Protein-Protein Interaction Confidence Assessment and Network Clustering Computational Analysis

Mathieu Lavallée-Adam

Doctor of Philosophy

School of Computer Science

McGill University

Montréal, Québec

August 2013

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

For Martine Lavallée, Roger Adam, and Fleurette Daigle.

ACKNOWLEDGEMENTS

I thank my parents for their love and support throughout my academic path. I am grateful to Justine Delisle for being my source of strength, for her everlasting understanding, and most of all, for her love. I could not have been supervised by better people than Mathieu Blanchette who made me discover the fascinating world of bioinformatics and will forever be a mentor in all aspects of my life and Benoit Coulombe whose help, teaching, and advising was invaluable to the accomplishment of my Ph.D. Finally I thank my Ph.D. progress committee for helpful feedback, and labmates who contributed in significant ways to the successful completion of my degree and shared amusing moments with me.

ABSTRACT

Protein-protein interactions represent a crucial source of information for the understanding of the biological mechanisms of the cell. In order to be useful, high quality protein-protein interactions must be computationally extracted from the noisy datasets produced by high-throughput experiments such as affinity purification. Even when filtered protein-protein interaction datasets are obtained, the task of analyzing the network formed by these numerous interactions remains tremendous. Protein-protein interaction networks are large, intricate, and require computational approaches to provide meaningful biological insights. The overall objective of this thesis is to explore algorithms assessing the quality of protein-protein interactions and facilitating the analysis of their networks. This work is divided into four results: 1) a novel Bayesian approach to model contaminants originating from affinity purifications, 2) a new method to identify and evaluate the quality of protein-protein interactions independently in different cell compartments, 3) an algorithm computing the statistical significance of clusterings of proteins sharing the same functional annotation in protein-protein interaction networks, and 4) a computational tool performing sequence motif discovery in 5' untranslated regions as well as evaluating the clustering of such motifs in protein-protein interaction networks.

ABRÉGÉ

Les interactions protéine-protéine représentent une source d'information essentielle à la compréhension des divers méchanismes biologiques de la cellule. Cependant, les expériences à haut débit qui identifient ces interactions, comme la purification par affinité, produisent un très grand nombre de faux-positifs. Des méthodes computationelles sont donc requises afin d'extraire de ces ensembles de données les interactions protéine-protéine de grande qualité. Toutefois, même lorsque filtrés, ces ensembles de données forment des réseaux très complexes à analyser. Ces réseaux d'interactions protéine-protéine sont d'une taille importante, d'une grande complexité et requièrent des approches computationelles sophistiquées afin d'en retirer des informations possédant une réelle portée biologique. L'objectif de cette thèse est d'explorer des algorithmes évaluant la qualité d'interactions protéine-protéine et de faciliter l'analyse des réseaux qu'elles composent. Ce travail de recherche est divisé en quatre principaux résultats: 1) une nouvelle approche bayésienne permettant la modélisation des contaminants provenant de la purification par affinité, 2) une nouvelle méthode servant à la découverte et l'évaluation de la qualité d'interactions protéine-protéine à l'intérieur de différents compartiments de la cellule, 3) un algorithme détectant les regroupements statistiquement significatifs de protéines partageant une même annotation fonctionnelle dans un réseau d'interactions protéine-protéine et 4) un outil computationel qui a pour but la découverte de motifs de séquences dans les régions 5' non traduites tout en évaluant le regroupement de ces motifs dans les réseaux d'interactions protéine-protéine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD: Activating Domain

AP: Affinity Purification

AP-MS: Affinity Purification Coupled to Mass Spectrometry

AP-MS/MS: Affinity Purification Coupled to Tandem Mass Spectrometry

CTD: Carboxyl-Terminal Domain

DBD: DNA Binding Domain

DNA: Deoxyribonucleic Acid

GO: Gene Ontology

GSEA: Gene Set Enrichment Analysis

GST: Glutathione-S-Transferase

LC: Liquid Chromatography

LESMoN: Local Enrichment of Sequence Motifs in Biological Networks

MCC-AP-MS/MS: Multiple Cell Compartment Affinity Purification Coupled to Tandem Mass Spectrometry

mChIP: Modified Chromatin Immunopurification

MCL: Markov Clustering

MCODE: Molecular Complex Detection

mRNA: Messenger RNA

MS: Mass Spectrometry

MS/MS: Tandem Mass Spectrometry

PCA: Protein Fragment Complementation Assay

PPI: Protein-Protein Interaction

PSM: Peptide-Spectrum Match

PTM: Post-Translational Modification

RBP: RNA Binding Protein

RNA: Ribonucleic Acid

RNAP II: RNA Polymerase II

SAINT: Significance Analysis of Interactome

TAP: Tandem Affinity Purification

UAS: Upstream Activating Sequence

UTR: Untranslated Region

WCE: Whole Cell Extract

Y2H: Yeast-Two-Hybrid

CHAPTER 1
Introduction

Most biological processes taking place in the living cell involve protein-protein interactions (PPIs). This thesis introduces computational tools assessing the quality of experimentally obtained PPIs. It then presents computational analyses of the networks formed by these PPIs with the goal of improving the understanding of various biological mechanisms occurring in the cell. In this chapter, background information about PPIs, methodologies used to study them, and classic computational approaches analyzing these interactions are discussed.

1.1 Central dogma of molecular biology

The central dogma of molecular biology states that the biological information of the cell is stored in three different types of molecules: deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins [43]. The information is originally stored in the DNA (genome) of the cell. DNA, with the assistance of certain helper proteins, can replicate itself in the genome. However, its main purpose in the cell is to be copied, using a mechanism called transcription, into RNA (Figure 1–1). Transcription is generally regulated by DNA binding proteins (transcription factors). Some transcribed RNA molecules called messenger RNAs (mRNAs) are then spliced and finally translated into proteins (Figure 1–1). mRNAs are also often transported before protein translation to a given cell compartment by RNA binding proteins (RBPs) recognizing a specific nucleotide motif in their sequences. Some RNAs (microRNAs) can regulate other RNAs through degradation, sequestering, or translational repression. Newly translated proteins then interact with other proteins, RNA, or DNA molecules to perform their function(s). The major steps in the biological information transfer described above (transcription, splicing, and translation) are performed by proteins

sometimes associated with RNAs (e.g. RNA polymerase, spliceosome, and ribosome). Proteins can also play a regulation role by repressing and degrading both other proteins and RNAs (e.g. proteasome and ribonucleases). Finally, certain proteins such as methyltransferases and kinases will modify other proteins by adding small molecules to one of their specific amino acid (post-translational modification (PTM)).

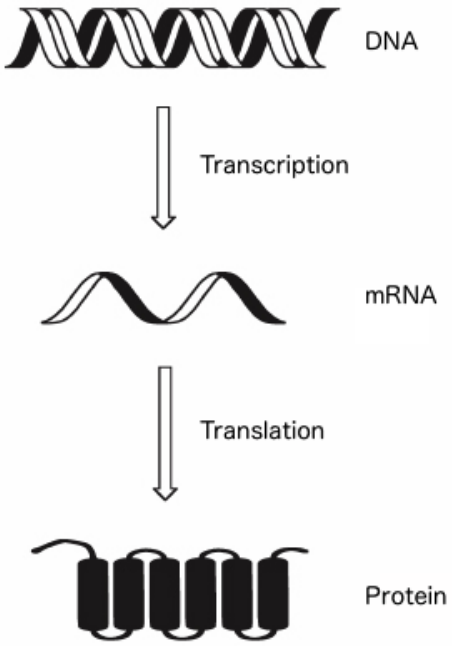

Figure 1–1: Central dogma of molecular biology (adapted from a figure from an article by David A. Omahen [167]).

1.2 Protein-protein interactions

In order to perform their various functions in the cell, proteins almost always interact with each other. Through these PPIs, proteins can for example fold other proteins, form protein complexes, and perform PTMs. However, even though a large fraction of yeast and human proteins have been observed in different experimental setups, very little is known about their respective interactions. Furthermore, the function of the vast majority of these proteins remains unknown, not to mention that numerous proteins perform multiple functions. Therefore, proteins that were thought to be well characterized, are still today associated to novel biological functions. PPIs

can reveal much about protein functions especially when these are analyzed as a network. If multiple proteins interact together and accomplish the same function in the cell and another uncharacterized protein interacts with these, then it is likely that this uncharacterized protein is involved in the same process ("guilt-by-association" principle) [166].

1.2.1 Types of protein-protein interactions

There are several several types of PPIs. First are the interactions forming protein complexes. Protein complexes are groups of two or more proteins interacting together non-covalently over a certain period of time to perform a certain biological function. An example of such PPIs are those forming the RNA polymerase II protein complex. This complex is formed of 12 proteins (subunits) in human and yeast: POLR2A, POLR2B, POLR2C, POLR2D, POLR2E, POLR2F, POLR2G, POLR2H, POLR2I, POLR2J (itself formed of three subunits in human), POLR2K, and POLR2L [160] (see Figure 1–2). These subunits interact directly or indirectly to form the RNA polymerase II complex. While the main transcriptional role of this complex is well understood, the mechanisms and co-factors mediating its assembly and import into the cell nucleus still remain a source of debate [17, 86]. Our own work sheds some light on this matter [37, 69, 70].

There are many more protein complexes in the cell performing a vast array of functions such as splicing (spliceosome), translation (ribosome), proteolysis (proteasome), transcriptional coactivation (mediator). They also vary in complexity. While some complexes contain only two subunits (CAP350 and FOP form a centrosomal complex required for microtubules anchoring [229]), some such as the ribosome can be much larger (79 subunits) [227].

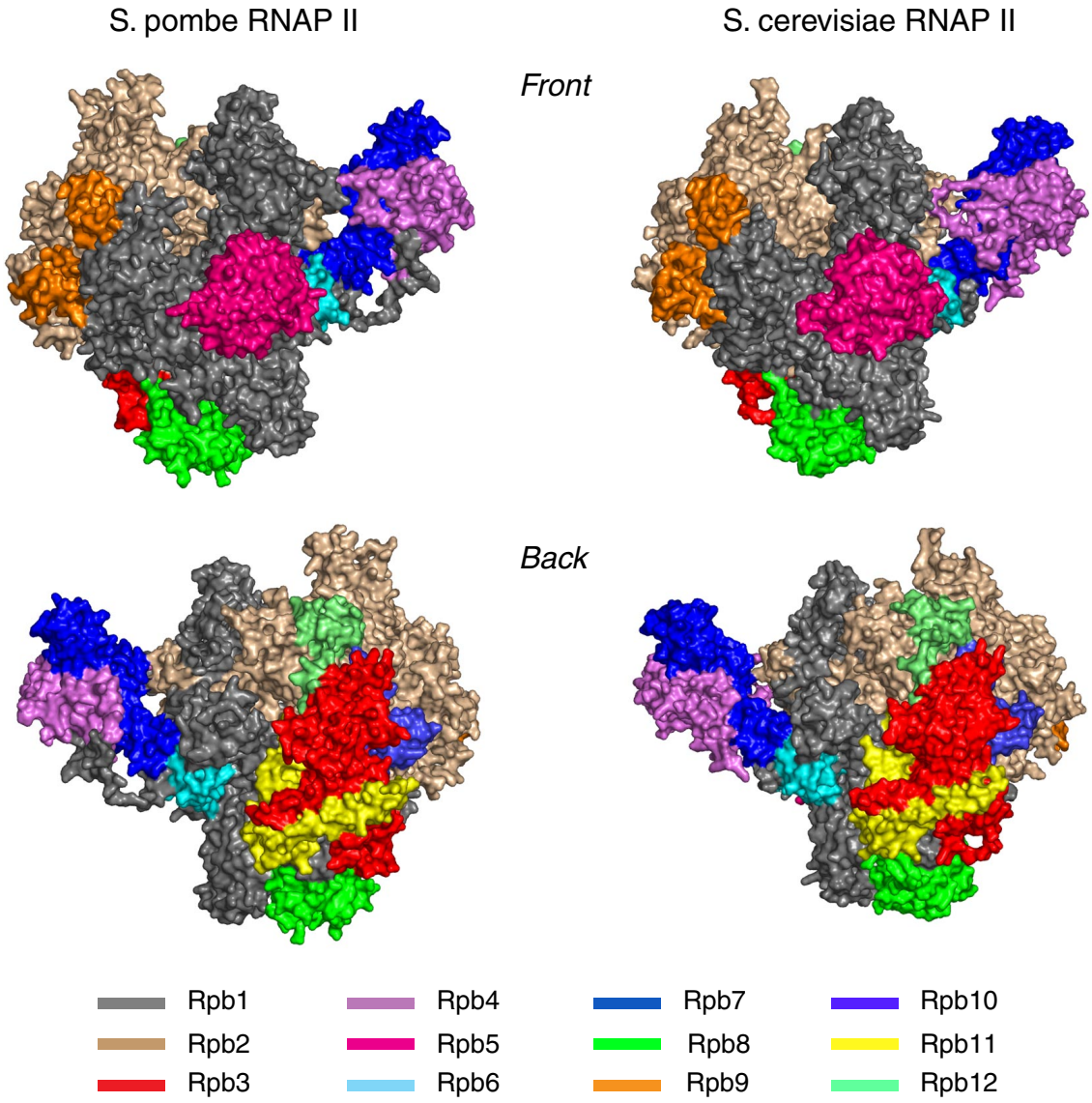

Figure 1–2: Surface representation of the structures of RNA polymerase II in *S. pombe* and *S. cerevisiae*. Each protein (Rpb1, ..., Rpb12) is color-coded and corresponds to a human protein homolog (from an article by Spåhr et al. [207]).

Another type of interaction occurs when a protein binds very briefly another to modify one of its specific amino acids by adding or removing a molecule (e.g. phosphate, methyl, or acetyl groups). Reactions like methylation, acetylation, phosphorylation and their counter parts (e.g. demethylation) are crucial regulation mechanisms. Such modification on an amino acid may repress the action of a protein, change its folding, and even make it gain or lose the ability to interact with other proteins. For

example, kinases are proteins that will bind other proteins to add a phosphate group on the amino acid of its target. For instance, PKA is a well characterized kinase involved among other things in the regulation of glycogen, sugar, and lipid metabolism that targets CREB [79]. Such proteins will usually recognize a given sequence motif on their target in order to bind them and perform their catalytic action.

A protein can also bind to another to transport it to a specific cell compartment. Some of these proteins are called importins, which are part of a large family of proteins named karyopherins. They bind temporarily certain proteins in order to transport them from the cytoplasm of the cell into its nucleus. These importins will bind to the nuclear localization signal (a specific sequence of amino acids of a protein) of their targets and will transport the targets to the nucleus [80].

These are only some of the major types of PPIs that can be observed in the living cell. Often PPI studies will focus on a given type of interactions or on a given biological process that will involve various types of interactions. Either way, all these types of PPIs play in one way or another a crucial role in various biological processes and are often essential to the survival of the cell.

1.2.2 Challenges in protein-protein interaction identification

Still today a large part (80-90%) of the human interactome (i.e. the set of human PPIs) remains unknown [220]. The intrinsic differences between interaction types make some harder to identify than others. Obviously, transient interactions such the ones implicated in PTMs are difficult to map and their detection requires very sensitive approaches. Similarly, proteins that are part of a protein complex but only for a brief moment to form an intermediary structure necessary for the proper complex assembly are also hard to detect. In addition to the short lifespan of certain interactions, some PPIs are PTM-dependent. Since PTMs are highly dynamic, identification of such PPIs is challenging [138]. Often the same protein will have

different PTM isoforms with very different stochiometries. The identification of PPIs specific to isoforms of lower abundances will therefore also be demanding. Different experimental methods used to identify PPIs will be introduced in Section 1.3.

1.2.3 Clinical implications

Several PPIs are known to be associated with various diseases [187]. Even though the exact PPIs involved in a particular disease are often unknown, the lack of expression, the over-expression, or the mutation of a given protein or set of proteins are likely to change their PPI profiles and cause this disease. The vast majority of the time, such abnormality will cause the protein to lose or gain interactions, which will disrupt the normal mechanisms of the cell. A classic example was the discovery of the implication of an abnormal interaction between *Htt* and a GTPase-activating protein GIT1 [78], which causes *Htt* to aggregate in insoluble neuronal inclusion bodies leading to neuronal degeneration and Huntington's disease [147]. This discovery was possible thanks to the identification of PPIs taking place in Huntington's disease cases. GIT1 was later validated as a potential drug target for Huntington's patients [54]. In this example a gain of a PPI was deleterious, but often a loss of a PPI will result in an abnormal functionality and therefore in a disease status. For example, our group has recently shown that VCP mutants (R155H, R159G, and R191Q) were not methylated at Lysine 315 by METTL21D as the wild type normally is [38]. These mutants are known to cause inclusion body myopathy with Paget's disease of bone and frontotemporal dementia and familial amyotrophic lateral sclerosis [38]. These examples show the importance of PPI identification in human health research. The discovery of novel disease-associated PPIs remains a very active field of research.

1.3 Identification of protein-protein interactions

Several moderate to high-throughput methods have been proposed to identify PPIs in a given organism, each with its advantages and its drawbacks. The following

sections will introduce the most popular approaches mapping PPIs, which are still being used at the time of submission of this thesis.

1.3.1 Yeast-two-hybrid

A method widely used to identify PPIs is Yeast-Two-Hybrid (Y2H) [64]. The basic idea behind this approach is to fuse two halves or domains of a transcription factor onto two candidate interacting proteins typically using yeast as the host organism. When those two domains come into close physical proximity (i.e. when the two proteins are interacting), the transcription of a reporter gene is activated. The expression of the reporter gene can be observed by a resistance to a chemical or the emission of a fluorescence under a certain type of light (see Figure 1–3). More precisely, one of the candidate interacting proteins (bait) is fused to the DNA binding domain (DBD) of the transcription factor, while the other (prey) is fused to the activating domain (AD). The DNA binding domain binds an upstream activating sequence (UAS) in the reporter gene promoter and if the prey interacts with the bait, the activating domain will come in close enough proximity that the transcription of the reporter gene will be induced. Nowadays, assays are built so that for a given bait, thousands of proteins are tested as potential interactors in a short time and at a low cost. Large scale PPI networks can therefore be mapped using such technology by repeating the procedure for multiple protein of interests (baits) [231, 186, 96, 191, 224]. This method however suffers from several drawbacks including the necessity of the interaction to happen in the nucleus to be detectable. Moreover, Y2H assays often require the simultaneous over-expression of the fusion proteins potentially causing interactions not occurring under normal *in vivo* conditions. The reporter gene fused to the candidate proteins can be fairly large and therefore physically inhibit the interactions between the candidate proteins by blocking their docking sites or changing their folding conformations. Finally, testing for candidate PPIs originating from organisms different than yeast may be problematic, since yeast might be lacking for example the chaperones ensuring the proper folding of the proteins in question and prevent them from interacting.

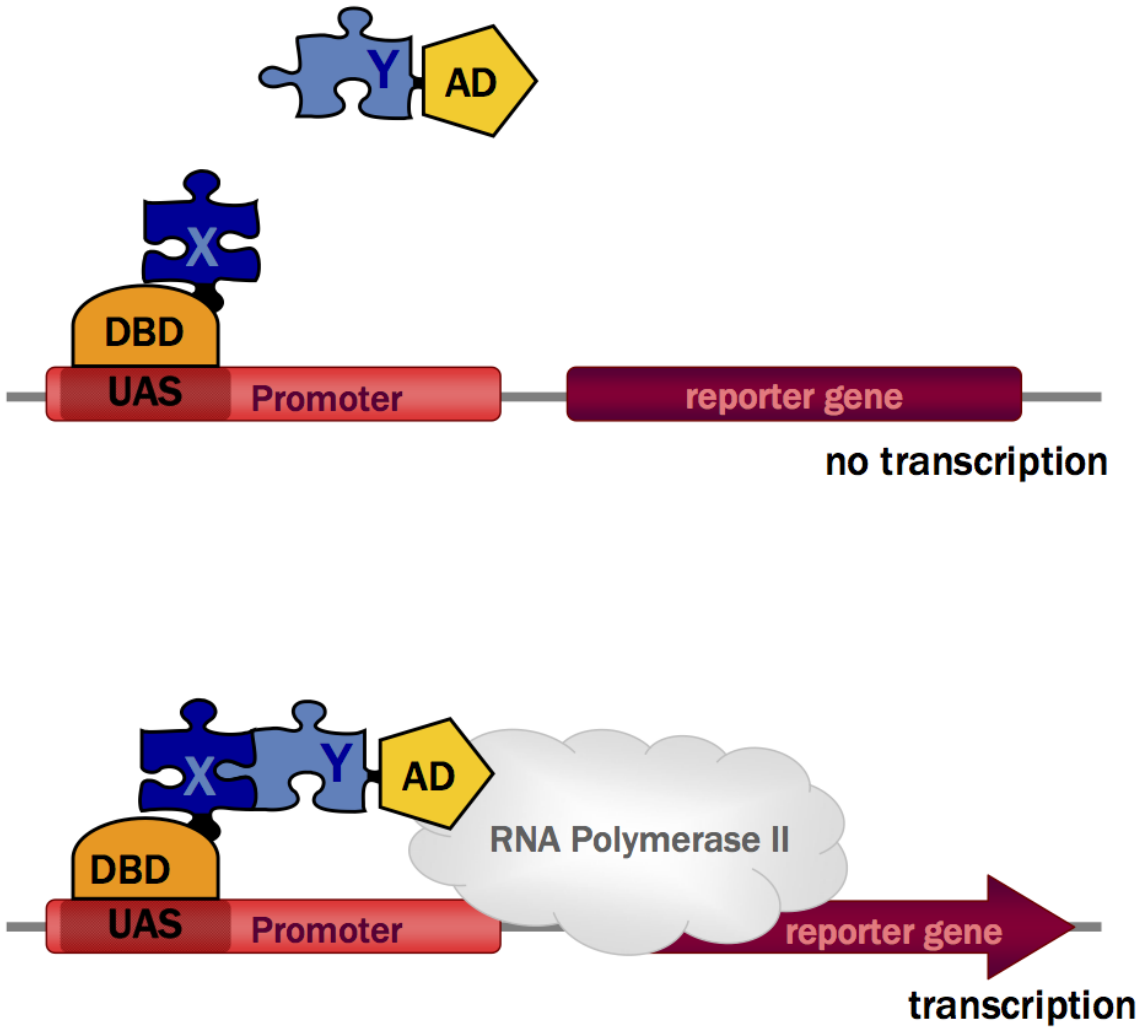

Figure 1–3: Graphical representation of the Y2H protocol (from an article by Brückner et al. [25]).

1.3.2 Protein fragment complementation assay

Protein fragment complementation assay (PCA) [72] is a technology that uses a idea similar to Y2H. Basically, candidate proteins are each covalently linked to a fragment of a reporter protein. When the candidates come into close proximity, the reporter protein becomes functional (see Figure 1–4). This reporter can take several forms such as: β-lactamase [72], dihydrofolate reductase (*DHFR*) [215], luciferase [28], and many more, each with its advantages and drawbacks. For instance DHFR

confers to its hosts a resistance to an antibiotic called trimethoprim. In this case, if the interaction occurs, the cells will survive upon treatment. Like Y2H, PCA experiments can be performed at a genome-wide scale, have low cost, and have been used to map large scale PPI networks [215]. In addition, PCA can typically identify PPIs with proteins being expressed at their endogenous level, hence limiting the number of false positive interactions. It however shares the disadvantage with Y2H that the reporter gene fragments may prevent the proper binding of the candidate interacting proteins.

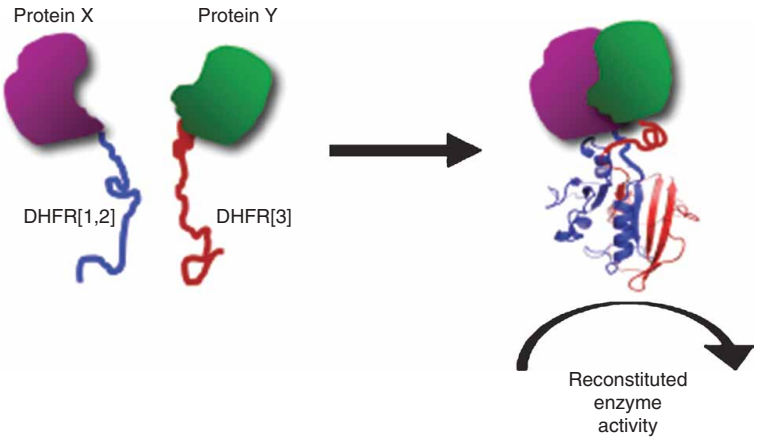

Figure 1–4: Graphical representation of the PCA method (from an article by Remy et al. [182]).

1.3.3 GST fusion protein pull-down

While the last two approaches are of high-throughput, there exist low-throughput experiments that can very accurately confirm direct PPIs, where the proteins come into physical contact. These experiments are often required when one wants to confirm with high confidence an interaction between two proteins detected to be interacting in a large scale mapping, which may contain several false positives. Indeed, the use of glutathione-s-transferase (GST) fusion protein pull-downs has been popularized to identify direct interactions [104, 168]. In this setup, a recombinant protein (bait) is fused to GST and purified. The GST tagged bait is then incubated *in vitro* with the highly purified interacting candidate (prey) with glutathione-agarose beads. The proteins recovered from the beads are then typically analyzed through western

blotting to assess their putative interaction. Obviously, such approach is of very low-throughput, but it compensates by being extremely specific and capable of detecting direct interactions.

1.3.4 Affinity purification coupled to mass spectrometry

Finally, an alternative for performing large scale PPI identifications is affinity purification coupled to mass spectrometry (AP-MS) [37, 74, 73, 119, 90, 20]. In this protocol, a molecular tag is fused to a protein of interest (bait) in order to discover its interactors (preys). Beads binding the tag are then used to purify the bait and the preys interacting with it directly or indirectly. The preys are then identified using mass spectrometry (see Figure 1–5).

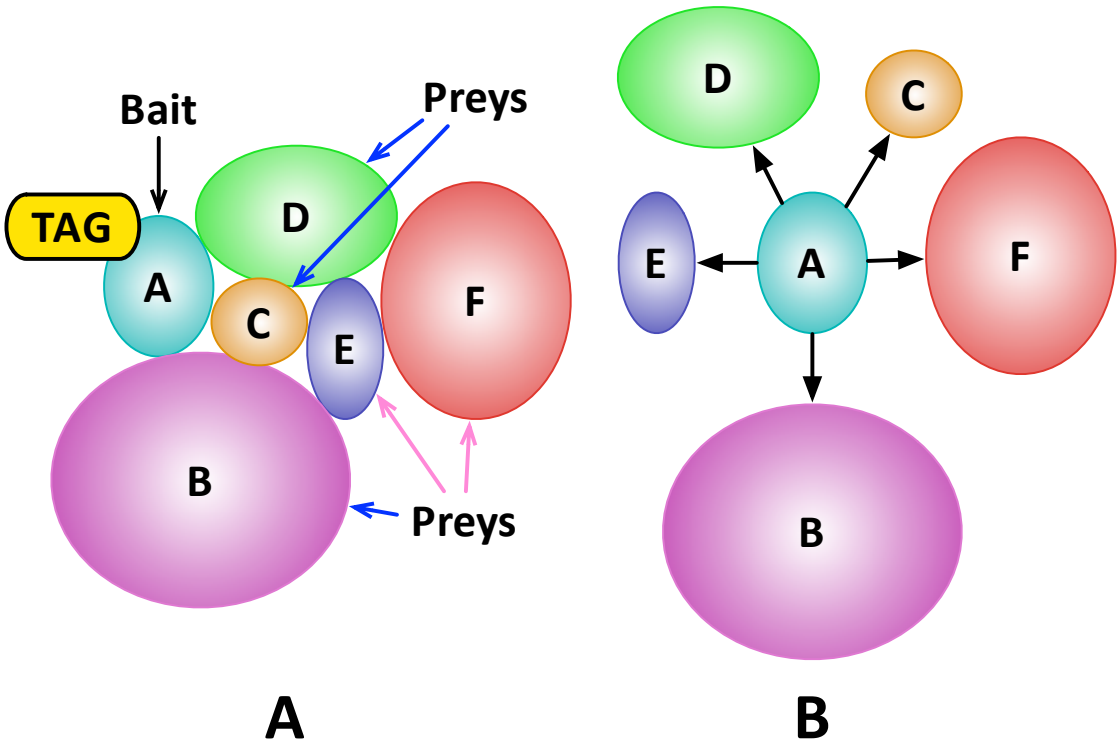

Figure 1–5: (A) Fictitious protein complex of 6 subunits, where subunit A is tagged. (B) Interactions obtained upon affinity purification of protein A.

1.3.4.1 Affinity purification

Different types of tags can be fused to the bait to perform affinity purification (AP). A commonly used tag is the tandem affinity purification (TAP) protocol [184] (see Figure 1–6). In a TAP experiment, the bait is fused to a tag consisting of two parts. A first purification is made with *IgG* beads binding the exposed part of the tag constituted of Protein A. Then, the tag is cleaved and a calmodulin binding peptide is exposed and purified using calmodulin beads. These two sequential purifications are used to minimize the number of non-specific bait binders obtained in the purified solution, at the cost of a lower sensitivity. The popular alternative to TAP, the FLAG-tag, only requires a single purification. FLAG purifications tend to be more sensitive and are great to detect transient PPIs that can be lost in the two stringent purifications of TAP. However, they are reputed to produce a large fraction of interactions that are the result of non-specific binding [20, 33].

1.3.4.2 Mass spectrometry

Once the bait and its preys are purified using either tags, mass spectrometry (MS) is then typically used to accurately identify and sometimes quantify the preys (see Figure 1–7). In order to process the proteins obtained in AP with the mass spectrometer, they first need to be enzymatically digested (usually with trypsin) into small peptides [143] (human tryptic peptide average length is 10 amino acids [158]). The resulting peptide mixture is then separated with liquid chromatography (LC) to favorise peptide detection [230]. LC uses various peptide chemical properties such as hydrophobicity and charge to separate as much as possible the elution times of the different peptides in the mixture and maximize the sensitivity of peptide detection at the MS level. An alternative to LC is to perform a gel based separation on one or two protein properties such as mass and isoelectric point [150]. In this context, in-gel digestion of proteins is often applied [198]. Tandem mass spectrometry (MS/MS) is then usually utilized to identify and potentially quantify the peptides present in the

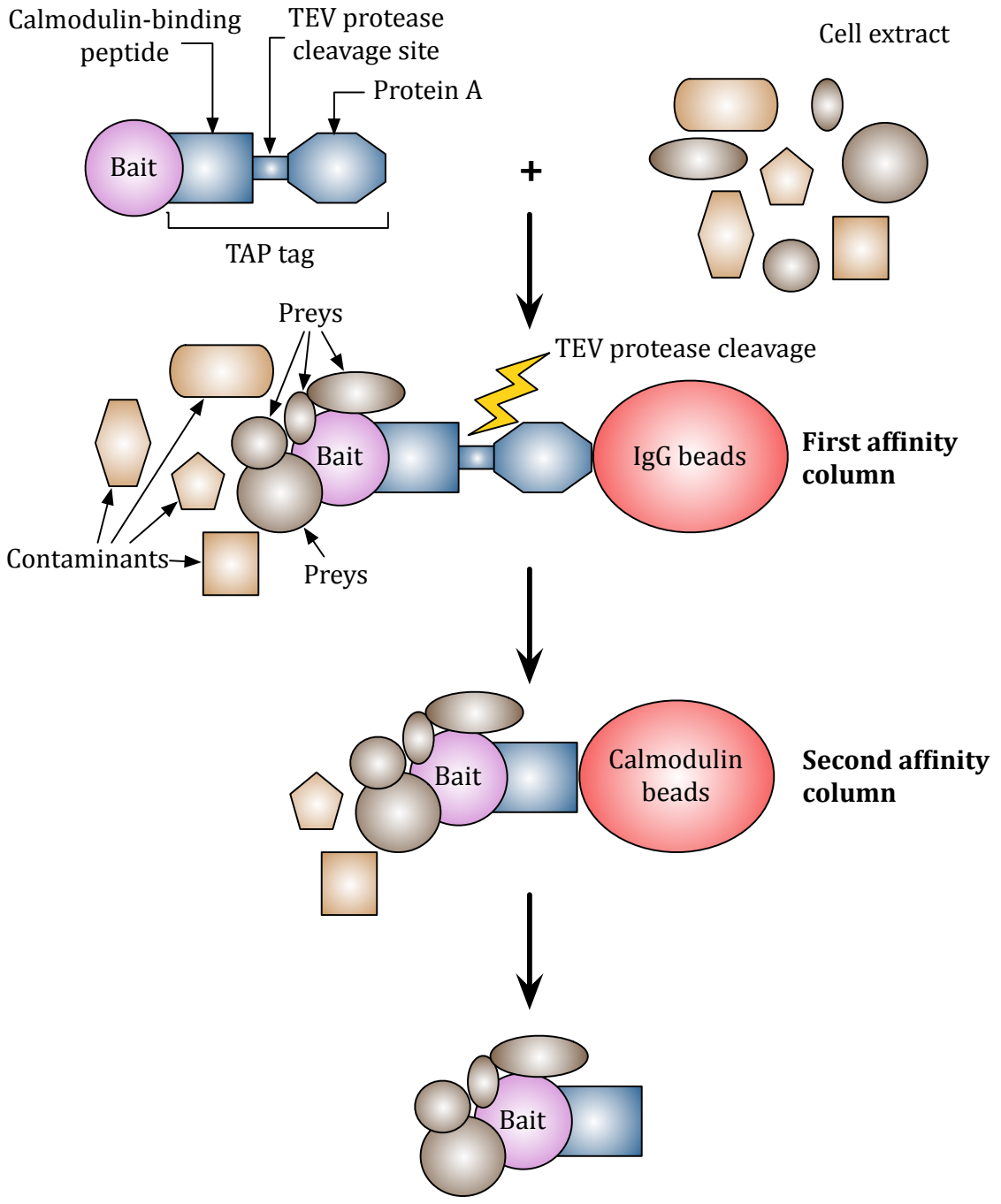

Figure 1–6: Graphical representation of the TAP procedure (inspired from a figure of an article by Lukas A. Huber [93]).

peptide mixture. In the first MS phase, the mass of each peptide is resolved. Selected peptides are then fragmented into smaller ions and analyzed in the second MS phase

[230]. The resulting spectra can then be associated to a given peptide sequence by either matching it to theoretical spectra from a database [174, 63, 58] or by using *de novo* sequencing if the genome of the organism analyzed is not available [146, 46]. The resulting sequence identification is then assigned a confidence score and proteins are identified based on their high scoring peptide sequences (technique described in Chapter 2). Even if protein abundances can be obtained using techniques such as selected reaction monitoring (SRM) [49, 126], a simple approach using the number of spectra associated to each peptide can provide a good quantification estimation if needed [32]. This approach, called spectral counting, makes the assumption that the number of acquired spectra for a given peptide type will correlate with the abundance of that peptide [32]. Proteomics pipelines are long and comprise several steps, many of which become sources of false positive protein identifications. Such sources will be explored in details in Chapter 2, where we propose computational approaches to address them.

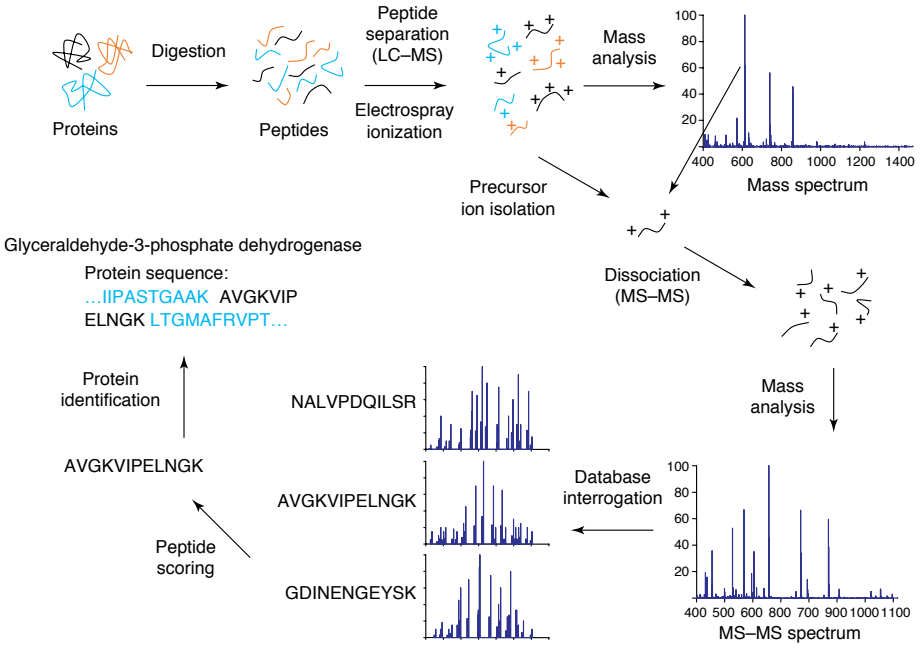

Figure 1–7: LC-MS/MS pipeline with spectral library searching (from an article by Kolker et al. [114]).

Even though this is less the case these days, AP-MS remains costly mainly because of the use of MS. As mentioned before, this approach will detect large numbers of interactors for a given bait because of its ability to detect indirect interactions [184, 189]. Some might classify this as a disadvantage as AP-MS will rapidly produce large datasets making analyses cumbersome. Nevertheless, with proper bioinformatics analysis this feature can prove very powerful. Indeed, to map an entire protein complex, only the purification of a single complex subunit may be necessary as the bait's direct and indirect interactors will be obtained (see Figure 1–5). While on the other hand, multiple screens would be needed to accomplish this task using PCA or Y2H. Another advantage of AP-MS is that the experiment can be performed in cell lines of any organisms. This allows the bait protein to be expressed in its endogenous host with the endogenous chaperones that helps its folding and the PTMs that normally affects it. Even more, PPIs can be detected in different cell compartments independently [131, 176, 125, 124, 68, 53].

The methods introduced above all produce PPI datasets containing a large fraction of false positive interactions. However, with proper bioinformatics and statistical analysis, it is possible to tackle this issue and identify the vast majority of these problematic interactions. This thesis will introduce among other things a novel innovative computational approach addressing this issue.

1.4 Protein-protein interaction networks

As mentioned previously, high-confidence PPIs can be viewed as a network. Such network can provide crucial information for protein function inference, protein complex discovery, and prediction of protein-disease associations. In addition, these networks may be useful to distinguish the true interactors of a protein from the false positives when performing an experiment to discover the interactions of a given protein. This last aspect will be discussed in more details in Chapter 2. PPI networks

have grown very large with time. For example, one of the most popular PPI repositories, BioGRID, contained $130,292$ unique interactions involving $17,373$ proteins at the time of submission of this thesis. Evidently, to process and extract information from such large networks, computational approaches quickly became a necessity.

1.4.1 Graph theory representation

A PPI network can be represented as a graph G where the proteins in the network consist of the set of vertices (nodes) V and the interactions form the set of edges E. Edges can be unweighted or weighted with a real-value weight representing for example a MS confidence score. They could also be directed from the bait to the prey if the experimental setup provides such information.

1.4.2 Properties of PPI networks

There are various graph theory properties that can be evaluated in the context of PPI networks. These properties often correlate with biological features of proteins. Some properties can tell much about the importance of certain proteins. These include the degree and the centrality of a vertex.

1.4.2.1 Vertex degree

We define the degree of a vertex to be $|N(v)|$, where $N(v)$ is the set of vertices adjacent to v. An interesting property to study in PPI networks is the degree distribution. Let $\Pr[k]$ be the probability that a randomly selected vertex has a degree equal to k. In Erdős-Rényi random networks [60], a popular random model, $\Pr[k]$ follows a Poisson distribution [11]. However, biological networks like PPI networks tend to resemble more like scale-free networks for which the vertex degrees are power-law distributed: $\Pr[k] = ck^{-\gamma}$, where c and γ are constants [11].

Several studies have been performed to verify the essentiality for the cell survival of different structures in a PPI network. Gene knockout experiments with the yeast model organism have shown that high degree proteins in the yeast PPI network tend to be more critical than low degree ones for the survival of the organism [100, 85, 12]. On the other hand, when a protein that possesses a sibling (protein sharing the same interacting partners) is deleted, it does not tend to be lethal. This can be explained by the presence of alternative paths, provided by the sibling protein in the network that might be performing a similar role as the protein suppressed [181].

1.4.2.2 Centrality measures

Other ideas have been proposed to assess the importance of a given vertex. Many of them revolve around the theme of centrality. A basic centrality measure of a vertex is its degree [234]. The higher the degree, the more central a vertex is considered to be in the network. However, it is easy to find an instance of a graph where a very high degree vertex could be connected with a very long path to a much larger subgraph and therefore not be central in the network.

A variety of distance-based centrality measures have also been proposed. Among them we count the betweenness centrality [95], which is defined as:

$$C_B(v) = \sum_{s \neq t \neq v \in V} \rho_{st}(v)/\rho_{st}$$

where $\rho_{st}(v)$ is the number of shortest paths from a source s to a target t passing through v and ρ_{st} is the number of shortest paths between s and t. This normalization, by the number of shortest paths between s and t, avoids that the centrality value of vertex v is biased by the number of shortest paths between s and t.

Another type of centrality, the feedback-based centrality [103], is based on the idea that a vertex becomes more central in tandem with the centrality of its neighbours. An example of such centrality is the famous PageRank score [21], which scores

a vertex based on the score of its neighbouring vertices. This scoring is mainly applicable in the context of directed networks. This recursive scoring can be written as follows:

$$C_{PR}(v) = (1 - d) + d(C_{PR}(t_1)/C_{(t_1)} + ... + C_{PR}(t_n)/C_{(t_n)})$$

where $t_1...t_n$ are the vertices with a directed edge towards v, while $C(v)$ is the out-degree of v, and finally d is a damping factor between 0 and 1.

In conclusion, several links have been found between certain vertex properties and specific biological roles that proteins play in the cell. However, it remains unclear if all these observations will still hold when PPI network mapping efforts will be completed. It is quite possible that the degree distribution or the enrichment in hubs for essential proteins are artifacts of the incompleteness of PPI networks and will potentially change in the future.

1.4.3 Protein annotation inference

A classic PPI network problem is to infer functions to uncharacterized proteins. The interactions of a given protein can tell much about its potential functions. These functions can take the form of disease-associations, molecular functions, biological processes, or pathways. Various methods have been developed in order to take advantage of known protein annotations to discover novel ones. A basic principle that is often shared among these approaches is called "guilt by association"; if two proteins P_1 and P_2 are interacting, and the function of P_1 is known, but the function of P_2 is unknown, then P_2 is likely to be associated to a function related to the one associated with P_1 [166].

Although such annotation inference could seem rather simple to perform, there are several challenges associated with this task. As it was mentioned before, PPI networks contain several false positive interactions complicating the annotation prediction by adding noise in the network. Also, proteins can have more than a single

annotation. For instance, RPAP2 was recently reported to be a phosphatase [57] and a protein involved in the import of RNA polymerase II [70].

1.4.3.1 Protein-disease association

A strategy being used to discover individual proteins or subset of proteins (pathways) associated with a specific disease is to analyze PPI networks using the "guilt by association" principle described earlier [218]. Some approaches will use the local network information to infer a disease association to a given protein. For instance, some will infer a disease association score to a protein based on the disease associations of its network neighbours [122]. More sophisticated approaches, where disease information is propagated in the PPI network to infer new disease causing genes, have now been developed [218]. These approaches all share the common ground that they highlight the crucial role that PPI data can play in protein-disease association inference.

1.4.4 Protein complex discovery

Since PPI networks are very large, clustering proteins into smaller components is essential in order to understand the biological processes represented in them. An interesting question to look at, from both a computational and a biological perspective, is the discovery of protein complexes in PPI networks. Complexes are represented by dense subgraphs in a network (i.e. group of vertices with several edges interconnecting them). Again, identifying such clusters may seem trivial. However, since proteins can be part of multiple complexes, they sometimes perform several functions. In this scenario, network clusters are expected to be overlapping, causing clustering algorithms outputting disjoint clusters to fail to identify some protein complexes in such datasets [234]. Obviously, the important amount of noise in PPI networks also complicates protein complex identifications.

1.4.4.1 Clique finding

The most intuitive way to find protein complexes is to find maximal cliques [234]. A clique is a set of vertices where each vertex is connected by an edge to every other vertex in the clique. A clique is maximal when no other adjacent vertices can be added to it while respecting the clique definition. However, finding all maximal cliques in a graph is NP-hard [132] and as mentioned before, the number of proteins in the networks analyzed is usually quite large. Also, protein complexes are very seldom found as cliques. It is often observed that a protein complex will not form a perfect clique as some of its interactions will not be observed because of the lack of sensitivity of the experimental protocol [234].

1.4.4.2 Dense subgraph identification

Rather then searching for maximal cliques, one can identify dense subgraphs in a network. Such subgraphs are often identified by finding the set of vertices of size n, where n is fixed, that minimizes the sum of all pair shortest paths within it [208]. This permits the identification of protein complexes not completely mapped in PPI networks. However, identifying all dense subgraphs in a network is computationally expensive. Approximation approaches have therefore been proposed to address this problem.

Among them is the Markov Chain Monte Carlo approach. In this context, the Markov Chain Monte Carlo is used to identify vertex sets with minimum all pair shortest paths as follows. Starting at time $t = 0$, a random set P of n vertices is selected and for each pair of vertices $i,j \in P$ the shortest path L_{ij} is computed. The sum of all initial L_{ij} is L_0. At each time step, one of the n vertices is randomly selected and replaced by one of its randomly selected neighbours letting the sum of the shortest paths with the new vertex to be L_1. If $L_1 < L_0$ the replacement is kept, otherwise the replacement is only kept with probability $e^{-(L_1-L_0)/T}$, where

T is a constant. This process attempts to avoid getting stuck in a local minimum that could be reached by the deterministic version of the algorithm. In addition, at every tenth time step, a vertex not connected to any vertices in P is attempted as a replacement vertex with the same rules. This procedure gives the opportunity to the algorithm to explore vertex sets in different disconnected graphs. The procedure is repeated until no vertices are replaced for a certain number of consecutive iterations or a time limit has been reached. The same Monte Carlo approach can be applied to optimize another subgraph density measure:

$$Q(P) = \frac{2m}{n(n-1)}$$

where m is the number of edges in the subgraph P [208].

1.4.4.3 Molecular complex detection algorithm

An alternative to these approaches is the molecular complex detection (MCODE) algorithm [8]. It finds densely connected subgraphs by weighting vertices with their local neighbourhood density. The weight $w = kd$ is computed for a vertex v where k is the maximal k-core of v and its direct neighbours and d is the edge density of these vertices. A k-core of a graph is defined as the subgraph created by the removal of all vertices with degree less than k and their incident edges. Vertices with degree less than k are iteratively removed, until the k-core definition is satisfied. k-cores are often used to identify interesting dense subnetworks inside large PPI networks that are usually associated to functional modules. The vertices with a high weight are chosen as cluster seeds. Clusters are then expanded by recursively attempting to add to them the neighbours of the seeds if their respective weights are above a certain percentage of the weight of the seed. Vertices added to a cluster are marked as visited and are only explored once. The recursion stops when an explored vertex weight is less than a certain percentage of the initial weight of the cluster seed. Finally, there is a post-processing step where clusters that do not contain at least a 2-core are removed. Vertices that do not belong to a cluster (not marked as visited) are added to a cluster

if their neighbourhood density is not below a certain fluff parameter (a value between 0.0 and 1.0). These vertices are not marked as visited so they can be added to multiple clusters. The clusters (or complexes) outputted by MCODE, because they can overlap, are more biologically relevant than those produced by other methods since protein complexes often share protein subunits. However, this approach is time consuming and different seeds can lead to the generation of very similar clusters.

1.4.4.4 Socio-affinity index

Another interesting approach for complex detection was proposed by Gavin et al. [74]. They derived what they called a Socio-affinity index. It quantifies the tendency for two proteins to purify each other or to be co-purified in an AP experiment. Basically, it computes the log-odds of the number of co-occurrences of the two proteins in AP experiments against their expected number of co-occurrences based on their dataset frequency. While this index is useful to identify false positives in an AP dataset, it is also a good basis for a clustering analysis in order to perform protein complex discovery. To do so, a matrix of Socio-affinity indices for all pairs of proteins in the dataset can be given as input to a clustering algorithm [74]. As it was described before, proteins can be members of numerous protein complexes. To account for this, the method applies a small penalty to the indices in the matrix after the initial clustering and then performs another clustering to report a new set of complexes. This step is iteratively repeated. Again, this approach has the advantage to have the capability to report overlapping clusters. However, in order to compute meaningful Socio-affinity indices, a large number of APs is required.

1.4.4.5 Markov clustering algorithm

None of the methods presented above were designed (although some can be adapted) to take weighted graphs as input. Such weights, which can be derived from

mass spectrometry data for example, often represent the confidence that a given interaction is a true positive. These confidence scores can help to better partition PPI networks. The Markov clustering (MCL) algorithm [59] was designed to be applied on simple weighted graphs. The algorithm takes as input a weighted similarity matrix of edge weights M, which is transformed in a Markov transition matrix where the diagonal is set to neutral values and each column is normalized to 1. The matrix is then transformed iteratively using two operations: expansion and inflation. The matrix is first "expanded" by squaring it and then inflated using the following equation

$$(\Gamma_r M)_{pq} = \frac{(M_{pq})^r}{\sum_{i=1}^{k}(M_{iq})^r}$$

where $r > 1$ and Γ_r is the inflation operator. Each operation is performed iteratively until there is no significant changes in M or after a given number of iterations. Finally, a threshold value is chosen to remove edges to form connected components. These connected components correspond to clusters, which in turn map to protein complexes in the PPI network.

All the above clustering approaches are however limited by the level of unreliability of the data and missing connectivity among the different modules in the network. It is hypothesized that with the growing size of PPI networks, protein complex connectivity will increase and therefore facilitate the clustering of proteins [112]. Methods filtering false positive interactions before or after the clustering of PPI networks are also likely to improve the quality of the clusterings obtained by computational approaches.

1.5 Thesis roadmap

This chapter introduced the biological background and importance of the approaches used to map and analyze protein-protein interactions. The four following chapters describe novel computational approaches for the analysis of PPIs and will

constitute the research contribution of this thesis. Each chapter corresponds to a specific project, with the addition of background notions related to the research question tackled in it. Chapter 2 introduces a novel Bayesian algorithm to model contaminants in AP-MS experiments. Building from the previous chapter, Chapter 3 describes a new approach derived from AP-MS to identify and computationally assess PPIs independently in three different cell compartments of the cell. With the tools developed in the last two chapters, the resulting high-confidence PPIs can be analyzed as a network. Chapter 4 presents an algorithm identifying Gene Ontology (GO) terms that are clustered in PPI networks. Chapter 5 then pushes the limit of the methods presented in the previous chapter and poses a novel approach for motif discovery in 5' untranslated region (UTR) sequences using PPI network information and links biological functions to RNA sequence motifs. Finally, Chapter 6 summarizes these research contributions and presents discussions on future works.

1.6 Publications and author contributions

This thesis comprises the full text and figures of four scientific articles, three of which have been published and one is in preparation for publication. These articles are listed below in the order they appear in this thesis. I am the first author of each of them.

- **Chapter 2:**

  M. Lavallée-Adam, P. Cloutier, B. Coulombe, and M. Blanchette. Modeling contaminants in AP-MS/MS experiments. *Journal of proteome research*, 10(2):886–895, 2010

  The design and implementation of the computational tool in this publication was performed by me under Dr. Mathieu Blanchette's and Dr. Benoit Coulombe's supervision and biological discussion was written by Philippe Cloutier and me under Dr. Benoit Coulombe's supervision.

- **Chapter 3:**

  M. Lavallée-Adam, J. Rousseau, C. Domecq, A. Bouchard, D. Forget, D. Faubert, M. Blanchette, and B. Coulombe. Discovery of cell compartment specific protein-protein interactions using affinity purification combined with tandem mass spectrometry. *Journal of proteome research*, 12(1):272–281, 2012

  The design and implementation of the computational tool in this publication was performed by me under Dr. Mathieu Blanchette's and Dr. Benoit Coulombe's supervision. The biological methodology and results were produced by Justine Rousseau, Céline Domecq, Annie Bouchard, Diane Forget under Dr. Benoit Coulombe's supervision, while the mass spectrometry analysis was performed by Dr. Denis Faubert.

- **Chapter 4:**

  M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of locally over-represented GO terms in protein-protein interaction networks. *Journal of Computational Biology*, 17(3):443–457, 2010

  The design and implementation of the computational tool in this publication was performed by me under Dr. Mathieu Blanchette's and Dr. Benoit Coulombe's supervision.

- **Chapter 5:**

  M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of functional sequence motifs in human 5' UTRs based on local enrichments in a protein-protein interaction network. *Manuscript in preparation.*

  The design and implementation of the computational tool in this manuscript was performed by me under Dr. Mathieu Blanchette's and Dr. Benoit Coulombe's supervision.

CHAPTER 2
Modeling contaminants in AP-MS/MS experiments

2.1 Preface

AP-MS is among the most popular methods to identify PPIs. Nevertheless, it still suffers from important specificity issues. Typically, the majority of interactions reported by unfiltered AP-MS experiments are false positives. The AP-MS experimental pipeline is long and requires several manipulations that create multiple sources of contamination of the results. These include, among other things, contamination of samples with human keratins, non-specific binding of proteins to purification antibodies, and carry-over of proteins from one LC run to another. However, such contamination events, which are discussed in this chapter are not the only ways false positives are added in AP-MS data.

The other main sources of false positives, which are not described in great details in this chapter, are misidentifications at the MS level when performing the database search for peptide-spectrum matches (PSMs). Misidentifications often originate from MS instrument noise, contamination from non-peptide molecules, peptides with PTMs not specified in the database search parameters, or incorrect charge-state determination [82]. Distinguishing correct from incorrect PSMs is not a trivial task. While some researchers rely on manual verification for small datasets, most turn themselves to various filtering criteria to process the thousands of spectra produced by mass spectrometers in AP-MS protocols. However, the latter strategy has unknown performances and limited portability as it heavily depends on the sample preparation and the type of MS instrument used [109]. More elegant statistical models have been developed to address this problem. One of the most popular tools is called Peptide Prophet [109]. It uses a Bayesian approach to evaluate each PSM by computing a

probability that it is correct using MS database search scores and the number of tryptic termini of matched peptides. It is then possible to compute the probability that a protein is present in a sample using its corresponding PSM confidence scores [161]. The method presented in this chapter does not (and was not designed to) tackle such problem. It could however benefit from either pre-processing the dataset analyzed with tools like Peptide Prophet or use their output as a basis for its input.

The approach introduced in this chapter, Decontaminator, uses a small number of biological context specific AP-MS negative controls to model the contaminants present in the AP-MS experimental setup. Soon after Decontaminator was made publicly available, a generalized version of a piece of software performing a similar task, Significance Analysis of Interactome (SAINT) 2.0, was published [34]. This version now addresses problems that are described later in this chapter. SAINT 2.0 can now use control experiments when available and only requires a limited number of experiments to accurately assess PPIs without any manual labelling of hub proteins in the input network [34]. Although Decontaminator and SAINT 2.0 are built from the same principles and achieve similar goals, their methods differ. We noted, after the publication of SAINT 2.0, that both Decontaminator and SAINT 2.0 slightly outperformed each other depending on the dataset analyzed. Decontaminator could be improved in the future by including the use of purification replicates in its model (a capability of SAINT 2.0 [34]).

In this chapter, Decontaminator was benchmarked against other approaches using the union of the BioGRID [210] and HPRD [179] databases to maximize the size of the reference dataset. An alternative approach would have been to compare approaches using a small but very high quality PPI dataset composed of interactions supported by multiple publications such as the one that can be obtained from the iRefWeb database [217]. Decontaminator was also benchmarked against other methods using GO terms [4]. While the GO database may contain false positive annotations and that these annotations may be biased towards proteins that have been

the subject of numerous publications, it remains an interesting validation tool since it is expected that proteins that are truly interacting are more likely to share a GO term annotation than those involved in a false positive PPI. Although GO semantic similarity scores [97] of the PPIs in the predicted sets of valid interactions were not used to compare each approach, they consist in another interesting benchmarking method.

On another note, reproducibility of AP-MS experiments is often challenged by the complexity and high variability of the methods (see Figure 2–1). To address this issue from a computational perspective, a program, ROCS, was recently developed to compute reproducibility indices for AP-MS experiments in order to discriminate reproducible experiments from outliers [48]. Such method could be used in combination with Decontaminator to identify the most reproducible controls and leave out the outliers in order to maximize the modeling accuracy of contaminants in a set of experiments.

Protein expression levels in the biological sample under study may also be used as complementary data to assess the quality of PPIs. Indeed, proteins with high expression levels are likely to be contaminants in a certain sample if their affinity for the antibody used for the AP-MS experiment is reasonably high. Such data could be obtained from protein expression measurements or from predictions made using mRNA concentrations and sequence signatures [221].

Lately, a repository of AP-MS control experiments, CRAPome, was launched [153]. This database contains 343 controls of various types. CRAPome could reveal to be a great resource for laboratories performing small-scale AP-MS experiments. Small numbers of controls are often not sufficient to capture the majority of possible contaminants in a given experimental setup. Since AP-MS controls are typically bait-independant, CRAPome controls can be used to complement controls from a given laboratory to better identify contaminants present in this laboratory's dataset.

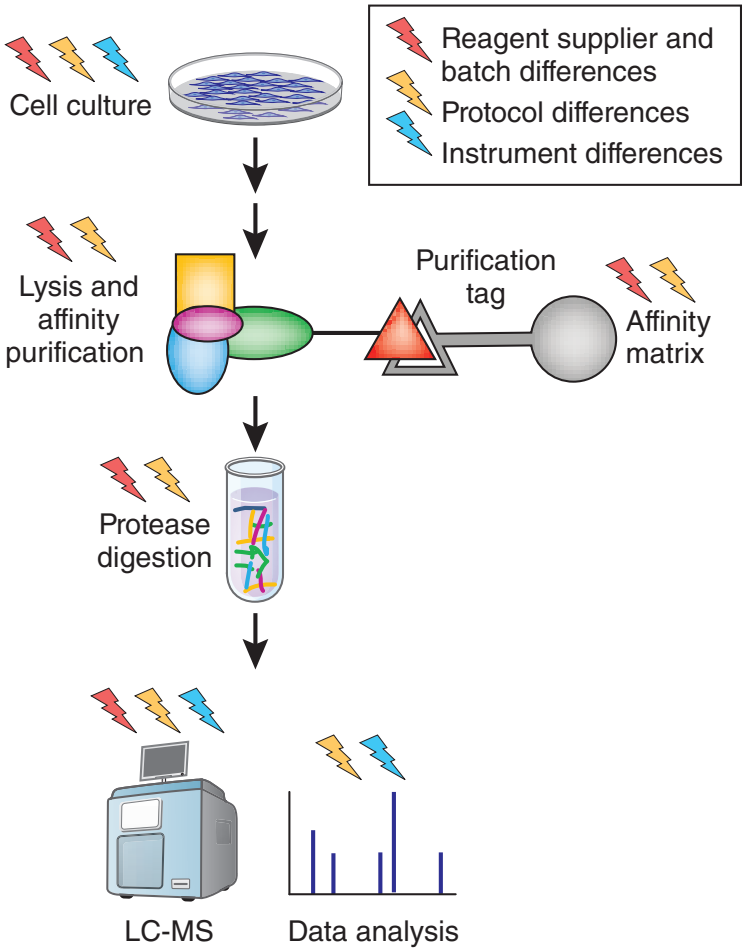

Figure 2–1: Graphical representation of the multiple sources of variation in the AP-MS protocol (adapted from a figure from an article by Pascal Braun [19]).

Protein MS sensitivity has been dramatically increasing over the last few years. In the context of AP-MS, this gain does not only allow the detection of less abundant PPIs, but also results in the observation of less abundant contaminants. This increases the number of contaminants in the results and demonstrates the need of approaches, such as Decontaminator, which are capable of modeling contamination events.

The remaining content of this chapter is reprinted with permission from:

- M. Lavallée-Adam, P. Cloutier, B. Coulombe, and M. Blanchette. Modeling contaminants in AP-MS/MS experiments. *Journal of proteome research*, 10(2):886–895, 2010

2.2 Abstract

Identification of protein-protein interactions (PPI) by affinity purification coupled to tandem mass spectrometry (AP-MS/MS) produces large datasets with high rates of false positives. This is in part because of contamination at the AP level (due to gel contamination, non-specific binding to the TAP columns in the context of tandem affinity purification, insufficient purification, etc.). In this paper, we introduce a Bayesian approach to identify false positive PPIs involving contaminants in AP-MS/MS experiments. Specifically, we propose a confidence assessment algorithm (called Decontaminator) that builds a model of contaminants using a small number of representative control experiments. It then uses this model to determine whether the Mascot score of a putative prey is significantly larger than what was observed in control experiments and assigns it a p-value and a false discovery rate. We show that our method identifies contaminants better than previously used approaches and results in a set of PPIs with a larger overlap with databases of known PPIs. Our approach will thus allow improved accuracy in PPI identification while reducing the number of control experiments required.

2.3 Introduction

The study of protein-protein interactions (PPI) is crucial to the understanding of biological processes taking place in cells [222]. Affinity purification (AP) combined with mass spectrometry (MS) is a powerful method for the large scale identification of PPIs [37, 73, 74, 119, 90, 20]. The experimental pipeline of AP consists in first tagging a protein of interest (bait) by genetically inserting a small peptide sequence

(tag) onto the recombinant bait protein. The bait protein is affinity purified, together with its interacting partners (preys), which are identified using MS. However, this type of experiment is prone to false positive identifications for various reasons [76], which can seriously complicate the downstream analyses. In the context of affinity purification, contamination of manually-handled gel bands, inadequate purification, purification of specific complexes from abundant proteins, and non-specificity of the tag antibody used are some of the many ways contaminants can be introduced in the experimental pipeline before the mass spectrometry (MS) phase. These contaminants, added to the already large set of valid preys of a given bait, create even longer lists of proteins to analyze. While common contaminants can be identified easily by a trained eye, sporadic contaminants can be considered erroneously as true positive interactions. In addition to contaminants, false positive PPIs can be introduced at the tandem mass spectrometry phase (MS/MS) step [18]. For example, peptides of proteins with low abundance or involved in transient interactions can be difficult to identify because of the lack of spectra. Such peptides can be misidentified by database searching algorithms such as Mascot [174] or SEQUEST [58]. Although many approaches have been proposed in order to limit the number of mismatched MS/MS spectra (e.g. Peptide Prophet [109] and Percolator [105]), the modeling and detection of contaminants, which is the problem we consider in this paper, has received much less attention.

2.3.1 Related work

A number of experimental and computational approaches have been proposed to reduce the rate of false positive PPIs. Several steps in the experimental pipeline can be optimized to minimize contamination. In-cell near physiological expression of the tagged proteins is preferred to over-expression to prevent spurious PPIs. Also, additional purifications could be performed in order to remove contaminating proteins from affinity purified eluate. The drawback of an increased number of purifications is a loss of sensitivity, as transient or weak PPIs will be more likely to be disrupted

[37]. When performing gel-based sample separation methods before MS/MS, manual gel band cutting can introduce contaminants such as human keratins in the sample. This can be addressed by robot gel cutting, although this increases equipment cost. As an alternative, gel-free protocols simply use liquid chromatography to separate the peptide mixture before MS/MS. However, depending on the complexity of the mixture, less separation might result in an important decrease in sensitivity. Finally, liquid chromatography column contamination from previous chromatographic runs is also important to consider. Although it is possible to wash the column to eluate peptides from previous chromatographic runs, very limited washing is typically done because of its time consumption.

Several computational methods have been used to identify the correct PPIs from AP-MS/MS data [190, 206]. Some involve the use of the topology of the network formed by the PPIs (e.g. number of times two proteins are observed together in a purification to assign a Socio-affinity index [74] or a Purification Enrichment score [39]. Others used various combinations of data features such as mass spectrometry confidence scores, network topology features and reproducibility data with machine learning approaches in order to assign probabilities that a given PPI is a true positive [119, 61, 101, 37]. However, with each of these methods, contaminants would often be classified as true interactions because of their high database matching scores and reproducibility. Such sophisticated machine learning procedures can be prone to over-fitting and the use of small, manually curated, but often biased training set, such as MIPS complexes [156] as used by Krogan et al. [119] or a manually selected training set as used by our previous approach [101, 37], can be problematic depending on the nature of the data analyzed. Finally, Chua et al. combined PPI data obtained from several different experimental techniques in an effort to reduce false positives [36]. Although such methods will be very efficient at filtering contaminants, they will typically suffer from poor sensitivity.

All of these methods attempt to model simultaneously several sources of false positive identifications including contamination but also, for example, misidentification of peptides at the mass spectrometry level. For instance, scoring methods relying on the topology of PPI networks will tend to assign low scores to proteins being observed as preys in several experiments because they are likely contaminants, while also scoring poorly proteins largely disconnected from the network, which are potential database misidentifications. However, none of these attempt to directly model and filter out contaminants resulting from AP experiments. Deconvoluting the modeling of false identification into AP contaminants modeling and database matching of MS/MS spectra will potentially lead to methods identifying PPIs with higher accuracy. To date, most computational methods aiming specifically at filtering out likely contaminants have been quite simplistic. Several groups maintain a manually assembled list of contaminants and then systematically reject any interactions involving these proteins [76]. However, it is possible that a contaminant for one bait is a true interaction for another, suggesting that a finer model of contaminant level would be beneficial. Recently, the Significance Analysis of Interactome (SAINT), a sophisticated statistical approach attempting to filter out contaminant interactions resulting from AP-MS/MS experiments was introduced [20]. SAINT assesses the significance of an interaction according to the semi-quantitative peptide count measure of the prey. It discriminates true from false interactions using mixture modeling with Bayesian statistical inference. However, the currently available version of SAINT (1.0) lacks the flexibility to learn contaminant peptide count distributions from available control data and requires a considerable number of baits (15 to 20) in order to yield optimal performances. Moreover, although not necessary, manual labeling of proteins as hubs or known contaminants is required to achieve the best possible accuracy.[1]

[1] It is worth noting that another version of SAINT is currently under development and promises to address many of these issues.

Once the tagging has been performed, some affinity purification methods require the vector of the bait to be induced so that the tagged protein is expressed. An alternate method to identify likely contaminant PPIs for a given bait is to perform a control experiment where the expression of the tagged protein is not induced prior to immunoprecipitation. It is then possible to compare mass spectrometry confidence scores (e.g. from Mascot [174]) for the preys from both the control and induced experiments. For example, in Jeronimo et al. [101], only preys with Mascot score at least 5 times larger in the induced experiment than the control experiment were retained, the others being considered as likely contaminants. There are limitations to such false positive filtering procedure. First, this method is expensive in terms of time and resources, since the cost is doubled for each bait studied. Second, because of the noisy nature of MS scores for low-abundance preys, comparing a single induced experiment to a single non-induced experiment is problematic. Pooling results from several non-induced experiments could be beneficial. Third, some baits will show leaky expression of the non-induced vector. Depending on the level of leakiness, several true interactions may be mistakenly categorized as false positives.

Here, we propose a confidence assessment algorithm (Decontaminator) using only a limited number of high quality controls sufficient to the proper identification of contaminants obtained from AP-MS/MS experiments without prior knowledge about neither hubs nor contaminants. By pooling control experiments, one-to-one comparisons of induced and non-induced experiment Mascot scores are avoided. Our fast computational method thus provides accurate modeling of contaminants while limiting resource usage.

2.4 Methods

We propose a Bayesian approach called Decontaminator that makes use of a limited number of non-induced control experiments in order to build a model of contaminant levels as well as to analyze the noise in the measurements of Mascot

scores. Decontaminator then uses this model to assign a p-value and an associated false discovery rate (FDR) to the Mascot score obtained for a given prey. We start by describing the AP-MS/MS approach used to generate the data, and then describe formally our contaminant detection algorithm. Alternate approaches are also considered and their accuracy is compared in the Results section.

2.4.1 Biological data set

The Proteus database contains the results of tandem affinity purification (TAP) combined with MS/MS experiments performed for a set E of 89 baits, both in non-induced and induced conditions [69, 37, 101, 120]. The baits selected revolve mostly around the transcriptional and splicing machineries. The set of proteins identified as preys by at least one bait (in either the induced or non-induced experiments) consists of 3619 proteins. Detailed TAP-MS/MS methodology has been described elsewhere [37]. Briefly, a vector expressing the TAP-tagged protein of interest was stably transfected in HEK 293 cells. Following induction, the cells were harvested and lysed mechanically in detergent-free buffers. The lysate was cleared of insoluble material by centrifugation and the tagged protein complexed with associated factors were purified twice using two sets of beads each targeting a different component of the TAP tag. The purified protein complexes were separated by SDS-PAGE and stained by silver nitrate. The acrylamide gel was then cut in its entirety in about 20 slices that were subsequently trypsin-digested. Identification of the tryptic peptides obtained was performed using microcapillary reversed-phase high pressure liquid chromatography coupled online to a LTQ-Orbitrap (Thermo Fisher Scientific) quadrupole ion trap mass spectrometer with a nanospray device. Proteins were identified using the Mascot software [174] (Matrix Science) (see Appendix for software information). For some of the baits tested, the promoter was leaky, which resulted in the expression, at various levels, of the tagged protein, even when it was not induced. These baits were identified by detection of the tagged protein in the non-induced samples and these samples were excluded from our analysis. A set $B \subset E$ of 14 non-leaky baits was

selected for this study: $B = \{b_1, ..., b_{14}\} = \{SFRS1, NOP56, TWISTNB, PIH1D1, UXT, MEPCE, SART1, RP11-529I10.4, TCEA2, PDRG1, PAF1, KIAA0406, POLR1E, KIN\}$. The set P of preys they detected in at least one of these 28 induced and non-induced experiments contains 2415 proteins. Out of these, 1067 proteins were unique to a single non-induced bait and 808 were detected more than once in the set of 14 controls, while 540 were only observed in induced experiments. We denote by $M^{NI}_{b,p}$ the Mascot score obtained for prey p in the experiment where bait b is not induced, and by $M^{I}_{b,p}$ the analogous score in the induced experiment. Note that for most pairs (b, p), where $b \in B$ and $p \in P$, p was not detected as a prey for b, in which case we set the relevant Mascot score to zero.

In the non-induced experiments, the number of preys detected for each bait varies from 206 to 626, with a mean of 316. These preys are likely to be contaminants, as the tagged bait is not expressed. In induced experiments, the number of preys per bait goes from 5 to 516, with a mean of 135. It may appear surprising that there are on average more preys detected in the non-induced experiments than in the induced ones. This is likely due to the fact that the presence of high-abundance preys in the induced experiments masks the presence of lower-abundance ones, including contaminants. Still, a significant fraction of the proteins detected in the induced condition are likely contaminants.

2.4.2 Computational analysis

An ideal model of contaminants would specify, for each prey p, the distribution of the MS scores (in our case, Mascot score [174]) in non-induced experiments, which we call the null distribution for p. However, accurately estimating this distribution would require a large number of non-induced experiments and the cost would be prohibitive. Instead, we use a small number of non-induced experiments and make the assumption that preys with similar average Mascot scores have a similar null distribution (this assumption is substantiated in the Discussion section). This allows

us to pool non-induced scores from different preys (if they have similar Mascot score averages) in order to build a more accurate noise model. Thus, results from a few control experiments are sufficient to build a contaminant model that can then be used to analyze the results of any number of induced experiments performed under the same conditions. Figure 2–2 summarizes our approach.

Our goal is now to use the data gathered from induced and non-induced AP-MS/MS experiments in order to build a model of noise in Mascot score measurements and eventually be able to assess the significance of a given Mascot score $M^I_{b,p}$. Let $\bar{M}^{NI}_p$ be the unobserved true mean of Mascot scores for prey p in non-induced experiments, which is defined as the mean of the Mascot scores of an infinite number of non-induced biological replicates. $\bar{M}^{NI}_p$ can never be observed, but its posterior distribution can be obtained if a few samples are available. Similarly, define $\bar{M}^I_{b,p}$ as the average of the Mascot scores of p of an infinite number of biological replicates of experiments where b is induced. The essence of our approach is to calculate the posterior distribution of $\bar{M}^{NI}_p$, the true mean Mascot score of prey p given our data from non-induced experiments, and to compare it to the posterior distribution of $\bar{M}^I_{b,p}$, the true mean Mascot score of prey p in an induced experiment, given our observed data $M^I_{b,p}$. If the second distribution is significantly to the right of the first, prey p is a likely bona fide interaction of bait b. If not, p is probably a contaminant and should be discarded.

Before describing our method in details, we give a few examples that illustrate how it works. Figure 2–3(a) shows an example of an interaction accepted by Decontaminator. POLR2E obtained a Mascot score of 320 in the induced experiment of RPAP3 and was only detected twice in control experiments (Mascot scores 38 and 48). From this figure, it can clearly be seen that the resulting posterior distribution of $\bar{M}^I_{RPAP3,POLR2E}$ is significantly to the right of the posterior distribution of $\bar{M}^{NI}_{POLR2E}$. Therefore, the RPAP3-POLR2E interaction obtains a small p-value (0.00018) and FDR (0.013) and is considered a valid interaction. This prediction is consistent with the literature about this interaction [101], which is present in

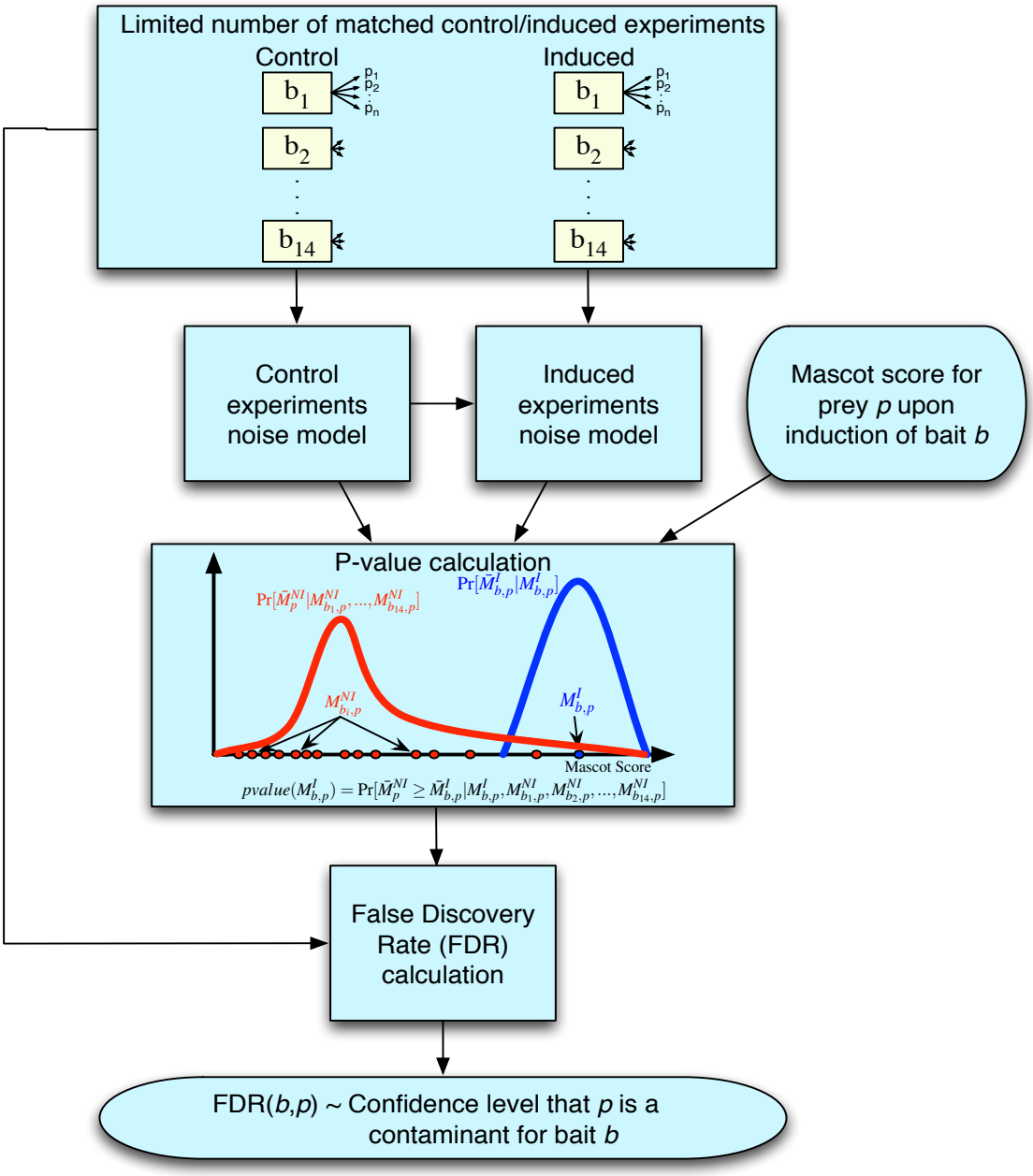

Figure 2–2: Decontaminator workflow. First, control and induced experiments are pooled to build a noise model for each. These noise models are then used to assign a p-value to each prey obtained upon the induction of a bait. Finally, a false discovery rate is calculated for each p-value.

databases such as BioGRID [210]. Conversely, Figure 2–3(b) shows the posterior distribution of $\bar{M}^{I}_{RPAP3,SAMD1}$ of SAMD1 for the induced experiment of RPAP3 (Mascot score: 102). The distribution is largely overlapping with the posterior distribution of $\bar{M}^{NI}_{SAMD1}$, which obtained Mascot scores of 95, 53, 147 and 164 in control experiments but was unobserved in 10 other controls. Clearly the RPAP3-SAMD1 interaction is not very reliable because its Mascot score does not exceed by enough some of the observed non-induced Mascot scores. Decontaminator assigns a large p-value (0.18) and FDR(> 0.9) and labels the interaction as a contaminant. Of note, this interaction would often have been incorrectly classified by methods based on a direct comparison of Mascot scores between matched induced and non-induced experiments (as used, for example, in Jeronimo et al.[101]), as SAMD1 shows up in control experiments only four out of 14 times. Finally, Figure 2–3(c) shows the results for the KPNA2-ACTB interaction. ACTB, like other actin proteins, is a known contaminant for our experimental pipeline, obtaining Mascot scores varying from 35 to 210 for eight of the controls. However, the Mascot score observed for this interaction (793) is sufficiently higher to allow the interaction to be predicted as likely positive (p-value: 0.005, FDR: 0.17). Beta-actin (ACTB) is known to shuttle to and from the nucleus [223], a process which may directly require karyopherin alpha import protein KPNA2. Beta-actin and actin-like proteins have been shown to be part of a number of chromatin remodeling complexes [165] and analysis of the KPNA2 purification revealed a strong presence for such complexes (TRRAP/TIP60 & SWI/SNF-like BAF complexes). Therefore, it is possible that these remodelers are assembled in the cytoplasm and imported together to the nucleus via KPNA2. This example shows how the method can differentiate specific interactions from classic contamination.

The Decontaminator approach involves four steps (see Figure 2–2):

1. Use non-induced experiments for different baits as biological replicates and obtain a noise model defined by $\Pr[M^{NI}_{b,p}|\bar{M}^{NI}_{p}]$ and $\Pr[M^{I}_{b,p}|\bar{M}^{I}_{b,p}]$.

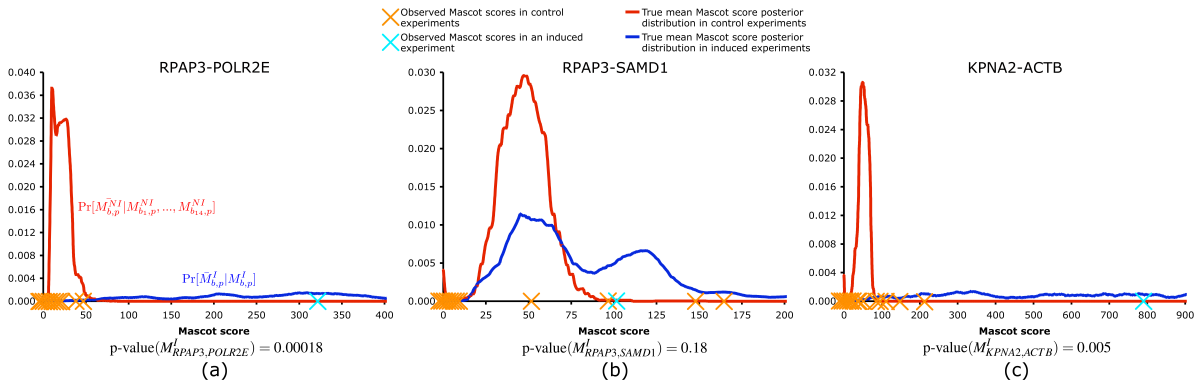

Figure 2–3: Plots of both the posterior distributions of $\bar{M}^{NI}_p$ and posterior distributions of $\bar{M}^I_{b,p}$ for three different interactions. The blue curve is the posterior distribution of $\bar{M}^{NI}_p$ and the red curve is the posterior distribution of $\bar{M}^I_{b,p}$. Orange X's positions on the x-axis are observed Mascot scores of the prey in control experiments. When the prey was not detected for a given control, an X is drawn on the x-axis close to the origin. The blue X's position on the x-axis is the Mascot score of the prey in the induced experiment for the corresponding bait. (a) The RPAP3-POLR2E interaction is an example of an interaction considered valid by the algorithm (b) The RPAP3-SAMD1 interaction is an example of an interaction where the prey is considered as a contaminant. (c) KPNA2-ACTB is an example of an interaction that is predicted as positive, even though ACTB is often observed as a contaminant (but with lower Mascot scores) in control experiments.

2. For each protein $p \in P$, calculate the posterior distribution of $\bar{M}^{NI}_p$ given $M^{NI}_{b_1,p}, M^{NI}_{b_2,p}, ..., M^{NI}_{b_{14},p}$. Similarly, calculate the posterior distribution of $\bar{M}^I_{b,p}$, given $M^I_{b,p}$.
3. For each pair $(b,p) \in B \times P$, assign a p-value to $M^I_{b,p}$:

$$p\text{-value}(M^I_{b,p}) = \Pr[\bar{M}^{NI}_p \geq \bar{M}^I_{b,p} | M^I_{b,p}, M^{NI}_{b_1,p}, M^{NI}_{b_2,p}, ..., M^{NI}_{b_{14},p}]$$

4. Using the non-induced and full induced data sets, assign a false discovery rate (FDR) to each p-value.

Each step is detailed further below.

2.4.2.1 Step 1: Building a noise model from non-induced experiments

The set of 14 TAP-MS/MS experiments where the bait's expression was not induced can be seen as a set of biological replicates of the null condition. We use these measurements to assess the amount of noise in each replicate, i.e. to estimate $\Pr[M_{b,p}^{NI}|\bar{M}_p^{NI}]$, the probability of a given observation $M_{b,p}^{NI}$, given its true mean Mascot score $\bar{M}_p^{NI}$. This distribution is estimated using a leave-one-out cross-validation approach on the set of 14 non-induced experiments. Specifically, for each bait $b \in B$, we compare $M_{b,p}^{NI}$ to $\mu_{\neq b,p}$, the corrected average (see Appendix) of the 13 Mascot scores of p in all non-induced experiments except where bait b was used. $\mu_{\neq b,p}$ provides a good estimate of $\bar{M}_p^{NI}$. Let $C(i,j)$ be the number bait-prey pairs for which $\lfloor M_{b,p}^{NI} \rfloor = i$ and $\lfloor \mu_{\neq b,p} \rfloor = j$. Then, a straight-forward estimator is

$$\Pr[M_{b,p}^{NI} = x|\bar{M}_p^{NI} = y] = C(x,y)/\sum_{x'} C(x',y).$$

Note that C is a fairly large matrix (the number of rows and columns is set to 1000; larger Mascot scores are culled to 1000). In addition, aside from the zero-th column $C(*,0)$, it is quite sparsely populated, as the sum of all entries is 40306. Thus, the above formula yields a very poor estimator. Matrix C therefore needs to be smoothed to matrix C^s using a k-nearest neighbors smoothing algorithm [66]. Specifically, let $N_\delta(i,j) = \{(i',j') : |i-i'| \leq \delta, |j-j'| \leq \delta\}$ be the set of neighboring matrix cells to entry i,j, for some distance threshold δ. For each entry (i,j) in the matrix, we choose δ in such a way that $S_\delta(i,j) = \sum_{(i',j') \in N_\delta(i,j)} C(i',j') \leq k$ and $S_{\delta+1} = \sum_{(i',j') \in N_{\delta+1}(i,j)} C(i',j') > k$. Then, C^s is obtained as:

$$C^s(i,j) = \frac{\sum\limits_{(a,b) \in N_\delta(i,j)} w_{a,b} \cdot C(a,b)}{\sum\limits_{(a,b) \in N_\delta(i,j)} w_{a,b}}$$

where

$$w(a,b) = \begin{cases} 1 & \text{if } (a,b) \in N_\delta(i,j) \\ \frac{k - S_\delta(i,j)}{S_{\delta+1}(i,j) - S_\delta(i,j)} & \text{if } (a,b) \in N_{\delta+1}(i,j) \setminus N_\delta(i,j) \end{cases}$$

In our experiments $k = 10$ has produced the best results.

Finally, we obtain our estimate of the noise in each measurement as

$$\Pr[M_{b,p}^{NI} = x | \bar{M}_p^{NI} = y] \quad = \quad C^s(x,y) / \sum_{x'} C^s(x',y). \tag{2.1}$$

The approach proposed so far is appropriate to model noise in non-induced experiments. However, it does not correctly model noise in induced experiments, for the following reason. The distributions of Mascot scores observed in non-induced and induced experiments differ significantly, with many more high scores observed in the induced experiments. High Mascot scores in non-induced data are rare and, when they do occur, relatively often correspond to cases where a prey p obtains very low scores with most non-induced baits, but a fairly high score with one particular bait, resulting in a large difference between $M_{b,p}^{NI}$ and $\bar{M}_p^{NI}$. Large Mascot scores are thus seen as particularly unreliable, which is the correct conclusion for non-induced experiments, but not for induced experiments. In a situation where high $\bar{M}_{b,p}^{I}$ are frequent, e.g. when b is induced, this does not properly reflect the uncertainty in the true Mascot score. To address this problem, a correction factor is computed. Let $I(i)$ be the fraction of (b,p) pairs such that $\lfloor M_{b,p}^{I} \rfloor = i$ and let $NI(i)$ be the fraction of preys with $\lfloor \bar{M}_p^{NI} \rfloor = i$. The I and NI distributions are first smoothed using a k-nearest neighbor approach, then the corrected matrix C^c for the noise model of Mascot scores for induced experiments is obtained as:

$$C^c(i,j) = C^s(i,j) \cdot \frac{I(j)}{NI(j)}$$

The correction is performed by multiplying each entry (i,j) of C^s by the ratio of the smoothed probabilities of the induced mascot score i and of the true mean Mascot score j. This correction allows a better estimation of the noise in induced experiments

by putting more weight in the matrix C^c for larger Mascot scores. (see Appendix for correction factor derivation.)

2.4.2.2 Step 2: Posterior distribution of $\bar{M}_p^{NI}$ and $\bar{M}_{b,p}^I$.

We are now interested in obtaining the posterior distribution of the mean Mascot score $\bar{M}_p^{NI}$, given the set of observations $M_{b_1,p}^{NI}, ..., M_{b_{14},p}^{NI}$. This is readily obtained using Bayes rule and the conditional independence of the observations, given their means:

$$\Pr[\bar{M}_p^{NI} | M_{b_1,p}^{NI}, ..., M_{b_{14},p}^{NI}] \quad = \quad \Pr[\bar{M}_p^{NI}] \cdot \prod_{i=1}^{14} \Pr[M_{b_i,p}^{NI} | \bar{M}_p^{NI}] / \zeta, \tag{2.2}$$

where ζ is a normalizing constant and $\Pr[\bar{M}_p^{NI} = \alpha] = NI(\alpha)$.

Since observed Mascot scores of induced experiments are also noisy, we estimate the noise of each Mascot score from an induced experiment as:

$$\Pr[\bar{M}_{b,p}^I = j | M_{b,p}^I = i] = C^c(i,j) / \sum_{j'} C^c(i,j')$$

2.4.2.3 Step 3: p-value computation

Given an observed Mascot score $M_{b,p}^I$ for prey p, Decontaminator can now assign a p-value to this score, which represents the probability that $\bar{M}_p^{NI}$, the true Mascot score of p in non-induced experiments, is larger than or equal to $\bar{M}_{b,p}^I$, the true Mascot score for the induced experiment:

$$\begin{aligned} &p\text{-value}(M_{b,p}^I) \\ &\quad = \quad \Pr[\bar{M}_p^{NI} \geq \bar{M}_{b,p}^I | M_{b,p}^I, M_{b_1,p}^{NI}, M_{b_2,p}^{NI}, ..., M_{b_{14},p}^{NI}] \\ &\quad = \quad \sum_{x \geq y} \Pr[\bar{M}_p^{NI} = x | M_{b_1,p}^{NI}, M_{b_2,p}^{NI}, ..., M_{b_{14},p}^{NI}] \cdot \Pr[\bar{M}_{b,p}^I = y | M_{b,p}^I], \end{aligned} \tag{2.3}$$

where the required terms are obtained from Steps 1 and 2.

2.4.2.4 Step 4: False discovery rate estimation

Although our p-values are in principle sufficient to assess the confidence in the presence of a given PPI, they may be biased if some of the assumptions we make were violated. A hypothesis-free method to assess the accuracy of our predictions is to measure a false discovery rate (FDR) for each PPI. For a given p-value threshold t, $FDR(t)$ is defined as the expected fraction of predictions (interactions with p-values below t) that are false positives (i.e. due to contaminants). We use a leave-one-out strategy to estimate the FDR: For every bait b in B, we compute p-value($M^{NI}_{b,p}$) for all preys $p \in P_{b^{NI}}$ (the set of preys detected when b was the control bait) and p-value($M^{I}_{b,p}$) for all preys $p \in P_{b^{I}}$ (the set of preys detected when b was induced), based on the set of non-induced experiments excluding bait b. Then, we obtain

$$FDR(t) = \frac{\frac{\sum_{b\in B}\sum_{p\in P_{bNI}} \mathbf{1}_{p\text{-value}(M^{NI}_{b,p})<t}}{|B\times P_{bNI}|}}{\frac{\sum_{b\in B}\sum_{p\in P_{bI}} \mathbf{1}_{p\text{-value}(M^{I}_{b,p})<t}}{|B\times P_{bI}|}}$$

where $\mathbf{1}_c = 1$ if c is true and 0 otherwise.

2.4.3 Five alternate approaches

We considered five alternate approaches to compare to Decontaminator:

1. Significance Analysis of Interactome (SAINT) [20], without any prior knowledge about hubs or contaminants. SAINT was executed with abundance, sequence length and bait coverage normalization and the following parameters: burn-in period: 2000, iterations: 20000, and empirical frequency threshold: 0.1.
2. SAINT with the same set of parameters except that hubs were manually labelled. All 24 tagged proteins in the dataset that were identified as preys in 10 or more other induced bait experiments were labeled as hubs.

3. The *MScore method* simply computes

$$r(b,p) = \begin{cases} M^I_{b,p} & \text{if } M^I_{b,p} > 5 \cdot M^{NI}_{b,p} \\ 0 & \text{otherwise} \end{cases}$$

and reports the interaction as true positive if $r(b,p) > t$, for some threshold t. This is the approach that was used as primary filtering by Jeronimo et al. [101] and Cloutier et al. [37].

4. The *MRatio method* computes $r'(b,p) = M^I_{b,p}/(1+M^{NI}_{b,p})$. It reports the interaction as true positive if $r'(b,p) > t'$, for some threshold t'. We note the MScore and MRatio approaches are only applicable if both induced and non-induced experiments are performed for *all* baits of interest.

5. The *Z-score method* assigns a Z-score to each Mascot score, as compared to the prey-specific mean $\bar{M}^{NI}_p$ and standard deviation $\sigma(M^{NI}_p)$ for prey p across all baits: $Z\text{-score}(b,p) = (M^I_{b,p} - \bar{M}^{NI}_p)/\sigma(M^{NI}_p)$. Compared to our Bayesian approach, the Z-score method may be advantageous if the Mascot score variance of contaminants with similar averages varies significantly from prey to prey. Our approach, by pooling all non-induced Mascot score observations from different preys, makes the assumption that such variation is low. If this is not the case the modeling accuracy of contaminants will be negatively affected. However, the Z-score approach works under the assumption that noise in Mascot scores is normally distributed which is not always the case. In addition, the mean and variance estimates are obtained from only as many data points as there are control experiments available, as no pooling is performed.

2.4.4 Implementation and availability

The proposed methods are implemented in a fast, platform independent Java program. Given a representative set of AP-MS/MS control experiments our software computes FDRs for all interactions identified in induced experiments as we describe

here. Note that any protein identification mass spectrometry confidence scores (SEQUEST Xcorr, spectral counts, peptide counts, ...) could replace the Mascot scores used in our software package by applying simple modifications. The Java program is available at: `http://www.cs.mcgill.ca/~blanchem/Decontaminator`.

2.5 Results

2.5.1 Contaminant detection accuracy

We start by studying the ability of Decontaminator to tease out contaminants from true interactions. Because neither contaminants nor true interactions are known before hand, one way to assess our method and compare it to others is to consider the number of bait-prey pairs from induced experiments that achieve a p-value at most t to the number of such pairs in non-induced experiments. The ratio of these two numbers, the false discovery rate $FDR(t)$, indicates the fraction of predictions from the induced experiments that are expected to be due to contaminants. One can thus contrast two prediction methods by studying the number of bait-prey pairs that can be detected at a given level of FDR. Figure 2–4 shows the cumulative distributions of FDRs for data from induced experiments, for Decontaminator as well as the Z-score approach described in the Methods section. Since the MScore and MRatio approaches require matched induced and non-induced experiments for every bait, FDRs cannot be computed for them. FDRs comparison is also not possible with SAINT because it scores the entirety of the input dataset without possibility of leaving out a subset of the data. It can be observed that at low FDRs, Decontaminator always yield a larger set of predicted interactions than the Z-score approach. For example, at 1%, 3%, and 15% FDRs, our approach yields 140%, 74%, and 43% more interactions than the Z-score approach. Alternatively, for the same number of interactions predicted, say 2000, the FDR of Decontaminator ($\sim$1%) is more than three times lower than that of the Z-score approach ($\sim$3.4%).

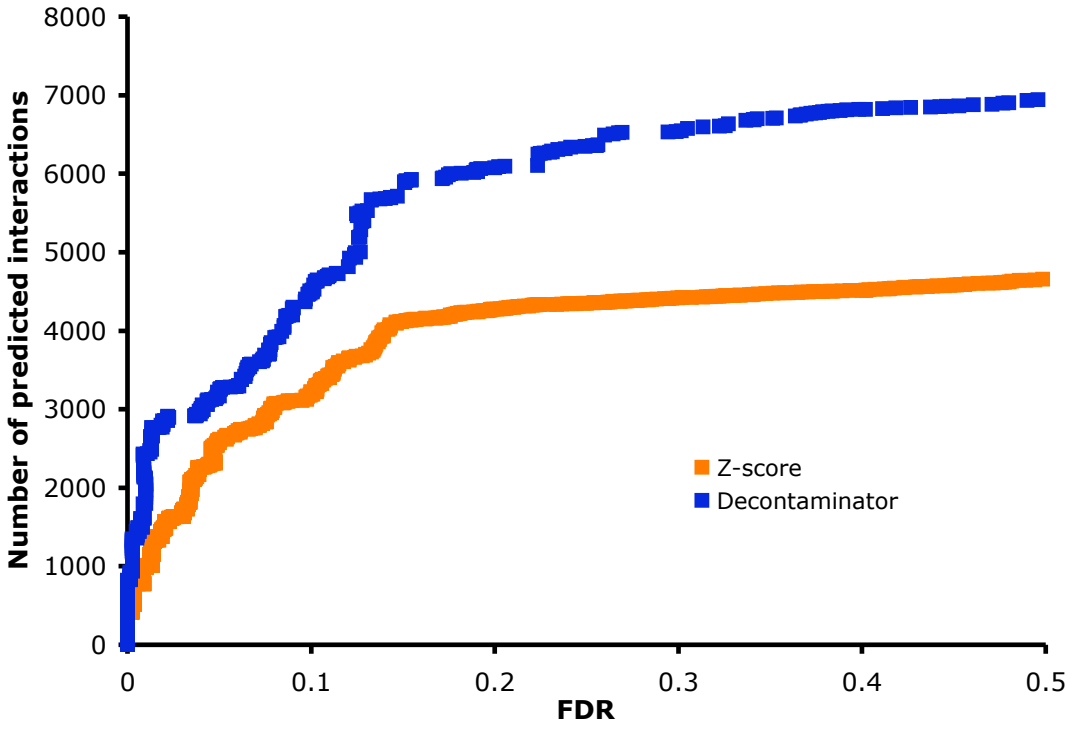

Figure 2–4: Cumulative distributions of the FDRs obtained from Decontaminator and the Z-score approaches. Each curve shows the number of interactions that can be predicted positive, as a function of the false discovery rate tolerated.

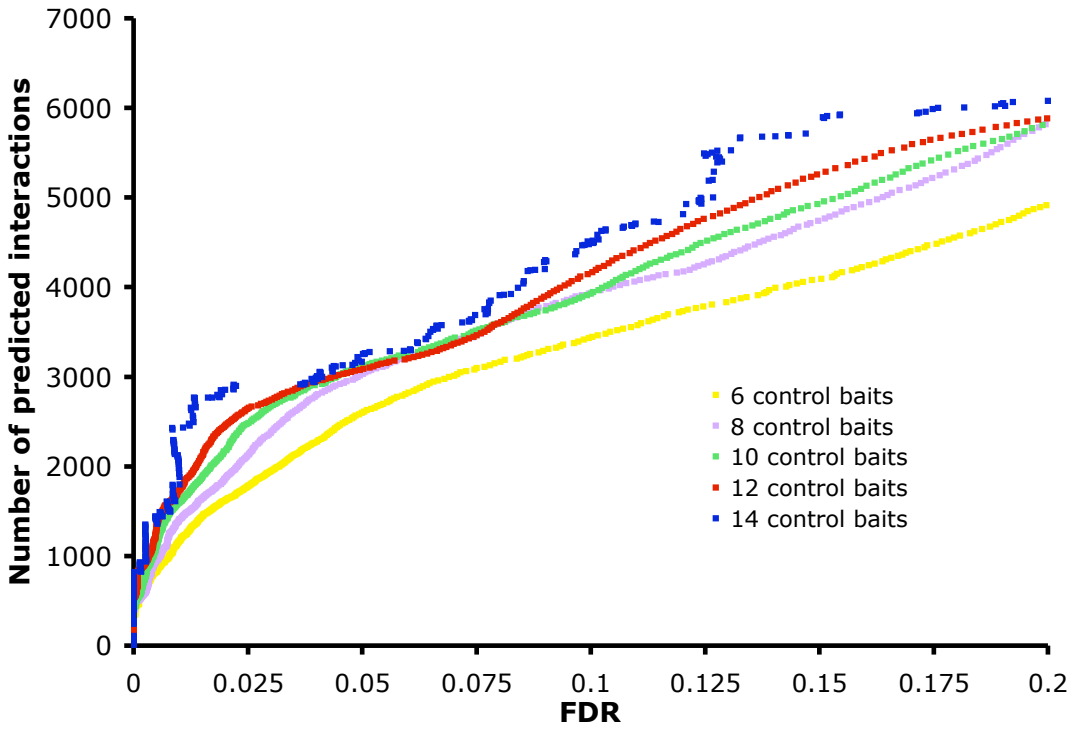

Figure 2–5: Cumulative distributions of the FDRs obtained from Decontaminator with 14, 12, 10, 8, and 6 controls. For sets of controls of size smaller than 14, we show the average cumulative distributions over 100 randomly selected subsets of controls.

2.5.2 Number of control experiments required

The required number of control non-induced experiments needed to build an accurate model of contaminants is an important issue to consider. So far we used all 14 available controls to filter out contaminants. Figure 2–5 shows the cumulative distributions of the FDRs obtained from Decontaminator with different number of controls. It can be observed that there is an increase in the number of high confidence ($< 20\%$ FDR) predictions when using a larger number of control baits. For example at 2% FDR we obtain 2855 interactions using the set of 14 control experiments but only 2440 with 12 controls, 2169 with 10 controls, 1858 with 8 controls and 1620 with 6 controls. These results show that the more high quality control baits are available, the better our algorithm will perform. It can also be observed that the prediction accuracy decreases significantly when fewer than 8 controls are used. However, this deterioration is not striking and therefore shows that even with very few control experiments it is possible to use Decontaminator to detect contaminants in AP-MS/MS experiments reasonably accurately. This aspect is particularly important for small scale studies where only a few baits are analyzed and for laboratories where the number of control experiments that can be performed is limited.

2.5.3 External validation

In order to evaluate the quality of our contaminant detection method we compared it to the five other methods described above on the basis of their ability to detect known PPIs or PPIs involving pairs of proteins of similar function.

We first used the union of two high-quality PPI databases, BioGRID [210] and HPRD [179] (both downloaded on Feb 1^{st} 2010), to assess the quality of the predictions made by each method. Note that BioGRID PPIs that originated from the curation of our own previously published dataset [101] were excluded from the analysis. Since a small fraction of all PPIs are known, we do not expect a large fraction

of our predictions to be found in these databases. However, clearly, the size of the overlap is a good indication of the accuracy of the methods. Figure 2–6 shows the fraction of positively predicted PPIs present in the merged database for the six methods described above, as a function of the number of PPIs predicted. For all six methods, high-confidence predictions overlap the two databases significantly more than low-confidence predictions. However, at any high confidence FDR level, the set of predicted interactions produced by Decontaminator always has a larger overlap than any of the alternate methods. The Z-score method also outperforms the other four alternate methods. For example, with a predicted set of 2000 PPIs (FDR of 1% for Decontaminator), 7.1% of the interactions overlap with the merged database for Decontaminator, 6.6% for the Z-score approach, 5.8% for SAINT, 5.7% for SAINT without manual labeling of hubs, 5.8% for the MScore, and only 4.9% MRatio methods. On average, Decontaminator results in a predicted set of PPIs with $\sim$ 40% more overlap with known PPIs than that obtained with the MRatio methods from Jeronimo et al. [101], $\sim$ 20% more than both types of runs of SAINT, $\sim$ 15% more than the MScore method, and $\sim$ 5% more than the Z-score method. We assessed the statistical significance of the differences between the overlaps obtained by each method using a two sample z-test. Decontaminator's performance significantly exceed those of the two SAINT variants, MScore, and MRatio (p-value ≤ 0.05), but is not significantly better than that of the Z-score approach (p-value $= 0.26$).

The Gene Ontology (GO) annotation database [4] associates a set of terms to characterized proteins, describing their functions, localization, and the biological processes in which they are involved. We say that a GO term is x%-specific if less than x% of the proteins in our data set are annotated with this term [37]. We assume that if two interacting proteins are sharing a x%-specific term (for x relatively small), they have greater chances to be truly interacting. Therefore, we applied the six filtering methods on our data set and compared the fraction of the positively predicted interactions for which the proteins were sharing at least one 10%-specific GO term

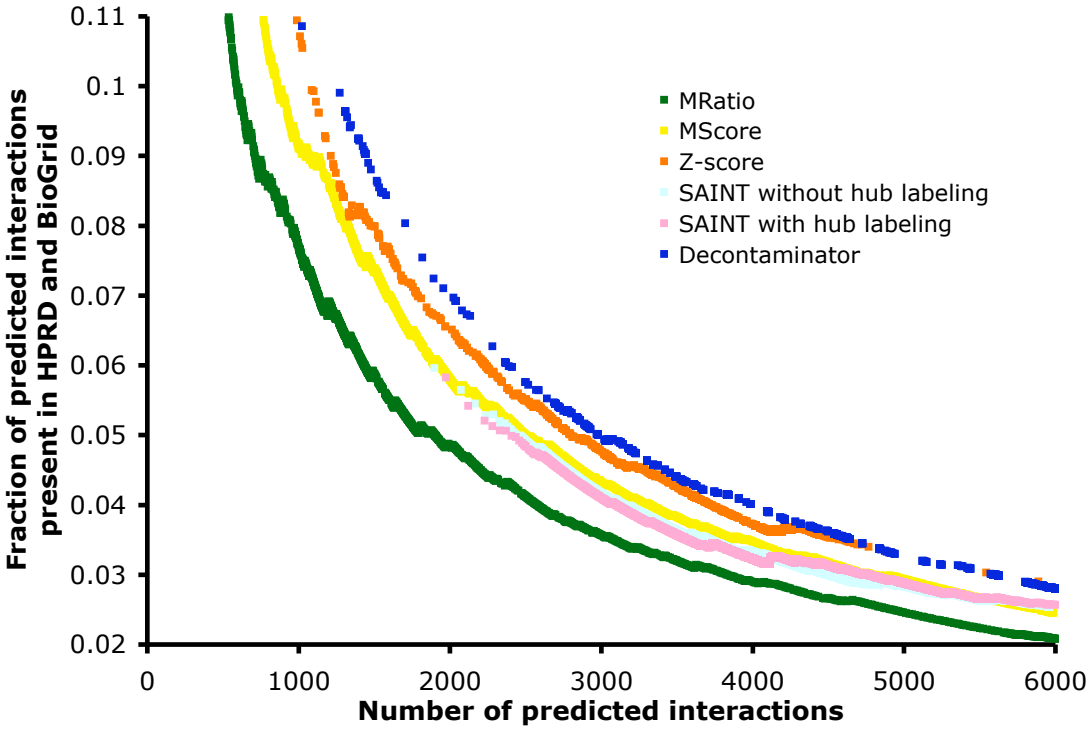

Figure 2–6: Fraction of positive predictions present in the HPRD and BioGRID merged databases (y-axis), for a varying number of predicted interactions (x-axis) by six filtering methods(MRatio, MScore, Z-score, SAINT without hub labeling, SAINT with hub labelling and Decontaminator.

(Figure 2–7). The PPIs predicted by our Bayesian approach are consistently more supported by shared GO annotations than those made by the five other approaches. For a set of 2000 positively predicted PPIs, 28.6% of the PPIs predicted by Decontaminator had both proteins sharing at least one 10%-specific GO term, compared to 26.7% with the Z-score, 26.3% with the MScore, 23.8% with SAINT and 23.7% with SAINT without hub labeling and 23.5% with the MRatio. The advantage of Decontaminator is statistically significant for all comparisons (p-value < 0.05; two-sample z-test), except against the Z-score approach, for which the advantage is marginal (p-value $= 0.095$). Similar patterns are observed for most other values of GO term specificity (x). Even though sharing a specific molecular function, biological process or cellular component does not guarantee that two proteins are interacting, we believe that this improved enrichment for shared GO terms reflects the higher quality of our contaminant detection method.

Overall, one might argue that the differences between Decontaminator and the Z-score approach are not striking. However, when referring back to Figure 2–4, it is clear that Decontaminator yields a much larger set of predictions for a given FDR when compared to the Z-score approach. It is possible that the external validation sets chosen are too small to demonstrate the importance of the improvement of Decontaminator over the Z-score approach as easily as it can be seen in Figure 2–4.

2.6 Discussion

Decontaminator shows an improvement in accuracy for the detection of contaminant PPIs in our dataset when compared to currently used alternate approaches. We expect that this decrease in false positive interactions will facilitate the analysis of PPI networks and ease the characterization of novel biological pathways. At the same time, our approach will greatly reduce experimental costs by cutting the number of most experimental manipulations almost in half. This expense reduction is due to the much smaller number of control experiments needed by our algorithm compared to

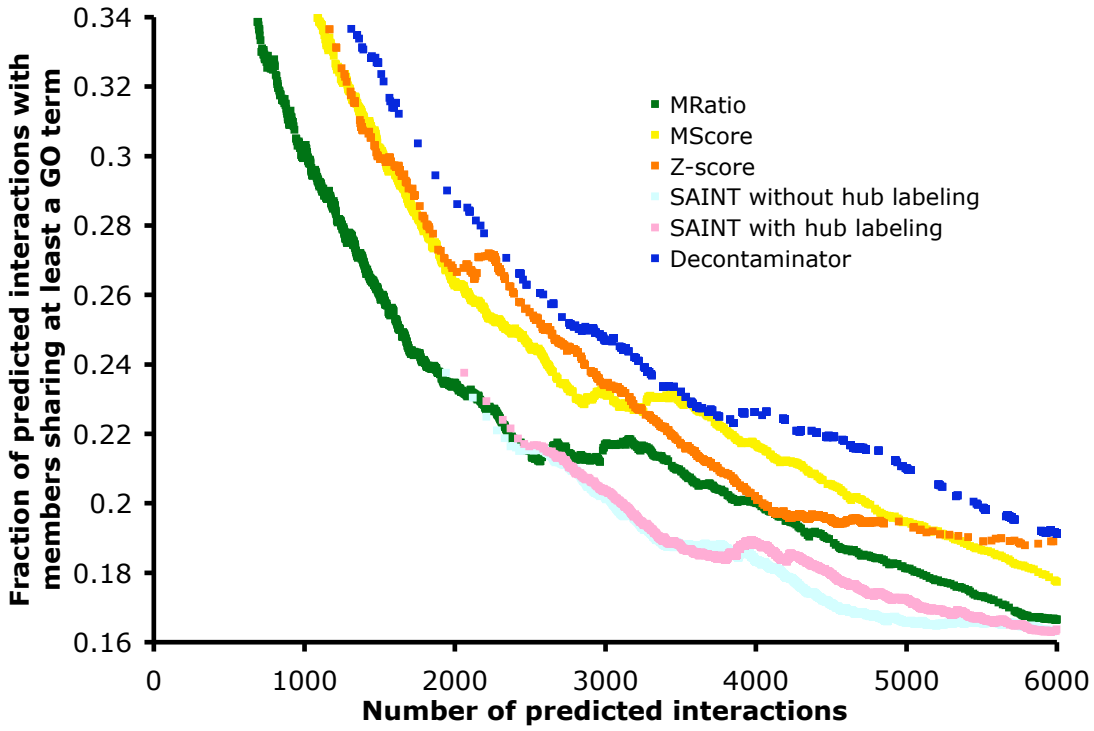

Figure 2–7: Fraction of positively predicted PPIs for which the interacting partners share at least a 10%-specific GO term (y-axis), for a given number of predicted interactions (x-axis) by six filtering methods(MRatio, MScore, Z-score, SAINT without hub labeling, SAINT with hub labeling and Decontaminator.

methods such as that described in Jeronimo et al. [101], where each induced experiment requires a matched non-induced experiment for its interactions to be classified. It is also worth noting that in theory, the control experiments provided as input to the algorithm could all be performed with the same bait protein. However, we used non-induced experiments produced from different baits, by different experimentalists at different time periods. These biological and technical replicates allow us to factor in the noise resulting from the change of baits in TAP-MS/MS experiments and technical variation.

2.6.1 On the impact of the protein-protein interaction discovery experimental design

PPIs are often viewed and studied as a network. Different strategies have been applied in order to decide which proteins should be tagged first in order to build this network. Some groups use a molecular function centric approach and choose as baits a set of proteins performing a specific type of molecular functions, such as kinases, methyltransferases, or phosphatases [20]. Others use a complex-centric approach and orient their study around specific biological processes and complexes, such as the RNA polymerase II [159, 101, 37], tagging several interacting partners to obtain a dense, focussed network. Other impressive projects focused on whole interactome mapping of particular organisms by tagging most of their known proteins to analyze the protein complexes present in their interaction networks [119, 74, 61]. The PPIs obtained with those strategies form networks that have drastically different topologies. Networks obtained by complex-centric or whole interactome mapping approaches will tend to be more connected than those obtained by a molecular function centric strategy. This will largely influence the type of computational analysis that should be used in order to filter contaminants and assign confidence scores to interactions. Several algorithms use the topology of PPI networks to determine whether an interaction is a likely true or false positive [101, 74, 39], reasoning that truly interacting proteins are likely to interact with similar sets of other proteins. However, this method only applies

to networks obtained from complex-centric and whole interactome mapping studies, as molecular function centric approaches yield loosely connected networks, making shared neighborhood a rarity. Thus, for small or sparse networks, topology is of little use, making approaches such as SAINT or Decontaminator the only alternative.

2.6.2 Advantages

Decontaminator is an alternative that does not rely on topological data and is not affected by the type or size of the network being analyzed. Obviously, in cases where the network topology is informative, our Bayesian FDR scores can be integrated into more complex predictors such as our Interaction Reliability Scores (IRS) [37], the Socio-affinity index [74] or the Purification Enrichment score [39]. Even though we applied our algorithm on a large scale data set in this paper, we showed that it can be applied to much smaller scale studies since a small number of controls are required to accurately model contaminants and score interactions.

Another important factor to consider is the need for prior knowledge in training the prediction program. Methods such as the Interaction Reliability Scores (IRS) [37] or the machine learning pipeline used by Krogan et al. [119] require a fairly large training set of examples of true positive and possibly true negative interactions. These training sets are difficult to assemble, prone to errors, and often not representative of the set of interactions one is really seeking (e.g. true positive interactions come from very strong complexes such as RNA polymerase II, whereas interactions of interest are weaker or more transient). On the other hand, SAINT requires no labeled training set, but it is reported to perform better when known contaminants and hubs are manually labeled. Such labeling might be hard to perform on smaller networks where hubs and contaminants are hard to differentiate. Thus, methods requiring no prior knowledge, such as Decontaminator, offers significant advantages.

It should also be noted that Decontaminator can easily be incorporated in a mass spectrometry analysis computational pipeline. Because it provides, for each predicted PPI a FDR, which can be interpreted as the probability that the PPI is involving a contaminant, our method can be added as a contaminant filter at the end of the typical computational pipelines involving identification of the preys via database searches (Mascot [174] and SEQUEST [58]) and validation of the identification (Percolator [105] and Protein Prophet [161]).

2.6.3 Limitations

Our approach works under the assumption that all experiments are performed using the same protocol, in such a way that the distribution of contaminants does not change over time (note that this does not mean that the observed levels of contaminants are constant, but rather that the true levels are). Clearly, whenever the experimental pipeline of the approach is modified (e.g. changes to the experimental protocol, the mass spectrometer, the tag, or the chromatography column used), new control experiments need to confirm the validity of the model or to build a new one. Therefore, in a context where the experimental pipeline would be rapidly evolving or when bait-specific antibodies are used for the pull down, Decontaminator may yield little or no cost benefit compared to the approach where each induced experiment is paired with its control counterpart. However, our results indicate that the number of control experiments required to obtain a good contaminant model is relatively modest, so unless changes are extremely common, significant benefits should be achievable. Finally, even when no obvious changes to the experimental pipeline have taken place, periodic control experiments should be performed in order to ensure that the set of control experiments used by the Bayesian approach remains representative.

Another limitation is that, Decontaminator cannot differentiate true positive interactions from contaminants obtained from inefficient purifications. Instead, the Mascot scores obtained for contaminants that are usually excluded at the purification

step are likely to be significantly larger than what was observed in control experiments. However, other approaches can be used to detect such problematic purifications. For example, typically, the distribution of Mascot scores observed in a faulty purification is significantly shifted to the right (greater number of high Mascot scores). This can be explained by the high abundance of various proteins, causing more peptides to be detected and therefore increasing the overall Mascot score of each protein. Also, a GO analysis can be performed to accomplish the same goal: preys interacting with a bait for which the purification was faulty will often not be enriched for specific GO terms. On the contrary, one would expect that when an efficient purification is performed, the preys interacting with a given bait will be more likely to share functions, biological processes or cellular components with it.

Mascot scores, which are used in the current version of Decontaminator, are not always accurate in the identification of peptides and quantification of their abundance. Peptide misidentifications will affect the accuracy of our method in several ways. For example, a contaminant that would have been misidentified in many of the control experiments may be predicted as a true interaction when correctly identified in an induced experiment. In addition, Mascot scores, like peptide and spectral counts, are influenced by the length of the protein and the detectability of its peptides, resulting in some true interactions obtaining relatively low scores. These problems can be overcome with the utilization of software like Peptide/Protein Prophet [109, 161] which give a probability of the protein presence based on database search scores like Mascot. Other measures like peptide retention time and precursor ion intensity have also shown to be valuable information and could be used in conjunction with database search scores in order to yield more accurate protein identifications [211, 108, 87]. These more accurate measures of protein identification/abundance could easily replace the Mascot score in Decontaminator.

Finally, our approach will work best if different preys with the same true mean Mascot score have the same Mascot score distribution in the non-induced condition,

i.e. that the variability in the observed Mascot scores is not dependent on the identity of the prey. Manual inspection confirms that this is the case for the vast majority of the preys in our dataset. Should this assumption fail, i.e. if the Mascot score of a certain protein had a much larger variance than other proteins with the same mean Mascot score, the consequences would be that its p-value would tend to be assessed incorrectly (above average Mascot scores would obtain unduly low p-values, and below average score unduly high p-values). Indeed, some of the errors made by our approach are due to this type of proteins, such as CKAP5, INF2, and TUBB2A. This large variability could be in part explained by the large size ($>$ 1200 a.a.) of these proteins. These preys could also be suffering of an undersampling by the mass spectrometer and be masked due to the presence of more abundant proteins. However, those are rare cases that can be treated separately by flagging, from the set of non-induced experiments, proteins with unexpectedly high variance, and analyzing them separately.

2.7 Conclusion

Several methods have been proposed to identify false positive PPIs in AP-MS/MS experiments. However, very few have considered the modeling of contaminants resulting only from AP experimental pipeline. We hypothesized that a more accurate model of contaminants would yield higher accuracy of PPI identification. We have shown here that using Decontaminator, only a few representative control experiments are necessary to accurately discard the vast majority of contaminants while allowing the detection of true PPIs involving preys that, in other experiments, may be contaminants. These findings will allow significant reductions in expenses and a greater number of experiments to be conducted with higher accuracy.

2.8 Acknowledgements

This work was funded by a CIHR operating grant to BC and MB and NSERC Vanier and CGS scholarships to MLA. The authors thank Christian Poitras for his help in the software development process, Denis Faubert for the mass spectrometry analysis, and Ethan Kim for useful suggestions.

2.9 Appendix

2.9.1 Protein and peptide identification software information

Peaklists were created using extract_msn.exe version 2005-02-15 (Thermo Xcalibur) with the following parameters: minimum mass: 600, maximum mass: 6000, minimum number of fragment ions: 10, no grouping of MS/MS spectra was performed, and precursor charge was set to automatic. Mascot 2.2.04 (Matrix Science) was used for protein database searching with precursor-ion mass tolerance set to 10 ppm and fragment-ion mass tolerance set to 0.6 Da. The modifications allowed were carbamidomethylation and oxidation of methionine. Finally, the digestion enzyme used was trypsin and 2 missed cleavages were allowed. Database searching was performed on the human NCBI nr protein database (version 2009-04-02), which contains 10 427 007 sequences.

2.9.2 Computation of corrected averages for Mascot scores

We note that in most AP-MS/MS applications, preys with Mascot scores below a certain threshold m (e.g. a fixed value $m = 20$, or the Mascot Identity Threshold) are discarded and not reported, as being likely protein identification errors. In our approach, when a protein p is not reported as a possible partner of bait b, we arbitrarily set its Mascot score $M_{b,p}^{NI}$ to zero. The set of observed Mascot scores for a given prey thus follow a type I censored distribution [16]. Let B'_p be the set of control experiments for which $M_{b,p}^{NI} < m$. Assuming the uncensored $\bar{M}_p^{NI}$ follows a normal distribution, a better estimate of $\mu_{\neq b,p}$ is thus obtained from the Persson-Rootzen method [175]:

$$\mu_{\neq b,p} = \frac{1}{|B'_p|} \sum_{b \in B'_p} M_{b,p}^{NI} - \gamma_p \sigma',$$

where $\gamma_p = \phi(\lambda_{|B'_p|/|B|})|B|/|B'_p|$, ϕ is the probability density function of the standard normal distribution,

$$\sigma' = \frac{1}{2}\left[\lambda_{|B'_p|/|B|}\frac{1}{|B'_p|}\sum_{b\in B'_p}(M^{NI}_{b,p}-m)+\left\{\left(\lambda_{|B'_p|/|B|}\frac{1}{|B'_p|}\sum_{b\in B'_p}(M^{NI}_{b,p}-m)\right)^2+\frac{4}{|B'_p|}\sum_{b\in B'_p}(M^{NI}_{b,p}-m)^2\right\}^{\frac{1}{2}}\right],$$

and where $\lambda_{|B'_p|/|B|}$ denotes the upper $(|B'_p|/|B|)^{th}$ quantile of the standard normal distribution. If $M^{NI}_{b,p} = 0\ \forall b \in B$ for a given p, we set arbitrarily one $M^{NI}_{b,p}$ to be equal to $m+1$.

2.9.3 C^c correction factor derivation

In order to correct the C^s matrix for induced experiments noise modeling, we used the following correction:

$$C^c(i,j) = C^s(i,j)\cdot\frac{I(j)}{NI(j)}$$

The above correction was derived the following way. The matrix C^s corresponds to the joint probability of $\bar{M}^{NI}_p$ and $M^{NI}_{b,p}$ given that the data was generated from control experiments.

$$C^s(i,j) = \Pr[\bar{M}^{NI}_p = j, M^{NI}_{b,p} = i|control] = \Pr[\bar{M}^{NI}_p = j|control]\cdot\Pr[M^{NI}_{b,p} = i|\bar{M}^{NI}_p = j, control]$$

We make the assumption that the noise of a Mascot score is independent of whether the experiment was induced or not. Therefore:

$$\Pr[M^{NI}_{b,p} = i|\bar{M}^{NI}_p = j, control] = \Pr[M^{NI}_{b,p} = i|\bar{M}^{NI}_p = j]$$

Similarly, let C^c correspond to the joint probability of $\bar{M}^{NI}_p$ and $M^{NI}_{b,p}$ given that the data was generated from induced experiments.

$$C^c(i,j) = \Pr[\bar{M}_p^{NI} = j, M_{b,p}^{NI} = i|induced] = \Pr[\bar{M}_p^{NI} = j|induced] \cdot \Pr[M_{b,p}^{NI} = i|\bar{M}_p^{NI} = j]$$

Following from the above assumption,

$$C^c(i,j) = \Pr[\bar{M}_p^{NI} = j|induced] \cdot \frac{\Pr[\bar{M}_p^{NI} = j, M_{b,p}^{NI} = i|control]}{\Pr[\bar{M}_p^{NI} = j|control]}$$

which give us the correction factor for C^s in order to get C^c.

$$C^c(i,j) = C^s(i,j) \cdot \frac{\Pr[\bar{M}_p^{NI} = j|induced]}{\Pr[\bar{M}_p^{NI} = j|control]}$$

or as described in the Methods section:

$$C^c(i,j) = C^s(i,j) \cdot \frac{I(j)}{NI(j)}$$

CHAPTER 3
Discovery of cell compartment specific protein-protein interactions using AP-MS/MS

3.1 Preface

Once the reliability of individual PPI predictions is assessed with a tool such as the one presented in Chapter 2, it is possible to select a subset of PPIs that will satisfy a certain confidence score threshold. This high quality dataset can provide much information on the proteins it contains. Approaches identifying protein complexes and inferring protein functions benefit from such removal of noisy PPIs. On the other hand, these methods also gain from an increase in sensitivity of the experiments identifying PPIs. Sensitivity improvements, in the context of AP-MS, can be obtained, for example, by performing only a single purification of the bait (FLAG instead of TAP) [2] to detect more transient interactions, which could be lost in subsequent purifications. Gains can also be made with the use of MS instruments of higher sensitivity that allow the detection of low abundance proteins. Clearly, sensitivity improvements also lead to an increased number of contaminants being identified. However, tools like Decontaminator lessen this problem and facilitate the implementation of methodologies bringing sensitivity increases to the AP-MS pipeline.

Another way to improve sensitivity in MS is at the sample preparation stage. As discussed previously in Chapter 2, separation of the input protein/peptide mixture with techniques such as gel separation or LC is crucial. Both methods can be performed in a multidimensional system in order to separate proteins/peptides according to different molecular properties (charge, hydrophobicity, etc.) [113]. This allows the resolution of complex protein/peptide mixtures and improves detection at the MS

level [143]. In addition to separation, MS sensitivity can be improved at the sample preparation through fractionation of cells into different compartments (nucleus, endoplasmic reticulum, mitochondria, cytoplasm, etc.) [117, 42]. This fractionation leads to simpler mixtures and therefore, to better detection. Cell fractionation improves sensitivity simply by allowing the mixture to be analyzed in multiple MS runs (one run per fraction). As a result, the entire original mixture will be allocated more time in MS analysis than if it was analyzed as a whole. Moreover, with this scenario, fractions containing low abundance proteins are less likely to be masked and undetected in MS by fractions with proteins of higher abundances. The methodology presented in this chapter is inspired by the cell fractionation approach. Basically, it identifies the PPIs of a given protein independently in three cell compartments (cytoplasm, chromatin, and nucleoplasm) using a modified AP-MS pipeline. The approach presented here includes a improved version of Decontaminator, which can make the use of controls obtained from all three cell compartments to assess the quality of an interaction detected in a given cell fraction. We show in this chapter that this strategy improved the performances of Decontaminator. Although this behaviour was rarely observed, it could however have resulted in a loss of sensitivity if an important number of contaminants had drastically different abundances across the three different cell compartments.

Interestingly, the approach presented here does not only improve the sensitivity of AP-MS. It also helps deconvoluting PPI datasets and PPI networks. Such datasets are often very large, complex, and therefore difficult to analyze. One explanation for such complexity resides in the fact that PPI datasets include all (or a subset of) PPIs happening in different cell compartments. More precisely, if a given protein interacts with a set of proteins in the cytoplasm, then it is possible that the same protein will interact with a disjoint set of proteins in the nucleus. Nevertheless, the PPIs reported for such protein would be the union of these two sets without ways to track the compartment where the interactions take place. The approach introduced

in this chapter aims, among other things, at deconvoluting PPI datasets using a spatial dimension. Our methods address this problem by identifying interactions for a given protein independently from the cytoplasm, the chromatin, and the nucleoplasm. Using this approach, a PPI network can be built for each cell fraction. These three networks constitute a precious resource to understand cellular mechanisms that span multiple cell compartments. They are also an important source of information for the functional characterization of proteins by providing their localization as well as their interactions, as it will be demonstrated in this chapter.

The remaining content of this chapter is reprinted with permission from:

- M. Lavallée-Adam, J. Rousseau, C. Domecq, A. Bouchard, D. Forget, D. Faubert, M. Blanchette, and B. Coulombe. Discovery of cell compartment specific protein-protein interactions using affinity purification combined with tandem mass spectrometry. *Journal of proteome research*, 12(1):272–281, 2012

3.2 Abstract

Affinity purification combined with tandem mass spectrometry (AP-MS/MS) is a well established method used to discover interaction partners for a given protein of interest. Because most AP-MS/MS approaches are performed using the soluble fraction of whole cell extracts (WCEs), information about the cellular compartments where the interactions occur is lost. More importantly, classical AP-MS/MS often fails to identify interactions that take place in the non-soluble fraction of the cell, e.g. on the chromatin or membranes, and, consequently, protein complexes that are less soluble are underrepresented. In this paper, we introduce a method called multiple cell compartment affinity purification coupled to tandem mass spectrometry (MCC-AP-MS/MS), which identifies the interactions of a protein independently in three fractions of the cell: the cytoplasm, the nucleoplasm, and the chromatin. We

show that this fractionation improves the sensitivity of the method when compared to the classical affinity purification procedure using soluble WCE, while keeping a very high specificity. Using three proteins known to localize in various cell compartments as baits, the CDK9 subunit of transcription elongation factor P-TEFb, RNA polymerase II (RNAP II)-associated protein 4 (RPAP4), and the largest subunit of RNAP II, POLR2A, we show that MCC-AP-MS/MS reproducibly yields fraction-specific interactions. Finally, we demonstrate that this improvement in sensitivity leads to the discovery of novel interactions of RNAP II carboxyl-terminal domain (CTD) interacting domain (CID) proteins with POLR2A.

3.3 Introduction

Large-scale mapping of human protein-protein interactions (PPIs) not only leads to the discovery of new protein functions, but also to a better understanding of several biological processes. Medium to high throughput PPI detection approaches are essential for comprehensive network mapping. These include the Y2H method [96, 224], the PCA [183] and the AP-MS/MS [119, 69, 226, 101, 20, 74]. Of those, AP-MS/MS has a number of advantages. First, this method accurately detects entire complexes when tagging only one protein (bait) [184]. PPIs obtained through AP-MS/MS can be direct or indirect interactions resulting from a co-complex bait-prey association. Second, it can be used in any organism or cell type of interest. Finally, since it does not use a heterogeneous host, post-translational modifications necessary for a given interaction to take place, may occur normally in the cell where the tagged protein is expressed.

Classical AP-MS/MS experiments, performed using whole cell extracts, purify proteins present in the soluble fraction of the cell [69, 226, 101], which includes the cytoplasm and the nucleoplasm, but not proteins tightly bound to chromatin or membranes. Consequently, this approach yields no information about the exact compartment where detected interactions are taking place. Over the years, some purification

approaches have been proposed to circumvent this limitation. Techniques have been developed to target interactions of membrane proteins [176]. These are based on sonication and sedimentation of membranes through differential centrifugation followed by solubilization of membrane proteins [176]. Recently, approaches to identify interactions for chromatin-bound proteins have been developed [125], including the modified chromatin immunopurification (mChIP) method that purifies protein-DNA macromolecules in yeast through mild sonication in order to minimize chromatin fragment precipitation [124]. Another way to solubilize chromatin makes use of nucleases. Foltz et al. used affinity purification on soluble small DNA fragments digested with a micrococcal nuclease [68], while Du et al. used DNase I, digesting DNA completely [53]. Lambert et al. addressed the impact of chromatin fragment size on affinity purification. They have shown that larger DNA fragments will favor indirect protein interactions [124], thereby complicating the interpretation of the results. Thus, the method of Du et al., with its complete DNA digestion, seems to be the best approach to minimize indirect protein interactions through DNA. However, they used the FLAG affinity purification technique [53], which may be more sensitive to detect transient interactions, but can also lead to more non-specific interactions when compared to tandem affinity purification (TAP), because of the number of purifications performed (i.e. a single affinity purification step for FLAG as opposed to two for TAP [2]). Although these methods have their respective advantages and disadvantages, there is a need for a method detecting PPIs occurring on the chromatin, while yielding as few as possible indirect interactions through DNA and protein contaminants.

Accurate identification of the cellular compartment(s) where an interaction takes place is often critical to understand the function of that interaction. This information can also be crucial for the experimentalist as it can guide validation experiments and speed up discovery. For instance, we recently established a role of the RNAP II-Associated Protein 3 and 4 (RPAP3 and RPAP4) in assembly and nuclear import of the RNAP II enzyme [37]. Because the study was initiated on the basis of AP-MS/MS

data from soluble WCE, educated guesswork followed by laborious experimental work was required to characterize the role of these two proteins. As we will show, the technique introduced here localizes the interactions between RPAP3/RPAP4 and RNAP II to the cytoplasm and nucleoplasm, which would have immediately suggested a putative role in some cytoplasmic functions such as RNAP II assembly and/or nuclear import. Similarly, discovering interactions taking place on chromatin could quickly direct hypotheses to a DNA binding role for the PPIs in question.

We therefore propose a new approach to map and localize PPIs in a more comprehensive manner. The method, called multiple cell compartment affinity purification coupled with tandem mass spectrometry (MCC-AP-MS/MS), can detect PPIs independently, from the same starting material, through TAP in three different cell compartments: in the cytoplasm, the nucleoplasm, and on the chromatin based on several centrifugations and a complete digestion of DNA. We show that separating a typical WCE into these three fractions yields crucial information about the detected interactions, as well as an increase in sensitivity through fractionation of the sample. We show that MCC-AP-MS/MS identifies fraction-specific PPIs and increases the sensitivity over WCE AP-MS/MS, while keeping high specificity and reproducibility. Moreover, we report the discovery of novel, compartment-specific, potentially biologically relevant PPIs for POLR2A.

3.4 Experimental procedures

We propose an approach for performing independent AP-MS/MS on three different cell compartments: cytoplasm, nucleoplasm, and resolubilized chromatin. The method then distinguishes bait specific interactions from contamination with a new version of our Decontaminator software [127]. In this section, we describe the MCC-AP-MS/MS methodology and its associated computational component. Figure 3–1 summarizes the steps of the procedure.

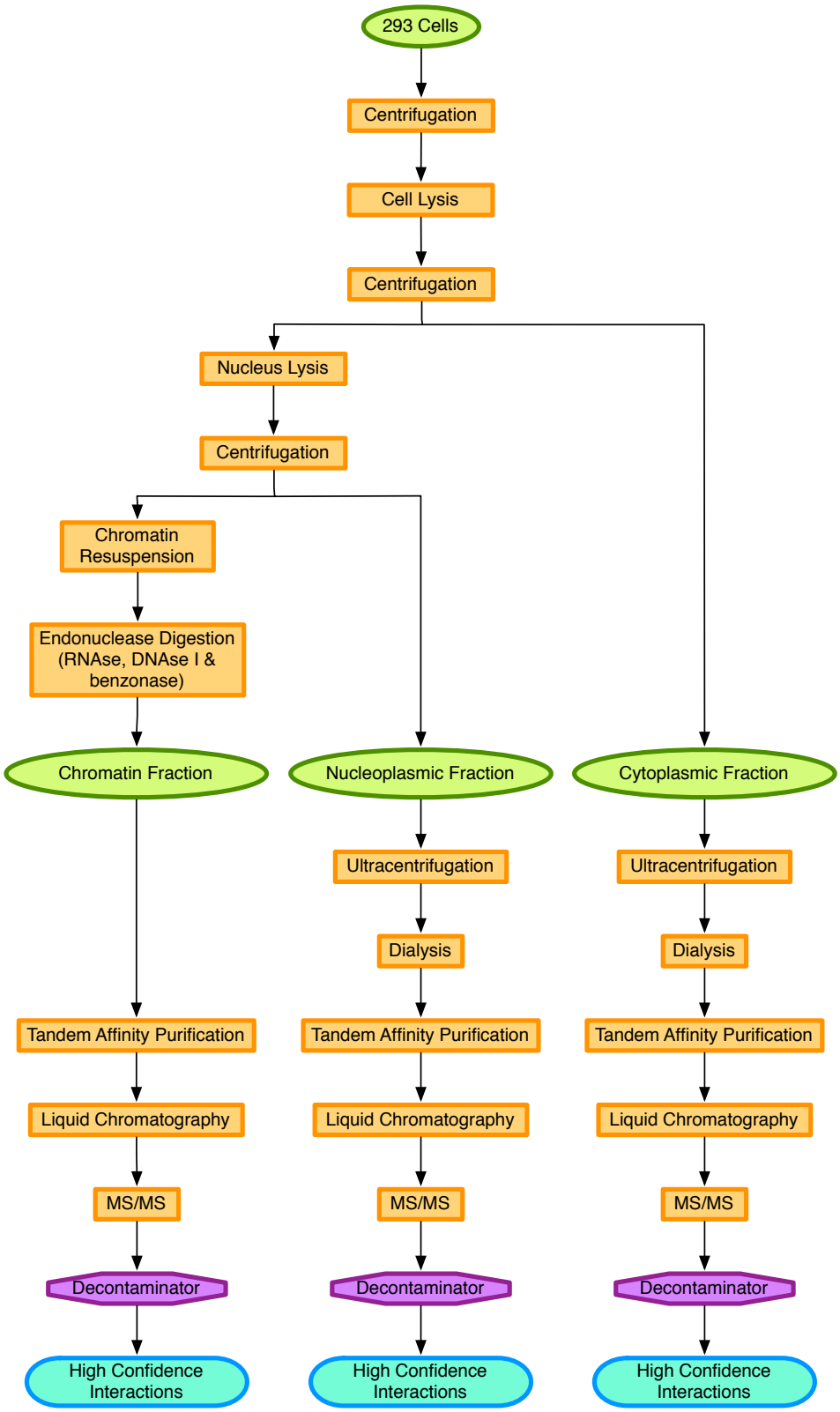

Figure 3–1: Pipeline of MCC-AP-MS/MS and its associated computational methods.

3.4.1 Cytoplasmic, nuclear, and chromatin fractions

ORFs encoding human POLR2A, CDK9, PPARG2, RPAP4, KLF14, FTO, IRS1, RPAP2, RPAP3, and JAZF1 were cloned into the mammalian expression vector pMZI [233] containing the protein A-Calmodulin Binding Peptide tag at the 3' end of the MCS in order to generate a fusion protein with a TAP tag at its C-terminus [184]. EcR-293 (Invitrogen) inducible stable cell lines carrying these constructs were produced as previously described [102] and selected clones were grown to obtain $2g$ of cell pellet. To generate the cytoplasmic, nucleoplasmic, and chromatin fraction, WCE was prepared as previously described [5] with some modifications. Cells were lysed by mechanical homogenization in lysis buffer [10 mM Tris-HCl (pH8), 0.34 M sucrose, 3 mM $CaCl_2$, 2 mM MgOAc, 0.1 mM EDTA, 1 mM DTT, 0.5% Nonidet P-40 and protease inhibitors]. WCE was centrifuged at $3,500$ x g, 15 min and, the supernatant, which represents the cytoplasmic fraction, was saved. The pellet containing the nuclei was resuspended, lysed by mechanical homogenization in lysis buffer [20 mM HEPES (pH 7.9), 1.5 mM $MgCl_2$, 150 mM KOAc, 3 mM EDTA, 10% glycerol, 1 mM DTT, 0.1% Nonidet P-40 and protease inhibitors] and, centrifuged at $15,000$ x g, 30 min. The supernatant, which corresponds to the nucleoplasmic fraction, was saved. The chromatin pellet was then minced with a scalpel in nuclease incubation buffer [150 mM Hepes (pH 7.9), 1.5 mM $MgCl_2$, 150 mM KOAc, 10% glycerol, and protease inhibitors] and disrupted mechanically using a glass homogenizer. Nuclease was then added and the chromatin fraction was digested overnight [0.15 $unit/\mu L$ benzonase (Novagen), 0.44 $unit/mL$ RNase A and 6.25 $units/mL$ DNaseI]. Cytoplasmic and nucleoplasmic fractions were centrifuged at $124,000$ x g and dialyzed overnight in dialysis buffer [10 mM Hepes (pH 7.9), 0.1 mM EDTA (pH 8), 0.1 mM DTT, 0.1 M KOAc and 10% glycerol]. The following day, the three fractions were clarified by centrifugation at $20,000$ x g for 30 min, and the supernatants containing the solubilized proteins were collected.

3.4.2 Tandem affinity purification

WCE prepared from induced stable EcR-293 cell lines was subjected to purification by the TAP procedure as previously described [102]. The eluates were precipitated with trichloroacetic acid and stored at -80^oC until analysis by LC-MS/MS. In parallel for some TAP experiments, part of the eluate (starting material was adjusted accordingly) were concentrated, loaded in $4-12\%$ bis-Tris gradient PAGE, and colored by silver staining.

3.4.3 Protein digestion with trypsin

Protein extracts were then re-solubilized in 10 μL of a 6 M urea buffer. Proteins were reduced by adding 2.5 μL of the reduction buffer (45 mM DTT, 100 mM ammonium bicarbonate) for 30 min at 37^oC, and then alkylated by adding 2.5 μL of the alkylation buffer (100 mM iodoacetamide, 100 mM ammonium bicarbonate) for 20 min at 24^oC in the dark. Prior to trypsin digestion, 20 μL of water was added to reduce the urea concentration to 2 M. 10 μL of the trypsin solution (5 $ng/\mu L$ of trypsin sequencing grade from Promega, 50 mM ammonium bicarbonate) was added to each sample. Protein digestion was performed at 37^oC for 18 h and stopped with 5 μL of 5% formic acid. Protein digests were dried down in vacuum centrifuge and stored at -20^oC until LC-MS/MS analysis.

3.4.4 LC-MS/MS

Prior to LC-MS/MS, protein digests were re-solubilized under agitation for 15 min in 10 μL of 0.2% formic acid. Desalting/cleanup of the digests was performed using C_18 ZipTip pipette tips (Millipore, Billerica, MA). Eluates were dried down in vacuum centrifuge and then re-solubilized under agitation for 15 min in 10 μL of 2% ACN / 1% formic acid. The LC column was a C18 reversed phase column packed with a high-pressure packing cell. A 75 μm i.d. Self-Pack PicoFrit fused silica capillary column (New Objective, Woburn, MA) of 15 cm long was packed with

the C18 Jupiter 5 μm 300 Å reverse-phase material (Phenomenex, Torrance, CA). This column was installed on the Easy-nLC II system (Proxeon Biosystems, Odense, Denmark) and coupled to the LTQ Orbitrap Velos (ThermoFisher Scientific, Bremen, Germany) equipped with a Proxeon nanoelectrospray ion source. The buffers used for chromatography were 0.2% formic acid (buffer A) and 100% acetonitrile / 0.2% formic acid (buffer B). During the first 12 min, 5 μL of sample was loaded on the column to a flow rate of 600 nL/min and, subsequently, the gradient went from $2-80\%$ buffer B in 110 min at a flow rate of 250 nL/min and then came back to 600 nL/min and 2% buffer B for 10 min. LC-MS/MS data acquisition was accomplished using a eleven scan event cycle comprised of a full scan MS for scan event 1 acquired in the Orbitrap. The mass resolution for MS was set to $60,000$ (at m/z 400) and used to trigger the ten additional MS/MS events acquired in parallel in the linear ion trap for the top ten most intense ions. Mass over charge ratio range was from 380 to 2000 for MS scanning with a target value of $1,000,000$ charges and from $\sim 1/3$ of parent m/z ratio to 2000 for MS/MS scanning with a target value of $10,000$ charges. The data dependent scan events used a maximum ion fill time of 100 ms and 1 microscan. Target ions already selected for MS/MS were dynamically excluded for 25 s. Nanospray and S-lens voltages were set to 0.9-1.8 kV and 50 V, respectively. Capillary temperature was set to 250^oC. MS/MS conditions were: normalized collision energy, 35 V; activation q, 0.25; activation time, 10 ms.

3.4.5 Protein identification

Protein database searching was performed with Mascot 2.2 (Matrix Science) against the human NCBI nr protein database. The mass tolerances for precursor and fragment ions were set to 15 ppm and 0.6 Da, respectively. Trypsin was used as the enzyme allowing for up to 2 mis-cleavages. Carbamidomethylation of cysteine residues was set as a fixed modification and oxidation of methionine was allowed as a variable modification.

3.4.6 Dilution experiments and tandem mass spectrometry

Due to the increasing sensitivity of mass spectrometers, and the fairly low complexity of typical AP-MS/MS or even less complex MCC-AP-MS/MS experiments, the amount of protein digests injected in the mass spectrometer is now a concern. If a too large amount is injected in the LC-MS/MS system, peptide saturation is likely to occur. This could cause very abundant proteins to mask less abundant ones. When saturation happens, the mass spectrometer sampling frequency is not sufficiently fast to analyze the entirety of the sample, therefore causing a great variability between replicated experiments. We therefore tested sample dilutions to ensure accurate modeling of the contaminants present with a reasonable number of controls. Dilution experiments aimed at verifying saturation of the LC-MS/MS system and were performed by analyzing 6 different dilutions of a protein solution, (1/2, 1/4, 1/8, 1/16, 1/32, 1/64). Upon visual inspection of the results, the highest volume yielding no saturation was chosen and used for all LC-MS/MS runs. 1/2 of the eluates obtained from the chromatin and cytoplasmic fractions were used for LC-MS/MS, while the entirety of the nucloplasmic eluate was kept.

3.4.7 Dataset

A total of 15 MCC-AP-MS/MS experiments have been performed with the following baits: POLR2A (x2), CDK9 (x2), RPAP4 (x2), PPARG2 (x2), KLF14 (x2), FTO, IRS1, RPAP2, RPAP3, and JAZF1, where (x2) signifies that the experiment has been done in two biological replicates. A set of 9 MCC-AP-MS/MS experiments of empty expression vector pMZI were performed as controls (Supplementary Table 3-6). Three of them were excluded from the main training set since they showed an unexpectedly high variance in both the set of observed proteins and their abundance.

3.5 Computational analysis

Many tools have been proposed to identify high quality PPIs in AP-MS/MS data [20, 190, 206, 48]. We elected to use a modified version of our Decontaminator software [127] to calculate the false discovery rate (FDR) for each individual bait-prey interaction. The existing version of Decontaminator required matched induced and non-induced expression vector AP-MS/MS experiments for its training procedure. We generalized the implementation of Decontaminator so that any number of empty pMZI vector controls can be used. The only requirement is that both controls and experiments are performed under the same conditions and that the number of high quality controls is sufficient (at least 3, but 6 were used in our case) to model the contaminants.

3.5.1 Control training set

To improve the modeling of contaminants and thus the specificity of our approach, a number of modifications were made to the Decontaminator algorithm. Previously, each interaction was assigned a p-value based on the Mascot score [174] obtained by the prey in the pull-down compared to those obtained for the same protein in control experiments performed under the same conditions (in our case, the same cellular fraction). Because the set of controls available is sometimes relatively small, a second set of controls is also considered, which consists of the union of the controls obtained for each of the three fractions, including the controls with very large variance mentioned previously. A second p-value is then computed based on this larger set of controls. The final p-value reported is the largest of the fraction-specific p-value and the pooled p-value. This approach has the advantage of accurately modeling contaminants that are not fraction-specific (based on a large set of pooled controls), while allowing those that are fraction-specific to also be identified. This approach maximizes the specificity of the predictions by making the best use of all available control

experiments. Although this may in theory be at the cost of a loss of sensitivity, this loss seems to be negligible in our dataset (data not shown).

3.5.2 FDR estimation

For the purpose of estimating FDRs for each cellular fraction, we used a set of selected baits that are known to be localized to (at least) that fraction, based on the Gene Ontology [4] (Cellular Component GO terms: cytoplasm, nucleus (used as a proxy for chromatin), and nucleoplasm). GO electronic annotations were disregarded, except in the case of the cytoplasm cellular component where too few curated annotations were available. This process ensures an automatic unbiased selection of the baits used to compute the FDRs in each cell fraction. The baits selected for the FDR estimation of the chromatin fraction were: CDK9 (x2), POLR2A (x2), PPARG2 (x2), FTO, and IRS1. Those chosen for the nucleoplasmic fraction were: POLR2A (x2), CDK9 (x2), and PPARG2 (x2). Finally, the experiments selected for the cytoplasmic fraction were: RPAP4 (x2), CDK9 (x2), PPARG2 (x2), and IRS1.

3.5.3 Implementation and availability

The proposed computational methods are implemented in a platform-independent Java program (Decontaminator). Given a set of control MCC-AP-MS/MS experiments, Decontaminator assigns FDRs to all interactions in all cell fractions according to the methods described previously. Of note, Mascot scores can in principle be interchanged by any other mass spectrometry derived confidence scores (i.e. spectral counts, SEQUEST Xcorr [58], peptide counts). Also, since the FDR calculation is performed independently for each p-value, it is possible that the function mapping p-values to FDRs is not monotonic. This was addressed by setting the FDR associated to a given p-value p to the minimum between its calculated FDR and the minimal FDR of all p-values larger than p. Decontaminator is available for download at: `http://www.cs.mcgill.ca/~blanchem/MCC_Decontaminator`.

3.6 Results

3.6.1 Cell compartment specific interactions

We started by assessing the ability of the MCC-AP-MS/MS method to reliably identify interactions that are specific to each cell compartment. In order to do so, we used three proteins with relatively well studied interactions as baits: (i) POLR2A, a subunit of RNAP II, expected to be present in all three cell compartments, but mainly in the chromatin fraction [69, 111, 145], (ii) CDK9, a mainly nucleoplasmic cyclin-dependent kinase that is a subunit of the positive transcription elongation factor P-TEFb, involved in transcription with potential additional localization to the cytoplasm and the chromatin [144, 71], and (iii) RPAP4, a mainly cytoplasmic protein involved in the nuclear import of RNAP II, through a mechanism that involves RPAP4 shuttling between the nucleus and the cytoplasm, implying its transit in the nucleoplasm [69]. Interaction partners were identified using the MCC-AP-MS/MS experimental/computational pipeline for each bait in each of the three fractions, in biological duplicates (Supplementary Table 3-7). For comparison, interactions were also detected using soluble WCE to identify interactions taking place in the soluble fraction. Figure 3–2 shows SDS gels of affinity purifications using a soluble WCE, a cytoplasmic fraction, a nucleoplasmic fraction, a chromatin fraction, and their associated controls for all three baits. Visual inspection revealed dramatic differences in the band distribution across the different fractions for any given bait. For instance, the chromatin fraction for RPAP4 is poorly populated when compared to its cytoplasmic and nucleoplasmic counterparts, as expected based on the literature [69]. As for CDK9 and POLR2A, it can be seen that each fraction, excluding the WCE, has at least one exclusive band, which is not observed in its control. These gels clearly demonstrate that different interaction partners can be identified in each fraction.

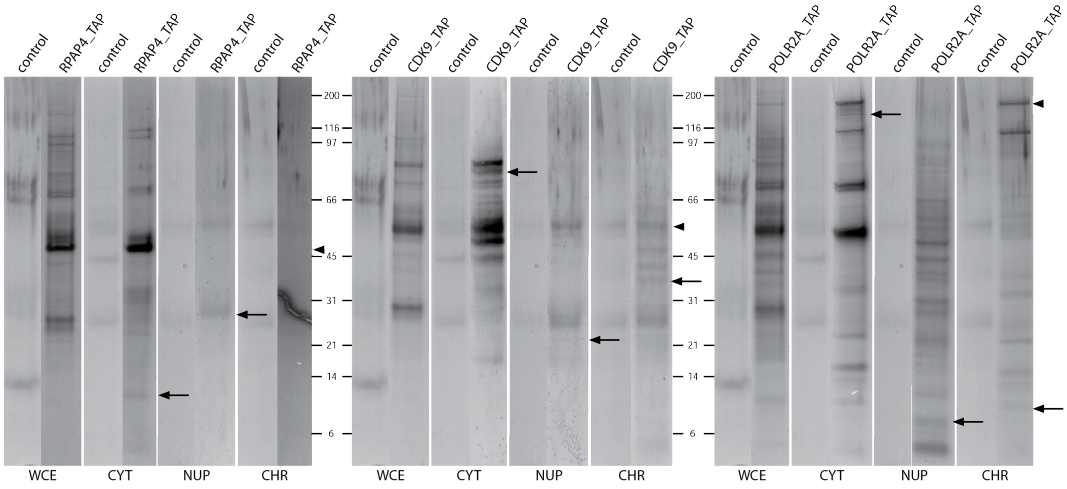

Figure 3–2: Precast acrylamide $4-12\%$ gels of eluates from MCC-AP-MS/MS and AP-MS/MS using RPAP4, CDK9 and POLR2A. WCE: Whole-Cell Extract, CYT: Cytoplasm, NUP: Nucleoplasm, CHR: Chromatin. Arrowheads point to bands corresponding to the baits. Full arrows indicate examples of exclusive bands in the various fractions for each bait.

To further evaluate these differences, we compared the set of high confidence PPIs found in each fraction for RPAP4, POLR2A and CDK9. Figure 3–3 shows the overlap between the sets of high confidence partners identified in each fraction (at least one replicate with FDR $< 10\%$). Due to an increased sensitivity in the cytoplasmic fraction (see below), the number of interactors identified in this particular fraction exceeds that of the other two fractions (see Discussion). However, as expected, of the three baits, POLR2A has a larger number of interactors in the chromatin fraction (37), as compared to non-chromatin associated CDK9 and RPAP4 (12 and 8 interactors respectively). To further confirm the fraction specificity of our approach, we analyzed interactions that are well documented and whose localization has been characterized. Most of the RNAP II subunits were found in the chromatin, but also in the cytoplasmic and nucleoplasmic fractions (Supplementary Table 3–3). This is in agreement with the transcriptional function of RNAP II, but also with the findings revealing POLR2A interactions with other RNAP II subunits in the cytoplasm [69, 163]. Also, for example, 18 of the 26 subunits of the mediator complex were

found to interact with POLR2A in the chromatin fraction, but almost none were seen in the other two fractions. This is consistent with the previously documented role of the mediator complex [214]. Finally, POLR2A interaction partners RPAP3, PIH1D1, UXT, and WDR92, all members of the RPAP3/R2TP/PFDL complex, were found to be exclusive to the cytoplasmic fraction and, by the same mean, to match previous results showing that RPAP3 is involved in the assembly/nuclear import of RNAP II [69, 37].

On the other hand, the mainly nucleoplasmic protein CDK9 has a large number of interactors in the nucleoplasmic fraction (50). This is significantly more than what is observed for the mainly chromatin-bound POLR2A (23) and slightly more than nucleoplasm-cytoplasm shuttling RPAP4 (35). As for POLR2A, our results are in agreement with the literature for the bait CDK9 (Supplementary Table 3–4), which interacts with BRD4, a positive regulator of the P-TEFb complex, with high confidence on the chromatin [235, 98], but not in the nucleoplasm and cytoplasm. CDK9 is also found to interact with very high confidence with HEXIM1 and HEXIM2, LARP7 and MEPCE in the nucleoplasmic and cytoplasmic fractions, but not in the chromatin fraction. These proteins are known to interact with P-TEFb and the 7SK snRNA to inhibit P-TEFb function, therefore supporting our observations [101, 120]. Even though some interactors of RPAP4 are not well characterized proteins, our findings that RPAP4 interacts with RNAP II subunits in the cytoplasm and nucleoplasm, as well as with RPAP3 in the cytoplasm (Supplementary Table 3-7), are in agreement with our previous results showing that RPAP4 plays a role in the nuclear import of RNAP II [69].

To confirm that the preys discovered by MCC-AP-MS/MS were indeed specific to the fraction in which they were identified, we considered the set of all high-confidence preys identified for at least one bait in a given cellular fraction and calculated the proportion of those proteins that are known to be localized to a certain GO cellular compartment (Figure 3–4). High-confidence interactors (FDR $< 10\%$) identified by

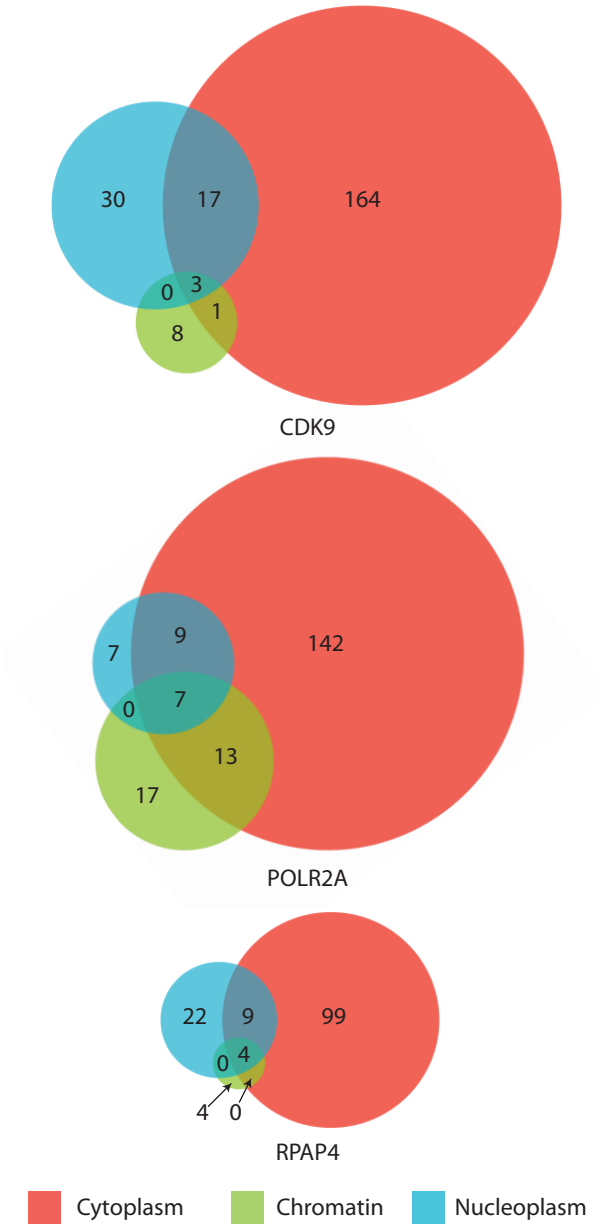

Figure 3–3: Number of preys obtained in each cell compartment (FDR $< 10\%$ in at least one of the duplicates) for each bait (CDK9, POLR2A, and RPAP4).

MCC-AP-MS/MS in the chromatin fractions are indeed generally annotated as being localized to the nucleus. Similarly, the interactors obtained in the nucleoplasmic fraction are annotated as nuclear, but to a lesser extent (Figure 3–4A). The results for the nucleoplasm cellular compartment were very similar to those of the nucleus (data not shown). Finally, those identified in the cytoplasmic fraction are often annotated as such (Fig. 3–4B). On the other hand, interactors found in the cytoplasmic fractions are much more rarely localized to the nucleus according to GO, but much more often in the cytoplasm than those identified in the two non-cytoplasmic fractions. Interestingly, the fraction of proteins obtained by MCC-AP-MS/MS of the nucleoplasmic fraction that are annotated to be cytoplasmic is also always higher than for the chromatin fraction at the same FDR threshold, suggesting that multiple proteins present in the nucleoplasm are shuttling to the cytoplasm. It is expected that an important portion of the nucleoplasmic proteins are related to the import or export of proteins to the nucleus. Shuttling of proteins is also a major way to regulate the activity of a protein in time and space and the outcome of signaling pathway activation. Also, as the FDR thresholds are allowed to increase, the level of false positives becomes higher, resulting in the drop of enrichments to background level.

3.6.2 MCC-AP-MS/MS is reproducible

An important aspect of a PPI detection approach is its reproducibility. We analyzed biological replicates of MCC-AP-MS/MS for all three baits (POLR2A, CDK9, and RPAP4) (Table 3–1). We define an interaction as being strictly reproduced if it obtained a FDR below 10% in both replicates and partially reproduced if it obtained a FDR below 10% in one replicate and below 20% in the other. A prey that obtains FDRs above 10% in both replicates is considered a likely contaminant. The data show very high levels of reproducibility for POLR2A, especially in the chromatin fraction (92%). However, when a bait is not localized to the fraction under consideration (e.g. RPAP4 in the chromatin fraction (43%)), we generally detected a smaller number of interactions, that tend to be less reproducible. We also observed that reproducibility

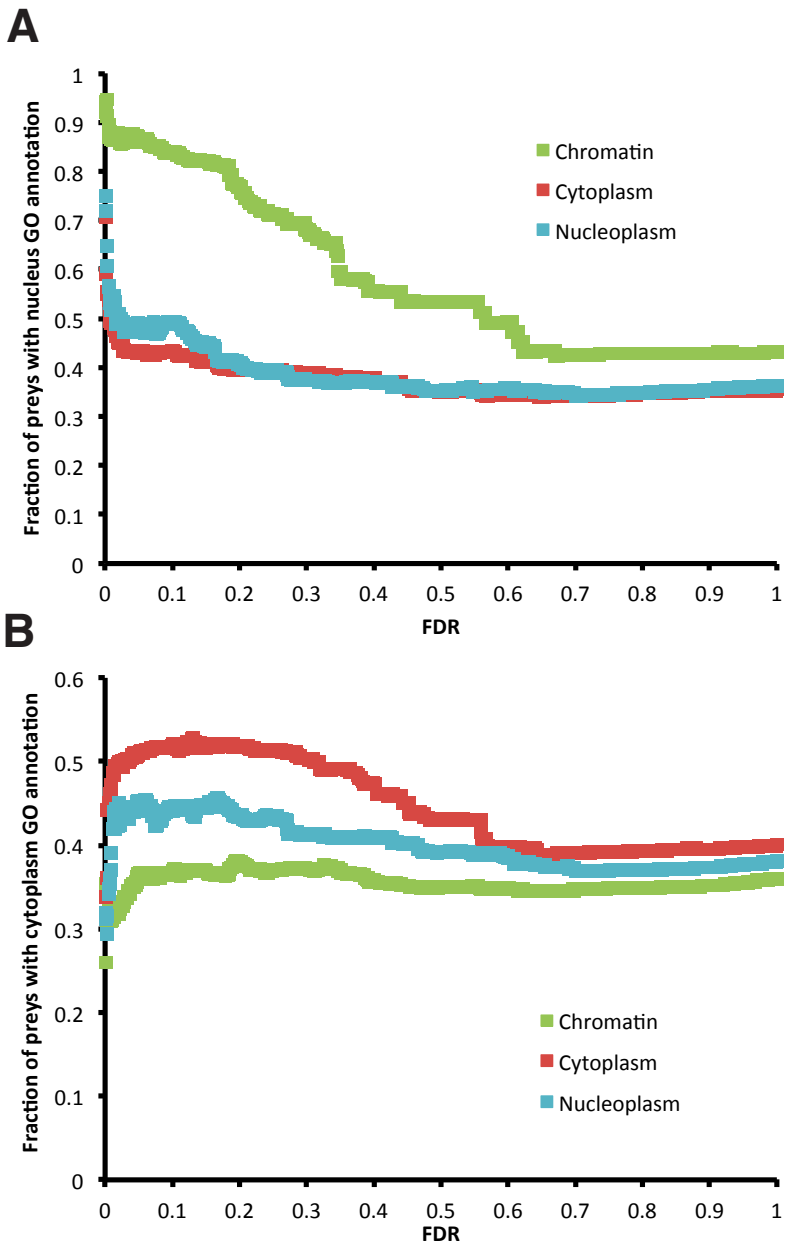

Figure 3–4: Proportion of the set of preys identified for the 15 baits in each cellular fraction (chromatin, cytoplasm, and nucleus) that are annotated with a GO cellular component term "Nucleus" (panel A) and "Cytoplasm" (panel B). Because "Chromatin" is a poorly populated cellular localization annotation in GO, "Nucleus" was used as a surrogate. The unexpectedly large proportion of preys from the cytoplasmic fractions with a "Nucleus" GO cellular component is caused by the important number of POLR2A interactions that take place in both the nucleus and the cytoplasm. Similarly, the sudden decrease of the proportion for all compartments at low FDRs for the GO term "Cytoplasm" can be explained again by the large number of POLR2A interactions that take place in both the nucleus and the cytoplasm, but for which the preys are not annotated to be localized in the cytoplasm in GO.

is generally lower for the nucleoplasmic fraction. Fewer proteins were identified in general in that fraction for all three baits when compared to the other fractions (Supplementary Table 3–5). In addition, the abundances of the nucleoplasmic proteins were also lower than those of the proteins identified in the other cell fractions, as indicated by the peptide counts (Supplementary Table 3–5). This hints towards the fact that the nucleoplasmic fraction may contain less material, which may therefore affect mass spectrometry detectability of the proteins contained in it, ultimately leading to a reduced reproducibility.

Table 3–1: Reproducibility results between duplicate MCC-AP-MS/MS experiments of POLR2A, CDK9, and RPAP4.†

	POLR2A			**CDK9**			**RPAP4**		
Fraction	Chromatin	Cytoplasm	Nucleoplasm	Chromatin	Cytoplasm	Nucleoplasm	Chromatin	Cytoplasm	Nucleoplasm
Strictly Reproduced Interactions	21	69	6	5	42	9	1	43	7
Partially Reproduced Interactions	13	28	3	1	16	2	2	11	2
Non Reproduced Interactions	3	74	14	6	127	39	4	58	26
Reproduced Contaminants Detected	2057	2274	1901	3142	2722	2165	1802	2271	1684
Fraction of Very High Confidence Interactions Strictly Reproduced	0.57	0.40	0.26	0.42	0.23	0.18	0.14	0.38	0.20
Fraction of Very High Confidence Interactions Partially Reproduced	0.92	0.57	0.39	0.50	0.31	0.22	0.43	0.48	0.26

† Results are shown for the resolubilized chromatin, cytoplasmic, and nucleoplasmic fractions.

3.6.3 MCC-AP-MS/MS has greater interactome coverage than whole cell extract AP-MS/MS

Having established the specificity and reproducibility of our method, we compared the sensitivity of the classic AP-MS/MS (based on WCE) and MCC-AP-MS/MS approaches under the same experimental conditions (see Experimental procedures). Figure 3–5 shows that the vast majority of the high confidence interactions obtained through AP-MS/MS are also recovered by MCC-AP-MS/MS. To ensure that very high confidence interactions were being compared, only proteins found in each replicate of both protocols were used to produce this figure. Indeed, if one were to perform AP-MS/MS after MCC-AP-MS/MS, it would only yield an increase in the

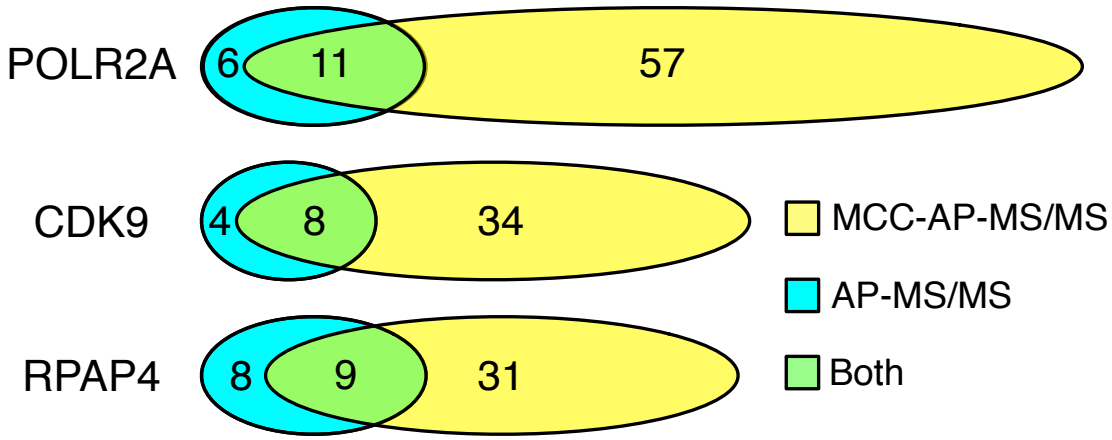

Figure 3–5: Number of interacting partners found in both replicated experiments of MCC-AP-MS/MS (FDR $< 10\%$) and AP-MS/MS (FDR $< 20\%$) for each bait (POLR2A, CDK9, and RPAP4). We allowed a higher FDR threshold for AP-MS/MS derived interactions to match previous studies [37]. This only advantages AP-MS/MS over MCC-AP-MS/MS in this comparison.

number of interactors by 9% for POLR2A, 20% for CDK9, and 10% for RPAP4. Conversely, MCC-AP-MS/MS yields a $180 - 335\%$ increase in the number of interactions detected when compared to AP-MS/MS alone. This gain in sensitivity is in part due to the separation of the sample into three fractions, which improves the sensitivity of MS/MS protein identification for each of the three fractions. In AP-MS/MS using WCE, low abundance PPIs localized to the nucleoplasm are likely to be masked by higher abundance cytoplasmic interactions. By separating samples in three different fractions, sample complexity is reduced, which improves the performance of the mass spectrometer, much the same way as sample fractionation through gel or liquid chromatography improves classical AP-MS/MS sensitivity. Remarkably, the set of 28 interactions detected by both AP-MS/MS and MCC-AP-MS/MS were all found in (at least) the cytoplasmic fraction by MCC-AP-MS/MS, suggesting that AP-MS/MS based on WCE is largely confined to identifying cytoplasmic interactions.

To further test the sensitivity of MCC-AP-MS/MS, we calculated the recall values for each of the three baits against human protein-protein interactions deposited

in the BioGRID database (Release 3.1.93) [209] that were obtained through affinity capture methods coupled to mass spectrometry (Figure 3–6). For each bait, MCC-AP-MS/MS obtains significantly higher recall values than AP-MS/MS at any FDR threshold, therefore showing that MCC-AP-MS/MS not only detects more PPIs than AP-MS/MS, but also that these PPIs were found independently by other laboratories. We also measured the recall of both methods using the top X preys for each bait ranked by their FDRs, for a varying value of X (Figure 3–6). MCC-AP-MS/MS shows an improvement in recall over AP-MS/MS for any value of X for both POLR2A and RPAP4, and comparable recall for CDK9. The latter observation may suggest that the AP-MS/MS FDRs for this bait may have been overestimated. Interestingly, recall values of MCC-AP-MS/MS for all three proteins keep increasing with the FDR threshold, reaching almost 1.0 at a very high FDR threshold. This shows that MCC-AP-MS/MS has the potential for excellent sensitivity, and that the current limitations are at the level of the experimental and computational filtering of contaminants, and mass spectrometry detectability.

3.6.4 MCC-AP-MS/MS improved sensitivity leads to discovery of new protein-protein interactions

The increased sensitivity of MCC-AP-MS/MS and its ability to detect fraction-specific interactions allow it to discover new potentially biologically important interactions and hint at the mechanisms/processes they may be involved in. Here, we discuss one such example. Among the interactors of POLR2A in the chromatin fraction are 5 proteins with RNAP II carboxyl-terminal domain (CTD) interacting domains (CID): RPRD1A, RPRD1B, RPRD2, PCF11, and SCAF4 (Table 3–2). Ni and colleagues recently reported the discovery of the interaction of the first three with RNAP II through AP-MS/MS in HEK293 cells [162], but interactions with PCF11 and SCAF4 were not detected. PCF11 was computationally predicted to interact with POLR2A [151], but this interaction like the one involving SCAF4, had not been detected *in vivo*. Strikingly, the 3 proteins that were identified in both the Ni et al. study and our

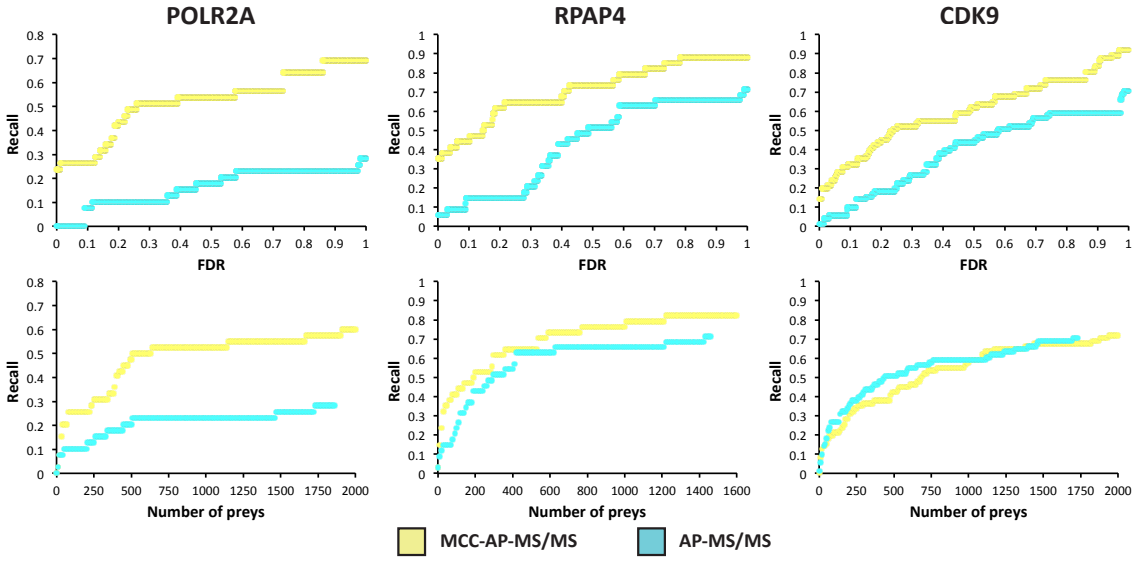

Figure 3–6: Recall values of MCC-AP-MS/MS and AP-MS/MS for varying FDR thresholds and number of top preys (ranked by their FDRs) against human PPIs listed in BioGRID that were obtained through affinity capture coupled to mass spectrometry. Recall values were calculated by taking the union of the preys identified in each bait replicated experiment. When a prey was detected in both replicates its smallest FDR was used.

own study were observed in the cytoplasmic fraction of MCC-AP-MS/MS. However, all 5 proteins were found with high or very high confidence in the chromatin fraction, with PCF11 and SCAF4 only being detected in this fraction. This may explain why Ni et al. could not identify PCF11 and SCAF4 as interactors of RNAP II. A classic AP-MS/MS could potentially simply not reach this space of the interactome for POLR2A.

3.7 Discussion and conclusion

Over the years, the AP-MS/MS methodology has proven to be successful at discovering interaction partners for a large number of proteins [119, 69, 226, 101, 20, 74]. To date, large-scale AP-MS/MS-based PPI mapping efforts used the soluble fraction of the cell, limiting the discovery and interpretation of compartment-specific interactions. We introduce here the multiple cell compartment AP-MS/MS (MCC-AP-MS/MS) experimental/computation pipeline to detect interactions occurring in

Table 3–2: CID proteins found in union of the MCC-AP-MS/MS experiments of POLR2A.†

Prey	Chromatin	Cytoplasm	Reported by Ni et al.
RPRD1A	0.07	0.00	Yes
RPRD1B	0.00	0.00	Yes
RPRD2	0.00	0.37	Yes
PCF11	0.00	N/O	No
SCAF4	0.12	N/O	No

FDR ≤ 0.1

$0.1 <$ FDR ≤ 0.2

FDR > 0.2

† Minimum FDR scores of the preys obtained in the two chromatin and cytoplasmic fraction experiments are color-coded.

the cytoplasmic, the nucleoplasmic, and the chromatin fraction, while using the same starting material. To minimize the number of contaminants and indirect interactions that may occur through DNA binding for the chromatin fraction, we performed a complete DNA digestion combined with tandem affinity purification. We have shown that MCC-AP-MS/MS generates a significant gain in sensitivity over classical AP-MS/MS, identifies compartment-specific interactions, and is reproducible. As an illustration, we demonstrated that MCC-AP-MS/MS reveals novel interactions for POLR2A, despite the fact that interactions for this protein have been intensively analyzed in the past [101, 162, 121, 1, 170].

The type of compartment-specific assays performed by MCC-AP-MS/MS would be impossible to perform in the context of a Y2H system due to the nature of the protocol. PCA techniques could potentially be modified to localize interactions, but this would not be scalable to a large-scale study. Therefore, to our knowledge, MCC-AP-MS/MS is the first PPI mapping technique that can both accurately map interactions and specify their localization in the cell. Given the fact that cell fractionation always leads to some cross-contamination of the fractions, which is estimated to be minimal according to our western blotting analysis (see Supplementary Figure 3–7), MCC-AP-MS/MS cannot be used to decisively conclude on the presence or absence of an interaction (or interactor) in any given compartment. However, as described above,

MCC-AP-MS/MS presents major advantages as compared to classical AP-MS/MS. When larger scale MCC-AP-MS/MS PPI mapping data become available, the development of computational approaches predicting protein complex components in different cell fractions will be possible.

Another aspect of our results deserves further discussion. One might be surprised by the fact that a bait such as POLR2A detects more interactors in the cytoplasmic than the chromatin fraction. However, a number of recent studies revealed the complexity of the cellular machinery required for the assembly and nuclear import of RNAP II [69, 37, 45, 27]. This machinery comprises several proteins that interact with the RNAP II subunits in the cytoplasm. It is therefore not surprising that an important number of high confidence cytoplasmic partners are identified. Moreover, most newly synthesized proteins are in one way or another present in the cytoplasm where they critically interact with proteins such as chaperones, transporters, inhibitors or activators.

Knowing in which compartment an interaction is occurring deconvolutes the complex PPI networks produced by AP-MS/MS and provides useful information on the context in which it is taking place. We believe that methods such as MCC-AP-MS/MS will significantly change the sensitivity and interpretability of future protein-protein interactions network mapping efforts.

3.8 Acknowledgements

We are grateful to the members of our laboratories for helpful discussions and comments. This work is supported by grants from the Canadian Institutes for Health Research (CIHR) and Fonds de la recherche en santé du Québec (FRSQ). M.L.A. holds a Vanier studentship from NSERC.

3.9 Appendix

Table 3–3: FDR obtained for a selected set of interaction partners of POLR2A in the cytoplasmic, nucleoplasmic, and resolubilized chromatin fractions.†

Prey	Chromatin	Nucleoplasm	Cytoplasm
GTF2B	N/O	N/O	0.00
GTF2F1	0.06	N/O	0.00
GTF2F2	0.12	N/O	0.00
MED1	0.00	N/O	0.73
MED4	0.12	N/O	0.67
MED6	0.12	N/O	N/O
MED8	0.12	N/O	0.73
MED10	0.14	N/O	N/O
MED11	0.44	N/O	N/O
MED12	0.22	N/O	N/O
MED14	0.00	0.18	0.04
MED15	0.03	0.59	0.67
MED16	0.00	N/O	N/O
MED17	0.00	N/O	0.06
MED18	0.23	N/O	N/O
MED19	0.22	N/O	N/O
MED20	0.12	N/O	N/O
MED23	0.00	N/O	0.67
MED24	0.00	0.33	0.10
MED25	0.20	0.86	N/O
MED26	0.00	N/O	N/O
MED27	0.06	0.59	N/O
MED28	0.18	N/O	N/O
MED29	0.19	N/O	N/O
MED30	0.19	N/O	N/O
MED31	0.44	N/O	N/O
POLR2A	0.00	0.00	0.00
POLR2B	0.00	0.00	0.00
POLR2C	0.00	0.00	0.00
POLR2D	0.44	N/O	0.02
POLR2E	0.03	0.00	0.00
POLR2F	0.90	N/O	0.08
POLR2G	0.00	0.00	0.00
POLR2H	0.05	0.00	0.00
POLR2I	0.08	0.18	0.01
POLR2J	0.23	N/O	0.16
POLR2K	0.46	N/O	0.67
TAF3	0.92	N/O	0.86
TAF4	0.19	N/O	N/O
TAF5	0.39	N/O	N/O
TAF6	0.25	N/O	N/O
PIH1D1	N/O	N/O	0.00
RPAP3	N/O	0.86	0.00
UXT	N/O	N/O	0.16
WDR92	N/O	0.94	0.01

FDR $\leq$ 0.1

0.1 $<$ FDR $\leq$ 0.2

FDR $>$ 0.2

† Minimum FDR scores of the preys obtained in the two replicates are color-coded.

Table 3–4: FDR obtained for a selected set of interaction partners of CDK9 in the cytoplasmic, nucleoplasmic, and resolubilized chromatin fractions.†

Prey	Chromatin	Nucleoplasm	Cytoplasm
BRD4	0.12	N/O	N/O
CCNT1	0.03	0.00	0.00
CCNT2	0.12	0.00	0.00
CDK9	0.03	0.00	0.00
HEXIM1	0.32	0.00	0.00
HEXIM2	N/O	0.00	0.00
LARP7	0.12	0.00	0.00
MEPCE	0.18	0.00	0.00

FDR ≤ 0.1

$0.1 < \text{FDR} \leq 0.2$

FDR > 0.2

† Minimum FDR scores of the preys obtained in the two replicates are color-coded.

Table 3–5: Averages of the numbers of proteins and peptides detected in the two replicate MCC-AP-MS/MS experiments performed for each bait.

	POLR2A			CDK9			RPAP4		
Fraction	Chromatin	Cytoplasm	Nucleoplasm	Chromatin	Cytoplasm	Nucleoplasm	Chromatin	Cytoplasm	Nucleoplasm
Proteins	1345.5	1611.5	1243	1942.5	1806	1392.5	1151	1547	1101.5
Peptides	3546	3949	2782.5	4257	3934.5	3182.5	3020	4199.5	2734.5

Supplementary Tables 3-6 and 3-7 are available via the Internet at:

http://pubs.acs.org/doi/abs/10.1021/pr300778b

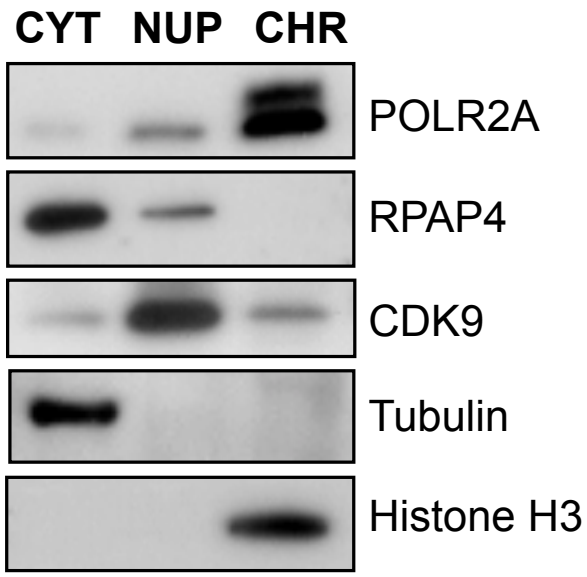

Figure 3–7: Western blotting analysis of the three baits (POLR2A, CDK9, and RPAP4) and two additional cytoplasmic (Tubulin) and chromatin-bound (Histone H3) markers in the various fractions used in our MS analysis. CYT=cytoplasmic, NUP=nucleoplasmic, and CHR=chromatin).

CHAPTER 4
Detection of locally over-represented Gene Ontology terms in protein-protein interaction networks

4.1 Preface

The approaches presented in Chapter 2 and 3 take as input very complex and noisy AP-MS data and output a high confidence PPI dataset. Such collections of PPIs can then be assembled to form PPI networks. As described in the introduction, such networks can provide much information on various biological mechanisms and protein functions. Even though these networks contain only high confidence PPIs, they still remain very large and complex.

For several groups building and studying PPI networks such as ours, an important aspect of their work consists in characterizing certain proteins or subnetworks in these networks. There is however often only very little hope that meaningful biological conclusions will be drawn from manual inspection of large and intricate PPI networks. Manual analysis of PPI networks is challenging despite good visualization tools such as Cytoscape [203], VisANT [91], or NAViGaTOR [24]. Computational methods have been developed to either identify protein complexes or infer protein functions (see Chapter 1). This chapter introduces a novel computational approach (GoNet) that detects sets of proteins annotated with the same GO term (e.g. biological process, molecular function), which are surprisingly clustered in a given PPI network. We are the first to formalize the problem addressed by GoNet. We also show in this chapter that the problem tackled by GoNet is reducible to the k-clique problem. Even though cliques of size k can be found in polynomial time whenever k is a fixed constant, k is often very large and therefore an approximation method was required to address the problem presented here.

GO terms are often used to analyze components of PPI networks [101, 7] and even to apply confidence scores on experimentally obtained or predicted PPIs [97]. The GO project will be described in further detail later in this chapter. GoNet therefore outputs the sets of proteins that are significantly clustered and that share a same GO term. With the overwhelming quantity of data in PPI networks, such sets are useful when comes the need to target certain regions of interest for smaller scale protein characterization experiments. GoNet can also help prioritizing regions of the network that should be studied first for laboratories interested in certain families of biological processes. In addition, an uncharacterized protein that co-clusters with a protein set detected by GoNet is likely to be linked with the GO term of this protein set. Such protein can represent a very interesting target for functional characterization studies.

A preliminary version of the work presented in this chapter was published in the proceedings of RECOMB 2009.

- M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of locally over-represented GO terms in protein-protein interaction networks. In *Research in Computational Molecular Biology*, pages 302–320. Springer, 2009

As it was discussed in the introduction, edges in PPI networks are often weighted based on the confidence that an interaction is a true positive (scores from SAINT [34] or Decontaminator [127]), the estimated abundance (e.g. spectral counts), or the identification confidence score (e.g. Mascot [174] or Protein Prophet [161]) of the prey of the interaction. The distance and similarity measures used to evaluate protein clusterings that were presented at RECOMB were defined on unweighted graphs. We therefore adapted the approach to allow, if the information is available, the calculation of weighted distance and similarity measures. The use of weighted measures penalizes GO terms for which the proteins are connected with low confidence scores and boosts those where the proteins are connected with large weights.

The RECOMB publication also argued that even though the proteins associated to a given GO term may be significantly clustered, it does not mean that all these proteins belong to a dense cluster. The article presented a greedy strategy to identify the core of a GO term that was significantly clustered. We replaced this algorithm by a faster and more adequate hierarchical clustering approach, which generally succeeds at identifying the various cores of a clustered set of proteins.

The remaining content of this chapter is taken from:

- M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of locally over-represented GO terms in protein-protein interaction networks. *Journal of Computational Biology*, 17(3):443–457, 2010

4.2 Abstract

High-throughput methods for identifying protein-protein interactions produce increasingly complex and intricate interaction networks. These networks are extremely rich in information, but extracting biologically meaningful hypotheses from them and representing them in a human-readable manner is challenging. We propose a method to identify Gene Ontology terms that are locally over-represented in a subnetwork of a given biological network. Specifically, we propose several methods to evaluate the degree of clustering of proteins associated to a particular GO term in both weighted and unweighted PPI networks, and describe efficient methods to estimate the statistical significance of the observed clustering. We show, using Monte Carlo simulations, that our best approximation methods accurately estimate the true p-value, for random scale-free graphs as well as for actual yeast and human networks. When applied to these two biological networks, our approach recovers many known complexes and pathways, but also suggests potential functions for many subnetworks.

4.3 Introduction

Gene ontologies provide a controlled, hierarchical vocabulary to describe various aspects of gene and protein function. The Gene Ontology (GO) Annotation project [4] is a literature-based annotation of a gene's molecular function, cellular component, and biological processes. GO analyses have become a staple of a number of high-throughput biological studies that produce lists of genes behaving interestingly with respect to a particular experiment. For example, a microarray experiment may result in the identification of a set of genes that are differentially expressed between normal and disease conditions. A GO term (or category) τ is said to be over-represented in a given list if the number of genes labeled with τ contained in the list is unexpectedly large, given the size of the list and the overall abundance of genes labeled with τ in the species under consideration (see tools like GoMiner [232], Fatigo [3], or GoStat [15]). Statistical over-representation is an indication that the GO category is directly or indirectly linked to the phenomenon under study. We say that this kind of set of differentially expressed genes is *unstructured*, in the sense that all genes within the list contribute equally to the analysis. A slightly more structured approach consists of considering an *ordered* list of genes, where genes are ranked by their "interest" with respect to a particular experiment (e.g. degree of differential expression). There, we seek GO terms that are surprisingly enriched near the top of the ranked list. This is the approach taken by the highly popular GSEA method [212], which generalizes this to include many kinds of gene annotations other than GO.

We propose taking this type of analysis one step further and applying GO term enrichment analysis to even more highly structured gene sets: biological networks. In such networks, genes (or their proteins) are vertices and edges represent particular relationships (protein-protein interactions, regulatory interactions, genetic interactions, etc.). Given a fixed biological network G and a gene ontology annotation database, our goal is to identify every term τ such that the genes labeled with τ are unexpectedly clustered in the network (i.e. they mostly lie within the same "region" of the

network). This local over-representation indicates that τ is likely to be linked to the function of that sub-network[1] . Indeed, and unsurprisingly, GO term clustering has been observed to occur in most types biological networks [47, 139], and has been used as a criterion to evaluate the accuracy of computational complex or module prediction [154]. However, to our knowledge, the problem of identifying locally over-represented GO terms in a network has never been formulated or addressed before.

This problem has a number of applications. High-throughput technologies generate large networks (thousands of proteins and interactions) that are impossible to analyze manually. Graph layout approaches (reviewed in [213]), integrated in many network visualization packages such as VisANT [92] and Cytoscape [195], can help humans extract biological meaning from the data, but revealing all aspects of a complex data set in a single layout is impossible and, often, key components of the network remain unstudied because the layout used did not reveal them visually. Various approaches have been proposed to ease the analysis of biological networks, including packages performing graph clustering and path analysis (e.g. NeAT [22, 195]). Several methods have been proposed to identify pathways [199] within PPIs or combine expression data with PPI networks to infer signaling pathways [193]. Expression data was also used to identify functional modules in PPI networks with a solution based on an integer-linear programming formulation [52]. Another popular strategy starts by identifying dense subnetworks within the network (using, for example, MCL [59]), and then evaluates various biological properties of the subnetwork, including GO term enrichment [194].

Our proposed approach identifies subsets of genes that share the same GO annotation and are highly interconnected in the network, thus formulating the hypothesis

[1] We note that in cases where the GO annotations themselves may be based on the PPI network, our analysis would form circular argument. However, GO annotations are based on a wide range of evidence and are rarely based on PPIs alone.

that the function of the subnetwork is related to that GO annotation. This reduces the complexity of the data and allows easier grasp by human investigators. Our approach could be extended to help function prediction: genes with incomplete functional annotation that are found to be highly interconnected with a set of genes of known function can be expected to share that function [36, 196].

In this paper, we define formally the problem of identification of locally-enriched GO categories for unweighted and weighted undirected interaction networks. We start by defining two measures of clustering of a set of genes within a given weighted or unweighted network. We then discuss the critical question of assessing the statistical significance of the local clustering scores using analytical approaches of a given GO term within the network, under a null hypothesis where vertices are selected randomly (empirical approaches for shortest path distance significance have been proposed previously [188]). We show that the exact computation of this probability is NP-hard, but we provide several efficient approximation methods. These p-value approximation methods are shown to be accurate on random scale-free graphs, as well as on large-scale yeast [119] and human [41, 101] protein-protein interaction networks. We then refine each significant gene sets to core subsets that contribute the most to its statistical significance. Our analysis identifies regions of these two networks with known function. It also suggests interesting functions for regions of the network that are currently poorly understood.

4.4 Methods

We are looking for GO terms whose distribution across a given network is non-random. In particular, we are interested in finding terms that are tightly clustered within the network. Let $G = (V, E)$ be an undirected, unweighted graph, where V is a set of n proteins and E is a set of pairwise interactions between them. The Gene Ontology project assigns to each gene a set of functional annotations, using a controlled vocabulary. For a given GO term τ, let $V(\tau) \subseteq V$ be the subset of the

proteins annotated with that term. Our goal is to investigate, for every possible term τ, whether $V(\tau)$ is particularly clustered in G, which would hint to the fact that τ is particularly relevant to the function of that subgraph. To this end, we introduce in Section 4.4.1 two measures of clustering, as well as their generalizations to weighted graph, and show in Section 4.4.2 how to measure their statistical significance.

4.4.1 Measures of clustering in a network

A number of approaches have been proposed to measure the clustering of a set of vertices within a given graph, and to identify dense clusters (e.g. MCL [59]; see [23] for a review). We focus on two simple but effective clustering measures, for which the statistical significance can be accurately approximated analytically.

4.4.1.1 Total pairwise distance

Given two vertices u and v in V, let $d_G(u, v)$ be the length of a shortest path from u to v in G. Since G is undirected, d_G is symmetric. The distance matrix d_G can be computed in time $O(|V|^3)$ using the Floyd-Warshall algorithm [67, 225]. Let W be a subset of V. Then, the *total pairwise distance* (TPD) for W is defined as

$$TPD(W) \quad = \quad \sum_{u,v \in W, u<v} d_G(u, v). \tag{4.1}$$

If most of the vertices in $V(\tau)$ are in the same region of the graph (e.g. the gray or black vertices in Figure 4–1), then $TPD(V(\tau))$ will be smaller than that of most random subsets of $|V(\tau)|$ vertices and τ will be reported as potentially interesting.

4.4.1.2 Random-walk based similarity

One issue with the TPD clustering measure is that it does not take into consideration the degree of the nodes on the path between the two proteins, in such a way that, for example, the two sets of proteins shown in black and gray in Figure 4–1 will get the same total pairwise distance (and, eventually, the same p-value), although

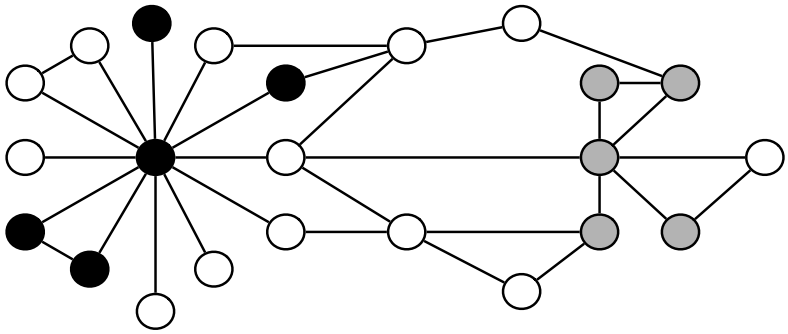

Figure 4–1: Example of a toy PPI network. The black and gray subsets of vertices obtain the same Total Pairwise Distance (13), but the gray subset obtains a higher Probability of Stopping within the Family (PSF).

intuitively the gray cluster appears more interesting. In addition, if the vertices in W form more than one dense subgraph, and these clusters are far away from each other, the TPD measure may not reveal anything unusual. We introduce an alternative to the total pairwise distance, which we call the Probability of Staying within the Family (PSF) clustering measure. This random-walk based similarity measure shares a relationship with diffusion kernels [115]. The PSF for a subset of vertices W is defined based on the following random process (similar to that modeled by MCL [59]), parameterized by a user-defined probability p: (i) Randomly select a vertex from W as a starting point; (ii) when at vertex u, stop with probability p, or, with probability $1-p$, continue to a vertex v uniformly chosen from the neighbors of u. Then, $PSF_p(W)$ is defined as the probability that the vertex where the process stops is an element of W. We note first that this process does make a difference between the two subsets in Figure 4–1 and will also assign a high score to a subset W that would consist of several dense but widely separated clusters.

If A_G is the adjacency matrix of G and $deg_G(u)$ is the degree of vertex u, then the transition probability matrix T_G for this random walk is defined as

$$T_G(u,v) \;=\; A_G(u,v) / \sum_{w \in V} A_G(u,w), \tag{4.2}$$

and the probability $P_{u,v}$ of stopping at vertex v, starting from vertex u, is given by

$$P_{u,v} = \sum_{i=0}^{+\infty} p(1-p)^i((T_G)^i)(u,v).$$

Thus,

$$PSF_p(W) = \sum_{u,v \in W} P_{u,v}/|W| = \sum_{u,v \in W} s_G(u,v),$$

and we obtain that, as for the total pairwise distance, the PSF measure is a sum of pairwise scores, with $s_G(u,v) = P_{u,v}/|W|$.

4.4.1.3 Generalization to weighted graphs

Edge weights are often used in protein-protein interaction networks to reflect the confidence that a given interaction is a true positive. Such scores can be provided by mass spectrometry analysis programs (e.g. Mascot [174] or PeptideProphet [35]). Both the TPD and PSF clustering measures can be adapted in the context of a weighted graph. The weighted TPD (WTPD) measure is obviously generalized using the weighted shortest path distances for d_G in Equation 4.1. In this case, edges are weighted as

$$w(e) = \begin{cases} 1 + \max(0, \log_{10}(maxMascot) - \log_{10}(mascot(e))) & \text{if } e \in E \\ +\infty & \text{otherwise} \end{cases}$$

where $mascot(e)$ is the Mascot score associated with edge e and $maxMascot = 500$. The obtained weighted distance matrix will be referred as d_{G_W}. This measure penalizes paths with low confidence scores, therefore when the vertices in $V(\tau)$ are located in the same region of the graph and edges connecting those vertices have high confidence, $WTPD(V(\tau))$ will be small.

A generalization of PSF to WPSF is obtained by replacing the adjacency matrix A_G by the weighted adjacency matrix A_{G_W} in Equation 4.2, where

$$A_{G_W}(e) = \begin{cases} \log_{10}(mascot(e)) & \text{if } e \in E \\ 0 & \text{otherwise} \end{cases}$$

The resulting weighted similarity matrix will be referred as s_{G_W}.

The methods proposed in Section 4.4.2 to assess statistical significance apply to both TPD and PSF and their respective weighted versions.

4.4.2 Measuring the statistical significance

Given a matrix $M_{|V|\times|V|}$ containing pairwise distances (d_G or d_{G_W}), or similarities (s_G or s_{G_W}), we consider the random variable obtained as follows. Let $R = \{r_1, r_2, ..., r_k\} \subseteq \{1, ..., n\}$ be a randomly selected subset of proteins of cardinality k. We are interested in the distribution of the random variable $S_k = \sum_{i,j\in R, i<j} M_{i,j}$. When using the weighted or unweighted TPD clustering measure, the p-value for GO term τ will be obtained as $p\text{-value}_{TPD}(\tau) = \Pr[S_{|V(\tau)|} \leq TPD(V(\tau))]$, whereas when using the PSF clustering measure, the p-value will be obtained as $p\text{-value}_{PSF}(\tau) = \Pr[S_{|V(\tau)|} \geq PSF_p(V(\tau))]$. Note that there is no need to adjust the p-values for k since we are analyzing a different distribution for each S_k.

A note on complexity. We first observe that computing the exact distribution of S_k when $M = d_G$ is NP-hard. Indeed, $\Pr[S_k = \binom{k}{2}]$ is non-zero if and only if G contains a k-clique. Therefore, we cannot expect an exact polynomial time algorithm. Although more difficult to prove, the same is likely true for PSF. We thus investigate three approaches that give approximations to the desired probability distributions.

4.4.3 Normal approximation

Being a sum of $\binom{k}{2}$ random variables, the distribution of S_k should converge to a normal distribution as k and $|V|$ become large (Central Limit Theorem), if these random variables were independent. Although these variables are clearly not independent (for example, in the case of TPD, they must satisfy the triangle inequality), it turns out that the normality assumption sometimes yields a useful approximation to the true distribution. The expectation of S_k can be calculated exactly in time $O(|V|^2)$. Let $E[S_2] = \frac{\sum_{1 \le a < b \le n} M_{a,b}}{\binom{n}{2}}$ be the average pairwise score in M. Then

$$E[S_k] = \binom{k}{2} \cdot E[S_2]$$

The variance of S_k is more challenging to obtain. We have $Var[S_k] = E[S_k^2] - E[S_k]^2$, where

$$\begin{aligned}
E[S_k^2] &= E[(\sum_{a,b \in R, a<b} M_{a,b})^2] \\
&= E[\left\{ \sum_{a,b \in R, a<b} (M_{a,b})^2 \right\} + \left\{ 2 \sum_{a,b,c \in R, a<b<c} M_{a,b}M_{a,c} + M_{a,b}M_{b,c} + M_{a,c}M_{b,c} \right\} + \\
&\quad 2 \left\{ \sum_{a,b \in R, a<b} \sum_{c<d \in R, c \neq a,b, d \neq a,b} M_{a,b}M_{c,d} \right\}] \\
&= \frac{\binom{k}{2}}{\binom{n}{2}} \sum_{1 \le a < b \le n} (M_{a,b})^2 + 2\frac{\binom{k}{3}}{\binom{n}{3}} \left\{ \sum_{1 \le a < b < c \le n} M_{a,b}M_{a,c} + M_{a,b}M_{b,c} + M_{a,c}M_{b,c} \right\} + \\
&\quad 2 \left\{ \sum_{a,b \in R, a<b} M_{a,b} \sum_{c<d \in R, c \neq a,b, d \neq a,b} M_{c,d} \right\}
\end{aligned}$$

The running time of the variance computation is thus $O(n^4)$, which, in many cases, is prohibitive. However, when $a \neq b \neq c \neq d$, $M_{a,b}$ is nearly independent from

$M_{c,d}$, so

$$E[S_k^2] \approx \frac{\binom{k}{2}}{\binom{n}{2}} \sum_{1\le a<b\le n} (M_{a,b})^2 + 2\frac{\binom{k}{3}}{\binom{n}{3}} \left\{ \sum_{1\le a<b<c\le n} M_{a,b}M_{a,c} + M_{a,b}M_{b,c} + M_{a,c}M_{b,c} \right\} + 2\binom{k}{2} \left\{ \binom{k}{2} - 2k + 3 \right\} E[S_2]^2$$

We call this approach the normal approximation method.

4.4.4 Convolution-based approaches

Considering again a random subset of vertices $R = \{r_1, r_2, ..., r_k\}$, we define the random variables $Z_{i,j} = M_{r_i,r_j}$, for $1 \le i < j \le n$ and $Y_i = \sum_{j=1}^{i-1} M_{r_j,r_i}$, for $i = 2...k$ (refer to Figure 4–2) . In this section, we assume that the scores in M are integers. This will always be the case when $M = d_G$. When $M = s_G$, $M = s_{G_W}$, or $M = d_{G_W}$, we assume that elements of M has been appropriately discretized to integers. Observe that $S_k = \sum_{i=1}^{k} \sum_{j=1}^{i-1} Z_{i,j} = \sum_{i=2}^{k} Y_i$. The random variable S_k is a sum of $\binom{k}{2}$ random but dependent variables. If we ignored the dependencies, the distribution of S_k could be obtained as the $\binom{k}{2}$-fold self-convolution of the discrete distribution f_G, where $f_G(a) = \sum_{1\le i<j\le |V|} \mathbf{1}_{a=M_{i,j}} / \binom{|V|}{2}$ is the fraction of entries in M with value a. This turns out to produce a very poor approximation of the distribution of S_k, severely underestimating the correct probability for small values of S_k. We can improve the situation by modeling some of the dependencies. Again, the family of Y random variables are dependent: in particular, if $S_{k-1} = \sum_{i=2}^{k-1} Y_i$ is small, i.e. $r_1, ..., r_{k-1}$ form a tight cluster, then the variance of Y_k is increased, because the variables $Z_{*,k}$ are highly dependent on each other (e.g. if $Z_{i,k}$ is small, then $Z_{i',k}$ is also likely to be small, because i and i' belong to the same tight cluster). We consider two approaches to the problem: the first calculates nearly exactly the distribution of the Y_i's but ignores their dependencies, while the second models the dependencies more accurately but is less accurate at the level of each distribution.

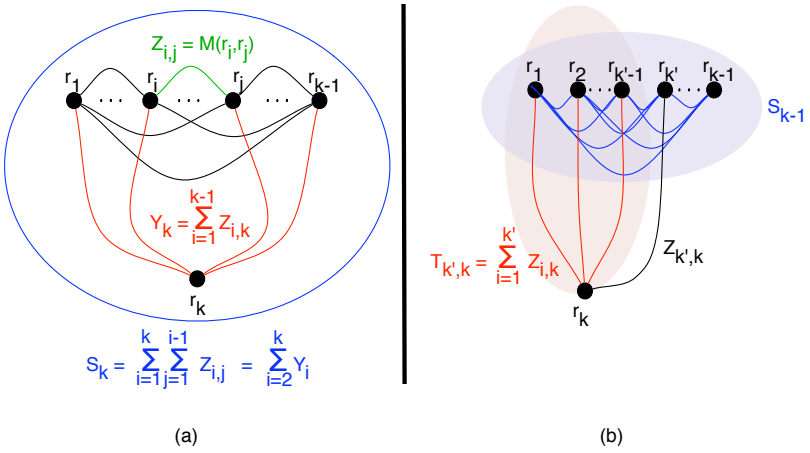

Figure 4–2: Definition of the variables used in the Convolution approaches.

4.4.4.1 The Y-convolution method

Let $g_i(a) = \sum_{j=1}^{n} \mathbf{1}_{a=M_{i,j}}/(n-1)$ be the fraction of pairs of vertices $(i,*)$ with score a and let $g_i^{(l)}$ be the l-fold self-convolution of g_i. Then, $y_l(a) = \Pr[Y_l = a] \approx 1/n \sum_{i=1}^{n}(g_i^{(l-1)})(a)$ (this is an approximation because the convolution models a situation where the random subset R would be allowed to repeatedly pick the same pair of vertices). Assuming the independence of the Y_i's, the distribution of S_k would be obtained by the convolution $y_2 * y_3 * ... * y_k$. We will refer to this approximation as the *Y-convolution* method. Its running time is $O(|V|^2k^2d^2)$, where d is the diameter of G, although the use of Fast Fourier transforms to compute convolutions may yield significant improvements. In the context of WTPD, PSF, or WPSF the running time becomes $O(|V|^2k^2(\delta \cdot \kappa)^2)$, where δ is the discretization factor and κ is $\max_{u,v \in V, u<v} d_{G_W}(u,v)$, $\max_{u,v \in V, u<v} s_G(u,v)$, or $\max_{u,v \in V, u<v} s_{G_W}(u,v)$ respectively.

4.4.4.2 The triangle decomposition methods

An alternate approach is to use a dynamic programming algorithm to better model dependencies (refer to Figure 4–2 (b)):

$$\Pr[S_k = a] = \Pr[\sum_{i=2}^{k} Y_i = a] =$$

$$\begin{cases} \sum_{a'=1}^{a} \Pr[S_{k-1} = a'] \cdot \Pr[Y_k = a - a' | S_{k-1} = a'] & \text{if } k > 1 \\ 0 & \text{if } k = 1, a \neq 0 \\ 1 & \text{if } k = 1, a = 0 \end{cases}$$

Define $T_{k',k} = \sum_{j=1}^{k'} Z_{j,k}$, for $1 \leq k' < k$, so that $Y_k = T_{k-1,k}$. The term of the form $\Pr[Y_k = b | S_{k-1} = c] = \Pr[T_{k-1,k} = b | S_{k-1} = c]$ is calculated using another a convolution-based dynamic programming algorithm.

$$\Pr[T_{k',k} = b | S_{k-1} = c] =$$

$$\begin{cases} \sum_{d=1}^{b} \Pr[T_{k'-1,k} = d | S_{k-1} = c] \cdot \Pr[Z_{k',k} = b - d | S_{k-1} = c, T_{k'-1,k} = d] & \text{if } 2 \leq k' < k \\ \Pr[Z_{1,k} = b | S_{k-1} = c] & \text{if } k' = 1 \end{cases}$$

It is most likely impossible to calculate exactly and in polynomial time $\Pr[Z_{k',k} = b - d | S_{k-1} = c, T_{k'-1,k} = d]$, as otherwise the derivation above would give the exact probability distribution for S_k, which we have shown to be an NP-hard problem. Instead, we boil down the information in the condition $(S_{k-1} = c, T_{k'-1,k} = d)$ to a simpler condition for which the conditional probability is easier to compute. Notice that if $S_{k-1} = c$, the average pairwise distance among $r_1, ..., r_{k-1}$ is $l_1 = c/\binom{k-1}{2}$. Also, if $T_{k'-1,k} = d$, then the average pairwise distance between r_k and $r_1, .., r_{k'-1}$ is $l_2 = d/(k' - 1)$.

4.4.4.3 Rounding approach

We assume that the desired condition can be represented as the condition $Z_{1,k'} = Z_{2,k'} = ... = Z_{k'-1,k'} = [l_1], Z_{1,k} = Z_{2,k} = ... = Z_{k'-1,k} = [l_2]$, where $[l_1]$ is the rounding of l_1, and similarly for l_2 . The information on $Z_{k',k}$ thus comes in the form of $k'-1$ nearly independent pairs $(Z_{i,k'} = [l_1], Z_{i,k} = [l_2])$. Let $t(a,b,c)$ be the number of triplets $1 \leq i < j < k \leq n$ such that $M(i,j) = a, M(i,k) = b, M(j,k) = c$. Assuming the independence of the $k'-1$ conditions, the desired posterior probability of $Z_{k',k}$ is obtained as:

$$\begin{aligned}
&\Pr[Z_{k',k} = b-d | S_{k-1} = c, T_{k'-1,k} = d] \\
&\quad = \Pr[S_{k-1} = c, T_{k'-1,k} = d | Z_{k',k} = b-d] \cdot \Pr[Z_{k',k} = b-d]/\zeta \\
&\quad \approx \left(\frac{t([l_1],[l_2],b-d)}{t(*,*,b-d)} \right)^{k'-1} \cdot f_G(b-d)/\zeta,
\end{aligned}$$

where ζ is a normalizing constant that does not need to be computed (it is sufficient to normalize the distribution to make it sum to 1).

4.4.4.4 Interpolation approach

The rounding procedure yields a rather crude modeling of the actual posterior probability, especially when l_1 or l_2 are far from $[l_1]$ or $[l_2]$ respectively. A better modeling may be obtained as follows. Instead of assuming that all $k'-1$ condition pairs have the same values $[l_1]$ and $[l_2]$, we assume $N_{00} = frac(l_1) \cdot frac(l_2) \cdot (k'-1)$ pairs have values $(\lfloor l_1 \rfloor, \lfloor l_2 \rfloor)$, $N_{01} = frac(l_1) \cdot (1 - frac(l_2)) \cdot (k'-1)$ pairs have values $(\lfloor l_1 \rfloor, \lceil l_2 \rceil)$, $N_{10} = (1 - frac(l_1)) \cdot frac(l_2) \cdot (k'-1)$ pairs have values $(\lceil l_1 \rceil, \lfloor l_2 \rfloor)$, and $N_{11} = (1 - frac(l_1)) \cdot (1 - frac(l_2)) \cdot (k'-1)$ pairs have values $\lceil l_1 \rceil, \lceil l_2 \rceil)$. We thus approximate:

$$\Pr[Z_{k',k} = b - d | S_{k-1} = c, T_{k'-1,k} = d] \approx$$
$$\left(\frac{t(\lfloor l_1 \rfloor, \lfloor l_2 \rfloor, b-d)}{t(*,*,b-d)}\right)^{N_{00}} \cdot \left(\frac{t(\lceil l_1 \rceil, \lfloor l_2 \rfloor, b-d)}{t(*,*,b-d)}\right)^{N_{10}} \cdot$$
$$\left(\frac{t(\lfloor l_1 \rfloor, \lceil l_2 \rceil, b-d)}{t(*,*,b-d)}\right)^{N_{01}} \cdot \left(\frac{t(\lceil l_1 \rceil, \lceil l_2 \rceil, b-d)}{t(*,*,b-d)}\right)^{N_{11}} \cdot f_G(b-d)/\zeta$$

Both triangle convolution approaches run in time $O(k^6 d^3 + |V|^3)$ in the case of TPD, where d is the diameter of G. For WTPD, PSF, or WPSF the running time is $O(k^6(\delta \cdot \kappa)^3 + |V|^3)$, where δ is the discretization factor and κ is $\max_{u,v \in V, u<v} d_{G_W}(u,v)$, $\max_{u,v \in V, u<v} s_G(u,v)$, or $\max_{u,v \in V, u<v} s_{G_W}(u,v)$ respectively.

4.4.5 Identification of core subgraphs

If a GO term τ obtains a small p-value from one of the methods described above, this means that the genes in $V(\tau)$ are unexpectedly clustered within G. This does not, however, mean that every gene in $V(\tau)$ belongs to that dense cluster, but only that a significant subset of $V(\tau)$ does. We call the $core(\tau) \subseteq V(\tau)$ the set of mutually exclusive subsets of $V(\tau)$ that contributes the most to its statistical significance, i.e. the set of one or more subsets of genes in $V(\tau)$ that are the most significantly clustered. $core(\tau)$ may consist of a single dense cluster, or of several dense but distant clusters. In most situations, it is $core(\tau)$, rather than $V(\tau)$, that sheds the most light on the function of a portion of a network. We use a simple partitioning algorithm to reduce $V(\tau)$ to $core(\tau)$, by first building a hierarchical clustering tree of the proteins using the average linkage algorithm and the TPD, PSF, WTPD, or WPSF measures (Algorithm 1). Each node of the tree represents the set of proteins below it in the tree and p-values can be assigned to each node using one of the approaches proposed in Section 4.4.2. We then recursively traverse the tree starting from the root exploring, and deciding to keep the current cluster or to split it into two subclusters corresponding to the left and right subtrees, based on the p-values at the current node and the two children

(see Algorithm 2). The construction of the tree with hierarchical clustering runs in $O(|V(\tau)|^2 \cdot \log(|V(\tau)|))$ and identifying the cores runs in $O(|V(\tau)|)$. This heuristic algorithm does not guarantee optimality but generally succeeds at identifying the key components of $V(\tau)$. The results presented in Section 4.5.2 are the cores of the GO terms that obtained good p-values.

Algorithm 1 Find Core subgraph

Input: Distance matrix d_G (or d_{G_W}, or derived from s_G or s_{G_W}),
Vertex subset $V(\tau)$, maximum p-value of interest mpv
Output: Vertex subset $core(\tau)$

```
root ←HierarchicalClustering(V(τ), d_G)
return DivideCluster(root, mpv)
```

Algorithm 2 DivideCluster

Input: Root r, maximum p-value of interest mpv
Output: A set of subsets of vertices of the subtree rooted at r that form significant clusters

```
if p-value(r)> mpv then
  return φ
else
  if p-value(r) < p-value(leftChild(r))  and  p-value(r) < p-value(leftChild(r))
  then
    return {V(r)}  % where V(r) is the set of vertices in the subtree rooted at r.
  else
    return DivideCluster(leftChild(r), mpv) ∪ DivideCluster(rightChild(r), mpv)
  end if
end if
```

4.4.6 Implementation considerations

The implementation of some of the four approximation schemes described in this section proves quite technically challenging, with issues of numerical precision arising for the two triangle convolution. Our crude approach to the problem is to make sure that, at every step, the intermediate probability distributions are properly normalized to sum to 1, although more subtle approaches would certainly improve our accuracy. Another issue is the time and memory required for the computations of the triangle convolution approaches, which require the storage of numerous large intermediate

tables, currently limiting their utilization to the computation of p-values for values of k less than 25. Program optimizations were required to accelerate the running time for the triangle convolution approaches. They consist in stopping the computations of a distribution for a given S_k when the only probabilities left to compute are those at the right tail of the distribution that are smaller than the 64-bit double precision. The discretization level chosen to be applied for the PSF, WTPD and WPSF methods was also an important aspect to consider in the implementation. A coarse discretization of the distributions can accelerate the running time for methods like the Y-convolution or both triangle convolutions, but provide a rather inaccurate estimation of the final distributions of S_k. On the other hand, a much finer discretization would require much more computational time but would yield a more accurate approximation. The challenge resides in determining the degree at which the distribution will be discretized in order to compute in reasonable time an accurate distribution approximation.

4.5 Results

4.5.1 Accuracy of p-value approximation methods

The accuracy of our four p-value approximation schemes can be assessed by Monte Carlo simulations: for a given graph G, repeatedly sample randomly a subset of k vertices and compute the sum of pairwise scores to eventually obtain an unbiased estimate of the true distribution. The limit of this approach is of course that the accuracy of the estimation depends on the number of samples, making small p-values difficult to estimate quickly.

We have measured the accuracy of our approximation approaches on both simulated and actual biological networks. Protein-protein interaction networks have been reported to be accurately modeled by scale-free random graphs [11], although geometric random graphs have also been used [180]. We randomly generated scale-free graphs with 1000 vertices and a number of edges ranging from 1000 to 3000. In

total, 2100 random graphs were generated. The distributions of the TPD and PSF scores were estimated empirically, using 10^6 samples, for each graph and each value of $k = 5, 10, 20, 50$. For each combination, critical values $Z_{0.1}, Z_{0.01}$, and $Z_{0.001}$ were estimated as being the value of TPD and PSF that obtains the empirical p-value 0.1, 0.01, and 0.001, respectively. Each of the four analytical approximation methods[2] were then used to estimate the p-values for $Z_{0.1}, Z_{0.01}$, and $Z_{0.001}$. Figures 4–3 and 4–4 report the accuracy of the p-values produced by each of our methods for the TPD and PSF clustering measures, for the target p-values 0.1, 0.01, and 0.001, and for $k = 5, 10, 20, 50$. We start by observing that although our p-value approximation methods apply in principle to both the TPD and PSF clustering measures, specificities of these datasets result in our methods behaving quite differently. This is due to the fact that the similarity scores that constitute the PSF clustering scores exhibit much stronger inter-dependencies than the pairwise distances that constitute the TPD clustering score, resulting in worse approximations when independence in assumed. Our observations are summarized below.

- **Y-convolution.** In the case of TPD, this method severely underestimates small p-values, by a factor ranging from 2 to 100 for $k = 5$ to more than 10^4 for $k = 50$. This due to the fact that dependencies in the graph are greatly underestimated. However, the approximation improves with the edge density. On the contrary, the method works quite well on PSF clustering for graphs with low edge density, but it severely underestimates p-values of highly connected graphs.
- **Normal approximation.** This approximation obtains much better results than the Y-convolution approximation in the case of TPD clustering, producing p-values that generally slightly over-estimate the correct p-value (1- to 3-fold for small k, 10- to 50-fold for k=50). Surprisingly, although, for small k, the

[2] Note that the triangle decomposition with interpolation approximation was not performed for PSF because of its high memory and running time requirements.

quality of the approximation improves with the edge density, the opposite trend is observed for larger k. However, for PSF clustering, this yields an extremely poor approximation for all values of k, erring by a factor ranging from 10^{10} to 10^{60} for a true p-value of 0.001.

- **Triangle decomposition with rounding.** We found that this method is an improvement to the Y-convolution approximation for TPD clustering since it does not underestimate as much p-values for small k (factor ranging from 2 to 10 for $k = 5$ and from 10 to 100 for $k = 10$). However, it behaves more irregularly for $k = 20$, underestimating the p-values by a factor greater than 100. This approach also yield good approximations for PSF clustering, overestimating small p-values for any k by a small margin. Interestingly, for both clustering measures, the accuracy of this approximation does not seem to be affected by the edge density of the network.
- **Triangle decomposition with interpolation.** The results obtained from this method on TPD clustering are comparable to the normal approximation estimation. For p-values 0.01 or less, computed p-values are slightly over-estimating the correct p-values (1- to 4- fold for small k). It sometimes even provides a tighter upper bound on the correct p-values. Again the accuracy of the p-value estimation for this method is not influenced by the edge density. We were unable to use this approximation for PSF because of high running time and memory requirements of the method.

Notably, all 4 methods behaved extremely similarly in terms of accuracy for both WTPD and WPSF compared to their respective unweighted version TPD and PSF. Overall, we conclude that given how quickly it can be computed, the normal approximation approach is the best tradeoff between running time and accuracy for TPD. However, the quality of that approximation degrades with the edge density, which is not the case for the two Triangle convolution approaches. This is an important point since we expect protein-protein interaction networks to gain in edge density as new

high-throughput assays become available. The Triangle convolution approach is also the most accurate for PSF. It is the only method providing tight upper bounds on p-values even for large k in highly connected graphs. However, because of its intensive use of memory and slow running time, it is hard to obtain p-value approximations for very large k. Since it produces p-value approximations in a much more reasonable time, the Y-convolution method can be used in this situation.

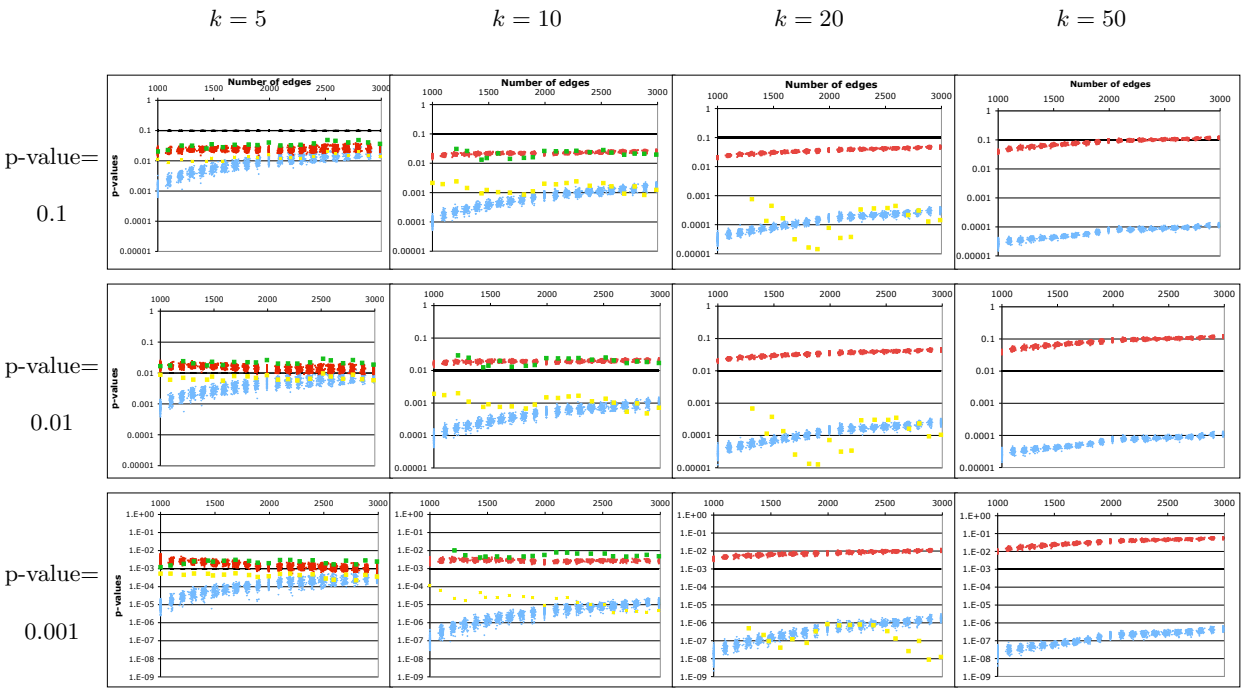

Figure 4–3: p-values predicted by our four approximation schemes (Normal: red; Y-convolution: blue; Triangle convolution with rounding: yellow; Triangle convolution with interpolation: green) for the TPD clustering measure. Each data point records the approximated p-value (y-axis) for the TPD score that obtained the given empirical p-value (0.001, 0.01, 0.1), on a random scale-free graph with 1000 vertices and the given number of edges (x-axis). The triangle convolution with rounding method was too slow to be evaluated for $k > 20$, and that with interpolation could only be run for $k = 5$ and $k = 10$.

Our results on two larger actual PPI networks in yeast [119] and human [101] (see Section 4.5.2) largely confirm our observations on random graphs. Figure 4–5 shows the complete TPD distributions (for $k = 10$) obtained by Monte Carlo simulations, as

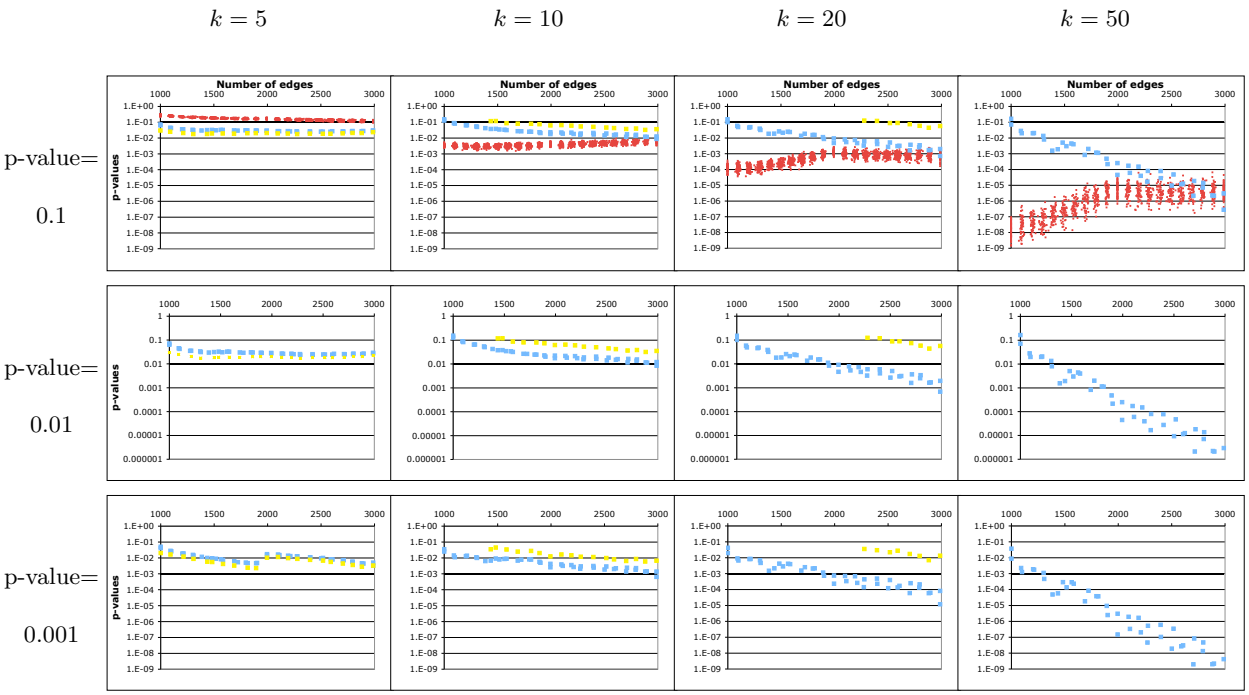

Figure 4–4: p-values predicted by three approximation schemes (Normal: red; Y-convolution: blue; Triangle convolution with rounding: yellow) for the PSF clustering measure. See caption of Figure 4–3. The triangle convolution with rounding method was too slow to be evaluated for $k > 20$ and some graphs for $k = 20$. The triangle convolution with interpolation was too slow for all k. The Normal approximation method produced p-value estimates that were too poor to show on these graphs, usually erring by a factor of 10^{10} or more.

well as each of our approximation methods, for the Krogan et al.'s yeast PPI network, which consists of more than 2500 proteins and 7000 interactions.

Of the four approximation methods proposed, the fastest is the normal approximation (Table 4–1). The Y-convolution method is approximately 10-fold slower, while the two triangle-based convolution approaches are several orders of magnitude slower. Note that for PSF, Triangle convolution with interpolation runs several order of magnitude slower than the values presented.

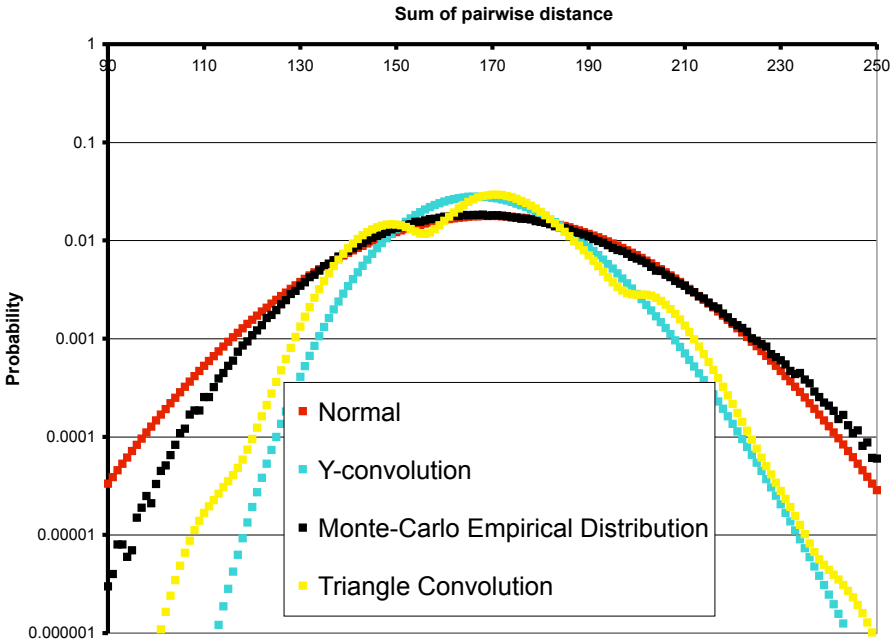

Figure 4–5: Empirical and approximated TPD distributions for the yeast PPI network, for k=10.

Table 4–1: Approximate running time, in minutes, to calculate one clustering p-value for a 1000-vertex scale-free graph with 2000 edges, for the TPD clustering measure.

	$k = 10$	$k = 20$	$k = 50$
Monte Carlo simulation[(a)]	2	5	20
Normal	0.3	0.7	1
Y-Convolution	1	5	15
Triangle with rounding	2	300	>1000
Triangle with interpolation	5	600	>1000

[(a)] 10^6 samplings were performed for the Monte Carlo simulations.

4.5.2 Biological analyses

We first applied our analysis using TPD to the yeast protein-protein interaction data set produced by Krogan et al. [119]. We analyzed the largest connected

component of their "core" network, which consists of 2559 proteins and 7037 interactions. Of the 299 GO terms present more than twice in the network, 91 obtained a normal approximation (conservative) p-value below 0.05 (corresponding to a FDR $= \frac{299\times0.05}{91} \approx 16\%$), and 42 obtain a p-value below 0.001 (FDR $= \frac{299\times0.001}{42} \approx 0.7\%$). As seen on Figure 4–6, the GO terms with significant p-values allow the automated annotation of much of the network. For many of the GO terms reported, our results reflect known protein complexes (e.g. ribosome, ribonuclease MRP, general pol-II transcription factors, etc.). Other clusters, often the larger, more diffused ones, do not correspond to complexes but rather contain proteins that interact with many of the same partners (e.g. the translation initiation factors or the signal sequence binding proteins). While most GO terms form a single, dense cluster, some, such as the structural components of the ribosome, the general RNA pol-II TFs, and the endopeptidases, are broken into two or three dense subgroups. Many of the fundamental functional interactions between groups of proteins of different function immediately stand out, for example the interplay between histone deacetylases (yellow), histone acetyltransferases (in cyan), and ATP-dependent 3'-5' DNA helicases (in green). The annotated network is clearly more interpretable and readily allows the formulation of specific hypotheses about the function of various unannotated proteins and of the various interactions observed. See Supplementary material for complete results.

Finally, we analyzed a human protein-protein interaction network published by Jeronimo et al. [41] using PSF and WPSF. The network contains 1053 proteins and 2014 interactions, built from 32 tagged proteins and their interactors in the soluble fraction of HEK293 cells. The tagged proteins are predominantly proteins related to the (extended) transcription machinery. As can be seen from Figure 4–7, the network is quite dense and existing automated layout systems fail to reveal much of the biological information contained in the graph. We ran our analyses on the network to identify which of the 135 GO categories present more than twice in the graph show unexpected clustering. 24 GO categories obtained p-values below 0.05

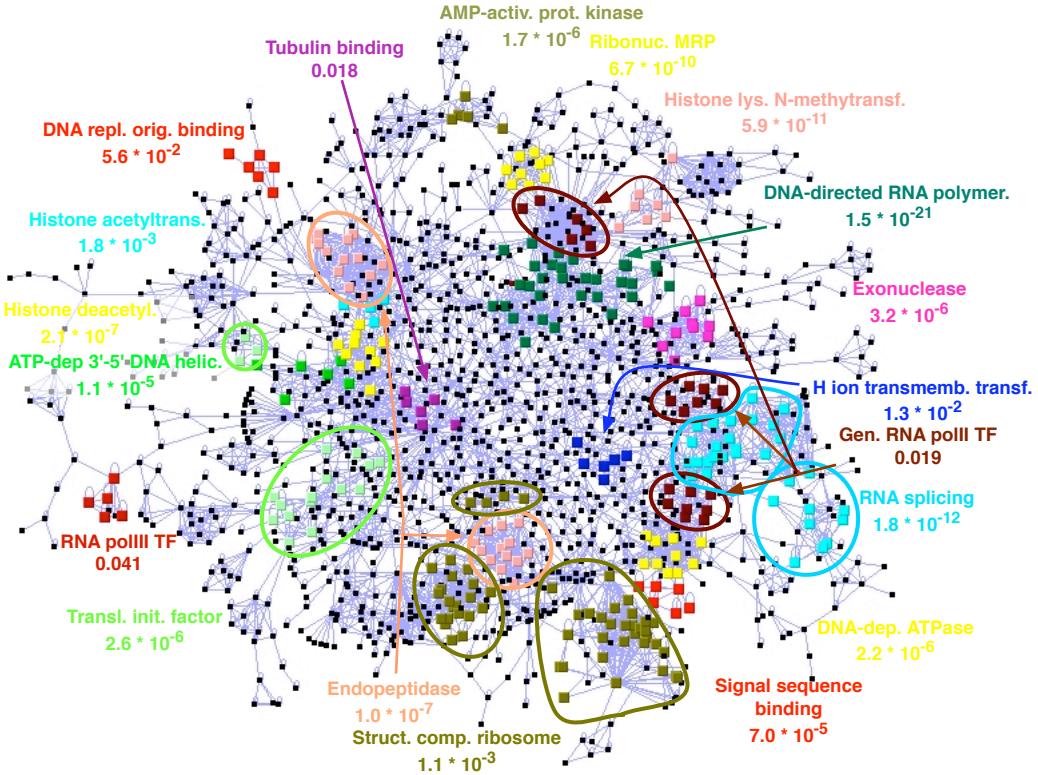

Figure 4–6: Yeast PPI network from Krogan et al. [119], annotated with the cores of some of the GO categories with significant clustering. The p-values given were obtained using the Normal approximation approach, which is almost always conservative. For readability, not all significant GO categories are shown. Subsets of $core(\tau)$ of size at least 3 are shown.

(FDR $= \frac{135\times0.05}{24} \approx 28.1\%$ for PSF and 19 for WPSF (FDR $= \frac{135\times0.05}{19} \approx 35.5\%$; see Supplementary material). Genes belonging to some of these categories are colored coded in Figure 4–7 (several categories are somewhat redundant; only one representative per group is shown). When the graph is manually laid out to highlight the connectivity among the selected protein groups (Figure 4–7), the role of several subnetworks is clearly revealed. For example, we can easily identify subunits of the RNA polymerase I, II and III, classified by GO as "DNA-directed RNA polymerase activity", which are clustered together. We also notice that RPAP1 is tightly connected to the POLR2 subunits within that cluster. This corroborates the observation of Jeronimo et al. where RPAP1, XAB1, C1ORF82, and FLJ21908 (now referred as

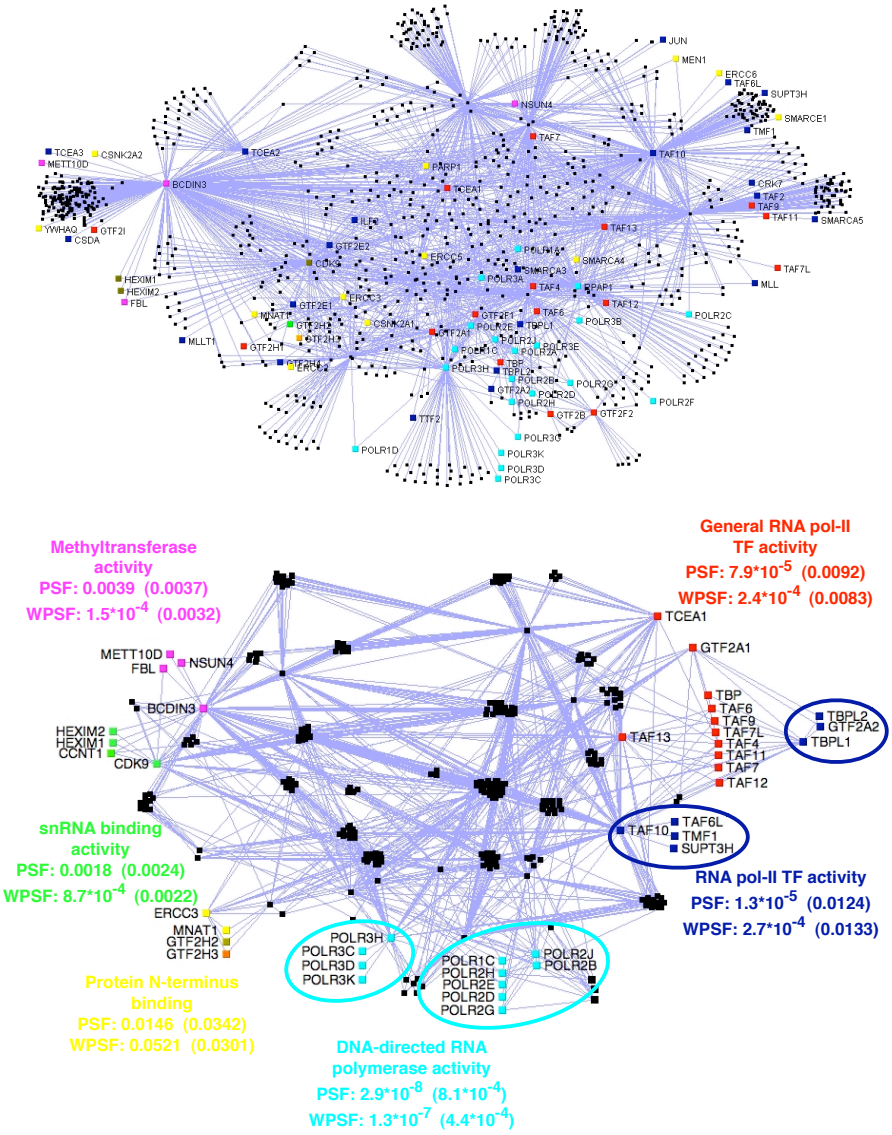

Figure 4–7: (Top) Human PPI network from Jeronimo et al. [101], laid out using the "relaxed" automatic layout procedure of VisANT [92]. (Bottom) Groups of protein with a significant PSF and WPSF clustering p-value are highlighted in colors. The Triangle convolution was used when the group size was small enough; otherwise, the Y-Convolution was used. Monte Carlo estimated p-values are between parentheses. Network laid out manually to highlight the connectivity of the proteins within each GO category reported (to improve readability, proteins that do not belong to any shortest paths between pairs of proteins of the selected groups are not shown). GTF2H3 (in orange) is part of both red and yellow groups. GTF2H2 (in khaki green) is part of both the yellow and blue groups. Subsets of $core(\tau)$ of size at least 3 are shown. Clearly, without the information provided by our GO clustering approach, the PPI network showed at the top would be hard to interpret.

RPAP2 and RPAP3 respectively) are forming an interface between the RNA polymerase II subunits and some molecular chaperone and prefoldins. We can also see that our method, by highlighting this GO term, facilitated the visualization of the interactions between the POLR2 subunits with the XAB1, RPAP2, and RPAP3 proteins. Hexamethylene bis-acetamide inducible (HEXIM) proteins were also found to be clustered with cyclin-dependent kinase 9 (CDK9) and cyclin T1 (CCNT1), both members of the P-TEFb complex [172]. All of these are associated with the GO term "snRNA binding". Interestingly, HEXIMs are known to be inhibitors of the cyclin-dependent kinase activity of P-TEFb [13, 26]. In addition, BCDIN3 (also known as MEPCE) and SART3, which are part of the 7SK snRNP complex, itself containing P-TEFb, are closely associated with HEXIMs and CDK9 [101, 41]. Finally, numerous TATA box binding protein (TBP)-associated factors (TAFs) and a general transcription factor II (GTF2A1), all sharing the "general RNA polymerase II transcription factor activity" GO function, were found to be significantly clustered. Many of these TAFs and GTF2A1 are interacting with TBPL1, another protein playing a key role in transcription [164].

4.6 Discussion and future work

The idea described in this paper, of seeking gene attributes that cluster within a given network, can be used to annotate PPI networks with any type of gene or protein features. Besides gene ontologies, we are currently expanding our tool to use protein domains from the PFAM database [65], pathways from the KEGG database [106], and gene expression data. Indeed, any annotation coming in the form of gene sets can be used to annotate the network, including, for example, those collected through the laudable efforts of the GSEA [212] team.

In the future, we will try to improve the accuracy and efficiency of our approximation algorithm. We will also seek provable approximation bounds for the p-value estimation problem. Currently, one of the main computational issues is that some of

our best approximation methods are quite slow and require a lot of memory. More efficient implementations would thus have a significant practical impact.

In this paper, we only studied the simplest version of a family of interesting problems. A number of extensions will be considered. One important generalization is to consider directed graphs. In these graphs, the edges directions represent biological information about the tag experiment that was performed. For instance, an edge would connect two proteins from the tagged protein to the purified protein. We are also considering the problem where gene annotations are not in the binary form (i.e. they belong to a given gene set or not) but are more quantitative measures, such as gene expression.

As we discussed previously, our method could be used for protein function prediction. For a given set of proteins sharing the same GO term that are surprisingly clustered, uncharacterized proteins co-clustering with the GO term could be expected to share the same GO annotation. Another exciting prospect is to use this type of local over-representation to search for sequence motifs. One would seek motifs that are locally enriched in a subnetwork of the graph. Locally over-represented motifs found in protein sequences may correspond to new domains or localization signals. Those found in the 5' or 3' UTRs of genes may contain mRNA localization signals or post-transcriptional regulatory elements relevant to the subnetwork, while those found in the regulatory regions (promoters and enhancers) would allow the coordinated transcription of the proteins in the subnetwork.

4.7 Acknowledgements

This work was funded by a CIHR operating grant to BC and MB and NSERC USRA and CGS scholarships to MLA. We thank Pablo Cingolani for his help on the GOA database and Ethan Kim and Ashish Sabharwal for useful suggestions.

4.8 Supplementary material

The Java program used to identify GO terms enriched in subnetworks is available at: `http://www.cs.mcgill.ca/~blanchem/GoNet`. All other supplementary files are available at the same location.

CHAPTER 5
Detection of functional sequence motifs in human 5' UTRs based on local enrichments in a protein-protein interaction network

5.1 Preface

In the previous chapter we have shown that GoNet has the capability to identify sets of proteins annotated with the same Gene Ontology (GO) term that are clustered in PPI networks. However, protein annotations may take many more forms than GO terms. For example, GoNet could be used to find proteins known to be involved in the mechanisms of a certain disease (e.g. from OMIM [84]) and that are clustered in the network. Many other types of annotations such as gene co-expression or co-methylation could also be used to perform a GoNet analysis.

One weakness of the use of GO terms as annotations for a GoNet analysis resides in the fact that some GO-protein associations may be derived from PPI networks themselves. For instance, a GO term may represent a given protein complex that was discovered using AP-MS. This circularity may hinder the biological discovery power of GoNet, which may report annotations derived from protein clusterings well characterized in PPI networks. The use of an approach similar to GoNet with annotations such as gene co-expression, co-methylation, or any annotations derived independently from PPI network data is likely to show a greater discovery potential. In this context, the relationships found by this approach between the clustered proteins and their shared annotations may be of greater biological interest as they may not be as trivial as those identified with GO terms (e.g. protein subunits of the ribosome that are clustered in the network as well as being annotated with the "ribosome" GO term.)

In this chapter, we propose to use a more unconventional type of protein annotation with a computational tool based on GoNet called: Local Enrichment of Sequence Motifs in biological Networks (LESMoN). We consider a set of proteins in a PPI network to share the same annotation if their sequences contain at least one occurrence of a given sequence motif.

An alternative approach to both GoNet and LESMoN consists in using a network clustering algorithm such as MCL [59] to identify clusters in an input network and then testing if those clusters are enriched for a certain annotation (GO term or sequence motif). The main problem with such strategy resides in the fact that most clustering strategies like MCL, when used on massive networks such as BioGRID [210], fail to identify large densely connected subgraphs that are embedded in dense subgraphs containing even more vertices. MCL, for instance, partitions the network into several very small clique-like subgraphs, while keeping more than half of the network as one large cluster.

The remaining content of this chapter is taken from:

- M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of functional sequence motifs in human 5' UTRs based on local enrichments in a protein-protein interaction network. *Manuscript in preparation.*

5.2 Abstract

Protein-protein interaction (PPI) networks are becoming increasingly large and complex with the accumulation of publications of high-throughput studies. We previously presented a computational method (GoNet) to identify Gene Ontology terms that are significantly clustered in PPI networks in order to help the interpretation of such intricate networks. We now propose to analyze the clusterings in PPI networks of proteins whose associated 5' UTR sequences contain a common motif. Specifically, we introduce a computational approach, LESMoN, to assess the statistical significance

of the clusterings of large sets of proteins sharing the same 5' UTR sequence motif in PPI networks and present several tools to evaluate the biological importance of the motifs identified. We show that when applied to the human BioGRID PPI network, our approach identifies several uncharacterized and known 5' UTR sequence motifs whose associated proteins are clustered in the network. The vast majority of these motifs are evolutionary conserved. Finally, we establish that the genes containing such motifs are significantly enriched with various Gene Ontology terms suggesting new associations between 5' UTR motifs and a number of biological processes.

5.3 Introduction

Enrichment analyses, where one identifies properties that are found in a set of genes of interest more often than expected by chance, have become an almost inescapable step in the analysis of high-throughput biological studies. There, sets of genes of interest may correspond to genes that are differentially expressed between conditions, cell types, or diseases, targeted by a given transcription factor or miRNA, or encoding a set of interacting proteins. The properties, or annotations, considered may originate from the controlled vocabulary functional annotations of the Gene Ontology (GO) project [4], pathway databases such as Kegg [107], or more comprehensively from gene sets from the MSigDB [140]. However, more generally, any function that separates genes into two sets - those that have the property and those that do not - can be used for gene set enrichment analysis. These include, among others, the presence of a given sequence motif in the protein sequence, in the gene encoding it, or in the regulatory regions of that gene. Irrespective of the nature of the annotation considered, a gene set enrichment suggests a direct or indirect relationship between the annotation and the property or behaviour of the gene set, provided appropriate controls and statistical approaches are used.

A typical strategy to test for the enrichment of an annotation (originating from GO or MSigDB) in a given set of genes S taken from the whole set of genes Ω of an

organism is to perform an hypergeometric or Fisher's exact test, which contrasts the proportion of genes with the property of interest in S to that in Ω and assigns a p-value reflecting the probability that an enrichment equal or greater would be observed by chance [232, 3, 15, 14]. Sequence motif discovery techniques, especially approaches that enumerate candidate motifs to identify those that are the most significantly enriched in S, fall under a similar umbrella. These include expectation-maximization algorithms (MEME [10]), Gibbs sampling (AlignACE [185]), word statistics approaches (YMF [202]), or ensemble approaches (SeSiMCMC [62], Amadeus [142]).

We have previously described such gene sets $S \subseteq \Omega$ as being "unstructured", since each gene in it contributes equally to the annotation enrichment in the analysis [130]. An extension to this enrichment analysis was explored by the famous Gene Set Enrichment Analysis (GSEA) computational tool [212]. Instead of separating genes into those that are "of interest" and those that are not and seeking enriched annotations in the former, GSEA takes as input a ranked list of genes based on their "level of interest" with respect to a particular measure (e.g. over-expression in a given condition) and identifies annotations whose distribution in the ranked list is non-uniform. In that sense, we could say that GSEA takes advantage of a "weak structure" defined on Ω by the (single) measure of interest, to identify annotations that are non-randomly distributed in this structured space.

Another type of annotation that has been considered is the presence of a given sequence motif (e.g. one represented by a regular expression) in a DNA sequence associated to a gene (e.g. its promoter). Here, for a given motif m, a gene is said to have the annotation m if its promoter contains one or more copies of m. Such approaches have been quite successful at identifying transcription factor binding site motifs based on ranked lists of differentially expressed genes [135, 136, 56] or sequences sorted by their affinities to a given transcription factor, as obtained by protein-binding arrays [31].

In previous work, we showed that GO enrichment analysis can be applied to much richer structures such as those defined by biological networks, e.g. PPI networks [129, 130]. In that case, annotations of interest were those where the genes (or proteins) with the property were non-randomly distributed in the network, i.e. more clustered than expected by chance, representing the so-called local enrichment of the property. In this paper, we introduce the Local Enrichment of Sequence Motifs in biological Networks (LESMoN) approach, which uses a similar principle for the detection of local enrichments of a different type of annotation, namely sequence motifs found in the mRNAs encoding the proteins in the network. In this paper, we focus on a particularly function-rich portion of the mRNA, the 5' untranslated region (UTR). Specifically, given a PPI network G with 5' UTR sequences associated to all proteins in G, our objective is to identify all sequence motifs, represented using simple regular expressions, for which the associated proteins are surprisingly clustered in G. The motifs identified are expected to be somehow linked to the biological function of the subnetwork where they reside, and we propose additional analyses that suggest what role this may be.

In RNAs, 5' UTR sequences play key roles in post-transcriptional regulation. Specific primary and secondary structure motifs regulate translation [116, 171, 99, 177, 178]. 5' UTRs are implicated for example in the translation regulation of ribosomal proteins and proteins involved in protein synthesis through a 5' TOP motif [169, 157]. Furthermore, 5' UTRs often contain intracellular localization elements, which are required for the binding of their mRNAs to certain cell structures such as membranes [192] and synapses [152]. In addition, riboswitches, a mechanism by which a section of mRNA adopts a certain structure to regulate the translation of its encoded protein, are known to be most often located in the 5' end of 5' UTRs of bacterial mRNAs [149]. Also, variations of the transcription start site and alternative splicing can lead to the production of different 5' UTR sequences for a gene. These sequence variations may contain different regulatory motifs affecting the mRNA of

the gene or protein translation [94]. Obviously, the DNA that encodes 5' UTRs can also host transcriptional regulatory regions such as transcription factor binding sites for which the binding motif may appear enriched in 5' UTRs; we therefore propose approaches to separate candidate transcriptional from post-transcriptional regulatory motifs.

As we show in this paper, LESMoN is capable of identifying a large set of sequence motifs that associate with specific functional subnetworks, including motifs involved in transcriptional regulation, translation regulation, splicing, and others. Whereas LESMoN recovers known functional motifs in 5' UTRs (e.g. the GGUG binding motif of the protein FUS or the CG rich binding motif of RBM4), the majority of the motifs identified appear uncharacterized. For most of them, additional evidences (inter-species conservation, position and strand biases, GO enrichment of corresponding proteins, etc.) point to specific functions.

5.4 Methods

The goal of our approach is to find 5' UTR sequence motifs for which the associated proteins are surprisingly clustered in a given PPI network. We enumerate all possible motifs over a given alphabet (described in Section 5.4.2) and test whether the motifs are clustered. To this end, we present a measure of protein clustering in PPI networks and methods to evaluate the clustering statistical significance. Should a motif be significantly clustered, it would suggest that the motif is linked directly or indirectly to the biological mechanism causing the clustering of the associated proteins in the PPI network. We also present in this section various tools and strategies to evaluate the biological significance of the motifs for which the proteins are clustered.

5.4.1 Protein-protein interaction network

We tested our approach on the human PPI network downloaded from the BioGRID database (version 3.2.97) [30, 210], one of the most comprehensive human PPI network

available. The network contains 14,113 proteins forming 127,433 unique interactions. Even if this network can be treated as directed because of the nature of the experiments used to build it, we decided to consider it as undirected since edge directionality is only an artefact of experimental procedures and is irrelevant when considering the real biological data in this context. Only the largest connected component $G = (V, E)$ of this network ($|V| = 14,021$ proteins and $|E| = 120,146$ interactions) was used in the present analysis to facilitate the use of our distance measure described in Section 5.4.3.

5.4.2 5' UTR motif enumeration

All 5' UTR exon sequences of the mRNAs of the proteins present in the human BioGRID PPI network were obtained from the RefSeq gene annotation through the UCSC Table Browser (28 Feb. 2013). When a protein was associated to multiple 5' UTR variants, their union was associated to that protein. To avoid the inclusion of a few misannotated 5' UTRs spanning over gene exons, which are in fact translated, we only considered the first 500 nucleotides (at most) of each 5' UTR. This includes the full length of more than 90% of 5' UTRs in our dataset. We then enumerated sequence motifs of length 8 over the alphabet $\sigma = \{A, C, G, U, R, Y, N\}$, where $R = [AG]$, $Y = [CU]$, and $N = [ACGU]$. Motifs of length smaller than 8 are represented in this enumeration with the inclusion of the N character at the beginning or the end of 8-mer motifs. Motifs of length 6, the typical size of RNA-binding protein motifs, are therefore considered [89]. A protein was annotated as containing a given motif if the corresponding 5' UTR had at least one match to it, considering only the forward strand (i.e. matches to the reverse complement are not considered).

5.4.3 Clustering measure

We previously used the following method to measure the clustering of a set of proteins in a PPI network [129, 130]. Let u and v be two vertices from the set of

vertices V. We define $d_G(u, v)$ to be the length of the shortest path in G between u and v. Floyd-Warshall's algorithm [67, 225] was used to compute the distance matrix d_G. Now let $V_m \subseteq V$ be the set of all proteins annotated with the motif m. We define the total pairwise distance (TPD) for V_m as

$$TPD(V_m) = \sum_{u,v \in V_m, u<v} d_G(u, v). \tag{5.1}$$

5.4.4 Clustering statistical significance

We previously showed how to evaluate the statistical significance of a given value of the TPD [129, 130]. However, that approach only works for small sets of proteins ($|V_m| < 100$) and uses a null model that is not appropriate here. The approach presented here is therefore slightly different. This strategy computes the distribution of the random variable $S_k = \sum_{i,j \in R, i<j} d_G(i, j)$, where $R = \{r_1, r_2, ..., r_k\} \subseteq \{1, ..., |V|\}$ is a randomly selected subset of k proteins. Contrary to our previous work where every node in the network was chosen with equal probability, the appropriate null model here is one where the probability that a given protein is selected in S_k is proportional to the length of its 5' UTR. To evaluate the statistical significance of the clustering of the proteins associated to a motif m, a p-value is then calculated as follows: p-value$(m) = \Pr[S_{|V_m|} \leq TPD(V_m)]$. In order to compute clustering p-values, we introduce two methods to calculate the distribution of S_k, one for small protein sets (≤ 300) and another for larger sets (> 300).

5.4.4.1 Monte Carlo sampling

We proved previously that the exact computation of the distribution of S_k is NP-hard [129]. We therefore cannot expect to perform this calculation exactly in polynomial time. Nevertheless, the statistical significance of the level of clustering of a set of proteins can be estimated using Monte Carlo sampling, where k proteins are repeatedly sampled and the TPD evaluated, in order to estimate the distribution of

S_k. Because the time required to compute $TPD(S_k)$ is $O(k^2)$ (once the full pairwise distance matrix d_G is computed) and this procedure needs to be repeated a large number of times (e.g. 10^6 times to obtain a p-value accuracy of approximately 10^{-6}), it is only feasible for values of k at most 300. However, for most motifs m, $|V_m| > 300$, so a faster approach is required.

5.4.4.2 Normal approximation

In a previous publication, we demonstrated that the distribution of S_k can be estimated using a normal distribution when k and $|V|$ are large [129]. We therefore propose to estimate the distribution of S_k when $k > 300$ with a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Although the mean μ can be computed exactly in time $\mathrm{O}(|V|^2)$, the exact computation of the variance σ^2 has a prohibitive running time ($\mathrm{O}(|V|^4)$). This approach is therefore not applicable. Instead, for each value of k between 2 and 1500, we estimate μ_k and σ_k^2 using a limited amount of Monte Carlo sampling. We assessed the quality of the p-value estimation of the normal approximation technique by estimating the distribution of S_k using 10^4, 10^5, 10^6, and 10^7 randomly chosen samples for $k = 300$ in the human BioGRID PPI network. We compared the p-values obtained for the TPDs with these approximations to our gold standard, which consists of the p-values obtained from the distribution of S_{300}, estimated using the Monte Carlo sampling procedure with 10^7 samples. Supplementary Figure 5–3 shows that excellent accuracy can be achieved using only 10^5 samples. There is practically no gain of using 10^6 or 10^7 samples, which would require much more computational time. We therefore opted to randomly choose 10^5 samples in order to estimate the mean and variance for the normal distribution approximation. The estimated normal distributions are then used to obtain the desired p-values for cases where $300 < k \leq 1500$. The significance of the clustering of motifs present in more than 1500 5' UTRs is not assessed, due to the excessive computational burden. This does not represent a big loss as these motifs are likely to be mainly composed of degenerate characters (R, Y, N) and yield very little biological significance. We also use this normal approximation for cases

where $k \leq 300$ and where the p-value estimated by the full Monte Carlo sampling from the previous section is too small to be estimated accurately ($< 10^{-6}$).

5.4.5 False discovery rate inference

Since an important number of 5' UTR motifs (at most 7^8) are tested for the clustering significance of their associated proteins, multiple hypothesis testing is a significant issue. These statistical tests are far from being independent, because many motifs tested are variants of each other, making a p-value correction such as Bonferroni correction [55] too stringent. To address this issue, we randomize the 5' UTR sequences in our dataset to estimate a false discovery rate (FDR) for every clustering p-values. More precisely, the order of the nucleotides of each 5' UTR sequences is permuted within non-overlapping windows of 10 nucleotides, in order to preserve local sequence properties such as GC content. Motif clustering p-values are then obtained for this randomized dataset, using the same procedure as described above. Let $M(p)$ be the number of motifs that obtained a p-value at most p in the actual set of sequences, and $N(P)$ be the number of such motifs in the permuted data set. We then calculate the FDR for a given p-value p as $FDR(p) = N(p)/M(p)$.

5.4.6 Gene Ontology enrichment analysis

To investigate the mechanisms in which the motifs identified by LESMoN may be involved, we used Ontologizer [14] (with the complete set of proteins V as background) to determine, for each motif, whether the set of associated proteins is enriched for particular Gene Ontology categories.

5.4.7 5' UTR motif conservation

To further explore the biological significance of the motifs detected by LESMoN, we evaluated their level of evolutionary conservation. For each motif, we obtained the number of 5' UTR matching sites whose middle position is contained within a highly

conserved genomic regions among placentals (phastConsElements46wayPlacental [201] from the UCSC Genome Browser). We then compared this overlap to the overall proportion of 5' UTR bases that are conserved among these same placentals and obtained a p-value using a simple binomial test. To avoid numerical instability, this binomial distribution was actually approximated using a normal distribution, with little loss of accuracy given the very large number of sites involved.

5.4.8 5' UTR motif strand specificity and 5' UTR positional enrichment evaluation

In order to evaluate the likelihood of a 5' UTR motif to play a functional role at the mRNA level rather than the DNA level, we measured its strand specificity, defined as the ratio of the number of occurrences of a motif to the number of occurrences of its reverse complement. One would expect post-transcriptional motifs to have a high strand specificity (> 1), whereas most transcriptional regulatory elements, whose function is often independent of strand orientation, may have a strand specificity close to 1. In addition to a high strand specificity, 5' UTRs that play a role post-transcriptionally, are expected to occur more often after the transcription start site than before (i.e. in promoter). To measure this occurrence bias for a given motif, we computed the difference between the averages of fractions of positions covered by the motif in 5' UTRs and in promoters. We performed then calculated the same difference for its reverse complement, which should be small for motifs for which the reverse complement is not functional. We then took the difference of these two calculations to discriminate motifs likely to be functional post-transcriptionally. A motif was judged as playing a role predominately in mRNAs if this difference was $> 8.0 \cdot 10^{-4}$ or $> 7.6 \cdot 10^{-4}$ and that its strand specificity ratio was > 1.4. These thresholds were chosen somewhat arbitrarily based on visual inspection of motif occurrences in promoters and 5' UTRs.

5.4.9 5' UTR motif families

To facilitate the analysis of our motifs, we used a hierarchical clustering approach to group motifs into families based on the similarity of the sets of proteins they are associated to. Specifically, let m_1 and m_2 be two motifs and V_{m_1} and V_{m_2} be their associated sets of protein. We define the similarity between m_1 and m_2 as

$$s(m_1, m_2) = \frac{|V_{m_1} \cap V_{m_2}|}{min(|V_{m_1}|, |V_{m_2}|)}$$

and turn this into a distance measure between using $d(m_1, m_2) = 1/s(m_1, m_2) - 1$. A hierarchical clustering tree is then constructed using the average linkage algorithm [204] (using the "cluster" R package [148]) with that distance measure. The resulting tree was displayed using the A2R R package: (http://addictedtor.free.fr/packages/A2R/). A cut in the tree is performed to identify a reasonable number of motif families. The motif that obtained the best clustering p-value among the members of its family is selected as the representative member of the family.

5.4.10 Implementation and availability

The proposed computational tools are implemented in a platform-independent Java program called LESMoN. LESMoN is available for download at:
`http://www.cs.mcgill.ca/~blanchem/LESMoN`.

5.5 Results

LESMoN is an approach that identifies short sequence motifs that occur in a set of sequences distributed non-randomly with respect to a given biological network. Specifically, LESMoN takes as input an undirected biological network $G = (V, E)$, where each node $v \in V$ is assigned to a sequence. In this paper, LESMoN is applied to a PPI network and the sequences associated to proteins are the 5' UTRs of the genes encoding them. However, other networks and types of sequences could also be considered (see Discussion). LESMoN enumerates all sequence motifs of a given

length (e.g. 5' UTR motifs of length 8) to identify those that occur in the sequences associated to a set of proteins that appears unexpectedly clustered in the network. Specifically, let $V_m \subseteq V$ be the subset of nodes whose associated sequences contain a match to m. The level of clustering of V_m in G is measured using the total pairwise distance $TPD(V_m)$, defined as the sum of the shortest path distances in G between all pairs of nodes in V_m. The key methodological challenge solved by LESMoN is the estimation of the statistical significance of $TPD(V_m)$: i.e. under a null model where matches to m are distributed randomly in the entire set of sequences considered, with what probability would the TPD of the matching nodes be smaller or equal to $TPD(V_m)$ (see Methods). An unbiased estimator of this probability can be obtained using Monte Carlo sampling, but the running time is prohibitive for very small p-values and large $|V_m|$. We have previously shown that when $|V_m|$ is sufficiently large, these p-values can be estimated based on a normal distribution, in time $O(|V|^4)$. Because this remains prohibitive for the size of the network considered and the number of different protein set cardinalities to be evaluated, we use a Monte Carlo sampling approach to estimate the mean and variance of this normal distribution, rather than calculating them analytically. As shown in Supplementary Figure 5–3, this results in only a minimal loss of accuracy.

5.5.1 Clustering significance of 5' UTR motifs

We used LESMoN to identify locally enriched motifs in the 5' UTR sequences of human genes encoding proteins whose interactions form the network G that structures the gene space. The PPI network, obtained from the BioGRID database [30, 210], contains 14,113 proteins and 127,433 unique pairwise interactions identified using various technologies and experimental protocols.

A set of $3,363,621$ mRNA motifs of length 8 were evaluated for clustering in G. Figure 5–1 shows the number of motifs identified, at various p-value thresholds. As a control and to estimate our false discovery rates, we locally permuted the nucleotides

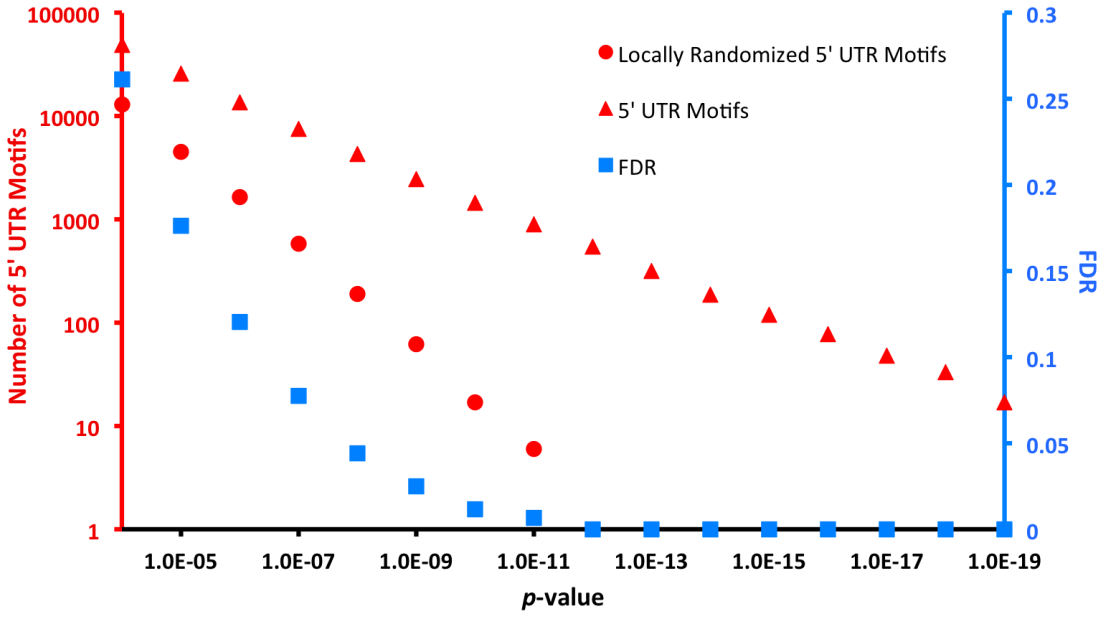

Figure 5–1: Number of motifs originating from both actual and locally randomized 5' UTR sequences (in red) and FDR for a given clustering p-value threshold (in blue, on the secondary axis).

of each 5' UTR sequence (see Section 5.4.5). A clustering p-value of 10^{-8} yields 4277 motifs, at a FDR smaller than 5%. We selected this set of motifs for further analyses. We note however that 545 motifs obtain a p-value below 10^{-12}, which corresponds to a FDR (< 0.0018).

Although this does not affect the correctness of the FDR estimates, we note that the distribution of p-values obtained in the locally randomized sequences is not quite uniform (Supplementary Figure 5–4). For example, in these locally randomized sequences, we found 1642 motifs with p-values below 10^{-6}, whereas only ~ 3 would have been expected. This appears to be due to differences in sequence compositions (most likely GC content) between the 5' UTRs of genes encoding proteins in different portions of the network. Indeed, when the procedure is repeated on 5' UTR sequences that are completely randomized (equal occurrence probability of each nucleotide), the distribution of p-values is very close to uniform (Supplementary Figure 5–4).

To reduce the redundancy in the set of 4277 motifs identified by LESMoN, we used hierarchical clustering based on the similarity of the sets of proteins for which the 5' UTR sequences contain the motifs (Section 5.4.9). Although there was no clear choice of the number of clusters to be obtained, we selected a similarity threshold that resulted in the identification of 269 motif families, ranging in size from 273 motifs to a single one (Figure 5–2). For each family, the motif with the best PPI clustering p-value was retained as representative.

5.5.2 LESMoN identifies evolutionarily conserved 5' UTR motifs

Interspecies sequence conservation is generally evidence of function [173, 81, 110] and functional portions of 5' UTRs have been mapped based on this principle [133, 118, 134]. To assess the biological relevance of each of the motifs identified, we determined the fraction of matching sequences that overlapped regions that are highly conserved within placental mammals (PhastCons elements [201]) and compared it to the overall fraction of 5' UTR bases that are highly conserved (27%; see Section 5.4.7). More than 83% of our 4277 motifs had a highly significant overlap with PhastCons elements (p-value < 0.0001). For 470 motifs, more than 50% of matching sites overlapped these elements, suggesting a very strong selective pressure. Notably, the motifs with the best clustering p-values are often those with the most conserved sites (Pearson's correlation coefficient of 0.43 between the two sets of log p-values of the 4277 motifs and of 0.58 for the 269 motif family representatives; see Supplementary Figure 5–5), suggesting that frequently, the more the proteins associated to a motif are clustered in the network, the more their associated 5' UTR motif is evolutionary conserved and therefore likely to be biologically functional.

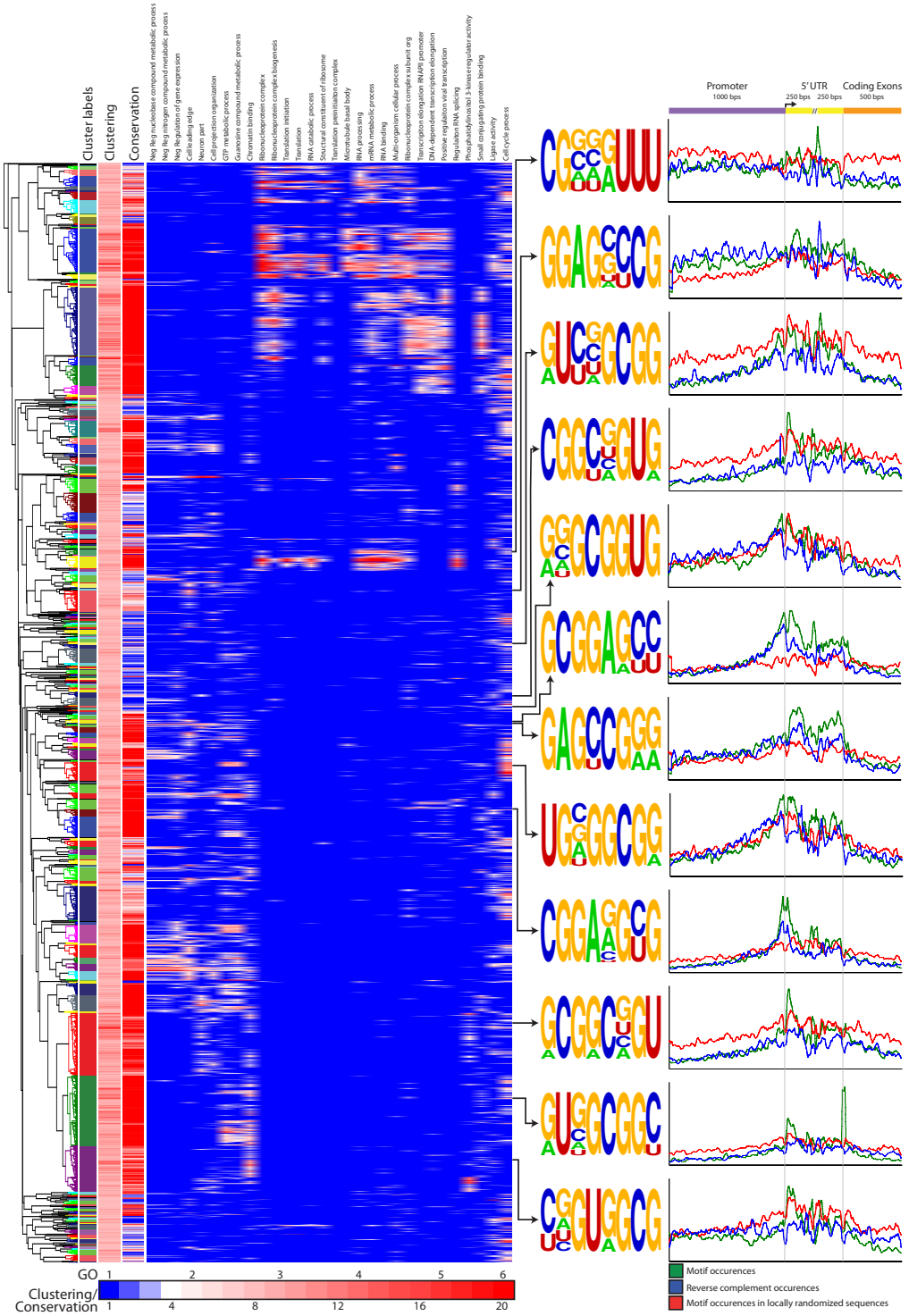

Figure 5–2: LESMoN identified 4277 motifs represented in a hierarchical clustering tree with clustering p-values $< 10^{-8}$ in the BioGRID network. Clustering, conservation, and GO enrichment p-values for each motif are colour-coded. GO enrichment p-values were computed with Ontologizer [14] using a Fisher's exact test. The 30 GO terms shown here are those that are significantly (p-value $< 10^{-6}$) associated to the most motifs, considering only terms that include ≤ 1000 human genes. 12 family representative motifs chosen based on the criteria described in Section 5.4.8 are shown as sequence logos (generated by Weblogo [44]), where nucleotide heights are proportional to their frequencies in 5' UTRs. For these 12 motifs, the motif and its reverse complement occurrences in promoters, 5' UTRs and coding exons in actual and locally randomized sequences are shown.

5.5.3 Proteins associated to the 5' UTR motifs detected by LESMoN are often enriched for specific biological functions

To further investigate the biological significance of each motif identified by LESMoN, we asked if GO terms were surprisingly enriched in the set of associated proteins. Figure 5–2 shows a subset of the GO terms that were found to be enriched in the set of genes associated to the 4277 motifs identified (see Supplementary material for a full list of results). Slightly more than half of the motifs selected in Figure 5–2 are associated to proteins enriched for at least one GO term (corrected enrichment p-value < 0.001). As expected, motifs within the same family are generally enriched for the same GO terms. Again, motifs with the strongest clustering in the PPI network are often those with strongest enrichment for GO terms. Many of these results are discussed in more details below.

5.5.4 5' UTR motifs involved in transcriptional and post-transcriptional regulation

We next attempted to separate motifs that may be involved in transcriptional regulation from those involved in post-transcriptional regulation. Indeed, even though motifs found by LESMoN are present in 5' UTRs, their primary function may still be as transcriptional regulators at the DNA level. We posit that motifs that possess high strand specificity (i.e. many more occurrences than their reverse complement) and whose density is higher in 5' UTRs than in flanking promoters (see Figure 5–2) are more likely to be involved in post-transcriptional regulation. Figure 5–2 and Table 5–1 show 12 such motifs, while motifs with a more likely involvement at the transcriptional regulation level are listed in Table 5–2. While the post-transcriptional implication of a few motifs shown in Figure 5–2 might be difficult to judge from their occurrence profiles, their strand specificity ratios provide a better evidence (Table 5–1). The 12 putative post-transcriptional motifs show positional enrichment, often very near the start of the 5' UTRs, but sometimes near its end, together with strong strand specificity.

5.5.5 Conserved motifs with potential post-transcriptional roles are significantly associated to multiple GO terms

Each of the 12 motifs that are conserved and occur more frequently in 5' UTRs than in promoters and coding exon sequences show enrichments for specific GO terms for their associated proteins (Table 5–1). It is worth noting that for the majority of the GO terms reported the fraction of the occurrences of the motif that overlap highly conserved regions within placental mammals (PhastCons elements [201]) improves when only the occurrences of the motif in the 5' UTRs of the gene annotated with the GO term are considered (Table 5–1). This represents another evidence of the potential functionality of the motifs discovered by LESMoN. Complete results for all 12 motifs of all GO term enrichments computed by Ontologizer are provided in the Supplementary material.

The motif CGGANGYG is enriched in the first 36 bases of 5' UTRs (Figure 5–2). Genes whose 5' UTRs contain this motif are enriched, among others, for GO terms related to "Ubiquitin-ubiquitin ligase". The genes annotated with this GO term share a more specific and longer version of this motif in their 5' UTRs: CGGARGUGR. All occurrences of this motif in the 5' UTRs of these genes are significantly conserved among placentals. Interestingly, CGGANGYG is very similar to the motifs SCGGAAGY and VCCGGAAGNGCR (where S = [GC] and V = [ACG]) that are bound respectively by ELK1 and GABPA [140], which are well characterized transcription factors of the ETS family [200, 197]. This may argue against a post-transcriptional role of this motif, for which the 5' UTR enrichment of the motif versus that in the promoter was moderate, but the strand specificity of the motif (1.41) was much higher than 1. The motif is also enriched for proteins annotated with GO terms related to the transcription elongation process. It is worth mentioning that among the 12 selected motifs figures a variation of the one presented here: CGGYNGUR. The main difference between the two resides at the fourth position where a Y replaces the A. This motif also appears to be linked to transcription, but in a different fashion, since

Table 5–1: Result summary for a set of 5' UTR motifs with putative biological interest† and potential post-transcriptional involvement.

5' UTR motif	Clustering p-value	Conservation p-value	Conservation ratio‡	Strand specificity	GO enrichment summary	Enrichment p-value	GO conservation ratio‡
CGGANGYG	$3.5 \cdot 10^{-17}$	$7.4 \cdot 10^{-115}$	0.54	1.41	Transcription elongation	$1.0 \cdot 10^{-4}$	0.67
					Ubiquitin-ubiquitin ligase	$7.9 \cdot 10^{-4}$	1.00
					Preassembly of GPI anchor	0.001	0.57
RUNGCGGY	$4.3 \cdot 10^{-19}$	$6.5 \cdot 10^{-57}$	0.45	1.34	Serine/threonine phosphatase inhibitor	0.001	0.33
					Negative regulation insulin secretion	$2.8 \cdot 10^{-4}$	0.38
YNGURGCG	$6.4 \cdot 10^{-19}$	$4.9 \cdot 10^{-16}$	0.37	1.42	K48-linked ubiquitination	$6.1 \cdot 10^{-5}$	0.63
					DNA repair	$2.8 \cdot 10^{-4}$	0.41
					RNA polymerase complex	$1.3 \cdot 10^{-4}$	0.40
					MCM complex	$2.3 \cdot 10^{-4}$	0.21
GAGYCGRR	$9.2 \cdot 10^{-10}$	$4.2 \cdot 10^{-15}$	0.35	1.46	Wnt receptor signaling pathway	$9.3 \cdot 10^{-6}$	0.49
					Heat shock protein binding	$9.8 \cdot 10^{-5}$	0.65
					Chromatin disassembly	$4.8 \cdot 10^{-4}$	0.77
RCGRCNGU	$1.0 \cdot 10^{-19}$	$6.3 \cdot 10^{-16}$	0.37	1.48	Cell cycle	$1.6 \cdot 10^{-4}$	0.40
					Kinetochore	$6.3 \cdot 10^{-4}$	0.50
CGGYNGUR	$6.7 \cdot 10^{-17}$	$1.4 \cdot 10^{-4}$	0.31	1.65	Transcription regulation response to oxidative stress	0.001	0.80
GCGGARYY	$1.1 \cdot 10^{-13}$	$4.5 \cdot 10^{-28}$	0.40	1.51	Proton-transporting two-sector ATPase complex	0.001	0.38
					Lung morphogenesis	$1.3 \cdot 10^{-4}$	0.54
					Response topologically incorrect protein	$6.6 \cdot 10^{-4}$	0.54
GGAGNYCG	$2.5 \cdot 10^{-9}$	$3.2 \cdot 10^{-15}$	0.36	1.26	Cell leading edge	$5.3 \cdot 10^{-6}$	0.45
					Sexual reproduction	$5.9 \cdot 10^{-4}$	0.50
					Actin filament	$9.4 \cdot 10^{-4}$	0.35
					Hormone-mediated signaling pathway	$9.6 \cdot 10^{-4}$	0.45
					Cell projection organization	0.002	0.51
					Locomotory behavior	0.001	0.36
UGNGGCGR	$3.0 \cdot 10^{-17}$	$1.2 \cdot 10^{-27}$	0.40	1.38	Catalytic step 2 spliceosome	0.003	0.64
					Cell cycle	$7.0 \cdot 10^{-7}$	0.43
CGNNRUUU	$3.2 \cdot 10^{-10}$	$2.7 \cdot 10^{-10}$	0.36	1.52	Cell division	$3.8 \cdot 10^{-4}$	0.36
					Chromosome segregation	0.001	0.47
					Spindle	0.001	0.32
					Microtubule cytoskeleton	$1.1 \cdot 10^{-4}$	0.36
RUYNGCGG	$1.8 \cdot 10^{-14}$	$4.5 \cdot 10^{-5}$	0.32	1.43	Endosome to lysosome transport	$8.4 \cdot 10^{-5}$	0.50
					Vacuolar transport	$1.3 \cdot 10^{-4}$	0.45
RNGCGGUG	$2.2 \cdot 10^{-10}$	$8.1 \cdot 10^{-9}$	0.35	1.34	Iris morphogenesis	$3.4 \cdot 10^{-4}$	0.60
					Activation innate immune response	$6.8 \cdot 10^{-4}$	0.50
					Regulation of interleukin-8 biosynthetic	$5.0 \cdot 10^{-4}$	0.00

† GO terms in the table were chosen based on their enrichment significance and level of biological interest.
‡ Conservation ratios are the fraction of the occurrences of the motif that overlap with regions that are highly conserved within placental mammals (PhastCons elements [201]) over the total number of occurrences of the motif. GO conservation ratios represent the same fraction but only for the occurrences of the motifs in the 5' UTRs of the genes annotated by the GO term specified on the same line in the table.

the proteins associated to it are enriched for "regulation of transcription from RNA polymerase II promoter in response to oxidative stress" (p-value = 0.001).

The motif YNGURGCG is also enriched at the beginnings of 5' UTRs. The proteins associated to that motif are overrepresented among others by the "K48-linked ubiquitination" and "DNA repair" GO terms. This motif resembles the binding motifs of the poorly characterized RBP ZC3H10 and of the RNA biding factor RBM4 according to the CISBP-RNA Database (in press). RBM4 is known to be implicated in the alternative splicing of 5' UTRs and in the exon selection of reporter pre-mRNAs [141, 123]. This strengthens our belief of the biological functionality of the YNGURGCG motif in 5' UTRs. Among the proteins associated to the CGNNRUUU motif, a surprisingly large number of proteins are linked to the cell cycle and cell division. Unlike the previously presented motifs, CGNNRUUU occurs more often in the second half of the 5' UTRs than at the beginning. It shares similarities with the motif KCCGNSWTTT (where K = [GT] and W = [AT]), which was found to be enriched within +/-2000 bases of transcription start sites but not associated to any transcription factors [140].

Another motif, RNGCGGUG shows similarities with a number of RBP binding motifs. Among those figure RBM4 again, but also SAMD4A, a translational repressor [9], and RBM8A, a splicing related factor [205]. The last 4 characters of this motif match exactly a FUS binding motif reported by Cook et al. [40]. FUS is a multi-functional RBP that is involved in (but not exclusively) RNA splicing and genome maintenance [88]. These similarities argue in favour of a post-transcriptional role of the motif.

Among the other motifs presented in Table 5–1, some motifs occur at both ends of 5' UTRs such as GGAGNYCG and GAGYCGRR (Figure 5–2). The latter is particularly interesting since for the proteins associated to it and that are annotated with the "chromatin disassembly" GO term, the motif seems to be extending at the

5' end with a G rich region (Supplementary Figure 5–6). Others are present towards the 5' end of 5' UTRs. These include RUNGCGGY and RCGRCNGU (Figure 5–2). On the other hand, the motif RUYNGCGG seems to occur mainly at the 3' end of 5' UTRs (Figure 5–2). Finally, GCGGARYY and UGNGGCGR appear to occur somewhat uniformly along 5' UTR sequences (Figure 5–2). More details about all of these motifs are provided in the Supplementary material.

5.5.6 Biological significance of motifs with potential transcriptional implications

The 12 motifs presented above were selected among other things because of their likelihood to play a role post-transcriptionally. Nevertheless, LESMoN identified motifs that are well conserved, but show no strand preference (strand specificity ≈ 1.0), which is expected for transcription factor binding sites. Examples of such motifs are reported in Table 5–2. Again it can be observed that for the vast majority of the GO terms described in Table 5–2 there are more occurrences of the motif that are conserved within the genes annotated with the GO term. When inspecting the occurrence profile of RNCGGAAR in promoters and 5' UTRs, it can be seen that it occurs mainly before and after the transcription start site of genes and that the occurrences of its reverse complement follow the same pattern (see Supplementary Figure 5–7). However, the 3' end of this motif is very similar to the RBP eukaryotic translation initiation factor 4B (EIF4B) biding motif "RGAM", where M denotes a A or a C [40]. EIF4B is known to bind the 5' cap of mRNAs to unwind their RNA secondary structure and promote the ribosome binding to perform translation [155]. This argues against a transcriptional role of the motif. Nevertheless, it also resembles the motif SCGGAAGY that is bound by the transcription factor ELK1 [140], this time arguing for a transcriptional role. Another similar motif with analogous properties is NCGRAARY (see Supplementary Figure 5–7). Its associated proteins are overrepresented among other things for protein transport GO terms. Finally, the motif YUUYCGGN putatively appears also to play a role transcriptionally and shows

Table 5–2: Result summary for a set of 5' UTR motifs with putative biological interest† and potential transcriptional involvement.

5' UTR motif	Clustering p-value	Conservation p-value	Conservation ratio‡	Strand specificity	GO enrichment summary	Enrichment p-value	GO conservation ratio‡
RNCGGAAR	$2.7 \cdot 10^{-21}$	$3.6 \cdot 10^{-128}$	0.57	0.97	Establishment protein localization	$5.4 \cdot 10^{-7}$	0.67
					Cellular response to stress	$6.3 \cdot 10^{-10}$	0.62
					Pos regulation viral transcription	$1.4 \cdot 10^{-6}$	0.72
YUUYCGGN	$1.4 \cdot 10^{-21}$	$3.7 \cdot 10^{-90}$	0.51	1.14	RNA polymerase complex	$1.5 \cdot 10^{-5}$	0.42
					Structural constituent ribosome	$1.5 \cdot 10^{-4}$	0.75
					Translation	$1.4 \cdot 10^{-5}$	0.58
					Spliceosomal complex	$2.1 \cdot 10^{-4}$	0.60
NCGRAARY	$1.9 \cdot 10^{-21}$	$1.0 \cdot 10^{-159}$	0.56	0.96	Establishment of protein localization	$1.5 \cdot 10^{-5}$	0.69
					Protein transport	$2.0 \cdot 10^{-5}$	0.67
					Golgi vesicle transport	$2.1 \cdot 10^{-4}$	0.76

† GO terms in the table were chosen based on their enrichment significance and level of biological interest.
‡ Conservation ratios are the fraction of the occurrences of the motif that overlap with regions that are highly conserved within placental mammals (PhastCons elements [201]) over the total number of occurrences of the motif. GO conservation ratios represent the same fraction but only for the occurrences of the motifs in the 5' UTRs of the genes annotated by the GO term specified on the same line in the table.

many significant enrichments of GO terms (see Supplementary Figure 5–7 and Table 5–2). Table 5–2 presents the complete computational results for these three motifs and GO enrichments complete results are provided in the Supplementary material.

5.6 Discussion and conclusion

Several computational approaches have been developed to identify sequence motifs with biological implications. Some, like the one presented by Xie et al. [228] rely on motif recurrence and evolutionary conservation. These techniques may be capable of identifying some of the motifs found by LESMoN. It is however much more likely that an important fraction of the motifs identified by LESMoN could not be discovered by such approaches. Indeed, a motif occurring with a high frequency but with only a few evolutionary conserved instances is most likely going to be ignored these approaches. However, LESMoN will report this motif if these instances are in the 5' UTRs of the RNAs of proteins forming a protein complex.

5.6.1 Limitations

Even though our method seems sensitive enough to detect numerous 5' UTR sequence motifs of potential biological interest, it could be improved. In the present

state, the approach only explores motifs constituted with 8 characters. With such length, random occurrences in 5' UTRs of motifs are fairly likely, making the set of proteins associated to it noisy. By increasing this length, longer 5' UTR motifs with a biological role may emerge from this noise, by being associated to a smaller set of proteins that possesses a clustering of greater significance in the network. The main drawback of such modification and the reason why we opted not to follow this direction is the amount of computational time required to test for the clustering of the important number of motifs with length larger than 8. Nevertheless, we are currently exploring approaches to accelerate this process such as methods evaluating the clustering of the motifs of size larger than 8 containing motifs of size 8 or subsequences thereof already known to be significantly clustered in the network.

Another aspect that could improve the sensitivity of LESMoN resides in the use of degenerate characters such as K, W, and S, which were neglected in this article. This could once again improve our sensitivity as some biologically functional motifs whose associated proteins are clustered in the network may be mainly composed of these omitted degenerate characters. Nevertheless, increasing the alphabet size of our motifs also increases the running time of LESMoN.

The discovery potential of our approach is largely affected by the quality and coverage of the PPI network analyzed. The approach is more likely to produce a significant number of discoveries when applied to large networks composed of PPIs of high quality. If the network is too small, few significant clusterings will be present. In addition, if the network contains a large number of noisy interactions, the clusterings identified by LESMoN may have little biological significance. Finally, PPI networks often contain hub proteins, which have sometimes hundreds if not thousands of protein interactions. These cause the majority of pairwise distances of proteins in the network to be small. When all pairwise distances in a network are relatively small, only very tight protein clusterings in such network will be judged significant by LESMoN, while slightly clustered proteins will not be considered as statistically significant. This

could therefore potentially limit our ability to detect 5' UTR motifs regulating the localization in the cells of their corresponding mRNAs, since the proteins associated to these motifs are likely to only be slightly clustered in the network.

While a more comprehensive PPI network such as the one extracted from the iRefWeb database [217] could have been used with LESMoN, we chose to performed our analysis on the curated BioGRID PPI network [210]. We opted for such network since its quality may be greater than some of the automated databases included in the iRefWeb network. Furthermore, at this time, the BioGRID network consisted in the major part of the iRefWeb network and therefore no great gain are expected from the use of iRefWeb. We do however acknowledge that the use of iRefIndex (i.e. the number of publications supporting an interaction) allows the extraction of a dataset of high quality from iRefWeb.

5.6.2 Extensions

RNA molecules are known to form various secondary structures in order to perform their functions, which often consist in binding proteins or other RNAs. In this article, we opted to only consider the primary structure of 5' UTRs, but our approach could be extended to study RNA secondary structure motifs. More precisely, LESMoN could evaluate the clusterings of all proteins in a PPI network for which their respective 5' UTRs contains a given secondary structure motif, such as a 6 nucleotide hairpin loop or a bulge of 2 nucleotides. This approach could be beneficial since RNA sequences may differ but still form similar RNA secondary structures, which are, for example, necessary for their binding to a regulatory protein.

Our method could also be extended to perform protein function prediction. Often, several proteins among those found by LESMoN to be clustered and associated to the same 5' UTR motif are uncharacterized. LESMoN provides crucial pieces of

information to infer the function of these uncharacterized proteins and brings an additional dimension to the "guilt by association" approach for protein function prediction [166, 196, 219, 51]. A strategy could be implemented to compute likelihoods for such uncharacterized proteins to perform a certain function based on their co-clusterings with already functionally annotated proteins and the occurrence of a given 5' UTR motif.

This manuscript only explored one of the many applications of LESMoN. Obviously, besides 5' UTRs, 3' UTRs could also be analyzed in the same fashion. In addition, we could, in the future, analyze different types of sequences, such as coding exons, promoters, and amino acid sequences. The latter could be interesting especially for the discovery of transcription factor binding sites regulating the transcription of proteins interacting in the cell.

Finally, another interesting extension of LESMoN is to study clusterings of sequence motifs in biological networks other than PPI networks (e.g. co-methylation, co-expression, gene regulation networks, etc.). For instance, our method could be applied to gene co-expression networks, where nodes represent genes and edges link genes whose expressions are correlated. LESMoN could be used to identify various types of sequence motifs (in promoters, UTRs, coding exons, etc.) associated to these genes that are clustered in the network. One could expect that for example, binding sites of transcription factors regulating the expression of a set of genes present (and therefore clustered) in the network could be identified using this approach. It could also be used to find DNA sequence motifs that are clustered in co-methylation networks, where nodes correspond to methylation sites and edges connect sites whose methylation patterns are correlated. A node could be annotated with a given motif if the motif occurs within a certain window size around the methylation site. In this context, applying LESMoN to such network could allow the discovery of DNA motifs that are linked directly or not to the mechanisms causing the associated sites to be clustered in the co-methylation network.

5.7 Acknowledgements

We are grateful to the members of our laboratories for helpful discussions and comments. This work is supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to M.B and by grants from the Canadian Institutes of Health Research (CIHR) and the Fonds de recherche du Québec-Santé (FRQS). M.L.A. held a Vanier studentship from NSERC.

5.8 Supplementary material

Complete result tables and all GO enrichment results for the 4277 motifs with clustering p-values $< 10^{-8}$ can be downloaded at:
`http://www.cs.mcgill.ca/~blanchem/LESMoN`.

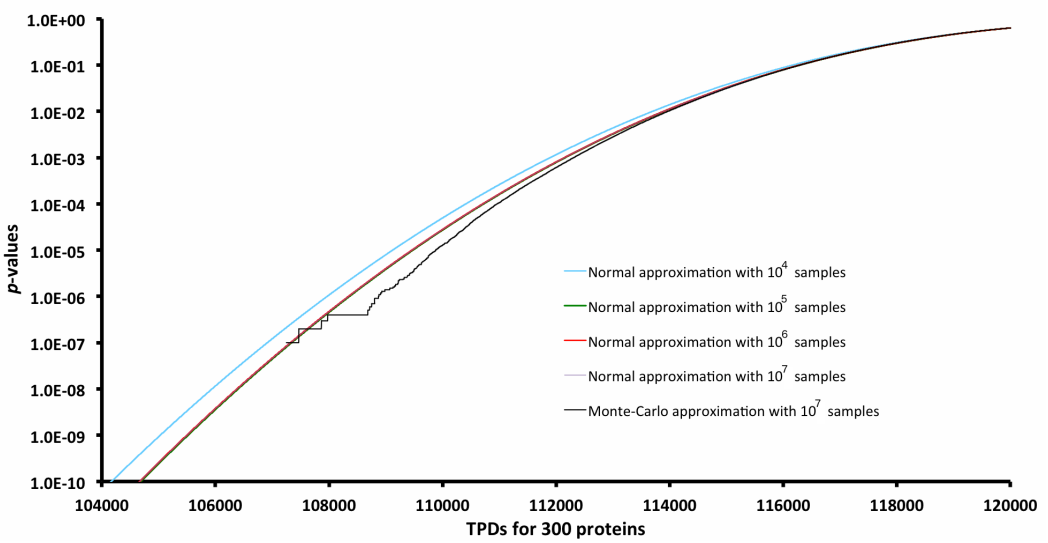

Figure 5–3: p-values of TPDs for 300 proteins using the normal distribution approximation with 10^4, 10^5, 10^6 and 10^7 samples and the Monte Carlo approach with 10^7 samples. Of note, curves of the normal distribution approximation with 10^5, 10^6, and 10^7 are hard to distinguish because they are heavily overlapping.

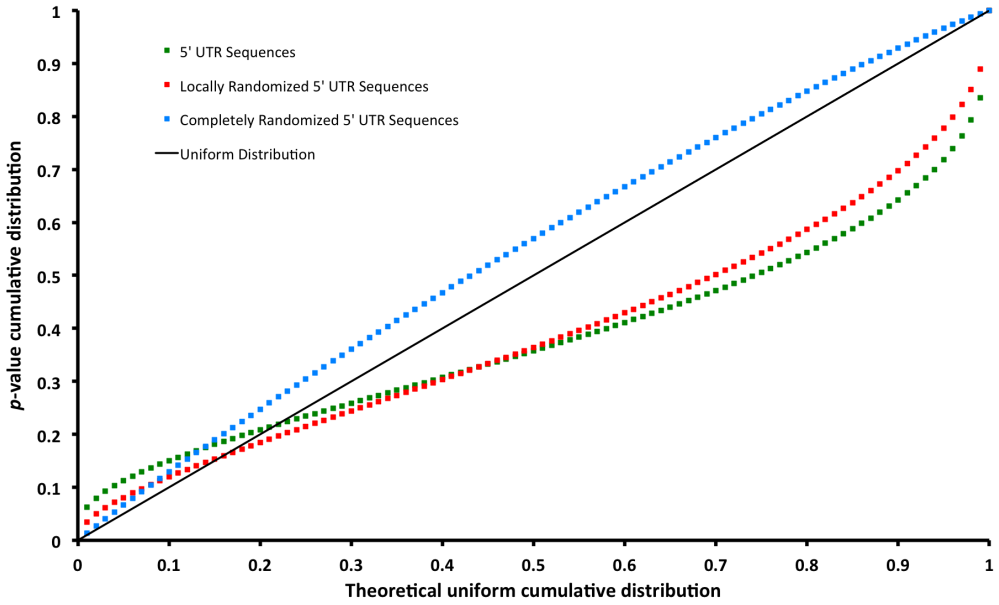

Figure 5–4: Cumulative distributions of clustering p-values computed by LESMoN for motifs obtained from completely randomized, locally randomized, and unmodified 5' UTR sequences compared to a theoretical uniform distribution.

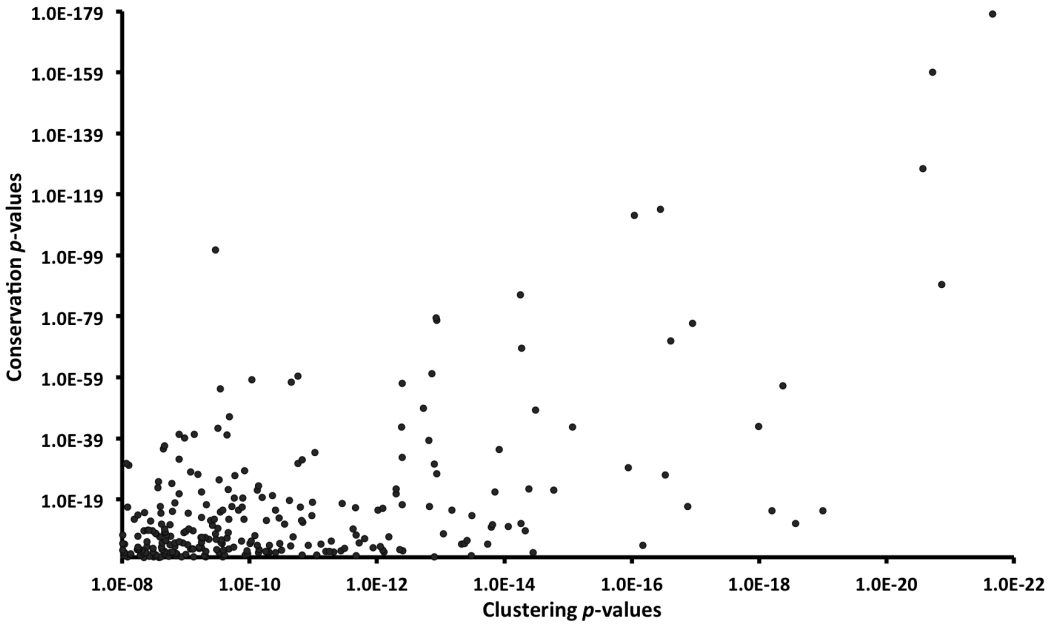

Figure 5–5: Evolutionary conservation p-value and clustering p-value of each of the 269 motif family representatives.

Figure 5–6: Extended sequence logo of the 5' UTR motif of proteins involved in chromatin disassembly, which were associated by LESMoN to the GAGCCGRR motif. Information content is plotted as a function of nucleotide position. The sequence logo was generated using Weblogo [44].

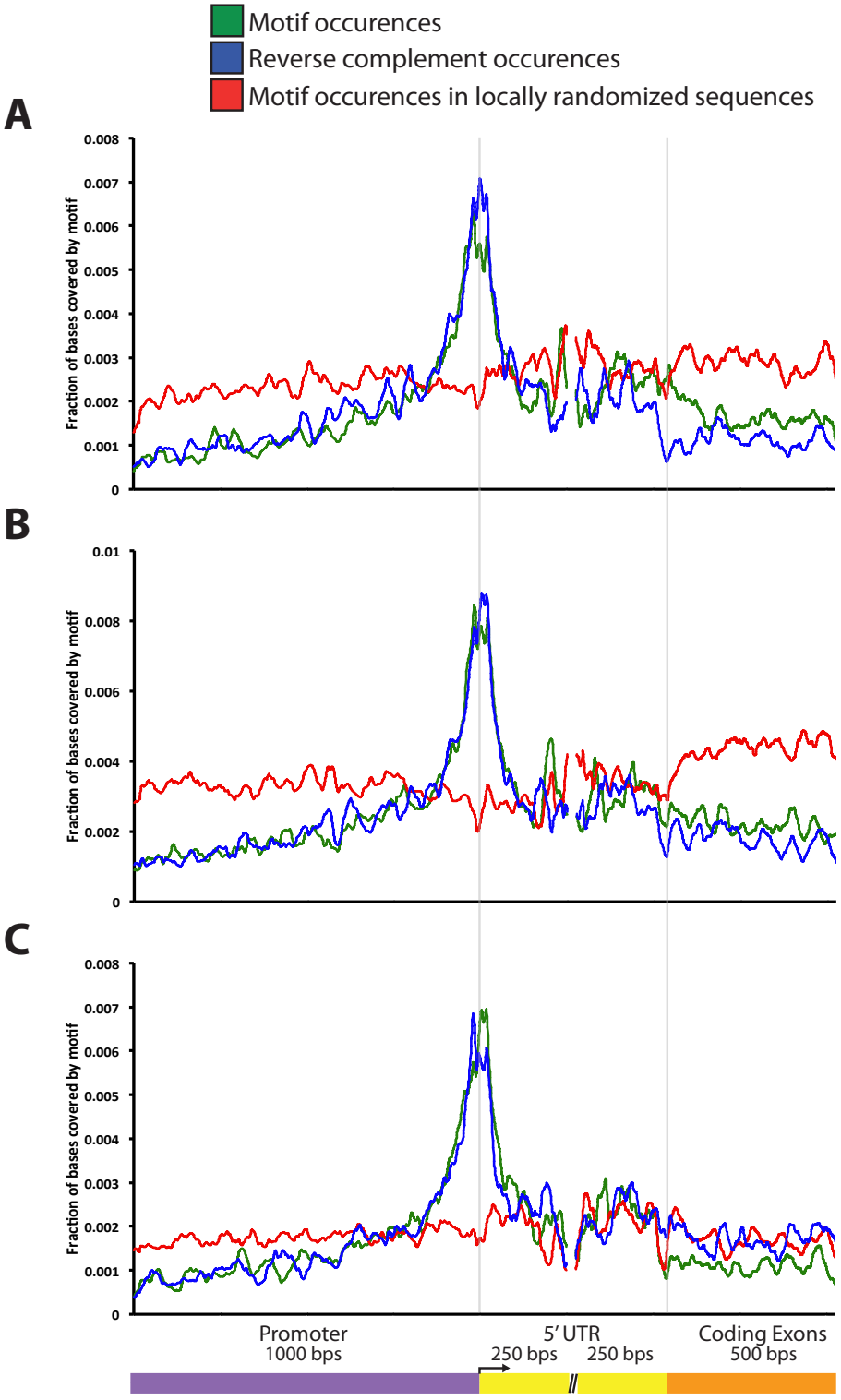

Figure 5–7: Fraction of bases at each position in promoters, 5' UTRs and coding exons covered by the motifs (A) RNCGGAAR, (B) NCGRAARY, and (C) YUUYCGGN.

CHAPTER 6
Conclusion

6.1 Contributions

The study of PPIs is critical to the understanding of the various mechanisms defining the behaviour of the cell. However, the poor specificity of the high-throughput experiments used to build PPI datasets has proved to be a real challenge. A stringent filtering is required to extract from these datasets the biologically relevant PPIs. Besides the difficulty of obtaining high-quality PPIs, one objective remains: the development of experimental pipelines allowing for a more sensitive detection of PPIs for a given protein. Moreover, with the accumulation of high-throughput PPI studies, PPI networks of organisms such as human and yeast have expanded to the point where sophisticated computational approaches are necessary to draw forth relevant biological information. This thesis is comprised of four significant contributions addressing these issues that hinder the analysis of PPI data.

AP-MS is among the most popular approaches to identify PPIs. It however produces datasets that are largely composed of false positives [20, 48, 34]. Chapter 2 presented a Bayesian approach (Decontaminator) to model contaminants in PPI data produced by AP-MS. The approach uses a limited number of AP-MS controls to assess the likelihood of a PPI to be originating from a contamination event. At the time of publication, this method was the only one that utilized a Bayesian inference to detect contaminants in AP-MS datasets when provided a small number of controls. Whereas other methods tended to confuse proteins co-purified in many AP-MS experiments (hub proteins) with contaminants because of their multiple interacting partners if not provided a list of user-defined list of hub proteins, ours does not. This represented a clear advantage over the other approaches published at the time especially when

performing AP-MS on uncharacterized proteins. A similar approach is now used in SAINT 2.0 [34], which uses spectral counts to assess the quality of PPIs.

Since we demonstrated our capability to filter out contaminants from AP-MS datasets using Decontaminator, we explored ways to improve the sensitivity of AP-MS experiments. A large number of interactions that are known to occur in the cell are very difficult to recover using a classic AP-MS protocol on whole cell extracts. One reason explaining this lack of coverage consists in the fact that typically, not all cell compartments are analyzed in AP-MS experiments. For example, PPIs occurring on the chromatin or the cell membrane often remain undetected with this method [124, 6]. Chapter 3 introduced a novel method to identify and computationally assess the quality of PPIs of a given protein independently in three different cell compartments (cytoplasm, nucleoplasm, and chromatin). Our approach produces a dataset of cell compartment specific PPIs. It provides significant coverage improvements by accessing the chromatin, but also by fractionating the samples analyzed into three fractions to maximize peptide detection at the mass spectrometry level. Furthermore, the computational approach associated to this pipeline (based on the methods presented in Chapter 2) allows for a greater flexibility in terms of the number of AP-MS experiments and controls needed for the training of the Bayesian procedure. In addition, it permits, if needed, the use of controls obtained in experimental conditions other than the one for which it assesses PPIs. This is particularly useful when there are very few controls available for a given experimental condition. Such approach was also recently taken in a publication by Mellacheruvu et al. [153].

High quality PPIs obtained using tools such as Decontaminator are typically analyzed and visualized as a network. PPI networks are often very large, extremely complex, and hard to investigate by visual inspection. Numerous strategies have been implemented to help the visualization of these networks [203, 91, 24]. Chapter 4 proposed a novel approach to efficiently identify subgraphs in a PPI network that are

overrepresented with a certain GO annotation. It introduced a distance and a similarity measure evaluating the clustering of a set of proteins in a PPI network. It also presented four computational strategies to assess the statistical significance of such protein clusterings. These four methods include a normal distribution approximation as well as three different approaches using convolution of probability distribution, each with its benefits and drawbacks. The results of our program (GoNet) helped the organization and visualization of yeast and human PPI networks. GoNet is, to our knowledge, the only available computational approach that identifies annotations that are significantly clustered in protein-protein interaction networks.

Nucleotide sequence motif discovery is a classic problem of computational biology. A common version of the problem is to find sequence motifs that are significantly overrepresented in a set of nucleotide sequences. In Chapter 5, we proposed to tackle a modified version of this problem with the help of PPI networks. Our method identifies 5' UTR sequence motifs for which the associated proteins are significantly clustered in a given PPI network. Our approach (LESMoN) is inspired by GoNet. It uses sequence motifs rather than Gene Ontology to define gene sets. Computationally, it improves on GoNet by allowing for the statistical assessment of clusterings of large protein sets in PPI networks. Chapter 5 also presented a set of statistical tests to evaluate the biological relevance of the 5' UTR motifs highlighted by LESMoN. Our approach discovered several previously uncharacterized 5' UTR motifs and associated them to biological processes taking place in PPI networks. We hope that this novel approach for sequence motif discovery will set a new basis for the analysis of not only 5' UTR sequences, but also 3' UTR, promoter, and coding sequences.

These four contributions have different applications: assessing the quality of PPIs obtained from AP-MS, increasing the sensitivity of the AP-MS protocol, highlighting regions of PPI networks that are of potential biological interest, and discovering sequence motifs and associating them with biological processes represented in such networks. However, they share the same goal, which is the deconvolution of large

and noisy PPI datasets. As discussed previously, each of these chapters contributed significantly to the field by providing novel algorithms yielding results with important biological implications. Networks formed by PPI datasets are commonly compared to hairballs because of their noisy and intricate nature [77]. The techniques presented here are one step forward to untangle these balls.

6.2 Perspectives on future work

The technologies leading to the discovery of PPIs have evolved at a fast pace in the last decade. This period saw the appearance of sophisticated computational tools, more sensitive and specific purification methods, specialized sample preparation techniques, and mass spectrometers with high resolution and speed. The last two aspects of this list are of particular interest because they might be the ones, in combination with the development of appropriate computational tools, where the most progress can be accomplished in the future for the mapping and deconvolution of PPI datasets.

6.2.1 Deconvolution of PPI networks based on time

As it was discussed in Chapter 3, PPI datasets produced by most studies represent the union of the interactions occurring in the different cell compartments of the cell. However, they also contain the union of all PPIs in a population of cells, which are at different phases of the cell cycle. PPI datasets are therefore constituted of the union of PPIs over the entire cell cycle. More precisely, if a protein interacts with a set of proteins in G1 phase, then it is quite possible that this protein will interact with disjoint sets of proteins in S and G2 phase. However, the union of these interactions would be reported for that protein without distinction of the phase in which each interaction happens. This inclusion of a hidden time component causes PPI datasets to be very large and complicated. A decomposition of networks over the different phases of the cell cycle would be desirable. One could imagine that for

a given organism, a PPI network could be built for each phase of the cell cycle. Still today, to our knowledge, and as reported by Levy et al. [137], a large-scale approach mapping PPIs independently in different cell cycle phases has not been presented. In order to simplify and understand PPI networks, several groups have combined them with time series of mRNA expression data [50, 85, 29]. The basic idea is that co-expressed proteins sharing the same interaction partners are likely to interact at the same time (cell cycle phase) with those partners. It is however widely accepted that the correlation of mRNA and protein expression varies significantly [83] since several biological mechanisms may alter the translation of mRNAs into functional proteins. Integrative algorithms using time series of mRNA expression and protein quantification data from mass spectrometry might yield better results and are an alternative that I believe should be explored in the future.

6.2.2 PPI quantification

Chapter 2 and 3 introduced the notion of modeling the abundance of contaminants in PPI datasets. Their modeling procedures used Mascot scores as proxies to measure the abundances of prey proteins. Even if Mascot scores or spectral counts (a correlated scoring method) are sufficient to accurately estimate protein abundances, they still suffer from a very important issue. The Mascot score of a peptide depends on the abundance of all peptides analyzed at the same period (elution time) in the mass spectrometer. Peptides with low abundance are often masked out when analyzed at the same time as much more abundant peptides. Improvements to the peptide detection dynamic range of mass spectrometers have been made in the last years, but not to the point where this issue is resolved. The dynamic range of protein abundances in certain protein mixtures can be much greater than that of mass spectrometers. Nevertheless, very recently, improvements to the spectral acquisition speed and resolution of mass spectrometers made data independent acquisition of peptide spectra more accessible. This strategy, instead of randomly selecting peptide ions in the ion population for the analysis, allows to analyze all peptide ions

within a given mass window [75, 216]. This removes the sampling bias towards highly abundant peptides that exists in classic mass spectrometry approaches. Such data independent approaches are promising and show the potential to provide unbiased Mascot scores and spectral counts yielding better protein abundance estimates. Data independent approaches could therefore provide accurate PPI quantification when used in combination with AP-MS. It is clear that Decontaminator would benefit from more accurate protein abundance measurements, which would allow for a more precise modeling of contaminants. The spectra produced by such technique are however much more convoluted than the ones acquired with classic mass spectrometry analyses. The computational approaches that should be used to correctly analyze and associate properly each spectrum to their corresponding peptide remain unclear and leave several interesting open problems.

6.2.3 Protein function inference

Achieving a more accurate PPI quantification can help solving another important computational biology problem: the inference of the function of uncharacterized proteins. Indeed, as it was mentioned previously, the function of an uncharacterized protein is often predicted based on the known functions of its interacting partners [166]. However, with PPI quantification, the contribution of the functions of the interacting partners in the function prediction model, can be weighted by the PPI abundances. Such approaches would be likely to produce more accurate protein function predictions, which often suffer from low specificity. PPI quantification could also be beneficial to the approaches presented in Chapter 4 and 5. We explained previously that GoNet can use as input a weighted PPI network. Such edge weights often represent the confidence that we have that an interaction is a true positive. However, these could also consist in measures of PPI absolute abundances. PPI abundances are likely to help to the identification of protein clusterings. GoNet could therefore evaluate clusterings of proteins sharing the same annotation in abundance weighted PPI

networks. We believe that such network input is likely to help GoNet and LESMoN to produce results with greater biological significance.

Protein mass spectrometry is a young field and analysis of protein-protein interactions remains in its infancy. Even at such young age, the latter has already revolutionized the way we think about biological mechanisms in the cell. It has been evolving at a breathtaking speed and its progression has shown no signs that it will slow down in the near future. It has demonstrated that it possesses a tremendous potential for protein functional inference and biomarker discovery. The technology improvements that are yet to come will only make these discoveries more impressive. I truly believe that the path toward a better understanding of cell biology and human health must inevitably pass through protein-protein interaction studies. The future of this field is full of promises. Several fascinating problems remain to be tackled in order to transform these protein-protein interaction hairballs into complex but well organized and understood knittings.

References

[1] J. Acker, M. de Graaff, I. Cheynel, V. Khazak, C. Kedinger, and M. Vigneron. Interactions between the human RNA polymerase II subunits. *Journal of Biological Chemistry*, 272(27):16815–16821, 1997.

[2] G. Adelmant and J. A. Marto. Protein complexes: the forest and the trees. *Expert Review of Proteomics*, 6(1):5–10, 2009.

[3] F. Al-Shahrour, R. Daz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, Mar 2004.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[5] O. Aygün, J. Svejstrup, and Y. Liu. A RECQ5–RNA polymerase II association identified by targeted proteomic analysis of human chromatin. *Proceedings of the National Academy of Sciences*, 105(25):8580–8584, 2008.

[6] M. Babu, J. Vlasblom, S. Pu, X. Guo, C. Graham, B. D. Bean, H. E. Burston, F. J. Vizeacoumar, J. Snider, S. Phanse, et al. Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, 2012.

[7] G. Bader and C. Hogue. Analyzing yeast protein–protein interaction data obtained from different sources. *Nature biotechnology*, 20(10):991–997, 2002.

[8] G. Bader and C. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2, 2003.

[9] M. V. Baez and G. L. Boccaccio. Mammalian Smaug is a translational repressor that forms cytoplasmic foci similar to stress granules. *Journal of Biological Chemistry*, 280(52):43131–43140, 2005.

[10] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl 2):W202–W208, 2009.

[11] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[12] A. Barabási and Z. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[13] M. Barboric, J. Kohoutek, J. Price, D. Blazek, D. Price, and B. Peterlin. Interplay between 7SK snRNA and oppositely charged regions in HEXIM1 direct the inhibition of P-TEFb. *EMBO J.*, 24(24):4291–303, 2005.

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

[15] T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[16] B. Bernholtz. Type I censoring and the structural approach. *Statistical Papers*, 18(1):2–12, 1977.

[17] S. Boulon, B. Pradet-Balade, C. Verheggen, D. Molle, S. Boireau, M. Georgieva, K. Azzag, M.-C. Robert, Y. Ahmad, H. Neel, et al. HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. *Molecular cell*, 39(6):912–924, 2010.

[18] K. Boutilier, M. Ross, A. Podtelejnikov, C. Orsi, R. Taylor, P. Taylor, and D. Figeys. Comparison of different search engines using validated MS/MS test datasets. *Analytica Chimica Acta*, 534(1):11–20, 2005.

[19] P. Braun. Reproducibility restored-on toward the human interactome. *Nature methods*, 10(4):301–303, 2013.

[20] A. Breitkreutz, H. Choi, J. R. Sharom, L. Boucher, V. Neduva, B. Larsen, Z.-Y. Lin, B.-J. Breitkreutz, C. Stark, G. Liu, et al. A global protein kinase and phosphatase interaction network in yeast. *Science*, 328(5981):1043–1046, 2010.

[21] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[22] S. Brohée, K. Faust, G. Lima-Mendez, G. Vanderstocken, and J. van Helden. Network Analysis Tools: from biological networks to clusters and pathways. *Nat Protoc*, 3(10):1616–1629, 2008.

[23] S. Brohée and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.

[24] K. R. Brown, D. Otasek, M. Ali, M. J. McGuffin, W. Xie, B. Devani, I. L. van Toch, and I. Jurisica. NAViGaTOR: network analysis, visualization and graphing Toronto. *Bioinformatics*, 25(24):3327–3329, 2009.

[25] A. Brückner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6):2763–2788, 2009.

[26] S. Byers, J. Price, J. Cooper, Q. Li, and D. Price. HEXIM2, a HEXIM1-related protein, regulates positive transcription elongation factor b through association with 7SK. *J Biol. Chem.*, 280(16):16360–7, 2005.

[27] C. Carré and R. Shiekhattar. Human GTPases associate with RNA polymerase II to mediate its nuclear import. *Molecular and cellular biology*, 31(19):3953–3962, 2011.

[28] P. Cassonnet, C. Rolloy, G. Neveu, P.-O. Vidalain, T. Chantier, J. Pellet, L. Jones, M. Muller, C. Demeret, G. Gaud, et al. Benchmarking a luciferase complementation assay for detecting protein complexes. *Nature Methods*, 8(12):990–992, 2011.

[29] X. Chang, T. Xu, Y. Li, and K. Wang. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Scientific reports*, 3, 2013.

[30] A. Chatr-aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. ODonnell, et al. The BioGRID interaction database: 2013 update. *Nucleic acids research*, 41(D1):D816–D823, 2013.

[31] X. Chen, T. R. Hughes, and Q. Morris. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics*, 23(13):i72–i79, 2007.

[32] H. Choi, D. Fermin, and A. I. Nesvizhskii. Significance analysis of spectral count data in label-free shotgun proteomics. *Molecular & Cellular Proteomics*, 7(12):2373–2385, 2008.

[33] H. Choi, S. Kim, A.-C. Gingras, and A. I. Nesvizhskii. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Molecular systems biology*, 6(1), 2010.

[34] H. Choi, B. Larsen, Z.-Y. Lin, A. Breitkreutz, D. Mellacheruvu, D. Fermin, Z. S. Qin, M. Tyers, A.-C. Gingras, and A. I. Nesvizhskii. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature methods*, 8(1):70–73, 2010.

[35] H. Choi and A. Nesvizhskii. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of proteome research*, 7(01):254–265, 2007.

[36] H. Chua, W. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623, 2006.

[37] P. Cloutier, R. Al-Khoury, M. Lavallée-Adam, D. Faubert, H. Jiang, C. Poitras, A. Bouchard, D. Forget, M. Blanchette, and B. Coulombe. High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods*, 48(4):381–386, 2009.

[38] P. Cloutier, M. Lavallée-Adam, D. Faubert, M. Blanchette, and B. Coulombe. A newly uncovered group of distantly related lysine methyltransferases preferentially interact with molecular chaperones to regulate their activity. *PLoS genetics*, 9(1):e1003210, 2013.

[39] S. Collins, P. Kemmeren, X. Zhao, J. Greenblatt, F. Spencer, F. Holstege, J. Weissman, and N. Krogan. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 6(3):439, 2007.

[40] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes. RBPDB: a database of RNA-binding specificities. *Nucleic acids research*, 39(suppl 1):D301–D308, 2011.

[41] B. Coulombe, M. Blanchette, and C. Jeronimo. Steps towards a repertoire of comprehensive maps of human protein interaction networks: the Human Proteotheque Initiative (HuPI). *Biochem Cell Biol*, 86(2):149–156, Apr 2008.

[42] B. Cox and A. Emili. Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. *Nature protocols*, 1(4):1872–1878, 2006.

[43] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[44] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.

[45] E. Czeko, M. Seizl, C. Augsberger, T. Mielke, and P. Cramer. Iwr1 directs RNA polymerase II nuclear import. *Molecular cell*, 42(2):261–266, 2011.

[46] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology*, 6(3-4):327–342, 1999.

[47] N. Daraselia, A. Yuryev, S. Egorov, I. Mazo, and I. Ispolatov. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. *BMC Bioinformatics*, 8:243, 2007.

[48] J.-E. Dazard, S. Saha, and R. M. Ewing. ROCS: a Reproducibility Index and Confidence Score for Interaction Proteomics Studies. *BMC bioinformatics*, 13(1):128, 2012.

[49] E. de Hoffmann. Tandem mass spectrometry: a primer. *Journal of mass spectrometry*, 31(2):129–137, 1996.

[50] U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, 2005.

[51] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10(6):947–960, 2003.

[52] M. Dittrich, G. Klau, A. Rosenwald, T. Dandekar, and T. Muller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223, 2008.

[53] Y.-C. Du, S. Gu, J. Zhou, T. Wang, H. Cai, M. A. MacInnes, E. M. Bradbury, and X. Chen. The dynamic alterations of H2AX complex during DNA repair detected by a proteomic approach reveal the critical roles of Ca2+/calmodulin in the ionizing radiation-induced cell cycle arrest. *Molecular & Cellular Proteomics*, 5(6):1033–1044, 2006.

[54] M. L. Duennwald, S. Jagadish, F. Giorgini, P. J. Muchowski, and S. Lindquist. A network of protein interactions determines polyglutamine toxicity. *Proceedings of the National Academy of Sciences*, 103(29):11051–11056, 2006.

[55] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[56] E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. Discovering motifs in ranked lists of DNA sequences. *PLoS computational biology*, 3(3):e39, 2007.

[57] S. Egloff, J. Zaborowska, C. Laitem, T. Kiss, and S. Murphy. Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. *Molecular cell*, 45(1):111–122, 2012.

[58] J. K. Eng, A. L. McCormack, and J. R. Yates Iii. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[59] A. Enright, S. Van Dongen, and C. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575, 2002.

[60] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[61] R. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. Robinson, L. O'Connor, M. Li, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular systems biology*, 3(1), 2007.

[62] A. V. Favorov, M. S. Gelfand, A. V. Gerasimova, D. A. Ravcheev, A. A. Mironov, and V. Makeev. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–2245, 2005.

[63] D. Fenyö and R. C. Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, 75(4):768–774, 2003.

[64] S. Fields and O.-k. Song. A novel genetic system to detect protein protein interactions. 1989.

[65] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–D288, Jan 2008.

[66] E. Fix and J. Hodges Jr. Project 21-49-004, Report No. 11, under Contract No. AF41(148)-31E. 1952.

[67] R. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, 1962.

[68] D. R. Foltz, L. E. Jansen, B. E. Black, A. O. Bailey, J. R. Yates, and D. W. Cleveland. The human CENP-A centromeric nucleosome-associated complex. *Nature cell biology*, 8(5):458–469, 2006.

[69] D. Forget, A.-A. Lacombe, P. Cloutier, R. Al-Khoury, A. Bouchard, M. Lavallée-Adam, D. Faubert, C. Jeronimo, M. Blanchette, and B. Coulombe. The protein interaction network of the human transcription machinery reveals a role for the conserved GTPase RPAP4/GPN1 and microtubule assembly in nuclear import and biogenesis of RNA polymerase II. *Molecular & Cellular Proteomics*, 9(12):2827–2839, 2010.

[70] D. Forget, A.-A. Lacombe, P. Cloutier, M. Lavallée-Adam, M. Blanchette, and B. Coulombe. Nuclear import of RNA polymerase II is coupled with nucleocytoplasmic shuttling of the RNA polymerase II-associated protein 2. *Nucleic acids research*, 2013.

[71] T.-J. Fu, J. Peng, G. Lee, D. H. Price, and O. Flores. Cyclin K functions as a CDK9 regulatory subunit and participates in RNA polymerase II transcription. *Journal of Biological Chemistry*, 274(49):34527–34530, 1999.

[72] A. Galarneau, M. Primeau, L.-E. Trudeau, and S. W. Michnick. β-Lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein–protein interactions. *Nature biotechnology*, 20(6):619–622, 2002.

[73] A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat, and et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[74] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[75] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, 11(6), 2012.

[76] A. Gingras, R. Aebersold, and B. Raught. Advances in protein complex analysis using mass spectrometry. *The Journal of Physiology*, 563(1):11, 2005.

[77] A.-C. Gingras and B. Raught. Beyond hairballs: The use of quantitative mass spectrometry data to understand protein–protein interactions. *FEBS letters*, 586(17):2723–2731, 2012.

[78] H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, A. Droege, K. S. Lindenberg, M. Knoblich, C. Haenig, et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Molecular cell*, 15(6):853–865, 2004.

[79] G. A. Gonzalez and M. R. Montminy. Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133. *Cell*, 59(4):675–680, 1989.

[80] D. Görlich, S. Prehn, R. A. Laskey, and E. Hartmann. Isolation of a protein that is essential for the first step of nuclear protein import. *Cell*, 79(5):767–778, 1994.

[81] B. Göttgens, L. M. Barton, M. A. Chapman, A. M. Sinclair, B. Knudsen, D. Grafham, J. G. Gilbert, J. Rogers, D. R. Bentley, and A. R. Green. Transcriptional regulation of the stem cell leukemia gene (SCL)Comparative analysis of five vertebrate SCL loci. *Genome research*, 12(5):749–759, 2002.

[82] V. Granholm and L. Käll. Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics*, 11(6):1086–1093, 2011.

[83] M. Gry, R. Rimini, S. Strömberg, A. Asplund, F. Pontén, M. Uhlén, and P. Nilsson. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC genomics*, 10(1):365, 2009.

[84] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517, 2005.

[85] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, 2004.

[86] U. Hardeland and E. Hurt. Coordinated nuclear import of RNA polymerase III subunits. *Traffic*, 7(4):465–473, 2006.

[87] M. Havilio, Y. Haddad, and Z. Smilansky. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem*, 75(3):435–444, 2003.

[88] G. G. Hicks, N. Singh, A. Nashabi, S. Mai, G. Bozek, L. Klewes, D. Arapovic, E. K. White, M. J. Koury, E. M. Oltz, et al. Fus deficiency in mice results in defective B-lymphocyte development and activation, high levels of chromosomal instability and perinatal death. *Nature genetics*, 24(2):175–179, 2000.

[89] M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, 34(17):e117–e117, 2006.

[90] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, and et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[91] Z. Hu, Y.-C. Chang, Y. Wang, C.-L. Huang, Y. Liu, F. Tian, B. Granger, and C. DeLisi. VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic acids research*, 2013.

[92] Z. Hu, J. Mellor, and C. DeLisi. Analyzing networks with VisANT. *Curr Protoc Bioinformatics*, Chapter 8:Unit 8.8, Dec 2004.

[93] L. A. Huber. Is proteomics heading in the wrong direction? *Nature Reviews Molecular Cell Biology*, 4(1):74–80, 2003.

[94] T. A. Hughes. Regulation of gene expression by alternative untranslated regions. *Trends in Genetics*, 22(3):119–122, 2006.

[95] W. Hwang, T. Kim, M. Ramanathan, and A. Zhang. Bridging centrality: Graph mining from element level to group level. pages 336–344, 2008.

[96] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[97] S. Jain and G. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC bioinformatics*, 11(1):562, 2010.

[98] M. K. Jang, K. Mochizuki, M. Zhou, H.-S. Jeong, J. N. Brady, and K. Ozato. The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Molecular cell*, 19(4):523–534, 2005.

[99] S. Jang, H. Kräusslich, M. Nicklin, G. Duke, A. Palmenberg, and E. Wimmer. A segment of the 5'nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *Journal of Virology*, 62(8):2636–2643, 1988.

[100] H. Jeong, S. Mason, A. Barabási, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[101] C. Jeronimo, D. Forget, A. Bouchard, Q. Li, G. Chua, C. Poitras, C. Thérien, D. Bergeron, S. Bourassa, J. Greenblatt, et al. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Molecular cell*, 27(2):262–274, 2007.

[102] C. Jeronimo, M.-F. Langelier, M. Zeghouf, M. Cojocaru, D. Bergeron, D. Baali, D. Forget, S. Mnaimneh, A. P. Davierwala, J. Pootoolal, et al. RPAP1, a novel human RNA polymerase II-associated protein affinity purified with recombinant wild-type and mutated polymerase subunits. *Molecular and cellular biology*, 24(16):7043–7058, 2004.

[103] B. H. Junker and F. Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2008.

[104] W. G. Kaelin Jr, D. C. Pallas, J. A. DeCaprio, F. J. Kaye, and D. M. Livingston. Identification of cellular proteins that can interact specifically with the T/ElA-binding region of the retinoblastoma gene product. *Cell*, 64(3):521–532, 1991.

[105] L. Kall, J. Canterbury, J. Weston, W. Noble, and M. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–926, 2007.

[106] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, Jan 2008.

[107] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.

[108] T. Kawakami, K. Tateishi, Y. Yamano, T. Ishikawa, K. Kuroki, and T. Nishimura. Protein identification from product ion spectra of peptides validated by correlation between measured and predicted elution times in liquid chromatography/mass spectrometry. *Proteomics*, 5(4):856–864, 2005.

[109] A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem*, 74(20):5383–5392, 2002.

[110] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.

[111] E. Kershnar, S.-Y. Wu, and C.-M. Chiang. Immunoaffinity purification and functional characterization of human transcription factor IIH and RNA polymerase II from clonal cell lines that conditionally express epitope-tagged subunits of the multiprotein complexes. *Journal of Biological Chemistry*, 273(51):34444–34453, 1998.

[112] A. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 2004.

[113] T. Kislinger, K. Rahman, D. Radulovic, B. Cox, J. Rossant, and A. Emili. PRISM, a generic large scale proteomic investigation strategy for mammals. *Molecular & Cellular Proteomics*, 2(2):96–106, 2003.

[114] E. Kolker, R. Higdon, and J. M. Hogan. Protein identification and expression analysis using mass spectrometry. *Trends in microbiology*, 14(5):229–235, 2006.

[115] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *In Proceedings of the ICML*, pages 315–322, 2002.

[116] M. Kozak. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic acids research*, 15(20):8125–8148, 1987.

[117] K. Krapfenbauer, M. Fountoulakis, and G. Lubec. A rat brain protein expression map including cytosolic and enriched mitochondrial and microsomal fractions. *Electrophoresis*, 24(11):1847–1870, 2003.

[118] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, et al. Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500, 2005.

[119] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[120] B. J. Krueger, C. Jeronimo, B. B. Roy, A. Bouchard, C. Barrandon, S. A. Byers, C. E. Searcey, J. J. Cooper, O. Bensaude, É. A. Cohen, et al. LARP7 is a stable component of the 7SK snRNP while P-TEFb, HEXIM1 and hnRNP A1 are reversibly associated. *Nucleic acids research*, 36(7):2219–2229, 2008.

[121] S. A. Krum, G. A. Miranda, C. Lin, and T. F. Lane. BRCA1 associates with processive RNA polymerase II. *Journal of Biological Chemistry*, 278(52):52012–52020, 2003.

[122] K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3):309–316, 2007.

[123] M.-C. Lai, H.-W. Kuo, W.-C. Chang, and W.-Y. Tarn. A novel splicing regulator shares a nuclear import pathway with SR proteins. *The EMBO journal*, 22(6):1359–1369, 2003.

[124] J.-P. Lambert, L. Mitchell, A. Rudner, K. Baetz, and D. Figeys. A novel proteomics approach for the discovery of chromatin-associated protein networks. *Molecular & Cellular Proteomics*, 8(4):870–882, 2009.

[125] J.-P. Lambert, T. Pawson, and A.-C. Gingras. Mapping physical interactions within chromatin by proteomic approaches. *Proteomics*, 12(10):1609–1622, 2012.

[126] V. Lange, P. Picotti, B. Domon, and R. Aebersold. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology*, 4(1), 2008.

[127] M. Lavallée-Adam, P. Cloutier, B. Coulombe, and M. Blanchette. Modeling contaminants in AP-MS/MS experiments. *Journal of proteome research*, 10(2):886–895, 2010.

[128] M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of functional sequence motifs in human 5' UTRs based on local enrichments in a protein-protein interaction network. *Manuscript in preparation.*

[129] M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of locally over-represented GO terms in protein-protein interaction networks. In *Research in Computational Molecular Biology*, pages 302–320. Springer, 2009.

[130] M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of locally over-represented GO terms in protein-protein interaction networks. *Journal of Computational Biology*, 17(3):443–457, 2010.

[131] M. Lavallée-Adam, J. Rousseau, C. Domecq, A. Bouchard, D. Forget, D. Faubert, M. Blanchette, and B. Coulombe. Discovery of cell compartment specific protein-protein interactions using affinity purification combined with tandem mass spectrometry. *Journal of proteome research*, 12(1):272–281, 2012.

[132] E. L. Lawler, J. K. Lenstra, and A. Rinnooy Kan. Generating all maximal independent sets: NP-hardness and polynomial-time algorithms. *SIAM Journal on Computing*, 9(3):558–565, 1980.

[133] I. Lee, S. S. Ajay, J. I. Yook, H. S. Kim, S. H. Hong, N. H. Kim, S. M. Dhanasekaran, A. M. Chinnaiyan, and B. D. Athey. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome research*, 19(7):1175–1183, 2009.

[134] E. A. Leibold and H. N. Munro. Cytoplasmic protein binds in vitro to a highly conserved sequence in the 5'untranslated region of ferritin heavy-and light-subunit mRNAs. *Proceedings of the National Academy of Sciences*, 85(7):2171–2175, 1988.

[135] L. Leibovich, I. Paz, Z. Yakhini, and Y. Mandel-Gutfreund. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic acids research*, 2013.

[136] L. Leibovich and Z. Yakhini. Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic acids research*, 40(13):5832–5847, 2012.

[137] E. D. Levy and J. B. Pereira-Leal. Evolution and dynamics of protein interactions and networks. *Current opinion in structural biology*, 18(3):349–357, 2008.

[138] X. Li, E. A. Foley, S. A. Kawashima, K. R. Molloy, Y. Li, B. T. Chait, and T. M. Kapoor. Examining post-translational modification-mediated protein-protein interactions using a chemical proteomics approach. *Protein Science*, 2013.

[139] Y. Li, P. Agarwal, and D. Rajagopalan. A global pathway crosstalk network. *Bioinformatics*, 24(12):1442–7, 2008.

[140] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[141] J.-C. Lin and W.-Y. Tarn. Exon selection in α-tropomyosin mRNA is regulated by the antagonistic action of RBM4 and PTB. *Molecular and cellular biology*, 25(22):10111–10121, 2005.

[142] C. Linhart, Y. Halperin, and R. Shamir. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome research*, 18(7):1180–1189, 2008.

[143] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates. Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology*, 17(7):676–682, 1999.

[144] H. Liu and C. H. Herrmann. Differential localization and expression of the Cdk9 42k and 55k isoforms. *Journal of cellular physiology*, 203(1):251–260, 2005.

[145] G. Lolli. Binding to DNA of the RNA-polymerase II C-terminal domain allows discrimination between Cdk7 and Cdk9 phosphorylation. *Nucleic acids research*, 37(4):1260–1268, 2009.

[146] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.

[147] M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993.

[148] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2013. R package version 1.14.4 — For new features, see the 'Changelog' file (in the package source).

[149] M. Mandal and R. R. Breaker. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology*, 5(6):451–463, 2004.

[150] M. Mann, R. C. Hendrickson, and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annual review of biochemistry*, 70(1):437–473, 2001.

[151] M. D. McDowall, M. S. Scott, and G. J. Barton. PIPs: human protein–protein interaction prediction database. *Nucleic acids research*, 37(suppl 1):D651–D656, 2009.

[152] E. J. Meer, D. O. Wang, S. Kim, I. Barr, F. Guo, and K. C. Martin. Identification of a cis-acting element that localizes mRNA to synapses. *Proceedings of the National Academy of Sciences*, 109(12):4639–4644, 2012.

[153] D. Mellacheruvu, Z. Wright, A. L. Couzens, J.-P. Lambert, N. St-Denis, T. Li, Y. V. Mitev, S. Hauri, M. E. Sardiu, T. Y. Low, et al. The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data. *Journal of Biomolecular Techniques: JBT*, 24(Suppl):S50, 2013.

[154] M. Mete, F. Tang, X. Xu, and N. Yuruk. A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics*, 9 Suppl 9:S19, 2008.

[155] N. Méthot, M. S. Song, and N. Sonenberg. A region rich in aspartic acid, arginine, tyrosine, and glycine (DRYG) mediates eukaryotic initiation factor 4B (eIF4B) self-association and interaction with eIF3. *Molecular and cellular biology*, 16(10):5328–5334, 1996.

[156] H. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, and et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic acids research*, 32(Database Issue):D41, 2004.

[157] O. Meyuhas. Synthesis of the translational apparatus is regulated at the translational level. *European Journal of Biochemistry*, 267(21):6321–6330, 2000.

[158] H. Molina, D. M. Horn, N. Tang, S. Mathivanan, and A. Pandey. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, 104(7):2199–2204, 2007.

[159] A. Mosley, S. Pattenden, M. Carey, S. Venkatesh, J. Gilmore, L. Florens, J. Workman, and M. Washburn. Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Molecular cell*, 34(2):168–178, 2009.

[160] V. E. Myer and R. A. Young. RNA polymerase II holoenzymes and subcomplexes. *Journal of Biological Chemistry*, 273(43):27757–27760, 1998.

[161] A. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, and et al. A statistical model for identifying proteins by tandem mass spectrometry. *ANALYTICAL CHEMISTRY-WASHINGTON DC-*, 75(17):4646–4658, 2003.

[162] Z. Ni, J. B. Olsen, X. Guo, G. Zhong, E. D. Ruan, E. Marcon, P. Young, H. Guo, J. Li, J. Moffat, et al. Control of the RNA polymerase II phosphorylation state in promoter regions by CTD interaction domain-containing proteins RPRD1A and RPRD1B. *Transcription*, 2(5):237–242, 2011.

[163] D. Nikolov and S. Burley. RNA polymerase II transcription initiation: a structural view. *Proceedings of the National Academy of Sciences*, 94(1):15–22, 1997.

[164] T. Ohbayashi, Y. Makino, and T. A. Tamura. Identification of a mouse TBP-like protein (TLP) distantly related to the drosophila TBP-related factor. *Nucleic Acids Res*, 27(3):750–755, Feb 1999.

[165] I. A. Olave, S. L. Reck-Peterson, and G. R. Crabtree. Nuclear actin and actin-related proteins in chromatin remodeling. *Annu Rev Biochem*, 71(NIL):755–81, 2002.

[166] S. Oliver. Proteomics: guilt-by-association goes global. *Nature*, 403(6770):601–603, 2000.

[167] D. A. Omahen. MicroRNA and diseases of the nervous system. *Neurosurgery*, 69(2):440–454, 2011.

[168] J. R. Orlinick and M. V. Chao. Interactions of cellular polypeptides with the cytoplasmic domain of the mouse Fas antigen. *Journal of Biological Chemistry*, 271(15):8627–8632, 1996.

[169] U. A. Ørom, F. C. Nielsen, and A. H. Lund. MicroRNA-10a binds the 5' UTR of ribosomal protein mRNAs and enhances their translation. *Molecular cell*, 30(4):460–471, 2008.

[170] G. Pan, T. Aso, and J. Greenblatt. Interaction of elongation factors TFIIS and elongin A with a human RNA polymerase II holoenzyme capable of promoter-specific initiation and responsive to transcriptional activators. *Journal of Biological Chemistry*, 272(39):24563–24571, 1997.

[171] J. Pelletier and N. Sonenberg. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature*, 334(6180):320–325, 1988.

[172] J. Peng, Y. Zhu, J. Milton, and D. Price. Identification of multiple cyclin subunits of human P-TEFb. *Genes Dev.*, 12(5):755–62, 1998.

[173] L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, 2006.

[174] D. Perkins, D. Pappin, D. Creasy, J. Cottrell, and et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[175] T. Persson and H. Rootzen. Simple and highly efficient estimators for a type I censored normal sample. *Biometrika*, 64(1):123, 1977.

[176] N. B. Pestov and J. Rydström. Purification of recombinant membrane proteins tagged with calmodulin-binding domains by affinity chromatography on calmodulin-agarose: example of nicotinamide nucleotide transhydrogenase. *Nature Protocols*, 2(1):198–202, 2007.

[177] B. M. Pickering and A. E. Willis. The implications of structured 5' untranslated regions on translation and disease. In *Seminars in cell & developmental biology*, volume 16, pages 39–47. Elsevier, 2005.

[178] C. Polychronakos. Gene expression as a quantitative trait: what about translation? *Journal of medical genetics*, 49(9):554–557, 2012.

[179] T. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, and et al. Human Protein Reference Database–2009 update. *Nucleic acids research*, 2008.

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

[181] N. Przulj, D. Wigle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340, 2004.

[182] I. Remy, F. Campbell-Valois, and S. W. Michnick. Detection of protein-protein interactions using a simple survival protein-fragment complementation assay based on the enzyme dihydrofolate reductase. *Nature protocols*, 2(9):2120–2125, 2007.

[183] I. Remy and S. W. Michnick. A highly sensitive protein-protein interaction assay based on Gaussia luciferase. *Nature methods*, 3(12):977–979, 2006.

[184] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, 17(10):1030–1032, 1999.

[185] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

[186] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al. Towards a

proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[187] D. P. Ryan and J. M. Matthews. Protein-protein interactions in human disease. *Current opinion in structural biology*, 15(4):441–446, 2005.

[188] M. Said, T. Begley, A. Oppenheim, D. Lauffenburger, and L. Samson. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A.*, 101:52:18006–11, 2004.

[189] L. Salwinski and D. Eisenberg. Computational methods of analysis of protein-protein interactions. *Current opinion in structural biology*, 13(3):377–382, 2003.

[190] M. E. Sardiu, Y. Cai, J. Jin, S. K. Swanson, R. C. Conaway, J. W. Conaway, L. Florens, and M. P. Washburn. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences*, 105(5):1454–1459, 2008.

[191] T. Sato, M. Hanada, S. Bodrug, S. Irie, N. Iwama, L. H. Boise, C. B. Thompson, E. Golemis, L. Fong, and H.-G. Wang. Interactions among members of the Bcl-2 protein family analyzed with a yeast two-hybrid system. *Proceedings of the National Academy of Sciences*, 91(20):9238–9242, 1994.

[192] C. Saunders and R. S. Cohen. The role of oocyte transcription, the 5' UTR, and translation repression and derepression in *Drosophila gurken* mRNA and protein localization. *Molecular cell*, 3(1):43–54, 1999.

[193] J. Scott, T. Ideker, R. M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13(2):133–144, Mar 2006.

[194] T. Z. Sen, A. Kloczkowski, and R. L. Jernigan. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, 7:355, 2006.

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

[196] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(1), 2007.

[197] A. D. Sharrocks, A. L. Brown, Y. Ling, and P. R. Yates. The ETS-domain transcription factor family. *The international journal of biochemistry & cell biology*, 29(12):1371–1387, 1997.

[198] A. Shevchenko, J. H. Henrik Tomas, J. V. Olsen, and M. Mann. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature protocols*, 1(6):2856–2860, 2007.

[199] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7:199, 2006.

[200] P. Shore, L. Bisset, J. Lakey, J. P. Waltho, R. Virden, and A. D. Sharrocks. Characterization of the Elk-1 ETS DNA-binding domain. *Journal of Biological Chemistry*, 270(11):5805–5811, 1995.

[201] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[202] S. Sinha and M. Tompa. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic acids research*, 31(13):3586–3588, 2003.

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

[204] R. R. Sokal and C. D. Michener. *A statistical method for evaluating systematic relationships*. University of Kansas, 1958.

[205] S. Solier, J. Barb, B. R. Zeeberg, S. Varma, M. C. Ryan, K. W. Kohn, J. N. Weinstein, P. J. Munson, and Y. Pommier. Genome-wide analysis of novel splice variants induced by topoisomerase I poisoning shows preferential occurrence in genes encoding splicing factors. *Cancer research*, 70(20):8055–8065, 2010.

[206] M. E. Sowa, E. J. Bennett, S. P. Gygi, and J. W. Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403, 2009.

[207] H. Spåhr, G. Calero, D. A. Bushnell, and R. D. Kornberg. Schizosacharomyces pombe RNA polymerase II at 3.6-Å resolution. *Proceedings of the National Academy of Sciences*, 106(23):9185–9190, 2009.

[208] V. Spirin and L. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12123, 2003.

[209] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, et al. The BioGRID

interaction database: 2011 update. *Nucleic acids research*, 39(suppl 1):D698–D704, 2011.

[210] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[211] E. Strittmatter, L. Kangas, K. Petritis, H. Mottaz, G. Anderson, Y. Shen, J. Jacobs, D. Camp II, and R. Smith. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *Journal of Proteome Research*, 3(4):760–769, 2004.

[212] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[213] M. Suderman and M. Hallett. Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659, Oct 2007.

[214] D. J. Taatjes. The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends in biochemical sciences*, 35(6):315–322, 2010.

[215] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. S. Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S. W. Michnick. An in vivo map of the yeast protein interactome. *Science*, 320(5882):1465–1470, 2008.

[216] S. Tate, B. Larsen, R. Bonner, and A.-C. Gingras. Label-free quantitative proteomics trends for protein-protein interactions. *Journal of proteomics*, 2012.

[217] B. Turner, S. Razick, A. L. Turinsky, J. Vlasblom, E. K. Crowdy, E. Cho, K. Morrison, I. M. Donaldson, and S. J. Wodak. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database: the journal of biological databases and curation*, 2010, 2010.

[218] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. PMID17353930. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

[219] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697–700, 2003.

[220] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, et al. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, 2008.

[221] C. Vogel, R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular systems biology*, 6(1), 2010.

[222] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *NATURE-LONDON-*, pages 399–404, 2002.

[223] A. Wada, M. Fukuda, M. Mishima, and E. Nishida. Nuclear export of actin: a novel mechanism regulating the subcellular localization of a major cytoskeletal protein. *The EMBO Journal*, 17(6):1635–1641, 1998.

[224] A. J. Walhout and M. Vidal. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, 24(3):297–306, 2001.

[225] S. Warshall. A theorem on boolean matrices. *Journal of the ACM*, 9(1):11–12, 1962.

[226] M. P. Washburn, D. Wolters, and J. R. Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology*, 19(3):242–247, 2001.

[227] I. G. Wool. The structure and function of eukaryotic ribosomes. *Annual review of biochemistry*, 48(1):719–754, 1979.

[228] X. Xie, J. Lu, E. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.

[229] X. Yan, R. Habedanck, and E. A. Nigg. A complex of two centrosomal proteins, CAP350 and FOP, cooperates with EB1 in microtubule anchoring. *Molecular biology of the cell*, 17(2):634–644, 2006.

[230] J. R. Yates. Mass spectrometry: from genomics to proteomics. *Trends in Genetics*, 16(1):5–8, 2000.

[231] H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[232] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003.

[233] M. Zeghouf, J. Li, G. Butland, A. Borkowska, V. Canadien, D. Richards, B. Beattie, A. Emili, and J. F. Greenblatt. Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. *Journal of proteome research*, 3(3):463–468, 2004.

[234] A. Zhang. *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, 2009.

[235] M. Zhou, K. Huang, K.-J. Jung, W.-K. Cho, Z. Klase, F. Kashanchi, C. A. Pise-Masison, and J. N. Brady. Bromodomain protein Brd4 regulates human immunodeficiency virus transcription through phosphorylation of CDK9 at threonine 29. *Journal of virology*, 83(2):1036–1044, 2009.