235

Ratcliffe PJ (2007). HIF-1 and HIF-2: working alone or together in hypoxia? *J Clin Invest* **117**: 862–865

Reguly T et al. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* **5**: 11

Reiss DJ, Baliga NS and Bonneau R (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7**: 280

Remy I and Michnick SW (2006). A highly sensitive protein-protein interaction assay based on Gaussia luciferase. *Nat Methods* **3**: 977–979

Remy I and Michnick SW (2007). Application of protein-fragment complementation assays in cell biology. *Biotechniques* **42**: 137, 139, 141

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

Rhodes DA, Ihrke G, Reinicke AT, Malcherek G, Towey M, Isenberg DA and Trowsdale J (2002). The 52 000 MW Ro/SS-A autoantigen in Sjögren's syndrome/systemic lupus erythematosus (Ro52) is an interferon-gamma inducible tripartite motif protein associated with membrane proximal structures. *Immunology* **106**: 246–256

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Séraphin B (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**: 1030–1032

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

Rivals I, Personnaz L, Taing L and Potier MC (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**: 401–407

Royer L, Reimann M, Andreopoulos B and Schroeder M (2008). Unraveling protein networks with power graph analysis. *PLoS Comput Biol* **4**: e1000108

Royer L, Plake C and Schroeder M (2009). 'Identification of cancer and cell-cycle genes with protein interactions and literature mining'. In: *Proceedings of the German Conference on Bioinformatics. GCB 2009*. Vol. 8

Royer L, Stewart AF and Schroeder M (2010). 'Compressibility as a Novel Systemic Measure for Coverage and Accuracy of Protein Interaction Networks'. submitted

Royer L, Linse B, Wächter T, Furch T, Bry F and Schroeder M (2007). Querying Semantic Web Contents : 31–52

Rual JF et al. (2005). Towards a proteome-scale map of the human protein-

protein interaction network. *Nature* **437**: 1173–1178

Rzhetsky A and Gomez SM (2001). Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* **17**: 988–996

Safran M et al. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford)* **2010**: baq020

Saito I et al. (1999). Fas ligand-mediated exocrinopathy resembling Sjögren's syndrome in mice transgenic for IL-10. *J Immunol* **162**: 2488–2494

Salomon D, Motta G and Bryant D (2007). *Data compression: the complete reference*. Springer-Verlag New York

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**: D449–D451

Sanchez C, Lachaize C, Janody F, Bellon B, Röder L, Euzenat J, Rechenmann F and Jacq B (1999). Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Res* **27**: 89–94

Sato S, Shimoda Y, Muraki A, Kohara M, Nakamura Y and Tabata S (2007). A large-scale protein protein interaction analysis in Synechocystis sp. PCC6803. *DNA Res* **14**: 207–216

Schaefer AM, Phoenix C, Elson JL, McFarland R, Chinnery PF and Turnbull DM (2006). Mitochondrial disease in adults: a scale to monitor progression and treatment. *Neurology* **66**: 1932–1934

Schlessinger J (1994). SH2/SH3 signaling proteins. *Curr Opin Genet Dev* **4**: 25–30

Schlicker A, Domingues FS, Rahnenführer J and Lengauer T (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**: 302

Schneider DA, French SL, Osheim YN, Bailey AO, Vu L, Dodd J, Yates JR, Beyer AL and Nomura M (2006). RNA polymerase II elongation factors Spt4p and Spt5p play roles in transcription elongation by RNA polymerase I and rRNA processing. *Proc Natl Acad Sci U S A* **103**: 12707–12712

Scholtens D, Vidal M and Gentleman R (2005). Local modeling of global interactome networks. *Bioinformatics* **21**: 3548–3557

Schreiber F, Dwyer T, Marriott K and Wybrow M (2009). A generic algorithm for layout of biological networks. *BMC Bioinformatics* **10**: 375

Shannon CE (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27**: 379–423 and 623–656

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

Shen X, Mizuguchi G, Hamiche A and Wu C (2000). A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406**: 541–544

Sheng Q, Moreau Y and Moor BD (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19 Suppl 2**: ii196–ii205

Shevchenko A, Zachariae W and Shevchenko A (1999). A strategy for the characterization of protein interaction networks by mass spectrometry. *Biochem Soc Trans* **27**: 549–554

Shevchenko A, Roguev A, Schaft D, Buchanan L, Habermann B, Sakalar C, Thomas H, Krogan NJ, Shevchenko A and Stewart AF (2008). Chromatin Central: towards the comparative proteome by

accurate mapping of the yeast proteomic environment. *Genome Biol* **9**: R167

Shoemaker BA and Panchenko AR (2007a). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol* **3**: e42

Shoemaker BA and Panchenko AR (2007b). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* **3**: e43

Simonis N et al. (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat Methods* **6**: 47–54

Singh AM and Dalton S (2009). The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. *Cell Stem Cell* **5**: 141–149

Sivachenko AY and Yuryev A (2007). Pathway analysis software as a tool for drug target selection, prioritization and validation of drug mechanism. *Expert Opin Ther Targets* **11**: 411–421

Smith L et al. (2008). Overview of BioCreative II gene mention recognition. *Genome Biol* **9 Suppl 2**: S2

Sneath P (2005). Numerical taxonomy. *Bergey's Manual® of Systematic Bacteriology* : 39–42

Sorribas A, Hernández-Bermejo B, Vilaprinyo E and Alves R (2007). Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations. *Biotechnol Bioeng* **97**: 1259–1277

Sousa SR, Lamghari M, Sampaio P, Moradas-Ferreira P and Barbosa MA (2008). Osteoblast adhesion and morphology on TiO2 depends on the competitive preadsorption of albumin and fibronectin. *J Biomed Mater Res A* **84**: 281–290

Spirin V and Mirny LA (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**: 12123–12128

Sproule DM and Kaufmann P (2008). Mitochondrial encephalopathy, lactic acidosis, and strokelike episodes: basic concepts, clinical phenotype, and therapeutic management of MELAS syndrome. *Ann N Y Acad Sci* **1142**: 133–158

Stanyon CA, Liu G, Mangiola BA, Patel N, Giot L, Kuang B, Zhang H, Zhong J and Finley RL (2004). A Drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol* **5**: R96

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A and Tyers M (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**: D535–D539

Stelzl U et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**: 957–968

Stiehler C et al. (2010). Tantalum coating induces earlier differentiation of mesenchymal stem cells compared with titanium surface. *Submitted* :

Stiehler M, Lind M, Mygind T, Baatrup A, Dolatshahi-Pirouz A, Li H, Foss M, Besenbacher F, Kassem M and Bünger C (2008). Morphology, proliferation, and osteogenic differentiation of mesenchymal stem cells cultured on titanium, tantalum, and chromium surfaces. *J Biomed Mater Res A* **86**: 448–458

Strassburger K and Bretz F (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Stat Med* **27**: 4914–4927

Strogatz SH (2001). Exploring complex networks. *Nature* **410**: 268–276

Stumpf MPH, Wiuf C and May RM (2005). Subnets of scale-free networks are not

scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A* **102**: 4221–4224

Stumpf MPH, Thorne T, De Silva E, Stewart R, An HJ, Lappe M and Wiuf C (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci U S A* **105**: 6959–6964

Suderman M and Hallett M (2007). Tools for visually exploring biological networks. *Bioinformatics* **23**: 2651–2659

Suh MR et al. (2004). Human embryonic stem cells express a unique set of microRNAs. *Dev Biol* **270**: 488–498

Sun J, Bollt EM and Ben Avraham D (2008). Graph compression—save information by exploiting redundancy. *Journal of Statistical Mechanics: Theory and Experiment* **2008**: P06001

Suthram S, Shlomi T, Ruppin E, Sharan R and Ideker T (2006). A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7**: 360

Syropoulos A (2001). 'Mathematics of Multisets'. In: *Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View*. Ed. by Cristian S. Calude, Gheorghe Paun, Grzegorz Rozenberg and Arto Salomaa. Vol. 2235. Lecture Notes in Computer Science. Springer-Verlag, Berlin, Germany, pp. 347–358

Tan PN, Steinbach M and Kumar V (2005). *Introduction to Data Mining*. 1st ed. Addison Wesley. ISBN: 0321321367

Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L and Weinstein JN (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* **27**: 1210–4, 1216–7

Tanabe L, Xie N, Thom LH, Matten W and Wilbur WJ (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics* **6 Suppl 1**: S3

Tanay A, Sharan R and Shamir R (2005). Biclustering algorithms: A survey. *Handbook of computational molecular biology* **9**: 26–1

Tanay A, Sharan R and Shamir R (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18 Suppl 1**: S136–S144

Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H and Michnick SW (2008). An in vivo map of the yeast protein interactome. *Science* **320**: 1465–1470

Taylor JS and Raes J (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**: 615–643

Tedder M, Corneil D, Habib M and Paul C (2008). Simpler linear-time modular decomposition via recursive factorizing permutations. *Lecture Notes in Computer Science* **5125**: 634–645

Teichmann SA and Babu MM (2004). Gene regulatory network growth by duplication. *Nat Genet* **36**: 492–496

Tetko IV, Facius A, Ruepp A and Mewes HW (2005). Super paramagnetic clustering of protein sequences. *BMC Bioinformatics* **6**: 82

Thomas A, Cannings R, Monk NAM and Cannings C (2003). On the structure of protein-protein interaction networks. *Biochem Soc Trans* **31**: 1491–1496

Titz B, Rajagopala SV, Goll J, Häuser R, McKevitt MT, Palzkill T and Uetz P (2008). The binary protein interactome of Treponema pallidum–the syphilis spirochete. *PLoS One* **3**: e2292

Torbergsen T, Aasly J, Borud O, Lindal S and Mellgren SI (1991). Mitochondrial myopathy in Marinesco-Sjögren syn-

drome. *J Ment Defic Res* **35 (Pt 2)**: 154–159

Torreira E, Jha S, López-Blanco JR, Arias-Palomo E, Chacón P, Cañas C, Ayora S, Dutta A and Llorca O (2008). Architecture of the pontin/reptin complex, essential in the assembly of several macromolecular complexes. *Structure* **16**: 1511–1520

Tschernitschek H, Borchers L and Geurtsen W (2005). Nonalloyed titanium as a bioinert metal–a review. *Quintessence Int* **36**: 523–530

Tsuzaka K, Matsumoto Y, Sasaki Y, Abe T, Tsubota K and Takeuchi T (2007). Down-regulation of Fas-ligand mRNA in Sjögren's syndrome patients with enlarged exocrine glands. *Autoimmunity* **40**: 497–502

Tutte WT (1998). *Graph Theory As I Have Known It*. Oxford University Press

Uetz P et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**: 623–627

Valencia A and Pazos F (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* **12**: 368–373

Van Kooten TG, Klein CL and Kirkpatrick CJ (2000). Cell-cycle control in cell-biomaterial interactions: expression of p53 and Ki67 in human umbilical vein endothelial cells in direct contact and extract testing of biomaterials. *J Biomed Mater Res* **52**: 199–209

Van der Geer P and Pawson T (1995). The PTB domain: a new protein module implicated in signal transduction. *Trends Biochem Sci* **20**: 277–280

Venkatesan K et al. (2009). An empirical framework for binary interactome mapping. *Nat Methods* **6**: 83–90

Venn J (1880). On the Diagrammatic and Mechanical Representation of Propositions and Reasonings. *Dublin Philosophical Magazine and Journal of Science* **9**: 1–18

Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399–403

Von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**: 258–261

Wagner A (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**: 1283–1292

Wang L, Azad N, Kongkaneramit L, Chen F, Lu Y, Jiang BH and Rojanasakul Y (2008). The Fas death signaling pathway connecting reactive oxygen species generation and FLICE inhibitory protein down-regulation. *J Immunol* **180**: 3072–3080

Wang Z and Zhang J (2007). In search of the biological significance of modular structures in protein networks. *PLoS Comput Biol* **3**: e107

Ward Jr J (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**: 236–244

Watts DJ and Strogatz SH (1998). Collective dynamics of 'small-world' networks. *Nature* **393**: 440–442

Weiss O, Jiménez-Montaño MA and Herzel H (2000). Information content of protein sequences. *J Theor Biol* **206**: 379–386

Wernicke S and Rasche F (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics* **22**: 1152–1153

White J, Southgate E, Thomson J and Brenner S (1986). The structure of the nervous system of the nematode Caen-

orhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **314**: 1

White S and Smyth P (2005). A spectral clustering approach to finding communities in graph :

Whitty A (2008). Cooperativity and biological complexity. *Nat Chem Biol* **4**: 435–439

Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T and Harrow JL (2008). The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36**: D753–D760

Wingender E (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform* **9**: 326–332

Winter C, Henschel A, Kim WK and Schroeder M (2006). SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res* **34**: D310–D314

Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP and Hautaniemi S (2009a). Integrated network analysis platform for protein-protein interactions. *Nat Methods* **6**: 75–77

Wu WH, Wu CH, Ladurner A, Mizuguchi G, Wei D, Xiao H, Luk E, Ranjan A and Wu C (2009b). N Terminus of Swr1 Binds to Histone H2AZ and Provides a Platform for Subunit Assembly in the Chromatin Remodeling Complex. *J Biol Chem* **284**: 6200–6207

Wu Z et al. (1999). Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. *Cell* **98**: 115–124

Wuchty S, Oltvai ZN and Barabási AL (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* **35**: 176–179

Xin X, Rual JF, Hirozane-Kishikawa T, Hill DE, Vidal M, Boone C and Thierry-Mieg N (2009). Shifted Transversal Design smart-pooling for high coverage interactome mapping. *Genome Res* **19**: 1262–1269

Yang H, Lin CH, Ma G, Orr M, Baffi MO and Wathelet MG (2002). Transcriptional activity of interferon regulatory factor (IRF)-3 depends on multiple protein-protein interactions. *Eur J Biochem* **269**: 6142–6151

Ye J, McGinnis S and Madden TL (2006). BLAST: improvements for better sequence analysis. *Nucleic Acids Res* **34**: W6–W9

Yeh A, Morgan A, Colosimo M and Hirschman L (2005). BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics* **6 Suppl 1**: S2

Yoo AS, Staahl BT, Chen L and Crabtree GR (2009). MicroRNA-mediated. *Nature* **460**: 642–646

Yook SH, Oltvai ZN and Barabási AL (2004). Functional and topological characterization of protein interaction networks. *Proteomics* **4**: 928–942

Yu H et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science* **322**: 104–10

Zhang B, Park BH, Karpinets T and Samatova NF (2008). From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* **24**: 979–986

Zhong J, Zhang H, Stanyon CA, Tromp G and Finley RL (2003). A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating. *Genome Res* **13**: 2691–2699

Zhou G, Shen D, Zhang J, Su J and Tan S (2005). Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics* **6 Suppl 1**: S7

Zitter H and Plenk H (1987). The electrochemical behavior of metallic implant materials as an indicator of their biocompatibility. *J Biomed Mater Res* **21**: 881–896

Zotenko E, Mestre J, O'Leary DP and Przytycka TM (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**: e1000140