# Mathematical and Statistical Models for the Analysis of Protein Interactions

Tony Chiang

King's College

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

European Molecular Biology Laboratory,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SD,
United Kingdom.

Email: tchiang@ebi.ac.uk

March 18, 2011

*There’s a divinity that shapes our ends,*
*Rough-hew them how we will.*
**Act 5, Scene II**

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

No part of this dissertation has every been submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This dissertation has been typeset in 12 pt Palatino using $\LaTeX 2\varepsilon$ according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

March 18, 2011 Tony Chiang

# Mathematical and Statistical Models for the Analysis of Protein Interactions

## Summary

Tony Chiang
King's College

March 18, 2011

Protein interactions, both amongst themselves and with other molecules, are responsible for much of the work within the cellular machine. As protein interaction data sets grow in number and in size, from experiments such as Yeast 2-Hybrid or Affinity Purification followed by Mass Spectrometry, there is a need to analyze the data both quantitatively and qualitatively. One area of research is determining how reliable a report of a protein interaction is - whether it could be reproduced if the experiment were repeated; or, if it were tested using an independent assay. One might aim to score each reported interaction using a quantitative measure of reliability. Studies that have tried to answer this question generally have been affected by both systematic and stochastic errors that occur in the experiments. Ultimately, protein interactions need to be addressed at the systems level where both the dynamic and functional nature of protein complexes and other types of interactions is ascertained.

In this dissertation, I present two methodological developments that are useful towards elucidating the nature of protein interaction graphs in the model organism *Saccharomyces cerevisiae*. The first one aims to estimate the *sensitivity* and *specificity* of a protein interaction data set, and does that, as much as possible, by looking at the data set's internal consistency and reproducibility. The second method aims to estimate the node degree distribution, using a multinomial model which is fit by maximum likelihood.

In the development of the methods for the analysis of the protein interactions, computational tools were built in the statistical environment R. Such tools are necessary for the implementation of each analytic step, for rendering visualisations of intermediate and conclusive results, and for the construction of optimal work-flows so as to make our research reproducible and extensible. We have also included such a work-flow in this dissertation as well as the software engineering component of the research.

# Preface

This manuscript encompasses work done in the period between April 2005 and October 2009. Much of the research was carried out at the European Molecular Biology Laboratory's (EMBL) outstation, the European Bioinformatics Institute (EBI) in Cambridge, England. The project was supported by a predoctoral fellowship from EMBL, the Oversea's Research Studentship from the University of Cambridge, and a King's College Bursary.

Much of this dissertation is the direct result of collaborations with other researchers, and without their help, I could not have come this far. My main collaborator is Denise Scholtens who has worked with me in every step of my graduate career; our mutual work has resulted in three journal publications and culminated in my doctoral thesis. Robert Gentleman and Wolfgang Huber have also been instrumental in my graduate research having served as the principal investigators for two of these three papers and also as co-advisers for my dissertation.

My other collaborators include: Deepayan Sarkar, Nianhua Li, Marc Carlson, Sujay Datta, Patrick Paddison (FHCRC); Sandra Orchard, Samuel Kerrien, Henning Hermjakob, Florence Cavalli (EBI); Li Wang (University of Southern California); Jitao David Zhang (German Cancer Research Center, DKFZ); Adrian Vetta (McGill University).

My appreciation to Lars Steinmetz, Gos Micklem, and Paul Bertone for their advice in and beyond my thesis advisory committee meetings.

Some noteworthy thanks for people who have helped me in these last few years: members of the Huber group at the EBI and the Gentleman group at the FHCRC. To both CUCC (Cambridge University Canoe and Cycling Clubs) for giving me some methods of venting when I became frustrated with my work. These clubs would later be replaced by the Cucina Fresca Cycling Team and the Washington Alpine Club. DS for working with someone who e-mailed out of the blue one day from Seattle, for helping me with so many projects, for helping me with this thesis, and for a great deal of e-mail and telephone encouragements. PR for being my best mate in the UK, for helping me navigate King's and Cambridge, and for giving me advice about working towards a DPhil. GP for talking science and life with me to give me some perspective about working outside of maths. OS for giving me an often humourous view on an Ukranian's perspective of life. JT for supplying these incredible LaTeX files for writing a Cambridge thesis (albeit unknowing to you that I filched it). MC for always being able to see the positive even if it is all in his head. CS for giving me a distraction and talking US politics with me via skype when I needed to take the piss. AS for calling me almost daily even while he was moving from Berkeley to UCSD. JW for believing in me and always encouraging me. GC for visiting his MIT roomie in Seattle on a day's notice, coming from Texas when I needed someone to help me get grounded again. RC for advice on academics, love, life, sailing, cycling, and so many random things. JD for becoming my only friend in Seattle when I moved back from the UK...now my best friend in Seattle. AM for forgiving my past bouts of moodiness and teaching me how to climb rocks again! MP for saving my life a couple-a-few times. My mum, dad, and brother for almost everything else.

Lastly, I am in great debt to those who have made an impact and helped me to decide and then to appreciate my chosen path. I work at the Fred Hutchinson Cancer Research Center, but my work often tends to be far removed from the devastating effects of cancer itself. I have seen it in others:

- In Marah who has just had a parent diagnosed with lung cancer; Marah has attributed her own kindness and values to the influence of her parents, and she has shown that same kindness towards me on several occasions.

- In Brett who has lost his older sister Chrissy to a brain tumor. And in the last few days of Chrissy's short life, he told me that "despite all of that, she has become more and more peaceful and happy in the past few weeks. She has always been one to do things the hard way, so being hopeful in the face of all this just shows that the tumor can't ever take away who she is."

- In Ric who successfully battled lymphoma and continues to live life in a positive way! He is still the same guy that supervised a confused 17 year old intern at Motorola 14 years ago and helped guide this kid away from Corporate America and into a life of academics and non-profits.

- In Helen who recently lost her 5 year battle with colorectal cancer. Helen was diagnosed 5 years ago whilst working as an Executive Assistant at the EBI. Her optimism led to her trekking Peru to raise money for cancer research, founding her own successful company, and becoming a motivational speaker for others fighting cancer. I received a note a few months ago about her passing, and now I can no longer say that I have never been directly affected by cancer.

- In Andrea who is fighting stage 4 lung cancer as I am about to submit the final draft of my dissertation. I can only be a spectator in this fight, though I want to be fighting at her side each moment of each day. With Andrea, I shall practice *ānāpānasati* and be mindful of each breath I take.

I have been inspired in big ways and small by these folks. Sometimes I cry in sadness and other times in joy, but even after my tears have dried and faded, I will continue to look onto these people to help guide my decision and choices.

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| **AD** | activation domain |
| **APMS** | affinity purification - mass spectronomy |
| **bp** | base pair |
| **BF** | Bayes factor |
| **BD** | binding domain |
| **CoIP** | co-immunoprecipitation |
| **CC** | correlation coefficient |
| **cDNA** | complementary deoxyribonucleic acid |
| **ChIP** | chromatin immunoprecipitation |
| **ER** | Erdös-Rényi |
| **FDR** | false discovery rate |
| **FN** | false negative |
| **FP** | false positive |
| **GO** | gene ontology |
| **GS** | gold standard |
| **i.i.d.** | independently identically distributed |
| **ORF** | open reading frame |
| **Pol I, II, III** | RNA Polymerase I, II, III |
| **RMSE** | Root Mean Squared Errors |
| **SGD** | Saccharomyces Genome Database |
| **TF** | transcription factor |
| **VB** | viable bait |
| **VBO** | viable bait only |
| **VBP** | viable bait/prey |
| **VP** | viable prey |
| **VPO** | viable prey only |

# Glossary

The following terms are used throughout this dissertation.

*Yeast 2-Hybrid* (Y2H) and the *Affinity Purification - Mass Spectronomy* (APMS) are interaction assays that test for direct physical interactions and protein complex co-membership respectively. In both technologies, there is an asymmetry between the sets of proteins involved, and they can be divided into *bait* and *prey*. The Y2H system is based on the gene expression regulation activity of a known TF, commonly GAL4. The TF is divided into two portions, one with the BD and the other with the AD. By convention, the bait protein is expressed with the BD attached; the prey protein, with the AD. Interaction between the two hybrid proteins can reconstitute the functional TF, resulting in the expression of a reporter gene. In the APMS system, a pre-defined bait is selected and a tag (e. g. an antibody epitope) is attached to the N-terminus of the protein. This bait protein is expressed in vivo and free to bind to other proteins in order to form multi-protein complexes. The cell is then opened and the bait protein (along with all members of the protein complex to which it belongs) is obtained by attracting the tag (via e. g. an affinity bead). Lastly, the proteins forming the complex are digested and the peptides are placed into a mass spectrometer for identification. It is important to note that the prey proteins are not altered in APMS; they are simply those proteins found to be co-complexed with the bait. In general, the *bait* is the protein whose interaction partners we are seeking; the *prey* proteins are those proteins detected to interact with a particular bait.

The *interactome* is the set of all interactions that occur in vivo. We can consider all direct binary interactions, which we model by an ordinary undirected graph, or we can consider complex membership interactions, which we model by a hypergraph. Both the graph and hypergraph models are simplifications of what is going on in nature, since interactions can be transient or stable, of varying degree of affinity, dependent on the presence of co-factors and/or the absence of competitors, and on post-translational modification of the proteins.

A *multi-protein complex* is a collection of proteins that combine into one coherent structure in vivo with some particular function within the cell. Protein complexes may share structure with other protein complexes (i.e. they might have common sub-components); the existence and nature of protein complexes can cover a wide spectrum of stability from complete and total integrity to being very transient in nature.

**Concerning gene and protein names** I follow the organism-specific conventions for writing gene and protein names in this dissertation. With *Saccharomyces cerevisiae*, gene locus names are given in all upper case letters (YNL309W); gene symbols are italicised, and all letters are in upper case as well, for instance, as with *STB1* for a gene whose description is *Sin Three Binding protein*. *S. cerevisiae* protein names are the same as the gene name, but are not italicised, and the first letter is upper case and the remaining letters are lower case (Stb1)[1].

| | **Gene locus name** | **Gene name** | **Protein name** |
|---|---|---|---|
| *S. cerevisiae* | YNL309W | *STB1* | Stb1 |

[1] Saccharomyces Genome Database (SGD) naming guidelines:
`http://www.yeastgenome.org/gene_guidelines.shtml`

# Chapter 1

# Introduction

Our research focused on the statistical analysis of protein interaction data derived from two types of assays: the Yeast 2-Hybrid (Y2H) and Affinity Purification - Mass Spectronomy (APMS) technologies. Because it offers the largest number of data sets, we have worked almost exclusively with data from the organism *Saccharomyces cerevisiae* for this thesis.

In this section, we shall give overviews of the technologies used to derive the protein interaction data sets, from which various protein interactomes have been estimated. In addition to the technologies, we will survey the biological features being queried, how the technologies may fall short in identifying these features, and some mathematical models we used to help overcome these difficulties.

Subsequent sections covers individual studies with which we have been involved, and each section shall contain a background section describing the research project, background, and the objectives. Briefly, these sections are:

Chapter 2 describes the systematic and stochastic errors affiliated with both Y2H and APMS like technologies. A *Methods of Moments* model is adopted and used to estimate the per-experiment stochastic error rates (Chiang et al. [1]).

Chapter 3 describes the use of the *Multinomial Model* in conjuction with the *Methods of Moments* for estimating node degree in protein interaction graphs (Scholtens et al. [2]).

Chapter 4 presents a workflow pipeline and the software necessary to conduct our research. We focus on particular software packages that we have designed and implemented for this dissertation (Chiang and Scholtens [3]).

Appendix A gives the background and details to the formulae and methods developed in Chapter 2.

Appendix B gives the background and details to the methods developed in Chapter 3.

Each chapter of this disseration has been published as an independent peer-reviewed journal manuscript, and the results have already been referenced and applied in other publications.

## 1.1 Protein Interactions

Much of the work accomplished within the cellular machine is done through protein interactions. Proteins can interact with a variety of different molecules, but it is their interactions with one another that account for many of the basic functions of each cell. Protein interactions have been studied extensively for the last decade, and several studies have been able to annotate new functions to proteins by their associations with other proteins. These studies, however, used the term interaction in a broad context encapsulating several different types of relationships between genes and gene products including genetic interactions [4–6], genomic context (such as fusion, neighbourhood, and phylogenetic profile) [7], and co-expression [8]. It is useful for us to bear in mind the inherent differences between many of these interactions. For instances, two genes that are known to genetically interact need not physically bond to one another; keeping track of these differences allows us to infer certain types of biological features in a more efficient manner. In this dissertation, we focus on two specific types of interactions: *physical interaction* and *complex membership interaction*. These two types of interactions are closely related, but not interchangeable, especially when attempting to estimate certain properties and features of the cellular protein interaction network.

## 1.2 Graph Theory and Protein Interactions

Much of the research in studying protein interactions has been done with the use of mathematical graph theory. Here we shall give a brief overview of some of the terminology and structures that have made their way from mathematics into biology.

First, we very briefly outline some of the basics of graph theory. More comprehensive treatments are given in [9] and [10]. A graph can be represented in many different ways; we will use $G = (V, E)$, the representation in terms of two sets, as our main representation. $V$ denotes a set of vertices, or nodes, and $E$ represents the set of edges that exist between pairs of elements of $V$. The edges represent relationships between the nodes. We denote the individual edges $e_i$. We let $|V|$ denote the cardinality of $V$ and similarly for $E$.

A walk between two nodes, $v_1$ and $v_2$, in a graph, $G$, is an alternating sequence of nodes and edges in $G$ that begins with $v_1$ and ends with $v_2$ and in which every edge joins two adjacent vertices. A path is a walk with no repeated edges and no repeated vertices (except possibly the first and last). A shortest path is the shortest of all possible paths. The diameter of a graph is the longest shortest path between any two members of the graph.

In some cases the relationships are directed while in others they are not. A graph in which all edges are not directed is often called an *undirected* graph, while any graph with one or more directed edges is a *directed* graph. If the graph is directed, we define two functions, $\phi(e_{i,j}) = i$ and $\psi(e_{i,j}) = j$ on the edges in $E$, taking values in $V$ with the implication that the relationship (or edge) is asymmetric and originates at the *from* node and ends at the *to* node. In particular, the $\phi$ function returns the *from* vertex and the $\psi$ returns the *to* vertex.

The *degree* of a node is the number of edges incident on that node. If the graph is directed then the *in degree* refers to the number of edges which end at the node and the *out degree* refers to the number of edges which originate at the node.

A multi-graph is a graph with multiple edge sets, where each set can represent a different type of relationship between the nodes of the graph. A bipartite graph is a graph where the nodes can be partitioned into two sets such that all edges go from nodes in one set to nodes in the other. Bipartite graphs can, for instance, be used to represent the relationship between proteins and protein complexes; one set of nodes represents the proteins, the other the complexes, and an edge indicates membership of a protein in a complex. Directly corresponding to the bipartite graph is the hypergraph. A hypergraph consists of a vertex set, $V$, together with a set of hyperedges, $\mathcal{H}$. A hyperedge, $H$ is any subset of the vertex set. Hence, we see that a hypergraph directly corresponds to a bipartite graph where one set of nodes is $V$ and the other the set of hyperedges.

A subgraph of $G = (V, E)$ can be defined by a subset of the nodes, $V_S \subset V$, where the nodes of the subgraph consist of the nodes $V_S$ and the edges are $E_V \subset E$ where $(i, j) = e_{i,j} \in E_V : i, j \in V_S$. We can denote such a subgraph as $G_S = (V_S, E_V)$. Similarly a subgraph of $G = (V, E)$ can be defined in terms of a subset of the edges, $E_S \subset E$, where the nodes in the subgraph are determined by the endpoints of the edges in $E_S$ and hence are denoted $V_E$. Thus, we write $G_E = (V_E, E_S)$. The two subgraphs are similar, but it is not, in general the case that the subgraph defined by the nodes $V_E$ is identical to that defined by $E_S$ since there can be edges, in $G$, between nodes in $V_E$ that are not in $E_S$.

There are a number of important concepts that can be used to describe cohesive subunits of a graph. A clique is a subset of the nodes of $V_s$ such that all members of the clique are connected by an edge to all other members. A maximal clique is a clique which is not a proper subset of any other clique. A $k$-core is a subgraph in which each node is adjacent to at least $k$ of the other members of the subgraph ($k$-cores were studied in [11]).

A graph is said to be connected if there is a walk between every pair of distinct vertices. A graph can consist of one or more connected components. A cut-set is a collection of vertices in a graph, $G$, such that their removal results in more connected components than in $G$.

## 1.3 Direct Interaction and Yeast 2-Hybrid

### 1.3.1 Background

The first type of interaction in which we are interested is a direct physical interaction between two proteins. A physical interaction is a direct and specific contact between a pair of proteins [12]. One method of detecting a direct physical interaction between two proteins is via *Yeast 2-Hybrid* technology [13–17]. In this technology, two protein encoding genes are chosen for examination; the sequence for a DNA binding domain (BD) is added to one of the genes whilst the sequence for the corresponding DNA activation domain (AD) is added to the other gene.

Expression of these two hybrid proteins allows for the possible reconstitution of a functional transcription factor (TF) if the two proteins physically interact. In theory, both hybrids should be brought into the nucleus by chaperone proteins. If the bait-prey pair interacts and mimics the function of the TF, then transcription of a reporter gene allows inference of an interaction between the two proteins.

### 1.3.2 Limitations to Detecting Direct Interactions

There are several known limitations to ascertaining direct physical interactions via Y2H technology [13, 14, 17–19]. Clearly, the most severe limitation to Y2H technology is that it detects protein interactions under the conditions of the assay rather than typical (i.e. wild type) conditions, and so the interaction must be further analyzed and studied to determine if they are meaningful *de facto* protein interactions. A number of other implications arise from this limitation: time, space, and competition constraints against observed interactions in vivo. For two proteins $(p_1, p_2)$ to interact, they must be found in the same cellular component at the same time. In addition, if protein $p_2$ can interact with $p_1$, but is usually hindered from doing so by the presence of another protein $p_3$, then a positive observation with Y2H technology does not represent a valid functional protein interaction. Many describe this type of observation as *false positive* (FP) interaction measurement or data point. The Y2H technology itself can also have the opposite effect as well, failing to detect a true interaction between two proteins $(p_1, p_2)$.

Attaching either the BD or the AD might change either structural or functional components of the protein, and so inference between $p_1$ and $p_2$ results in a false negative observation.

There are other limitations arising from Y2H technology. In some instances, the addition of either the BD or the AD to the gene $g$ sequence might change properties of the gene in such a way that it can no longer be expressed. This has the effect that inference between the products of $g$ and any other protein cannot be made because these relationships have never been tested. Other proteins might become what is known as *auto-activators*; that is, the inclusion of either a BD or an AD induces them to transcribe the reporter gene without actually binding to the other hybrid and re-forming the intended TF. Some proteins, such as chaperone proteins and kinases, are known to interact with a large number of other proteins. These proteins are often referred to as *sticky* in the literature, and often, these proteins are either removed from the experimental design already, or are removed from the experimental dataset, because these interactions are not thought to give many new insights to the functional organization of the cell. Caution needs to be maintained in differentiating between the auto-activators and the so-called sticky proteins because both types of nodes will have very large degrees though for very different reasons.

One more limitation of particular interest to us is an inferred interaction by way of the *bridging effect* or mediation of a third protein. In this instance, two proteins $(p_1, p_2)$ do not interact directly but rather through a third protein $p_3$ which simultaneously binds to both. This scenario also falls under the category of a FP measurement though it is quite hard to ascertain. In general, structure information or detailed biochemical experimentation is needed to elucidate these types of indirect interactions.

A prominent portion of this dissertation is to identify consequences of these limitations *in and from data*, and to systematize and quantify how they might contribute to the measurement errors present in the experiment. Ascertaining the error rates allows for a more meaningful interpretation of the data and more robust subsequent inference and model parameter estimation.

### 1.3.3 Graph Representation

Direct protein interactions are amenable to being modeled by ordinary graphs. The set $V$ represents either the genes or the gene products while the set $E$ represents a direct physical interaction between the two nodes. Caution must be taken, however, when using graphs to model interaction data because while interactions between proteins are symmetric in nature, the interactions between the bait and the prey hybrids are inherently asymmetric. The limitations referenced in section 1.3.2 point to the fact that attaching the BD to a coding gene region might have drastically different effects than attaching the AD. We need to be mindful of such asymmetric effects when we model the data.

Using a directed graph approach, we can adopt the convention that the bait is the node of origin and the prey is the node of termination. Thus, a directed graph is more amenable in representing the protein interaction data upon which analysis can be conducted. Even the directed graph, however, is not representative enough for protein interaction data most of the time. Either through under-sampling or the failure of proteins to express under the conditions of an Y2H assay, many relationships between protein pairs will not have been queried. This makes the absence of an directed edge between two nodes ambiguous. Was the relationship tested and not found or simply not tested at all? For this reason, we have defined the notion of *viability* for the hybrids. A viable bait (VB) is a bait hybrid that has been observed to have at least one interacting prey partner. A viable prey (VP) is similarly defined. A viable bait/prey (VBP) is a protein that has served both as a VB and VP. It is important to note that viability of a protein is defined within the scope of a single experimental data set, since different conditions might arise in different experiments. We also note that the subgraph induced by the set of all VBPs in one experiment is a graph where all relationships between protein pairs have been tested twice, i.e. in each of the two directions.

The distributions of unreciprocated in- and out-edges for nodes in the VBP graph are used to diagnose nodes prone to systematic bias using the Binomial method of [1]. Systematically biased nodes are excluded from further analysis in this report.

The VBP-subgraph will be an important tool for this dissertation.

### 1.3.4 Binary Protein Interactome

From the Y2H data, we may try to infer the undirected graph of direct protein interactions, or the direct *Binary Protein Interactome*. This graph structure designates the possible direct physical interactions between each protein pair in a given cell state or tissue type (for multi-cellular systems). The graph depends on cell state, for instance, since proteins that can in principle interact may not be co-expressed and co-localised in certain cell states or conditions. It is also understood that its notion of *interaction* is more precisely *the type of interaction that is detectable by the Y2H assay*. The role of this graph is to model the underlying interactome within the context of the experiment.

One aspect of our work is to obtain an estimate for this interactome. From the Y2H data (either with or without information from other types of technologies), we would like to estimate this undirected graph from the directed graph that describes the experimental data. There are a number of uses for such a graph since knowing the possible interaction partners of a particular protein $p$ allows us to infer: 1. the possible biological process in which $p$ might be involved; 2. $p$'s physical structure or folding pattern, if we know those of the interacting proteins [20]; 3. possible small molecules that might interact with $p$ to possibly inhibit its function (drug therapies) by looking at induced phenotypes of such molecular interactions [21, 22]. Such inference allows for many areas of new research at the systems biology level.

## 1.4 Indirect Interaction and Affinity Purification

### 1.4.1 Background

The second type of protein interaction in which we are interested is that of protein complex membership. We have an interest in studying the known multi-protein complexes and discovering new ones. Proteins rarely work alone. Some proteins belong to at least one multi-protein complex to complete some task within the cell [23–28]. Some protein complexes can be quite stable and exist for the life of the cell while others can be transient, assembled for some purpose and quickly broken down when no longer needed. In other instances, the core of the protein complex is stable and certain components (sub-complexes) are transient; in such a case, the entire protein complex need not be constantly assembled and disassembled, and different sub-components can be added and deleted if the complex participates in a multitude of different processes.

In order to discover novel protein complexes, it is necessary to decipher the constitutive members for each new protein complex. Thus we can define a protein complex membership interaction as an indirect interaction between a set of proteins that all belong to at least one protein complex. Ascertaining such a relationship can be done with a number of different technologies. The first is *X-ray Crystallography* where a protein complex is obtained and the complete crystal structure of the complex can be ascertained. In such a technology, we can not only find the constitutive members of the protein complex but we can also find the direct physical interactions between the member proteins. This technique is limited in scope since obtaining crystal structures is a very difficult process.

A widely used method to discover protein complexes is to purify protein complexes via precipitation (such as co-immunoprecipitation, CoIP) or affinity chromatography (or more generally affinity purification) wherein molecules in solution (mobile phase) are separated based on differences in chemical or physical interaction with a stationary material (solid phase) [23, 24]. There are a number of different types of such assays (antibody precipitation, pull-down assays, fusion tag protein purification, et al.), and in this dissertation we shall umbrella all of these technologies under affinity purification - mass spectronomy (APMS) and refer to a specific technology when needed [29].

Currently, the collective APMS technologies are widely used to query for protein complexes within the cellular system. We have also used data derived from such technologies both to try and estimate protein complexes and also to infer other features such as a protein's affiliation with multiple complexes and hence a protein's functional annotations. We shall see that the ability of a protein to belong to more than one distinct protein complexes can confound many of the APMS technologies.

### 1.4.2 Limitations to Finding Protein Complexes

As we have mentioned, the fact that a single protein might belong to a number of distinct multi-protein complexes is one of great importance, but it is also the biggest obstacle that the APMS technologies need to confront. Unlike crystallography, the APMS technologies do not simply query the constitutive members of a single pre-determined multi-protein complex; rather, a pre-determined protein $p$ (or an encoding gene $g$, an antibody $a$, etc) is selected so that all the protein complexes $C_1, \ldots, C_n$ to which $p$ belongs can be obtained. The protein complexes are then reduced to separate proteins and each protein is further reduced to peptides [22]. Mass spectroscopy detects the peptide signatures by molecular weight, and so the proteins *pulled down* with $p$ can be inferred and identified. Thus each pull-down $P$ yields a set of proteins:

$$P = \{\phi : \phi \in C_i \text{ for some } i \ \in \{1, 2, \ldots n\}\} \tag{1.1}$$

that rarely gives the make-up for each protein complex $C_i$. We define the interaction between $p$ with each protein $\phi \in P$ as a co-membership interaction, or simply as an indirect interaction. With perfectly sensitive and specific measurements, $P$ is simply the union over all proteins found in the complexes $C_1, \ldots, C_n$; to identify these complexes from the pull down data is a major goal for the computational biologist. This goal is not a trivial endeavor.

Much like the Y2H technology, the pull down technologies also suffer from detecting spurious co-membership interactions (FP) or fail to detect true ones (FN). False positive interactions can be generated because the purification process fails to remove proteins that are apart of the ambient complex mixture but do not interact with the protein of interest (i.e. the bait protein). Because all proteins obtained from the APMS experiment are digested and then analyzed via the mass spectrometer, proteins can remain part of the mixture that do not interact with the bait protein. False negatives can also occur, and a major reason for FN interactions is because modification of the bait protein (such as the addition of a TAP tag) may hinder the protein's natural interaction abilities.

Much like the Y2H data, the indirect interaction data from APMS technologies suffers from imperfect sensitivity and specificity. We have implemented a new model to estimate the error rates so as to estimate features within the data, such as multi-protein complexes from the co-membership APMS data.

### 1.4.3 Graphs and Hypergraphs

The data generated from the APMS technologies can be readily modeled by ordinary directed graphs where the directed edges originate from the bait protein and terminate at prey proteins. The underlying graph, however, is undirected; the edges of the undirected graph represents indirect co-membership interactions rather than direct physical interactions as in the binary interactome. The undirected graph is not the final entity we seek. We recall that the co-membership relationship between two proteins $p_1, p_2$ implies that there exists at least one multi-protein complex $C$ that contains both these proteins. From the undirected graph, we will still need to estimate the multi-protein complexes.

Because the membership within a protein complex is no longer a one-to-one relationship, ordinary graphs are not amenable to modeling this type of structure. Being part of a protein complex is a relationship between a number of different proteins, and so we need to employ a model that can record this type of relationship. The simplest structure is just to group the members of a protein complex as a subset of all the available proteins. These subsets can be thought of as a hyperedge, or a generalisation of an ordinary edge. Hence, a better graph structure to model protein complexes is that of a hypergraph.

### 1.4.4 Protein Complex Interactome

In order to estimate and catalogue the unique multi-protein complexes that might arise in vivo, it is necessary and sufficient to estimate the hyperedges for the hypergraph model of the protein complex interactome. Hypergraphs offer the flexibility of hyperedges to overlap, sometimes to great extent, so that protein complexes that have a great deal of shared structure will be amenable to this model. An example is RNA polymerase I, II, III 1.1; all three complexes share several protein members. Many analyses of the APMS co-membership data have used some type of clustering methodology, typically, partitioning clustering, to infer protein complex composition. However, such an approach is problematic. For instance, if several protein complexes have shared subunits, a partitioning clustering algorithm might return a single large complex; or alternatively, it might estimate only one of the complexes correctly and miss the others.

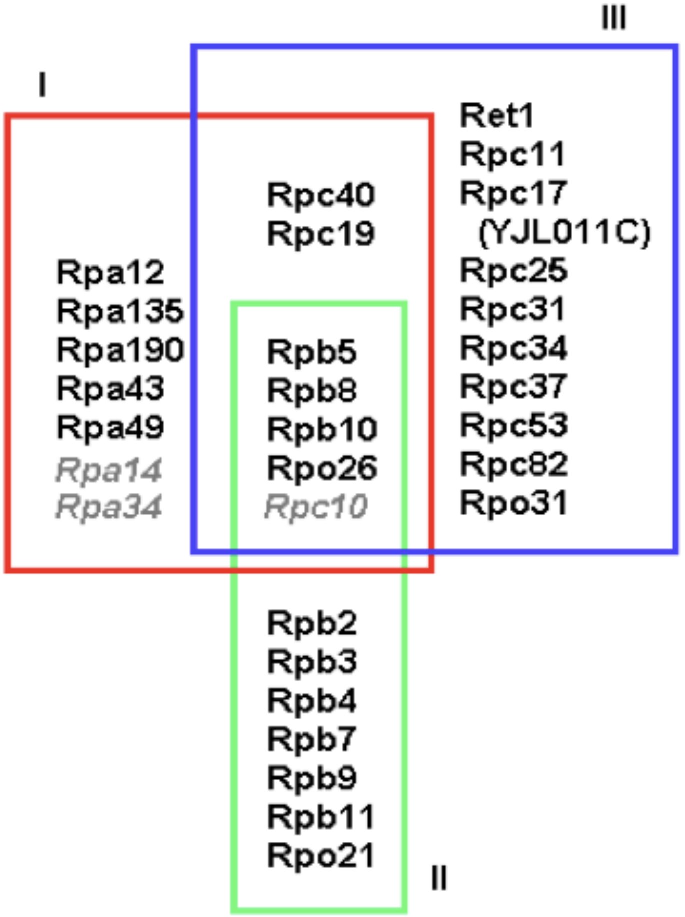


Figure 1.1: The Venn diagram shows the overlap between RNA Polymerase I, II, III. Each of the colored boxes displays the composition of a polymerase complex, and so each box also represents a unique hyperedge in a hypergraph model. A hypergraph is amenable to overlap between complexes, as opposed to partitioning clustering of the data, the usual method of grouping interaction data.

# Chapter 2

# Coverage and Error Models

## 2.1 Introduction

Using a directed graph model for bait to prey systems and a Multinomial error model, we assessed the error statistics in all published large-scale data sets for *Saccharomyces cerevisiae* and characterized them by three traits: the set of tested interactions, artifacts that lead to false positive or negative observations, and estimates of the stochastic error rates that affect the data. These traits provide a prerequisite for the estimation of the protein interactome and its modules.

## 2.2 Background

Within the last decade, a large amount of data on protein-protein interactions in cellular systems has been obtained by the high-throughput scaling of technologies such as the yeast two-hybrid system (Y2H) and affinity purification – mass spectrometry (AP-MS) [15,23,24,26–28,30–38]. This opens the possibility for molecular and computational biologists to obtain an understanding of cellular systems and their modules [39]. There are many references in the literature, however, to the apparent noisiness and low quality of high-throughput protein interaction data. Evaluation studies have reported discrepancies between the data sets, large error rates, lack of overlap, and contradictions between experiments [16,18,19,40–50].

A grand challenge for computational biology is the interpretation and integration of these large sets of protein interaction data.

In essence, inference of an interaction between two proteins is made on the measured data, and such inference can either be right or wrong. Most publicly available data are stored as positive measured results, and therefore, most analyses have employed the most obvious method to infer interactions: a positive observation indicates an interaction while lack of observation indicates no interaction. This method, while useful and sometimes unavoidable, does not make use of other indicators for the presence or absence of interactions. The most useful and yet seldom used indicator is the information about which set of interactions were tested. As mentioned, most studies report positively measured interactions but few report the negative measurements. It is quite often the case that untested protein pairs and negative measurements are not distinguished. A second possible indicator of the presence of an interaction is reciprocity. bait-to-prey systems allow for the testing of an interaction between a pair of proteins in two directions. If bi-directionally tested, we anticipate the result as both positive or both negative. Failure to attain reciprocity indicates some form of error. A third indicator is the type of interaction being assayed; direct physical interactions need to be differentiated from indirect interactions, and this difference plays an important role in inference. In the Y2H system, two proteins are modified so that a physical interaction between the two can reconstitute a functioning TF. In AP-MS, a single protein is chosen and modified, and each pull down detects proteins that are in some complex with the selected one but may not necessarily directly interact with the chosen protein.

Restricting our attention to bi-directionally tested interactions, we can use a Binomial model to identify proteins that either find a disproportionate number of prey relative to the number of baits that find them or vice versa. For the AP-MS experiments, there is an association between whether a protein shows this discrepancy and its relative abundance in the cell. For the Y2H system, analyses conducted separately by *Walhout et al.* [51], *Mrowka et al.* [41], and *Aloy and Russell* [52] have reported on this type of artifact and have discussed a likely relationship between it and some bait proteins' propensity to act alone as activators of the reporter gene. Our methods provide a simple test to identify proteins likely affected by such systematic errors. Such diagnostics can aid in the interpretation of the data and in the design of future experiments. By restricting attention to proteins not affected by this artifact, we can refine the error modeling and the subsequent biological analysis.

## 2.3 Results and Discussion

### Tested Interactions and Their Representations

We recall that in Y2H, the bait is the protein tagged with the DNA binding domain, and the prey is the hybrid with the activation domain. Only those constructs resulting in a functional fusion protein will be tested as bait or as a prey. In AP-MS, a piece of DNA encoding a *tag* is inserted into a protein-coding gene so that yeast cells express the tagged protein; these are the baits. The prey are unmodified proteins expressed under the conditions of the experiment. The set of tested baits, even in experiments intended to be genome-wide, can be quite restricted. For example, *Gavin et al.* [27] designed their experiment to employ the 6,466 ORFs that were at that time annotated to the *S. cerevisiae* genome, but successfully obtained tandem affinity purification for 1,993 of those. The remaining 4,473 (69%) failed at various stages, because, for example, the tagged protein failed to express, or the bands resulting from the gel electrophoresis were not well-separated.

It is difficult to give an accurate enumeration of the sets of *tested baits* and *tested prey* in an experiment, and often, the published data are not sufficiently detailed as to identify these sets. As a proxy, we introduce the concepts of *viable baits* and *viable prey*; the first is the set of baits which were reported to have interacted with at least one prey, and the latter is similarly defined. These quantities are unambiguously obtained from the reported data and provide reasonable surrogate estimates for what are the tested baits and tested prey. The set of ordered pairs, one being a viable bait and the other a viable prey, are interactions for which we have a level of confidence that they were experimentally tested and could, in principle, have been detected. The failure to detect an interaction between a *viable bait* and a *viable prey* is informative, whereas the absence of an observed interaction between an untested bait and prey is not. This approach over-emphasizes positive interactions. Potentially valid data on tested proteins that have truly no interactions with any other tested protein will be discarded.

Protein interactions have been generally modeled by ordinary graphs [53]. The proteins correspond to the nodes of the graph, and edges between protein pairs indicate an interaction (either physical interaction or complex co-membership). For measured data from bait-to-prey systems, protein pairs are ordered $(b, p)$ to distinguish a bait $b$ from a prey $p$. There are three types of relationships between protein pairs of an experimental data set: 1. tested with an observed interaction, 2. tested with no observed interaction, and 3. untested. An adequate representation for this type of data would be a directed graph with edge attributes. A directed edge $(b, p)_+$ signals testing with an observed interaction while a directed edge $(b, p)_-$ signals testing without an observed interaction. Interactions between proteins that are not adjacent were not tested. In those cases where all protein pairs were reciprocally tested, we can suppress the $(b, p)_-$ edges, and a directed graph (digraph) is an adequate representation.

As mentioned, information on which protein pairs were tested for an interaction is rarely explicitly published, and so we represent the current data by a directed graph with node attributes. Using *viability* as a proxy for testing, the nodes with non-zero out-degree are presumed to be the set of viable baits, and similarly, the nodes with non-zero in-degree are presumed to be the viable prey. Isolated nodes become identified as the set of untested proteins (both as baits and prey). We make use of such a digraph data structure in this paper (See Figure 2.1).

### 2.3.1 Interactome Coverage

Given the experimental data, one can partition the proteins into four different sets: 1. viable baits only (VB), 2. viable prey only (VP), 3. viable bait/prey (VBP), and 4. the untested proteins. Figure 2.2 shows the proportions of the yeast genome as measured by each experiment. For most experiments, relatively large portions of the proteome were untested by the assay (gray area), thereby rendering an incomplete picture of the overall interactome [16,42,46,54].

We asked whether the sets of viable bait and viable prey showed a coverage bias in

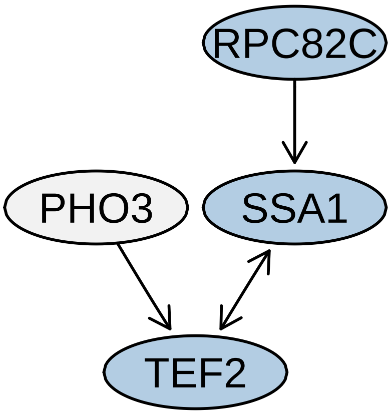


Figure 2.1: Measured protein interaction data are represented by a directed graph. The graph shows the interaction data between four selected proteins from the report by *Krogan et al.* [28]. The bi-directional edge between the ATPase SSA1 and the translational elongation factor TEF2 indicates that either one as a bait pulled down the other one as a prey. The directed edge from RPC82, a subunit of RNA polymerase III, to SSA1 indicates that RPC82 as a bait pulled down SSA1, but not vice versa. Another unreciprocated edge goes from the phosphatase PHO3 to TEF2. An investigation of the data set shows that PHO3, which localizes in the periplasmatic space, was not reported in any interaction as a prey, whereas RPC82C was. In the interpretation of the data, we would have most confidence that there is a real interaction between SSA1 and TEF2. We can differentiate between the two unreciprocated interactions; the one between RPC82C and SSA1 has been bi-directionally tested, but only found once, whereas the other one has only been uni-directionally tested and found.

the experimental assays. Applying a conditional Hypergeometric test [55] on the terms within the Cellular Component branch of Gene Ontology (GO), we found that proteins annotated to categories such as *nucleus* (primarily Y2H), *cytoplasm*, and *protein complex* were over-represented among the viable protein population relative to the yeast genome. This is not surprising since both Y2H and AP-MS assays two kinds of interactions in protein complexes. The Y2H technology is more successful in generating viable proteins within the *nucleus* because this is the cellular location where the test is performed, and so native proteins tend to work more successfully.

The conditional Hypergeometric tests can also identify portions of the cellular

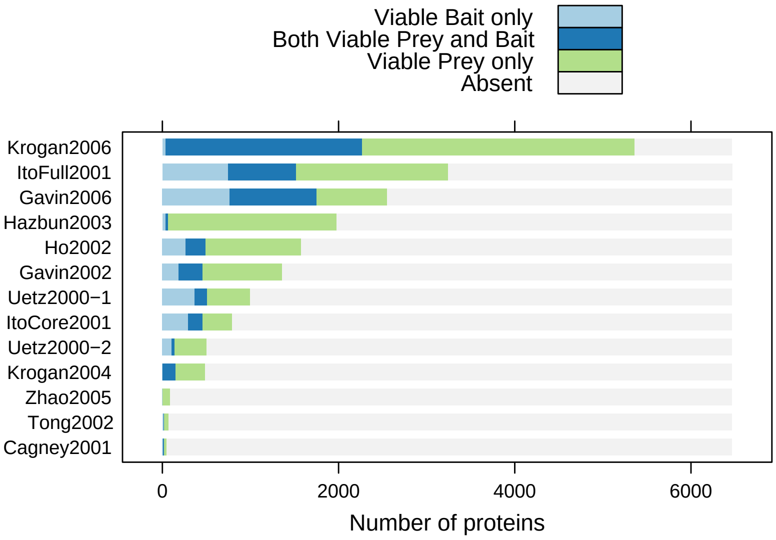


Figure 2.2: Proteins sampled across data sets. This bar chart shows the proteins sampled either as a viable bait (VB), a viable prey (VP), or as both (VBP). With the exception of the data report by *Krogan et al.* [28], the other 11 data sets show large portions of the yeast genome that did not participate in any positive observations. Without additional information, there is little we can do to elucidate whether these proteins were tested but inactive for all tests, or whether these proteins were not tested.

component missed by either Y2H or AP-MS. For the Y2H technology, terms associated with the *mitochondrion*, *ribosome*, and *integral to membrane* were under-represented by viable proteins. Like the Y2H systems, the viable proteins from AP-MS assays were also under-represented with respect to terms associated with *mitochondrion*, *integral to membrane*, but instead of the *ribosome*, AP-MS showed under-representation in the *vacuole*. These under-represented categories are limited by the technologies since all data sets were derived before progress had been made to probe membrane-bound proteins.

Every data set, whether Y2H or AP-MS, showed under-representation for the term *cellular component unknown*. One possible explanation for this phenomenon can be attributed to the correlation between different technologies. It seems that proteins which are problematic in the Y2H and AP-MS systems might also be problematic in systems to determine their cellular localization. Ultimately, further experiments are needed to determine why certain GO categories are under-represented. A gene set enrichment analysis for each data set can be found in Appendix A.

These results point to the fact that the subset of the interactome is either non-randomly sampled or non-randomly covered by the experiment. Either effect limits the type of inference that can be conducted on the resulting data. For instance, inference on statistics such as the degree distribution or the clustering coefficient of the overall graph is less meaningful as long as the direction and magnitude of the coverage or sampling biases are not well understood [18, 56, 57].

### 2.3.2 Systematic Bias – Per-Protein and Experiment-Wide

The interactions between VBP proteins were tested in both directions, and a surprising, yet useful, observation is that there is a large number of unreciprocated edges in the data [52]. These unreciprocated interactions can be used to better understand the experimental errors.

Each VBP protein $p$ has $n_p$ unreciprocated edges, and under the assumption of randomness, we expect the number of unreciprocated in-edges and out-edges to be similar. More precisely, under the assumption that the direction of the edge is random, the number of unreciprocated in-edges is distributed as the number of heads obtained by tossing a fair coin $n_p$ times. Based on this coin-tossing model, we used a per-protein Binomial error model (cf section 2.5) by which to test the statistical significance for the number of unreciprocated in-edges (heads) against the number unreciprocated out-edges (tails). Figure 2.3 shows a partition of the VBP proteins from the data of *Krogan et al.* [28] based on the two-sided statistical test derived from the Binomial model with a p-value threshold of 0.01. Those proteins falling outside the diagonal band are considered to be affected by a systematic bias.

It is interesting to note that the proportion of VBP proteins identified by the Binomial error model as potentially affected by bias is quite small for the Y2H experiments and the smaller-scale AP-MS experiments (less 3%) while the two larger-scale AP-MS experiments showed relatively larger proportions (greater than 14%). It is equally important to note that while these proportions still constitute a minority of VBP proteins, these proteins (within the large-scale AP-MS experiments) participate in a relatively large number of observed interactions, most of which are unreciprocated.

Having identified sets of proteins that are likely to have been affected by this systematic bias, we asked if these proteins could be associated with biological properties. To this end, we fit logistic regression models (cf Appendix A) to predict this effect, and in the AP-MS system, we found evidence that the codon adaptation index (CAI) and protein abundance are associated with the highly unreciprocated in-degree of VBP proteins (i.e. proteins that were found by an exceptionally high number of baits relative to the number of prey they found themselves when tested as baits). The CAI is a per-gene score that is computed from the frequency of synonymous codons in a gene's sequence, and can serve as a proxy for protein abundance [58].

To visualize the association between such proteins and CAI, we plotted diagrams of the adjacency matrix. If the value of CAI is associated with the tendency of a protein having a large number of unreciprocated edges, then we should see a pattern in the adjacency matrix when the rows and columns are ordered by ascending CAI values. We do this for the *Gavin et al.* [27] data in Figure 2.4. We see a dark vertical band in Figure 2.4b representing a relatively high volume of prey activity. There is no corresponding horizontal band in Figure 2.4a which suggests that CAI's relationship to the AP-MS system is primarily reflected in a protein's in-degree.

Next, we standardized the in-degree for each protein by calculating its $z$*-score* (cf section 2.5) and then plotted the distributions of these $z$-scores by their density estimates. Four experiments seem to exhibit particularly distinct distributions (*Ito-Full*, *Ito-Core*, *Gavin et al. 2006*, *Krogan et al. 2006*) [27,28,30] (cf Figure 2.4). The *Ito-Full* [30] data set shows the largest mean (approximately 2-4 times the mean of the other Y2H distributions). This is consistent with reports that there were many auto-activating baits in the *Ito-Full* data sets [52]: If a relatively small number of baits auto-activate, resulting in the cell's expression of the reporter gene, then this artificially increases the number of in-edges for a large number of prey proteins. Auto-activation would cause a shift of the $z$-score distribution in the positive direction. This effect, however, is not seen in the *Ito-Core* data.

While *Ito et al.* have tried to eliminate systematic errors by generating the *Ito-Core* subset of interactions, it is noteworthy to recall that they only used reproducibility as a criterion for validation without considering reciprocity. Consequently, almost half of the reciprocated interactions were not recorded in the *Ito-Core* set. While reproducibility is a necessary condition for validation, it is insufficient because systematic errors are often reproducible.

Among the AP-MS data sets, both *Gavin et al.'s* [27] and *Krogan et al.'s* [28] data display negative means. A possible interpretation for this effect can be attributed to the abundance of the prey under the conditions of the experimental assay. The AP-MS system is more sensitive in detecting the complex co-members of a particular bait than in reverse. For instance, if a lowly expressed protein $p$ is tagged and expressed as a bait and pulls down proteins $p_1, \dots, p_k$ as prey, then the reverse tagging of each protein of $p_1, \dots, p_k$ will have a smaller probability of finding $p$. Even if the low abundance protein $p$ is pulled down in the reverse tagging, the mass spectrometry may fail to detect $p$ within the complex mixture [59, 60]. Both these observations could explain why we observed proteins having an overall slightly higher out-degree than in-degree, and, therefore, an overall slightly negative mean for the $z$-score distribution.

Finally, we wanted to cross-compare the systematic errors between experiments. Only two experiments had sufficient size to give reasonable statistical power. Thus, to compare systematic errors of *Gavin et al. 2006* [27] against those of *Krogan et al. 2006* [28], we generated two-way tables (cf Appendix A). While the concordance is not complete, there is evidence that overlapping sets of proteins are affected. This indicates that both experiment–specific and more general factors could be at work resulting in these unreciprocated edges.

### 2.3.3 Stochastic Error Rate Analysis

There has been confusion in the literature when analyzing error statistics, as different articles have used different definitions for the same statistic. Gentleman and Huber describe the prevelant use of FDR when the FPR is needed [61]. Many statistical models make use of the FPR and less so the FDR; confusion between these two statistics can lead to inaccurate interpretation of the data. Proteins pairs can either interact or not, and so the pairs themselves can be partitioned into two distinct sets: the set of interacting pairs, $I$, and the set of non-interacting pairs, $I^c$. False negative (FN) interactions and true positive (TP) interactions can only occur within the set $I$, and therefore, the false negative probability ($p_{\mathrm{FN}}$) and the true positive probability ($p_{\mathrm{TP}}$) are properties on $I$. Similarly, the false positive ($p_{\mathrm{FP}}$) and true negative ($p_{\mathrm{TN}}$) probabilities are properties on $I^c$ [62]. These standard definitions, along with the values $n = |I|$ and $m = |I^c|$, allow us to set up equations for the values of three random variables: the number of reciprocated edges ($X_1$), the number of protein pairs between which no edge exists ($X_2$), and the number of unreciprocated edges ($X_3$).

$$
\begin{aligned}
E[X_1] &= n\,(1-p_{\mathrm{FN}})^2 + m\,p_{\mathrm{FP}}^2 && (2.1)\\
E[X_2] &= n\,p_{\mathrm{FN}}^2 + m\,(1-p_{\mathrm{FP}})^2 && (2.2)\\
E[X_3] &= 2n\,p_{\mathrm{FN}}(1-p_{\mathrm{FN}}) + 2m\,p_{\mathrm{FP}}(1-p_{\mathrm{FP}}) && (2.3)
\end{aligned}
$$

We recall that if $N$ is the number of proteins then $n + m = \binom{N}{2}$, which is the number of all pairs of proteins. Any two of these three equations imply the third, and therefore, there are three unknowns and two independent equations. By the *Method of Moments* [63], we replace the left hand side of Equations (2.1)–(2.3) with the observed values for the number of reciprocated interactions $(x_1)$, for the number of reciprocally non-interacting protein pairs $(x_2)$, and for the number of unreciprocated interactions $(x_3)$; it follows that knowledge of any one of $(p_{\mathrm{FP}}, p_{\mathrm{FN}}, n)$ yields the other two through an application of the quadratic formula (cf Appendix A). Otherwise, if none of these three parameters is known from other sources, Equations (2.1)–(2.3) define a family of solutions, i.e. a one-dimensional manifold of solutions in a space of three variables (cf Figure 2.6).

The variability, or stochastic error, that affects a bait-to-prey system can thus be characterized by a one-dimensional curve in a three-dimensional space, $\{(p_{\mathrm{FP}}, p_{\mathrm{FN}}, n)\}$, which depends on the experiment and can be estimated from the three experiment-specific numbers $x_1$, $x_2$ and $x_3$. If we can identify portions of the data that appear to be affected by a systematic bias, such as that described in the previous section, we can set these aside and focus the characterization of the experimental errors on the remaining filtered set of data, typically, with lower estimates for $p_{\mathrm{FP}}$ and $p_{\mathrm{FN}}$.

To gain insight on the prevalence of false positive and negative stochastic errors, we calculated estimates of the expected number of FP and FN observations using Equations (2.1)–(2.3) and show the results in Table A.3. The upper part of Table A.3 considers the worst case scenario for false positive errors, setting $p_{\mathrm{FN}} = 0$, and hence assuming that all errors are false positives. We discuss the first row, corresponding to the data of *Ito-Full* [30], as an example. 720 proteins were not rejected in the two-sided Binomial test, and there are $\binom{720}{2} = 258,840$ protein pairs, excluding homomers. This gives us an upper limit for $m$. From the solution manifold shown in Figure 6d, we see that an estimate for $p_{\mathrm{FP}}$ is approximately 0.0008. From this it follows that the expected number of unreciprocated FP interactions is 414 and of reciprocated FP interactions is 0.17. The actual data contain 435 unreciprocated interactions and 68 reciprocated ones. So, even in the estimated worst case when all errors are false positive observations, all reciprocated observations are still most likely due to true interactions.

It is useful to contrast the nature of the stochastic error rates. From Figure 2.6, the solution curves gives an estimate for the $p_{\mathrm{FP}}$ rate at 0.0008 conditioned on the *Ito-Full* VBP data and conditioned on $p_{\mathrm{FN}} = 0$; a similar estimate for the *Ito-Core* data set yields $p_{\mathrm{FP}}$ at 0.0025. The reason for this is because the number of non-interacting protein pairs in the former is estimated to be approximately $250,000$ while this number is $8,000$ for the latter. Table A.3 shows that the number of expected false positively identified unreciprocated interactions for *Ito-Full* is 414 and for the *Ito-Core* is 41. Thus, while the $p_{\mathrm{FP}}$ rate of *Ito-Full* is three times smaller than that of *Ito-Core*, the expected number of of falsely discovered interactions is an order of magnitude greater. Therefore, an interaction contained within *Ito-Core* is much more likely to be true than one from *Ito-Full*. Comparing the $p_{\mathrm{FP}}$ rate from *Ito-Full* with the $p_{\mathrm{FP}}$ rate from *Ito-Core* is unreasonable when the underlying sets of non-interacting proteins pairs are entirely different. The false discovery rate is more intuitive, and this statistic has often been confused in the literature with the false positive rate.

We also considered the worst case scenario for false negative errors. By setting $p_{\text{FP}} = 0$, we calculated the expected number of unreciprocated and reciprocated false negatives in the absence of false positive errors. These numbers are in the lower part of Table A.3. Due to the size of $p_{\text{FN}}$, we find that a large number of protein pairs between which no edge was reported in either direction may still, in truth, interact.

Ultimately, an observed unreciprocated interaction in the data indicate that either a false positive or a false negative observation was made. Computational models cannot definitively conclude which of these two occurred, but these models indicate the magnitude and nature of the problem and can be used to compare experiments, as those with relatively higher error rates should be discounted in any downstream analyses.

## 2.4 Conclusion

We have shown that protein interaction data sets can be characterized by three traits: the coverage of the tested interactions, the presence of biases in the assay that systematically affect certain subsets of proteins, and stochastic variability in the measured interactions. In turn, these three characteristics can benefit the design of future protein interaction experiments.

The set of interactions tested is important since data sets usually report positive results, but tend to be ambiguous on the significance of the unreported interactions: is it because the interaction was tested and not detected, or because it was not tested in the first place? Distinguishing the two cases is important for inference and for integration across data sets. For the currently available data sets from Y2H and AP-MS, a practical estimate of what is the set of tested interactions are all pairs of tested bait and tested prey. A comprehensive list of tested proteins is usually not published. We can, however, obtain a useful approximation for the tested baits and prey using the notion of *viability*. This assumption, however, does introduce some bias, especially for experiments with relatively few bait proteins, as proteins that were tested, but did not interact with any bait protein will not be counted, falsely raising the proportion of interactions. On the other hand, when complete data are not reported, the presumption that interactions were tested, when they were not, introduces bias in the other direction.

There has been substantial interest in cross-experiment analysis, or in integrating data from multiple sources [19, 41, 44, 45, 50]. The possible pitfalls of naive comparisons between two experimental data sets are depicted in Figure 2.7. The interactions in the intersection of the rectangles (red) were tested by both; the interactions in the green and purple areas were tested by one experiment but not the other; and the interactions in the light grey areas were tested by neither experiment. Any data analysis that does not keep track of these different coverage characteristics risks being misled. Therefore, coverage must be taken into consideration when integrating and comparing multiple data sets. Additionally, systematic bias due to the experimental assay affects the detection of certain interactions between protein pairs, and these systematic errors should be isolated from the data set before estimating stochastic errors. Ultimately, there are still many more steps needed to integrate data sets, and we discuss a few necessary components.

If the assay system were perfect, all bi-directionally tested protein pairs would be reciprocally adjacent or not. In practice, unreciprocated edges are observed and can be used to better understand the sources of error. Measurement error can be divided into two categories: systematic and stochastic. We have shown that there are proteins with an inordinate imbalance between unreciprocated in- and out-edges, and they behave in a systematically different way when used as a bait than when found as a prey. This is an indication that the interaction data involving these proteins contain either a large number of false positives or false negatives. Further data are needed to differentiate between these two alternatives. The mode in which they fail is distinct from the unspecific stochastic errors that we model via the FP and FN rates, and hence, they should be excluded from these analyses.

It is useful to distinguish between the concepts of *stochastic* and *systematic* measurement error. Systematic errors are due to imperfections or biases of the experimental system, and they occur in a correlated or reproducible manner. Stochastic errors occur at random in an irreproducible manner; in principle, they can be *av-*

*eraged* out by repeating the experiment often enough. There are many benefits to an analysis that identifies and separates these two types of measurement error. We have identified one type of systematic error in bait-to-prey systems which seems to be associated with artifacts of the technologies.

The occurrence of unreciprocated edges also points to some of the aspects of the technologies that could be improved. In AP-MS experiments, this artifact shows a strong association to CAI and protein abundance. Because mass spectrometry techniques are known to be, at times, less sensitive in identifying proteins with low abundance in a complex mixture, refinements of such methodology could potentially yield more accurate measurements.

The methods we have described are useful towards future application of Y2H or AP-MS. Newer experiments can, and should, take into consideration relative protein abundance when assaying protein interactions. Besides this, the GO category analysis for under-representation (Appendix A) shows certain proteins and protein complexes which do not work as intended under the conditions of the assay system. Knowing which categories are under-represented allows experimenters to adjust the technologies or create new technologies (such as the Y2H test for membrane-bound proteins [64]).

These elementary questions of data preprocessing, quality assessment and error modeling may seem far removed from the systems-level modeling of biological systems. Such a modeling, however, requires the use and integration of multiple different data sets, to increase the breadth and depth of the data compared to those from a single experiment. This can only be done if the error statistics and possible patterns in the errors are sufficiently understood. We believe that the methods and tools developed in this work provide a step in this direction.

## 2.5 Materials and Methods

### 2.5.1 Protein Interaction Data

We investigated twelve publicly available data sets for *S. cerevisiae* of which seven were assayed by Y2H and five were assayed by AP-MS. We obtained [15, 27, 30–34] from the *IntAct* [65] repository and [23, 24, 26, 28] from their primary sources. All data sets have two key properties: 1. information on the bait-to-prey directionality is retained and 2. the prey population is documented as genome-wide. A table with an overview of the data sets can be found in *Appendix A*.

### 2.5.2 Statistical Analysis

**Binomial Error Model - Detecting Bias**

The Binomial error model assumes that in- and out-degrees are equally likely among unreciprocated edges of a bi-directionally tested protein. Thus, we presume that the number of unreciprocated out-edges for any bi-directionally tested protein $p$ is distributed as $B(n_p, \frac{1}{2})$ where $n_p$ is the total number of unreciprocated edges of $p$. Under this hypothesis, we can compute the p-value for the observed measured directed degree for each protein $p$. The null hypothesis is rejected at the 0.01 threshold. Proteins for which we reject the null hypothesis are deemed likely to be affected by a systematic bias in the assay.

**Multinomial Error Model**

Let $N$ be the number of proteins in an interactome of interest, then the total number of distinct protein pairs, excluding homomers, is $\binom{N}{2}$. Denote the set of all unique interacting protein pairs among the $N$ proteins by $I$ and its complement by $I^c$. Recall that $\binom{N}{2} = n + m$ where $n = |I|$ and $m = |I^c|$.

Only two of the three Equations (2.1)–(2.3) are independent, any two of them imply the third. Explicitly, because there are $\binom{N}{2}$ total number of possible interactions, knowing (for example) the expected number of reciprocated and unreciprocated interactions would clearly imply the number of reciprocated non-interacting pairs. We parameterize the one-dimensional solution manifold by $n$ $(0 \leq n \leq \binom{N}{2})$. Relevant solutions are those for which $0 \leq p_{\mathrm{FP}}, p_{\mathrm{FN}} \leq 1$. Consider $n$ given, then we can solve for $p_{\mathrm{FN}}$ in terms of $p_{\mathrm{FP}}$:

$$p_{\mathrm{FN}} = \frac{1}{2n}(\Delta + 2mp_{\mathrm{FP}}) \tag{2.4}$$

where we have defined $\Delta = (x_2 - m) - (x_1 - n)$. Here, $x_1$ is the observed number of reciprocated interactions and $x_2$ is the number of reciprocated non-interacting protein pairs. Making a substitution for $p_{\mathrm{FN}}$ in Equation (2.2), the problem reduces to a quadratic equation in one parameter, $p_{\mathrm{FP}}$:

$$\underbrace{(n+m)}_{a} p_{\mathrm{FP}}^2 + \underbrace{(\Delta - 2n)}_{b} p_{\mathrm{FP}} + \underbrace{n + \frac{\Delta^2}{4m} - \frac{n}{m} x_2}_{c} = 0. \tag{2.5}$$

An application of the quadratic formula gives two solutions for $p_{\mathrm{FP}}$:

$$(p_{\mathrm{FP}})_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \tag{2.6}$$

where $a$, $b$, and $c$ are as defined in Equation 2.5. Then substituting an estimate of $p_{\mathrm{FP}}$ back into Equation 2.4 gives a solution for $p_{\mathrm{FN}}$.

A similar argument carries through given any one of the three parameters $\{n, p_{\mathrm{FP}}, p_{\mathrm{FN}}\}$. Thus an estimate of one of the parameters generates estimates of the other two.

### Conditional Hypergeometric, Logistic Regression Tests, and Two-Way Tables

We grouped the yeast genome into several defined subsets (VB, VP, VBP, and those proteins appearing to be affected by bias), and we wanted to determine whether or not the subsets showed over/under representation among biological categories such as GO, KEGG, and Pfam. We used the conditional Hypergeometric testing as described by [55] to probe for such over/under representation at a p-value threshold of 0.01. A list of such GO categories and Pfam domains can be generated by the R scripts `hgGO.R` and `hgPfam.R` contained within the `Bioconductor` software package `ppiStats`.

For those proteins affected by a systematic bias of each experiment, we fitted a logistic regression on these sets against 31 protein properties reported in [66] and set a p-value threshold at 0.01.

Let $S_i$ be the set of proteins identified to be affected by a systematic bias in data set $i$, and suppose we want to compare $S_i$ against $S_j$; we define two methods of generating $S_i$ and $S_j$ for such a comparison. One method is the application of the Binomial test on the $VBP$-subgraph of each data set $i$ exclusively to determine each $S_i$. The second method aims to streamline the experimental conditions of $i$ with that of $j$. First, we compute $X = VBP_i \cap VBP_j$; then we apply the Binomial test on the $X_i$-subgraph as well as the $X_j$-subgraph (since the edge-sets will be different). Obtaining such subsets allows us to generate a two-way table, $T$, to compare $S_i$ against $S_j$. If the first method is used to generate the subsets $S_i$ and $S_j$, then we must still restrict to $X$ when computing $T$.

$$\begin{aligned}
T_{(2,2)} &= \mid S_i \cap S_j \mid \\
T_{(1,2)} &= \mid S_i \setminus S_j \mid \\
T_{(2,1)} &= \mid S_j \setminus S_i \mid \\
T_{(1,1)} &= \mid S_i^c \cap S_j^c \mid
\end{aligned}$$

We can apply Fisher's Exact Test to test the independence of these two sets at a designated p-value threshold.

### Per Protein In-Degree $z$-score and Cross Experimental Comparisons

Let $o_p$ be the unreciprocated out-degree for a protein $p$ and $i_p$ its unreciprocated in-degree. Then denote the number of unreciprocated edges by $n_p = i_p + o_p$. Assuming the distribution $i_p \sim B(n_p, \frac{1}{2})$, we can compute the standardized in-degree ($z$-score) for $p$:

$$z_p = \frac{i_p - o_p}{\sqrt{i_p + o_p}} \tag{2.7}$$

### Estimating the Number of Stochastic FP/FN Observations

We used the filtered data after setting aside proteins rejected by the two-sided Binomial tests to calculate the results of Table A.3. In the first case, we set $p_{\mathrm{FN}} = 0$, and $p_{\mathrm{FP}}$ is the maximal value in the solution curve shown in Figure 2.6. $m$ is estimated as $\binom{N}{2}$. The expected number of unreciprocated FP observations is

$p_{\text{FP}}(1 - p_{\text{FP}})m$ and of reciprocated FP observations $p_{\text{FP}}^2 m$. In the second case, we set $p_{\text{FP}} = 0$ and obtain $n$ from the solution curve. The expected number of unreciprocated FN observations is $p_{\text{FN}}(1 - p_{\text{FN}})n$ and of reciprocated FN observations $p_{\text{FN}}^2 n$.

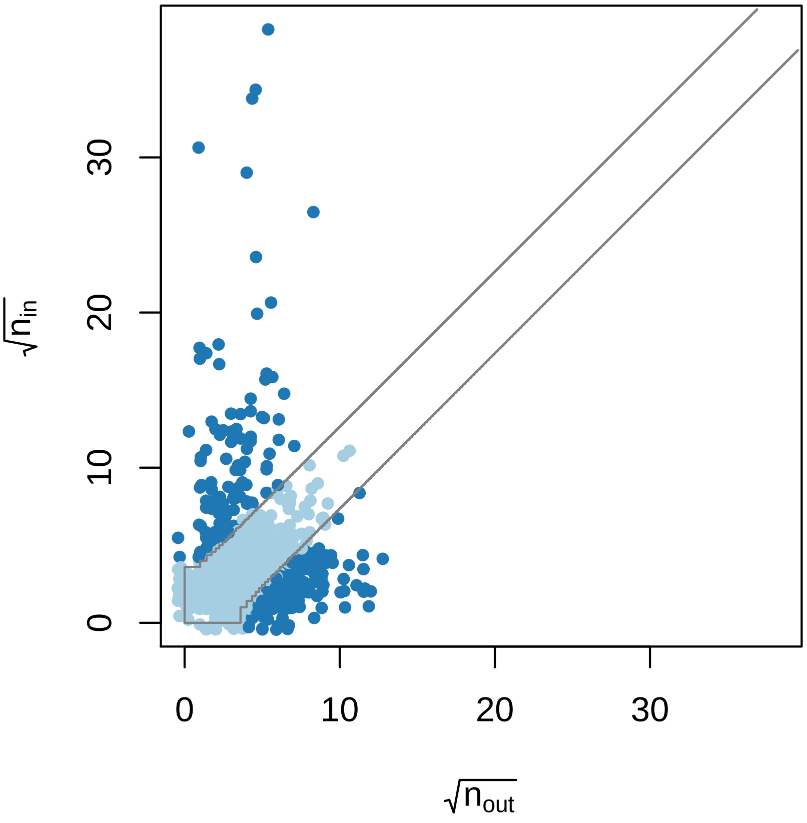


Figure 2.3: Two-sided binomial test on the data from *Krogan et al.* [28]. The scatter-plot shows the out-degree and in-degree, $n_{\text{out}}$ (or denoted $o_p$) and $n_{\text{in}}$ (also $i_p$), for each protein in the set of viable bait/prey from the report by *Krogan et al.* [28] (axes are scaled by the square root). The proteins that fall outside of the diagonal band exhibit relatively skewed directed degree (skewed in either in or out-degree). This figure shows a graphical representation of a two-sided binomial test. The points above and below the diagonal band are proteins for which we reject the null hypothesis that the distribution of unreciprocated edges is governed by $B(n_p, \frac{1}{2})$. For the purpose of visualization, small random offsets were added to the discrete coordinates of the data points by the R function `jitter`.

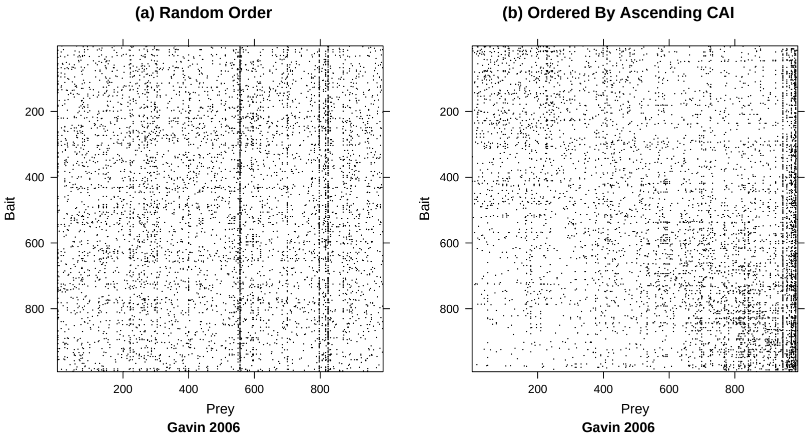


Figure 2.4: Adjacency matrices: random versus ascending codon adaptation index (CAI). These plots present a view of the adjacency matrix for the viable bait/prey (VBP) derived from the report from *Gavin et al.* [27]. An interaction between bait $b$ and prey $p$ is recorded by a dark pixel in $(b, p)^{th}$ position of the matrix. (a) Rows and columns are randomly ordered; (b) rows and columns are ordered by ascending values of each protein's CAI. Contrasting these two figures, we can ascertain that there is a relationship between bait/prey interactions and CAI. The relationship is based on proteins with large un- reciprocated in-degree because panel b shows a dark vertical band. Had unreciprocated out-degree also been associated with CAI, then there would be a similar horizontal band reflected across the main diagonal of the matrix.

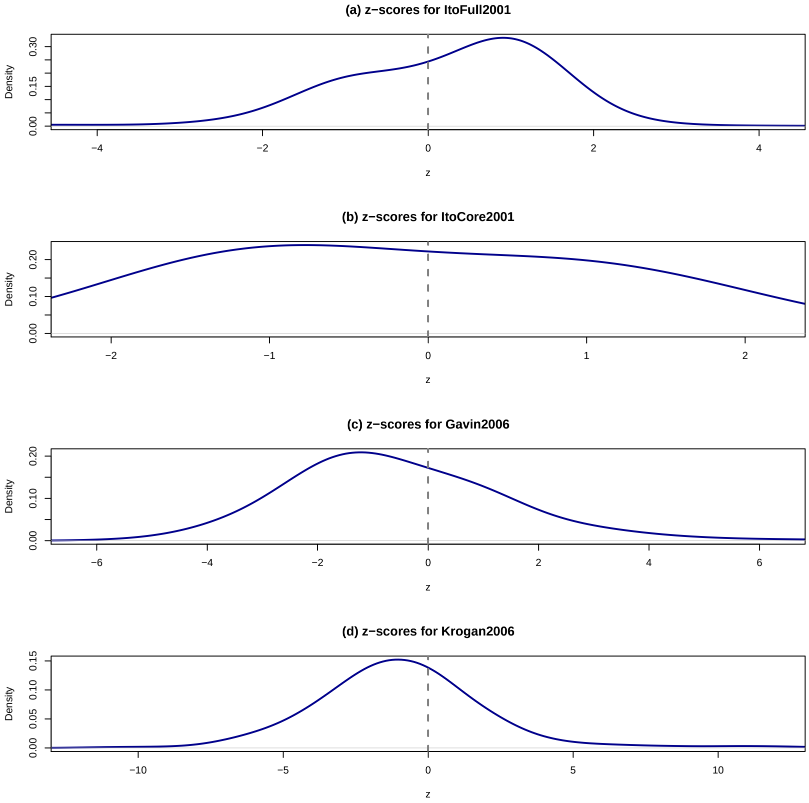


Figure 2.5: Density plots of the in-degree $z$-scores. The plots show the density estimates of the in-degree $z$-scores for three data sets [27, 28, 30]. The zero line is present to distinguish between positive and negative $z$-scores. The distribution reported by *Ito et al.* [30] shows a high concentration of data points that have positive $z$-scores, whereas the data reported by textitGavin et al. [27] and *Krogan et al.* [28] have maximal density for negative z. Systematic artifacts such as auto-activators in the yeast two-hybrid (Y2H) system and protein abundance in affinity purification-mass spectrometry (AP-MS) might play a role in off-zero mean of these density plots. Restricting to the Ito-Core set appears to eliminate the effect from the Ito-Full set.

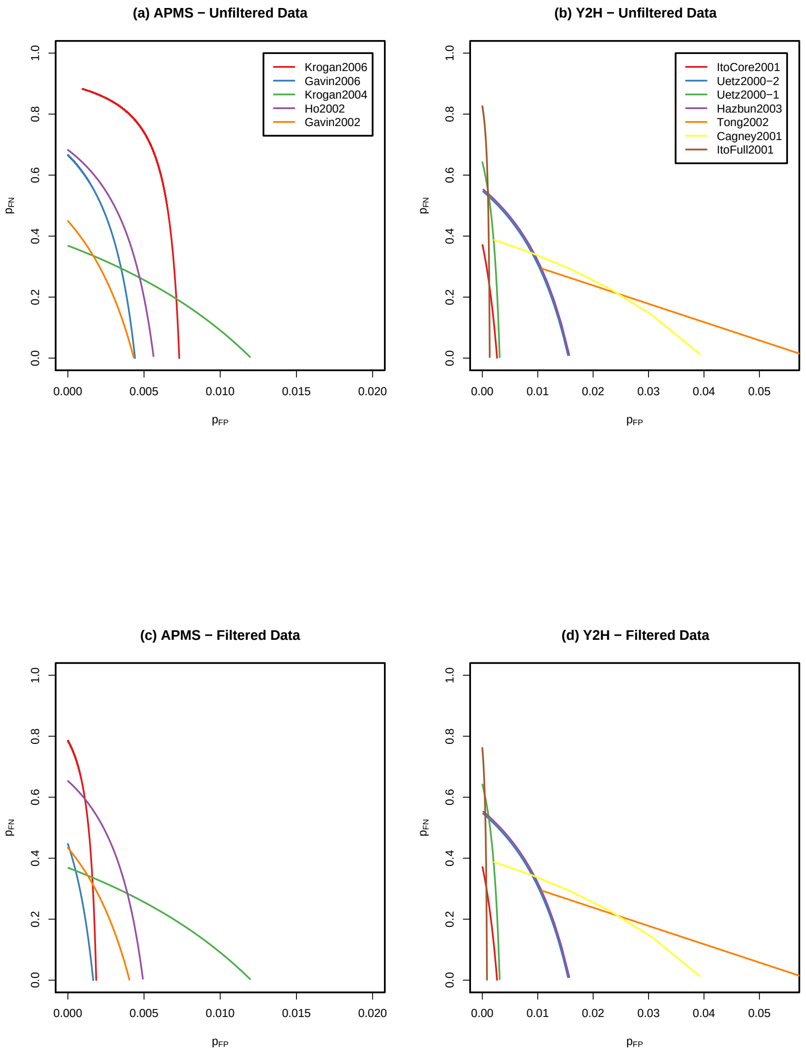


Figure 2.6: Geometric visualization of the solution curves from the algebraic equations (1) to (3). (a) Plot of ($p_{\mathrm{FP}}$, $p_{\mathrm{FN}}$) parameterized by $n$ for the APMS data sets and (b) Y2H datasets. (c) APMS data filtered for the proteins rejected by the binomial test and (d) Y2H data with the analogous filters. These curves give upper bounds for the values of ($p_{\mathrm{FP}}$,$p_{\mathrm{FN}}$) in the multinomial error model for each experiment. Each point on any of the curves represents 3 distinct values restricted to the VBP proteins: the true number of interactions between the VBP proteins, the $p_{\mathrm{FP}}$ rate, and the $p_{\mathrm{FN}}$ rate. Any one of these values determines the other two.

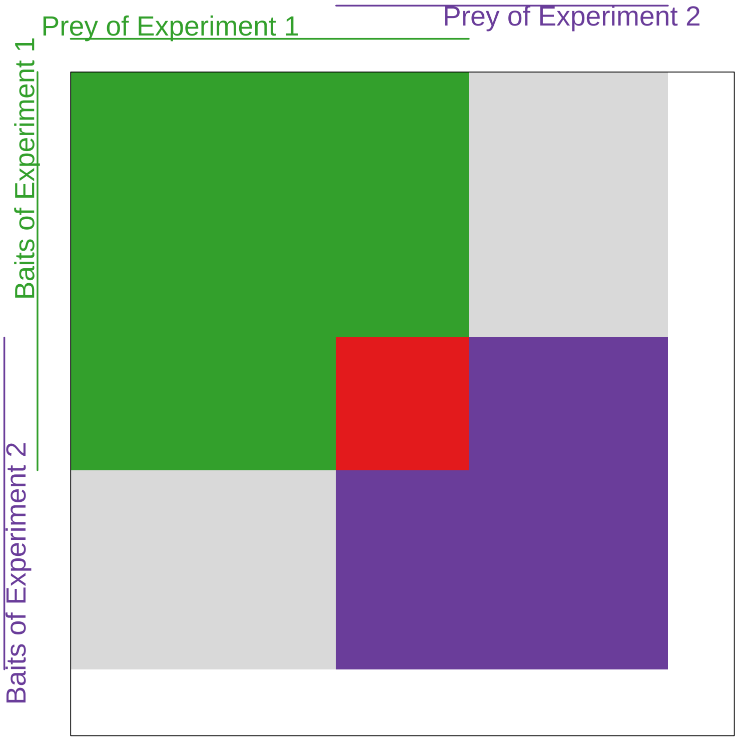


Figure 2.7: Matrix representation on two separate bait-to-prey data sets. A schematic representation of the interactome coverage of two protein interaction experiments. The adjacency matrix of the complete interactome is represented by the large square. Experiment 1 covers a certain set of proteins as baits (rows covered by the green vertical line) and as prey (columns covered by the green horizontal line). The tested interactions for experiment 1 are contained within the green rectangle. Similarly, experiment 2 covers another set of proteins and tests for a set of interactions contained in the purple rectangle. In the intersection of the rectangles, the red area, are the bait-to-prey interactions tested by both experiments, and in the union are the interactions tested by at least one of the experiments. Note that the interactions in the light gray area were tested by neither experiment, either because there are missing tested prey (upper right corner) or missing tested baits (lower left corner). The interactions in the white region are also tested by neither experiment because the baits and the prey were not tested.

# Chapter 3

# Modeling Interactions in Bait-Prey Systems

Summarizing global and local properties of the set of protein interactions, the *interactome*, is necessary for describing cellular systems. We consider a relatively simple per-protein feature of the interactome: the number of interaction partners for a protein, which in graph terminology is the *degree* of the protein. Using data subject to both stochastic and systematic sources of false positive and false negative observations, we develop an explicit probability model and resultant likelihood method to estimate node degree on portions of the interactome assayed by bait-prey technologies. This approach yields substantial improvement in degree estimation over the current practice which naïvely sums observed edges. Accurate modeling of observed data in relation to true but unknown parameters of interest gives a formal point of reference from which to draw conclusions about the system under study.

## 3.1 Introduction

Understanding the contribution of proteins to cellular systems requires knowledge of the various interactions they have with other proteins [15, 23, 24, 26–28, 30]. Global and local statistics on the topology of *interactome* graphs aim to infer the nature and behavior of protein interactions, and can provide a basis for planning and interpretation of experiments [67]. These measures capture simple but informative features of graphs, for example the number of interactors for each protein. There are caveats to the use of these measures, including the dynamic nature of the *in vivo* interactome as opposed to the static data yielded by currently available technologies [61]. But these summary statistics capture characteristics of high-throughput observations in tractable form and accurate estimation is paramount to making correct conclusions about interactome behavior.

We understand that the APMS and Y2H technologies probe undirected, symmetric relationships in a directed manner from bait-to-prey. Inference on the features of APMS and Y2H graphs is complicated by this directedness, as well as sampling bias, incomplete coverage, and stochastic and systematic errors leading to both false positive (FP) and false negative (FN) observations. These issues have often been overlooked when analyzing protein interaction data [1]. In this study, we apply a statistical modeling approach to ameliorate these difficulties for APMS and Y2H.

This report specifically demonstrates the use of statistical likelihood for estimating node degree to obtain a substantial improvement over the naïve approach in which degree is estimated by summing observed interactions. The implications are widespread since many other graph statistics, for example the clustering coefficient and node degree distribution, are functions of node degree and special biological interpretations are often assigned to nodes of particularly high degree. Furthermore, knowing the number of interactions for any protein is helpful for identifying its true interactors given a set of reported possibilities. Node degree estimation is one example of interactome-based statistical modeling. The paradigm we propose applies generally to other bait-prey graph statistics and is critical for accurately describing interactome behavior.

## 3.2 Results

### 3.2.1 Multinomial Model for Node Degree

In concept, the set of edges in a bait-prey graph can be divided into distinct sets of doubly, singly, and untested edges. If all experiments for each bait work properly and all proteins in the cell are available for detection as prey, then all edges between pairs of baits are tested twice, all edges extending from baits to non-bait proteins are tested once, and all edges between pairs of non-baits are untested. Under realistic experimental conditions, some proteins fail as baits and some proteins are not detectable as prey, and so the distinctions between these collections of edges blur. [1] discuss a viable-bait-prey (VBP) subgraph of a full set of bait-prey experimental data induced by the subset of bait proteins that detect at least one prey and are detected as prey by at least one other bait. By focusing on the set of proteins with direct evidence of viability as both bait and prey and by eliminating proteins prone to systematic bias (see Section 1.3.3), the VBP graph only includes edges for which two bait-prey assays running in opposite directions between a pair of proteins can reasonably be viewed as replicate observations on the same underlying true edge. Different experimental conditions, and other factors, dictate that VBP nodes may differ even for experiments using the same original bait set (see Section 3.2.4).

Despite replicate testing of all edges in the VBP graph, the observations in each direction are not necessarily consistent due to measurement error. For each VBP node there is an observed number of reciprocated edges, $r$, an observed number of unreciprocated edges, $u$ (see Figure 3.1), and a true but unknown degree $\Delta$. The joint probability of observing specific values of $r$ and $u$ for any given $\Delta$ can be written as a function of $\Delta$, the TP probability ($p_{\text{TP}}$) and the FP probability ($p_{\text{FP}}$) using Multinomial models for both TP and FP observations; full model development is reported in Section 3.3.1. After adjusting the probability statement for only observing non-zero counts of interactions, statistical maximum likelihood estimation (MLE) can be used to arrive at a degree estimator, $\hat{\Delta}_{\text{MLE}}$, that accounts for restrictions on observed data in the VBP graph as well as $p_{\text{TP}}$ and $p_{\text{FP}}$. Current practice is to estimate degree naïve to $p_{\text{TP}}$, $p_{\text{FP}}$ and subtleties in data collection,

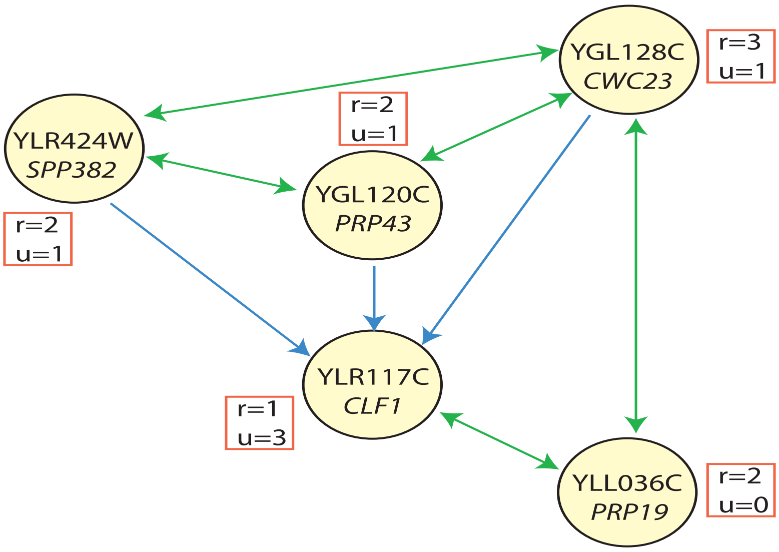


Figure 3.1: A subgraph of VBP nodes from the Gavin et al. (2006) APMS data. Green reciprocated edges were tested twice and observed twice. Blue unreciprocated edges were tested twice and observed once. All edges not shown were tested twice and not observed twice.

specifically $\hat{\Delta}_{\text{naïve}} = r + u$.

### 3.2.2 Estimating Degree when $p_{\text{TP}}$ and $p_{\text{FP}}$ are Known

**Local performance under misspecification of $p_{\text{TP}}$**

Calculation of $\hat{\Delta}_{\text{MLE}}$ depends on values of $p_{\text{TP}}$ and $p_{\text{FP}}$, and since true values of $p_{\text{TP}}$ and $p_{\text{FP}}$ are not generally known for any particular technology, they must also be estimated. Given the potential for misspecification of $p_{\text{TP}}$ and $p_{\text{FP}}$, we studied the accuracy of $\hat{\Delta}_{\text{MLE}}$ under deviations from the truth for these parameters. Since $p_{\text{TP}}$ applies to a small number of true edges relative to the total number possible, its effects are most evident at the per-node level.

For nodes with $\Delta$ ranging from 1 to 20, 500 observations were generated from graphs with 1000 nodes, doubly tested edges, $p_{\text{TP}}$=0.6, 0.7, and 0.8, and $p_{\text{FP}} = 0.001$. This $p_{\text{FP}}$ parameter results in a mean of 2 FP observations per node with FP observations occurring in either direction. To simulate the VBP paradigm, observations of zero incident edges were excluded from analysis. Given the simulated observations, $\hat{\Delta}_{\text{MLE}}$ was estimated under the true $p_{\text{TP}}$ and $p_{\text{FP}}$ parameters as well as incorrect assumed values for $p_{\text{TP}}$. Figure 3.2 demonstrates the mean relative error for $\hat{\Delta}_{\text{MLE}}$ using both correct and misspecified values of $p_{\text{TP}}$. For these simulation parameters, relative bias in the estimator corresponds roughly with the amount of under- or overestimation of $p_{\text{TP}}$; when $p_{\text{TP}}$ is under- or overestimated by 0.10 (0.20), degree is estimated on average within ten (twenty) percent of the true degree. Greater relative bias is apparent for nodes of lower degree.

### Global performance under correct specification of $p_{\text{TP}}$ and $p_{\text{FP}}$ for varying graph topologies

A series of Erdös-Renyi (Er) random graphs containing 1000 nodes and 2000 edges were examined to explore global performance of $\hat{\Delta}_{\text{MLE}}$ versus $\hat{\Delta}_{\text{naïve}}$. For 100 of these ER graphs, edges were doubly tested and observed with $p_{\text{TP}} = 0.6$ and 0.7 and $p_{\text{FP}} = 0.0008$ and 0.001. Table 3.1 records mean, minimum, and maximum observations of the square root of the mean squared errors (RMSEs) for the naïve and MLE estimates on each of the 100 generated graphs. Interestingly, as $p_{\text{TP}}$ increases from 0.6 to 0.7 (i.e. the technology is more sensitive), RMSE for $\hat{\Delta}_{\text{naïve}}$ also increases for equal values of $p_{\text{FP}}$. The naïve approach to degree estimation simply adds FPs to TPs, hence an increase in sensitivity can lead to over-estimation of degree depending on the number of FPs per node. On the other hand, as $p_{\text{TP}}$ increases, RMSE for $\hat{\Delta}_{\text{MLE}}$ decreases, indicating improved estimation of degree for more sensitive technology as would be expected. Also of interest in Table 3.1 is the notable increase in RMSE for $\hat{\Delta}_{\text{naïve}}$ as $p_{\text{FP}}$ increases from 0.0008 to 0.001 for equal values of $p_{\text{TP}}$. In contrast, $\hat{\Delta}_{\text{MLE}}$ accounts for these FP observations and shows only a modest increase in RMSE for the larger value of $p_{\text{FP}}$.

Much debate has centered on whether graphs exhibit node degree distributions with heavier tails or higher variability than ER random graphs [68]. To explore the performance of $\hat{\Delta}_{\text{MLE}}$ in this setting, we generated a series of graphs with 1000 nodes and 1000 edges according to the preferential attachment model of [69] and observed edges according to $p_{\text{TP}} = 0.7$ and $p_{\text{FP}} = 0.001$. Log-log plots

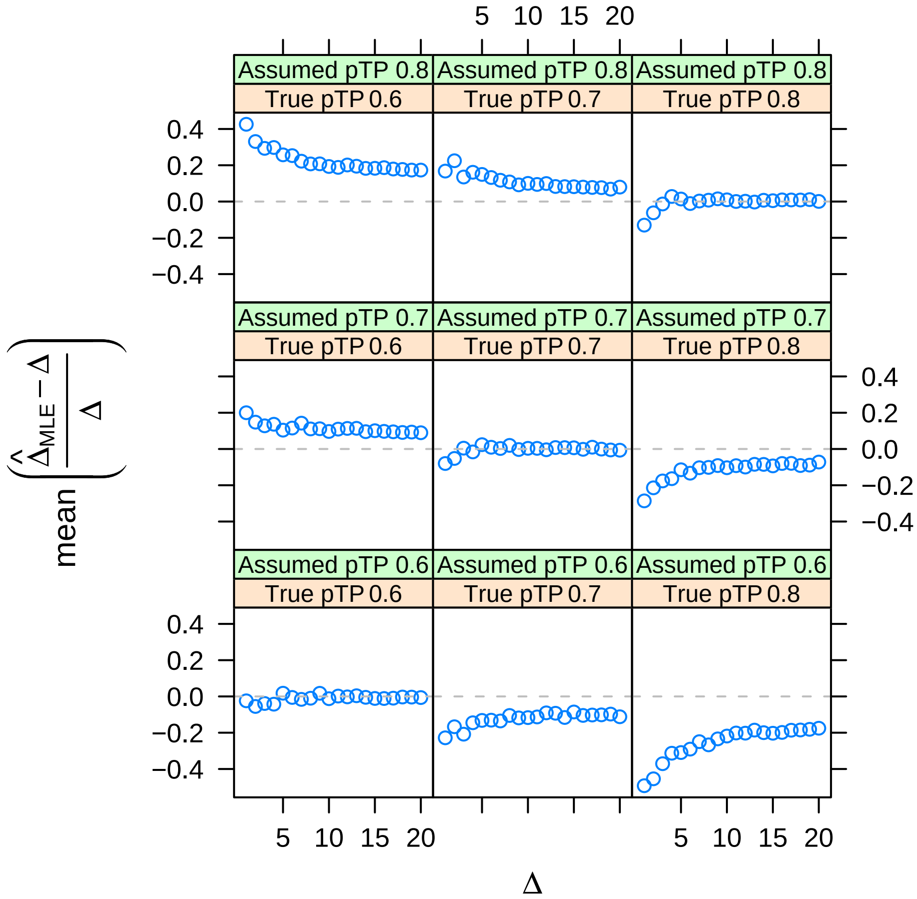


Figure 3.2: Mean relative bias for $\hat{\Delta}_{\text{MLE}}$ for 500 observations from graphs with 1000 nodes calculated as the mean difference between $\hat{\Delta}_{\text{MLE}}$ and $\Delta$, divided by $\Delta$. True $p_{\text{TP}}$ values are those used to generate observations from the simulated graphs and the assumed $p_{\text{TP}}$ values are those used in estimation of degree. Incorrect, assumed $p_{\text{TP}}$ values are used to study the accuracy of $\hat{\Delta}_{\text{MLE}}$ under $p_{\text{TP}}$ misspecification. In all simulations here, $p_{\text{FP}} = 0.001$ and is correctly specified for estimation purposes. Each panel represents a single simulation. Assumed $p_{\text{TP}}$ is constant within rows, and true $p_{\text{TP}}$ is constant down columns.

depicted in Figure B.1 demonstrate that the distribution of $\hat{\Delta}_{\text{MLE}}$ more closely resembles the true distribution than that of $\hat{\Delta}_{\text{naïve}}$. Goodness-of-fit of the naïve and MLE distribution estimates was also assessed using RMSE, this time comparing the true probability mass function with the estimated values at each degree. The mean ratio of RMSE for the $\hat{\Delta}_{\text{MLE}}$ and $\hat{\Delta}_{\text{naïve}}$ estimates on each graph was 0.716 (minimum=0.598, maximum=0.796), indicating a consistent reduction in RMSE between 20 and 40 percent when using the MLE approach. Large variability of degree is a general property of graph data [68], and these simulations suggest that the MLE approach improves coverage of the full range of true node degrees, particularly in the extremes of the distribution.

### 3.2.3 Estimating $p_{\text{TP}}$ and $p_{\text{FP}}$ Using a Gold Standard

For real data analysis, estimation of $\Delta$ requires estimates of $p_{\text{TP}}$ and $p_{\text{FP}}$ since their true values are generally unknown. A variety of techniques have been discussed for estimation of $p_{\text{TP}}$ and $p_{\text{FP}}$ [70–73] but these do not directly account for bait-prey viability in an experiment. The method we develop here for AP-MS data first aligns protein complex viability in the gold standard with the data under study. The observed interactions are then compared to the viable complexes in the gold standard to estimate values of $p_{\text{TP}}$. Results are reported here for five AP-MS data sets on *Saccharomyces cerevisiae*: *Gavin et al. 2002* [74], *Ho et al. 2002* [75], *Krogan et al. 2004* [26], *Gavin et al. 2006* [27], and *Krogan et al. 2006* [28].

Our source of candidate gold standard complex co-memberships is a collection of 335 protein complexes culled from the Munich Information Center for Pro-

| | $p_{\text{FP}}$ | $p_{\text{TP}} = 0.60$ | $p_{\text{TP}} = 0.70$ |
|---|---|---|---|
| $\hat{\Delta}_{\text{naïve}}$ | 0.0008 | 1.77 (1.61,1.91) | 1.86 (1.66,2.00) |
| | 0.001 | 2.11 (1.91,2.30) | 2.24 (2.04,2.42) |
| $\hat{\Delta}_{\text{MLE}}$ | 0.0008 | 1.63 (1.50,1.75) | 1.44 (1.34,1.52) |
| | 0.001 | 1.76 (1.62,1.88) | 1.48 (1.37,1.58) |

Table 3.1: Mean (minimum,maximum) values of RMSE for $\hat{\Delta}_{\text{naïve}}$ and $\hat{\Delta}_{\text{MLE}}$ for 100 ER graphs with 1000 nodes and 2000 edges.

tein Sequences (MIPS) [76] and Gene Ontology (GO) [77], specifically excluding complex estimates based on the high-throughput AP-MS data sets under investigation in this text (see Section 3.3.2). All pairs of proteins jointly annotated in one of these 335 complexes could, in principle, be detected as interactors by AP-MS technology, as long as all members of the complex are viable proteins in the experiment under consideration. For each AP-MS experiment, we determine the subset of the 335 candidates whose constituent members are all reported in the data set as viable prey and, when applicable, viable baits. Given the resultant number of true complex co-memberships and the observed data, a slightly modified version of the probability statement in Equation (3.1) can be used to estimate $p_{\text{TP}}$ and its variance for each data set (see Section 3.3.2). Specific estimates of $p_{\text{TP}}$ for each experiment are reported in Table 3.2. Figure B.2 demonstrates the trend in estimated $p_{\text{TP}}$ as the gold standard set of complexes centers on the set of viable proteins in each experiment.

The lack of a robust set of high confidence protein complex '*non*-co-memberships' prohibits estimation of $p_{\text{FP}}$ in the same manner as $p_{\text{TP}}$; however, with an estimate of $p_{\text{TP}}$ in place, a corresponding value for $p_{\text{FP}}$ can be calculated using a method of moments approach (see Section 3.3.3). Table 3.2 records the method of moments $p_{\text{FP}}$ estimates for the five AP-MS data sets. The effects of the specific estimates on node degree are discussed in Section 3.2.4.

### 3.2.4 AP-MS Data Results

We estimated node degrees via the statistical likelihood for the five *Saccharomyces cerevisiae* AP-MS data sets. As stated in Section 3.2.1, the Multinomial model assumes that baits prone to various types of systematic bias have been excluded from analysis and only stochastic errors, globally applicable to all VBP proteins, remain. We diagnosed systematically biased VBP proteins by examining the distribution of unreciprocated in- and out-edges as in [1], eliminating those with severe imbalance from further analysis. Table B.1 records the number of baits and prey originally reported for each data set as well as the number of VBP nodes both pre- and post-filtering for systematic bias. MLE estimates reported in Sections 3.2.4 and 3.2.4 were computed using the $p_{\text{TP}}$ and $p_{\text{FP}}$ estimates in Table 3.2.

|  | # VBP nodes | $p_{\text{TP}}$ | $p_{\text{FP}}$ | mean # FPs per node |
|---|---|---|---|---|
| Gavin02 | 268 | 0.63<br>(0.57,0.70) | $1.0 \times 10^{-3}$<br>$(9.0 \times 10^{-06}, 1.8 \times 10^{-3})$ | 0.54<br>$(4.8 \times 10^{-3}, 0.98)$ |
| Ho02 | 226 | 0.67<br>(0.54,0.79) | 3.6e-3<br>$(2.7 \times 10^{-3}, 4.3 \times 10^{-3})$ | 1.6<br>(1.2,1.9) |
| Krogan04 | 149 | 0.84<br>(0.74,0.94) | $3.7 \times 10^{-3}$<br>$(1.3 \times 10^{-3}, 5.6 \times 10^{-3})$ | 1.1<br>(0.4,1.7) |
| Gavin06 | 852 | 0.70<br>(0.66,0.74) | $7.9 \times 10^{-4}$<br>$(6.1 \times 10^{-4}, 9.6 \times 10^{-4})$ | 1.4<br>(1.0,1.6) |
| Krogan06 | 1505 | 0.52<br>(0.46,0.59) | $8.9 \times 10^{-4}$<br>$(6.3 \times 10^{-4}, 1.1 \times 10^{-3})$ | 2.7<br>(1.9,3.3) |

Table 3.2: Number of VBP nodes, estimated values of $p_{\text{TP}}$ and $p_{\text{FP}}$, and mean number of FPs per node for the five AP-MS data sets. Numbers in parentheses below the $p_{\text{TP}}$ estimates are 95% confidence intervals using the variance estimate for $\hat{p}_{\text{TP}}$ discussed in Section 3.3.2. Numbers in parentheses below the $p_{\text{FP}}$ estimates are corresponding method of moments estimates of $p_{\text{FP}}$ for the estimated range of $p_{\text{TP}}$ estimates. The expected number of FPs per node is roughly the product of $p_{\text{FP}}$ and the number of VBP nodes multiplied by two to account for FPs occurring as either in- or out-edges.

### Local analysis

Figure 3.3 plots estimated degree versus $u$ for increasing values of $r$. The left panel reveals the exact linearity of $\hat{\Delta}_{\text{naïve}}$ as $r$ and $u$ increase. The middle and right panels relevant to the Gavin06 and Krogan06 data, respectively, demonstrate the additional dependence of $\hat{\Delta}_{\text{MLE}}$ on $p_{\text{TP}}$ and $p_{\text{FP}}$. In practical terms, this means that the same numbers of observed reciprocated and unreciprocated interactions in different data sets can map to different values of $\hat{\Delta}_{\text{MLE}}$. In contrast, the naïve approach equates observations across all experiments without regard to divergent error probabilities.

Figure 3.4 illustrates the differences in local node degree estimates for the MLE

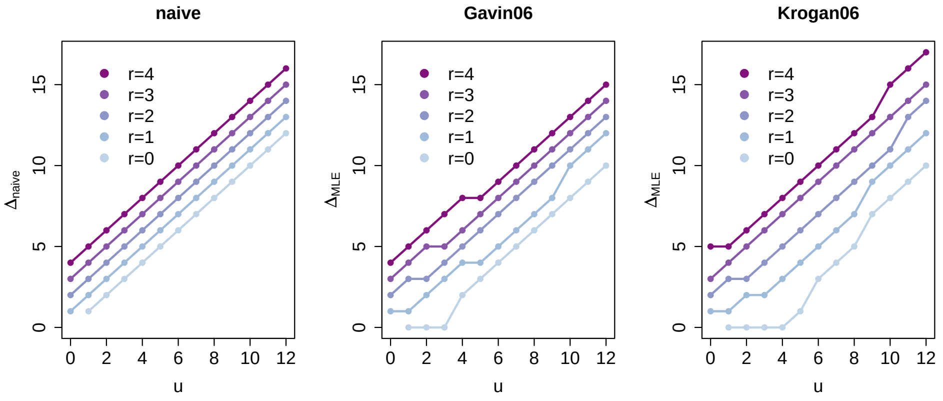


Figure 3.3: Estimated degree versus $u$ for increasing values of $r$. The leftmost panel plots $\hat{\Delta}_{\text{naïve}}$ versus $u$. The middle and right panels plot $\hat{\Delta}_{\text{MLE}}$ for the same values of $r$ and $u$ using the $p_{\text{TP}}$ and $p_{\text{FP}}$ parameters for the Gavin06 and Krogan06 data sets, respectively.

and naïve techniques. Actual numeric estimates are reported in Supplementary Table III. In these figures, colors map to negative or positive values of $\hat{\Delta}_{\text{MLE}} - \hat{\Delta}_{\text{naïve}}$. These figures illustrate important points regarding between-experiment variability in node degree estimation. First, the pattern with which estimated degree is higher or lower than the naïve sum varies dramatically from experiment to experiment. While the Gavin02 data set had a low mean estimate of 0.541 FPs per node, it also only had moderate sensitivity. Estimates of Gavin02 degree remain largely consistent with the naïve approach in the lower range, and then consistently increase. On the other hand, the Krogan04 data had high sensitivity as well as a mean of 1.1 FPs per node, so the MLE estimates are consistently less than the naïve estimates for most of the reciprocated and unreciprocated pairs studied here. Second, Figure 3.4 illustrates that the largest disparities in node degree estimates tend to exist in the extremes of the observations, that is, for large or small numbers of reciprocated and unreciprocated interactions.

In addition to between-experiment variability, Figures 3.3 and 3.4 also illustrate within-experiment variability. Observations within a data set that would be equivalent under a naïve paradigm are not when estimated via MLE. For example, in the Krogan06 data $r = 2$ and $u = 0$ yield $\hat{\Delta}_{\text{MLE}} = 2$, $r = 1$ and $u = 1$ yield $\hat{\Delta}_{\text{MLE}} = 1$, and $r = 0$ and $u = 2$ yield $\hat{\Delta}_{\text{MLE}} = 0$. While a node with two incident

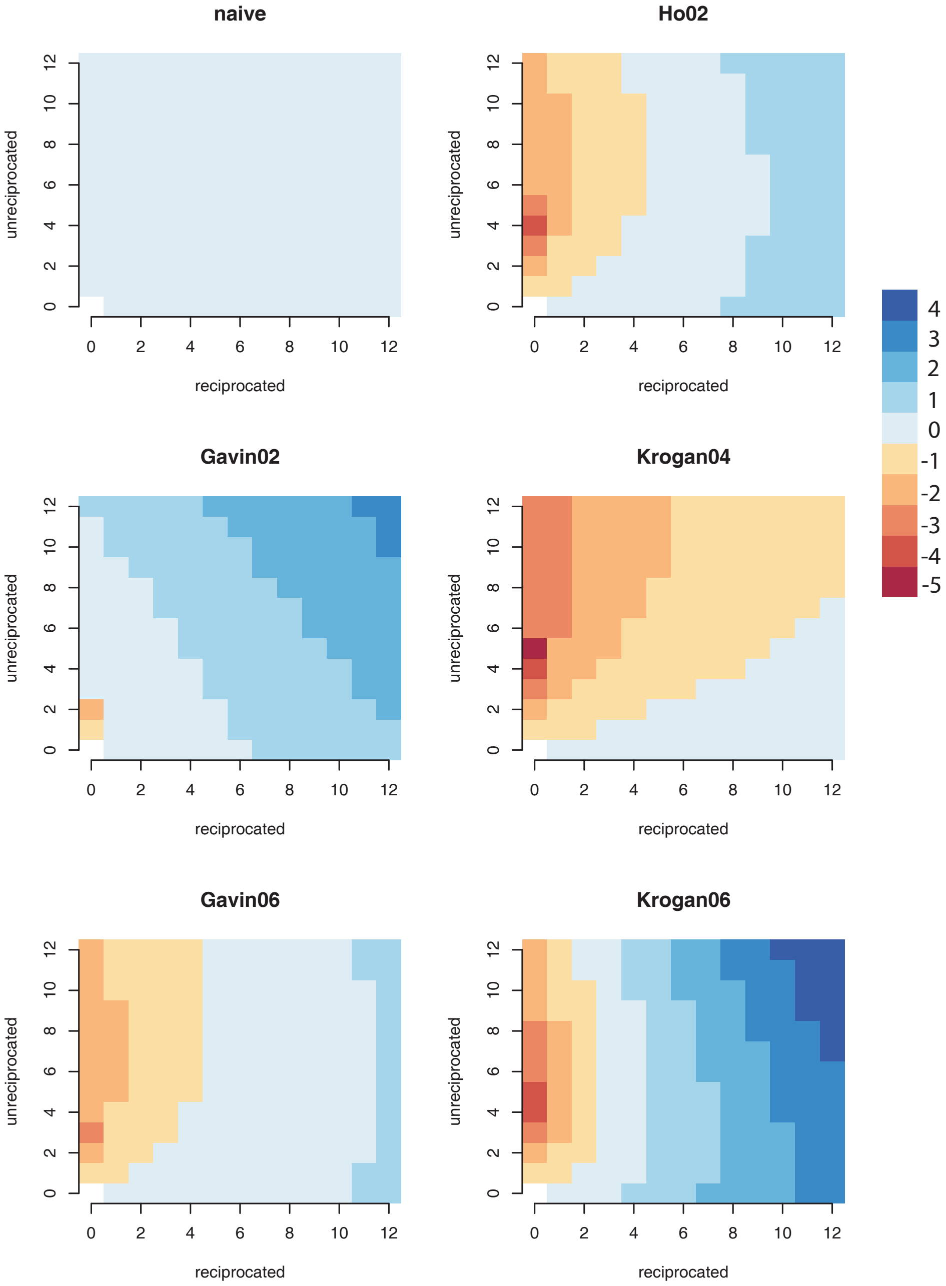


Figure 3.4: Differences in degree estimates, $\hat{\Delta}_{\text{MLE}} - \hat{\Delta}_{\text{naïve}}$, under the stochastic error probability parameters in each of the AP-MS data sets. The gradation from orange to red marks negative differences ranging from -1 to -5. The gradation from light blue to dark blue marks non-negative differences ranging from 0 to 4. The pattern in under- and overestimation varies across experiments and depends on $p_{\text{TP}}$ and the number of FPs per node. Differences are not plotted for $r = u = 0$ since degree estimation is not performed for nodes without any observed incident edges.

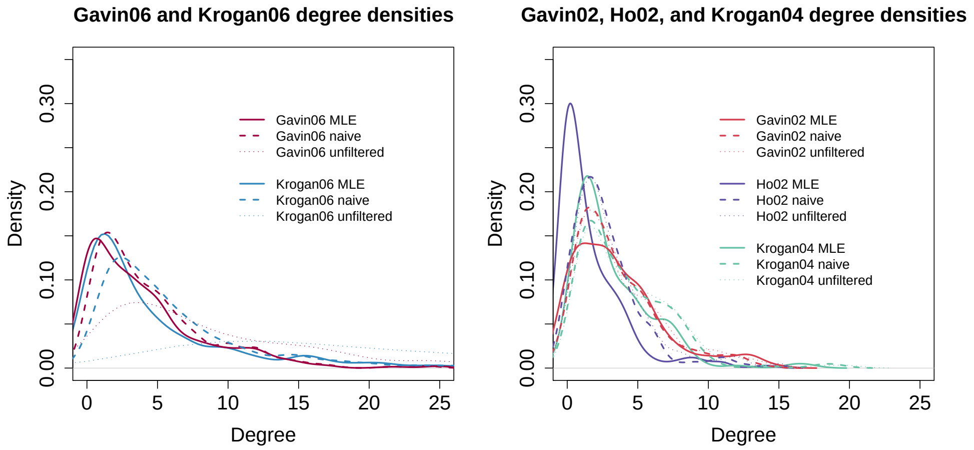


Figure 3.5: Degree density estimates of $\hat{\Delta}_{\text{MLE}}$ and $\hat{\Delta}_{\text{naïve}}$, pre- and post-filtering the VBP graph for baits prone to systematic bias.

unreciprocated edges would naïvely be assumed to interact with two members of the viable prey population, in fact, the observed interactions could quite possibly be due to stochastic error and the protein may not contribute to the system under study at all. Further wet lab experiments are still required to confirm hypotheses along these lines, but modeling observed interactions according to reciprocity status and stochastic error probabilities points to the reliability of observed interactions and can foster well-informed experimental design and resource allocation.

### Global degree analysis

Figure 3.5 contains density plots for estimates of AP-MS degree using $\hat{\Delta}_{\text{naïve}}$ prior to filtering baits prone to systematic bias, $\hat{\Delta}_{\text{naïve}}$ post-systematic-filtering, and $\hat{\Delta}_{\text{MLE}}$. The left panel compares degree density estimates for the most recent and larger Gavin06 and Krogan06 AP-MS data sets, and the right panel compares density estimates for the three smaller Gavin02, Ho02, and Krogan04 data sets.

Degree density estimates for the Gavin06 and Krogan06 data suggest that both experiments were prone to systematic error, the latter more so than the former. Degree densities for $\hat{\Delta}_{\text{naïve}}$ change drastically after the removal of biased baits with lower degrees being more prominent than in the raw unfiltered data sets. Even after removal of systematic error and modeling of stochastic error, the degree densities for $\hat{\Delta}_{\text{MLE}}$ are not identical for the Gavin06 and Krogan06 data. Although both of these experiments were intended to be genome wide, experimental conditions differentially affected both bait and prey.

In contrast to the genome-wide surveys, the earlier, smaller data sets appear far less subject to systematic error and in general have much lower degree. Although the graphs contained a similar number of nodes, the degree density for Ho02 weights low degree nodes more heavily than the degree density for the Gavin02 data. Krogan04 baits were selected based on involvement in RNA-processing and prefractionization by high-speed centrifugation was used as an intentional means of investigating smaller protein complexes. Relatively low degrees are expected due to the small complexes under investigation, but some connectivity is observed because of the functional commonality of the baits under study.

The interplay between total graph size and local node degree must be considered when drawing biological conclusions about graph data, particularly in the face of measurement error. For example, the Krogan06 data had the highest expected number of FPs per individual node (Table 3.2), but the global degree densities for Ho02 and Krogan04 actually experienced the greatest shrinkage toward zero after estimation with $\hat{\Delta}_{\text{MLE}}$. Estimated degree for Krogan06 is in general much larger than Ho02 or Krogan04, hence the impact of modeling FPs on a local level plays out differently when the graphs are considered from a global point of view. In a second example, the number of nodes and global $p_{\text{TP}}$ and $p_{\text{FP}}$ estimates for Gavin02 are closest to those for Ho02. However, after modeling of stochastic error, the node degree distribution for Gavin02 in fact much more closely resembles that for the larger Gavin06 and Krogan06 data sets, indicating larger local degree estimates. For practical data analysis, both local degree estimates and global degree distributions provide complementary information about interactome graphs.

### 3.2.5 Y2H Data Results

Rather than select a particular pair of $p_{\text{TP}}$ and $p_{\text{FP}}$ parameters for estimating node degree via MLE, it is also of interest to explore a range of plausible error probability estimates and their effect on estimated node degree. [1] discuss a method of moments approach for estimating a family of solutions for $p_{\text{TP}}$ and $p_{\text{FP}}$. Af-

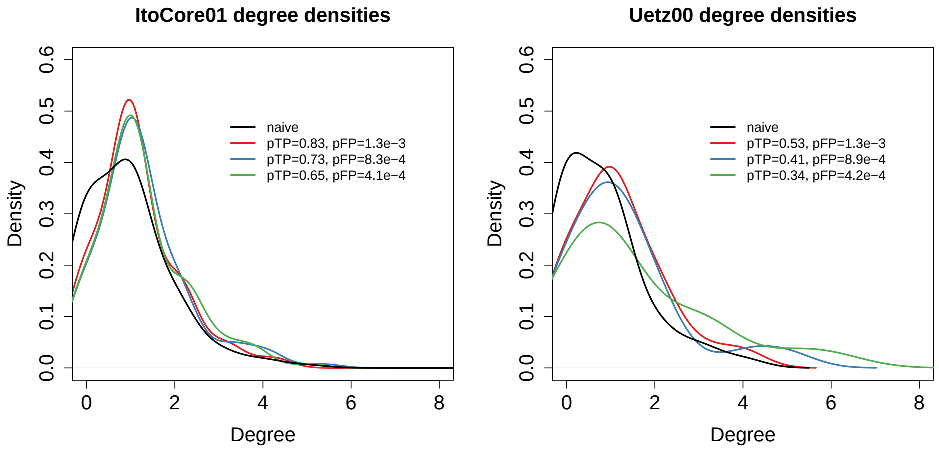


Figure 3.6: Degree density estimates of $\hat{\Delta}_{\text{MLE}}$ at the 25th, 50th, and 75th percentiles from the family of solutions for $p_{\text{TP}}$ and $p_{\text{FP}}$ using the method of [1].

ter removing VBP proteins prone to systematic bias from the ItoCore01 [14] and Uetz00 [78] data sets, we estimated $\hat{\Delta}_{\text{MLE}}$ using the 25th, 50th, and 75th percentiles for $(p_{\text{TP}}, p_{\text{FP}})$ pairs from the family of solutions for these two parameters and compared the resultant degree densities to the naïve estimate in Figure 3.6. Interestingly, for the ItoCore01 data the three parameter pairs make very little difference in node degree density (left panel of Figure 3.6). For small numbers of observed reciprocated and unreciprocated edges, the MLE estimates are equal for a wide range of error probability combinations. Degree estimate densities for the Uetz00 data are plotted in the right hand panel of Figure 3.6, and for these lower estimates of $p_{\text{TP}}$, we do see much more variation in resultant MLE density. For both Y2H data sets described here, the MLE estimates at all parameter combinations suggest that in general the nav̈e approach underestimates degree. We note that in the application of these methods to these two Y2H data sets, we assume the VBP paradigm as described in Section 1.3.3 is appropriate. These screens did use pooled tests and sampled a limited number of clones for analysis; were these parameters readily available, the bait-specific number of tests could easily be incorporated into analysis.

## 3.3 Methods

### 3.3.1 Multinomial Model for Node Degree

Let $R_{\mathrm{T}}$ and $U_{\mathrm{T}}$ be random variables representing the number of reciprocated and unreciprocated observations for true edges incident on the node of interest. Given $\Delta$ and the sensitivity of the technology, $p_{\mathrm{TP}}$, $R_{\mathrm{T}}$ and $U_{\mathrm{T}}$ are jointly distributed according to a Multinomial model, specifically

$$\begin{aligned} &Pr(R_{\mathrm{T}} = r_{\mathrm{T}}, U_{\mathrm{T}} = u_{\mathrm{T}} \mid \Delta \,, p_{\mathrm{TP}}) \\ = \; & \frac{\Delta!}{r_{\mathrm{T}}! u_{\mathrm{T}}! (\Delta - r_{\mathrm{T}} - u_{\mathrm{T}})!} \; p_{\mathrm{T2}}^{r_{\mathrm{T}}} \; p_{\mathrm{T1}}^{u_{\mathrm{T}}} \; p_{\mathrm{T0}}^{(\Delta - r_{\mathrm{T}} - u_{\mathrm{T}})}, \end{aligned} \tag{3.1}$$

where $p_{\mathrm{T2}} = p_{\mathrm{TP}}^2$, $p_{\mathrm{T1}} = 2p_{\mathrm{TP}}(1 - p_{\mathrm{TP}})$, and $p_{\mathrm{T0}} = (1 - p_{\mathrm{TP}})^2$. These probabilities arise when an individual edge is subjected to two independent assays, each with probability $p_{\mathrm{TP}}$ of producing a TP result. Under stochastic error, the direction of the unreciprocated edges is uninformative, so an FN or FP observation could be made in either direction with equal probability. Since bait-prey technologies are not perfectly sensitive, $0 < p_{\mathrm{TP}} < 1$.

Both reciprocated and unreciprocated FP edges may also arise, and these are also jointly distributed according to a Multinomial model. Let $R_{\mathrm{F}}$ and $U_{\mathrm{F}}$ be random variables representing the number of reciprocated and unreciprocated observations for edges incident on the node of interest that *do not* exist. Given $\Delta$ and the specificity of the technology, $1 - p_{\mathrm{FP}}$, $R_{\mathrm{F}}$ and $U_{\mathrm{F}}$ are jointly distributed according to

$$\begin{aligned} &Pr(R_{\mathrm{F}} = r_{\mathrm{F}}, U_{\mathrm{F}} = u_{\mathrm{F}} \mid \Delta, p_{\mathrm{FP}}) \\ = \; & \frac{(|V_{\mathrm{BP}} \mid - \Delta)!}{r_{\mathrm{F}}! u_{\mathrm{F}}! ((|V_{\mathrm{BP}}| \; - \Delta) - r_{\mathrm{F}} - u_{\mathrm{F}})!} \; p_{\mathrm{F2}}^{r_{\mathrm{F}}} \; p_{\mathrm{F1}}^{u_{\mathrm{F}}} \; p_{\mathrm{F0}}^{((|V_{\mathrm{BP}}| - \Delta) - r_{\mathrm{F}} - u_{\mathrm{F}})}, \end{aligned} \tag{3.2}$$

where $p_{\mathrm{F2}} = p_{\mathrm{FP}}^2$, $p_{\mathrm{F1}} = 2p_{\mathrm{FP}}(1 - p_{\mathrm{FP}})$, $p_{\mathrm{F0}} = (1 - p_{\mathrm{FP}})^2$, and $|V_{\mathrm{BP}}|$ is the total number of nodes in the VBP graph. Since bait-prey technologies are not perfectly specific, $0 < p_{\mathrm{FP}} < 1$.

Data arising from bait-prey technologies do not afford direct observation of $R_{\mathrm{T}}$, $U_{\mathrm{T}}$, $R_{\mathrm{F}}$, and $U_{\mathrm{F}}$ for a node. Rather, the observations are generated by $R = R_{\mathrm{T}} + R_{\mathrm{F}}$ and $U = U_{\mathrm{T}} + U_{\mathrm{F}}$ where $R$ and $U$ are random variables representing the total number (both TPs and FPs) of reciprocated and unreciprocated edges incident upon that node, respectively. The joint distribution of $R$ and $U$ is the convolution of the two distributions in (3.1) and (3.2), with a truncation factor that accounts for the fact that the VBP graph includes only nodes for which at least one in- and out-edge is observed. In particular, the convolution is divided by $1 - p_{\mathrm{T0}}^{\Delta} p_{\mathrm{F0}}^{(|V_{BP}|-\Delta)}$.

The current practice naïve estimate of degree is $\hat{\Delta}_{\text{naïve}} = r + u$. The maximum likelihood estimate of $\Delta$, $\hat{\Delta}_{\mathrm{MLE}}$, is that which maximizes the joint probability given estimates of $p_{\mathrm{TP}}$ and $p_{\mathrm{FP}}$.

We note that the assumption of independence is reasonable in deriving these two equations because each protein in the VBP graph has been selected and tested as a bait against all other proteins (nodes) within this graph. Each test (experiment) is assumed to be independent and identically distributed.

### 3.3.2 Gold Standard Complex Co-membership Edges and Estimation of $p_{\mathrm{TP}}$

Candidate complexes for the AP-MS gold standard data sets were obtained from both GO and MIPS. For the GO repository, the Cellular Component Ontology was searched to identify terms with the entire word *complex* or the suffixes *ase* or *some*. We performed manual curation to exclude the following terms that were not protein complexes but were rather subcellular locations with descriptions containing the words *chromosome*, *endosome*, *chitosome*, or *kinetochore*: GO:0000794, GO:0000780, GO:0000781, GO:0000784, GO:0000778, GO:0000942, GO:0031902, GO:0045009, GO:0000776. We also manually checked that all selected GO terms used the word *complex* as a noun. For the MIPS repository, the *Saccharomyces cerevisiae* genome database was parsed under the *complex* catalog. We included all terms containing the word *complex* in the description but excluded any containing the word *complexes* so as to avoid miscellaneous collections of multiple complexes. After collection and curation, the set of protein complexes were then merged to form one aggregate set of 335 complexes after deleting redundant protein complexes. The complete set is available in the `ScISIC` data set in the R package *ScISI* (version 1.9.7).

For each data set, let $\Delta_{\text{GS}}$ represent the total number of edges in the complex co-membership graph induced by the gold standard data set. Reciprocated and unreciprocated observations of this set of edges that we believe to exist in truth are then used to estimate $p_{\text{TP}}$. Equation (3.1) is maximized for $p_{\text{TP}}$ after replacing $\Delta$ by $\Delta_{\text{GS}}$ and letting $R_T$ and $U_T$ represent the number reciprocated and unreciprocated observations on the gold standard set, respectively. Truncation of observations is not a concern in this case since it is possible for none of the gold standard edges incident on any unique node to be observed. Specifically, the maximum likelihood estimate for $p_{\text{TP}}$ is $\hat{p}_{\text{TP}} = (2r_T + u_T)/(2\Delta_{\text{GS}})$. Variance for this estimator can be directly calculated to be $p_{\text{TP}}(1 - p_{\text{TP}})/(2\Delta_{\text{GS}})$, and is estimated by plugging in $\hat{p}_{\text{TP}}$.

### 3.3.3 Method of Moments Estimator for $p_{\text{FP}}$

[1] describe a method of moments approach for the problem of estimating $p_{\text{FP}}$. We briefly describe their method here. Let $|V_{\text{BP}}|$ be the number of nodes in the VBP graph, then the largest number of possible distinct interacting protein pairs is $\binom{|V_{\text{BP}}|}{2}$. Let $\Delta_{\text{VBP}}$ be the true number of unique interacting pairs and $\Delta^c_{\text{VBP}} = \binom{|V_{\text{BP}}|}{2} - \Delta_{\text{VBP}}$ the number of non-interacting protein pairs. The expected number of reciprocally adjacent pairs $R_{\text{VBP}}$ and unreciprocally adjacent pairs $U_{\text{VBP}}$ in the VBP graph are:

$$\begin{aligned} E[R_{\text{VBP}}] &= \Delta_{\text{VBP}}\, p^2_{\text{TP}} + \Delta^c_{\text{VBP}}\, p^2_{\text{FP}} && (3.3) \\ E[U_{\text{VBP}}] &= \Delta_{\text{VBP}}\, 2\, p_{\text{TP}}\, (1 - p_{\text{TP}}) + \Delta^c_{\text{VBP}}\, 2\, p_{\text{FP}}\, (1 - p_{\text{FP}}). && (3.4) \end{aligned}$$

These two independent equations in three parameters $\{(p_{\text{TP}}, p_{\text{FP}}, \Delta_{\text{VBP}})\}$ generate a family of solutions in which a value of any one of these parameters determines unique solutions for the other two [1].

## 3.4 Discussion

Experimental data from bait-prey technologies are prone to sampling bias, FP and FN observations. While these issues are widely recognized, little has been done to statistically model these errors when estimating global and/or local graph statistics. Straightforward application of likelihood techniques can be used to estimate node degree with more accuracy than the current practice of a naïve estimate. Simulation studies demonstrate the accuracy of the MLE approach even under misspecification of stochastic error probabilities, and depictions of graphs with known node degree distributions show the extent to which MLE degree estimation can be used to more closely resemble the true distribution. Node degree itself is useful for understanding the range of influence of an individual protein in the cell, and is a helpful contributor for elucidating the identities of a protein's interactors given the possibilities reported in a data set. Accurate estimation of node degree, and many other graph parameters, cannot be overlooked.

Our MLE approach limits analysis to the VBP graph and in this report we compare it to the naïve estimator also used on the VBP graph. In practice naïve summation is generally applied to the graph including all baits and prey-only proteins, regardless of the severe imbalance in the number of tested edges incident on each. Under perfect sensitivity and specificity, the number of detected interactions for a bait would equal its true degree since all incident edges are tested. On the other hand, for prey-only proteins, only the subset of edges connected to bait proteins are tested. The naïve sum approach generally treats the remaining untested edges as though they are tested but not observed. Hence, even under perfect sensitivity and specificity for tested edges, naïve summation actually only yields a lower bound for the true degree for prey-only proteins. Estimation of degree for nodes outside the VBP set could proceed quite naturally using the Hypergeometric distribution under the assumption that the VBP nodes are a random sample from the population of proteins. If the VBP nodes are not a random sample then it is not, in general, possible to make specific statements about the probability that a prey is observed. The APMS and Y2H data sets discussed here give no evidence that the set of VBP nodes is a representative random sample of proteins in the cell, hence extrapolation of the results to untested portions of the interactome using either the

MLE or naïve degree estimation approach would be largely irrelevant.

Focusing on the VBP graph for data analysis does therefore demand careful interpretation of the term *degree*. Degree in this context represents the number of interactions between viable baits that were also viable prey, and is therefore representative of both bait selection and the set of constitutive proteins under the experimental conditions. For intended genome-wide experiments such as Gavin06 and Krogan06 or for smaller-scale experiments with baits targeting specific cellular processes such as Krogan04, interpretation is more straightforward. Experiments between the two extremes of genome-wide assay and those targeting specific cellular functions, for example Gavin02, Ho02, ItoCore01, and Uetz00 make the interpretation of VBP degree slightly more difficult. That said, degree density estimates for these experiments do point to different characteristics of the detected interactions. While APMS and Y2H data are used here to demonstrate MLE degree estimation, the technique is equally applicable and similar conclusions can be made for data generated by other bait-prey technologies for which the VBP paradigm is appropriate.

One limitation of the MLE estimation method in its current formulation is that the same estimate of degree is made for all nodes with the same numbers of observed unreciprocated and reciprocated edges within a data set. In reality, protein-specific covariates, for example PFAM domains, likely contribute to variability in the observed data. As further relationships between these types of protein characteristics and node degree are uncovered, they can easily be accommodated into the MLE paradigm by introducing protein-specific FP and FN error probabilities.

Classic statistical techniques are readily applicable to graph feature analysis, and in this specific example yield accurate estimates of node degree. This work represents a first step toward rigorous handling of experimental noise in bait-prey data when estimating graph statistics. Likelihood methods are also appropriate for estimation of other features of *interactome* behavior, and a move beyond naïve summary statistics of observed data is both needed and warranted. Provided data are reported with bait-prey distinctions, sampling, systematic errors, and stochastic errors can be easily addressed. Such approaches promise greatly increased accuracy in estimation of global and local statistics, as well as more holistic model development.

# Chapter 4

# Statistical Assessment Protocol

Here we suggest a protocol for statistical analysis of node-and-edge graph representations of these data using R and Bioconductor, recognizing that steps may be added or omitted depending on the data set at hand. The fundamental purpose for such analyses is the estimation of data-type-specific biological features such as protein complex composition and the physical interaction integrity of known or estimated complexes. In preparation for Feature Estimation tasks, we outline a progression through three analytic components common to all bait-prey data types: Preliminary Set-up, Exploratory Analysis, and Quality Assessment. The end result is a collection of descriptive and inferred characteristics of the data, ready for biological interpretation in a computationally tractable form.

## 4.1 Introduction

Determining protein interaction partners is essential to deciphering cellular machinery since proteins often work together to form modular functional units. Different types of protein interactions are required to accomplish biological activities and can be detected by various bait-prey assay systems. Data from these bait-prey assays can be recorded in node-and-edge graphs with nodes representing genes/proteins and edges denoting observed interactions.

In response to the recent large-scale generation of high-throughput protein interaction data sets, statistical models have been designed to estimate relevant features given the observed data. Several models can be employed to quantify classic features of experimental data, including measurement error and bias, for bait-prey graphs [1]. The results can then be used to estimate specific biological features uniquely relevant to different bait-prey data types, for example protein complexes from AP-MS data [67].

This protocol provides computational guidelines for such analyses of bait-prey protein interaction data using the statistical environment R and the relevant software packages available through the Bioconductor project. We develop one specific pipeline in support of our proposed Feature Estimation methodologies; a wide variety of other software and analysis techniques for protein interaction data merit exploration and consideration by the user. We assume availability of R on the user's computer, basic familiarity with R programming, as well as an experimental design that yields bait-prey data recordable in a graph. After providing Preliminary Set-Up instructions for data compilation into a uniform data structure, we guide the user through Exploratory Analyses and Quality Assessment techniques that are generally applicable to all bait-prey data. Using the results from the exploratory analyses and quality assessment, we perform Feature Estimation depending on the type of protein interactions included in the data set of interest. Here we address specific options for AP-MS data and Y2H data. Other bait-prey systems, e.g. synthetic genetic technologies, have their own uniquely estimable features but are not addressed here.

## 4.2 Protocol Design

The analysis of protein interaction action data as outlined in this protocol can be partitioned into four distinct categories:

- Preliminary Set-up
- Exploratory Analysis
- Quality Assessment
- Feature Estimation

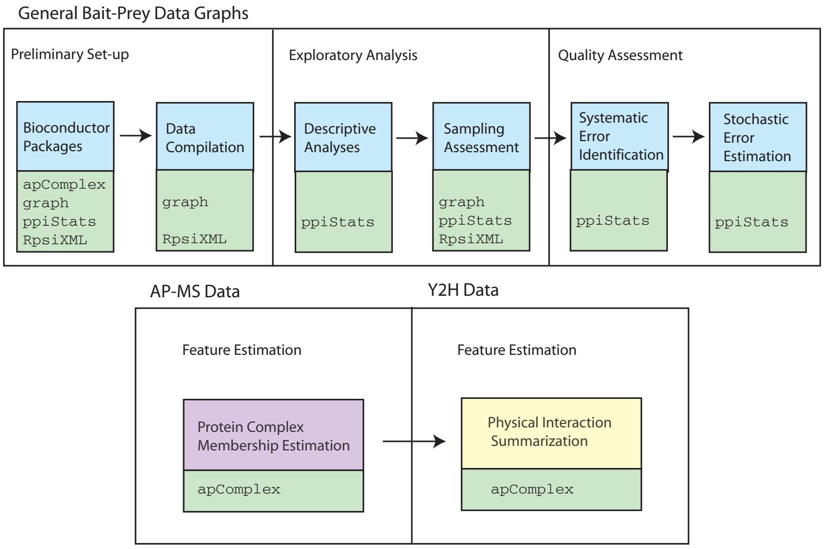


Figure 4.1: Flow chart describing the outline of analyses presented in this protocol. The top panel is relevant to analysis of general bait-prey graphs and each blue box contains the type of suggested analyses. The bottom panel outlines data-type-specific steps with the suggested analyses highlighted in different colors according to data type. The green boxes contain the names of the Bioconductor packages required to execute the analyses for each step as outlined in this protocol.

In our approach, Preliminary Set-up, Exploratory Analysis and Quality Assessment are uniformly applicable to all bait-prey data sets, while Feature Estimation depends on the nature of the data being analyzed. The flow chart in Figure 4.2 lays out the order of analysis followed in this protocol, differentiating between generally applicable and data-type-specific analysis steps. Before enumerating the step-by-step procedure for analysis, we first discuss the various components of our full bait-prey graph analysis.

### 4.2.1 Preliminary Setup

**Bioconductor Packages**: Four Bioconductor packages and their dependencies are used in the analyses presented in this protocol: apComplex, graph, ppiStats, and RpsiXML. Installation of these and R is addressed in the Materials section.

**Data Compilation**: Data must be obtained either from primary sources or from repositories in which they are stored and then transformed into a uniform data structure. We assume that a protein interaction data set consists of a uniform naming scheme for the interactors, clearly defines the type of interaction assayed (i.e. Y2H or AP-MS), and identifies the bait and prey for each interacting pair.

### 4.2.2 Exploratory Data Analysis

**Descriptive Analyses**: To characterize the sampled bait population and the covered prey population, we ascertain the following sets of proteins involved in an interaction in a data set: VB, VP and VBP. These sets serve as reasonable proxies for the bait and prey population when this information is not otherwise available.

Additionally, we ascertain the total number of directed interactions (TI) reported in the data set. Only edges extending from VB nodes to VP nodes are deemed tested. The TI statistic, therefore, must be interpreted against the number of edges queried in the graph, not the total number of possible edges.

**Sampling Assessment**: The selected bait proteins (and prey proteins depending on the technology) can give insight into the sampling scheme employed by each investigator. Protein viability summarizes which proteins are amenable to the technology and/or which proteins are expressed under the conditions of the experiment. Different data sets covering distinct sections of the interactome may result from intentional bait selection, or from differential response of baits and prey to ambient conditions if the sampled bait population is relatively invariant for the data sets.

Testing viability also allows detection of enrichment of genes or proteins with respect to certain functional annotations. Using a hypergeometric distribution, we may discover over- and under-representation within certain gene sets, for example those defined by Gene Ontology (GO) or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Such analyses may help elucidate bias for or against certain types of proteins for each experiment and technology.

**Quality Assessment**

**Systematic Error Identification**: The bait/prey distinction can further help us dissociate systematic bias from stochastic variation. Since two proteins either truly interact or do not in a given condition, with perfectly sensitive and specific technology, we should find reciprocated positive or negative observations if both proteins are VBPs. Actual technologies frequently yield un-reciprocated observations signaling some type of measurement error. If unreciprocated interactions were due only to random measurement error, then in-degree and out-degree for any VBP protein should be roughly equal. Using a binomial test to diagnose large disparities of in- and out-degree [1], we can segregate proteins into groups that might be biased in the assay system and groups that do not display this type of bias. Proteins affected by systematic errors should be removed prior to downstream analysis.

**Stochastic Error Estimation**: After diagnosing proteins prone to systematic error and removing them from further analysis, we can estimate the stochastic false negative probability ($p_{\text{FN}}$) and false positive probability ($p_{\text{FP}}$) of the technology yielding a particular data set using the remaining set of VBP proteins. Chiang et al. (2007) [1] describe a method of moments approach to estimate a family of solutions for $p_{\text{FN}}$ and $p_{\text{FP}}$ pairs given an observed VBP graph. Using gold standard (GS) sets of protein interactions, we can identify a single point estimate for $p_{\text{FN}}$ from the family of solutions, and then identify the corresponding $p_{\text{FP}}$ paired estimate (see Figure 6) [2]. Estimates of $p_{\text{FN}}$ and $p_{\text{FP}}$ provide useful comparative measures across experiments and are also necessary parameters for feature estimation.

### 4.2.3 Feature Estimation

**Protein Complex Membership Estimation**: Having identified the viable bait and prey populations, diagnosed bait proteins subject to systematic error, and estimated stochastic error probabilities for a data set, estimation of biological features can proceed. Given a set of protein complex co-membership data from an AP-MS or Co-IP experiment, one highly relevant estimation task is that of cataloguing protein complex membership. We employ a penalized likelihood approach for this estimation task [67]. Complex estimates can be sorted into three different types: multi-bait-multi-edge estimates (MBME), single-bait-multi-hit (prey) (SBMH), and un-reciprocated bait-bait (UnRBB) estimates.

**Physical Interaction Summarization**: Given a catalogue of protein complexes and their members (either known or estimated from AP-MS data), it is then natural to overlay the set of physical interactions reported by Y2H experiments. These analyses yield candidate physical structures for individual protein complexes.

## 4.3 Materials

**R and Bioconductor**: We have chosen to work within the R statistical environment (http://cran.r-project.org) and use the freely available and open source catalogue of computational libraries that comprise the Bioconductor Project (http://www.bioconductor.org). Here we reference the primary protein interaction data-related software packages that we use for analysis and list the relevant functions for each package in Table 4.1.

**R/Bioconductor Packages and Functions**:

- *apComplex*: Estimates protein complex membership from AP-MS/CoIP data; estimates the physical structures via Y2H data.
- *graph*: Instantiates the graphical data structures within R.
- *ppiStats*: Contains a variety of statistical methods and tests for protein interaction analysis.
- *RpsiXML*: Parses molecular interaction databases that implement a compliant PSI-MI XML2.5 schema [79].

### Protocol Structure

From section 4.3 Step 1 to Step 7 of this protocol, we describe the methodology as it pertains to the analysis of a collection of protein interaction data sets. These

| ppiStats | graph | RpsiXML | apComplex |
|---|---|---|---|
| createSummaryTables | graphNEL | psimi25Graph | findComplexes |
| estErrProbMethodOfMoments | ftM2graphNEL | getAbstractByPMID | plotComplex |
| estimateCCMErrorRates | removeSelfLoops | separateXMLDataByExpt | sortComplexes |
| findAdjacent | subGraph | translateID | |
| idSystematic | | | |
| idViableProteins | | | |
| inOutScatterCharts | | | |
| ppiBuildParams4GO | | | |
| ppiHGTest4GO | | | |
| viabilityCharts | | | |
| separateExptBySize | | | |

Table 4.1: This table details the relevant classes and functions for bait-prey interaction analysis contained in the packages (in red). A complete description of each class object (green) or function (blue) can be found within its corresponding *help page*. The use of all functions is demonstrated within the scope of this protocol.

steps allow the user to conduct meta-analysis on a number of data sets. From Step 8 onwards, we focus our attention to a single data set since the procedures and methodology become more technical. In principal, all the steps taken in this protocol may be conducted on either on a single data set or multiple data sets simultaneously.

# Procedure

## Preliminary Set-up: Bioconductor Packages

### 1. Load Bioconductor packages required for analysis using the library command.

```
> library(apComplex)
> library(graph)
> library(ppiStats)
> library(RpsiXML)
> library(RBGL)
> library(GOstats)
> library(org.Sc.sgd.db)
> options(error=recover)
```

Loading of these libraries first requires installation of the packages; see http://www.bioconductor.org/docs/install for installation instructions. Because R and Bioconductor are open source and freely available software tools maintained by a number of different contributors, only the most current versions of the software packages are intended for use; therefore, only the current stable released version of each software package is recommended for analyzing one's data. Installation of these packages depends on installation of the package *Rgraphviz*, which can be difficult. The Bioconductor Mailing List (http://www.bioconductor.org/docs/mailList.html) is a good resource for receiving help in installation.

# Preliminary Set-up: Data Compilation

**2. Obtain protein interaction data either from on-line databases or primary sources.**

A. Obtain protein interaction data from on-line databases.

*RpsiXML* communicates with databases that have implemented a compliant PSI-MI XML2.5 schema. Currently, these databases include *BioGRID, DIP, HPRD, IntAct, matrixDB,* and *MINT*. Within the scope of this protocol, we work exclusively with the *IntAct* repository. In the first example below, 20 XML files of *Saccharomyces cerevisae* protein interactions were downloaded and saved in a directory called *Intact/intactYeast/* prior to analysis. We created the character vector `filenames` that records the relative location of these XML files to the R working directory. We print out the first element of `filenames` as an example:

```
> filenames[1:2]

[1] "Intact/intactYeast/yeast_small-01.xml"
[2] "Intact/intactYeast/yeast_small-02.xml"
```

Create experiment-specific psimi25Graph objects from the set of PSI-MI XML2.5 files using the function `separateXMLDataByExpt` from the package *RpsiXML.* The output of `separateXMLDataByExpt` is a list of psimi25Graph objects indexed by pubmed IDs. Each psmi25Graph object will have vertices labeled with UniProtKB accession codes (AC) with directed edges representing detected bait-prey interactions among them. Databases such as IntAct, MINT, and DIP that adhere to the minimal information for molecular interactions schema (MIMIx) [80] will have bait and prey information necessary for this analysis.

```
> y2hGraphs <- separateXMLDataByExpt(filenames,
+                                    INTACT.PSIMI25,
+                                    type="direct",
+                                    directed=TRUE,
+                                    verbose=FALSE)
```

```
> apmsGraphs <- separateXMLDataByExpt(filenames,
+                                     INTACT.PSIMI25,
+                                     type="indirect",
+                                     directed=TRUE,
+                                     verbose=FALSE)
```

It is important to keep different interaction types separate; the type parameter can be set to *direct* for physical binary interactions or *indirect* for protein complex co-membership interactions. The analysis steps performed in this protocol should be conducted on individual data sets of one particular interaction type without combining data sets into single graph data structures. Informative summary statistics typically vary between experiments, and the type of assay will contribute to different types of systematic errors. Mixing edges of different types can lead to mis-estimation of biological features.

B. Obtain protein interaction data from primary sources.

Data from primary sources are generally recorded as a two column table of bait to prey interactions. We have obtained data from [28] and created the R dataframe bpTable to encapsulate the bait-prey interactions. The first three rows are printed below:

```
> bpTable[1:3,]
```

```
     Bait      Prey
[1,] "YAL001C" "YAL001C"
```

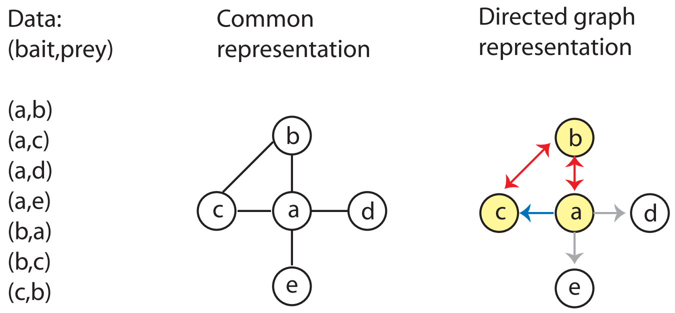


Figure 4.2: The seven ordered pairs represent protein interaction (bait,prey) pairs. The graph on the left is a common graphical representation of the data in which there is no distinction between baits and prey and thus ambiguity as to the number of assays performed on each edge. The directed graph representation on the right brings clarity to the reported data, differentiating between bi-directionally tested edges connecting pairs of VBP nodes, uni-directionally tested edges between VBs and VPOs, and untested edges between pairs of VPOs. The yellow nodes represent VBP nodes while white nodes represent VPO nodes. Red arrows indicate bi-directionally tested and observed edges; blue arrows indicate bi-directionally tested but uni-directionally observed edges; grey arrows indicate uni-directionally tested and observed edges.

```
[2,] "YBR103W" "YAL001C"
[3,] "YBR123C" "YAL001C"
```

We then transform the data frame into a directed graph object using the R function `ftM2graphNEL` from the graph package.

```
> krogan2006Graph <- ftM2graphNEL(bpTable)
```

When the data are given as such a table, we still need to make sure to distinguish the role of the proteins in each interacting pair. Distinguishing between bait and prey proteins in a set of edges is necessary for correctly understanding the data recorded in a representative directed graph (see Figure 4.3).

**3. Partition interaction data sets**: If desired, partition interaction data into small, medium, and large scales by the number of bait-prey interactions using the function `separateExptBySize` from the package *ppiStats*. In the example below, small-scale experiments are arbitrarily defined as having less than 50 interactions, large-scale experiments are those with more than 100 interactions, and medium-scale are those in between. The output of `separateExptBySize` is a three-element list where each element contains those graph objects that are deemed small, medium, or large, respectively. Splitting the data sets according to size can be helpful depending on the analyses being conducted. We focus on the large-scale data sets for this protocol as the statistical methods are most relevant to larger data sets.

```
>   sortedY2H <- separateExptBySize(y2hGraphs, 50, 100)
>   sortedAPMS <- separateExptBySize(apmsGraphs, 50, 100)
>   names(sortedAPMS)
>   y2hLargeScale <- sortedY2H[["largeScale"]]
>   apmsLargeScale <- sortedAPMS[["largeScale"]]
```

**4. UniProtKB IDs** If desired, exchange UniProtKB node names in graphs for alternative names using the function translateID from the package RpsiXML. Each database repository will support a limited number of gene/protein names, so it is necessary to find a common naming scheme across databases. For example, translate the nodes from each psimi25Graph from UniProtKB AC to ordered locus names (as given by the Comprehensive Yeast Genome Database) and then add the data of Krogan et al. (2006) manually:

```
> apmsLSLocusNamesNodes = lapply(apmsLargeScale,
+   function(x) {translateID(x,"cygd")})
> apmsLSLocusNamesNodes = lapply(apmsLSLocusNamesNodes,
+   function(x){nodes(x) = toupper(nodes(x));
+               return(x)})
> apmsLSLocusNamesNodes[[length(apmsLSLocusNamesNodes)+1]] =
+   krogan2006Graph
> y2hLSLocusNamesNodes = lapply(y2hLargeScale,
+   function(x) translateID(x,"cygd"))
> y2hLSLocusNamesNodes = lapply(y2hLSLocusNamesNodes,
+   function(x){nodes(x) = toupper(nodes(x));
+               return(x)})
```

We then replace the experiment IDs from each psimi25Graph from the pubmed ID to the first author and year by obtaining abstract information from NCBI, getting the identity of the first author, determining the date of publication, and concatenating the author and date information. This allows for easy access to each of the data sets.

```
> apmsAbstract = getAbstractByPMID(names(apmsLargeScale))
> fauthor <- sapply(apmsAbstract, function(x){authors(x)[1]})
> datePub <- sapply(apmsAbstract, pubDate)
> expNames <- paste(fauthor, datePub, sep=":")
```

Lastly, we manually add the name and date from the data obtained from primary source Krogan et al. (2006) and rename the list of large-scale AP-MS data sets.

```
> expNames[length(expNames)+1] <- "NJ Krogan:Mar 2006"
> names(apmsLSLocusNamesNodes) <- expNames
```

## Exploratory Analyses: Descriptive Analyses

**5. Protein Interaction Table** Calculate VB, VP, VBP, and TI measures for experiments of the same type using the function `createSummaryTables` from the package ppiStats. Table 4.2 contains output for AP-MS data generated by the code given below.

```
> aTable <- createSummaryTables(apmsLSLocusNamesNodes[
+                                sapply(apmsLSLocusNamesNodes,
+                                isDirected)])
```

**6. Protein Interaction Figure** Generate bar charts summarizing the VBP, VBO, and VPO sets relative to the entire yeast genome for each of the experimental data sets using the function `viabilityCharts` from the package *ppiStats*. Figure 4.3 contains the bar charts generated by the code below.

```
> viabilityCharts(apmsLSLocusNamesNodes[
+                 sapply(apmsLSLocusNamesNodes,
+                 isDirected)])
```

| | VB | VP | VBP | VBP/VB | VP/VB | TI | TI/VB |
|---|---|---|---|---|---|---|---|
| AC Gavin:Jan 2002 | 588 | 1436 | 554 | 0.94 | 2.44 | 3947 | 6.71 |
| Y Ho:Jan 2002 | 493 | 1315 | 232 | 0.47 | 2.67 | 3686 | 7.48 |
| MD Ohi:Apr 2002 | 7 | 44 | 7 | 1.00 | 6.29 | 167 | 23.86 |
| P Grandi:Jul 2002 | 13 | 66 | 12 | 0.92 | 5.08 | 405 | 31.15 |
| J Graumann:Mar 2004 | 22 | 376 | 5 | 0.23 | 17.09 | 563 | 25.59 |
| TR Hazbun:Dec 2003 | 33 | 415 | 11 | 0.33 | 12.58 | 640 | 19.39 |
| KK Baetz:Feb 2004 | 8 | 26 | 8 | 1.00 | 3.25 | 132 | 16.50 |
| R Zhao:Mar 2005 | 53 | 74 | 8 | 0.15 | 1.40 | 133 | 2.51 |
| AC Gavin:Mar 2006 | 1752 | 1790 | 991 | 0.57 | 1.02 | 19105 | 10.90 |
| NJ Krogan:Mar 2006 | 2264 | 5323 | 2226 | 0.98 | 2.35 | 63360 | 27.99 |

Table 4.2: A general overview of summary statistics for the large scale AP-MS data sets. The statistics within this table can differentiate between experiments for which estimation might not yield the same features. For example, the last column details the number of interactions per VB; the experiments of Ohi2002, Grandi2002, and Graumann2004 have remarkably high numbers of average interactions per VB, and so would produce markedly distinct features from that of Gavin2006.

# Exploratory Analyses: Sampling Assessment

**7. Interaction Partners**: We shall examine prey partners for specific baits of interest to note sampling variations among experiments using the function `findAdjacent` from the package *ppiStats*. Here we give one specific example of the differences in prey detected for the same bait *YDL047W* in the set of large-scale AP-MS experiments. First we specify a bait of interest and find its prey partners in the large-scale AP-MS experiments.

```
> myBait <- "YDL047W"
> baitSubgraphNodes <-lapply(apmsLSLocusNamesNodes,
+                             FUN=findAdjacent,
+                             bait=myBait)
```

Next we determine experiments where $YDL047W$ is a viable bait:

```
> baitSubgraphNodes[sapply(baitSubgraphNodes,
+                          function(x) {!is.null(x)})]

$`AC Gavin:Jan 2002`
```

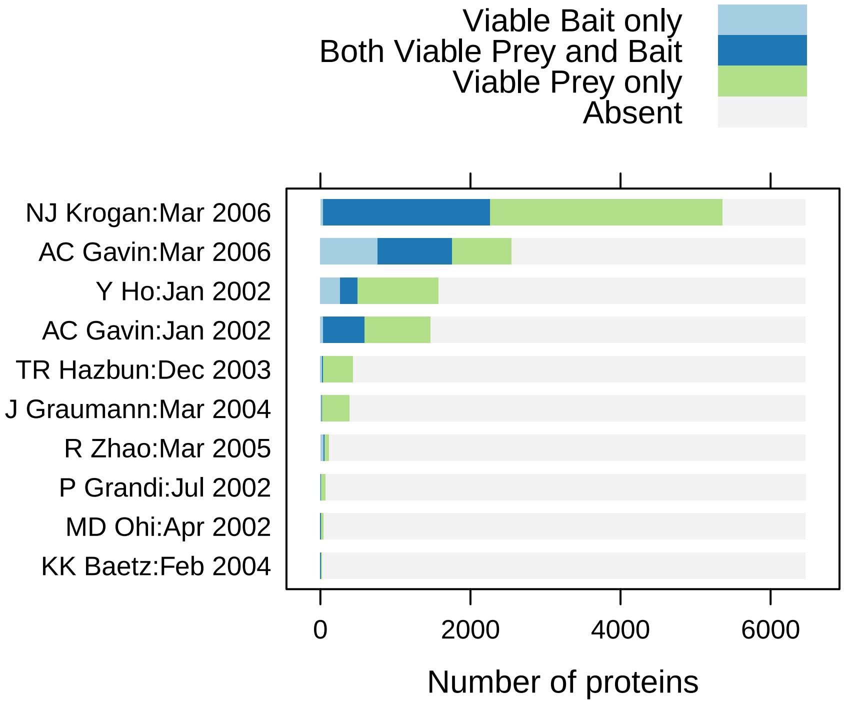


Figure 4.3: Viable protein levels encapsulated in bar charts for the AP-MS experiments. The relative proportion of VBP to VBO and VPO allows for not only more statistical power with which to determine systematic versus stochastic errors but also more quantitative estimation of features such as protein complexes. In addition, viable protein levels allow for an easy comparative analysis. For instance, with the exception of Krogan2006, the remaining large-scale interaction experiments cover at most half of the known yeast protein encoding genes.

```
 [1] "YDL047W" "YNR016C" "YMR024W" "YER155C" "YIL142W" "YLR310C" "YFR019W"
 [8] "YGL197W" "YFR040W" "YJL098W" "YKR028W" "YKL195W" "YNL101W" "YNL187W"
[15] "YOR267C"

$`Y Ho:Jan 2002`
 [1] "YDL047W" "YMR205C" "YDL029W" "YJL098W" "YFR040W" "YMR246W" "YJL026W"
 [8] "YBR039W" "YKR028W" "YNR016C" "YML048W" "YJR072C" "YOR317W" "YMR145C"
[15] "YGR282C" "YDR188W" "YJR105W" "YMR028W" "YHR096C" "YMR196W" "YDL171C"
[22] "YKL029C" "YDR465C" "YOR259C"

$`AC Gavin:Mar 2006`
```

```
 [1] "YDL047W" "YBR118W" "YAL005C" "YBR127C" "YPL106C" "YER155C" "YJL098W"
 [8] "YFR040W" "YGL197W" "YNR016C" "YOR267C" "YLL024C" "YDL229W" "YOL127W"
[15] "YGR085C" "YMR024W" "YKR028W" "YNL187W" "YNL101W" "YIL142W" "YLR310C"
[22] "YFR019W" "YKL195W"

$`NJ Krogan:Mar 2006`
[1] "YDL047W" "YFR040W" "YJL098W" "YLR097C" "YKR028W"
```

And we also determine experiments where $YDL047W$ is not a viable bait:

```
> names(baitSubgraphNodes[sapply(baitSubgraphNodes,
+                                is.null)])
```

```
[1] "M Ohi:2002" "P Grandi:2002" "J Graumann:2004"
[4] "T Hazbun:2003" "K Baetz:2004" "R Zhao:2005"
```

From the 4 experiments where $YDL047W$ is a viable bait, we can look at the pairwise intersections:

```
> intersect(baitSubgraphNodes[["AC Gavin:Jan 2002"]],
+           baitSubgraphNodes[["AC Gavin:Mar 2006"]])
```

```
 [1] "YDL047W" "YNR016C" "YMR024W" "YER155C" "YIL142W" "YLR310C" "YFR019W"
 [8] "YGL197W" "YFR040W" "YJL098W" "YKR028W" "YKL195W" "YNL101W" "YNL187W"
[15] "YOR267C"
```

```
> intersect(baitSubgraphNodes[["AC Gavin:Jan 2002"]],
+           baitSubgraphNodes[["NJ Krogan:Mar 2006"]])
```

```
[1] "YDL047W" "YFR040W" "YJL098W" "YKR028W"
```

```
> intersect(baitSubgraphNodes[["AC Gavin:Mar 2006"]],
+           baitSubgraphNodes[["NJ Krogan:Mar 2006"]])
```

```
[1] "YDL047W" "YJL098W" "YFR040W" "YKR028W"
```

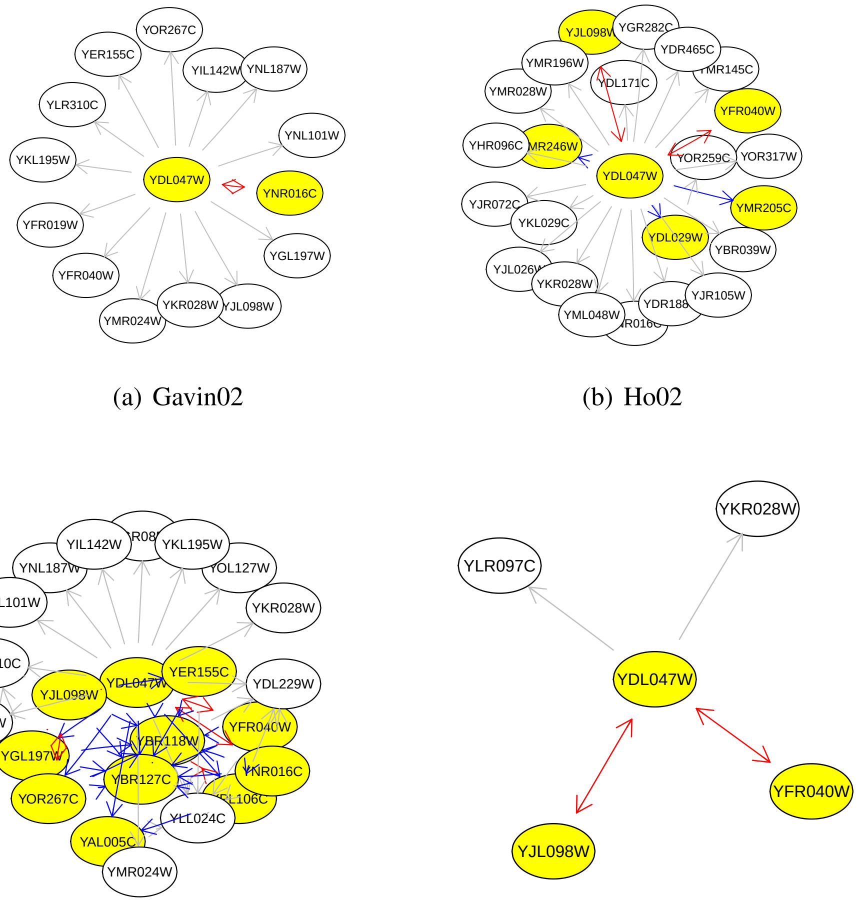
The intersection function only finds the common interactions that were reported but cannot determine meaningful statistics based upon the repetition of interactions. Since the experiments might vary in both proteins sampled and experimental conditions, it is not recommended to make qualitative conclusions from this function. Methods such as Venn diagram comparison may be misleading comparative tests. The pairwise intersection of the prey found by *YDL047W* allows the inference of similar experimental conditions. The output above shows not only a large overlap between the two Gavin data sets but also identical overlap when each Gavin data set is intersected with the Krogan 2006 data.

**8. Gene Ontology** Assess sampling with regard to protein function as described by Gene Ontology for individual experiments. Use the function `ppiBuildParams4GO` from *ppiStats* to gather the appropriate data sets, ontology, and other relevant parameters and then `ppiHGTest4GO` to execute the tests. Figure 4.5 contains a screenshot of the html page generated to summarize the results from `ppiHGTest4GO`. For this and all remaining analyses, we will focus on the single high throughput data set of Gavin et al. These analyses assume there are no edges pointing from a node into itself. The function `removeSelfLoops` can be used to remove any such edges before analysis.

```
> Gavin2006Graph = apmsLSLocusNamesNodes[["AC Gavin:Mar 2006"]]
> Gavin2006Graph = removeSelfLoops(Gavin2006Graph)
> viableProteins = idViableProteins(Gavin2006Graph)
> VB = viableProteins[["VB"]]
> VBP = viableProteins[["VBP"]]
> VPO = setdiff(viableProteins[["VP"]],VBP)
> yg = Lkeys(org.Sc.sgdGENENAME)
```

In the following example (code below), we specified use of the Cellular Component (CC) ontology by setting `ontology="CC"`. The biological process (BP) or molecular function (MF) ontologies from GO may also be used by specifying the parameter `ontology` in `ppiBuildParams4GO` accordingly. In addition, other categories beyond GO may be tested, e.g. KEGG or pFAM (cf *?ppiBuildParams4GO* in an R session). The hypergeometric tests require specification of the `geneSet` of interest as well as the `universe` against which enrichment will be evaluated. In our example, we specify all known *S. cerevisiae* encoding genes to comprise the universe. Other specifications could drastically change the results.

```
> param = ppiBuildParams4GO(geneSet=VB, universe=yg,
+   direction="over", annot = "org.Sc.sgd.db",
+   ontology="CC", cond=TRUE, pThresh=0.001)
> hg <- ppiHGTest4GO(parameter=param, filename = "Gavin2006",
+                    append=FALSE, label = "Gavin 2006",
+                    typeGeneSet = "Viable Baits", cs = 50)
```



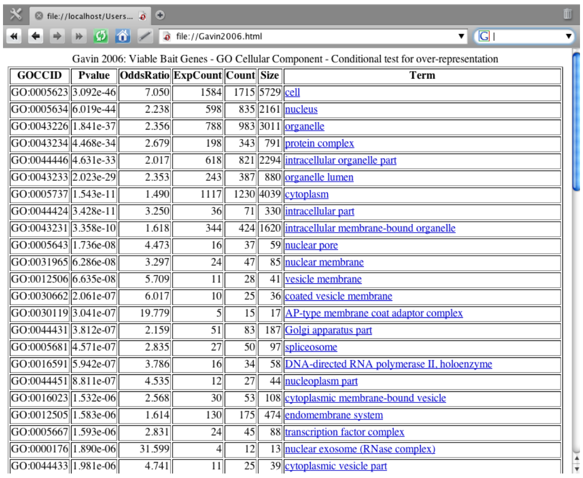
(a) Gavin02

(b) Ho02

(c) Gavin06

(d) Krogan06

Figure 4.4: Here we render the subgraphs of all detected interactions between YDL047W and its prey in four different AP-MS experiments. The subgraphs from Gavin02 and Gavin06 have the greatest number of common interactions, while all other pairings of the subgraphs yield very little overlap. The lack of overlap is due to a combination of measurement error and different experimental conditions. While it is important to minimize the effects of errors within the data, care must also be taken not to force two data sets with different ambient conditions to fit one particular model. A variation of features is expected across data sets.



Gavin 2006: Viable Bait Genes - GO Cellular Component - Conditional test for over-representation

| GOCCID | Pvalue | OddsRatio | ExpCount | Count | Size | Term |
|---|---|---|---|---|---|---|
| GO:0005623 | 3.092e-46 | 7.050 | 1584 | 1715 | 5729 | cell |
| GO:0005634 | 6.019e-44 | 2.238 | 598 | 835 | 2161 | nucleus |
| GO:0043226 | 1.841e-37 | 2.356 | 788 | 983 | 3011 | organelle |
| GO:0043234 | 4.468e-34 | 2.679 | 198 | 343 | 791 | protein complex |
| GO:0044446 | 4.631e-33 | 2.017 | 618 | 821 | 2294 | intracellular organelle part |
| GO:0043233 | 2.023e-29 | 2.353 | 243 | 387 | 880 | organelle lumen |
| GO:0005737 | 1.543e-11 | 1.490 | 1117 | 1230 | 4039 | cytoplasm |
| GO:0044424 | 3.428e-11 | 3.250 | 36 | 71 | 330 | intracellular part |
| GO:0043231 | 3.358e-10 | 1.618 | 344 | 424 | 1620 | intracellular membrane-bound organelle |
| GO:0005643 | 1.736e-08 | 4.473 | 16 | 37 | 59 | nuclear pore |
| GO:0031965 | 6.286e-08 | 3.297 | 24 | 47 | 85 | nuclear membrane |
| GO:0012506 | 6.635e-08 | 5.709 | 11 | 28 | 41 | vesicle membrane |
| GO:0030662 | 2.061e-07 | 6.017 | 10 | 25 | 36 | coated vesicle membrane |
| GO:0030119 | 3.041e-07 | 19.779 | 5 | 15 | 17 | AP-type membrane coat adaptor complex |
| GO:0044431 | 3.812e-07 | 2.159 | 51 | 83 | 187 | Golgi apparatus part |
| GO:0005681 | 4.571e-07 | 2.835 | 27 | 50 | 97 | spliceosome |
| GO:0016591 | 5.942e-07 | 3.786 | 16 | 34 | 58 | DNA-directed RNA polymerase II, holoenzyme |
| GO:0044451 | 8.811e-07 | 4.535 | 12 | 27 | 44 | nucleoplasm part |
| GO:0016023 | 1.532e-06 | 2.568 | 30 | 53 | 108 | cytoplasmic membrane-bound vesicle |
| GO:0012505 | 1.583e-06 | 1.614 | 130 | 175 | 474 | endomembrane system |
| GO:0005667 | 1.593e-06 | 2.831 | 24 | 45 | 88 | transcription factor complex |
| GO:0000176 | 1.890e-06 | 31.599 | 4 | 12 | 13 | nuclear exosome (RNase complex) |
| GO:0044433 | 1.981e-06 | 4.741 | 11 | 25 | 39 | cytoplasmic vesicle part |

Figure 4.5: Screenshot of the output HTML for hypergeometric over-representation testing. Over-representation (or enrichment) shows that the experimental assay has a preference for a group of genes/proteins. Using information about enriched categories such as GO terms allows us to use prior information in the estimation of new features. Conversely, categories which show under-representation may indicate sampling bias.

In addition to the output of `ppiHGTest4GO` in the R session (stored as `hg`), an html page containing results from the GO analysis will be created and stored in the working directory unless otherwise specified. The `filename` argument in `ppiHGTest4GO` is used to specify the file name, e.g. in this case 'Gavin2006.html'.

# Quality Assessment: Systematic Error Identification

**9. Directed Degree** Create scatter charts of indegree versus outdegree for VBP nodes to visualize the subset of nodes prone to systematic error using the function `inOutScatterCharts` from the package *ppiStats*. Nodes with large disparities in in- vs. outdegree are likely subject to some type of bias as they do not exhibit variation consistent with technology-related stochastic measurement error rates.

```
> gl = list("Gavin2006" = Gavin2006Graph)
> inOutScatterCharts(gl)
```

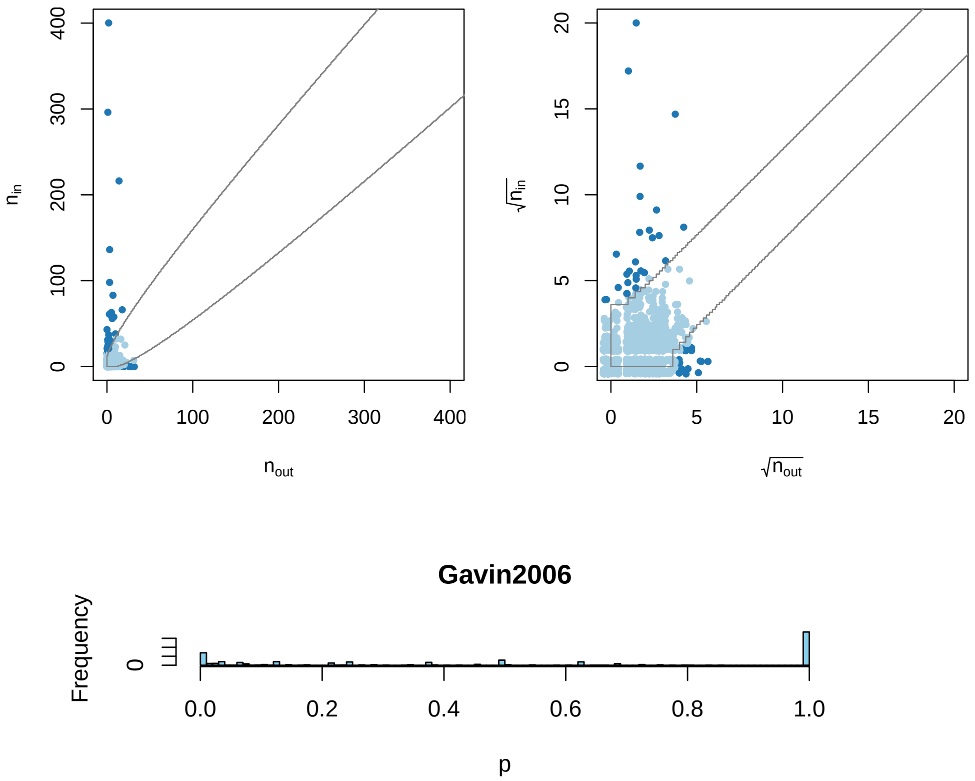


Figure 4.6: Scatter plots corresponding to the in- and out-degree for each VBP of Gavin 2006. Each point in the left plot details the in-degree on the x-axis and the out-degree on the y-axis; in the right plot, each degree is scaled by its square root. The points within the diagonal bands represents those proteins for which there is not enough evidence to reject the null hypothesis that the protein might be affected by some systematic bias while the other points are rejected at the 0.01 p-value. The bottom plot describes the distribution of the p-values for each data point. While the majority of proteins are not rejected, we can see that a substantial number of proteins are. Rejected proteins should not be analyzed using standard statistical methods to model stochastic variability nor should they be used to estimate underlying features within the data.

**10. Potential Bias** Identify specific VBP nodes prone to systematic error using the function `idSystematic` from the package *ppiStats*. The output of the `idSystematic` function is a character vector of those proteins/genes for which in-degree and out-degree is largely disproportionate based upon a Binomial test with $p = \frac{1}{2}$ (see [1]).

```
> systematicBaits = idSystematic(Gavin2006Graph,VB,bpGraph=TRUE)
```

**11. Stochastic Error Estimation** Estimate stochastic error probabilities on the subgraph induced by the set of VBP nodes not prone to systematic error using the function `estimateCCMErrorRates` from the package *ppiStats*. This function estimates stochastic $p_{TP}$ and $p_{FP}$ probabilities by first calculating the method of moments solution manifold proposed by [1] and then selected a single pair of point estimates according to a gold standard. In this example, the object `ScISIC` is a set of manually curated protein complexes culled from MIPS, GO, and IntAct that is used to create a set of gold standard positive complex co-membership interactions [2]. Figure 4.7 demonstrates where the gold standard point estimates lie on the two-dimensional manifold of possibilities.

```
> unbiasedBaits = setdiff(VB,systematicBaits)
> Gavin2006sub = subGraph(c(unbiasedBaits,VPO),
+   Gavin2006Graph)
> Gavin2006submat = as(Gavin2006sub,
+   "matrix")
> Gavin2006m = Gavin2006submat[unbiasedBaits,
+   c(unbiasedBaits,VPO)]
> data(ScISIC)
> errorRates = estimateCCMErrorRates(m=Gavin2006m,
+   GS=ScISIC)
> pTPestimate = errorRates$globalpTP
> pFPestimate = errorRates$globalpFP
```

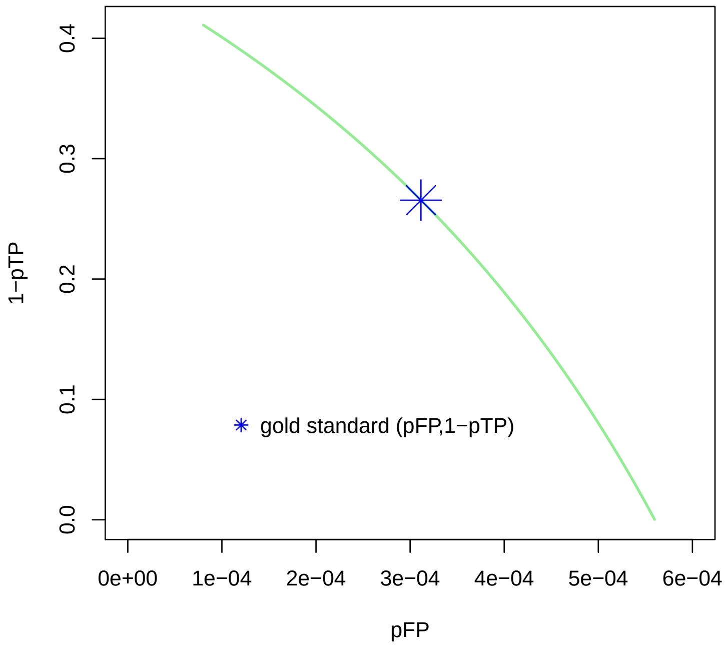


Figure 4.7: Estimate of $p_{TP}$ and $p_{FP}$ for Gavin06 AP-MS data using the gold standard `ScISIC` reference data set and a manifold of possible solutions. In particular, the `ScISIC` protein complexes were used as a bench mark set of positive co-membership interactions so that the $p_{FN} = 1 - p_{TP}$ rate could be estimated. With the $p_{FN}$ rate and the Methods of Moments, an estimate for the $p_{FP}$ rate (blue star) is recovered from the solution loci. Estimates of $p_{FN}$ and $p_{FP}$ are essential components for downstream feature estimation.

## 4.4 Feature Estimation: Protein Complex Membership Estimation

The feature estimation analyses that follow are specific to AP-MS and Y2H data. They make use of the previously discussed descriptive and quality assessment techniques and results. Other feature estimation approaches are possible and relevant to other bait-prey data types. We note that from here forward, we shall use the protein names rather than the gene locus names. Using the gene locus name made the analysis simpler; the protein names will be more helpful in the functional

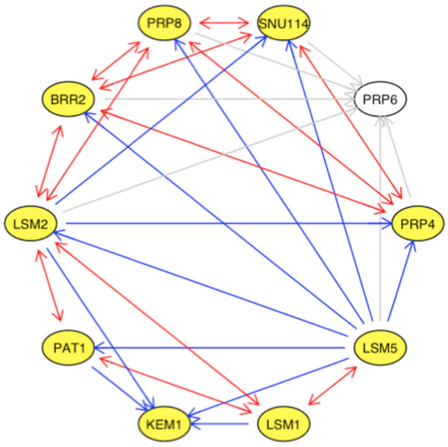


Figure 4.8: This graph renders all the proteins found to interact with the preselected bait protein LSM5 (YER146W) within the Gavin 2006 data set. Inspection of this subgraph shows 42 observed interactions between 10 proteins, which makes this a highly connected subgraph. Of the 10 proteins that make up this subgraph, 9 of these proteins (including LSM5) participated in the experiment as both viable baits and viable prey. Discounting self-interactions, there were 72 bilaterally tested relationships amongst the VBPs. Of these 72 tested interactions, Gavin and colleagues found 12 reciprocated and 12 unreciprocated interactions. In general, the feature of interest is to estimate whether or not these 10 proteins constitute a single functional unit or the amalgamation of several smaller units. Information such as reciprocity, error rates, and bias can all contribute to meaningful estimation.

analysis of multi-protein complexes.

**12. Protein Complexes** We estimate protein complex membership for AP-MS data using the function `findComplexes` from the package *apComplex*. The penalized likelihood approach employed by the *apComplex* package seeks to locate tightly connected local features within the global set of detected interactions, as depicted in Figure 4.8. The R function `findComplexes` requires specification of $p_{\text{FN}}$ and $p_{\text{FP}}$. In the working example below, we use the gold standard estimates generated in the Quality Assessment exercises. We note that descriptions of the bait/prey sampling scheme and functional representation of viable proteins as outlined in the Exploratory Analyses section will aid in the understanding of the resultant protein complex estimates.

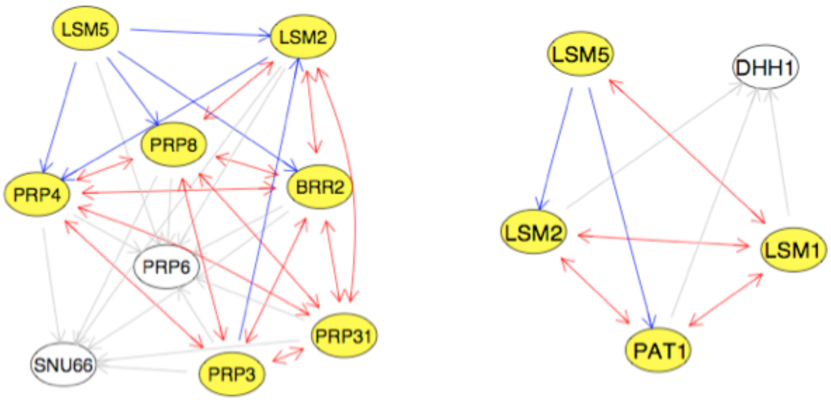


Figure 4.9: Two examples of complex estimates using the *apComplex* algorithm on the Gavin 2006 data set. This particular example demonstrates the assignment of Lsm5 to two of its characterized roles. The complex on the left depicts Lsm5 participating in a complex with other members of the U4/U6 x U5 tri-snRNP complex [81]. The complex on the right depicts its activity with the Lsm1p-7p–Pat1p complex [82]. Estimating functional modules via clustering algorithms cannot attain such resolution since clustering partitions the data.

```
> Gavin2006complexEstimates =
+ findComplexes(Gavin2006m,sensitivity=pTPestimate,
+                  specificity=1-pFPestimate)
```

We can also used the `plotComplex` function to obtain a pictorial representation of the protein complexes. It is useful to note that if the graphs have a large number of nodes with a dense set of edges, the rendering of these graphs will not be very useful.

```
> par(mfrow=c(1,2))
> for (i in c(7,31)){
+ plotComplex(complexMembers=Gavin2006sorted$MBME[[i]],
+                    g=Gavin2006sub,
+                    VBs = unbiasedBaits,
+                    VPs=viablePrey,
+                    geneName=FALSE)
+ }
```

## 4.5 Feature Estimation: Physical Interaction Summarization

Finally, we summarize the overlay of Y2H physical interaction data on protein complex estimates derived from AP-MS data. In this example, we use the function `plotComplex` from *apComplex* to illustrate the set of detected physical interactions from the [30] (Ito01) and [15] (Uetz00-1) Y2H data sets for one of the protein complex estimates from the Gavin06 AP-MS data.

```
> y2hLargeScaleIU = y2hLSLocusNamesNodes[c(1,3)]
> y2hviableProteins = lapply(y2hLargeScaleIU,
+   FUN=idViableProteins)
> par(mfrow=c(1,2))
> allComplexMembers = Gavin2006sorted$MBME[["MBME31"]]
> for (i in 1:2){
+
+     thisy2hGraph = y2hLargeScaleIU[[i]]
+     subComplexMembers = intersect(allComplexMembers,
+       nodes(y2hLargeScaleIU[[i]]))
+     plotComplex(complexMembers=subComplexMembers,
+                 g = y2hLargeScaleIU[[i]],
+                 VBs = y2hviableProteins[[i]]$VB,
+                 VPs = y2hviableProteins[[i]]$VP,
+                 geneName = FALSE)
+                 title(names(y2hLargeScaleIU[i]))
+ }
```

# Timing

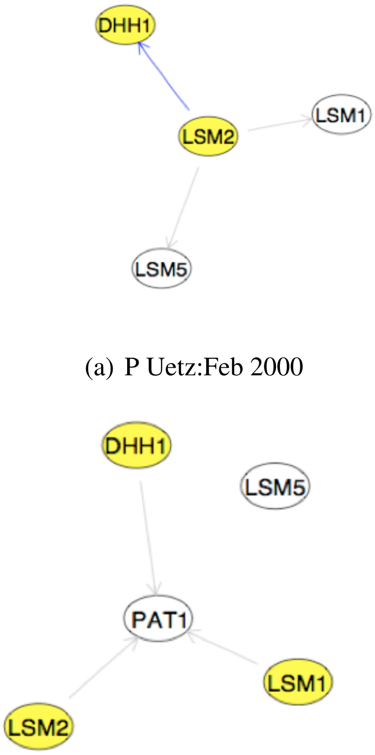
Depending on computational resources and the size of the data sets under study, some of the functions may take many minutes, possibly approaching hours, to run. The time needed to execute all the analytic commands within this protocol can be computed by calling the `proc.time` function twice, immediately prior and after the first and last commands. If we compute the difference as `time.elapsed`, we can see that:

```
> time.elapsed

   user  system elapsed
 71.252   9.062  84.533
```

The *user* and *system* times (in seconds) are determined by the operating system in use. The *elapsed* time documents the total time used for all processes. The `proc.time` function is inherently dependent on the computational power and memory available.

The function `findComplexes` from the package *apComplex* may be computationally intensive for larger data sets. All other functions in this protocol should yield results very efficiently.



(a) P Uetz:Feb 2000

(b) T Ito:April 2001

Figure 4.10: Here we show the second complex estimate from Figure 4.9 restricted to the direct interactions observed by a. Uetz et al. [15] and b. Ito et al [30]. A node for PAT1 is not included in the 'P Uetz:Feb 2000' over-lay graph since it was neither a viable bait a viable prey in this experiment. One possible interpretation for the discrepancy between the two over-laid graphs is that the interactions between LSM2 with both DHH1 and LSM1 within the 'P Uetz:Feb 2000' data is mediated via PAT1 though unobserved because of bait and prey sampling.

# Conclusions

Protein interactions are a central component of the organizational structure of a cell. Systematic protein interaction assays provide researchers with a method for elucidating the cellular and biological processes and the genes and proteins involved in them. In addition, combining this type of data with other genomic and proteomic information can help to understand and reconstruct the cellular machine. Their importance is underscored by the vast resources being consumed to collect, curate, and warehouse these data sets. It is an area of research that is crucial in the understanding of basic biology but still in its infancy.

At EMBL-EBI, I had the opportunity to work with a number of publicly available protein interaction data sets. While all past work tried to estimate the errors found within protein interaction data, confusion between the FPR and the FDR has lead to mis-interpretation of the data. In this dissertation, we developed a method to estimate the FPR and the FNR for protein interactions based on the data sets themselves. Using the lack of reciprocity as a measure of error within the data, we were able to simultaneously estimate the FPR, FNR, and the expected number of interactions given a sample of proteins. These three statistics lead us to a better understanding of many useful structures of the underlying protein interaction graphs such as a reasonable estimate of the distribution of each protein's interacting partners (its degree), and a better estimate of features such as protein complex composition.

My work points to the need for a sound statistical analysis of the reported protein interaction data before embarking on more high-level estimation tasks, such as integration of multiple, compatible datasets into a global estimate of the interactome, or using an estimated protein interaction network to facilitate the analysis of other data modalities, such as gene expression or RNAi data. While the eventual motivations of such analyses lie in a better understanding of biological phenomena, the emphasis of this work has been on statistical and computational methodology.

Indeed, as the amount of data increases, the computational component might prove an even greater challenge. Developing novel methods and implementing them as efficient algorithms will become increasingly more important in computational biology. The recent advent of high-throughput sequencing technologies is an example that currently generates much interest. Using such new data in addition to protein interactions to find structures in the protein interactomes such as maximal cliques or deciding if two interaction graphs align (graph isomorphisms) are computationally infeasible without approximation algorithms. This dissertation, therefore, is only a modest beginning.

# Appendix A

# Appendix - Coverage and Error Models

## A.1 Introduction

This chapter serves as an appendix for Chapter 2, *Coverage and Error Models*. It explains the steps taken to perform the analysis of protein interaction data described in that chapter. This chapter has been produced as a *reproducible document* which can be found in the extdata directory of the *ppiStats* package entitled *supp.Rnw*. That document contains all the computer instructions to reproduce the analysis and to create the figures, tables and numeric results of the paper. In addition, further analyses are produced that extend and support the main results described in the paper.

The production of the reproducible document employs the computational system and language R and the packages *ppiStats*, *ppiData*, and *yeastExpData*. You will need R version 2.11 or greater together with corresponding versions of the three packages and some other add-on packages that they depend upon and which can be obtain from CRAN or Bioconductor. To reproduce the computations shown here, you do not need to type them or copy-paste them from the PDF file; rather, you can take the file *supp.rnw* in the doc directory of the *ppiStats* package, open it in a text editor, run it using the R command Sweave; and if you wish, modify the program it to your needs. Alternatively, if you would simply like the code without the surrounding text, you can call Rtangle on the *supp.rnw* file to generate a script file called supp.R.

## A.2 Sampling and Coverage of the Interactome

### A.2.1 Analysis on the Bait/Prey Interactions

We addressed the issue of coverage initially by the viable bait and viable prey population observed in the experimental data sets. From the directed graphs, we created the two lists `viableBaits` and `viablePrey` by asking if each protein as a vertex had non-zero out- and in-degree respectively modulo self-loops (i.e homomers). From these two lists, we were able to find the set theoretic intersections of the viable baits (VB) and viable prey (VP) per experiment to ascertain the viable bait/prey (VBP) populations.

```
> getVBP <- function() {
+     vbp <- list()
+     for (g in bpExperimentNames) {
+         m = as(get(g), "matrix")
+         diag(m) = 0
+         stopifnot(identical(rownames(m), colnames(m)))
+         vbpEach = rownames(m)[(rowSums(m) > 0) &
+             (colSums(m) > 0)]
+         vbp[[g]] <- vbpEach
+     }
+     return(vbp)
+ }
> vbp <- getVBP()
```

Before we conduct any other statistical tests on the protein interaction data, we list the definitions of some standard statistical terms in Table A.1. Any of these terms used throughout this document (as well as the article *Coverage and Error Models in Protein Interaction Data by Directed Graph Analysis*) correspond to the given definitions.

| Error Statistics | | |
|---|---|---|
| True Positives | TP | Number of cases in which a true interaction is experimentally observed. |
| True Negatives | TN | Number of cases in which two proteins do not interact, their interaction is tested, but not observed. |
| False Positives | FP | Number of cases in which two proteins do not interact, but an interaction is reported by the experiment. |
| False Negatives | FN | Number of cases in which a true interaction is experimentally tested and not found. |
| True Tested Interactions | P | TP+FN |
| True Tested Non-interactions | N | TN+FP |
| False Positive Rate | $p_{\mathrm{FP}}$ | Probability that a truly absent interaction is detected. It can be estimated by FP / N. |
| False Negative Rate | $p_{\mathrm{FN}}$ | Probability that a true interaction is not detected. It can be estimated by FN / P. |
| Sensitivity | | Probability that a true interaction is detected. It can be estimated by TP / P. |
| Specificity | | Probability that a truly absent interaction is not detected, estimated by TN / N. |
| False Discovery Rate | FDR | Informally, the expected value of FP/(TP+FP) [83]. |
| Positive Predictive Value | PPV | Probability that an observed interaction is indeed true. It can be estimated by TP / (TP+FP). |
| Negative Predictive Value | NPV | Probability that an observed non-interaction is truly absent. It can be estimated by TN / (TN+FN). |

Table A.1: Standard definitions of various error terms [62]. The probabilities are conditional on that the interaction is tested.

.

### A.2.2 Hypergeometric Testing

We wanted to ascertain if the viably tested proteins showed signs of being affected by a coverage bias in the experimental assay. To investigate, we used the conditional Hypergeometric tests described by [55] to test for over/under representation in GO categories. Using the R software packages *Category* and *GOstats*, we were able to asses these questions. For our purposes, we used a p-value cutoff at the $10^{-2}$ threshold. We were only interested in GO categories which contained at least 50 unique annotations as well. Both these parameters can be set by the user, and those familiar with the R programming language are free to manipulate these parameters within the R scripts.

The code written to conduct these Hypergeometric tests can also be found in the Scripts directory of the ***ppiStats*** package. The file `hgGO.R` is a script file which computes the conditional Hypergeometric testing on the GO directed acyclic graph (DAG). The file `hgPfam.R` computes the Hypergeometric testing on Pfam domains. Running these scripts will generate several *.html* files which provide the results to the Hypergeometric analysis.

## A.3 Systematic Bias

### A.3.1 Probability model

For a protein $\rho$ from VBP, we want to construct a probability model for the joint distribution of $N_R$, the number of reciprocated edges, $N_I$, the number of unreciprocated in-edges, and $N_O$, the number of unreciprocated out-edges, given the true degree $\delta^*$ and the parameters $p_{\text{FP}}$, $p_{\text{FN}}$ and $N$, the number of interesting proteins.

We will use the shortcut $N_U = N_I + N_O$ for the total number of unreciprocated edges, and $\Theta = (\delta^*, p_{\text{FP}}, p_{\text{FN}}, N)$ for the parameters.

We consider

$$
\begin{aligned}
& P\left(N_R = n_r, N_I = n_i, N_O = n_o\,;\ \Theta\right) \\
= \quad & P(N_I = n_i, N_O = n_u - n_i \mid N_U = n_u, N_R = n_r\,;\ \Theta) \\
& \times P(N_U = n_u, N_R = n_r\,;\ \Theta) \qquad \text{(A.1)}
\end{aligned}
$$

The decomposition of $P$ on the right hand side will be useful.

For convenience, we suppress the index $\rho$ in our notation, but please keep in mind that the parameter $\delta^* \equiv \delta^*_\rho$ depends on $\rho$, and that $N_R$, $N_I$, $N_O$ and $N_U$ are random variables that depend on $\rho$. $N$ is an experiment-wide parameter, and we also consider $p_{\mathrm{FP}}$ and $p_{\mathrm{FN}}$ to be experiment-wide; although some of what follows might also apply to a model where $p_{\mathrm{FP}}$ and $p_{\mathrm{FN}}$ depend on $\rho$, if there were data to estimate them.

We will now make some modeling assumptions. If we find that the data for a particular protein does not concur well with these assumptions, we will consider it subject to systematic error.

### Symmetry

The first assumption is that of symmetry, that is, equality of the distributions of $N_I$ and $N_O$.

$$N_I =_d N_O \qquad \text{(A.2)}$$

and in particular

$$(N_I \mid N_U = n_u) \sim \mathrm{B}(n_u, \frac{1}{2}). \qquad \text{(A.3)}$$

This gives us the first term on the RHS of (A.1). The remarkable thing is that it depends on $n_u$, but not on any of the parameters! Now for the second term.

### Decomposition

We can decompose $N_R$ and $N_U$ into those counts that originate from real interactions (i. e. that are true) and those that originate from false positive measurements.

$$N_R = N_R^v + N_R^f \tag{A.4}$$
$$N_U = N_U^v + N_U^f \tag{A.5}$$

The false positives are easy:

$$\begin{aligned} N_R^f &\sim \mathrm{B}(N - \delta^* - 1,\, p_{\mathrm{FP}}^2) \\ N_U^f &\sim \mathrm{B}(N - \delta^* - 1,\, 2p_{\mathrm{FP}}(1 - p_{\mathrm{FP}})) \end{aligned} \tag{A.6}$$

The ones that originate from a real interaction follow a multinomial distribution

$$\begin{aligned} &P(N_R^v = n_r^v,\, N_U^v = n_u^v \,|\, \Theta) \\ &= (1-p)^{2n_r^v} \cdot (2p(1-p))^{n_u^v} \cdot p^{2n_{\mathrm{none}}^v} \cdot \frac{\delta^*!}{n_r^v! n_u^v! n_{\mathrm{none}}^v!} \end{aligned} \tag{A.7}$$

where for notational convenience we used the abbreviations $n_{\mathrm{none}}^v = \delta^* - n_r^v - n_u^v$ and $p \equiv p_{\mathrm{FN}}$.

The density function of the second term on the RHS of (A.1) can then be obtained by convolution of (A.6) and (A.7). For each value of the parameters $\Theta \equiv (N, \delta^*, p_{\mathrm{FP}}, p_{\mathrm{FN}})$, this is a 2D matrix with infinite numbers of rows and columns, corresponding to $n_r$ and $n_u$. Most of the probability mass, however, is concentrated within a bounded range. Furthermore, we will restrict our attention to values of $\delta^*$ between 0 and $\delta^*_{\max}$, depending on the data set. This is implemented in the function `nullDistDoublyTestedEdges` in the package *ppiStats*.

### Using in/out asymmetry to identify baits that are likely to be subject to systematic errors

We now use Equation (A.3) to assign a $p$-value to each protein. For a protein with unreciprocated degrees $(n_{out}, n_{in})$, the $p$-value is

$$\begin{aligned} p(n_{out}, n_{in}) &= P(\min\{N_I, N_O\} \leq \min\{n_i, n_o\}) \\ &= \max\{2P(N_I \leq \min\{n_i, n_o\}), 1\} \end{aligned} \tag{A.8}$$

This is computed by the function `assessSymmetry` which is also contained in the R package *ppiStats*. In addition, the function calculates the contours of the function $p$ in the $(n_{out}, n_{in})$-plane. These will be used in the plots.

Now we are ready to apply the symmetry $p$-values, and we will create an environment, `bpRed` containing the reduced set of data with only proteins with $p$-values larger than or equal to p-value threshold of $10^{-2}$.

Figures A.1–A.7 each show three characteristics of the specified data set. The top graph shows the out-degree ($n_{out}$) against the in-degree ($n_{in}$) for each VBP. The middle graph takes the root of these two statistics to decrease the spread for a better visual interpretation. Because several proteins have both the same out-degree and in-degree, we have used the `jitter` function to show the proportion of proteins for a given ($n_{out}$, $n_{in}$) point. The bottom plot shows the frequency of proteins at a given p-value based on the equation A.8.

### A.3.2 Logistic Regressions

For a protein with $n_i$ unreciprocated in-edges and $n_o$ unreciprocated out-edges, we expect

$$n_i \mid n_u \quad \sim \mathcal{B}\left(n_u, \frac{1}{2}\right)$$

if false positive and false negative errors are independent of a protein's properties. Let $p$ be the true probability ($H_0 : p = \frac{1}{2}$) for any particular protein.

- Perform binomial tests for $H_1 : p < \frac{1}{2}$ and $H_1 : p > \frac{1}{2}$ for each protein (in each experiment)
- Use test outcomes as responses to fit logistic regressions with abundance and CAI as predictor [1].

Regression is restricted to the subgraph of proteins that are VBP.

There has been a number of research articles that point to the relationship between CAI and protein abundance. To verify this fact, we computed both the Pearson and Spearman correlation coefficients between CAI and three sets of abundance data: 1. a general measure of abundance in the yeast cell, 2. the mean measure of abundance of a yeast cell in YEPD medium, and 3. the mean measure of abundance of a yeast cell in a SD medium (Table A.2). The Spearman correlation seems to be the more accurate measure, not because it is larger, but because the relationship between CAI and abundance is not linear [58] (cf Figures A.8 and A.9).

We plotted the values of each protein's CAI value (log) against the three sets of measured abundance data to visualize this association (cf Figure A.8 and Figure A.9). Those proteins which are likely affected by a systematic bias in the Gavin et al's [27] data are colored red in Figure A.8; proteins affected in Krogan et al's [28] are colored red in Figure A.9. The most interesting fact is that the measured protein abundances in SD medium have the highest correlation with CAI. This seems to suggest that the reference set of genes used to compute all

[1] We actually use logarithm (base 2) of abundance and CAI as predictor since that has a much more symmetric distribution.

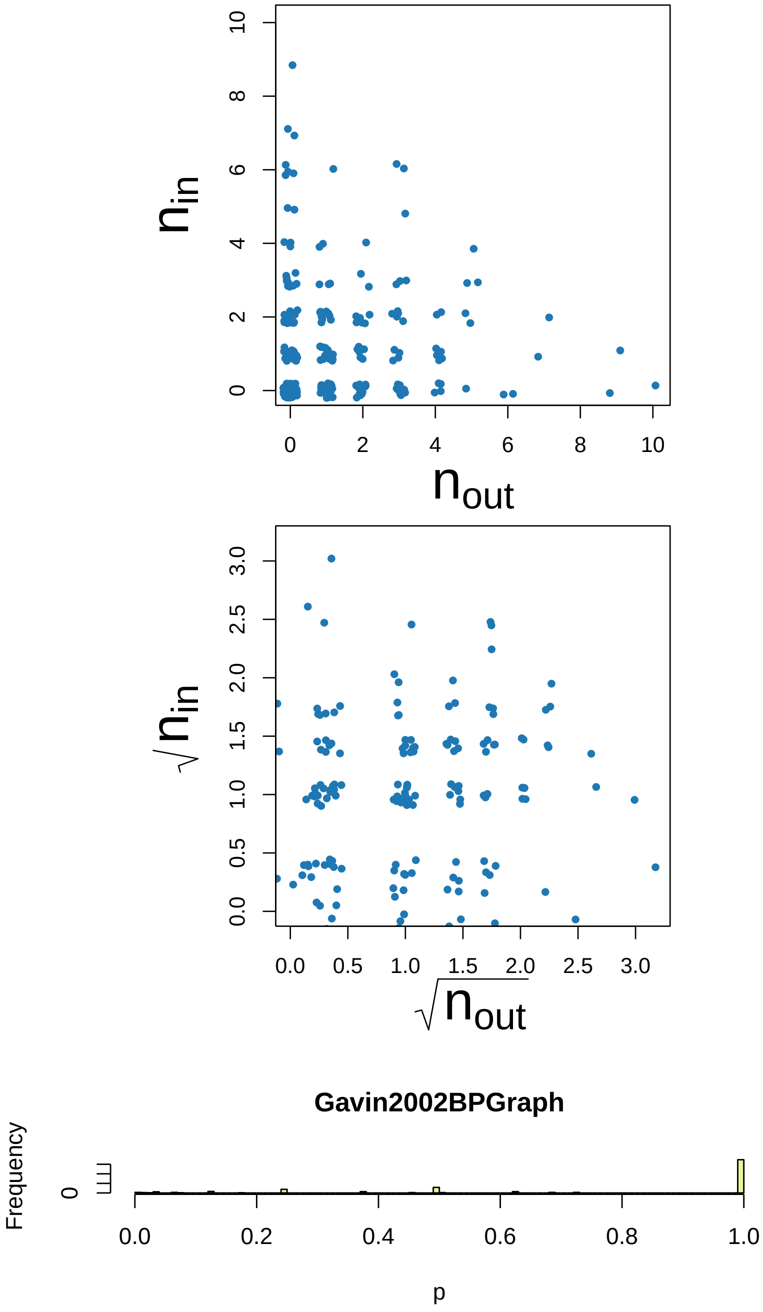


Figure A.1: Scatterplots of in- and out-degree and symmetry *p*-values for Gavin2002BPGraph

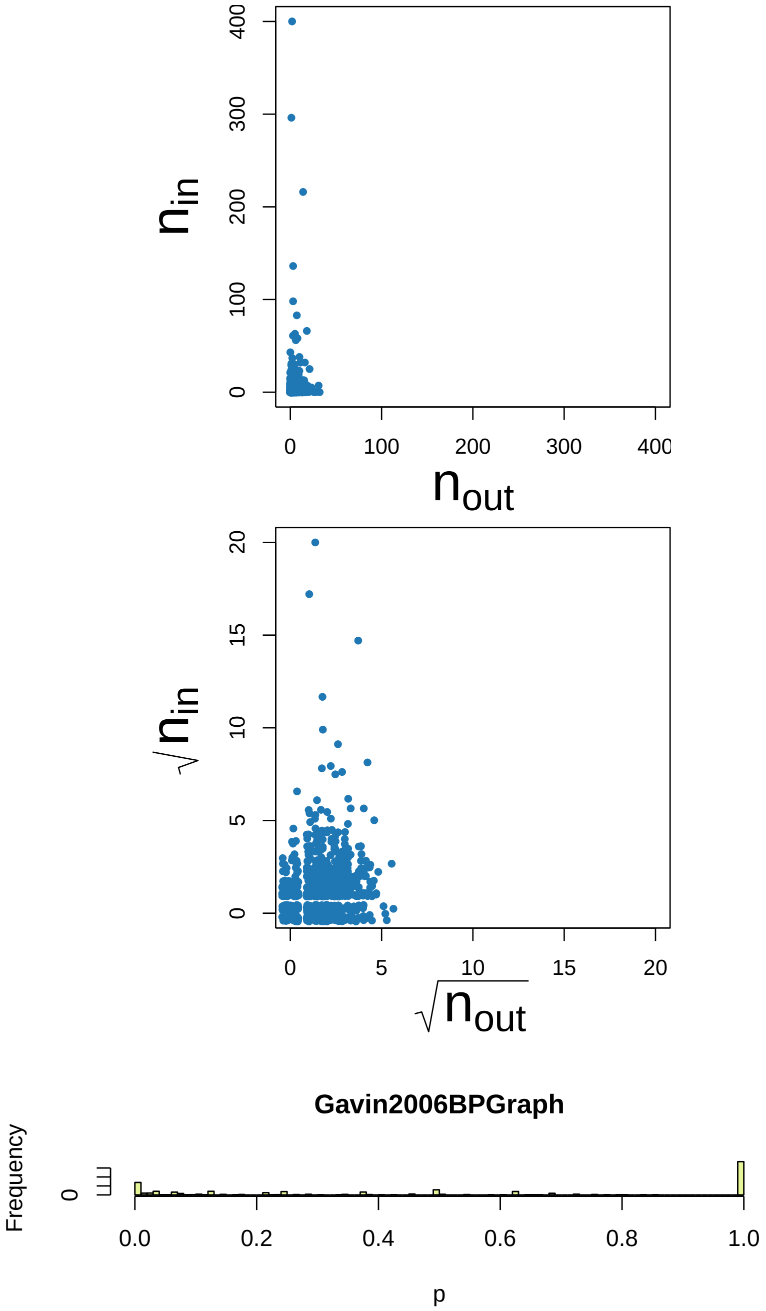


Figure A.2: Scatterplots of in- and out-degree and symmetry *p*-values for Gavin2006BPGraph

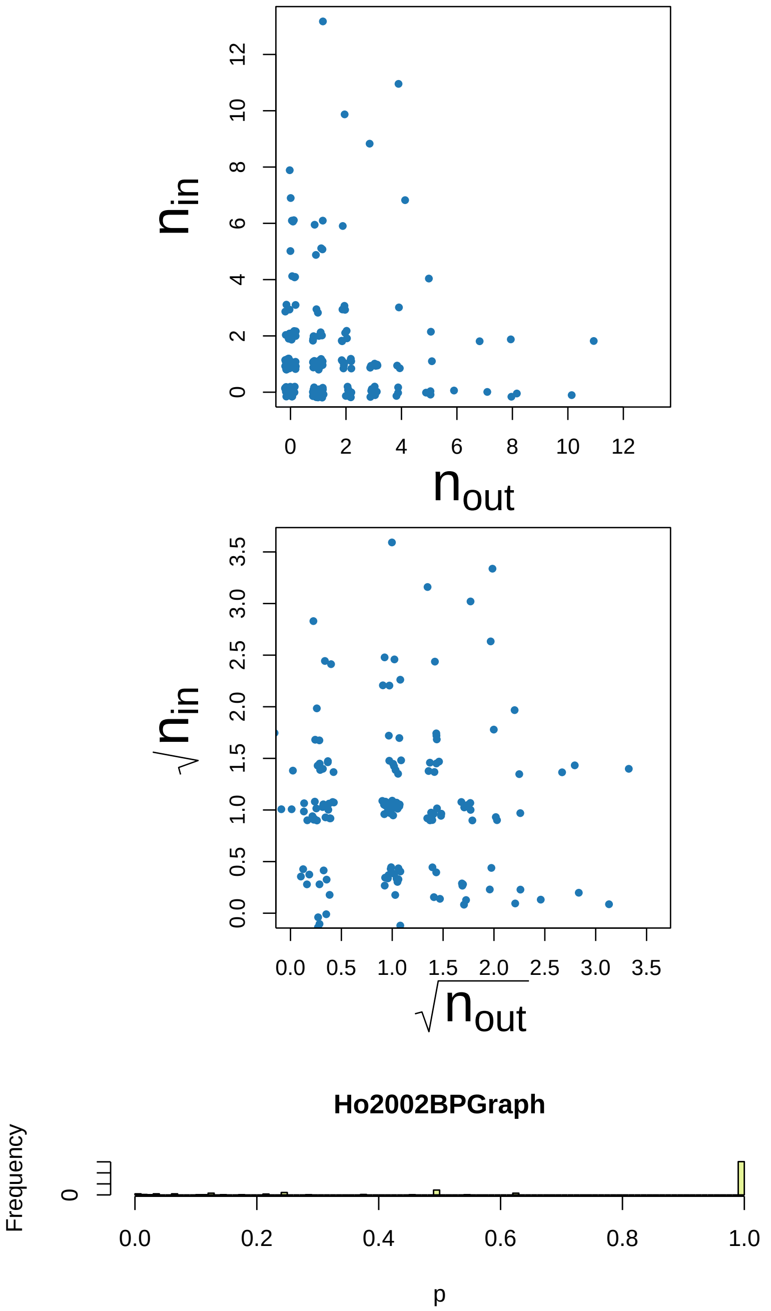


Figure A.3: Scatterplots of in- and out-degree and symmetry $p$-values for Ho2002BPGraph

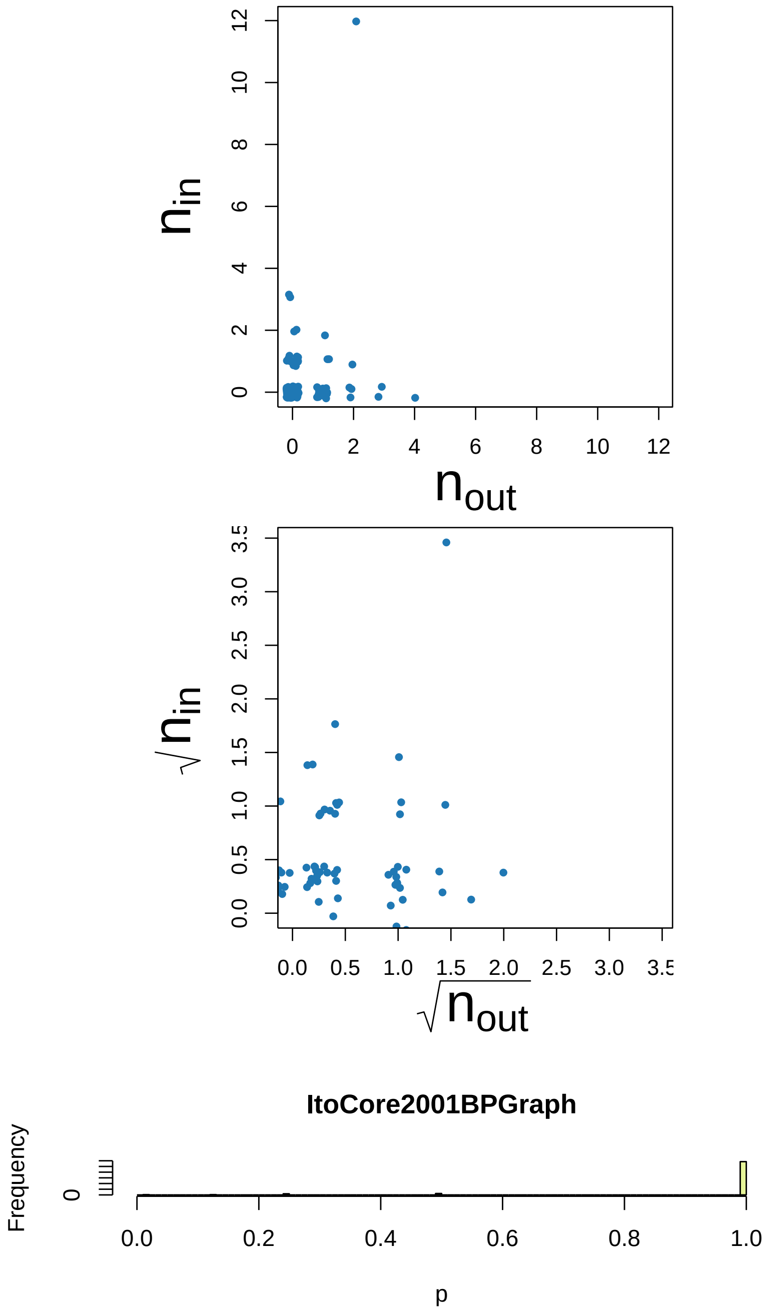


Figure A.4: Scatterplots of in- and out-degree and symmetry $p$-values for ItoCore2001BPGraph

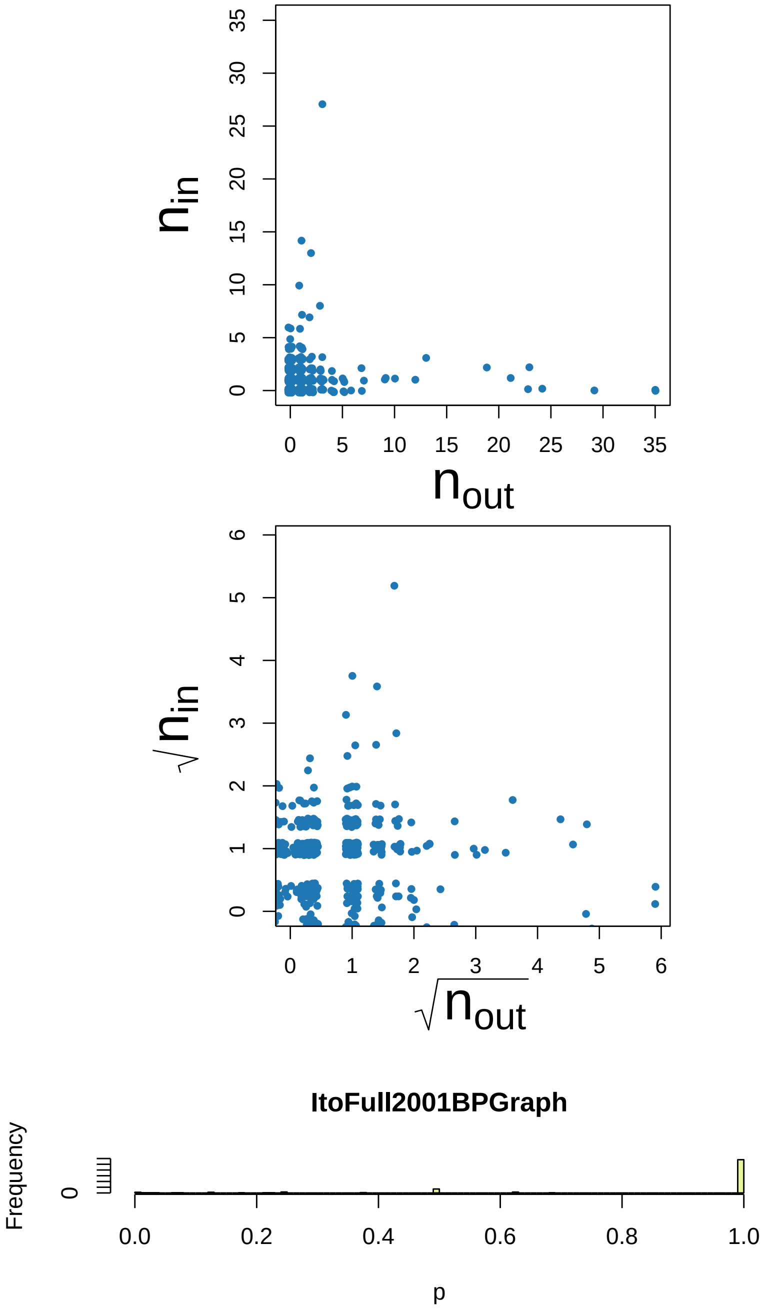


Figure A.5: Scatterplots of in- and out-degree and symmetry $p$-values for ItoFull2001BPGraph

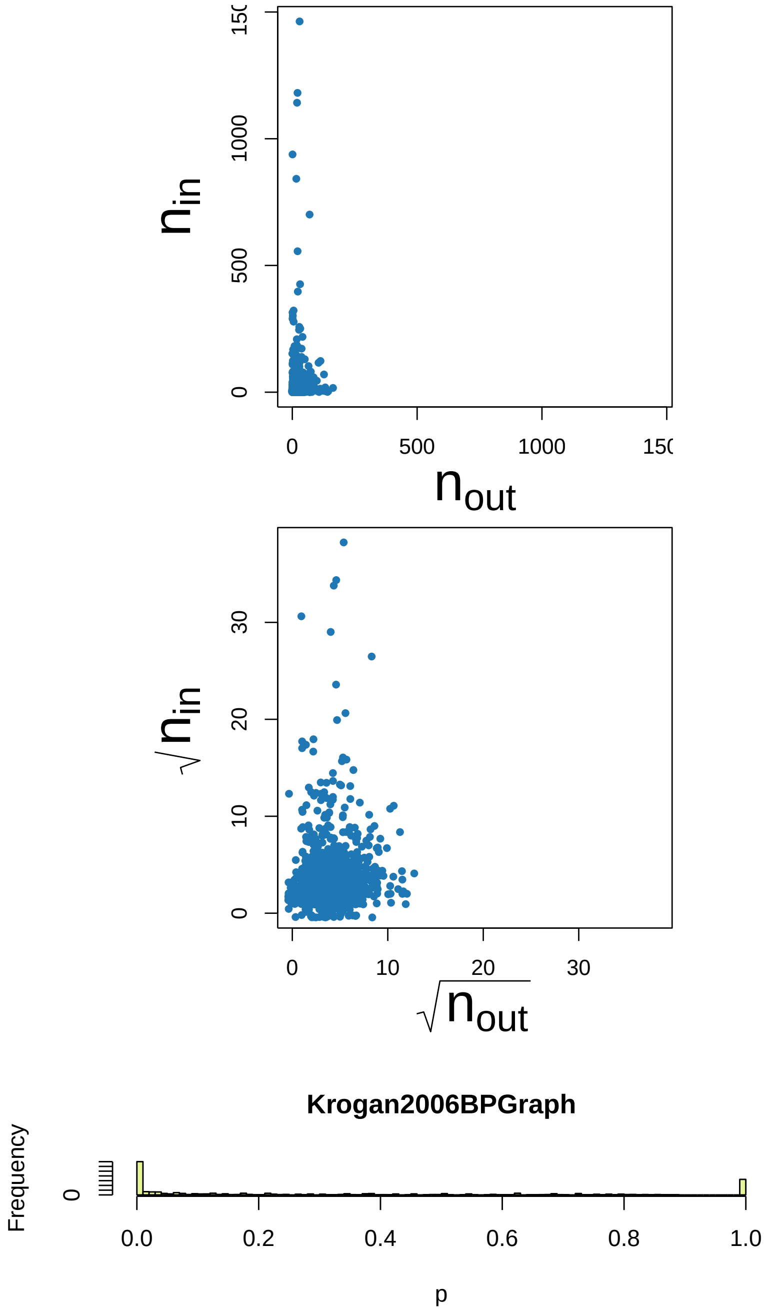


Figure A.6: Scatterplots of in- and out-degree and symmetry $p$-values for Krogan2006BPGraph

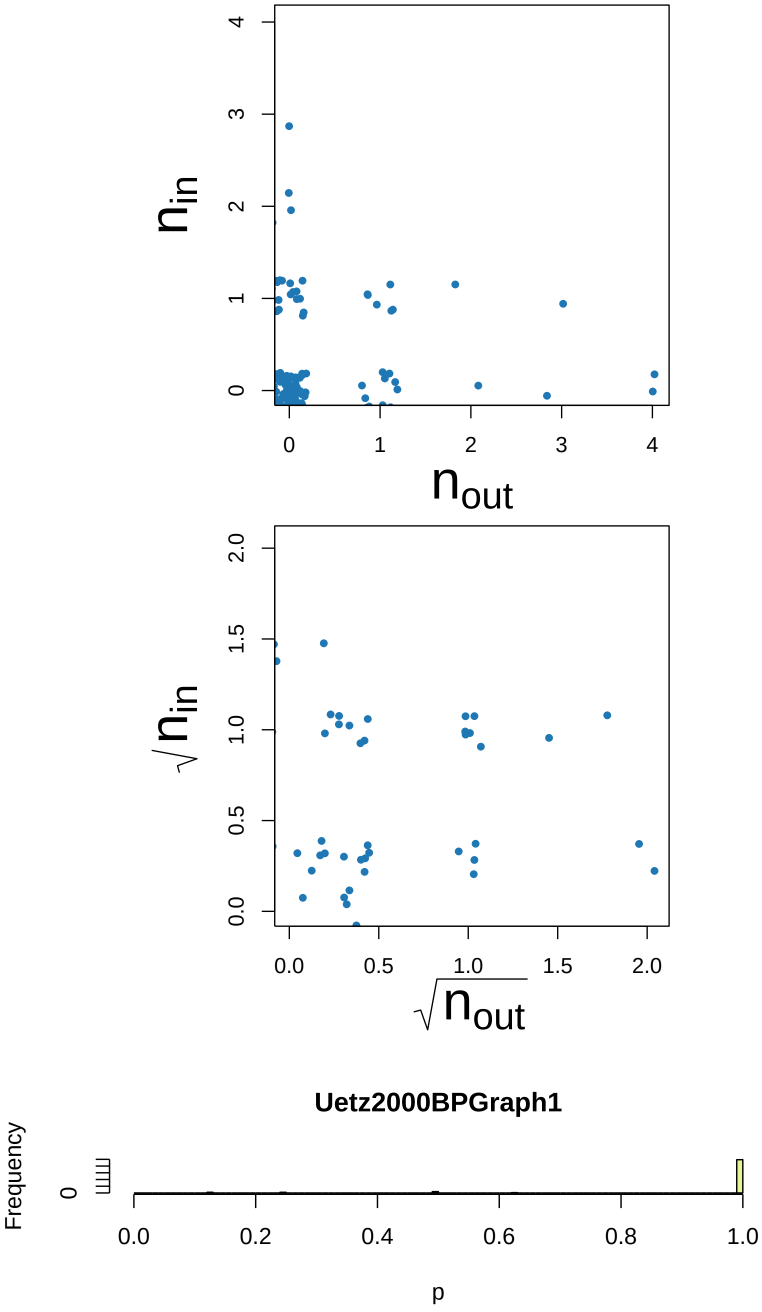


Figure A.7: Scatterplots of in- and out-degree and symmetry *p*-values for Uetz2000BPGraph1

|  | General Abundance | YEPD | SD |
|---|---|---|---|
| Pearson | 0.48 | 0.53 | 0.55 |
| Spearman | 0.54 | 0.65 | 0.66 |

Table A.2: This table gives both the Pearson and Spearman correlation between CAI against three distinct protein abundance data sets: 1. General Abundance, 2. Abundance in the YEPD medium, 3. Abundance in the SD medium. An interesting observation is that the highest correlation found is between CAI and protein abundance in the SD medium.

CAI might be highly expressed under SD medium. In addition, the relationship between the systematic bias with CAI and protein abundance becomes much more apparent (more so with [28]).

### A.3.3 Fisher's Exact Test Across Experiments

Next we wanted to ascertain if the protein subset ($S_1$) that was affected by a systematic bias in one experiment is related to the protein subset ($S_2$) affected by a systematic bias in another experiment. There are two ways to generate the subsets $S_1$ and $S_2$. The first methods generates these sets in an independent manner; the Binomial model is applied to each experimental data set generating a subset $S_i$ per experiment $i$. These subsets can be compared by restricting to the set of common $VBP$ of the two experiments. The second method generates $S_1$ and $S_2$ by first restricting to the common $VBP$ (denoted by $X$) of experiment 1 and experiment 2. Then the subset $S_1$ is generated by applying the Binomial model to the data set of Experiment 1 restricted only to $X$, or to use graph theoretic terms, using the node induced subgraph generated by $X$. $S_2$ is generated in the same manner with the data set of experiment 2. We compare the protein subsets $S_1$ and $S_2$ using both methods.

To investigate this relationship, we created three $2 \times 2$ tables. Only two data sets [27, 28] contained sufficient data points for this analysis. The $2 \times 2$ tables were created where the overall universe is restricted to $X = \text{VBP}_{[27]} \cap \text{VBP}_{[28]}$. [27] index the rows; [28], the columns. In the $(2, 2)$-entry of each table, we count the number of common proteins affected by a bias in both experiments ($|S_1 \cap S_2|$);

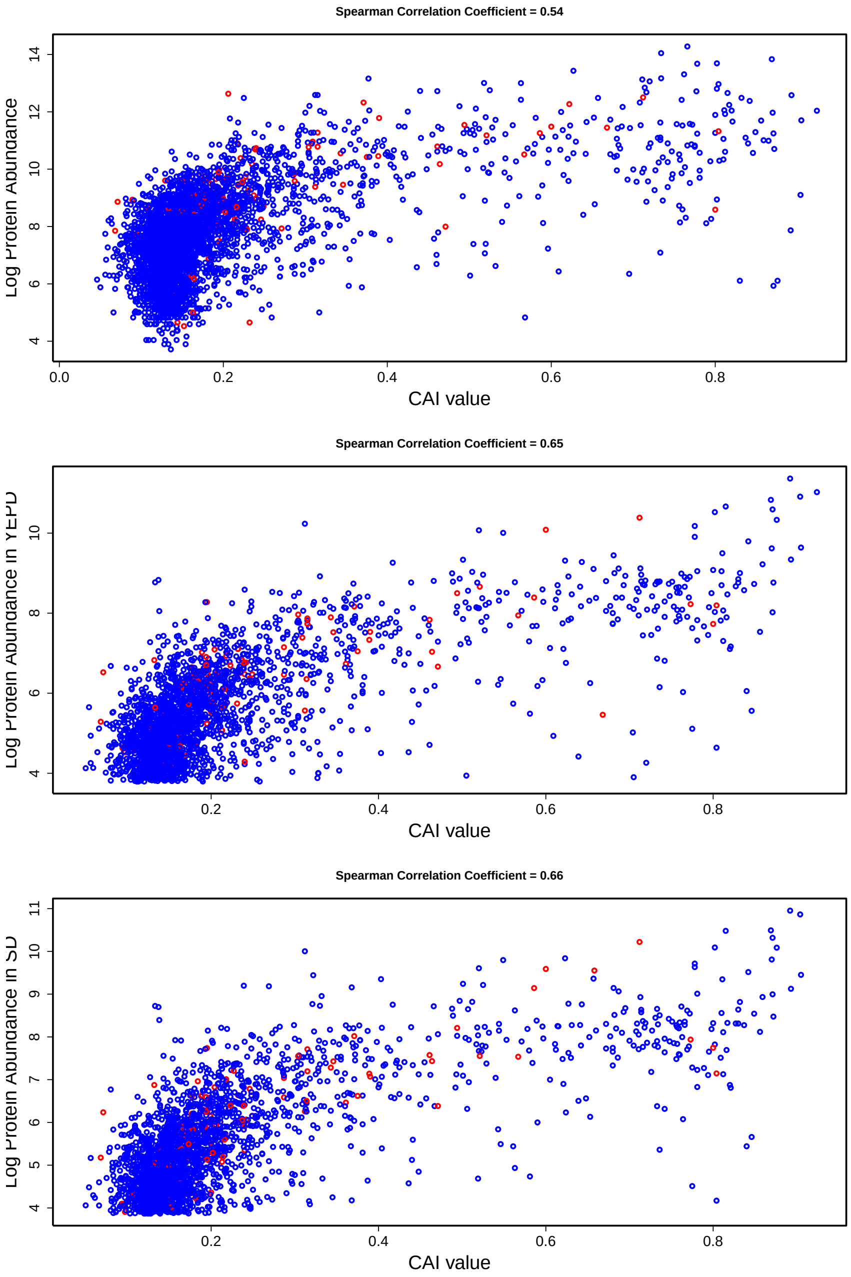


Figure A.8: Plots of CAI against log of the three measured abundance datasets (wild type, YEPD medium, and SD medium) for *Gavin et al.* [27] data. We colored those proteins found to be affected by a systematic bias (as described in the text) red and all other proteins blue. These plots show that the measured protein adundances in SD medium have the strongest correlation with CAI.

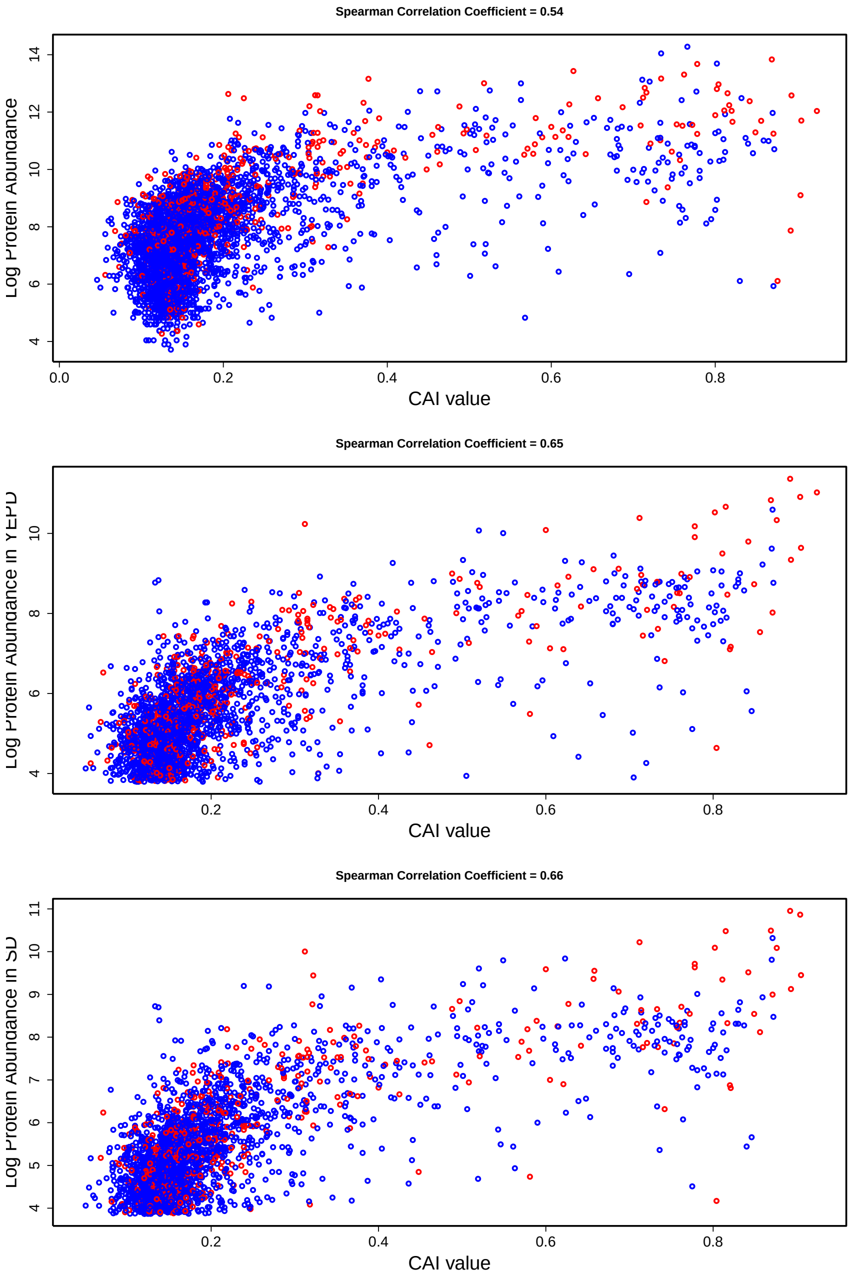


Figure A.9: Plots of CAI against log of the three measured abundance data sets for the *Krogan et al.* [28] data. We colored those proteins found to be affected by a systematic bias red and all other proteins blue. This plot shows a more apparent relationship between the systematic bias with the CAI as compared with the *Gavin et al.* data.

in $(1,2)$-entry, we count the number affected in [27] only ($|S_1 \setminus S_2|$); in $(2,1)$, the number in [28] only ($|S_2 \setminus S_1|$); and in $(1,1)$, the number not affected in both $(|S_1^c \cap S_2^c|)$. We can create three separate 2 × 2 tables based on which Binomial test we use:

- Number of proteins identified by the two-sided Binomial test.
- Number of proteins identified by the one-sided Binomial test where in-degree is much larger than out-degree.
- Number of proteins identified by the one-sided Binomial test where out-degree is much larger than in-degree.

```
> tab2Way

      0   1
  0 391 228
  1  57  57

> tab1WayIN

      0   1
  0 625  63
  1  33  12

> tab1WayOUT

      0   1
  0 483 181
  1  40  29
```

We can use these three tables as the parameters for Fisher's Exact test (again a Hypergeometric test), and see the results:

```
> fisher.test(tab2Way)
```

```
        Fisher's Exact Test for Count Data

data:  tab2Way
p-value = 0.009
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.123257 2.614963
sample estimates:
odds ratio
  1.713527

> fisher.test(tab1WayIN)

        Fisher's Exact Test for Count Data

data:  tab1WayIN
p-value = 0.0009893
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.608501 7.593418
sample estimates:
odds ratio
  3.597664

> fisher.test(tab1WayOUT)

        Fisher's Exact Test for Count Data

data:  tab1WayOUT
p-value = 0.01189
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.119609 3.304052
sample estimates:
```

```
odds ratio
  1.932813
```

The previous table used each individual distinct $VBP$ population to generate the sets $S_1$ and $S_2$. We then restricted to $X$ to calculate two way tables based on these protein subsets. In the following tables, we first restrict to the node induced subgraph of $X$ for each experiment, and then generate the sets $S_1$ and $S_2$. We then create two way tables based on these protein subsets to determine the level of independence.

```
> ta2Way

      0   1
  0 519 115
  1  70  29

> ta1WayIN

      0   1
  0 652  46
  1  27   8

> ta1WayOUT

      0   1
  0 590  79
  1  53  11

> fisher.test(ta2Way)

        Fisher's Exact Test for Count Data

data:  ta2Way
p-value = 0.01379
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
 1.114654 3.073070
sample estimates:
odds ratio
  1.867942

> fisher.test(ta1WayIN)

        Fisher's Exact Test for Count Data

data:  ta1WayIN
p-value = 0.002547
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.554209 10.167306
sample estimates:
odds ratio
  4.185891

> fisher.test(ta1WayOUT)

        Fisher's Exact Test for Count Data

data:  ta1WayOUT
p-value = 0.2297
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.699581 3.161824
sample estimates:
odds ratio
  1.548972
```

Because we are looking for reproducibility across experiments, the two-way tables `tab2Way` and (ta2Way) as well as the results from their corresponding Fisher's exact test are not particularly relevant. Instead, we should focus on the one-sided Binomial tests and see if we such artifacts are reproducible across experiments. For those proteins where in-degree dominates out-degree, i.e. proteins tested in `tab1WayIN` and `ta1WayIN`, we see an exceptionally small p-value for the former and a reasonably significant p-value for the latter, and so we should probably reject the null hypothesis that these $S_1$ is independent of $S_2$. For those proteins where out-degree dominates in-degree, `tab1WayOUT` and `ta1WayOUT`, we see a less significant p-value in the former but an incredibly small p-value in the latter, and so again we should reject independent null hypothesis.

## A.4 Stochastic Error Analysis: Estimation of $p_{\text{FP}}$ and $p_{\text{FN}}$ by the method of moments

### A.4.1 Derivation

| | | |
|---|---|---|
| $\binom{N}{2}$ | | The total number of possible interactions (excluding homomers) |
| $n$ | | the true number of interactions |
| $m$ | $= \binom{N}{2} - n$ | the true number of non-interactions |
| $X_1$ | | observed number of reciprocated edges |
| $X_2$ | | observed number of non-edges |
| $X_3$ | $= n + m - X_1 - X_2$ | observed number of unreciprocated edges |

We have

$$
\begin{aligned}
E[X_1] &= n\,(1-p_{\text{FN}})^2 + m\,p_{\text{FP}}^2 && \text{(A.9)}\\
E[X_2] &= n\,p_{\text{FN}}^2 + m\,(1-p_{\text{FP}})^2 && \text{(A.10)}\\
E[X_3] &= 2n\,p_{\text{FN}}(1-p_{\text{FN}}) + 2m\,p_{\text{FP}}(1-p_{\text{FP}}) && \text{(A.11)}
\end{aligned}
$$

Only two of the three equations (A.9)–(A.11) are independent, any two of them imply the third. Our goal is to estimate $p_{\text{FP}}$, $p_{\text{FN}}$. We can replace the expectation values on the left side of Equations (A.9)–(A.11) by the observed sample values $x_1$, $x_2$, $x_3$. Since we do not know $n$, the above system of two independent equations for three variables defines a one-dimensional solution manifold.

We will parameterize that manifold by $n$ ($0 \leq n \leq \binom{N}{2}$) in $(p_{\text{FP}}, p_{\text{FN}})$-space. Relevant solutions are those for which $0 \leq p_{\text{FP}}, p_{\text{FN}} \leq 1$.

Consider that $n$ is given. Let us solve Equations (A.9)–(A.11) for $p_{\text{FP}}$ and $p_{\text{FN}}$. First, subtracting (A.12) $=$ (A.9) $-$ (A.10), we have

$$\begin{aligned} x_1 - x_2 &= n\,(1-2p_{\mathrm{FN}}) - m\,(1-2p_{\mathrm{FP}}) && \text{(A.12)} \\ \Leftrightarrow \quad p_{\mathrm{FN}} &= \frac{1}{2n}\left((x_2-m)-(x_1-n)+2m\,p_{\mathrm{FP}}\right) \\ p_{\mathrm{FN}} &= \frac{1}{2n}\left(\Delta + 2m\,p_{\mathrm{FP}}\right), && \text{(A.13)} \end{aligned}$$

where we have defined $\Delta := (x_2 - m) - (x_1 - n)$ for convenience. We can plug this expression for $p_{\mathrm{FN}}$ into Equation (A.10) and obtain

$$\underbrace{(n+m)}_{=:a} p_{\mathrm{FP}}^2 + \underbrace{(\Delta - 2n)}_{=:b} p_{\mathrm{FP}} + \underbrace{n + \frac{\Delta^2}{4m} - \frac{n}{m}x_2}_{=:c} = 0. \qquad \text{(A.14)}$$

The equation $a(p_{\mathrm{FP}})^2 + b(p_{\mathrm{FP}}) + c = 0$ is solved by

$$(p_{\mathrm{FP}})_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \qquad \text{(A.15)}$$

Hence, the problem is solved: for data $N$, $x_1$, $x_2$ (from these, $x_3$ is implied) and for all possible (unknown) $n = 0, 1, \ldots, \binom{N}{2}$ we can calculate $p_{\mathrm{FP}}$ via Equation (A.15), then $p_{\mathrm{FN}}$ via Equation (A.13). Only some of the theoretically possible values of $n$ will lead to admissible solutions for $p_{\mathrm{FP}}$ and $p_{\mathrm{FN}}$. This is exemplified in the following section.

### A.4.2 Application to the PPI datasets

We now take the experimental data sets obtained from [15, 23, 24, 26–28, 30–33] to obtain the 1-dimensional manifolds.

We plot each data set individually to ascertain the range for $p_{\mathrm{FP}}$ and for $p_{\mathrm{FN}}$. The result of this is plotted in Figure A.10 for the unfiltered data and in Figure A.11 for the filtered data.

Next we wanted to superimpose all experiments of the same type (i.e. the same system was used to determine the interactions) so that we can compare the solutions curves across experiments. We do this first on the set of VBP for each data set, and these are rendered in the top two plots. After, we filtered out the proteins likely to be affected by a systematic bias and calibrated the solution curves. These are rendered in the bottom two plots. The result of this is plotted in Figure A.12.

## A.5 Stochastic Error Analysis: Estimation of Unreciprocated and Reciprocated FP/FN Errors within the Measured Data

Using the Multinomial error model, we can estimate the expected number of unreciprocated and reciprocated false positive as well as false negative interactions. For these estimates, we use only the filtered set of data on the protein interactions (let $N$ be the number of proteins in the filtered set). We begin by estimating the FP errors. To calculate the expected number of FP observations, we need estimates for $p_{\text{FP}}$ as well as $m$ (since FP is a property on $I^c$). To obtain these estimates, we assume that $p_{\text{FN}} = 0$ so that all errors will be strictly FP. This makes the estimate for $p_{\text{FP}}$ maximal, and the number of expected FP observations that we calculate will all be for the worst case scenario. Using the Multinomial error model, we generate $p_{\text{FP}}$'s for all the data sets. We can also generate the value for $m$ by assuming that $p_{\text{FN}} = 0$, but for the sake of convenience, we will approximate $m$ by $\binom{N}{2}$. Unreciprocated FP errors can be estimated by $2p_{\text{FP}}(1 - p_{\text{FP}})m$; reciprocated, by $p_{\text{FP}}^2 m$. Lastly we provide the number of observed unreciprocated and reciprocated interactions to serve as a reference. These estimates can be found in Table A.3

Similarly, we also use the Multinomial error model to estimate the expected number of unreciprocated and reciprocated FN observations. We estimate $p_{\text{FN}}$ and $n$ by assuming that $p_{\text{FP}} = 0$, and so, again we presume that all errors are strictly FN. This makes $p_{\text{FN}}$ maximal, and this also makes these estimates for FN errors in the worst case scenario. Similar to the FP estimates, unreciprocated FN errors

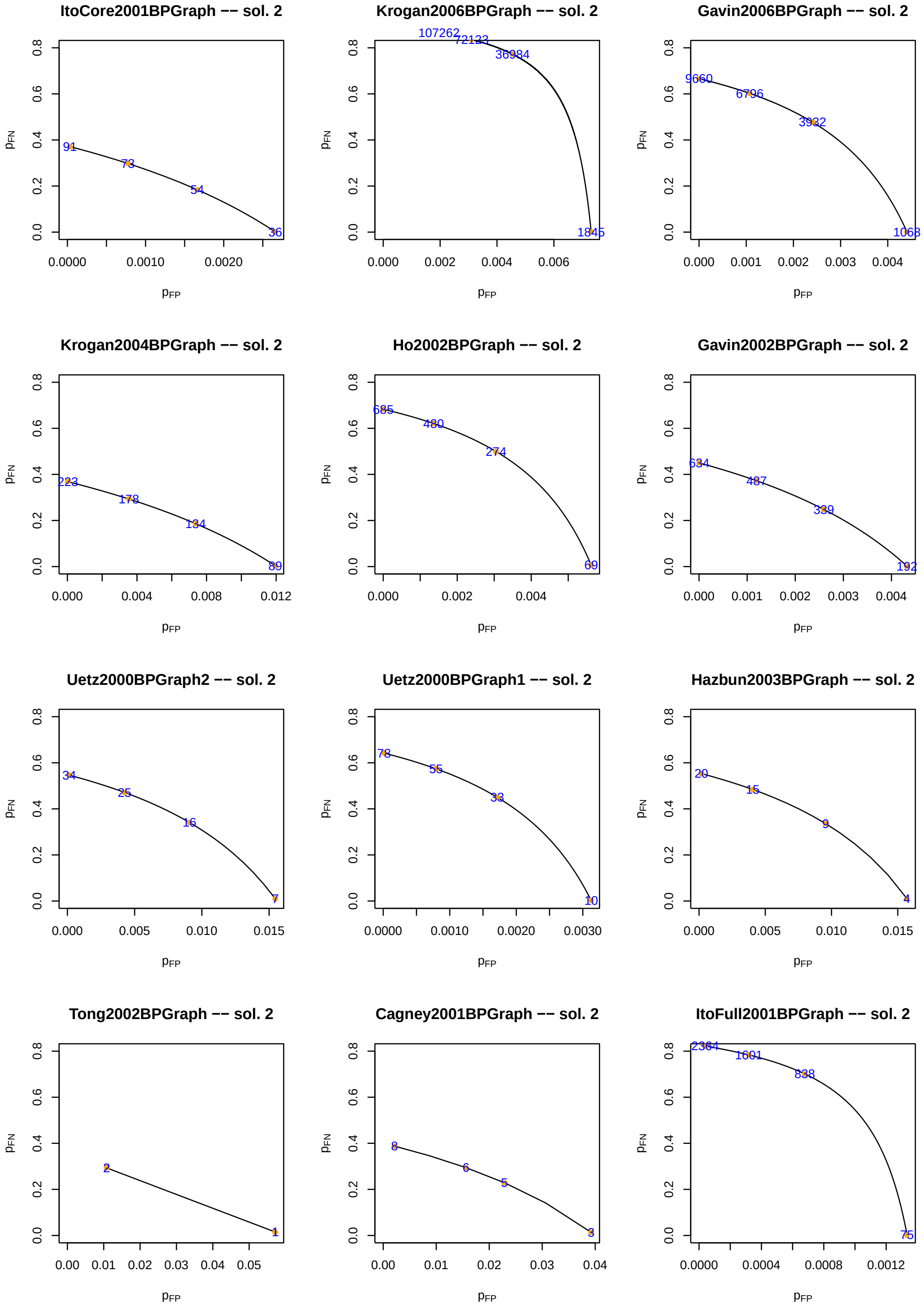


Figure A.10: The solution manifolds (One plot per data set, unfiltered data).

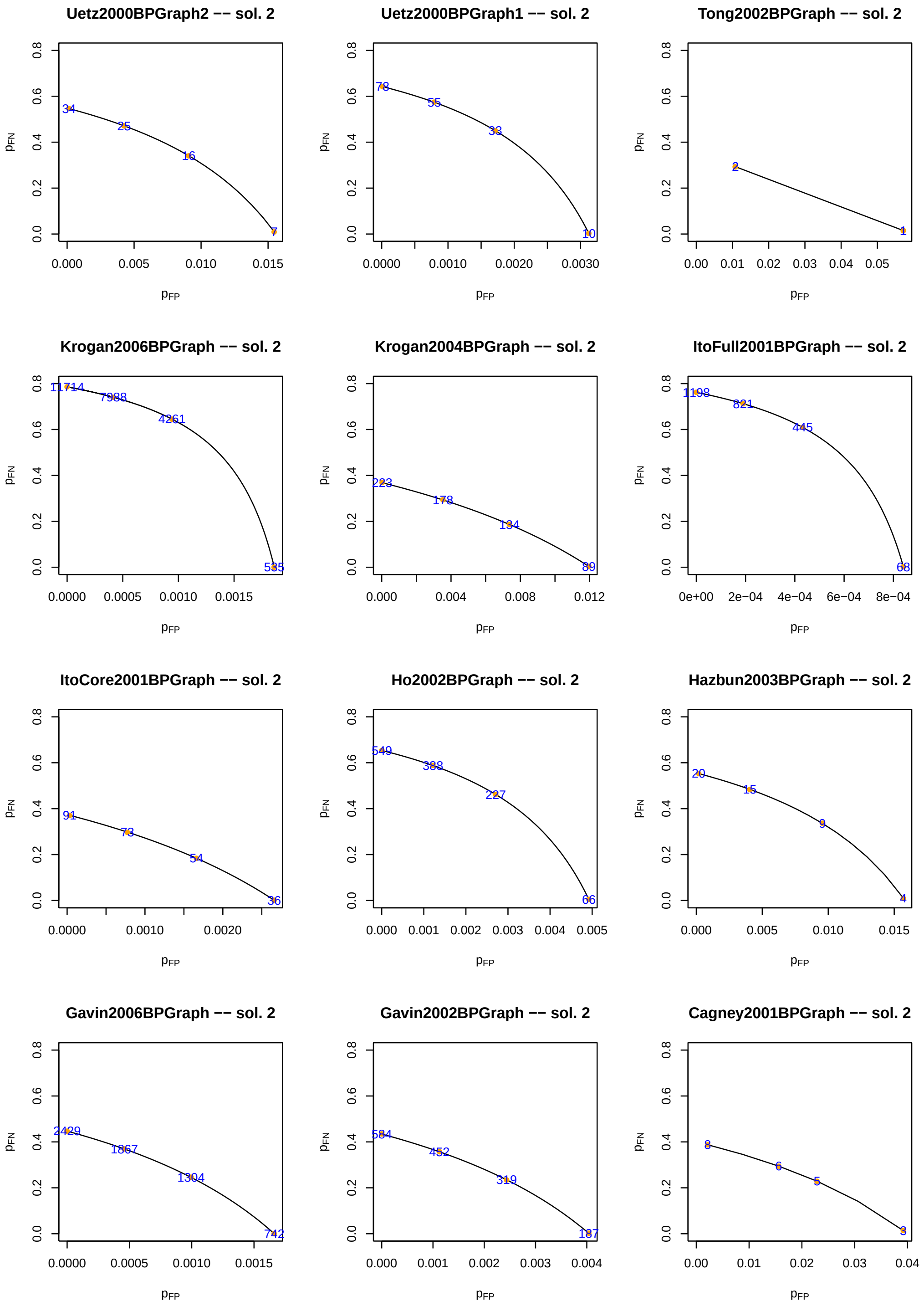


Figure A.11: The solution manifolds (One plot per data set, filtered data)

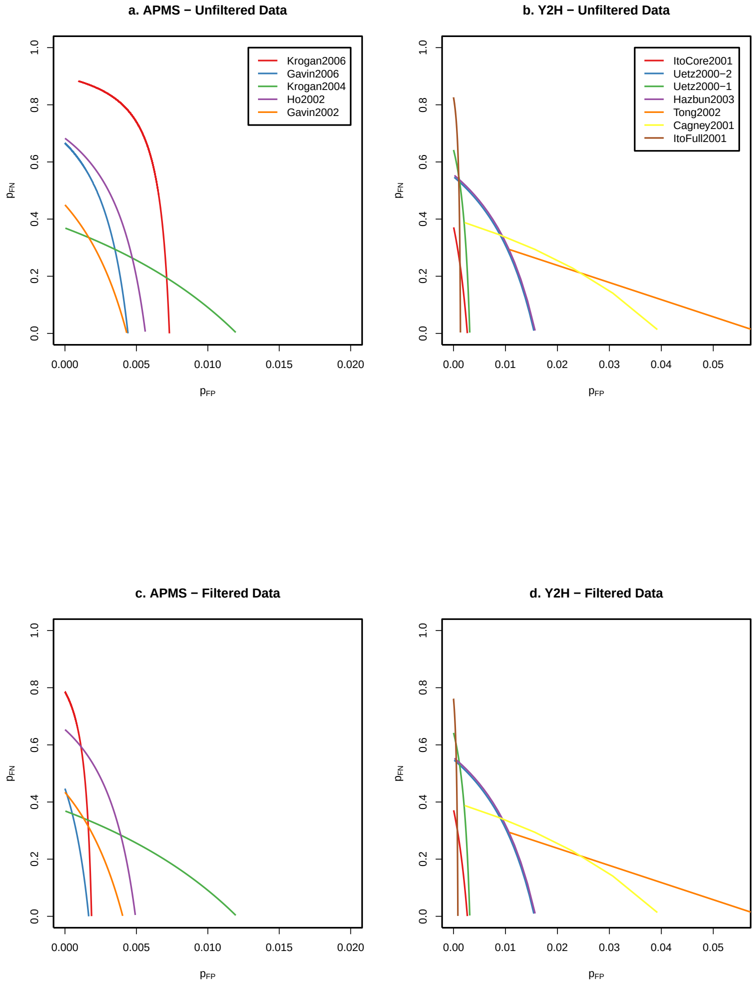


Figure A.12: These figures detail the error statistics for each of the data sets. Plot *a* generates the 1-dimensional solution curves for $(p_{\mathrm{FP}}, p_{\mathrm{FN}})$ parametrized by $n$ for the AP-MS data sets. Plot *b* generates similar 1-dimensional curves but for the Y2H data sets. Plots *c* and *d* recalculates these solution curves for the AP-MS and Y2H respectively. These recalculations are done by setting aside those interactions which appear to be affected by a systematic bias of the experimental assay. Having set aside those interactions, the range for $p_{\mathrm{FP}}$ is substantially constrained for the solution curves characterizing [27,28] implying that systematic errors may potentially have large effects on $p_{\mathrm{FP}}$.

|  | $N$ | $m$ | $p_{\mathrm{FP}}$ | $E[Z]$ | $E[Z_2]$ | $z$ | $z_2$ |
|---|---|---|---|---|---|---|---|
| ItoFull2001 | 720 | 258840 | 0.0008 | 414 | 0.17 | 435 | 68 |
| Cagney2001 | 11 | 55 | 0.04 | 4 | 0.09 | 4 | 3 |
| Tong2002 | 5 | 10 | 0.07 | 1 | 0.05 | 1 | 1 |
| Hazbun2003 | 26 | 325 | 0.016 | 10 | 0.08 | 10 | 4 |
| Uetz2000-1 | 108 | 5778 | 0.003 | 35 | 0.05 | 36 | 10 |
| Uetz2000-2 | 34 | 561 | 0.015 | 17 | 0.13 | 17 | 7 |
| Gavin2002 | 268 | 35778 | 0.004 | 285 | 0.57 | 287 | 187 |
| Ho2002 | 226 | 25425 | 0.005 | 253 | 0.64 | 249 | 66 |
| Krogan2004 | 95 | 4465 | 0.012 | 106 | 0.64 | 104 | 89 |
| Gavin2006 | 852 | 362526 | 0.0017 | 1230 | 1.1 | 1201 | 743 |
| Krogan2006 | 1458 | 1062153 | 0.0019 | 4029 | 3.8 | 3945 | 538 |
| ItoCore2001 | 128 | 8128 | 0.0025 | 41 | 0.05 | 43 | 36 |

Table A.3: Estimates of FP unreciprocated and reciprocated errors via the Multinomial error Model. $Z$ is the random variable associated with unreciprocated FP errors; $Z_2$ corresponds with reciprocated FP errors. $z$ and $z_2$ denote the observed number of unreciprocated and reciprocated interactions found in the data.

are estimated by $2p_{\mathrm{FN}}(1-p_{\mathrm{FN}})n$; reciprocated by $p_{\mathrm{FN}}^2 n$. We also provide the number of observed unreciprcated and reciprocated non-interacting protein pairs found within the data sets. These estimates can be found in Table A.4

|  | $N$ | $n$ | $p_{\text{FN}}$ | $E[W]$ | $E[W_2]$ | $w$ | $w_2$ |
|---|---|---|---|---|---|---|---|
| ItoFull2001 | 720 | 1200 | 0.76 | 438 | 693 | 435 | 259132 |
| Cagney2001 | 11 | 8 | 0.39 | 4 | 1 | 4 | 57 |
| Tong2002 | 5 | 2 | 0.30 | 1 | 0 | 1 | 11 |
| Hazbun2003 | 26 | 20 | 0.55 | 10 | 6 | 10 | 334 |
| Uetz2000-1 | 108 | 78 | 0.65 | 35 | 33 | 36 | 5822 |
| Uetz2000-2 | 34 | 34 | 0.55 | 17 | 10 | 17 | 571 |
| Gavin2002 | 268 | 584 | 0.44 | 288 | 113 | 287 | 35725 |
| Ho2002 | 226 | 649 | 0.68 | 282 | 300 | 249 | 25472 |
| Krogan2004 | 95 | 223 | 0.37 | 104 | 31 | 104 | 4423 |
| Gavin2006 | 852 | 2429 | 0.44 | 1197 | 470 | 1201 | 362209 |
| Krogan2006 | 1458 | 11744 | 0.80 | 3758 | 7516 | 3945 | 1062344 |
| ItoCore2001 | 128 | 100 | 0.38 | 47 | 14 | 43 | 8156 |

Table A.4: Estimates of FN unreciprocated and reciprocated errors via the Multinomial error Model. $W$ is the random variable associated with unreciprocated FP errors; $W_2$ corresponds with reciprocated FP errors. $w$ and $w_2$ denote the observed number of unreciprocated and reciprocated interactions found in the data.

|  | VB | CB | TB | VB/TB | VP | VBP | VP/VB | TI | TI/VB |
|---|---|---|---|---|---|---|---|---|---|
| ItoFull2001 | 1522 |  | 6604 | 0.23 | 2493 | 773 | 1.64 | 4524 | 2.97 |
| Cagney2001 | 19 |  | 31 | 0.61 | 40 | 11 | 2.11 | 54 | 2.84 |
| Tong2002 | 20 |  | 22 | 0.91 | 59 | 5 | 2.95 | 115 | 5.75 |
| Hazbun2003 | 66 |  | 100 | 0.66 | 1940 | 28 | 29.39 | 2524 | 38.24 |
| Zhao2005 | 1 |  | 1 | 1.00 | 90 | 0 | 90.00 | 90 | 90.00 |
| Uetz2000-1 | 508 |  | 6604 | 0.08 | 630 | 142 | 1.24 | 952 | 1.87 |
| Uetz2000-2 | 139 |  | 192 | 0.72 | 400 | 36 | 2.88 | 524 | 3.77 |
| Gavin2002 | 455 | 600 | 725 | 0.63 | 1178 | 271 | 2.59 | 3418 | 7.51 |
| Ho2002 | 493 | 589 | 1739 | 0.28 | 1316 | 231 | 2.67 | 3687 | 7.48 |
| Krogan2004 | 153 | 165 | 165 | 0.93 | 483 | 151 | 3.16 | 1132 | 7.40 |
| Gavin2006 | 1752 | 1993 | 6466 | 0.27 | 1790 | 991 | 1.02 | 19105 | 10.90 |
| Krogan2006 | 2264 | 2357 | 4562 | 0.50 | 5323 | 2226 | 2.35 | 63360 | 27.99 |
| ItoCore2001 | 455 |  | 6604 | 0.07 | 504 | 164 | 1.11 | 839 | 1.84 |

Table A.5: A general overview of seven Y2H and five AP-MS experiments: VB - the number of viable baits; CB - the number of cloned (hybridized) baits if available; TB - the total number of baits; VB/TB; VP - the number of viable prey; VBP - the number of proteins observed as both bait and prey; VP/VB, TI - the total number of interactions observed; TI/VB.

|  | Ito01 | Cagney01 | Tong02 | Hazbun03 | Zhao05 | Uetz00-1 | Uetz00-2 |
|---|---|---|---|---|---|---|---|
| Ito01 |  | 9 | 7 | 24 | 1 | 224 | 47 |
| Cagney01 | 28 |  | 0 | 0 | 0 | 7 | 3 |
| Tong02 | 34 | 0 |  | 0 | 0 | 4 | 7 |
| Hazbun03 | 855 | 14 | 25 |  | 0 | 15 | 12 |
| Zhao05 | 43 | 1 | 2 | 38 |  | 0 | 0 |
| Uetz00-1 | 389 | 14 | 22 | 272 | 15 |  | 36 |
| Uetz00-2 | 199 | 9 | 26 | 204 | 13 | 108 |  |

Table A.6: This table shows two distinct statistics on the pairwise comparison of the data sets. The values above the diagonal give the number of common viable baits, the values below the diagonal give the number of common viable prey.

|  | Gavin2002 | Ho2002 | Krogan2004 | Gavin2006 | Krogan2006 |
|---|---|---|---|---|---|
| Gavin2002 |  | 82 | 51 | 446 | 336 |
| Ho2002 | 517 |  | 25 | 223 | 286 |
| Krogan2004 | 300 | 246 |  | 122 | 151 |
| Gavin2006 | 1148 | 721 | 373 |  | 1131 |
| Krogan2006 | 1150 | 1277 | 478 | 1756 |  |

Table A.7: This table shows two distinct statistics on the pairwise comparison of the data sets. The values above the diagonal give the number of common viable baits, the values below the diagonal give the number of common viable prey.

# Appendix B

# Appendix - Degree Distribution

## Introduction

This chapter serves as the appendix for Chapter 3, *Modeling Interactions in Bait Prey Systems*. Contained here are mostly supplementary figures and tables that give supporting evidence to the results obtained in Chapter 3.

## Supplementary Figure B.1

### Node degree distribution simulation results

To test our hypothesis that the MLE approach better estimates degree distribution for bait-prey systems, we simulated studies on 100 graphs generated via the preferential attachment model of Barabasi et al. (1999) [69]. Figure B.1 shows the results for 16 such simulations. In every instance, the RMSE of the MLE estimate is smaller than the naïve estimates because the MLE estimate takes into account information such as the error rates in estimating the degree distribution. Because the naïve estimates generally view any positive observation as an edge, this estimate generally over-estimates the true number of interactions even with very small FPR.

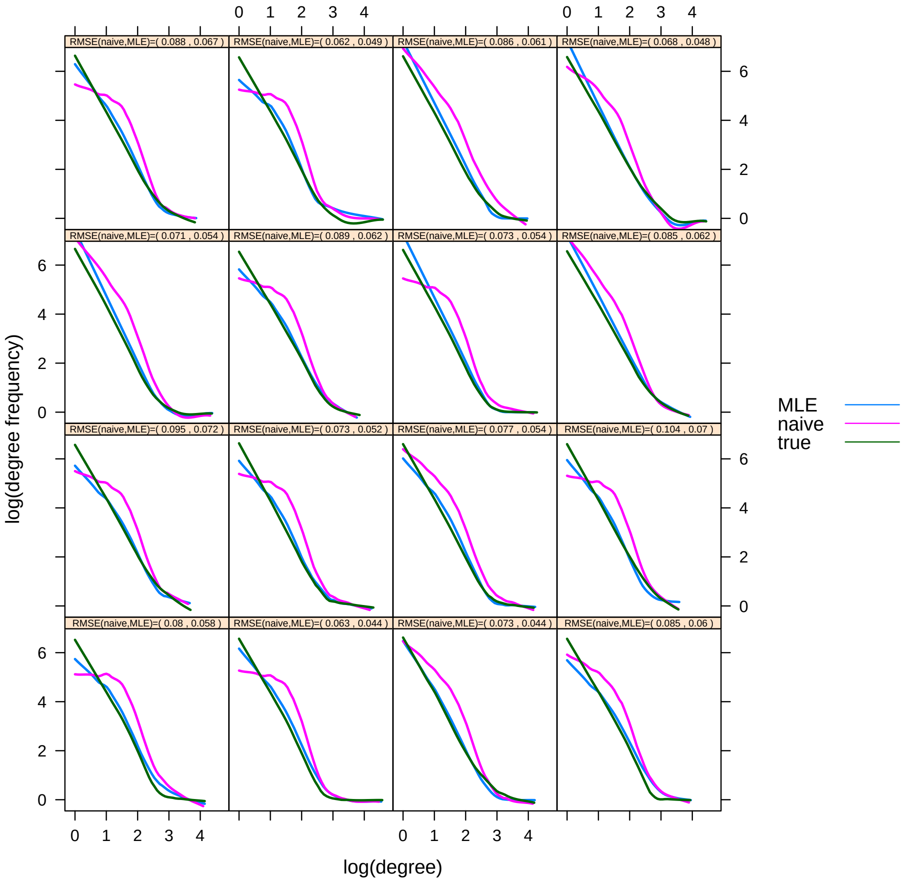


Figure B.1: Sixteen examples of true ($\Delta$), naïve ($\hat{\Delta}_{\text{naïve}}$), and MLE ($\hat{\Delta}_{\text{MLE}}$) node degree distributions from 100 simulations of graphs generated according to the preferential attachment model of Barabasi et al. (1999) [69]. Lines are loess-smoothed curves and RMSE measures goodness-of-fit of the MLE and naïve estimates compared with the true distribution.

## Supplementary Figure B.2

### Estimated $p_{\mathrm{TP}}$ values as gold standard centers on observed proteins

Figure B.2 demonstrates the trend in estimated $p_{\mathrm{TP}}$ as the gold standard set of complexes centers on the set of proteins observed in five AP-MS experiments, here referred to as Gavin02 [74], Ho02 [75], Krogan04 [26], Gavin06 [27], and Krogan06 [28]. Specifically, complexes with a threshold percentage of constituent members observed in each data set are included, and as the threshold increases from 0 to 100%, the gold standard set of complexes narrows only to those whose constituent proteins are viable under the experimental conditions for the data set of interest. There is a dramatic increase in $p_{\mathrm{TP}}$ as the threshold increases (see Supplementary Table B.2 for specific estimates), with the lower threshold values artificially penalizing the technologies for failing to detect interactions which can occur in the cell, but likely do not under the experimental conditions being observed. The steady-state estimates of $p_{\mathrm{TP}}$ between 0 and 40% in Figure B.2 result from limiting inference only to bait proteins; this is explained in detail in Supplementary Table B.2 and Supplementary Figures B.3 – B.8. Our final estimates of $p_{\mathrm{TP}}$ are those at the threshold value of 100% (see Table 2 in the main text for specific estimates and 95% confidence intervals).

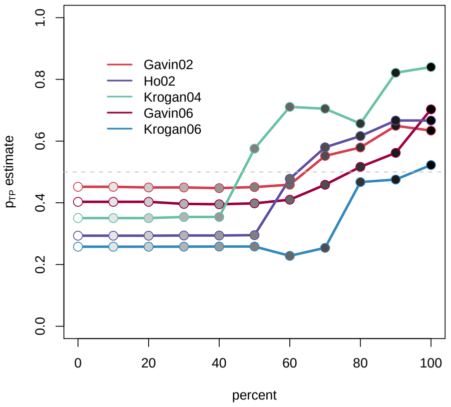


Figure B.2: Estimated values of $p_{\mathrm{TP}}$ at different percentage threshold values determining inclusion of candidate GO/MIPS complexes in the experiment-specific gold standard data sets. Each line represents a different experiment. The $x$ axis represents the minimum percentage of candidate complex members required for entry of any individual complex into the experiment-specific gold standard data set. The $y$ axis represents the resultant estimate of $p_{\mathrm{TP}}$ using the approach described in Section 2.3 in the main text.

## Supplementary Table B.1

### Number of AP-MS graph nodes

Much of the information in bait prey systems has been suppressed in the naïve estimates of degree distribution. For instance, a reciprocally observed interaction indicates a higher likelihood of a true interaction. Unreciprocated interactions are problematic for a number of reasons. An unreciprocated interaction can be the result of merely testing in a single direction between two proteins. Or it can be due to the observation of conflicting data. These two situations will effect how inference is made upon the data, and so it is crucial to know the difference in these two cases. Because negative interactions and tested pairs are often unreported, we have determined a proxy for such information. Table B.1 outlines this proxy information for the high throughput AP-MS data in S. cerevisiae.

| Data set | Number of baits in | | |
|---|---|---|---|
| | original data set (additional prey) | VBP graph pre-systematic filtering (% of original number reported as baits) | VBP graph post-systematic filtering (% of original number reported as baits) |
| Gavin et al. (2002) | 455 (909) | 271 (59.6%) | 268 (58.9%) |
| Ho et al. (2002) | 493 (1085) | 231 (46.9%) | 226 (45.8%) |
| Krogan et al. (2004) | 153 (330) | 151 (98.7%) | 149 (97.4%) |
| Gavin et al. (2006) | 1752 (799) | 991 (56.6%) | 852 (48.6%) |
| Krogan et al. (2006) | 2264 (3097) | 2226 (98.3%) | 1505 (66.5%) |

Table B.1: The first column contains the number of baits and additional proteins detected as prey but not used as baits in each AP-MS data set. The second column contains the number of proteins viable as both bait and prey, i.e. the number of nodes in the VBP graph previous to systematic filtering. The third column contains the number of nodes in the VBP graph after filtering baits prone to systematic error according to the method of Chiang et al. (2007) [1].

## Supplementary Table B.2, Supplementary Figures B.3–B.8

### Gold standard set of edges

Supplementary Table B.2 contains the number of eligible complexes and induced complex co-membership edges in each experiment-specific gold standard data set for five AP-MS data sets. This table demonstrates convergence toward experimental conditions for each data set. Candidate gold standard complexes are required to have a certain percentage of consistent members reported in the data set, and as the percentage threshold increases from 0% to 100%, the experimental conditions of the gold standard and experimental data sets align.

In addition, to ensure estimation of stochastic error unbiased with systematic error, candidate complexes are assessed for extreme distributions of missing edges. For each protein in the complex, observed in-degree and out-degree are calculated relative to the possible in-degree and out-degree for the data set under investigation and the entire set of observations for the complex proteins is tested using Fisher's Exact Test. If either test on in- or out-degree results in a p-value less than 0.01, the complex is not included in the experiment specific gold-standard set. For large complexes with small dynamic subunits, this approach is useful to detect whether or not that subunit was observable under the experimental conditions. If a certain subset of the candidate complex is consistently not detectable, but the remainder of the proteins show a consistent edge pattern, then it is not reasonable to penalize the technology for failing to detect a transient subunit that was not available under the experimental conditions.

The upper panel of Supplementary Figure B.3 illustrates the decrease in the number of complexes used in the gold-standard sets as the threshold increases. The lower panel plots the same values divided by the number of VBP baits in each experiment. At a threshold of 100%, the number of complexes in the gold-standard set is between roughly 5 and 20% of the number of VBP nodes for each experiment.

For all five data sets, the number of baits and therefore the number of induced edges in common to both the filtered VBP graph and the experimental data set remain constant for threshold values up to 30 or 40%. The candidate complex members are compared to the entire set of viable prey in an experiment, including the set of viable baits as a subset. The upper panels of Supplementary Figures B.4 – B.8 indicate that at low threshold values, very large complexes are retained in the gold-standard sets. However, the lower panels of these same figures illustrate that the percentage of the proteins in the gold standard sets that are also VBP proteins in the AP-MS data are quite low. At the low threshold values, many complexes are retained that are irrelevant to the data at hand. As the threshold increases, the percentage of gold standard proteins in the VBP sets also increases, hence narrowing the number of complex edges used in the estimation of $p_{\text{TP}}$. At a threshold of 100%, all proteins in the gold standard complexes are members of the viable prey population, and the subset that are also viable baits are used for analysis.

**Supplementary Table B.2**

| % | Number of Eligible Complexes | | | | | Number of Eligible Baits | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| threshold | G02 | H02 | K04 | G06 | K06 | G02 | H02 | K04 | G06 | K06 |
| 0% | 335 | 335 | 335 | 335 | 335 | 136 | 85 | 69 | 363 | 447 |
| 10% | 246 | 246 | 103 | 266 | 307 | 136 | 85 | 69 | 363 | 447 |
| 20% | 235 | 229 | 91 | 255 | 307 | 133 | 85 | 69 | 359 | 447 |
| 30% | 212 | 186 | 63 | 230 | 302 | 133 | 84 | 64 | 356 | 447 |
| 40% | 184 | 149 | 50 | 209 | 290 | 131 | 82 | 64 | 346 | 443 |
| 50% | 174 | 134 | 43 | 199 | 288 | 127 | 75 | 47 | 337 | 443 |
| 60% | 118 | 81 | 32 | 129 | 214 | 115 | 60 | 39 | 292 | 357 |
| 70% | 97 | 66 | 30 | 108 | 175 | 105 | 53 | 37 | 270 | 315 |
| 80% | 77 | 55 | 26 | 90 | 127 | 96 | 48 | 32 | 212 | 222 |
| 90% | 58 | 42 | 22 | 70 | 95 | 74 | 30 | 24 | 157 | 138 |
| 100% | 53 | 41 | 21 | 65 | 91 | 63 | 30 | 21 | 128 | 111 |
| % | Number of Induced Edges | | | | | Estimate of $p_{\mathrm{TP}}$ | | | | |
| threshold | G02 | H02 | K04 | G06 | K06 | G02 | H02 | K04 | G06 | K06 |
| 0% | 271 | 143 | 174 | 1194 | 1719 | 0.437 | 0.297 | 0.351 | 0.375 | 0.204 |
| 10% | 271 | 143 | 174 | 1194 | 1719 | 0.437 | 0.297 | 0.351 | 0.375 | 0.204 |
| 20% | 265 | 143 | 174 | 1192 | 1719 | 0.443 | 0.297 | 0.351 | 0.375 | 0.204 |
| 30% | 265 | 142 | 168 | 1189 | 1718 | 0.443 | 0.299 | 0.354 | 0.376 | 0.204 |
| 40% | 264 | 141 | 168 | 1150 | 1713 | 0.441 | 0.294 | 0.354 | 0.387 | 0.204 |
| 50% | 258 | 126 | 86 | 1141 | 1713 | 0.450 | 0.286 | 0.576 | 0.390 | 0.204 |
| 60% | 241 | 67 | 64 | 1015 | 1301 | 0.456 | 0.463 | 0.711 | 0.409 | 0.168 |
| 70% | 191 | 47 | 61 | 883 | 1014 | 0.550 | 0.574 | 0.705 | 0.422 | 0.170 |
| 80% | 169 | 40 | 51 | 561 | 605 | 0.577 | 0.613 | 0.657 | 0.516 | 0.207 |
| 90% | 113 | 24 | 28 | 401 | 261 | 0.642 | 0.667 | 0.821 | 0.565 | 0.299 |
| 100% | 96 | 24 | 25 | 231 | 119 | 0.630 | 0.667 | 0.840 | 0.710 | 0.492 |

Table B.2: The upper left panel contains the number of eligible complexes for the experiment-specific gold standard data sets that meet the % threshold criteria for the number of proteins reported as either viable baits or viable prey in the experiment under consideration. The upper right panel reports the number of proteins in the eligible gold standard complexes that were reported as viable baits in the experiment under consideration. The lower left panel contains the number of gold standard complex co-membership edges induced by the eligible baits. The lower right panel contains estimates of $p_{\mathrm{TP}}$ given the set of induced edges at each % threshold level for each experiment. G02=Gavin02, H02=Ho02, K04=Krogan04, G06=Gavin06, K06=Krogan06.

## Supplementary Figure B.3

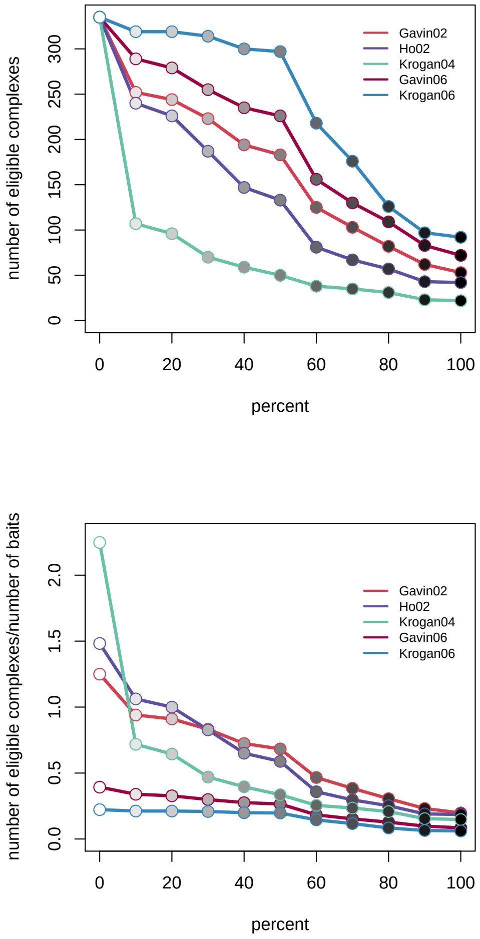
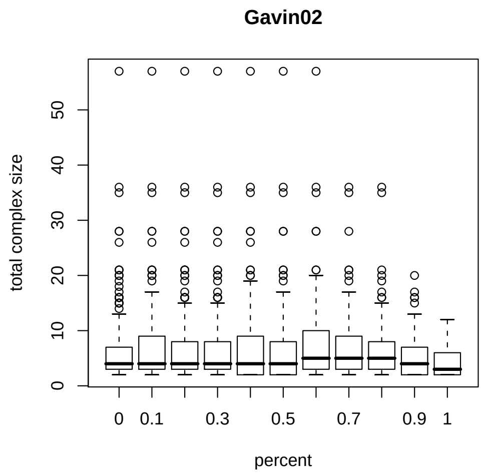


Figure B.3: The upper panel plots the number of eligible complexes in the gold standard data sets at each % threshold value for the five AP-MS data sets. The lower panel plots the same values, divided by the number of viable baits in each experiment.

## Supplementary Figure B.4



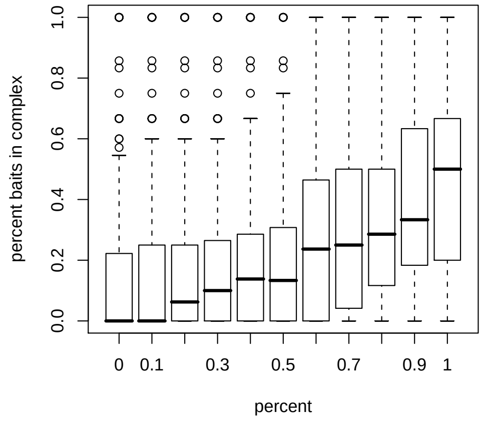
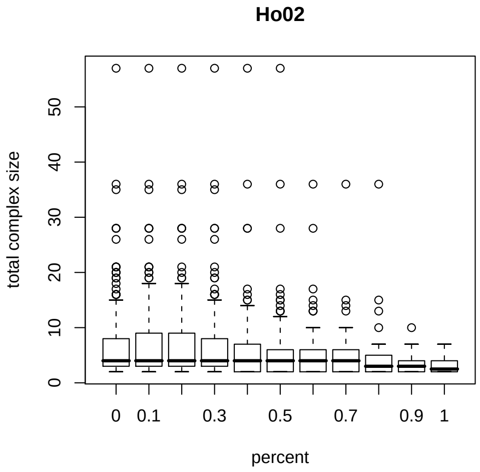


Figure B.4: The upper panel plots boxplots of complex sizes in the gold standard data sets at each threshold level for the Gavin02 data. The lower panel plots boxplots of the percent of viable Gavin02 baits in the complexes at each threshold level.

## Supplementary Figure B.5



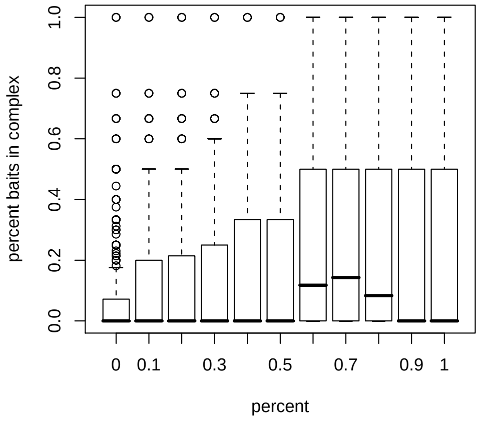


Figure B.5: The upper panel plots boxplots of complex sizes in the gold standard data sets at each threshold level for the Ho02 data. The lower panel plots boxplots of the percent of viable Ho02 baits in the complexes at each threshold level.

## Supplementary Figure B.6

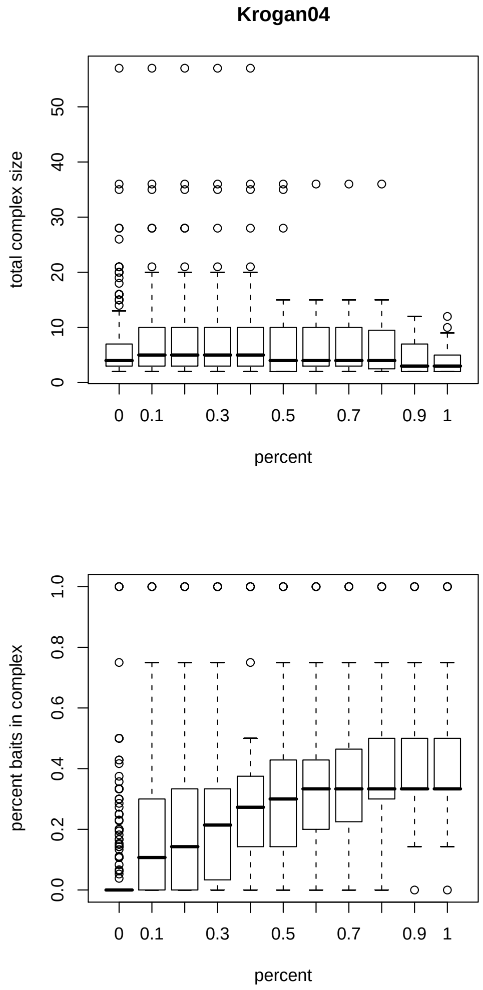


Figure B.6: The upper panel plots boxplots of complex sizes in the gold standard data sets at each threshold level for the Krogan04 data. The lower panel plots boxplots of the percent of viable Krogan04 baits in the complexes at each threshold level.

## Supplementary Figure B.7

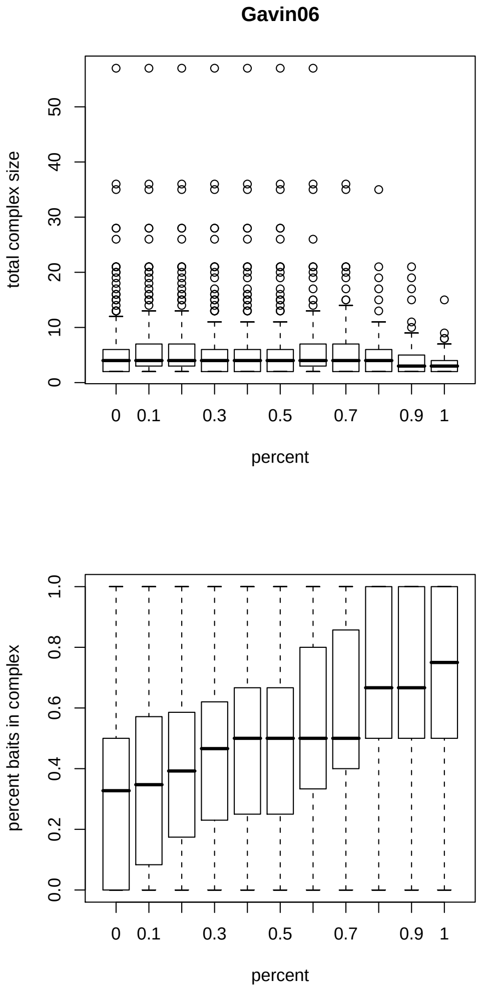


Figure B.7: The upper panel plots boxplots of complex sizes in the gold standard data sets at each threshold level for the Gavin06 data. The lower panel plots boxplots of the percent of viable Gavin06 baits in the complexes at each threshold level.

## Supplementary Figure B.8

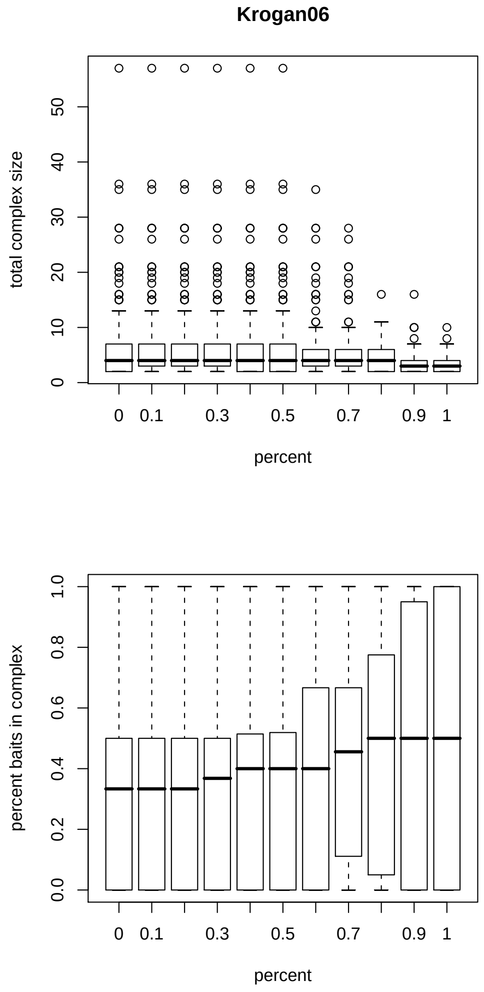


Figure B.8: The upper panel plots boxplots of complex sizes in the gold standard data sets at each threshold level for the Krogan06 data. The lower panel plots boxplots of the percent of viable Krogan06 baits in the complexes at each threshold level.

## Supplementary Table B.3

### Degree Estimates for AP-MS Data Sets

Finally, Table B.3 shows a comparison between the estimates obtained from the naïve estimation against the MLE estimates pertaining to the five large-large scaled AP-MS data sets. Because the MLE estimator uses an approximate sensitivity and specificity in determining the degree distribution, we can see from the table that the number of reciprocated observations clearly influences the degree estimates, more so than the unreciprocated observations.

| | $u$= | | | | | | | | | | | | | $u$= | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | Naïve | | | | | | | | | | | | | Ho02 | | | | | | | | | | | | |
| $r = 0$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 11 | 12 |
| 2 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 4 | 5 | 6 | 7 | 8 | 8 | 9 | 10 | 11 | 12 | 13 | 15 | 16 |
| 5 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 6 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 7 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 8 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 9 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 21 |
| 9 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 10 | 11 | 12 | 13 | 13 | 14 | 15 | 16 | 18 | 19 | 20 | 21 | 22 |
| 10 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| 11 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 12 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| | Gavin02 | | | | | | | | | | | | | Krogan04 | | | | | | | | | | | | |
| $r = 0$ | 0 | 0 | 0 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 13 | 14 | 1 | 1 | 2 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 13 | 14 | 15 | 2 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 11 | 12 | 13 | 14 | 15 | 16 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 4 | 4 | 5 | 6 | 7 | 8 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 4 | 5 | 5 | 6 | 7 | 8 | 9 | 9 | 10 | 11 | 12 | 13 | 14 |
| 5 | 5 | 6 | 7 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 19 | 5 | 6 | 7 | 7 | 8 | 9 | 10 | 11 | 12 | 12 | 13 | 14 | 15 |
| 6 | 6 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 19 | 20 | 6 | 7 | 8 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 18 | 19 | 20 | 21 | 7 | 8 | 9 | 10 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 18 | 19 | 20 | 21 | 22 | 8 | 9 | 10 | 11 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 9 | 10 | 11 | 12 | 13 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 10 | 11 | 12 | 13 | 14 | 15 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 10 | 11 | 12 | 13 | 14 | 15 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 11 | 12 | 13 | 14 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 26 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 17 | 18 | 19 | 20 | 21 | 22 |
| 12 | 13 | 14 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 25 | 26 | 27 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 19 | 20 | 21 | 22 | 23 |
| | Gavin06 | | | | | | | | | | | | | Krogan06 | | | | | | | | | | | | |
| $r = 0$ | 0 | 0 | 0 | 0 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 4 | 5 | 7 | 8 | 9 | 10 |
| 1 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 10 | 11 | 12 | 1 | 1 | 2 | 2 | 3 | 4 | 5 | 6 | 7 | 9 | 10 | 11 | 12 |
| 2 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 13 | 14 |
| 3 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 4 | 4 | 5 | 6 | 7 | 8 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 5 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 15 | 16 | 17 |
| 5 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 6 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 18 | 19 | 20 |
| 7 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 9 | 9 | 10 | 11 | 12 | 13 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 8 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 22 | 23 |
| 9 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 20 | 21 | 22 | 23 | 24 |
| 10 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 12 | 13 | 14 | 15 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 26 |
| 11 | 12 | 13 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 23 | 24 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 24 | 25 | 26 | 27 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 23 | 24 | 25 | 26 | 27 | 28 |

Table B.3: The upper left sub-table is the naïve estimate of degree for pairs of the numbers of reciprocated and unreciprocated edges incident on an individual node in a graph of protein interaction data. The remaining regions are the MLE estimates of node degree for the same numbers in each of the AP-MS data sets under study. Node degree estimates depend on estimated $p_{\mathrm{TP}}$ and the estimated number of FPs per node.

# Appendix C

# Publications

These are the publications that I contributed to while researching for this dissertation:

- Chiang T, Scholtens D, Sarkar D, Gentleman R, Huber W. (2007) Coverage and error models of protein-protein interaction data by directed graph analysis. [1]
- Chiang T, Li N, Orchard S, Kerrien S, Hermjakob H, Gentleman R, Huber W. (2008) Rintact: enabling computational analysis of molecular interaction data from the IntAct repository. [84]
- Scholtens D, Chiang T, Huber W, Gentleman R. (2008) Estimating node degree in bait-prey graphs. [2]
- Chiang T, Scholtens D. (2009) A general pipeline for quality and statistical assessment of protein interaction data using R and Bioconductor. [3]

# Bibliography

[1] Chiang T, Scholtens D, Sarkar D, Gentleman R, Huber W (2007) Coverage and error models of protein-protein interaction data by directed graph analysis. Genome Biology 9:R186.

[2] Scholtens D, et al. (2008) Estimating node degrees in bait-prey graphs. Bioinformatics 24:218–24.

[3] Chiang T, Scholtens D (2009) A general pipeline for quality and statistical assessment of protein interaction data using R and Bioconductor. Nature Protocols 4:535–46.

[4] Tong A, Lesage G, Bader G, Ding H, Xu H, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303:808.

[5] Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nature Biotechnology 23:561–566.

[6] Lehner B, et al. (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nature Genetics 38:896–903.

[7] Bork P, Jensen L, von Mering C, Ramani A, Lee I, et al. (2004) Protein interaction networks from yeast to human. Current Opinion Structural Biology 14:292–299.

[8] Stuart JM, et al. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302:249–255.

[9] Gross JL, Yellen J (1998) Graph theory and its applications. Boca Raton: CRC Press.

[10] Wasserman S, Faust K (1994) Social Network Analysis, Methods and Applications. Cambridge: Cambridge University Press.

[11] Tong AHY, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science :321–324.

[12] Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the USA 93:13–20.

[13] Fields S, Sternglanz R (1994) The two hybrid system: an assay for protein-protein interactions. Trends in Genetics 10:145–150.

[14] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of the USA 98:4569–4574.

[15] Uetz P, Giot L, Cagney G, Mansfield T, Judson R, et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627.

[16] Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nature Biotechnology 18:1242–3.

[17] Uetz P (2002) Two-hybrid arrays. Current Opinion Chemical Biology 6:57–62.

[18] Tucker C, Gera J, Uetz P (2001) Towards an understanding of complex protein networks. Trends Cell Biology 11:102–6.

[19] Goll J, Uetz P (2006) The elusive yeast interactome. Genome Biology 7:223.

[20] Aloy P, Russell R (2003) InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 19:161–2.

[21] Campillos M, Kuhn M, Gavin A, Jensen L, Bork P (2008) Drug target identification using side-effect similarity. Science 321:263–6.

[22] Bader S, Khner S, Gavin A (2008) Interaction networks for systems biology. FEBS Letter 582:1220–4.

[23] Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147.

[24] Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183.

[25] Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, et al. (2004) Structure-based assembly of protein complexes in yeast. Science 303:2026–2029.

[26] Krogan N, Lam M, Fillingham J, Keogh M, Gebbia M, et al. (2004) High-definition macromolecular composition of yeast RNA-processing complexes. Molecular Cell 13:225–239.

[27] Gavin A, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631–636.

[28] Krogan N, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440:637–643.

[29] Charbonnier S, Gallego O, Gavin A (2008) The social network of a cell: recent advances in interactome mapping. Biotechnology Annual Review 14:1–28.

[30] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of the USA 98:4569–4574.

[31] Cagney G, Uetz P, Fields S (2001) Two-hybrid analysis of the *Saccharomyces cerevisiae* 26s proteasome. Physiological Genomics 7:27–34.

[32] Tong A, Drees B, Nardelli G, Bader G, Brannetti B, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295:321–324.

[33] Hazbun T, Malmstrom L, Anderson S, Graczyk B, Fox B, et al. (2003) Assigning function to yeast proteins by integration of technologies. Molecular Cell 6:1353–1365.

[34] Zhao R, Davey M, Hsu Y, Kaplanek P, Tong A, et al. (2005) Navigating the chaperone network: An integrative map of physical and genetic interactions mediated by the Hsp90 chaperone. Cell 120:715–727.

[35] Giot L, Bader J, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302:1727–1736.

[36] Li S, Armstrong C, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan *Caenorhabditis elegans*. Science 303:540–543.

[37] Rual J, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437:1173–1178.

[38] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–968.

[39] Hartwell LH, Hopfield J, Leibler S, Murray A (1999) From molecular to modular cell biology. Nature 402:47–52.

[40] Walhout A, Boulton S, Vidal M (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. YEAST 17:88–94.

[41] Mrowka R, Patzak A, Herzel H (2001) Is there a bias in proteome research? Genome Research 11:1971–1973.

[42] Hazbun TR, Fields S (2001) Networking proteins in yeast. Proceedings of National Academy of Science 98:4277–8.

[43] Deane C, Salwinski L, Xenarios I, Eisenberg D (2002) Two methods for assessment of the reliability of high throughput observations. Molecular and Cellular Proteomics 1.5:349–56.

[44] Edwards A, Kus B, Jansen R, Greenbaum D, Greenblat J, et al. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends in Genetics 18:529–536.

[45] von Mering C, Krause R, Snel B, Cornell M, Oliver S, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399–403.

[46] Thomas A, Cannings R, Monk N, Cannings C (2003) On the structure of protein-protein interactions networks. Biochemical Scociety 31:1491–6.

[47] Lappe M, Holm L (2004) Unraveling protein interaction networks with rear-optimal efficiency. Nature Biotechnology 22:98–103.

[48] Poyatos J, Hurst L (2004) How biologically relevant are interaction based modules in protein networks. Genome Biology 5:R93.

[49] Vidalain P, Boxem M, Ge H, Li S, Vidal M (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. Methods 32:363–370.

[50] Gagneur J, David L, Steinmetz L (2006) Capturing cellular machines by systematic screens of protein complexes. Trends in Microbiology 14:336–339.

[51] Walhout A, Vidal M (1999) A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. Genome Research 9:1128–34.

[52] Aloy P, Russell RB (2002) Potential artefacts in protein-interaction networks. FEBS Letter 530:253–254.

[53] Stanley RP (1997) Enumerative Combinatorics I. New York: Cambridge University Press.

[54] de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, et al. (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. BMC Biology 4.

[55] Falcon S, Gentleman R (2006) Using GOstats to test gene lists for go term association. Bioinformatics 2:257–258.

[56] Han J, Dupuy D, Bertin N, Cusick M, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nature Biotechnology 23:839–844.

[57] Stumpf MPH, Wiuf C (2005) Sampling properties of random graphs: the degree distribution. Physics Review E, Statistical Nonlinear, and Soft Matter Physics 72:036118.

[58] Sharp PM, Li WH (1987) The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acid Res 15:1281–1295.

[59] Domon B, Aebersold R (2006) Mass spectrometry and protein analysis. Science 312:212–217.

[60] Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207.

[61] Gentleman R, Huber W (2007) Making the most of high-throughput protein-interaction data. Genome Biology 8.

[62] Kelsey J, Whittemore A, Evans A, Thompson W (1996) Methods in observational epidemiology. In: Monographs in Epidemiology and Biostatistics, New York: Oxford University Press.

[63] Bickel P, Doksum K (2001) Mathematical Statistics: Basic Ideas and Selected Topics. New Jersey: Prentice Hall, 556 pp.

[64] Miller J, Lo R, Ben-Hur A, Desmarais C, Stagljar I, et al. (2005) Large-scale identification of yeast integral membrane protein interactions. Proceedings of the National Academy of Sciences 34:12123–8.

[65] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, et al. (2006) IntAct – open source resource for molecular interaction data. Nucleic Acids Research 35:D561–D565.

[66] Saccharomyces Genome Database. URL `"http://www.yeastgenome.org"`.

[67] Scholtens D, Gentleman R (2004) Making sense of high-throughput protein-protein interaction data. Statistical Applications in Genetics and Molecular Biology 3:Article 39.

[68] Li L, Alderson D, Doyle J, Willinger W (2006) Towards a theory of scale-free graphs: Definition, properties, and implications. Internet Mathematics 2:4.

[69] Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512.

[70] Deng M, Sun F, Chen T (2003) Assessment of the reliability of protein-protein interactions and protein function prediction. Pacific Symposium on Biocomputing :140–151.

[71] D'haeseleer P, Church G (2004) Estimating and improving protein interaction error rates. In: Proc IEEE Comput Syst Bioinform Conf: August 16-19 2004; California. IEEE Computer Society, pp. 216–23.

[72] Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interactions networks. Genome Biology 7.

[73] Collins S, Kemmeren P, Zhao XC, Greenblatt J, Spencer F, et al. (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Molecular and Cellular Proteomics 6:439–450.

[74] Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147.

[75] Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183.

[76] Mewes H, Amid C, Arnold R, Frishman D, Guldener U, et al. (2004) MIPS: analysis and annotations of proteins from whole genomes. Nucleic Acids Research 32:D41–D44.

[77] Harris MA, et al. (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Research 32:258–261.

[78] Uetz P, Giot L, Cagney G, Mansfield T, Judson R, et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627.

[79] Kerrien S, et al. (2007) Broadening the horizon-level 2.5 of the hupo-psi format for molecular interactions. BMC Biology :5:44.

[80] Orchard, et al. (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nature Biotechnology 8:894–8.

[81] Stevens S, Abelson J (1999) Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins. Proceedings of the National Academy of Sciences of the USA 96:7226–7231.

[82] Chowdhury A, Mukhopadhyay J, Tharun S (2007) The decapping activator Lsm1p-7p-Pat1p complex has the intrinsic ability to distinguish between oligoadenylated and polyadenylated RNAs. RNA 13:998–1016.

[83] Storey J (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B 64:479–498.

[84] Chiang N T Li, et al. (2007) Rintact: enabling computational analysis of molecular interaction data from the IntAct repository. Bioinformatics 24:1100–1.