Universidad de Concepción
Dirección de Postgrado
Facultad de Ingeniería - Programa de Magíster en Ciencias de la Computación

CELL CYCLE AND PROTEIN COMPLEX DYNAMICS IN DISCOVERING SIGNALING PATHWAYS IN YEAST (DINÁMICAS DEL CICLO CELULAR Y COMPLEJOS DE PROTEÍNAS EN EL DESCUBRIMIENTO DE VÍAS DE SEÑALIZACIÓN EN LEVADURA)

Tesis para optar al grado de
MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN

POR
DANIEL IGNACIO INOSTROZA RODRÍGUEZ
CONCEPCIÓN, CHILE
Enero, 2018

Profesor guía: DIEGO SECO NAVEIRAS
CECILIA HERNÁNDEZ RIVAS
Departamento de Ingeniería Informática y Ciencias de la Computación
Facultad de Ingeniería
Universidad de Concepción

Abstract

Signaling pathways are responsible for the regulation of cell processes, such as, monitoring external environment, transmiting information across membranes, and making cell fate decisions. Given the increasing amount of biological data available and the recent discoveries showing that many diseases are related to the disruption of cellular signal transduction cascades, modeling signaling pathways in cell biology has become an active research topic in past years. However, reconstruction of signaling pathways remains a challenge mainly because of the need of systematic approaches for predicting causal relationships, like edge direction and activation/inhibition among interacting proteins in the signal flow. We propose an approach for predicting signaling pathways that integrate protein interactions, gene expression, phenotypes, and protein complex information. Our method first finds candidate pathways using an edge direction algorithm and then defines a graph model to include causal activation relationships among proteins in candidate pathways using cell cycle gene expression and phenotypes to infer consistent pathways. Then, we incorporate protein complex coverage for deciding the final predicted signaling pathways. We show that our approach improves the results of previous approaches, between 21 and 50 %, using different ranking metrics.

Resumen

Las vías de señalización son responsables de la regulación de los procesos celulares, tales como el monitoreo del ambiente externo, la transmisión de información a través de la membrana celular, y la toma de desiciones en torno al destino de la célula. Dada la cantidad creciente de información biológica disponible y los descubrimientos recientes sobre el hecho que muchas enfermedades están relacionadas con la interrupción de las señalizaciones en cascada, modelar vías de señalización en biología celular ha llegado a ser un tema activo de investigación en los últimos años. A pesar de esto, la reconstrucción de vías de señalización continúa siendo un gran desafío, principalmente debido a la falta de enfoques sistemáticos para predecir estas relaciones, tales como orientación de interacciones y la activación o inhibición entre interacciones de proteínas en el flujo de una vía. Este trabajo propone un enfoque para predecir vías de señalización que integran interacciones de proteínas, expresión de genes, fenotipos, e información de complejos proteicos. Primero, nuestro método encuentra vías candidatas, usando un algoritmo de orientación de aristas, y define un modelo de grafos para incluir las relaciones de activación entre proteínas en las vías candidatas, usando expresión de genes en el ciclo celular y fenotipos, para inferir pathways consistentes. Después, incorporamos un algoritmo de cubrimiento de complejos proteicos para elegir las vías de señalización predichas finales. Mostramos que nuestro enfoque mejora los resultados de los enfoques de previas investigaciones, entre un 21 y un 50%, usando diferentes métricas de ranking.

Index

Tables Index

Figures Index

Chapter 1

INTRODUCTION

Proteins are molecules formed by sequences of amino acids. They ussually interact with each other to perform specific functions in organisms. Discovering the protein roles in different functions is an important research area in the biological and biochemical field, because of the impact that such information may have in the creation of new treatments of several diseases and in the comprehension of functions in living systems. A kind of cell activity where several proteins work together in sequence is called “signaling pathway”. A signaling pathway can be seen as a linear path in cascade, where multiple proteins associate and/or modify each other to perform a specific function. In general, a signaling pathway has a set of proteins whose sequence interaction from a source to a target produces the activation of transcription factors, which regulate the gene expression or inhibition. Ahother kind of biological function, where multiple proteins work together, is called a “protein complex”, where there is a high level of interaction among the involved proteins, but there is not a linear dependency of their interactions.

Chapter 2

STATE OF ART

Diverse technologies of biological experimentation that have been developed through time, have made possible the compilation of Protein-Protein Interaction (PPI) networks, which have pairs of protein interactions in a determinated experimental context. There are several methods that aim to discover interactions between proteins, such as yeast-two hybrid (Y2H), affinity purification-mass spectrometry (AP-MS) approaches [1], or interaction reports inferred from mining information in scientific publications. Nevertheless, those techniques are not completely reliable, since they can show interactions that do not happen “in vivo” or are not able to detect interactions that really exist [14]. Those PPI network databases ussually are available to the scientific community and are used to extract information about how proteins interact in different organisms. On the other hand, there are databases that keep cured protein complexes (CYC2008 [24], SGD [4], MIPS [19], PCDq [15] and CORUM [26]) and there are several computational systems that predict protein complexes [10] [22]. Also, there are databases that collect information about cured signaling pathways (e.g. Wikipathways and KEGG) [3], as well as databases that store PPI networks related to a diverse number of diseases [18].

Aditionally, there is the “Gene Ontology Consortium” (GOC), which objective is to keep the information of biological systems of many organisms updated. The information is structured and storaged as ontologies (gene ontology) and annotations (gene annotation). One of the most interesting analysis that uses the GOC data is the enrichment analysis, which allows to determine the biological functions where a set of proteins works. Enrichment is a statistical method that relates a group of significantly enriched genes that coexpress in certain biological functions. This analysis provides a better understanding of biological processes in organisms by relating genes themselves with different ontologies. There are several software tools that calculate the

enrichment of a set of proteins to verify the degree of association between the genes in determinated biological functions, such as GO, DAVID and Enrichr [12] [17].

Despite the progress made to date, proteins are molecules that have different and multiple functions in every organism and, as biological systems are complex, there are still many gaps about their interactions, behavior and functions. In particular, the prediction of alternative or missing signaling pathways, as well as their relation to protein complexes and diseases, may provide insight on how regulation processes or crosstalk works [27].

From the computer science point of view, PPIs can be modeled as undirected graphs [21] and signaling pathways can be seen as paths in a directed graph. The problem gets more complex when PPIs include information about the confidence of the interaction's actual existence between every pair of proteins, which is usually modeled as undirected weighted graphs [33] [11]. One of the most interesting models is the use of hypergraphs, which is used to model and store specific functions of proteins and their interactions. There are other alternatives like adapted hypergraphs, also called signaling hypergraphs [25]. Such work models reactions as complex assembly and dissociation, combinatorial regulation, and protein activation/inactivation.

Prediction of biological signaling pathways from PPIs is a complex task, mainly because of the large size of PPI networks and also because signaling pathways are directed. Thus, there is a high number of possible signaling pathways to consider from a PPI database, which can produce high rates of false positives and false negatives in the results. This is why it is necessary to develop new methods capable of distinguishing, only the signaling pathways with a good chance of actually existing. The identification of signaling pathways is a critical point to understand biological processes, as well as pathological alterations of these functions that may trigger diseases. In this sense, several researches have showed that the irregular behaviour of certain proteins trigger many diseases [3, 8, 2, 16].

To date, this problem has been approached from many points of view. Gitter [8] finds pathways starting from a weighted PPI (to represent the degree of confidence between the proteins in every interaction) and predicts alternative pathways using a random orientation algorithm and a local search one. Their proposal aims to maximize

all weights from interactions in every pathway, since a pathway is more reliable if the multiplication of its weights is greater. The authors first build a high quality PPI network, where all the considered interactions are obtained from different databases and they take in consideration their correlation with scientific reports, so every interaction in the PPI is documented and takes a different weight value according to the number of publications that support it.

Cao [3] shows a pathway prediction tool that uses a distance based metric (DSD: Diffusion State Distance), measuring the topological similarity of proteins in a network, adding information from databases to make it more specific (e.g. experiments, number of researches that prove the interactions and reference pathways). Shin [29] uses a shortest path algorithm based on Dijkstra [6], choosing the best pathways as the ones that minimize the pathway length. Nguyen [20] uses the genetic algorithm [5], which optimizes the fitness (objective function) in terms of degree of confidence for each candidate pathway (like Gitter's algorithm) to solve the problem. Vinayagam et al. [32] proposed a computational model that predicts activation/inhibition performing phenotype correlation among proteins and build a signed PPI network for *Dropsophila melanogaster*, where the sign is positive for activation and negative for inhibition relationships. Even though there are some approaches that integrate some biological knowledge for signaling pathway predictions, genome scale reconstruction of signaling pathways is still challenging, mainly because causal relationships are difficult to infer [32].

As mentioned, there are many approaches to solve the problem, but the information used is not based on biological data, temporal relations, protein complexes association, enrichment or any other information that can add biological context knowledge to improve the penality of the predicted pathways. This idea was derived from [30], where it is studied the relation of temporality (via Cell cycle dynamics and protein expression) with protein complexes, in order predict protein complexes, showing that the use of that biological information improves the model for this kind of predictions.

In this work, we propose an algorithm that allows us to integrate the information about PPIs with biological protein knowledge available in public databases, in order to provide biological context for every pathway and its interactions. Our approach is based

on two steps. The first step applies an edge orientation and local search algorithm in the input PPI to find candidate signaling pathways, whereas the second step consists of defining a graph model and a decision algorithm that includes temporal biological data to determine which candidate pathways are biologically consistent. Information like Cell Cycle dynamics in pathways and protein complexes are the best choices. We evaluated our method using the ranking metrics proposed by Gitter et al. [8], included other centrality measures, and found that relating biological information with PPIs to predict signaling pathways using our method improves precision in the predictions.

Chapter 3

MAIN WORK

3.1 Biological knowledge

Research proposals have developed a wide range of biological knowledge describing biological processes, components, or structures in which individual genes and proteins are known to be involved, such as protein complexes, signaling pathways and crosstalks.

3.1.1 Cell cycle information

The cell cycle is a set of events where the cell grows and develops processes that lead to the duplication of its DNA and subsequently cell division. The cell cycle has four phases: G1, S, G2 and M, which happen in sequence and, in each one of them, the cell components fulfill specific functions. The G1 and G2 phases are gaps, the S phase represents the synthesis (replication of its DNA) and the M phase represents the mitosis and cytokinesis. Also, there are two checkpoints, where the cell verifies if its ready to continue with the next phase. Those checkpoints are G1/S (at the end of G1) and G2/M (at the end of G2) [9].

There are many datasets related to the mitotic cell cycle. These datasets include microarray-based time courses of mRNA expression, mass-spectrometry-based proteomics on protein expression during the cell cycle, systematic screens for cyclin-dependent kinase (CDK) sustrates and high-content screening for knockdown phenotypes. All these datasets provide a high detail of information on the mitotic cell cycle and its many regulatory layers. As combining and analyzing all this information is a complex task, Cyclebase [28] aims to address this problem by proccesing

different datasets, mapping them to common gene identifiers and normalizing experiments onto a common timescale, facilitating direct comparison of expression profiles between all experiments within an organism. Current Cyclebase content is updated with new mRNA and protein expression data, and integrated cell cycle phenotype information from high content screens and model organism databases. Cyclebase also provides an easy way of obtaining information about cell cycle peak gene expression and phenotypes of indivual genes. Figure 3.1 shows the information available for a gene (YBR088C) in yeast (*Saccharomyces cerevisiae*), where it is displayed the gene expression in the different phases of the cell cycle and the time course experiments, where it is observed the periodic behavior of the gene. In addition we show a sample of the information observed for each gene related to the peak expression and phenotype for a set of yeast genes, including YBR088C.

3.1.2 Protein Complex information

Proteins are known to participate in several biological processes such as transport, signaling, metabolic and enzymatic catalysis. Most proteins interact with others forming functional units, called protein complexes, which allows them to perform biological functions in a collaborative way. Many proteins participate in different protein complexes according to the function needs of the organism. Understanding the functions of proteins is important for many diseases since some research studies have shown that the deletion of some proteins in a network can have lethal effects on organisms [13]. This has been an important motivation for the research community to propose different prediction methods for protein functions, protein complexes and signaling pathways [8, 14, 31].

In the context of yeast, one of the first gold standards datasets of protein complexes was cataloged by Munich Information Center for Protein Sequences (MIPS). MIPS contains 203 curated protein complexes. Another known gold standard for yeast is the SGD (Saccharomyces Genome Database), which currently contains 323 yeast complexes. CYC2008 is the most recent gold standard dataset of protein complexes for yeast [24]. It contains 408 manually curated heteromeric protein complexes, which reliability is supported by the current literature. CYC2008 has been used as a reference

by many protein complex prediction tools [31]. The CYC2008 reference was compiled with the intention of defining an up-to-date gold standard that considers the new yeast complexes identified in small-scale and large-scale experiments as well as the protein complexes cataloged in MIPS and SGD.

3.1.3 Enrichment information

One of the most interesting approaches is the gene set enrichment analysis. In general, the enrichment analysis is an association of a set of proteins with a functional biological term [17]. To perform this analysis it has been developed a representation of genes and their attributes from several species, which is called the Gene Ontology. This representation keeps a vocabulary of genes and all their attributes, so all the species data are unified in the same nomenclature. All of this effort allows us to extract computational analysis of a set of genes, to see how much they are related. The ontology is formed by gene product properties and is defined by three domains:

- Biological process: role in operations needed to the functioning of living organisms.
- Molecular function: activities or events at a molecular level.
- Cellular component: localization of components in the cell.

One of the first forerunner tools available to make this analysis was the GO and then, many others have been developed, such as FatiGO, BiNGO, Enrichr and TermFinder. These tools are available for several species, but mostly for homo sapiens and yeast, and there are the three onthologies available so people can select any of them to make the analysis. Within the information used for the process, we can find chromosome location of genes, computationally predicted targets of microRNAs, transcription factors and membership of genes in pathway databases. All of the tools return a statistical value (called p-value) that measures the degree of relationship of the set of proteins according to the different biological criteria of each analysis. This value can take ranges between zero and one. The lower the value, the more related are the set of proteins analyzed.

3.1.4 Gold Standard Signaling Pathways

Several databases keep cured signaling pathways, such as KEGG, Wikipathways and Science Signaling Database of Cell Signaling. They keep relevant information about pathways, such as their names, the proteins involved, their interactions, directions, activation/inhibition relationships, localization within the cell, function in the cell and scientific reports related to their interactions. They are called Gold Standard as they are a experimentally tested reference and are the main objective to get in prediction researches.

3.2 Method

In this section, we define the signaling pathway discovery problem and propose a method that incorporates biological data to infer causal or temporal relationship in the signal flow. Our method consists of composing an edge orientation optimization heuristic with a graph model and decision algorithms based on biological data.

3.2.1 Problem formulation

Let $G(V, E, w)$ be a weighted undirected graph, which models a PPI network with proteins as vertices in V, protein interactions as edges in E, and weights, w, as the confidence of a real protein interaction. We denote $w(e)$ as the weight (confidence) of an edge e in $G(V, E, w)$. Let also assume a set of (s_i, t_i) pairs of $source$, $target$ proteins in V and a maximum path length h in G for each pair (s, t).

We formulate the problem of discovering signaling pathways in two steps. The first step has the goal of defining an edge orientation and a fast heuristic for detecting *candidate pathways*, which do not include temporal information of biological relationships among genes or proteins. The second problem is defined as decision rules, where each candidate pathway is tested using biological data that include temporal dynamics using cell cycle and protein complexes. As a result, we obtain a set of signaling pathways, which we called *consistent pathways*. Then, each consistent pathway is tested for protein complex coverage. As a result, we obtain predicted pathways.

Definition 3.2.1. Candidate pathway.
The first problem is to orient edges $e = (u, v) \in E$ in an undirected graph, $G(V, E, w)$, from vertices u to v or from v to u so that all possible paths between $sources$ and $targets$ of length at most h is maximized. Each path is defined as $P = \{s_i, v_1, v_2, \ldots, v_j, t_i\}$, where each pair of consecutive vertices in a path form an edge $(v_j, v_{j+1}) \in E$ and the first vertex is a $source$, s_i, and the last is a $target$, t_i. We denote these paths as *candidate pathways*.

Definition 3.2.2. Consistent pathway.
Given, $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ a set of candidate pathways, where a candidate pathway $P_i = \{v_1, v_2, \ldots, v_j\}$ is a path in G, $v_1 = s_i$, and $v_j = t_i$. A consistent pathway is a candidate pathway where all of its edges, formed by proteins, satisfy temporal cell-cycle dynamics, or protein complex rules (after a candidate pathway is processed by Algortithm 1).

Definition 3.2.3. Predicted pathway.
Given $\mathcal{P}_c = \{P_{c1}, P_{c2}, \ldots, P_{cm}\}$ a set of consistent pathways, and $covPercent$ a protein complex coverage percentage. A predicted pathway is a consistent pathway where all of its edges formed by proteins satisfy a certain percentage of coverage, given by $covPercent$. These pathways are obtained after the consistent pathways are processed by Algorithm 2.

3.2.2 Approach

Our approach finds candidate pathways using edge orientation of PPI networks and a fast heuristic for determining simple paths of length at most L. Gitter et al. [8] showed that determining the maximum edge orientation in a graph is NP-hard, and the authors proposed a random orientation algorithm with local search that provides the best results for finding signaling pathways. However, their proposal does not consider the temporal dynamics involved in signaling pathways. Our proposal uses Gitter's algorithm as a first step, but includes a second step adding decision rules to examine each candidate pathway to consider temporal dynamics. We model such dynamic using cell cycle phases and protein phenotypes as well as the protein involvement in protein

complexes.

We define a cell cycle graph $CG = (A, B)$, where A represents the cell cycle phases, B their transitions and $N(x)$ returns the transitions from one phase to others in CG.

$$\begin{aligned} A &= \{G1, G1/S, S, G2, G2/M, M\} \\ B &= \{(x, y) \in A \times A\} \\ N(x) &= \{y \in A/(x, y) \in B\} \end{aligned} \tag{3.1}$$

Figure 3.2 shows this graph representation. As the information extracted from databases such as Cyclebase include checkpoints (G1/S and G2/M), we used them as a part of the cycle, so that the sequence remains in order. Thus, if we can follow the sequence of proteins in cell cycle, the first protein being in a certain phase, the next one can be in the same phase or in the following one.

We incorporate temporality in the form of cell cycle phases to check if two consecutive proteins in a candidate pathway are likely to participate in a pathway. If all proteins forming edges in a candidate pathway satisfy the expected transitions in the cell cycle dynamics or the protein complex involment, the *candidate pathway* is a *consistent pathway*. We model the cell cycle transition as the function 3.2.

$$T : V \mapsto C \subseteq A \tag{3.2}$$

We define function T based on cell cycle peak expression, T_{pk}, and cell cycle phenotype T_{ph}, using the information available in Cyclebase 3.0 [1]; and using the domain V as we showed in section 3.1 *Problem formulation*

In addition, we consider the protein complex involment function

$$L : V \mapsto R \subseteq PC \tag{3.3}$$

to map proteins to protein complexes where they participate, where PC are gold standard complexes.

[1] https://cyclebase.org/CyclebaseSearch

Then, given $CG(A, B)$ and $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$, we decide whether a candidate pathway P is TRUE or FALSE as shown in Algorithm 1. Consistent pathways are candidate pathways that are TRUE after applying Algorithm 1.

Algorithm 1 Signaling consistent pathway decision using cell-cycle and protein complex rules.

Input: Candidate pathway P (where $P \in \mathcal{P}$), functions T_{pk}, T_p, L.
Output: Returns TRUE (consistent pathway) if a candidate pathway satisfies the cell-cycle dynamics and protein complex rules.

1: **for** $v_i, v_{i+1} \in P$ **do**
2: **if** $(|T_{pk}(v_{i+1}) \cap N(T_{pk}(v_i))|) \vee (|T_{ph}(v_{i+1}) \cap N(T_{pk}(v_i))|) \vee (|T_{pk}(v_{i+1}) \cap N(T_{ph}(v_i))|) \vee (|T_{ph}(v_{i+1}) \cap N(T_{ph}(v_i))|)$ **then**
3: continue
4: **else if** $(|L(v_i) \cap L(v_{i+1})|)$ **then**
5: continue
6: **else**
7: **return** FALSE
8: **end if**
9: **end for**
10: **return** TRUE

The first part of the Algorithm 1 (Cell Cycle rule) evaluates whether a protein-protein interaction, (v_i, v_{i+1}) is related by:

1. Peaks of expression: where v_{i+1} peak must be in the next phase of v_i peak.
2. Peaks and phenotypes: where v_{i+1} phenotype must be in the next phase of v_i.
3. Phenotypes and peaks: where v_{i+1} peak must be in the next phase of v_i phenotype.
4. Phenotypes: where v_{i+1} phenotype must be in the next phase of v_i phenotype.

The second part (complex rule), is evaluated only when the interactions are not satisfied by the cell cycle rule. Then, the algorithm verifies whether the proteins in the interaction belong to a protein complex in common. If none of these two parts are satisfied, we discard such interaction and, subsequently, the candidate pathway.

A third part was added to the algorithm. Taking advantage of the high relation between the proteins when they collaborate in a complex, we tested all the proteins

of each *consistent pathway* to calculate the highest percentage of coverage that the complexes had on the pathways. This can be seen in Algorithm 2, where we compute the number of common protein complexes where proteins in a *consistent pathway* participate in (lines 1-3 in Algorithm 2). Then, we compute the protein complex coverage r (line 4 in Algorithm 2) and choose over a threshold given by $covPercent$. This process helps us to see the level of cohesion within the pathway, showing that certain proteins in it carry out functions together, beyond temporality.

Algorithm 2 Signaling consistent pathway with protein complex coverage.

Input: Consistent pathway P_c and $covPercent$ (protein complex coverage percentage).
Output: Returns TRUE (Predicted pathway) if a consistent pathway satisfies the protein complex coverage.

1: **for** $v_i \in P_c$ **do**
2: $Freq[L(v_i)] + +$
3: **end for**
4: $r = max(Freq)/|P_c|$
5: **if** $r >= covPercent$ **then**
6: **return** TRUE
7: **else**
8: **return** FALSE
9: **end if**

Enrichment analysis

For this analysis, we build a PPI with modified weights of the interactions, based on enrichment analysis. In order to do this, we use the same interactions from the PPI used by Gitter et al. [8] and we calculate via BINGO all the p-values for every interaction in the PPI, using biological process. After that, we replace the weight of each interaction with the $p - value$ calculated for such interaction. As the p-values between a group of proteins is better if it is lower, we use the complement of this value, that is $1 - (p - value)$ so, in this way, we can still use the same metrics for evaluation, as they take higher values as better. After we generate the new enriched PPI, we use Gitter's algorithm to generate candidate pathways and then, we evaluate them.

3.2.3 Ranking Metrics

In order to evaluate if a predicted pathway is a true positive, we consider that at least three of the five interactions are consecutive in at least one reference pathway. Thus, true positive pathways are the ones that are a complete match (i.e. those that match all five interactions in sequence) or a partial match (i.e. those that match three or four interactions in sequence). As measuring pathways are biologically meaningful, true pathways are evaluated using different ranking metrics. We divide these ranking metrics as edge and path metrics, and vertex centrality metrics. The definition of true positives and ranking metrics is the same as the one used in Gitter et al. [8], with the exception of the *betweenness* and *closeness* vertex centrality metrics which we added in our the evaluation.

Edge and path metrics

The main path and edge metrics are displayed in Equation 3.4. The *Path Weight* is the multiplication of all the edge weight values ($w(e)$) from a pathway P. The *Edge Weight* returns the minimum, average or maximun weight of an interaction in a pathway. The *Edge Use* is the number of times a certain edge is present in all predicted pathways. Path weight was used to break ties when ranking by other metrics. *Edge Weight* and *Edge Use* are considered as in their maximum, average and minimum values.

$$
\begin{aligned}
PathWeigh &= \prod_{e \in P_p} w(e) \\
EdgeWeight &= \forall_{e \in P_p} w(e) \\
EdgeUse &= \forall_{e \in P_p} Freq(e)
\end{aligned}
\tag{3.4}
$$

Vertex centrality metrics

We also consider vertex centrality metrics as in degree, betweenness and closeness as seen in Equation 3.5. *Degree* measures the number of connections that a vertex has in the graph. Equation 3.5 provides its definition, where $|N(v)|$ is the number of neighbors of vertex v. *Betweenness* measures the number of shortest paths passing through

a vertex in a graph, Equation 3.5 shows its equation, where $\sigma_{st}(v)$ is the number of paths that pass through v and σ_{st} is the total number of shortest paths, and *Closeness* measures how many steps are required to access from a vertex v all other vertices of a graph, and is defined as the inverse of the average length of the shortest paths to all the other vertices in a graph where $d(v,i)$ is the number of steps from vertex v to vertex i.

$$
\begin{aligned}
Degree(v) &= |N(v)| \\
Betweenness(v) &= \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \\
Closeness(v) &= \frac{|V| - 1}{\sum_{i \neq v}(d(v,i))}
\end{aligned}
\tag{3.5}
$$

After computing all metrics in Equations 3.4 and 3.5 for all predicted pathways, we analyse the top-k pathways, to see which of the top k predicted pathways of a selected ranking, were true positives (predicted pathways in the gold standard pathways).

3.3 Evaluation

This section describes the experimental setup used for applying the proposed method. This setup consists of the input datasets, the gold standards, evaluation metrics and alternative methods used for comparison.

3.3.1 Experimental setup

We used the input PPI network defined in Gitter et al. [8], which consists of high confident interactions driven by the union and analysis of different PPI databases such as MINT, BioGrid, and IntAct. This PPI contains 3,446 proteins and 10,944 interactions.

To obtain a set of reference pathways, we use the Gold Standard Pathways obtained by Gitter et al [8], which are real pathways extracted from KEGG and Science Signaling Database of Cell Signaling. As they stated, signaling pathways from KEGG (MAPK signaling pathway) and the Science Signaling Database of Cell Signaling (Pheromone

Source standard name	Source systematic name	Target standard name	Target systematic name
SLN1	YIL147C	CDC42	YLR229C
YCK1	YHR135C	HOG1	YLR113W
YCK2	YNL154C	STE7	YDL159W
SHO1	YER118C	STE20	YHL007C
MF(ALPHA)2	YGL089C	DIG2	YDR480W
MID2	YLR332W	DIG1	YPL049C
RAS2	YNL098C	PBS2	YJL128C
GPR1	YDL035C	FUS3	YBL016W
BCY1	YIL033C	STE5	YDR103W
STE50	YCL032W	GPA1	YHR005C
MSB2	YGR014W	MSN1	YOL116W
SIN3	YOL004W	FKS2	YGR032W
RGA1	YOR127W	FUS1	YCL027W
RGA2	YDR379W	STE12	YHR084W
ARR4	YDL100C	SWI4	YER111C
MF(ALPHA)1	YPL187W	FLO11	YIR019C

Table 3.1: List of sources and targets used

pathway and High Osmorality Glycerol pathway) contain an average of 5 edges between a source and its closest target. This is why we set L=5 in our experiments. This value was calculated by getting the shortest path from any source to each target, with a PPI with only interactions from each of the 3 pathways in evaluation. This study is useful to bound the length of the reference pathways and the predicted pathways, since the longer the pathway, the more computational resources (memory and time) are needed. In all gold standard pathways, we discarded inhibition interactions and only considered activation interactions, because inhibition interactions lead to stopping in the cascade of interactions. From the reference PPI from KEGG and Science Signaling Database of Cell Signaling (where the gold standard is made from), there were only four inhibition interactions.

We generated the reference pathways from a list of 16 sources and 16 targets chosen in the same research, as a list of vertices without a father vertex (in the case of sources) and a list of vertices without children (in the case of targets), which can be seen in Table 3.1. These sources and targets were used to generate the candidate pathways in the initial step.

Config	Description
G	Gitter's method, using the best values from 10 tests (as it is a random model, results may vary).
GcovPercent	Gitter's method plus protein complex coverage (i.e. using candidate pathways as input and protein complex coverage in Algorithm 2).
GC	Gitter for candidate pathways plus applying Cell cycle rule (defined in part Algorithm 1).
GCC	Gitter for candidate pathways, applying Cell cycle rule and protein complex rule (defined in Algorithm 1).
GCCcovPercent	Gitter for candidate pathways, applying Cell cycle rule, protein complex rule (defined in Algorithm 1), and protein complex coverage (defined in Algorithm 2).
Enrich	The same procedure as Gitter, but with a Gene Set Enrichment Analysis based PPI.

Table 3.2: Method configurations.

Also, we include biological datasets that register the peak expression and phenotypes availabe in Cyclebase [28] and yeast protein complex gold standards CYC2008 [24], which we considered as input for our pathway prediction method. We use a ranking scheme choosing the top-k predicted pathways that considers different pathway metrics. These pathway metrics are based on path and edge (in Equation 3.4) and a centrality measure (*Degree*) as used in Gitter et al. [8]. We also include *betweenness* and *closeness* centrality measures as seen in Equation 3.5.

We used as parameters, the length of the pathways to consider, $covPercent$, intended to consider different ratio of protein complex coverage, top-k, using different k values. We compared our method using different configurations, as we can see in Table 3.2. Here we include Gitter's method to see every configuration tested.

3.3.2 Results

As mentioned in previous section, we used Gitter's PPI network, path length of 5 interactions and top-k as it is used in Gitter et al. [8]. We evaluated all configurations

described in Table 3.2, using $covPercent = 0, 10, 20, 30, 40, 50$ % for all metrics defined in Equations 3.4 and 3.5. In addition, the influence of enrichment on signaling pathways was tested with the p-value complement method. Table 3.3 presents the results of the total number of true positives by sorting the predicted pathways by the path and edge metrics and taken the top-100. We observed that the best results are achieved by using Cell cycle and protein complex rules using protein complex coverage of 20% and 30% (GCCcov20% and GCCcov30%) with *Path Weight* metric, where we are able to achieve 47 true positives over the first 100 predicted pathways. We can see that the metrics with Minimum and Average Edge Weight achieve the second and third best combinations. Using only cell cycle and protein complex rules also provide good quality, but not as good as using protein complex coverage. Furthermore, we observed that using centrality metrics do not perform as well as edge and path metrics as seen in Table 3.4. In any case, we can see that using cell cycle and protein rules as well as protein complex coverage are better alternatives than using only the random edge orientation with local search proposed by Gitter et al. [8], since the true positives increment between 21% and 27%. The results with the enriched PPI are similar in every edge and path metric, but the performance is lower than the other configurations, staying below Gitter's results. In addition, the influence of enrichment on signaling pathways was tested with the p-value complement method.

We also measure true positives for different number of top-k predicted pathways. Figure 3.3 shows the number of true positive pathways for a range between 10 and 100 top pathways. We observed that all our combined configurations of our method provides much better results than the pathways predicted by Gitter [8], and for small number of top pathways, between 10 and 50, we obtain a good number of true positives, like 8 for the top-10 and 31 for the top-50. Figure 3.4 includes the results for the top-500 pathways, where we observed that our method keeps improving over Gitter, and the best results are obtained when a coverage percentage between 0% and 30%. Also, it can be seen that with $k < 100$, Random + local (Gitter's approach) search is lower in the number of true positives than our method, but with $k > 100$ our method (between the coverage 0% to 30%) almost doubles the matches that are achieved by Gitter et al. [8].

Method	Path Weight	Max Edge Weight	Avg Edge Weight	Min Edge Weight	Max Edge Use	Avg Edge Use	Min Edge Use
G	31	7	31	**34**	0	0	0
Gcov10%	31	7	31	**34**	0	0	0
Gcov20%	32	7	32	**37**	0	1	8
Gcov30%	32	7	32	**37**	0	0	8
Gcov40%	**28**	3	**28**	26	0	0	0
Gcov50%	**28**	3	**28**	26	0	0	0
GC	41	12	**43**	36	7	4	3
GCC	**45**	17	44	42	6	2	0
GCCcov10%	**45**	17	44	42	6	2	0
GCCcov20%	**47**	17	45	46	0	1	5
GCCcov30%	**47**	17	45	46	0	1	5
GCCcov40%	40	10	**42**	41	0	0	0
GCCcov50%	40	10	**42**	41	0	0	0
Enrich	27	**24**	26	29	17	22	26

Table 3.3: Ranking results (path and edge measures) from the top100 consistent signaling pathways.

Method	Max Degree	Avg Degree	Min Degree	Max BETW	Avg BETW	Min BETW	Max Closeness	Avg Closeness	Min Closeness
G	3	0	0	4	1	**13**	7	0	0
Gcov10%	3	0	0	**7**	5	7	7	3	0
Gcov20%	3	0	0	**7**	5	7	7	3	0
Gcov30%	3	0	0	**7**	5	7	7	3	0
Gcov30%	3	0	0	**7**	5	7	7	3	0
Gcov40%	1	0	0	**7**	5	7	7	3	0
Gcov50%	1	0	0	**7**	5	7	7	3	0
GC	9	0	0	4	4	**13**	4	0	0
GCC	7	0	0	4	4	**13**	4	0	0
GCCcov10%	7	0	0	4	4	**13**	4	0	0
GCCcov20%	7	0	0	4	4	**13**	4	0	0
GCCcov30%	7	0	0	4	4	**13**	4	0	0
GCCcov40%	1	0	0	4	4	**13**	4	0	0
GCCcov50%	1	0	0	4	4	**13**	4	0	0
Enrich	17	15	**19**	2	2	10	2	0	0

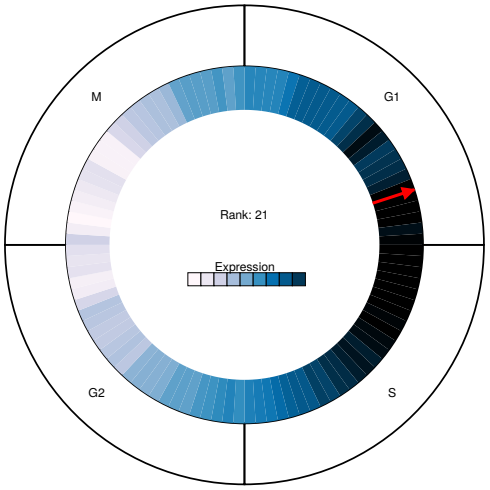
Table 3.4: Ranking results (vertex centrality measures) from the top100 consistent signaling pathways.

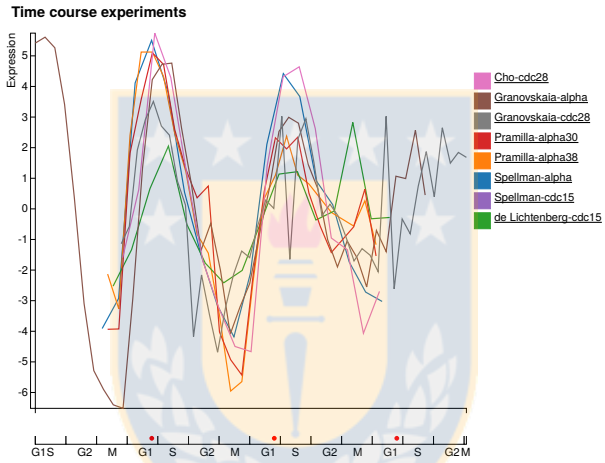
Figure 3.5 shows the best results using Gcov20% and Gcov30% (for Random edge orientation and random search) and our method with Gcov20% and GCCcov30% (Random edge orientation with random search and cell cycle and protein complex rules). We also observed that, when the percentage of coverage increases over 30%, the number of predicted pathways is decreased. Results from a test with coverage Percentage over 50% are not shown as the set of predicted pathways falls abruptly to 8, so we can not measure the matches with $k < 5$ (as we do with the rest of the experiments).

As an example, Figure 3.6 displays the top20 predicted pathways. There are 4 complete matched predicted pathways in the top20 (Figure 3.6a) and there are 13 partially matched pathways in the top 20 (Figure 3.6b). All of them share the same source protein (YCL032W) but they have different targets. We observe that, the interactions from YLR362W to YDR103W and from YDR103W to YDL159W, are the most repeated in the 13 predicted pathways from Figure 3.6b, which can be an indicator of a high level of certainty.

3.4 Visualization

In order to help the analysis of predicted pathways, we developed a visualization tool that processes the predicted pathways and displays them as a directed graph. This tool was developed using visNetwork package in R, and displays the pathways in HTML format, so it can be seen through any browser. As it can be seen in Figure 3.7, it can distinguish the reference from the predicted pathways, displaying true interactions in blue and interactions that are not in the gold standard in red. Also, users can select for highlighting individual proteins, or a group of proteins having a common attribute. Other allowed actions are zooming in and out, moving proteins (vertices) of place and show information about each protein.





Matched name	Primary name	Type	Identifier	Peaktime	rank	Phenotypes
POL30	POL30	Saccharomyces cerevisiae protein	YBR088C	G1	21	
MCD1	MCD1	Saccharomyces cerevisiae protein	YDL003W	G1	22	
HTA2	HTA2	Saccharomyces cerevisiae protein	YBL003C	S	23	
SVS1	SVS1	Saccharomyces cerevisiae protein	YPL163C	G1/S	24	
CLN2	CLN2	Saccharomyces cerevisiae protein	YPL256C	G1	25	G1
HTB1	HTB1	Saccharomyces cerevisiae protein	YDR224C	S	26	
MRC1	MRC1	Saccharomyces cerevisiae protein	YCL061C	G1	27	G1/S, S

Figure 3.1: Capture from Cyclebase 3.0. Information available for protein YBR088C (top), including temporal expression data from eight experiments (middle) with normalized results of all experiments in every part of the cycle, and a set of genes with corresponding peak expression phase and phenotype (bottom). Source: www.cyclebase.org

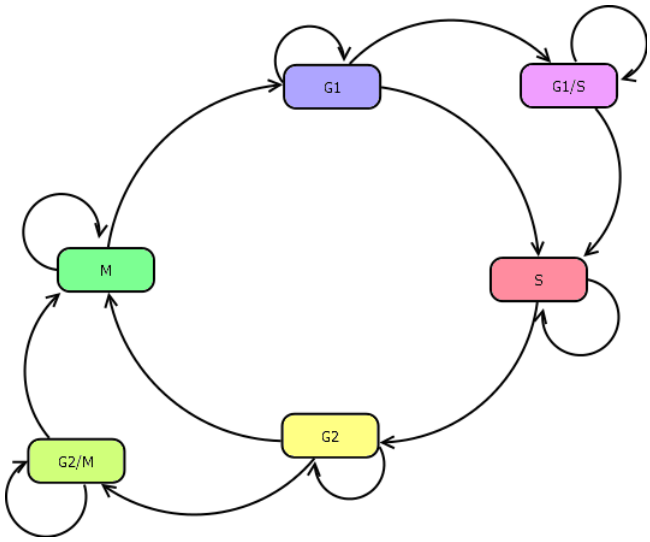


Figure 3.2: Graph represetantion of a cell cycle $CG(A, B)$. Source: Prepared by the authors.

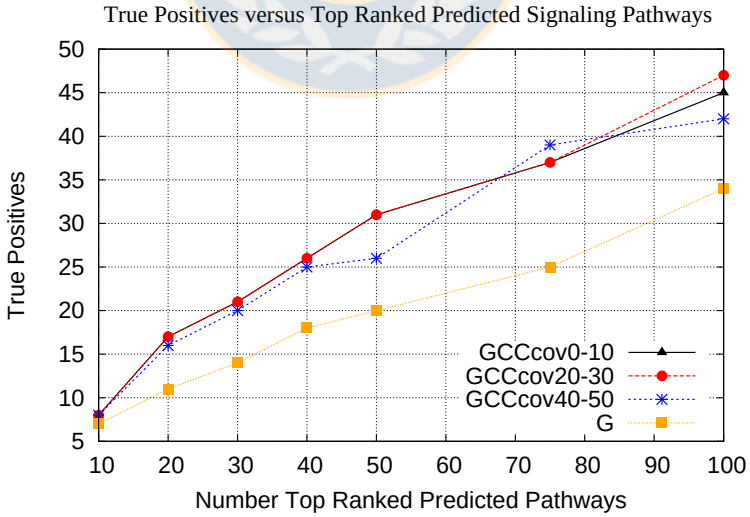


Figure 3.3: Variation of the first top-100 by principal experiments. Source: Prepared by the authors.

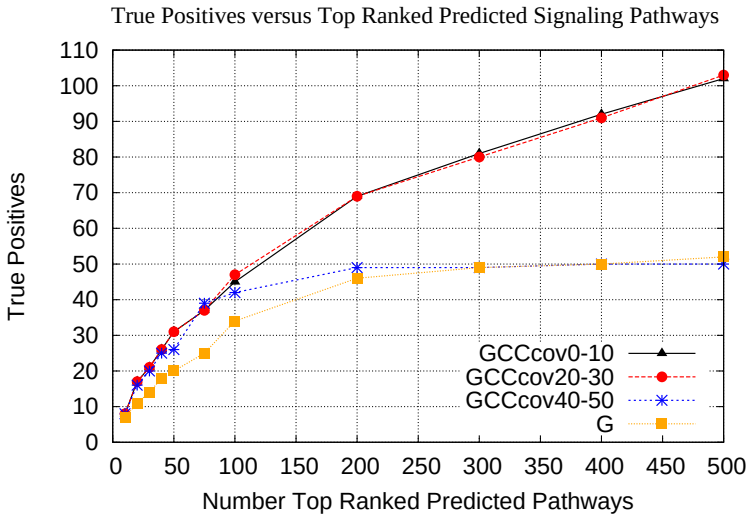


Figure 3.4: Variation of k in top-k by principal experiments. Source: Prepared by the authors.

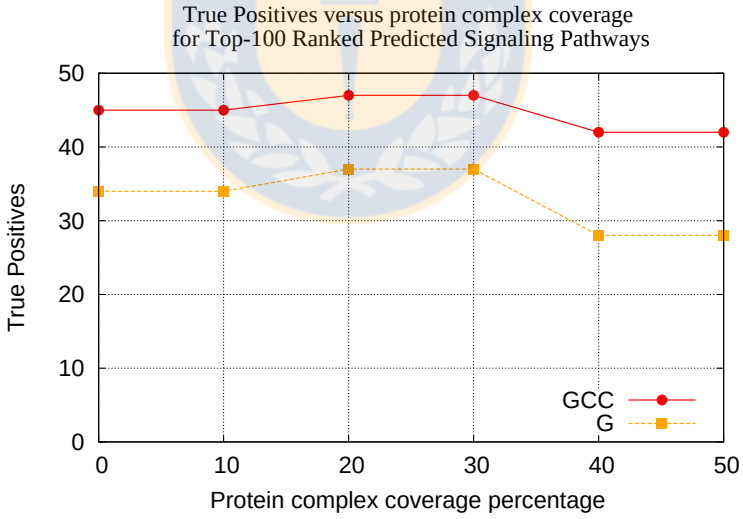


Figure 3.5: Variation in percentage of coverage for k=100 for Random + local search (Gitter) and GCCcovPercent. Source: Prepared by the authors.

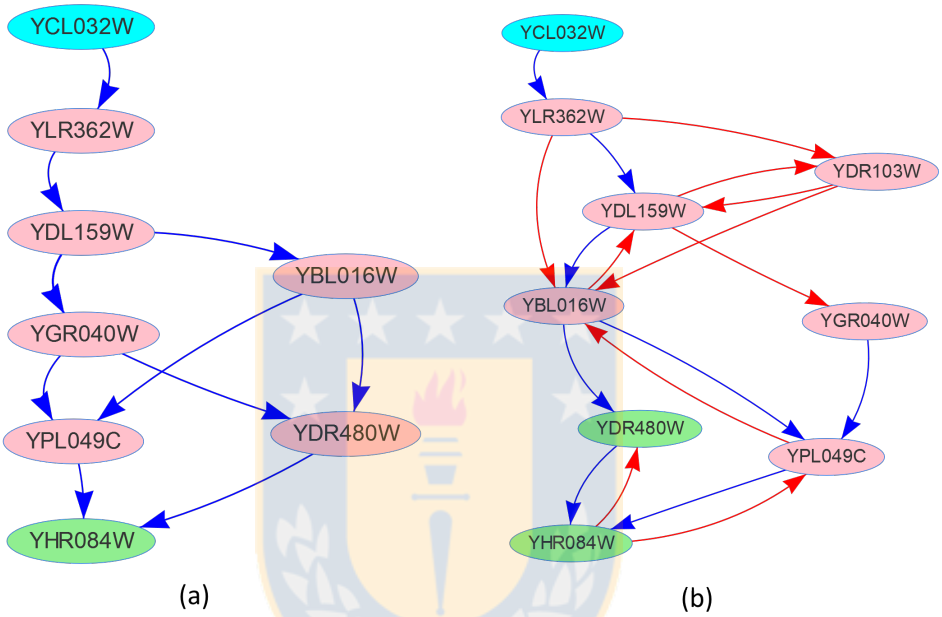


Figure 3.6: Top20 pathways from our GCCov20% method (a) Pathways completely matched (b) Pathways partially matched, with 3 or more interactions right. In red are interactions from partially matched pathways and in blue are interactions from completely matched pathways. Source: Prepared by the authors.

.

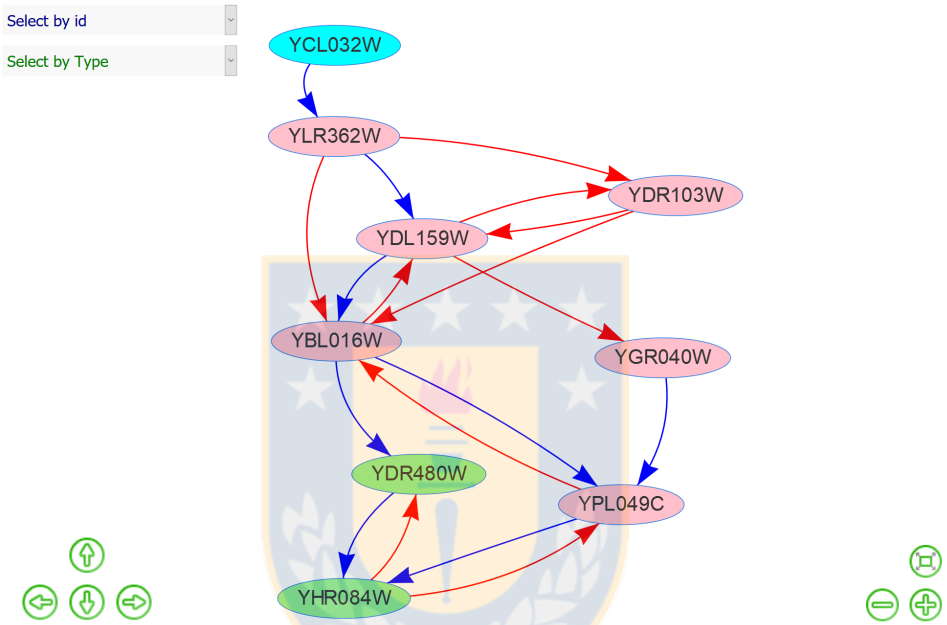


Figure 3.7: Caption of the visualization in a browser, where it is shown a reference pathway and predicted pathways (the same from Figure 3.6b). The reference pathway it is shown with blue edges and the red ones belong to predicted pathways. Source: Prepared by the authors.

Chapter 4

CONCLUSIONS AND FUTURE WORK

We study whether if adding biological knowledge to predict signaling pathways improves the results of matched predicted pathways over reference pathways. We showed that the most useful information is about cell cycle and protein complexes. As the use of cell cycle information was shown with the peaks of expression and phenotypes, it may be a good idea to test the cell cycle with a greater level of detail. That is, instead of using the expression peaks, all the experimental data of the proteins could be used to determine all the phases where the proteins express the most. This does not limit to a single peak, but considers all phases in which the protein has high levels of expression. This study should consider that the experiments have been carried out under different parameters, so that the expression values are not directly comparable. This can be managed by normalizing the values according to each type of experiment and they should be aligned onto a common time-scale (in percent of the cell cycle, where each phase represents a 25% out of a 100% that is the complete cell cycle) [7]. This level of detail, can help to discriminate the pathways in a more specific way, having more precise consistent pathways according to the biological behavior. Although the performance of our algorithm is better than other methods, there is still a task to be solved in this area, because the rate of true positives (47 out of 100) is still low for a prediction algorithm. A future implementation of our work can be tested for human PPIs and pathways. This could need more information than what it is in Cyclebase because the proteins of humans do not behave in the same way as those of yeast. In yeast, the behavior of proteins is periodic in the cycle, however in humans that is not always true, so the behavior of each human protein does not necessarily repeat itself in each cycle, which makes it difficult to detect expression peaks. Within the information that can be used for human testing there are protein databases related to diseases. There

are several databases that show which proteins influence certain diseases. These relations are important because it is the key to understand diseases medically and for drug development to treat them [23].

Bibliography

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[2] M Barton, R Marecek, I Rektor, P Filip, E Janousova, and M Mikl. Sensitivity of ppi analysis to differences in noise reduction strategies. *Journal of neuroscience methods*, 253:218–232, 2015.

[3] Mengfei Cao, Christopher M Pietras, Xian Feng, Kathryn J Doroschak, Thomas Schaffner, Jisoo Park, Hao Zhang, Lenore J Cowen, and Benjamin J Hescott. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*, 30(12):i219–i227, 2014.

[4] J Michael Cherry, Caroline Adler, Catherine Ball, Stephen A Chervitz, Selina S Dwight, Erich T Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, et al. Sgd: Saccharomyces genome database. *Nucleic acids research*, 26(1):73–79, 1998.

[5] Kalyanmoy Deb. An introduction to genetic algorithms. *Sadhana*, 24(4-5):293–315, 1999.

[6] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[7] Nicholas Paul Gauthier, Malene Erup Larsen, Rasmus Wernersson, Ulrik De Lichtenberg, Lars Juhl Jensen, Søren Brunak, and Thomas Skøt Jensen. Cyclebase. orga comprehensive multi-organism online database of cell-cycle experiments. *Nucleic acids research*, 36(suppl_1):D854–D859, 2007.

[8] Anthony Gitter, Judith Klein-Seetharaman, Anupam Gupta, and Ziv Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. *Nucleic acids research*, 39(4):e22–e22, 2010.

[9] Jane V Harper. Synchronization of cell populations in g1/s and g 2/m phases of the cell cycle. *Cell Cycle Control: Mechanisms and Protocols*, pages 157–166, 2005.

[10] Cecilia Hernandez, Carlos Mella, Gonzalo Navarro, Alvaro Olivera-Nappa, and Jaime Araya. Protein complex prediction via dense subgraphs and false positive analysis. *PloS one*, 12(9):e0183460, 2017.

[11] Lun Hu and Keith CC Chan. A density-based clustering approach for identifying overlapping protein complexes with functional preferences. *BMC bioinformatics*, 16(1):174, 2015.

[12] Jui-Hung Hung. Gene set/pathway enrichment analysis. *Data Mining for Systems Biology: Methods and Protocols*, pages 201–213, 2013.

[13] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[14] Junzhong Ji, Aidong Zhang, Chunnian Liu, Xiaomei Quan, and Zhijun Liu. Survey: Functional module detection from protein-protein interaction networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):261–277, 2014.

[15] Shingo Kikugawa, Kensaku Nishikata, Katsuhiko Murakami, Yoshiharu Sato, Mami Suzuki, Md Altaf-Ul-Amin, Shigehiko Kanaya, and Tadashi Imanishi. Pcdq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset. *BMC systems biology*, 6(2):S7, 2012.

[16] Daniel C Kirouac, Julio Saez-Rodriguez, Jennifer Swantek, John M Burke, Douglas A Lauffenburger, and Peter K Sorger. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC systems biology*, 6(1):29, 2012.

[17] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[18] Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R Sinha, Ryan Miller, et al. Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic acids research*, 44(D1):D488–D494, 2015.

[19] Hans-Werner Mewes, Dmitrij Frishman, Ulrich Güldener, Gertrud Mannhaupt, Klaus Mayer, Martin Mokrejs, Burkhard Morgenstern, Martin Münsterkötter, Stephen Rudd, and B Weil. Mips: a database for genomes and protein sequences. *Nucleic acids research*, 30(1):31–34, 2002.

[20] Hoai Anh Nguyen, Cong Long Vu, Minh Phuong Tu, and Thu Lam Bui. Discovery of pathways in protein–protein interaction networks using a genetic algorithm. *Data & Knowledge Engineering*, 96:19–31, 2015.

[21] Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4(1):10, 2011.

[22] Marco Pellegrini, Miriam Baglioni, and Filippo Geraci. Protein complex prediction for large protein protein interaction networks with the core&peel method. *BMC bioinformatics*, 17(12):372, 2016.

[23] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, 45(D1):D833–D839, 2017.

[24] Shuye Pu, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana J Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research*, 37(3):825–831, 2008.

[25] Anna Ritz, Brendan Avent, and T Murali. Pathway analysis with signaling hypergraphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2015.

[26] Andreas Ruepp, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Michael Stransky, Brigitte Waegele, Thorsten Schmidt, Octave Noubibou Doudieu, Volker Stümpflen, et al. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(suppl_1):D646–D650, 2007.

[27] Sarah A Sam, Joelle Teel, Allison N Tegge, Aditya Bharadwaj, and TM Murali. Xtalkdb: a database of signaling pathway crosstalk. *Nucleic acids research*, 45(D1):D432–D439, 2017.

[28] Alberto Santos, Rasmus Wernersson, and Lars Juhl Jensen. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic acids research*, 43(D1):D1140–D1144, 2014.

[29] Yu-Keng Shih and Srinivasan Parthasarathy. A single source k-shortest paths algorithm to infer regulatory pathways in a gene network. *Bioinformatics*, 28(12):i49–i58, 2012.

[30] Sriganesh Srihari and Hon Wai Leong. Temporal dynamics of protein complexes in ppi networks: a case study using yeast cell cycle dynamics. *BMC bioinformatics*, 13(17):S16, 2012.

[31] Sriganesh Srihari, Chern Han Yong, Ashwini Patil, and Limsoon Wong. Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes. *FEBS letters*, 589(19):2590–2602, 2015.

[32] Arunachalam Vinayagam, Jonathan Zirin, Charles Roesel, Yanhui Hu, Bahar Yilmazel, Anastasia A Samsonova, Ralph A Neumüller, Stephanie E Mohr, and Norbert Perrimon. Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nature methods*, 11(1):94–99, 2014.

[33] Xiao-Fei Zhang, Dao-Qing Dai, Le Ou-Yang, and Hong Yan. Detecting overlapping protein complexes based on a generative model with functional and topological properties. *BMC bioinformatics*, 15(1):186, 2014.