
# A2-D8

# Reasoning over networks with BioRevise and PowerGraphs

| | |
|---|---|
| Project title: | Reasoning on the Web with Rules and Semantics |
| Project acronym: | REWERSE |
| Project number: | IST-2004-506779 |
| Project instrument: | EU FP6 Network of Excellence (NoE) |
| Project thematic priority: | Priority 2: Information Society Technologies (IST) |
| Document type: | D (deliverable) |
| Nature of document: | R (report) |
| Dissemination level: | PU (public) |
| Document number: | IST506779/Dresden/A2-D8/D/PU/b |
| Responsible editors: | Gihan Dawelbait, Loic Royer |
| Reviewers: | Michael Schroeder |
| Contributing participants: | Dresden |
| Contributing workpackages: | A2 |
| Contractual date of deliverable: | 28 Febreuary 2008 |
| Actual submission date: | 28 Febreuary 2008 |

**Abstract**
The analysis of protein interactions is a fundamental application of bioinformatics. Previous deliverables discussed the BioCham pathway analysis engine developed in the Paris group, this deliverable introduces the BioRevise systems, which integrates a non-monotonic reasoning engine orginially developed in Lisbon with the application of reasoning over metabolic pathways. We discuss examples, modelling of the networks, reasoning, and the demonstrator. Complex networks such as the metabolic networks in BioRevise require sophisticated visualisation. We develop power graph analysis, which identifies modules in networks and visualises them in a compact fashion. The theoretical basis for power graphs are the identification of bicliques and cliques in graphs. Power graphs can be applied to any network including class dependencies and UML diagrams as discussed in groups I1 and I3. The deliverable introduces an algorithm and discusses a number of bioinformatics applications for power graphs.

**Keyword List**
Reasoning, revision, conflic resolution, metabolic network, protein interaction, visualisation

*Project co-funded by the European Commission and the Swiss Federal Office for Education and Science within the Sixth Framework Programme.*

# Reasoning over networks with BioRevise and PowerGraphs

**Bill Andreopoulos**[Dre], **Gihan Dawelbait**[Dre], **Vasco Pedro**[Dre], **Matthias Reimann**[Dre], **Loic Royer**[Dre], **Michael Schroeder**[Dre]

[Dre] Technische Universität Dresden, Germany

28 Febreuary 2008

**Abstract**

The analysis of protein interactions is a fundamental application of bioinformatics. Previous deliverables discussed the BioCham pathway analysis engine developed in the Paris group, this deliverable introduces the BioRevise systems, which integrates a non-monotonic reasoning engine orginially developed in Lisbon with the application of reasoning over metabolic pathways. We discuss examples, modelling of the networks, reasoning, and the demonstrator. Complex networks such as the metabolic networks in BioRevise require sophisticated visualisation. We develop power graph analysis, which identifies modules in networks and visualises them in a compact fashion. The theoretical basis for power graphs are the identification of bicliques and cliques in graphs. Power graphs can be applied to any network including class dependencies and UML diagrams as discussed in groups I1 and I3. The deliverable introduces an algorithm and discusses a number of bioinformatics applications for power graphs.

**Keyword List**

Reasoning, revision, conflic resolution, metabolic network, protein interaction, visualisation

# Contents

# Reasoning over networks: BioRevise

Biological pathways are complex networks that provide energy for the processes of life and synthesising new cellular material. Modelling the behaviour of these networks is a challenging yet a very important task. Here, we show a high level representation of inhibition in metabolic pathways, the model helps to identify reactions that are affected by metabolic disorders which are either genetic or acquired as a result of diet, toxins, or infections. In this work, we present a belief revision system -BioRevise- where the inhibition of enzyme-catalysed reactions is modelled using extended logic programing. The system provides possible explanations to justify the abnormal levels of observed metabolite concentrations. These explanations are lists of the enzymes that are affected and therefore might cause certain inhibition of reactions in the metabolic pathways.

KEGG (Kyoto Encyclopedia of Genes and Genomes) database is used as the knowledge base that contains the present state of knowledge of metabolic pathways. A user friendly and interactive visualisation is implemented to overcome the problem of complicated and condensed representation of the metabolic pathway maps.

**Results** We demonstrate the success of the BioRevise system with a metabolic disorder example. BioRevise successfully identified the inhibition of the enzyme glucose-6-phosphatase (EC:3.1.3.9) as responsible for the Glycogen storage disease type I, which according to literature is known to be the main reason for this disease.

# 1 Reasoning over networks: BioRevise

## 1.1 Introduction

Chemical reactions that take place in the cell tend to equilibrium, they are accelerated or catalysed by specialised enzymes. Enzymes are proteins that catalyse most reactions taking part in living organisms. They are considered as the main activators of different parts of the metabolic networks. Enzyme-catalysed reactions are usually connected in series, so that the product of one reaction becomes the substrate for the next. These connections of linear reactions constitutes pathways that are in turn linked to one another, forming a maze of interconnected reactions. These interconnected reactions enable the cell to survive, grow and reproduce, constituting metabolism [Alberts et al., 1998].

The inhibition of crucial enzymes can imply the interruption of a synthesis pathway and thus, to the lack of products or it can lead to the accumulation of an intermediate that is toxic when present in high concentration.

If an enzyme gets inhibited, affected metabolic pathways will lead to equilibrium loss and therefore the concentration of the metabolites changes.

The aim of this project is to utilise a belief revision system -REVISE- to model inhibition of reactions in metabolic pathways. The model describes in a high level the behaviour of the metabolic system when the concentration of the metabolites changes due to the inhibition of enzymes-catalysed reactions.

REVISE is built on top of the SLX proof-procedure for well founded semantics with explicit negation (WFSX). Since programs may be contradictory the paraconsistent version of WFSX is used. The top-down characterisation of WFSX relies on the construction of two types of AND trees (T and TU-trees), whose nodes are either assigned the status successful or failed. T-trees

compute whether a literal is true; TU-trees whether it is true or undefined. A successful (resp. failed) tree is one whose root is successful (resp. failed). If a literal $L$ has a successful T-tree rooted in it then it belongs to the paraconsistent well-founded model of the program (WFMp); otherwise, i.e. if all T-trees for $L$ are failed, $L$ does not belong to the WFMp. Accordingly, failure does not mean falsity, but simply failure to prove verity [Damásio et al., 1997].

Our work is motivated by the necessity to identify possible inhibited reactions caused by metabolic disorders which are either genetic, mainly inborn errors of metabolism, or acquired as a result of introduction of toxins into the system. These toxins usually contain substances that are designed to inhibit certain reactions or increase the concentration of some metabolites. However the main problem with this new substances, they might cause unforeseeable side effects by inhibiting other enzyme-catalysed reactions.

The KEGG PATHWAY database is a collection of manually drawn pathway maps representing the up to date knowledge on the molecular interaction and reaction networks for metabolism and other biological processes.

We extract the reactions topology knowledge from the KEGG database and use it as part of the background knowledge (background predicates) for our belief revision model, which given observed concentration levels of the metabolites will revise the metabolic pathways model to explain these observations.

The inhibition of enzyme-catalysed reactions in the metabolic pathways is modelled using extended logic programing (XLP), using the REVISE system and the background predicates are obtained from the KEGG database.

Modelling metabolic pathways is a challenging task. Besides the obvious complexity arising from the amount of data to be processed, metabolism exhibits some complex mechanisms such as negative feed back where the end product(s) of a pathway are often inhibitors of the committed step enzymes thus regulating the amount of end product made by the pathways.

There is a considerable effort to model metabolic pathways, In [Hurlebaus et al., 2002] the authors used differential equations in modelling the kinetics of the pathways where all real metabolites concentration are considered. Pathway logic is an implementation in [Eker et al., 2002] where they use a rule based model with algebraic syntax. Another rule based modelling effort is BIOCHAM which in addition implements temporal behaviour of the system, automated reasoning and an expressive query language [Chabrier-Rivier et al., 2004] and more others such as BioNet-Gen [L et al., 2004], Bio-ambients [Regev et al., 2004], Hybrid Petri Nets [Hofestadt and Thelen, 1998], and Hybrid Concurrent Constraint languages [Bockmayr and Courtois, 2002].

In our approach we use a high level representation of inhibition of enzyme-catalysed reactions, which reasons over the KEGG network to solve problems such as showing all reactions and pathways affected when one or more metabolites change their concentrations to abnormal levels.

To our knowledge it is the first time that Belief revision is used along with such a high level representation of inhibition to reason over metabolic networks. Our system does not depend on kinetic information which is not available for all known metabolic reactions. Our system can help scientist to limit their search space when working with this kind of complex networks, by suggesting a concise set of affected enzymes to be checked as a result of a metabolic disorder.

## Material and Methods

## 1.2 Definitions

### 1.2.1 Belief Revision

A belief revision occurs when a new piece of information that is *inconsistent* with the present belief system (or database) is added to that system in such a way that the result is a new consistent belief system. In a generic belief revision system the database is not viewed merely as a collection of logically independent facts, but rather as a collection of axioms from which other facts can be derived. It is the interaction between the updated facts and the derived facts that is the source of the problem.

### 1.2.2 Extended Logic Programming (XLP)

XLP extends logic programming by integrity constraints and two kinds of negation: *default negation* and *explicit negation*. Where default negation gives a way of expressing a kind of negation, based on a lack of knowledge about a fact, explicit negation, on the other hand, allows to explicitly assert the falsity of a fact.

### 1.2.3 Integrity Constraints (IC)

An integrity constraint has the form

$\bot \leftarrow L_1, ..., L_m, notL_{m+1}, ..., notL_n$ $(0 \leq m \leq n)$

where each $L_i$ is an objective literal $(0 \leq i \leq n)$, and $\bot$ stands for false. Syntactically, the only difference between the program rules and the IC is the head. A rule's head is an objective literal, whereas the constraint's head is $\bot$, the symbol for false. Semantically the difference is that program rules open the solution space, whereas IC limit it, as indicated in Figure 1 [Damásio et al., 1997]).

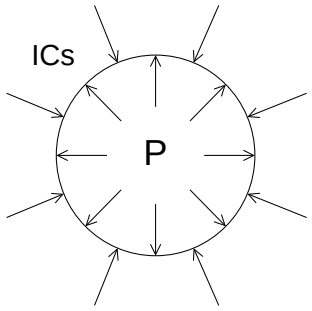


Figure 1: IC closing the solution space of the program $P$.

### 1.2.4 Metabolic Pathways

A metabolic pathway is a series of chemical reactions occurring within a cell, catalysed by enzymes, resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway (then called a *flux generating step*).

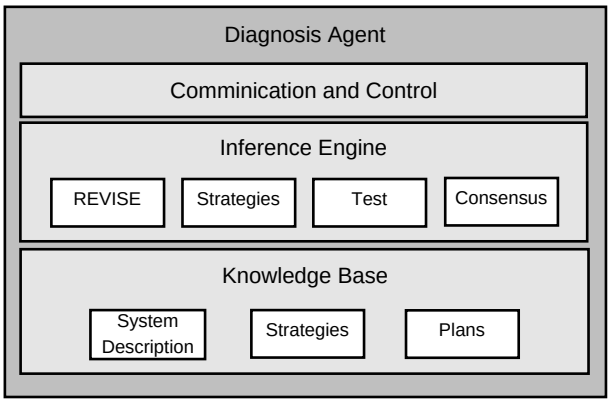


Figure 2: Components of REVISE.

Common properties of metabolic pathways includes reversibility of reactions, regulation of pathways using cycles or feedback inhibition.

## 1.3 Resources

### 1.3.1 The REVISE system

is a non-monotonic reasoning system that uses belief revision to revise extended logic programs. It is based on logic programming with explicit negation and IC. It provides two-valued revision assumptions to remove contradictions from the knowledge base. It has been tested on a spate of examples with emphasis on model-based diagnosis.

The system is embedded into an architecture for a diagnosis agent consisting of tree layers: a knowledge base, an interface layer, and on top a component for communication and control as shown in Figure 2 . The core of the inference machine is the REVISE system, which removes contradictions from extended logic programs with integrity constraints. Additionally, there is a strategy component to employ diagnosis strategies and compute diagnosis in a process. The test component solves conflicts among agents in a multi agent setting. This modular structure is general and allows to solve a variety of problems; depending on the requirements of the application, a suitable configuration can be set up. REVISE is described in detail in [Damásio et al., 1997].

### 1.3.2 KEGG

is a database resource for understanding higher-order functions and utilities of the biological system, such as the cell or the organism, from genomic and molecular information [Kanehisa et al., 2006]. The KEGG pathway database contains a collection of pathway maps each corresponding to a known network of functional significance [Kanehisa et al., 2004]. We extract all enzyme-catalysed reactions that are organised as connected networks. This logical representation of the network is used as the background predicates which are needed to perform the belief revision.

The metabolic pathway maps are build in three different levels of abstraction. The first level is the map representation of all the metabolism network, the second is the collection of

maps within the different groups of metabolism, and the third level contains the maps of the metabolic pathways, which are the most detailed maps also representing the pathway reactions and their components 3.

### 1.3.3 Model View Controller (MVC)

The MVC paradigm has been pursued for a clear design which separates different responsibilities within an interactive application. A *model* in this paradigm is a class which originates in a specific domain. It is an abstraction of a domain specific entity and has no knowledge about the graphic user interface (GUI). The representation of the *model* as GUI element is called *view*. A *view* can be seen as a wrapper around the *model*, which is capable of displaying a subset of the data that is encapsulated in the *model*. Each view has an associated *controller*. A *controller* is responsible for all possible actions that are defined in the *view* concerning the associated *model*. A *model* can have multiple *views* [Veit and Herrmann, 2003]. The *model* does not only captures the state of the system, but also how the system works.

The MVC separates the user interface from the core application data and functionality. With this separation one of the components can change without requiring changes in the others components. This is an alternative to the traditional (input, processing, output) applications. In the MVC perspective the keyboard and mouse inputs are handled by the *controller* that makes the proper connection with the other two components, the *view* and the *model*. The *model* is able to change state, answer about its state and manage the needed data structures. Presenting the data related to the state or changes in the *model* is the role of the *view* component in the MVC.

The *model* is built without any knowledge about *views* and *controllers*. The model must implement the functionality defined in the role "Subject" and introduce notifications in all places where the state of the *model* is changed. The update message is sent to every attached *view* or *controller* [Veit and Herrmann, 2003]. But the *model* itself is more than only data and functions, it is meant to serve as a computational approximation or abstraction of some real world process or system.

*View* and *controller* use an explicit association to the *model* to query its state. A *view* can define actions and can send events, if a specific action has happened. The *controller* is able to listen to specific events of the associated *view* and has to handle all those events, which have an impact on the associated *model*. A *controller* "translates" from specific GUI events to application logic [Veit and Herrmann, 2003].

## 1.4 BioRevise work flow

The starting point of the BioRevise system is the extraction of the background predicates from the KEGG database. The background knowledge consists of the background predicates and the metabolites concentrations observed (observable predicates). It is used by the *model* to write the system input files. These observations are marked in the maps of the metabolic pathways shown by the *view*.

BioRevise takes as input the extracted background knowledge and then uses it to revise the inhibition model, and as output it produces lists of possible inhibited enzyme-catalysed reactions that explain the abnormal levels of the observed metabolic concentration. The *view* shows the corresponding enzyme-catalysed reactions in the maps using the extracted coordinates of the enzymes. Through the *controller* the user is also able to *zoom in* from the maps of the different

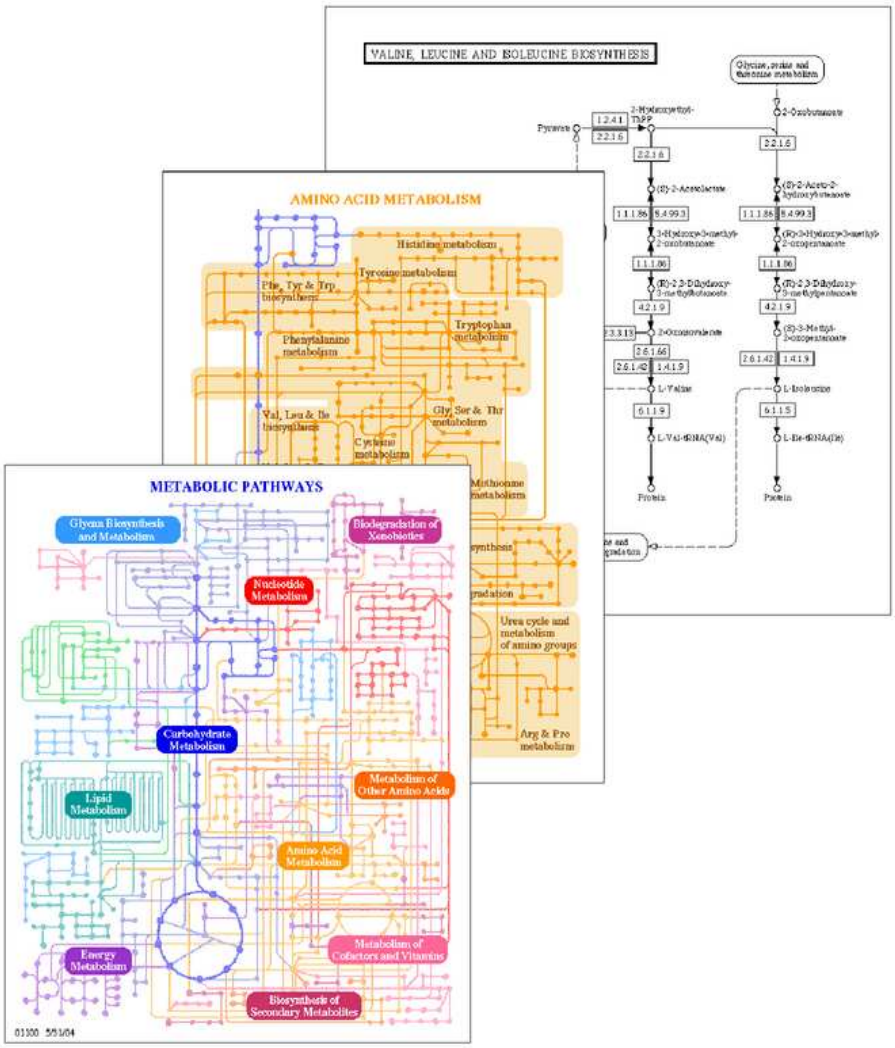


Figure 3: The three abstraction levels of the KEGG maps.

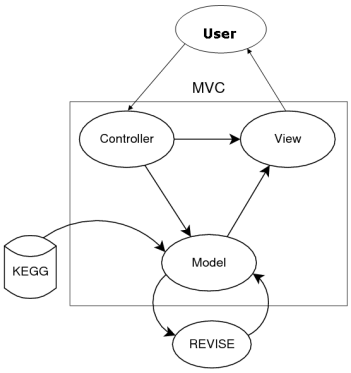


Figure 4: BioRevise system workflow. The connections between the main components of the BioRevise system.

metabolic families to the metabolic pathways maps with the inhibited reaction. Figure 4 shows the connection between the different components of the system.

The background knowledge, program (definitional knowledge), and IC representing the metabolism are modelled using XLP while the visualisation is implemented using Java.

### 1.4.1 REVISE

Is the core of the BioRevise system. It is used to perform belief revision over the knowledge representation of the metabolic pathways, it is integrated in the *model* component of the MVC. The REVISE uses three input files, the first file consisting of the observables (metabolite concentration levels) provided by the user, the second file contains the network topology rules and the third file contains the knowledge modulation describing the inhibition behaviour and the IC.

### 1.4.2 Knowledge Modulation

The knowledge modulation is the computational representation of the metabolism. By capturing the knowledge of the components of the metabolism as well as their connections and the behaviour of the metabolism system, some inferences about changes in the state of the BioRevise system can be made based on the new states due to abnormal observations obtained by the system. In this work we concentrate in modelling of inhibition. The modulation presented here is based on a previous work developed by [Tamaddoni-Nezhad et al., 2004].

The model of the metabolism can separate two disjoint sets of predicates: the *observable* predicates and the *abducible* predicates. The model can be *incomplete* in its description. To complete the description, the new information given by the observation can be used. The basic assumption is that all the incompleteness of the model can be isolated in its abducible predicates. The *background* predicates are auxiliary relations linking observable and abducible information. For this purpose a logic program is required which models how the concentration of metabolites is related to inhibition of enzymes [Tamaddoni-Nezhad et al., 2004].

To represent the changes on the metabolites concentration, the *observable* predicate $obs/2$ is used:

$obs(Metabolite, Concentration)$

This predicate encodes the observations made by the user, where variable *Concentration* can be either *up* or *down*.

The relational representation of metabolic networks that form the metabolism is represented by the background predicates $reaction/4$:

$reaction(ID, Sub, Enz, Prod).$

representing the fact that the enzyme-catalysed reaction occurs in a direct path from one node to the other, where *Sub* is a set of substrates and Prod is the product they produce. The product of one reaction becomes the substrate of another. For example the predicate:

$reaction(1, [' Acetate'],' 6.2.1.1',' Acetyl - CoA').$

represents the enzyme-catalysed reaction between *Acetate* and *Acetyl-CoA* catalysed by the enzyme *Acetate-CoA ligase* (EC:6.2.1.1) in the "Glycolysis/Gluconeogenesis" pathway.

For handling products, which can be produced by several reactions we introduce a new product representation as follows:

$reaction(ID1, Sub1, e1, (Prod, 1)).$
$reaction(ID2, Sub2, e2, (Prod, 2)).$
where ID1 and ID2 are the identifiers used for the two reactions producing the same product.

The enzyme-catalysed reactions from the metabolic pathways can be reversible, in order to compensate the variation of concentration in one of the metabolites.

The incompleteness of the model resides in the lack of knowledge of which metabolic reactions are adversely affected in the event of a metabolic disorder [Tamaddoni-Nezhad et al., 2004].

To predict the inhibition of one reaction and complete the model, the *abducible* predicate is used:

$inhibited(Enz, Prod, Sub)$

encoding the fact that the enzyme-catalysed reaction producing the product *Prod* from the substrate *Sub* is inhibited because of the inhibition of the enzyme *Enz*, due to a metabolic disorder of the system. For example:

$inhibited('6.2.1.1',' Acetyl - CoA',' Acetate')$

which captures the hypothesis that the reaction from *Acetyl-CoA* to *Acetate* that is normally catalysed by the enzyme 3.5.3.1. is inhibited due to a certain defect in this enzyme.

For describing the behaviour of the system due to inhibition of a certain reaction, the following rules are used:

$concentration(ID, Prod, down) \leftarrow$
$reaction(ID, Sub, Enz, Prod)$,
$inhibited(Enz)$.

$concentration(ID, Prod, V) \leftarrow$
$reaction(ID, Sub, Enz, Prod)$,
$not\ inhibited(Enz, Prod, Sub))$,
$andAll(Sub, V)$.

$concentration(ID, Prod, V) \leftarrow$
$multipleReactions(ID, Reactions, Prod)$,
$orAll(Reactions, Prod, V)$.

$andAll([], up)$.
$andAll([Substrate|Substrates], V) \leftarrow$
$andAll(Substrates, V1)$,
$val(\_, Substrate, V2)$,
$andTwo(V1, V2, V)$.

$orAll([], \_, down)$.
$orAll([Reaction|Reactions], Product, V) \leftarrow$
$orAll(Reactions, Product, V1)$,
$val(Reaction, (Product, Reaction), V2)$,
$orTwo(V1, V2, V)$.

orTwo(up,up,up).
orTwo(down,up,up).
orTwo(up,down,up).
orTwo(down,down,down).

andTwo(up,up,up).
andTwo(down,up,down).
andTwo(up,down,down).
andTwo(down,down,down).

$input(Metabolite) \leftarrow$
$reaction(ID, Substrates, Enzyme, Product)$,
$member(Metabolite, Substrates)$,
$notreaction(\_, \_, \_, Metabolite)$.

$obs(Metabolite, up) \leftarrow$
$input(Metabolite)$.

These rules describe the model in a simplified way, considering the fact that some metabolic disorders can change the concentration of metabolites by the inhibition of enzymes catalysing enzyme-catalysed reactions.

The above mentioned rules do not consider the possibility of the reversed enzyme-catalysed

reactions that can occur to compensate the changes in the metabolites concentrations.

To represent the fact that if the reaction that produces a product *Prod* is inhibited, this will cause a decrease in the concentration of the product *Prod*, the first rule is used.

The second rule represents the changes on the concentration caused through indirect effects, where a metabolite *Prod* can have *down*/*up* concentration due to the fact that some other substrate metabolite *Sub*, that produces *Prod* was caused to have decrease/increase concentration respectively (even when the reaction is not currently inhibited)

The third rule is used when a product is produced as a result of different reactions.

The IC of the model captures several *validity requirements* that must be satisfied by the abducible information of *inhibited*/3. In our model to express the fact that the concentration of a metabolite can not be *up* and *down* at the same time the following IC are used:

$\leftarrow obs(Metabolite, up), concentration(\_, Metabolite, down).$
$\leftarrow obs(Metabolite, down), concentration(\_, Metabolite, up).$

## 1.5 MVC

### 1.5.1 Model

The data structures populated by the KEGG extractor with the knowledge base of the BioRevise system are used by the *model* to write the REVISE observables input file. The file contains the observations marked by the user corresponding to the new concentration of metabolites. The concentration observed are either *up* or *down*. The REVISE system is then used to read the other input files and generate the output file containing the possible explanations. These explanations are the possible inhibited enzyme-catalysed reactions caused by the modification in the concentration of the metabolites. Using these results the *model* updates the inhibition state of the enzymes in the data structures from *not inhibited* to *inhibited*, in order to also display it in the *view*.

### 1.5.2 Visualisation

Due to the standard complex representation of metabolic networks (i.e Boehringer map), finding enzymes that are affected as a side effect of an inhibition of certain reaction is like searching for a needle in a hey sack.

To perform the visualisation of the metabolic pathway networks and to simplify finding the enzyme catalysed-reactions, the maps from KEGG are used.

The first picture of the metabolism is composed of the maps that represent the different groups of metabolism, biosynthesis and biodegradation in KEGG. The picture generated is outlined by putting together the maps according to the coordinates from the *xml* files. The coordinates of the *xml* files were manually built to present all maps in the same picture.

By clicking on any of the metabolic pathways in the first picture, it is possible to see the most detailed maps from KEGG. This maps contain the representation of the enzyme-catalysed reactions that constitute the metabolism. In both windows the user can perform the operations of *zooming in* and *out*, as well as *translate* the picture by moving it up or down. It is also possible to mark the observations made by the user in this second window. This corresponds to the change of the metabolites concentration to *up* and *down*. The concentration of the metabolites is marked using different colours, green for down and red for up as illustrated in Figure 5.

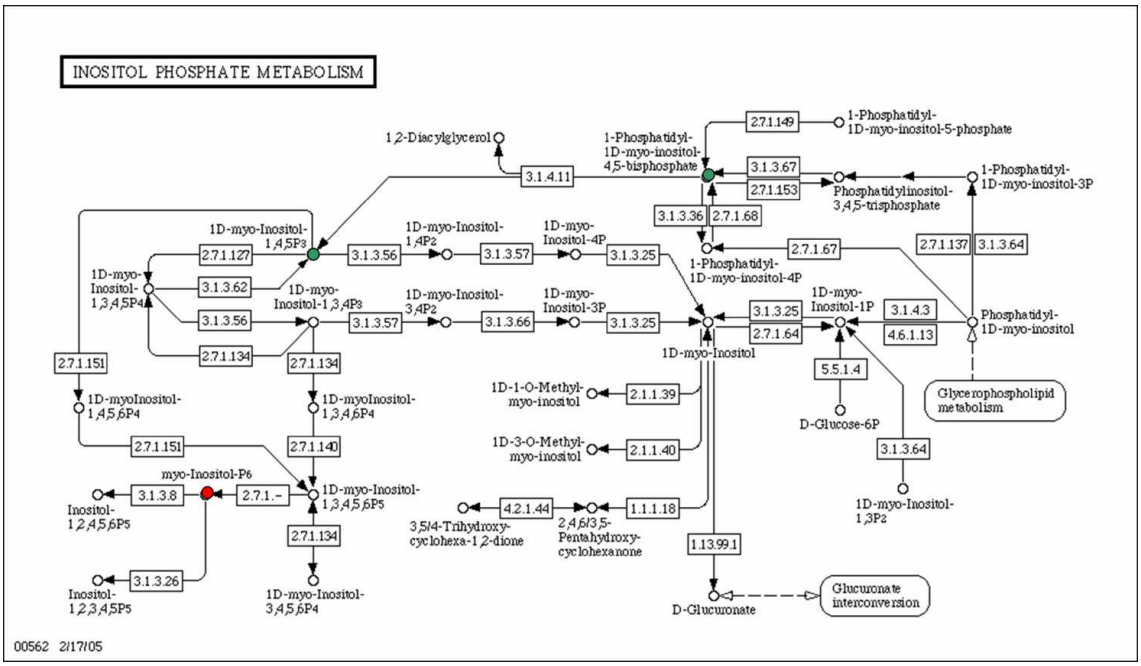


Figure 5: The concentration of the metabolites is marked using different colours, green for down and red for up.

For the visualisation of the inhibited reactions, the affected reactions are marked in the first window in the region that corresponds to a metabolic pathway, shown in Figure 6. At this stage, the user is able to see all the inhibited reactions in the metabolism. The user can then visualise the maps of the metabolic pathways in the second window to see each of the enzyme-catalysed reactions which are inhibited. To show these inhibitions on the second window, the enzymes that catalyse the reaction are highlighted with red, blue, green or pink rectangles, corresponding to the first, second, third or fourth solution respectively. Figure 7.

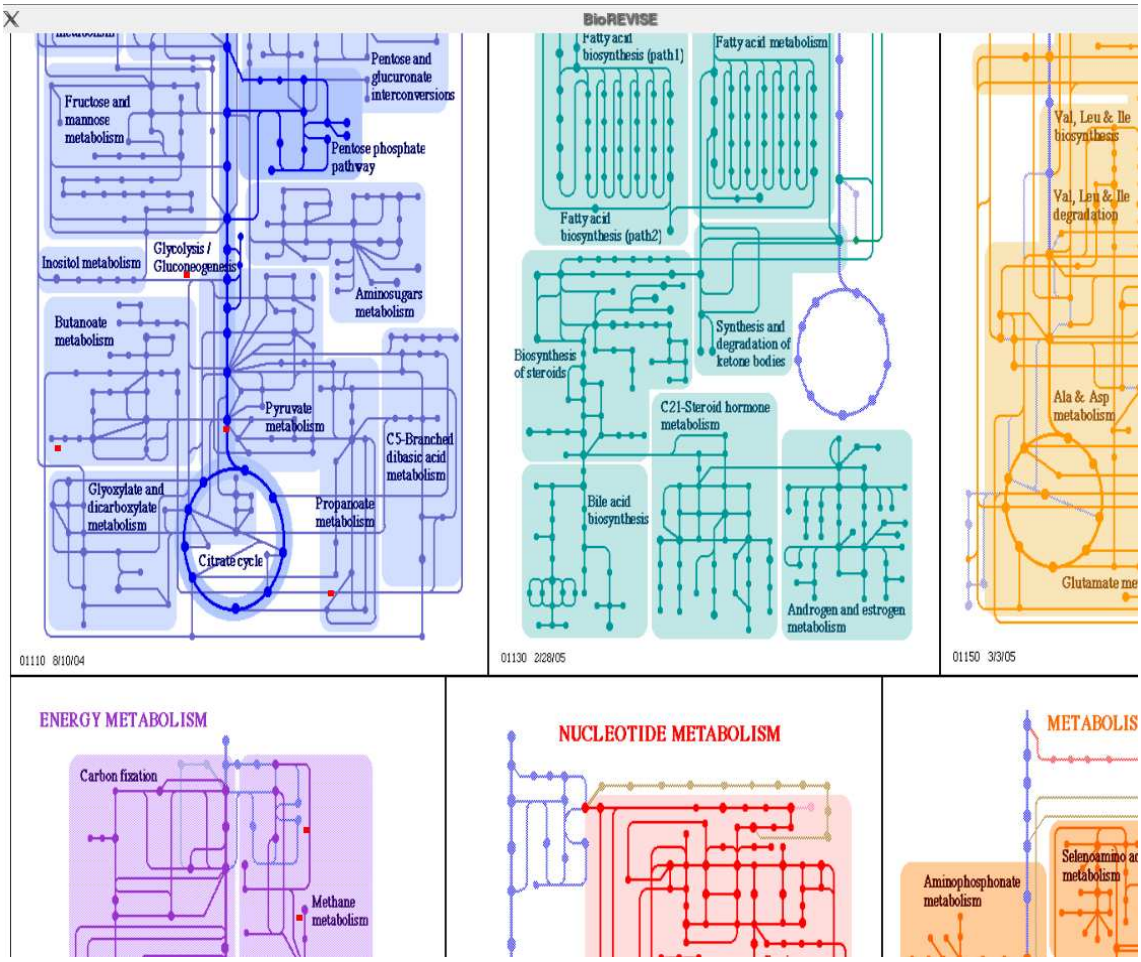


Figure 6: The visualisation of the inhibited reactions, the affected reactions are marked in the first window in the region that corresponds to a metabolic pathway

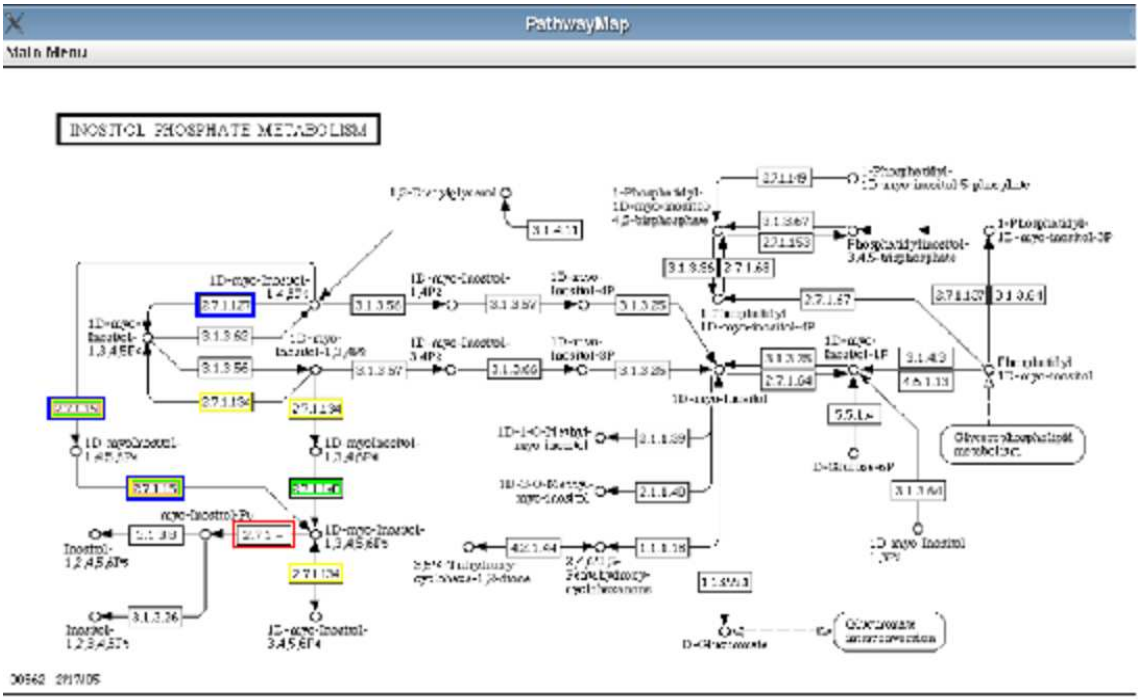


Figure 7: Inhibited reactions with the enzymes that catalyse the reaction are highlighted with a red, blue, green or pink rectangles, corresponding to first, second, third or fourth solution respectively at the second window.

### 1.5.3 Controller

The *controller* performs the connection between the *model* and the *view*, by reacting to the mouse and keyboard events. For each event the *controller* performs the corresponding data update in the data structures, used by the *model* or the *view*. Other data related to the *zoom* and *translation* operations made in the maps, is also updated.

## 1.6 Metabolic disorder example: Glycogen storage disease

Glycogen storage disease type I is a metabolic disorder that is caused by the deficiency in the glucose-6-phosphatase enzyme. This deficiency impairs the ability of the liver to produce free glucose from glycogen and from gluconeogenesis. Since these are the two principal metabolic mechanisms by which the liver supplies glucose to the rest of the body during periods of fasting, it causes severe hypoglycemia. Reduced glycogen breakdown results in increased glycogen storage in liver and kidneys, causing enlargement of both. [home page., ]

An obvious symptom of the GSD type I is the inability to maintain adequate blood glucose levels during fasting which results from the combined impairment of both glycogenolysis and gluconeogenesis.

According to the BioRvise model, there are many possible hypothesis that can explain the up- or down-regulation of the observed metabolites. However the system provides only the most comprehensive and short explanations.

We used the observed values of the metabolites caused by the GSD (the glucose down, pyruvate and lactate up) as input for the BioRevise system.

Glucose production and breakdown is associated with metabolic pathways that are crucial to human metabolism (i.e Glycolysis / Gluconeogenesis, Galactose metabolism, Starch and sucrose metabolism).

BioRevise could identify that the inhibition of the enzyme glucose-6-phosphatase (3.1.3.9) as a possible explanation for the abnormal levels of concentration of Gloucose, which according to literature is known to be the main reason for this disease.

## 1.7 Results and Discussion

We provide a system that uses belief revision to model reaction inhibition in metabolic pathways. Given abnormal concentration levels of metabolites, the system will reasons over the KEGG network to show all reactions and pathways affected due to the metabolic disorder that caused the metabolites levels to behave abnormaly. The system used a high level representation of inhibition, which make it independent of detailed kinetic information of metabolism modelling.

# Unraveling protein interaction networks with power graphs

We propose a novel network analysis paradigm – power graphs – in which modules and motifs are explicit in a compact representation without loss of information. We apply power graphs to high-throughput interaction networks and show that up to 85% – and on average 56% – of the topological information is redundant. The advantages of power graph analysis are demonstrated with four examples. We show how the quaternary structure of vacuolar type $H^+$ ATPase and casein kinase II complex is outlined by power graphs; how power graphs hint at the correlation between interaction profiles and sequence phylogeny of SH3 domains, and how false positive interactions in the nucleosome can be spotted.

Studying power graphs of synthetic scale-free networks, we find the general law $r = d^{-\frac{2}{3}}$, where $r$ is the edge reduction in power graphs and $d$ is the average degree of a network. Moreover, experimental networks are more compressible than synthetic ones, indicating that experimental networks are rich in biologically relevant structures. This leads to the proposal of a new measure for the deviation of a network from the preferential attachment model: the power graph index.

# 2 Unraveling protein interaction networks with power graphs

## 2.1 Introduction

In recent years, novel high-throughput methods such as yeast two-hybrid assays [Fields and Song, 1989] and affinity purification techniques [Rigaut et al., 1999, Mann et al., 2001] have been used to characterize protein interactions at a large scale and have produced a wealth of data in the form of networks of interacting proteins. Comprehensive protein interaction networks have been assembled for several species: S. cerevisiae [Gavin et al., 2006, Ito et al., 2001, Krogan et al., 2006], C. elegans [Li et al., 2004], D. melanogaster [Giot et al., 2003, Stanyon et al., 2004], H. pylori [Rain et al., 2001], H. sapiens [Stelzl et al., 2005, Rual et al., 2005], and P. falciparum [LaCount et al., 2005]. The challenge remains to discover biological insights through the analysis of these networks. The topology of the networks can be explored through clustering the proteins into groups that share the same biological function, are similarly localized in the cell, or are part of a complex. To this end several algorithms have been developed, such as socio-affinity clustering [Gavin et al., 2006], Restricted Neighborhood Search Clustering (RNSC) algorithm [King et al., 2004], the MCODE algorithm [Bader and Hogue, 2003], statistical sub-complexes [Hollunder et al., 2005], or the MULIC clustering algorithm [Andreopoulos et al., 2007].

How does the underlying biology manifest itself in the networks? Fig. 8 illustrates three recurrent motifs that have been reported in the literature. The first motif is the star, which often arises in protein interaction networks because of the scale-free statistics induced by the existence of a highly connected hub proteins [Li et al., 2006a]. These hub proteins have their roots in evolutionary models based on gene duplication and divergence [Taylor and Raes, 2004] and because of preferential attachment [Barabasi and Albert, 1999]. The second motif is the biclique, also referred to as complete bipartite graph: all proteins in one group interacting

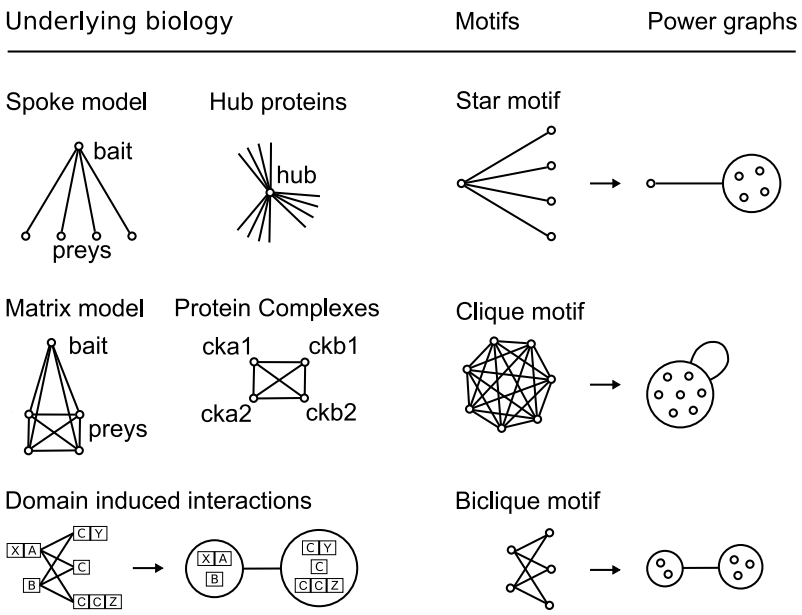


Figure 8: The three basic motifs: star, biclique, clique. Stars often occur because of hub proteins and when purification complexes are interpreted using the spoke model. Bicliques often arise because of domain-domain or domain-motif interactions inducing protein interactions [Morrison et al., 2006]. Cliques are really a special case of reflexive biclique which are often observed at the core of complexes, and as a result of the matrix interpretation of purified complexes. Similarly, stars are also a special case of biclique in which one node is connected to several other nodes.

with all proteins in another group. Domain interactions have been reported to induce the occurrence of bicliques. 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 interactions and protein interactions have been shown in many studies to be sufficiently correlated, to allow domain bindings to be used to predict protein interactions, and conversely, protein interactions 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]. The third motif is the clique also referred as complete graph: a set of completely interconnected proteins. In the core of molecular complexes, where the frontier between direct and indirect physical interactions is often blurred, protein interactions are observed to organize into cliques and bicliques. Indeed, the completion of quasi-cliques and quasi-bicliques has been shown to successfully predict missing interactions between proteins [Bu et al., 2003].

The abundance of stars, cliques, and bicliques suggests that modeling protein interaction networks as a collection of binary interactions is an inappropriate level of representation and an obstacle toward a detailed analysis of the wealth of information contained in high-throughput networks. These networks have many edges and many edge crossings that redundantly diffuse the information instead of highlighting it. In this study we propose a new visual language that not only groups proteins into biologically relevant modules, but also conveys in all detail – without loss of information – and with few edges, the subtle connection patterns within and between clusters.

## 2.2 Results and Discussion

## 2.3 A new visual language

Here we introduce the *power graph* language. It is a visual conceptualization of networks that relies on two novel symbols: *power nodes* and *power edges*. These language primitives allow for the succinct representation of stars, bicliques and cliques. *Power nodes* are a set of nodes brought together within circles, these circles are interpreted as Euler diagrams [Euler, 1772]. Euler diagrams use discs to represent sets, with the position and overlap of the discs indicating the relationships between the sets. *Power edges* are represented as segments connecting power nodes.

As Fig. 8 shows, a star is expressed as a node connected via a power edge to a power node, a biclique is expressed as two power nodes connected by a power edge, and a clique is a power node connected to itself by a power edge. In Fig. 8, the power graph representation reduces the number of edges needed to represent the network. In the following we will often use the notion of edge reduction that is the proportion of edges that are removed from the original network in the power graph representation. Many aspects of network analysis such as node clustering, module detection, network visualization, and network models can be recast in terms of power graph analysis. In the following we demonstrate how power graphs facilitates the task of uncovering the underlying biology.

## 2.4 Understanding interactions within molecular complexes with Power Graphs

Some recent large-scale experiments [Gavin et al., 2006] specifically aim at identifying complexes and not binary interactions. Complexes are difficult to interpret from the point of view of binary interactions: are two proteins $p_1$ and $p_2$ participating in a complex $C$, but not in direct physical contact, interacting?

This point is crucial for the interpretation of results from pull-down assays where whole complexes are identified rather than binary interactions [Rigaut et al., 1999,Mann et al., 2001]. In a pull-down assay, a purified and tagged protein, the bait, is used to capture other proteins: the preys. These observed complexes are either modeled as cliques in the *matrix* model, or as stars in the *spoke* model [Bader and Hogue, 2002]. In the case of the spoke model the bait is at the center of the star, and the preys are linked to it. But in the matrix model, all proteins are linked together, signifying that they belong to the same observed complex.

The problem with this perspective is that the spoke model underestimates, and the matrix model overestimates the number of true physical interactions between the members of a complex, and for both models the use of binary interactions cannot convey succinctly an otherwise simple connection pattern. Let $n$ be the number of proteins in the complex. The matrix model represents the complex with a quadratic number of interacting pairs: $\frac{n(n-1)}{2}$. The spoke model requires only $n-1$ interacting pairs to represent the same complex. Fig. 8 shows that the power graph representation mitigates this problem: in both cases only one power edge is needed. In the case of the matrix model all proteins are brought together in one power node, whereas in the spoke model the bait protein is on its own and all preys are together in a power node. Let us consider three examples.

### 2.4.1 Vacuolar type $H^+$ ATPase

In [Gunsalus et al., 2005] the authors discuss predictive models for the determination of the molecular machines involved in C. elegans embryogenesis. To evaluate our method we computed a power graph representation of the protein interactions therein discussed. Fig. 9 shows one group of proteins that form a regular structure and all proteins involved in this sub network are subunits of Vacuolar type $H^+$ ATPase. Moreover, the organization of the subunits in the power graph are reminiscent of the known quaternary structure of the complex.

In particular, the subunits B, E and F which are all in the cytoplasmic V1 subunit of the complex form a power node. The reason for this is that these three proteins have similar interaction partners, namely G, D, d, a, and c" subunits. This result is remarkable because B, E and F are only connected by three interactions, whereas they participate in 16 interactions with their neighbors. Connectivity or density based clustering methods are not able to convey the subtle pattern of interconnections which is conveyed by the power graph representation.

### 2.4.2 Casein Kinase II Complex

A recent survey of the yeast proteome investigated the modularity of the yeast cell machinery [Gavin et al., 2006]. Fig. 10 shows the Casein Kinase II Complex and its neighboring complexes. 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

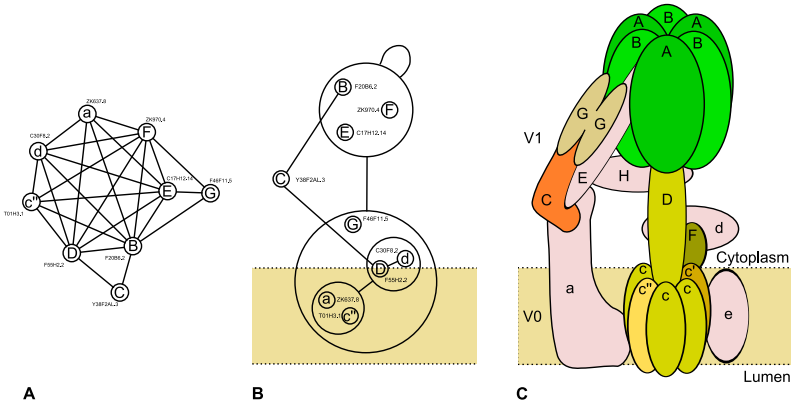


Figure 9: Vacuolar type $H^+$ ATPase. (A) Sub-network of interactions between 9 of the $H^+$ ATPase subunits. (B) Power graph representation of these subunits' interactions. (C) Known quaternary structure of the vacuolar $H^+$ ATPase complex [Beyenbach and Wieczorek, 2006]

.

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 being 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, and through the intermediary protein UTP22 to a power node consisting of proteins ROK1, RRP7 and YLR003C that do not correspond to a known complex but that are all related to RNA processing, possibly a small complex. In contrast, the alpha subunits do not interact with these two groups, but instead with YKL088W an uncharacterized enzyme.

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 Spt16p/Pob3p, a complex involved in the 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].

Overall we see that the power graph representation manages to 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 on the basis of 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, which is usually not the case of most clustering methods.

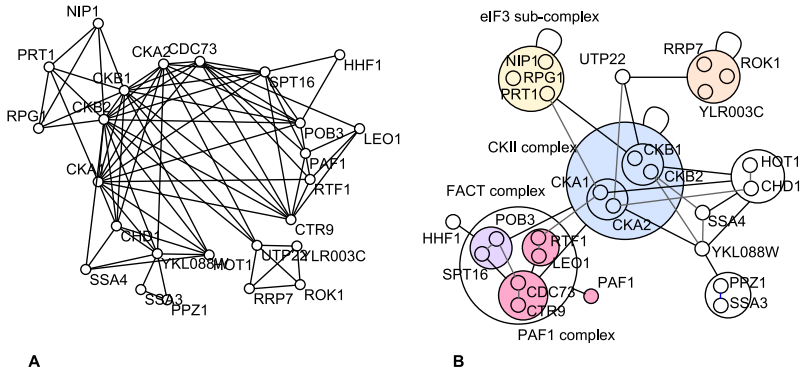


Figure 10: Casein Kinase II Complex. Two catalytic alpha subunits (CKA1, CKA2) and two regulatory beta subunits (CKB1, CKB2) interacting with the FACT complex, with sub-complex NIP1-RPG-PRT1, and with the PAF1 complex. The graph representation (A) consists of 80 edges whereas the power graph representation (B) has 30 power edges, thus an edge reduction of 62%

## 2.5 Understanding domain-mediated protein interactions with power graphs

### 2.5.1 Interaction profiles of motif binding domains

In [Landgraf et al., 2004] Landgraf et al. have used a combination of phage display and SPOT synthesis to discover peptides in the yeast proteome that have the potential to bind to eight SH3 domains. Fig. 11A 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 dramatic reduction in complexity by reducing the number of edges necessary for the representation by 80%. 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 motifs of class 1 (RxxPxxP) or in motifs of class 2 (PxxPxR) [Landgraf et al., 2004].

### 2.5.2 Domain-interaction profiles correlates to sequence similarity

Based on the previous results we wanted to investigate how the interaction profiles of these eight SH3 carrying proteins relate to the domain sequences. Fig. 11B shows a strong correlation between the phylogenetic tree of the SH3 domains sequences and the neighborhood similarity tree of interaction partners. The neighborhood similarity tree is computed using the proportion of common interaction partners as a similarity measure between two proteins (c.f. neighborhood similarity in methods).

The pair of SH3-carrying proteins Yhr016c/Yfr024 that are grouped in one power node in Fig. 11A are also close in the neighborhood similarity tree. Note how 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

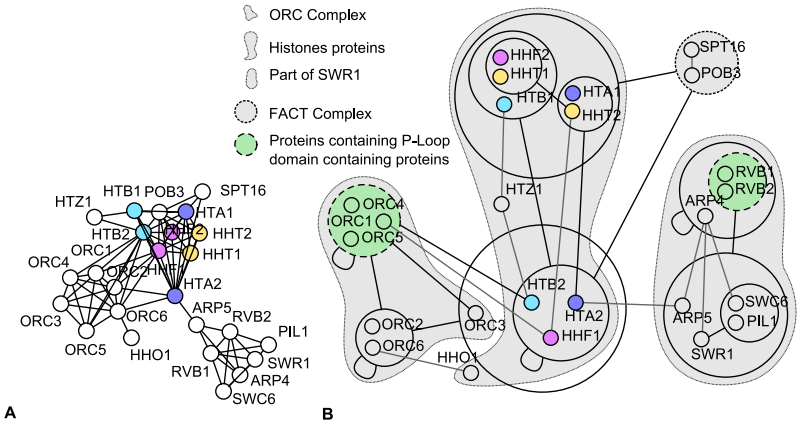


Figure 11: Interactions of SH3 carrying proteins. (A) Protein interaction network showing the 105 interaction partners of the SH3 domain carrying proteins: Sho1, Abp1, Myo5, Boi1, Boi2, Rvs167, Yhr016c and Yfr024. The underlying network consist of 182 interactions represented here as 36 power edges, thus a reduction of 80%. Class 1 motif (RxxPxxP) proteins are shown in green. Class 2 motif (PxxPxR) proteins are shown in pink [Landgraf et al., 2004]. (B) Phylogeny and interaction profiles. Comparison of the phylogenetic tree of the SH3 domains sequences with the neighborhood similarity tree of interaction partners.

in the two trees - RVS167 and Yhr016c/Yfr024 have similar interaction partners but dissimilar sequences.

### 2.5.3 Untangling the nucleosome

Similarly to the survey of the yeast proteome by Gavin et al., Krogan et al. [Krogan et al., 2006] have used Tandem affinity purification (TAP) to identify 7123 interactions between 2708 yeast proteins. Fig. 12 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. 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 inducing a complete bipartite subgraph (biclique) of interactions between proteins expressing duplicated genes. In yeast, HTA1 and HTA2 and 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 these histones we have again the FACT complex which is known to destabilize the interactions between the H3/H4 tetramer and the H2A/H2B dimer. It is clear from the power graph representation that this complex interacts with all histones, except for histone H1 and for histone HTAZ, a variant of the H2A histone. Also interacting with histones is the ORC Complex (Origin Recognition Complex) responsible for marking origin regions prior to DNA replication. This 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

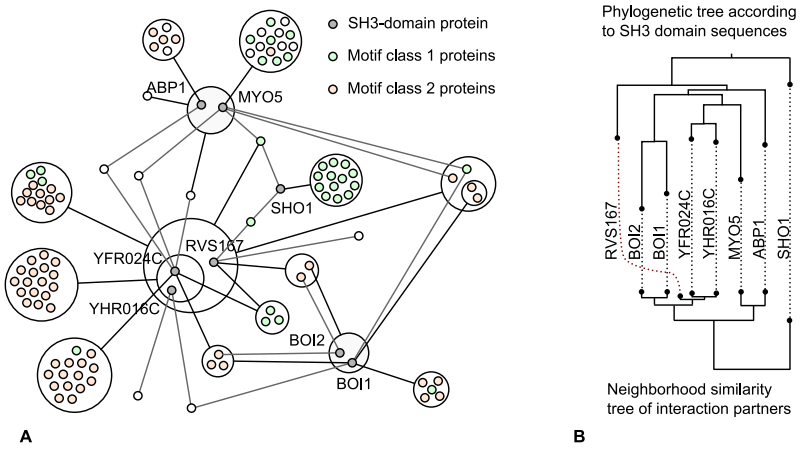


Figure 12: Histone protein interactions and neighboring proteins according to Krogan et al. [Krogan et al., 2006]. (A) Standard graph representation. (B) Power graph representation. The ORC complex is visible with a power node of proteins – ORC1/ORC4/ORC5 – carrying a nucleotide binding P-loop domain [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.

complex.

Surprisingly, histones HTA2, HTB2 and HHF1 are segregated from their twin subtypes HTA1, HTB1 and HHF2 in the power graph representation, as subunits ORC2 and ORC6 interact with HTA2, HTB2 and HHF1 and not with the HTA1, HTB1, and HHF2. This is contradictory to the identity/near identity of these pairs of histones.

In the case of H2A histones, each subtype has been shown to be sufficient for cell viability, and no clear functional difference were reported apart from homozygous strains for $hta1^{-}$ 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.

Yet, this hypothesis cannot 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 www.yeastgenome.org. Moran et. al. [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. Thus a possible explanation for not observing interactions between ORC2/ORC6 and HTA1/HTB1 is that under some circumstances – that might be triggered by the TAP methodology – the production of subtypes HTA1 and HTB1 is depressed. Moran et. al. argue that the same regulation feed-back takes place for HTA1 and all variants of HHT and HHF.

## 2.6 Power Graphs reveal hidden structures in networks

Power graphs can be used to analyse the topology of biological networks and compare it to theoretical models. We have conducted experiments in order to understand the behavior of the edge reduction for two important classes of networks: synthetic random networks generated according to the Erdös-Rényi model [Erdös and Rényi, 1960] and synthetic scale-free networks generated according to the preferential-attachment model of Barabási and Albert [Barabasi and Albert, 1999]. These models are compared with 13 protein interaction datasets as well as with networks of the IntAct database [Hermjakob et al., 2004]. The basis of the following analysis is the edge reduction which can be seen as a negentropic measure on networks. The more ordered a network, the less power edges are needed to represent it, and thus the higher the edge reduction will be. Algorithms that extract power graphs from networks perform a de facto information theoretic compression of the graphs.

### 2.6.1 Edge reduction is size invariant for scale-free networks

Fig. 13A shows that the edge reduction of synthetic scale-free networks generated using the preferential attachment model does not correlate with the number of nodes in the network, confirming indeed that such networks are scale-free. We observed that most of the edge reduction in these networks arises from star motifs, moreover the sizes of these motifs follow a power-law distribution, producing an approximately self-similar nesting and branching of these stars [Song et al., 2005, Goh et al., 2006]. In comparison, the edge reduction of random networks (generated via the Erdös-Rényi random network model) diminishes as the size of the networks increases.

### 2.6.2 Edge reduction is inversely correlated with average degree

A strong correlation exists between the average degree of the networks and the edge reduction: the edge reduction tends to decrease as the average degree increases. Fig. 13B shows a scatter plot of the log edge reduction versus the log average degree. For a given average degree the edge reduction is higher for synthetic scale-free networks than for random networks. More precisely, regression analysis shows that the edge reduction of both random and scale-free networks follows a power law of the form: $r = \alpha d^{-\beta}$ where $r$ is the edge reduction and $d$ is the average degree of the networks. Both have similar exponents of $\beta = 0.66 \pm 0.01$ for scale-free networks and $\beta = 0.49 \pm 0.006$ for random networks. On the other hand, the parameters $\alpha$ are significantly different with $\alpha = 1.0 \pm 0.01$ for scale-free networks and $\alpha = 0.52 \pm 0.003$ for random networks.

### 2.6.3 Power graph index discriminates network types

The previous analysis suggests that scale-free networks generated according to the preferential attachment model follow a power law: $r = d^{-\frac{2}{3}}$. Surprisingly, this exponent shows no correlation with the scale-free exponent or with the clustering coefficient [Watts and Strogatz, 1998], implying that the exponent $\frac{2}{3}$ is a signature specific to the preferential attachment model, and possibly of scale free networks in general. Erdös-Rényi random networks also fit a simple power law: $r = \frac{1}{2} d^{-\frac{1}{2}}$, although in this case the dispersion is higher.

This result suggests that an interesting discriminative measure on networks can be defined from the edge reduction and the average degree. We define the power graph index as $I =$

$ln\,(r) + \frac{2}{3} ln\,(d)$ which is essentially the deviation of a network from the preferential attachment model as depicted in Fig. 13B.

### 2.6.4 Influence of two sorts of noise on the edge reduction

To understand the influence of false positives and false negatives on the edge reduction, we also generated synthetic scale-free networks with different levels of noise. We used two different models for noise: the first we term Erdös-Rényi noise which consists of the random addition or removal of edges, the second we term spoke noise, which consists of reconnecting all neighbors of the neighbors of a node to the node itself. The spoke noise model mimics the confusion of bait-prey interactions for indirect interactions of chained prey-bait with bait-bait interactions. As seen in Fig. 13C the addition of Erdös-Rényi noise to scale-free networks decreases their level of edge reduction down toward that of pure random networks, which was to be expected considering that at the limit the addition of 100% Erdös-Rényi noise suppresses all scale-free caracteristics from the networks. In contrast, the addition of spoke noise increases the edge reduction.

### 2.6.5 Protein interaction networks have a higher edge reduction than expected.

The edge reduction, average degree and power graph index of various protein interaction networks were calculated as shown in Table I and in Fig. 13C.

| Protein Interaction Network | # nodes | # edges | $d$ | $r$ | $I$ |
|---|---|---|---|---|---|
| Ito et al. (2001) [Ito et al., 2000] | 3245 | 4367 | 2.69 | 53% | 0.01 |
| Hazbun et al. (2003) [Hazbun et al., 2003] | 1979 | 2514 | 2.54 | 83% | 0.44 |
| Li et al. (2004) [Li et al., 2004] | 2622 | 3955 | 3.01 | 54% | 0.12 |
| Stanyon et al. (2004) [Stanyon et al., 2004] | 232 | 491 | 4.23 | 50% | 0.28 |
| Gunsalus et al. (2004) [Gunsalus et al., 2005] | 281 | 534 | 3.80 | 64% | 0.45 |
| Rual et al. (2005) [Rual et al., 2005] | 1527 | 2529 | 3.31 | 48% | 0.07 |
| Stelzl et al. (2005) [Stelzl et al., 2005] | 1665 | 3083 | 3.70 | 50% | 0.18 |
| Lacount et al. (2005) [LaCount et al., 2005] | 1272 | 2643 | 4.15 | 37% | -0.03 |
| Gavin et al. (2006) [Gavin et al., 2006] | 902 | 1612 | 3.57 | 64% | 0.40 |
| Kim et al. (2006) [Kim et al., 2006] | 577 | 1090 | 3.77 | 61% | 0.40 |
| Lim et al. (2006) [Lim et al., 2006] | 380 | 467 | 2.45 | 85% | 0.43 |
| Krogan et al. (2006) [Krogan et al., 2006] | 2708 | 7123 | 5.26 | 43% | 0.28 |

Table I Higher than expected edge reduction. Average degree ($d$), edge reduction ($r$), and power graph index ($I$) of protein interaction networks

Most networks, exhibit a substantial higher edge reduction than is expected from the preferential attachment model. This higher-than-expected edge reduction may be explained in two ways: either the networks have suffered the addition of spoke noise, or they genuinely contain many cliques and bicliques that have a biological interpretation. Both of these explanations may be valid: spoke noise is known to be occurring typically as an artifact of the experimental methodology, especially in the case of Co-immunoprecipitation or Tandem affinity purification. However, for some networks spoke-like noise cannot explain the higher than expected edge reductions. For example, the Structural Interaction Network (SIN) is a set of interactions carefully

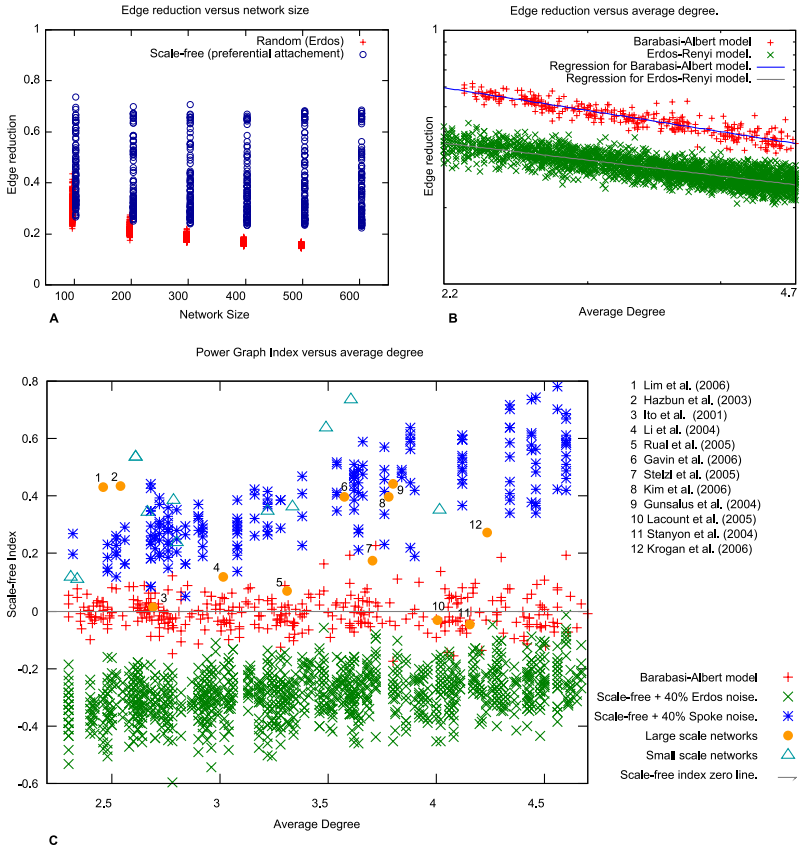


Figure 13: (A) Edge reduction versus network size. Edge reduction is size invariant for synthetic networks generated according to the preferential attachment model but is not for networks generated according to the Erdös-Rényi random network model. (B) Power-law correlation between edge reduction and average degree. The edge reduction is correlated to average degree for both preferential attachment and Erdös-Rényi random networks. In log-log space the correlation is linear, thus the edge reduction follows a power law with respect to the average degree. Moreover, for a given average degree the edge reduction is higher for synthetic scale-free networks than for random networks. (C) Power graph index versus average degree. The power graph index is essentially the logarithmic edge reduction corrected for average degree independence. This index is close to zero for scale free networks, is positive for scale-free networks to which spoke noise has been added, and is negative for scale-free networks with added Erdös-Rényi noise. Moreover, large scale networks from Table. I are also plotted and exhibit a higher than expected edge reduction.

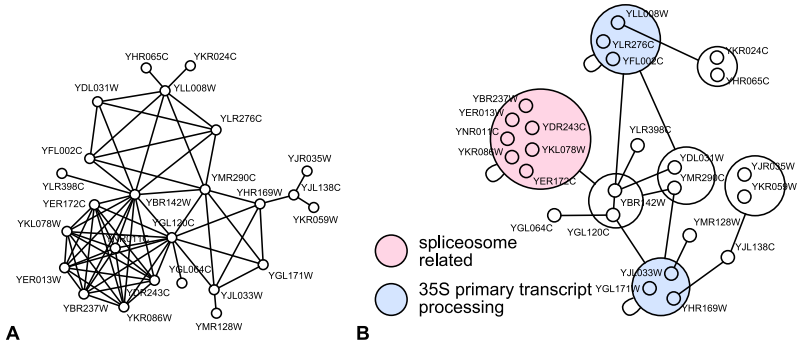


Figure 14: (A) Close-up of a 25 node, 68 edges, connected component of the Structural Interaction Network (SIN) [Kim et al., 2006]. (B) Power graph visualization consisting of 17 power edges, thus an edge reduction of 73%. Three cliques enriched in GO terms related to 35S primary transcript processing and to the spliceosome become explicit in the representation.

curated using structural information: all interactions reported are direct physical interactions explained by a known structural binding [Kim et al., 2006]. Yet this network exhibits an edge reduction which is at least three standard deviations higher than expected. Fig. 14 shows a close-up of a connected component of the SIN that illustrates its richness in structures: we see three cliques and two bicliques. The three cliques are enriched in Gene Ontology [consortium, 2005] terms related to the spliceosome and to 35S primary transcript processing, so essentially the proteins of this component are most likely part of the the ribosome and spliceosome machinery. Small-scale interaction networks – believed to be of higher quality than high-throughput datasets – are also characterized by a high power graph index. This supports the idea that the higher-than-expected edge reduction is indeed due to genuine biological phenomena.

This corroborates results from studies that looked at network motifs identified as functional units in the context of biological networks [Milo et al., 2002]. 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, further confirming their biological significance [Wuchty et al., 2003].

### 2.6.6 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 has also be 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, this bearing similarities to our notion of spoke noise [Stumpf et al., 2005, Han et al., 2005]. 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. However, the degree distribution of geometric random networks is a Poisson distribution whereas most protein networks follow a power-law.

Our results also tend to support the idea that protein interaction networks deviate from the

preferential attachment model, not just because of noise introduced by imperfect methods but also because of more significant biological phenomena.

## 2.7 Domain and gene ontology term enrichment of power nodes

To further support the idea that power nodes are not artifacts of the networks topology but have in fact a biological interpretation, we analysed the enrichment of power nodes in InterPro domains [Mulder et al., 2007] and in Gene Ontology (GO) terms [consortium, 2005]. Our null hypothesis is that "annotations are randomly distributed" according to a hypergeometric distribution. In order to take into account missing domain annotations, only power nodes for which more than 66% of the proteins are annotated with at least one domain are considered. Table II shows that sufficiently annotated power nodes are significantly enriched in domains, with most p-values below 1 per-thousand. Similarly, Table III shows the distribution of p-values for the enrichment in GO terms. Some power nodes are enriched in GO terms but less significantly than for domains, only 4% to 11% of power nodes are enriched at a level of 1 percent.

| PIN | $p < 0.001$ | $0.001 <= p < 0.01$ | $0.01 <= p < 0.1$ | $p >= 0.1$ | i.a. |
|---|---|---|---|---|---|
| Hazbun (2003) | 16% | 0% | 0% | 0% | 84% |
| Li et al. (2004) | 15% | 7% | 0% | 0% | 78% |
| Rual (2005) | 34% | 7% | 1% | 0% | 58% |
| Stelzl (2005) | 14% | 7% | 7% | 0% | 72% |
| Lacount (2005) | 25% | 25% | 13% | 0% | 38% |

Table II Domain enrichment. Power nodes of five large scale protein interaction networks (PIN) are significantly enriched in InterPro domains. Most sufficiently annotated power nodes turn out to be enriched at a level of statistical significance of 1 per-thousand. (i.a. stands for insufficiently annotated)

| PIN | $p < 0.001$ | $0.001 <= p < 0.01$ | $0.01 <= p < 0.1$ | $p >= 0.1$ | i.a. |
|---|---|---|---|---|---|
| Hazbun (2003) | 4% | 5% | 20% | 12% | 59% |
| Li et al. (2004) | 0% | 0% | 0% | 45% | 55% |
| Rual (2005) | 13% | 4% | 9% | 5% | 69% |
| Stelzl (2005) | 0% | 5% | 10% | 6% | 79% |
| Lacount (2005) | 0% | 11% | 11% | 0% | 78% |

Table III Gene Ontology terms enrichment. Enrichment in Gene Ontology terms of the power nodes of five large scale protein interaction networks (PIN). (i.a. stands for insufficiently annotated)

## 2.8 Beyond protein interactions

Other biological networks benefit from power graph representations. An example is protein homology networks [Medini et al., 2006] in which nodes are proteins and edges represent

BLAST E-values below a given threshold. These networks can be modeled as 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 raw gene regulatory networks that also benefits from the power graph representation - in particular since gene duplication events tend to create biclique motifs [Teichmann and Babu, 2004, Milo et al., 2002].

## 2.9 Methods

### 2.9.1 Formal Definition of Power Graphs

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* $G' = \left(V', E'\right)$ is a graph defined on the power set $V' \subseteq \mathcal{P}\left(V\right)$ of *power nodes* connected to each other by *power edges*: $E' \subseteq V' \times V'$. Hence power graphs are defined on the power sets of nodes. The semantics of power graphs are as follows: if two power nodes are connected by a power edge, this means that all nodes of the first power node are connected to all nodes of the second power node. Similarly, if a power node is connected to itself by a power edge, this signifies that all nodes in the power node are connected to each other by edges. The following two conditions are required for simplifying the representations:

**Power node hierarchy condition:** Any two power nodes are either disjoint, or one is included in the other.

**Power edge disjointness condition:** Each node of the original graph is represented by one and only one power edge.

Note that power graphs are not *hypergraphs* for which hyper-edges are *n-tuples* of nodes.

### 2.9.2 Minimal power graphs and edge reduction

For a given graph $G$ it is desirable to search for the power graph with the least number of edges. A power graph is *minimal* if there is no other power graph representation with fewer power edges. There can be several minimal power graphs for a given graph, for example a graph is in itself a power graph of singleton sets.

From a combinatorial point of view the problem of finding a minimal power graph for a graph $G$ amounts to finding a minimum partition of the set of edges into disjoint cliques and bicliques in a way that satisfies the hierarchy and disjointness conditions. The complexity of similar problems has already been investigated. For example the problem of finding the minimal partition of a graph into cliques is known to be NP-hard [R. and Fürer, 1997]. And the problem of finding the minimal biclique partition is NP-complete [Kratzke et al., 1988]. These results do not directly imply the NP-completeness of the minimal power graph problem because of the additional conditions used in the definition. The minimal power graph is to our knowledge an open problem. This is however not critical since in practice one is more interested in *near-minimal* power graphs and not in strict minimality.

Our definition of edge reduction is thus based on near-minimal power graphs. Moreover, it is only computed for connected components of at least 3 nodes, since for smaller components the edge reduction makes little sense.

In the following we present heuristic algorithms that find near-minimal power graphs $G'$ for a graph $G$.

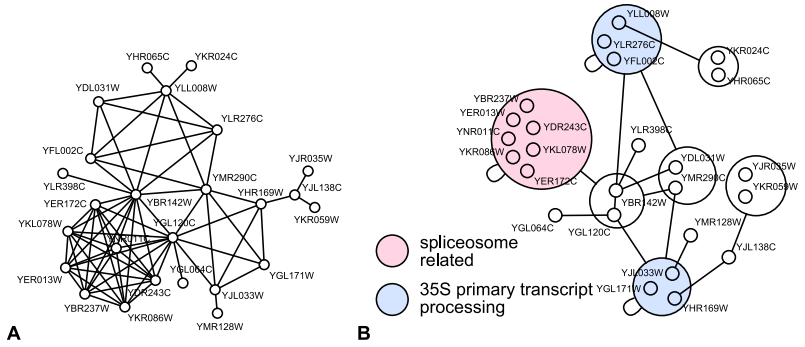


Figure 15: (A) Greedy-based power graph algorithm. First stars are enumerated and decomposed, giving priority to bigger stars. In a second step cliques and bicliques are detected. (B) Clustering-based power graph algorithms. First a neighborhood similarity clustering of the nodes is performed providing candidate power nodes. In a second step power edges are searched between nodes and candidate power nodes.

## 2.10 Near-minimal power graph extraction algorithms

We have developed three algorithms for computing near-minimal power graph representations from graphs. The first is a greedy algorithm that first detects hubs in the networks and then incrementally builds a power graph representation. The two other algorithms are based on node clustering that relies on neighborhood similarity measures.

### 2.10.1 Greedy power graph extraction algorithm

The minimal power graph problem is to be seen as an optimization problem in which the power graph achieving the highest edge reduction is searched. The greedy power graph extraction algorithm follows the heuristic of making the locally optimum decision at each step with the hope of finding the global optimum, or at least to come reasonably close to it [Cormen et al., 2001]. As shown in Fig. 15A, the algorithm proceeds as follows: In a first phase, candidate stars abstracting the most edges are added successively to the power graph until no more stars are added, cliques are detected as well. In a second step bicliques are identified.

### 2.10.2 Clustering-based extraction algorithms

Another approach to compute near-minimal power graphs is the use of clustering methods to first identify potential power nodes, and then use these as a heuristic for the search of power edges.

The MULIC clustering algorithm [Andreopoulos et al., 2007] that relies on the notion of *neighborhood similarity* was our starting point for investigating such an approach. We then further refined the neighborhood similarity to better suit the problem of finding potential power nodes. Other clustering techniques such as spectral clustering [Bu et al., 2003] can also be used.

*Neighborhood similarity clustering* is an intuitive way to identify candidate power nodes. Fig. 15B shows how clustering nodes having identical and similar neighborhoods can provides candidate sets for cliques and bicliques.

**Centrality-weighted neighborhood similarity** In order to apply a clustering algorithm, we need a similarity measure that captures the notion of similarity between neighborhoods of nodes. More generally we define the similarity of neighborhoods of sets of nodes. Given $U$ a set of nodes, let $N(U)$ be the set of its neighboring nodes.

The detection of stars and other highly asymmetric bicliques cannot be done without making highly connected nodes more significant neighbors. In essence, all nodes in a graph are not equally significant and the calculation of the neighborhood similarity must take these differences into account. In order to take into account the centrality of nodes, we use as centrality measure for a set of nodes $U$, the degree of this set, i.e. the cardinality of the neighbors's set:

$$c(U) = |N(U)|$$

We define the centrality-weighted neighborhood similarity of two sets $U$ and $V$ of nodes as:

$$s_f(U, V) = \frac{c(N(U) \cap N(V))}{c(N(U) \cup N(V))}$$

This similarity is a weighted version of the Jaccard index of two sets [Jaccard, 1901]. It is always between zero and one: it is zero if the sets $U$ and $V$ have no common neighbors, and one if both have identical neighborhoods.

We choose to use a hierarchical clustering algorithm [Eisen et al., 1998] for clustering the nodes of the original graph. Once the nodes have been clustered in sets constituting candidate power nodes, it suffices to search among all pairs of candidates the ones that correspond to power edges. Any remaining edges – not covered by a power edge – are added.

### 2.10.3 Comparison of power graph extraction algorithms

In the following we compare the three algorithms that compute near-minimal power graphs: greedy based power graph extraction, MULIC-based power graph extraction [Andreopoulos et al., 2007], and hierarchical clustering based power graph extraction.

We use 346 small and large scale networks from IntAct [Hermjakob et al., 2004] to compare the degree of edge reduction, the minimal p-values for domain enrichment, duration of computation, and the level of nesting of power nodes.

Throughout this paper we have used the neighborhood similarity based hierarchical clustering for all visualization examples and for the edge reduction analysis. The advantages of this algorithm are the high edge reduction together with a moderate level of nesting, the major disadvantage being that hierarchical clustering has the highest computational cost. Computing the similarity of two neighborhoods is linear in the number of nodes, and in our case the hierarchical clustering is quadratic, thus the time required for computing a power graph using hierarchical clustering is proportional to $n^3$, where $n$ is the number of nodes. In contrast, the complexity of the greedy and MULIC-based algorithms is quadratic in the number of nodes. In conclusion, the greedy based algorithm offers the best trade off between power graph quality (high edge reduction and low nesting) and computation time whereas the hierarchical based algorithm offers the best quality disregarding the computation time.

### 2.10.4 Quantifying near-minimality

How close are these algorithms to minimal power graphs? As we have already discussed, the problem of finding the minimal power graph of a graph is an open combinatorial problem. In-

stead, we investigated for a class of synthetic power graphs – chosen for their intricate structure and near minimality – how much edge reduction is lost by going from the power graph to the graph, and then recomputing the power graph (synthetic power graph → graph → computed power graph). The measured edge reduction loss for these networks is thus a tight lower-bound of the loss incurred by the minimal power graphs. For this experiment we start with synthetic power graphs that are made of cliques and bicliques linked by single edges, the underlying graph is computed, for which we apply the hierarchical clustering and greedy based algorithms in order to compare the original power graph with the result of the algorithms. The synthetic power graphs were designed to be near-minimal from the start by limiting the possibility of edge removal by the merging and/or splitting of power nodes. Sup. Fig. 1F shows the results: both algorithms come close to the original edge reduction with an average loss of at best 8 edge reduction points.

## 2.11 Noise Models

Both Erdös-Rényi noise and Spoke noise preserve the average degree of the original graphs. This is important since we wanted to measure the influence on the edge reduction independently of the average degree - whose influence we already know. A simple way to guarantee the conservation of the average degree is to alter the graphs by using *rewiring* operations by which an edge is detached from a pair of nodes and attached to another pair of nodes that does not already constitute an edge. The noise level, in percentage, corresponds to the proportion of rewired edges with respect to all edges present in the network. The two noise models thus only differ by the modus-operandi of the rewiring operations. The Erdös-Rényi noise is such that edges are chosen randomly for rewiring and are reattached randomly to another pair of nodes. The spoke noise consists in rewiring the edges of a node such that it becomes linked to the neighbors of its neighbors. This model simulates the typical errors introduced by affinity purifications.

## 2.12 Calculation of the p-value

We evaluate the enrichment of a cluster's proteins with domains using $p$-values assuming an hypergeometric distribution [King et al., 2004]. The $p$-value for a cluster of size $C$ containing $k \leq C$ proteins with domain $X$ is:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\left( \begin{array}{c} C \\ i \end{array} \right) \left( \begin{array}{c} G-C \\ n-i \end{array} \right)}{\left( \begin{array}{c} G \\ n \end{array} \right)}$$

This is the likelihood that the cluster has $k$ or more proteins with domain $X$, if the cluster's contents were drawn randomly from the set of known proteins. Where $G$ is the size of the set of known proteins among which $n \leq G$ have domain $X$.

## 2.13 Power graph visualization with Cytoscape

Visualization is a first step toward understanding networks. Several tools exist to visualize biological networks such as Cytoscape [Shannon et al., 2003], Pajek [Batagelj and Mrvar, 2003], Osprey [Breitkreutz et al., 2003], Navigator [Motamed-Khorasani et al., 2007], VisANT

[Hu et al., 2005], ProViz [Iragne et al., 2005], and GraphViz [Gansner and North, 2000]. We have implemented the described algorithms and made them available as a plugin to Cytoscape. (www.biotec.tu-dresden.de/schroeder/group/powergraphs)

# References


[Alberts et al., 1998] Alberts, B., Bray, D., Jonhson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (1998). *Essential cell biology: an introduction to the molecular biology of the cell*, chapter 3 Energy, Catalysis, and Biosynthesis, pages 77–106. Garland Publishing, Inc.

[Andreopoulos et al., 2007] Andreopoulos, B., An, A., Wang, X., Faloutsos, M., and Schroeder, M. (2007). Clustering by common friends finds locally significant proteins mediating modules. *Bioinformatics*.

[Bader and Hogue, 2002] Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–997.

[Bader and Hogue, 2003] Bader, G. D. and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2.

[Barabasi and Albert, 1999] Barabasi and Albert (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[Batagelj and Mrvar, 2003] Batagelj, V. and Mrvar, A. (2003). *Graph Drawing Software*, chapter Pajek - Analysis and Visualization of Large Networks, pages 77–103. Springer.

[Beyenbach and Wieczorek, 2006] Beyenbach, K. W. and Wieczorek, H. (2006). The v-type h+ atpase: molecular structure and function, physiological roles and regulation. *J Exp Biol*, 209(Pt 4):577–589.

[Bockmayr and Courtois, 2002] Bockmayr, A. and Courtois, A. (2002). Using hybrid concurrent constraint programming to model dynamic biological systems.

[Breitkreutz et al., 2003] Breitkreutz, B.-J., Stark, C., and Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol*, 4(3):R22.

[Bu et al., 2003] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., and Chen, R. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res*, 31(9):2443–2450.

[Chabrier-Rivier et al., 2004] Chabrier-Rivier, N., Fages, F., and Soliman, S. (2004). The biochemical abstract machine biocham. In Springer-Verlag, editor, *Proceedings of the Second International Workshop on Computational Methods in Systems Biology CMSB'04, Paris, France, May 2004)*.

[consortium, 2005] consortium, G. (2005). The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, 34(Database issue):D322–6.

[Cormen et al., 2001] Cormen, Leiserson, and Rivest (2001). *Introduction to Algorithms*.

[Damásio et al., 1997] Damásio, C. V., Pereira, L. M., and Schroeder, M. (1997). REVISE: Logic programming and diagnosis. In Dix, J., Furbach, U., and Nerode, A., editors, *Proocedings of the Fourth International Conference on Logic Programming and Non-Monotonic Reasoning*, number 1265 in Lecture Notes in Artificial Intelligence, pages 353–362. Springer-Verlag.

[Deng et al., 2002] Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–1548.

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868.

[Eker et al., 2002] Eker, S., Knapp, M., Laderoute, K., Lincoln, P., , and Talcott, C. (2002). Pathway logic: Executable models of biological networks. In *Fourth International Workshop on Rewriting Logic and Its Applications (WRLA'2002*, volume 71 of *Electronic Notes in Theoretical Computer Science*. Elsevier.

[Erdös and Rényi, 1960] Erdös, P. and Rényi, A. (1960). Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5.

[Euler, 1772] Euler, L. (1772). *Lettres à une Princesse d'Allemagne*, volume 2.

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

[Gansner and North, 2000] Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software — Practice and Experience*, 30(11):1203–1233.

[Gavin et al., 2006] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636.

[Giot et al., 2003] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. *Science*, 302(5651):1727–1736.

[Goh et al., 2006] Goh, K.-I., Salvi, G., Kahng, B., and Kim, D. (2006). Skeleton and Fractal Scaling in Complex Networks. *Physical Review Letters*, 96(1):018701.

[Guimaraes et al., 2006] Guimaraes, K. S., Jothi, R., Zotenko, E., and Przytycka, T. M. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biol*, 7(11):R104.

[Gunsalus et al., 2005] Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J.-D. J., Hao, T., Berriz, G. F., Bertin, N., Huang, J., Chuang, L.-S., Li, N., Mani, R., Hyman, A. A., Sönnichsen, B., Echeverri, C. J., Roth, F. P., Vidal, M., and Piano, F. (2005). Predictive

models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861–865.

[Han et al., 2005] Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 23(7):839–844.

[Hazbun et al., 2003] Hazbun, T. R., Malmström, L., Anderson, S., Graczyk, B. J., Fox, B., Riffle, M., Sundin, B. A., Aranda, J. D., McDonald, W. H., Chiu, C.-H., Snydsman, B. E., Bradley, P., Muller, E. G. D., Fields, S., Baker, D., Yates, J. R., and Davis, T. N. (2003). Assigning function to yeast proteins by integration of technologies. *Mol Cell*, 12(6):1353–1365.

[Hermjakob et al., 2004] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–D455.

[Hofestadt and Thelen, 1998] Hofestadt, R. and Thelen, S. (1998). Quantitative modeling of biochemical networks. *In Silico Biol*, 1(1):39–53.

[Hollunder et al., 2005] Hollunder, J., Beyer, A., and Wilhelm, T. (2005). Identification and characterization of protein subcomplexes in yeast. *Proteomics*, 5(8):2082–2089.

[home page., ] home page., W. http://en.wikipedia.org.

[Hu et al., 2005] Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., and Delisi, C. (2005). Visant: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res*, 33(Web Server issue):W352–W357.

[Hurlebaus et al., 2002] Hurlebaus, J., Buchholz, A., Alt, W., Wiechert, W., and Takors1, R. (2002). Mmt - a pathway modeling tool for data from rapid sampling experiments. *In Silico Biology*.

[Iragne et al., 2005] Iragne, F., Nikolski, M., Mathieu, B., Auber, D., and Sherman, D. (2005). ProViz: protein interaction visualization and exploration. *Bioinformatics*, 21(2):272–274.

[Ito et al., 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574.

[Ito et al., 2000] 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 combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 97(3):1143–1147.

[Jaccard, 1901] Jaccard, P. (1901). Bulletin del la société vaudoise des sciences naturelles. 37:241–272.

[Jothi et al., 2006] Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J Mol Biol*, 362(4):861–875.

[Kanehisa et al., 2006] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Research*, 34:D354–D357.

[Kanehisa et al., 2004] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The kegg resource for deciphering the genome. *Nucleic Acids Research*, 32:D277–D280.

[Kashtan et al., 2004] Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2004). Topological generalizations of network motifs. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(3 Pt 1):031909.

[Keller et al., 2001] Keller, D. M., Zeng, X., Wang, Y., Zhang, Q. H., 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(2):283–292.

[Khanin and Wit, 2006] Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *J Comput Biol*, 13(3):810–818.

[Kim et al., 2006] Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938–1941.

[Kim et al., 2002] Kim, W. K., Park, J., and Suh, J. K. (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform*, 13:42–50.

[King et al., 2004] King, A. D., Przulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020.

[Kolodrubetz et al., 1982] Kolodrubetz, D., Rykowski, M. C., and Grunstein, M. (1982). Histone h2a subtypes associate interchangeably in vivo with histone h2b subtypes. *Proc Natl Acad Sci U S A*, 79(24):7814–7818.

[Kratzke et al., 1988] Kratzke, T., Reznick, B., and West, D. (1988). Eigensharp graphs: Decomposition into complete bipartite subgraphs. *Trans. Amer. Math. Soc.*, 308(2):637–653.

[Krogan et al., 2006] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., Onge, P. S., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643.

[L et al., 2004] L, B. M., R, F. J., Byron, G., and S, H. W. (2004). Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291.

[LaCount et al., 2005] LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. E. (2005). A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, 438(7064):103–107.

[Landgraf et al., 2004] 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(1):E14.

[Li et al., 2006a] 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(2):456–461.

[Li et al., 2006b] Li, H., Li, J., and Wong, L. (2006b). Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996.

[Li et al., 2004] Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Heuvel, S. V. D., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans. *Science*, 303(5657):540–543.

[Lim et al., 2006] Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J.-F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabási, A.-L., Vidal, M., and Zoghbi, H. Y. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4):801–814.

[Liu et al., 2005] Liu, Y., Liu, N., and Zhao, H. (2005). Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21(15):3279–3285.

[Mann et al., 2001] Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem*, 70:437–473.

[Mason and Struhl, 2003] Mason, P. B. 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(22):8323–8333.

[Medini et al., 2006] 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(12):e173.

[Milo et al., 2002] 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(5594):824–827.

[Moran et al., 1990] Moran, L., Norris, D., and Osley, M. A. (1990). A yeast h2a-h2b promoter can be regulated by changes in histone gene copy number. *Genes Dev*, 4(5):752–763.

[Morrison et al., 2006] Morrison, J. L., Breitling, R., Higham, D. J., and Gilbert, D. R. (2006). A lock-and-key model for protein-protein interactions. *Bioinformatics*.

[Motamed-Khorasani et al., 2007] Motamed-Khorasani, A., Jurisica, I., Letarte, M., Shaw, P. A., Parkes, R. K., Zhang, X., Evangelou, A., Rosen, B., Murphy, K. J., and Brown, T. J. (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(2):198–214.

[Mulder et al., 2007] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J. A., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2007). New developments in the interpro database. *Nucleic Acids Res*, 35(Database issue):D224–D228.

[Ng et al., 2003] Ng, S.-K., Zhang, Z., and Tan, S.-H. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–929.

[Nye et al., 2006] Nye, T. M. W., Berzuini, C., Gilks, W. R., Babu, M. M., and Teichmann, S. (2006). Predicting the strongest domain-domain contact in interacting protein pairs. *Stat Appl Genet Mol Biol*, 5:Article5.

[Nye et al., 2005] Nye, T. M. W., Berzuini, C., Gilks, W. R., Babu, M. M., and Teichmann, S. A. (2005). Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21(7):993–1001.

[Patil and Nakamura, 2005] Patil, A. and Nakamura, H. (2005). Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6:100.

[Przulj et al., 2004] Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515.

[R. and Fürer, 1997] R., D. and Fürer, M. (1997). Approximation of k-set cover by semi-local optimization. In *Proc. 29th Ann. ACM Symp. on Theory of Comp.*, pages 256–265. ACM.

[Rain et al., 2001] Rain, J. C., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A., and Legrain, P. (2001). The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817):211–215.

[Regev et al., 2004] Regev, A., Panina, E. M., Silverman, W., Cardelli, L., and Shapiro, E. (2004). Bioambients: An abstraction for biological compartments. *Theoretical Computer Science*, pages 141–167.

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

[Rigaut et al., 1999] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Sèraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032.

[Riley et al., 2005] Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, 6(10):R89.

[Rual et al., 2005] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504.

[Song et al., 2005] Song, C., Havlin, S., and Makse, H. A. (2005). Self-similarity of complex networks. *Nature*, 433(7024):392–395.

[Stanyon et al., 2004] Stanyon, C. A., Liu, G., Mangiola, B. A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., and Finley, R. L. (2004). A drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol*, 5(12):R96.

[Stelzl et al., 2005] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968.

[Stumpf et al., 2005] Stumpf, M. P. H., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A*, 102(12):4221–4224.

[Tamaddoni-Nezhad et al., 2004] Tamaddoni-Nezhad, A., Kakas, A. C., Muggleton, S., and Pazos, F. (2004). Modelling inhibition in metabolic pathways through abduction and induction. In *ILP*, pages 305–322.

[Taylor and Raes, 2004] Taylor, J. S. and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*, 38:615–643.

[Teichmann and Babu, 2004] Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet*, 36(5):492–496.

[Thomas et al., 2003] Thomas, A., Cannings, R., Monk, N. A. M., and Cannings, C. (2003). On the structure of protein-protein interaction networks. *Biochem Soc Trans*, 31(Pt 6):1491–1496.

[Veit and Herrmann, 2003] Veit, M. and Herrmann, S. (2003). Model-view-controller and Object Teams: A perfect match of paradigms. In *AOSD'03*, pages 140–149.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[Wuchty et al., 2003] Wuchty, S., Oltvai, Z. N., and Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*, 35(2):176–179.