

A computational and informatics framework for the
analysis of affinity purification mass spectrometry data
and reconstruction of protein interaction networks

by

Dattatreya Mellacheruvu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2015

Doctoral Committee:

Associate Professor Alexey I Nesvizhskii, Chair

Professor Philip Andrews

Assistant Professor James Cavalcoli

Associate Professor Anuj Kumar

Professor Kerby Shedden

© Dattatreya Mellacheruvu, 2015

Table of Contents

List of Figures	vi
List of Tables	viii
List of Appendices	ix
CHAPTER 1. Introduction	1
Significance of protein interactions	1
Detecting protein interactions.....	2
Yeast two-hybrid.....	2
Affinity purification mass spectrometry	2
Affinity purification of protein complexes.....	3
Protein identification and quantitation using tandem mass spectrometry	5
Quantitation of peptide and protein abundance	6
Identifying <i>bona fide</i> interactions from AP-MS data	8
Overview of thesis	9
CHAPTER 2. SPInt: A Framework for Scoring Protein Interactions from Affinity Purification Mass Spectrometry Data	13
Introduction	13
Methods.....	14

Design and implementation of SPrint.....	14
Scoring modules.....	15
Interaction scoring: FC score	15
Interaction scoring: SAINT	16
Interaction scoring: EScore	17
Interaction scoring: CompPASS	17
Comparison to literature data	18
Preparation of test data.....	19
Data Formats.....	21
Access to SPrint.....	22
Results and Discussion	22
Creation of the SPrint processing pipeline	22
Analysis of small scale data sets	24
Analysis of medium/large scale data sets.....	26
Concluding Remarks.....	31
Contributions	32
CHAPTER 3. Pint: A tool for analysis of Protein Interaction Networks	33
Introduction	33
Methods.....	34
Design and implementation of Pint.....	34
Generating a database of interactions from literature	35
Generating the network context	35
Using bait-bait cluster gram to identify sub-networks.....	36

Preparation of test data.....	36
Results and Discussion	37
Creation of Plnt.....	37
Analysis of small scale networks.....	39
Analysis of medium/large scale networks	39
Contributions	40
 CHAPTER 4. CRAPome: The Contaminant Repository for Affinity Purification	
Mass Spectrometry Data.....	43
Introduction	43
Methods.....	44
Design and architecture of the CRAPome repository.....	44
Processing of mass spectrometry data and population of the CRAPome database	45
Quality control	47
Integrated scoring tool (SPrint)	47
Global analysis of CRAPome and reduced gene counts	47
Contaminant propensity as a function of protein abundance	48
Preparation of test data sets	49
Access to CRAPome	49
Results and Discussion	49
Creation of the CRAPome Repository.....	49
Graphical user Interface.....	50
Using CRAPome to score interactions	51
Characterization of CRAPome.....	53

Concluding Remarks.....	57
Contributions	58
CHAPTER 5. RePrint: A Repository of Protein Interactions Generated from Affinity Purification Mass Spectrometry Data	59
Introduction	59
Methods.....	60
Design and architecture of RePrint.....	60
Processing of mass spectrometry data and population of RePrint database	60
Interaction scoring: RePrint score	62
Preparation of test data.....	62
Access to RePrint	63
Results and Discussion	65
Creation of RePrint	65
Graphical user interface.....	66
Using RePrint to generate interaction networks.....	66
Concluding Remarks.....	70
Contributions	71
CHAPTER 6. Conclusions and Future Directions.....	72
Summary of the thesis	72
Impact on research community.....	74
Utility of interaction networks in biological data analysis.....	76
Future directions.....	78
References.....	92

List of Figures

Figure 1-1: Overview of affinity purification mass spectrometry (using epitope tagged baits). ...	3
Figure 1-2: Identifying <i>bona fide</i> interactions using negative controls.	9
Figure 1-3: Computational and Informatics framework for analysis of AP-MS data.	10
Figure 2-1: Input file formats for SPrint.	22
Figure 2-2: SPrint pipeline and the graphical user interface.	28
Figure 2-3: Enrichment scoring functions of SPrint illustrated on a four-bait data set.	29
Figure 2-4: Utility of FC B in filtering sporadic contaminants.	30
Figure 2-5: Specificity scoring functions of SPrint illustrated using two medium scale data sets.	31
Figure 3-1: PInt pipeline and the graphical user interface.	38
Figure 3-2: Analysis of a small scale network using PInt.	41
Figure 3-3: Analysis of a medium/large scale network using PInt.	42
Figure 4-1: Schema of the CRAPome database.	45
Figure 4-2: Creation of CRAPome database and the graphical user interface.	52
Figure 4-3: Scoring protein interactions using controls from the CRAPome (V 1.0) with SAINT.	53
Figure 4-4: Characterization of CRAPome (V 1.0).	55
Figure 5-1: Creation of RePrint and the graphical user interface.	64
Figure 5-2: Hippo pathway generated from three data sources.	68
Figure 5-3: Utility of topological filtering.	70
Figure 6-1: Impact on research community.	74
Figure A-1: Software design paradigm.	81
Figure A-2: Software architecture diagram.	82
Figure A-3: Overview of system integration.	83
Figure A-4: Overview of user job execution.	84

Figure B-1: Overview of annotation procedure.....	85
Figure B-2: Experiment view (annotator login)	86
Figure B-3: Protocol view (annotator login)	86
Figure B-4: Controlled vocabularies (annotator login).	87
Figure B-5: Adding data sets and baits (annotator login).	88
Figure B-6: Guidelines for adding new protocols (annotator login).	88
Figure B-7: Creating new protocol; part A: define controlled vocabulary (annotator login).	89
Figure B-8: Creating new protocol; part B: adding protocol details (annotator login).	89
Figure B-9: Experiment view (annotator login).	90
Figure B-10: Editing experiments to associate protocol (annotator login).	90
Figure B-11: Linking protocol to experiments using the 'Select Protocol' drop down menu (annotator login).	91

List of Tables

Table 4-1: Controlled vocabulary for annotating experiments (V 1.0).....	50
Table 4-2: Frequency of detection across CRAPome database (V 1.0).	54
Table 4-3: Most Frequently detected protein families across the CRAPome (V 1.0).....	57
Table 6-1: List of contributors for the CRAPome repository (V 1.0).....	76

List of Appendices

Appendix A: Software Manual	80
Appendix B: Annotator Manual	85

CHAPTER 1

Introduction

Significance of protein interactions

Proteins play a central role in the functioning of the cell. Most of the proteins function in collaboration with other proteins and bio-molecules (DNA, RNA, small molecules). Protein complexes, i.e. ensembles of interacting proteins, play a mechanistic role in several basic biological processes. For example, the transcription pre-initiation complex in eukaryotes, which comprises of several transcription initiation factors and RNA polymerase II, is responsible for transcribing DNA to mRNA [2]. The translation of mRNA to proteins in eukaryotes is carried out by the ribosomal complex, comprising several ribosomal sub-units [3]. Similarly, the proteasome complex that degrades damaged proteins in the cell also comprises of several intricately arranged protein sub-units. Such protein complexes are typically held together by 'stable' protein interactions. On the other hand, several weak and transient interactions are responsible for various signaling mechanisms in the cell. These signaling mechanisms play a 'regulatory' role by modulating basic biological processes. For example, histone de-acetylases (HDACs) regulate gene expression by de-acetylating histones [4]. The protein kinase Hpo (and other members of the Hippo pathway) control organ development in mammals by modulating the cell cycle and other related biological processes [5]. The Wnt family proteins transmit messages from outside the cell (through surface receptors) and control embryonic development by modulating cell proliferation and migration [6]. Suffice to say, that protein interactions are central to almost every biological process in the cell. Accordingly, protein interaction maps can help unravel the mechanisms of several underlying biological processes.

The molecular basis of disease can be understood by a comparative analysis of protein interactions in normal versus disease states. This approach is particularly relevant to cancer,

where cells undergo transformation and evolve by ‘rewiring’ the underlying protein interaction networks. Accordingly, identifying and inactivating the master regulators in the cancerous cells is a promising strategy for rational drug discovery. Kinases, which regulate several biological processes by phosphorylating proteins, are thus the favorite targets in cancer [7]. In summary, generating high quality protein interaction maps can help in developing new therapies for diseases such as cancer, in addition to providing a unique insight to underlying biological processes in the cell.

Detecting protein interactions

Several experimental approaches have been developed for detecting protein interactions. Here, we briefly discuss two popular high throughput methods – the yeast two-hybrid assay [8] and affinity purification mass spectrometry (AP-MS) [9] .

Yeast two-hybrid

This protein-fragment compensation assay reports the physical interaction between two interacting proteins through the expression of a reporter gene. Briefly, the binding domain and the activation domain of a transcription factor that activates the expression of a reporter gene are fused separately to each of the two proteins that are being tested. If the proteins interact physically, the binding and activation domains are placed in close proximity, thus driving the expression of the reporter gene. This simple approach can easily be multiplexed to generate a high experimental throughput. However, the approach also suffers from several issues including a high false positive rate. Although popular during the past decade, it has now been replaced by affinity purification mass spectrometry as the method of choice for detecting protein interactions.

Affinity purification mass spectrometry

In AP-MS, a protein complex is affinity purified in its native form and analyzed on a mass spectrometer (Figure 1-1). The typical experimental workflow involves a) affinity purification of protein complexes, b) protein identification and quantitation using tandem mass spectrometry and c) identifying *bona fide* interactions from AP-MS data. Advances in protein mass spectrometry and standardization of purification protocols have placed AP-MS as the method

of choice for identifying protein interactions. Each of the components of the experimental workflow is discussed in detail below.

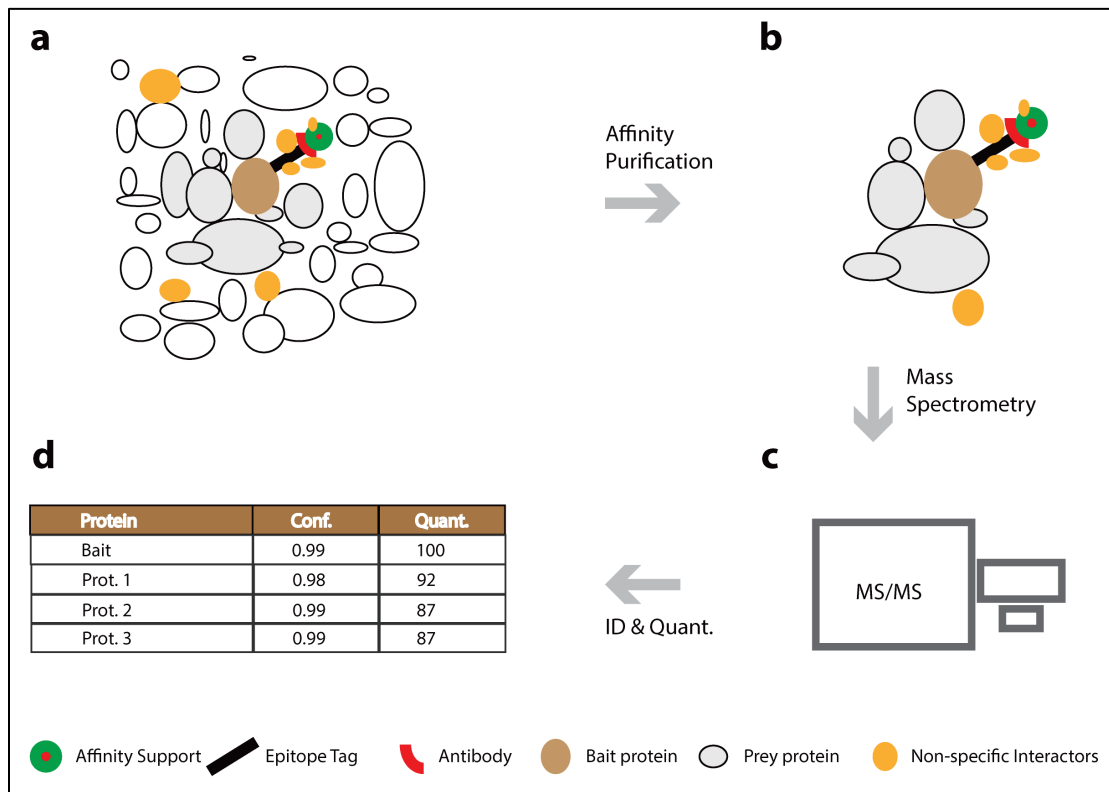


Figure 1-1: Overview of affinity purification mass spectrometry (using epitope tagged baits).

(a, b) Protein complexes of interest are purified using epitope tagged bait proteins. (c) The purified complex is analyzed using protein mass spectrometry. (d) Data is processed to identify and quantify proteins. Protein lists (prey) are scored to identify *bona fide* interactions (main text).

Affinity purification of protein complexes

The process involves purifying a protein complex using an antibody targeted against one of the members of the complex, called the ‘bait’. The antibody is immobilized on what is called the ‘affinity support’. After cell lysis, the sample is incubated with the immobilized antibody and washed under mild conditions. The protocol is optimized to ensure that the protein complex is not disrupted during sample processing. Washing removes unbound proteins, leaving the bait and its interacting partners (i.e., the protein complex) behind. Elution buffers are then used to release the complex back into solution. Co-eluting proteins are identified and quantified using protein mass-spectrometry (described below).

The two common flavors of affinity purification are a) immuno-purification and b) epitope tag-based affinity purification. In immuno-purification assays, the antibody is targeted directly against the bait protein. In epitope tag-based AP experiments, the bait protein is clonally modified to express an epitope tag and antibodies are targeted against the epitope tag. Epitope tags are short peptide sequences for which high-affinity antibodies are available. Popular epitope tags include FLAG, GFP, HA, His, TAP and c-Myc¹. In general, tag-based protocols do not have the problem of (antibody) cross reactivity, since the epitope tags have strong affinity to their antibodies.

Epitope tag based protocols can further be classified into two important sub-categories - single step and tandem affinity purification. In tandem affinity purification, elute from the first purification is subjected to a second affinity purification. Tandem AP typically uses a fusion tag that is cleavable at the junction. Single step protocols involve a single washing step, whereas tandem protocols involve two. Accordingly, tandem protocols tend to generate cleaner samples. However, additional washes eliminate weak and transient interactions and reduce the sensitivity of measurement. The phrases 'tag-based AP-MS' and 'AP-MS' are used interchangeably in literature. **Throughout this thesis 'AP-MS' means 'epitope tag-based AP-MS', unless specified otherwise.**

In theory, AP protocols are meant to capture 'native' complexes. However, a few caveats exist. It is a common practice to over-express the bait protein in order to amplify interactions. Whether over-expression preserves native conditions needs to be taken into consideration. Also, cell lysis brings together proteins that are otherwise localized to different organelles. Accordingly, proteins that do not natively interact may do so in the lysate. Notwithstanding these caveats, affinity purification can largely be considered as a powerful approach to profile native interactions.

¹ <http://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma-Aldrich/Brochure/1/epitope-tags-in-protein-research.pdf>

Protein identification and quantitation using tandem mass spectrometry

Peptides are more amenable for analysis on a mass spectrometer than intact proteins. Hence a shotgun approach is adapted, where proteins are enzymatically cleaved to generate peptides that are then analyzed on a mass spectrometer [10]. Trypsin is commonly used for protein digestion. Mass spectrometers cannot handle complex samples; hence the peptide mixture is typically processed through an online chromatographic separation before it is introduced into the mass spectrometer. Tandem mass spectrometry is the standard approach to sequence peptide ions. The first step (MS^1) involves isolating a single species of peptide ions using 'mass filters'. The second step involves fragmenting the isolated 'parent' ions and accurately measuring the mass (actually, mass to charge ratio) of the fragment ions (MS^2). The instrument alternates between MS^1 and MS^2 cycles and hence the name 'tandem mass spectrometry'. The duty cycle varies from instrument to instrument. It is easy to see that the instrument 'samples' peptide ions for sequencing. To avoid over sampling abundant peptide species, a data dependent acquisition (DDA) strategy is adapted. DDA strategy typically works by 'excluding' continuously eluting peptide ions for repeated sequencing beyond a certain number of times. DDA strategy can be subject to technical variation depending on chromatography conditions and the sample complexity. A more recent approach to protein mass spectrometry is based on a data independent strategy (DIA). This approach opens up the isolation window of MS^1 to include a wider range of peptide ion species, rather than trying to isolate a single species of peptide ions for fragmentation. While DIA approach can sequence greater number of peptides, it generates chimeric spectra that are difficult to interpret.

Spectral data (MS^1 and MS^2) for each sequenced peptide is computationally interpreted to identify the composition and sequence of amino acids [11]. The typical approach is referred to as the 'database search'. Here, each 'experimental' spectrum is compared to a set of 'theoretical' spectra generated from a database of known peptide sequences. The best hit is assigned to each spectrum. Popular search engines such as SEQUEST [12], X! Tandem [13] and Mascot² have their own scoring functions to compare experimental and theoretical spectra.

² <http://www.matrixscience.com/>

Peptide-to-spectrum matches (PSMs) are then filtered to identify the high confidence hits. PeptideProphet [14] is one of the popular algorithms that take a statistical approach to filter PSMs. It converts raw scores generated by the search engine to a discriminant score. The distribution of the discriminant score is modelled using the expectation maximization (EM) algorithm to derive the distribution of true and incorrect hits. It then uses Bayes theorem to assign a probability score for each PSM. In other words, PeptideProphet [15] puts peptide confidence scores on a standardized (probability) scale.

A question arises as to how to filter PSMs in order to generate a list of high confidence PSMs. Typically, the cut-offs are chosen to keep the overall false discovery rate (FDR) low, typically $\leq 5\%$. Decoy sequences are used to estimate the FDR for a given cut-off [16]. Decoys are hypothetical peptide sequences that are unlikely to be present in the sample and hence represent false hits. Decoy-based FDR is estimated by appending an equal number of decoy sequences to the protein/peptide sequence database used for database search (described earlier). PSMs are scored as described above and peptide probabilities are generated using PeptideProphet. The FDR for a selected probability cut-off is estimated as the percentage of decoy sequences that pass the cut-off.

In shotgun proteomics experiments, the presence/absence of proteins in the sample is inferred using the list of high confidence peptides. ProteinProphet is a popular algorithm for protein inference that treats each high confidence peptide as independent evidence of its parent protein and uses the following approach to derive a protein probability (i.e., probability that a protein was present in the sample). Again the error rates can be estimated using decoy sequences. An in-depth discussion of the error rates can be found here [16].

$$Protein Prob. = 1 - \prod (1 - peptide_i)$$

Quantitation of peptide and protein abundance

There are two principle approaches to peptide and protein quantitation – labelling based and label free[17]. ‘Stable isotope labeling by amino acids in cell culture (SILAC)’ [18] and ‘isobaric tags for relative and absolute quantitation (iTRAQ)’ [19] are examples of labeling based

approaches. Spectral counting and peptide ion intensity based quantitation are examples of label free quantitation.

In SILAC, the cells are metabolically labelled to incorporate 'heavy' amino acids, i.e., amino acids comprising higher isotopes of carbon (^{13}C). In iTRAQ, the peptides from each sample are chemically modified using isobaric tags containing a reporter molecule. In both approaches, the samples from two or more sources are mixed and analyzed in a single experiment. In the case of SILAC, differences in protein abundances between the groups being tested is proportional to the differences between 'heavy' and 'light' peptides obtained from each source. In case of iTRAQ, peptide fragmentation releases the reporter ion from each isobaric tag, the intensity of which can be used for relative quantitation. Multiplexing samples helps reduce the technical variability. Both approaches use ion intensities for quantitation, which is in general sensitive compared to the spectral counting approach (detailed below). However, labelling based approaches are resource intensive. Also, iTRAQ suffers from labelling biases.

Label free approaches attempt to directly estimate peptide and protein abundances from spectral data. The popular spectral counting approach simply uses the number of high quality spectra assigned to a peptide as a semi-quantitative estimate of the peptide abundance. In peptide ion intensity based quantitation, the area under the curve (AUC) of the extracted ion chromatogram (XIC) is typically used to estimate peptide abundances. Spectral counting is robust and easy to compute, but is less sensitive compared to intensity based quantitation. Peptide ion intensity based quantitation is more sensitive, but involves sophisticated computation that requires careful interpretation.

In all shotgun approaches, protein abundances are estimated using peptide abundances. It is often the case that several peptides map to multiple proteins, making the process of protein abundance estimation less than trivial. Such issues have been discussed by Fermin et al., in their software tool, ABACUS that processes the output of PeptideProphet and ProteinProphet to generate spectral abundance estimates. Due to inherent ambiguity in peptide identification and quantitation, juxtaposing data from multiple experiments may result in a sparse spectral count matrix. ABACUS implements heuristic protein selection criterion that helps overcome

some of those problems. Similar approaches can be applied to derive protein abundance using peptide ion intensity based quantitation. Throughout this thesis, we use spectral counting as the sole quantitation approach for estimating peptide and protein abundances. The methods presented here can, however, be extended to peptide ion intensity based quantitation.

Identifying *bona fide* interactions from AP-MS data

In an ideal world, affinity purification protocols are expected to selectively isolate protein complexes from the sample. However, what we see in practice is a modest to high ‘enrichment’ of the protein complex. This implies that the enriched sample has significant amount of interactions that are **not** *bona fide* members of the complex. Such ‘background’ interactions are largely due to proteins that stick to the affinity support, epitope tag and/or the matrix/column. Interactions that are not *bona fide* are also commonly referred to as ‘non-specific interactions’ or ‘non-specific background’. The phrase ‘non-specific’ is used to emphasize that they are not specific to the bait or the complex. As alluded to briefly earlier, stringent washing helps reduce such non-specific background, but it results in the loss of weak and transient interactions. On the other hand, less thorough washing leads to higher background and a proportionally higher false positive rate. However, it is better suited to capture weak and transient interactions. A plausible solution for this conundrum is to use a liberal purification regimen and develop alternative strategies to distinguish *bona fide* interactions from the non-specific background. We discuss several such strategies in this thesis.

Some of the early strategies for distinguishing *bona fide* interactions are a) application of frequency filters and b) simple background subtraction using negative controls. The background profile in AP experiments depends on the experimental protocol. In large scale studies comprising several AP-MS experiments performed using the same protocol; non-specific interactions are likely to be observed more frequently than *bona fide* interactions. Accordingly, the frequency of detection of a prey across multiple experiments can be used to distinguish *bona fide* interactions from non-specific background. The background subtraction approach requires negative controls (Figure 1-2). Negative controls are mock AP-MS purifications that do not contain the bait. In epitope tag-based approaches, tag-only purifications and experiments

performed using a control protein (e.g. green fluorescent protein) are typically used as negative controls. Both these scoring approaches are based on the presence/absence of prey proteins in a sample and do not utilize their abundance estimates. Accordingly, they incorrectly estimate the ‘enrichment’ / ‘specificity’ of interactions and discard some *bona fide* interactions.

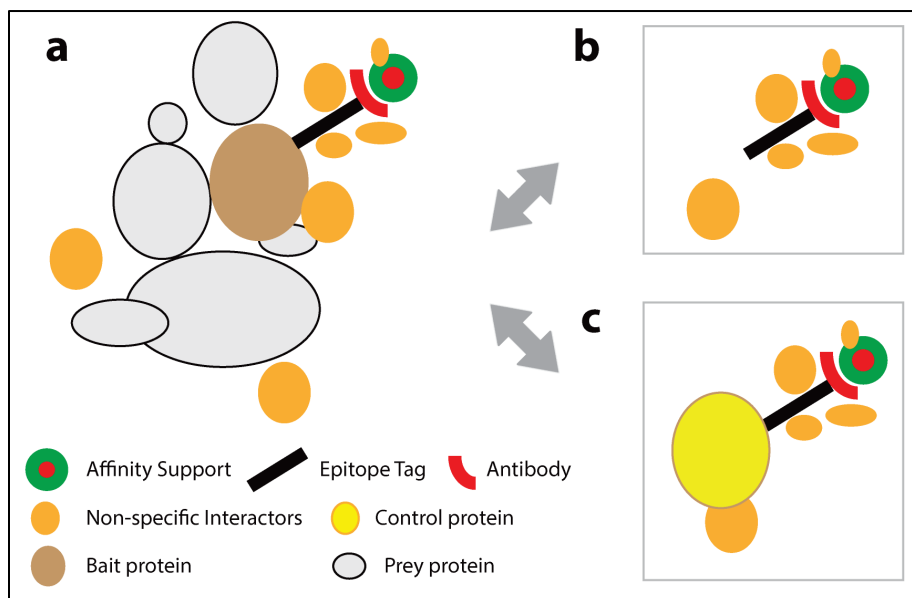


Figure 1-2: Identifying *bona fide* interactions using negative controls.

(a) Schematic representation of the composition of a protein complex comprising an epitope tagged bait protein and its interacting partners (prey). In addition to the ‘bait’ and ‘prey’ proteins, several non-specific, background interactors exist in the purified sample. (b) Schematic representation of a tag-only purification that profiles the background interactions. (c) Schematic representation of a mock AP-MS purification with a control protein. (b) and (c) are typically used as negative controls and are used for background subtraction from an AP-MS experiment.

Advances in mass spectrometry over the past two decades have resulted in an enhanced ability to quantitate peptides and proteins (see ‘Quantitation of peptides and proteins’). Accordingly, current approaches for distinguishing *bona fide* interactions from the background are based on ‘scoring’ interactions using abundance estimates. Several such scoring schemes are discussed in chapter 2.

Overview of thesis

This thesis is devoted to developing a computational and informatics framework for systematic analysis of AP-MS data (Figure 1-3). It comprises of two processing pipelines and two repositories as detailed below.

In **Chapter 2**, we present a pipeline for scoring **protein interactions** (SPrint). The ‘enrichment scoring module’ and the ‘specificity scoring module’ form the core of SPrint. The enrichment scoring module implements SAINT [20] and a newly developed empirical fold change (FC) score. Both these scoring functions estimate the enrichment of a prey in the true purification, compared to negative controls; highly enriched prey are potentially *bona fide*. The specificity scoring module implements CompPASS [21] and a newly developed enrichment specificity score (EScore). Both these scoring functions estimate the specificity of a prey to one or a few bait purifications in a data set; highly specific preys are potentially *bona fide*. While the computation of enrichment scores requires negative controls, specificity scores can only be calculated in medium/large scale data sets comprising several bait purifications. In summary, SPrint is a versatile tool that can score interactions from a wide variety of data sets. Integrated visualization tools help in filtering data and identifying high confidence interactions (HCIs). Such HCIs are pieced together to generate protein interaction networks.

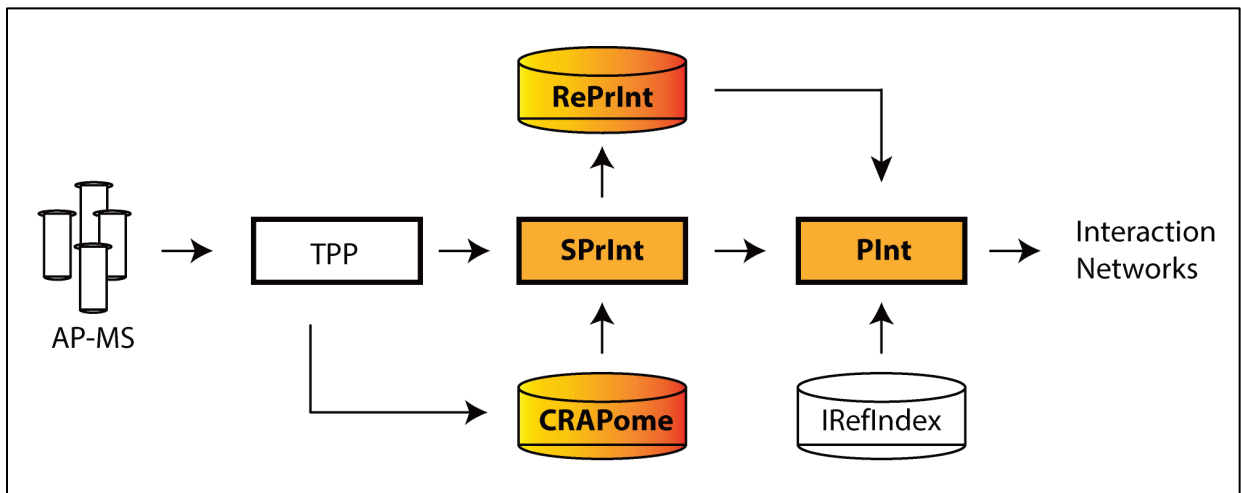


Figure 1-3: Computational and Informatics framework for analysis of AP-MS data.

(a) Protein complexes are purified using affinity purification and analyzed using mass spectrometry. (b) Protein identification and quantitation is performed using the transproteomic pipeline (TPP). (c) SPrint is a tool for scoring interactions (chapter 2). (d) CRAPome is a database of standardized negative controls that assist in scoring interactions. CRAPome is created using negative controls processed systematically through TPP and annotated using a controlled vocabulary (chapter 4). (e) Plnt is a tool for reconstruction of protein interaction networks using high confidence interactions generated by SPrint and prior knowledge from IRefIndex. (f) Several AP-MS data sets are processed through SPrint and scored interactions are stored in a repository that can assist in the reconstruction of protein interaction networks (RePrint). The database allows systematic aggregation and standardized processing of AP-MS data. The associated tools provide novel approaches for network reconstruction (chapter 5).

In **chapter 3**, we present Plnt, a tool for reconstruction and analysis of protein interaction networks. The tool provides several options for integrating prior knowledge with user-generated data. Such integration of prior knowledge facilitates enhanced interpretation of networks generated from small scale data sets. Medium and large scale networks are better understood when they are dissected to identify the constituent network modules. Plnt implements a streamlined approach to identify and zoom into subnetworks. In summary, Plnt is an integrated and versatile tool for network generation and analysis.

The performance of enrichment scoring schemes is dependent on the availability of good negative controls. In many small scale studies, such high quality negative controls are not available. Fortunately, negative controls are largely bait independent. Hence, aggregating negative controls from multiple AP-MS studies can increase coverage and improve the characterization of background associated with a given experimental protocol. This motivated the creation of the publicly available repository of standardized negative controls, called the Contaminant Repository for Affinity purification (CRAPome). The database and associated tools are presented in **Chapter 4**.

The availability of a standardized pipeline for systematic analysis of AP-MS data provides us with a unique opportunity to create a repository of protein interactions (RePrint). Almost all of the currently available public protein interaction databases follow a manual/automatic curation approach to aggregate protein interactions. Inherent limitations of both manual and automatic curation lead to high false positive rates and/or reduced scope of interactions. RePrint takes a completely data-driven approach that is based on aggregating raw spectral data and making available uniformly scored protein interactions. A novel network reconstruction algorithm has been developed to generate comprehensive interaction maps. The algorithm lays a particular emphasis on identifying *bona fide* interactions that are weak and transient. Such interactions, by their very nature, are not expected to be high scoring and may fail to qualify as *bona fide* in typical network reconstruction approaches. A new scoring function (RScore) has been developed to 'merge' evidence from multiple data sets, when available. RScore is particularly

useful in boosting the confidence of *bona fide* weak/transient interactions. The creation and utility of RePrint (and associated computational tools) is presented in **Chapter 5**.

CHAPTER 2

SPrInt: A Framework for Scoring Protein Interactions from Affinity Purification Mass Spectrometry Data

Parts of this chapter have been published by the author in Nature Methods [1].

Introduction

Affinity purification mass spectrometry (AP-MS) is a powerful technology for analyzing protein-protein interactions [22]. In the commonly used tag-based AP-MS approach, protein complexes are purified using antibodies directed against the epitope tag of a clonally modified member of the complex (also called the ‘bait’) and analyzed on a mass spectrometer. While the co-purifying proteins (also called the ‘prey’) comprise the *bona fide* members of the complex, the process also generates significant amount of non-specific background interactions [1]. Accordingly, identifying *bona fide* protein interactions from AP/MS data is a challenging task. Several informatics strategies have been developed to score interactions and filter out the background. In small and medium-scale data sets comprising of one or a few bait purifications, scoring is typically based on the comparison of the ‘true’ purification with negative controls generated in parallel. In large-scale data sets comprising multiple bait purifications, the specificity of prey to bait purification can also be used to distinguish genuine interactions from non-specific background contaminants. We present here SPrInt, an online tool for **scoring protein interactions** from AP/MS data. The ‘enrichment’ scoring module of SPrInt provides two models (empirical fold change FC Score and SAINT [20]) to compare the true purifications with negative controls. The ‘specificity/network’ scoring module also provides two scoring models (new EScore and previously published CompPASS [21, 23]) to generate metrics of specificity that assist in filtering out the background. SPrInt generates several visualizations of the scored

interactions that assist the user in filtering the data. An integrated network reconstruction tool (PInt, described in chapter 3) facilitates the creation and visualization of protein interaction networks using *bona fide* interactions and prior knowledge. SPInt is also integrated with the CRAPome database [1], a large repository of standardized negative controls that can assist in improving the scoring outcomes. The tool was originally developed as workflow 3 of the CRAPome database, but has been improved ever since its original publication. All tools are publicly available at www.crapome.org.

Methods

Design and implementation of SPInt

SPInt was designed in the model-view-controller architectural framework [24] and is available as a web service. The user interface (UI) was developed using Drupal, an open-source PHP-based web framework, and MySQL and SQLite databases. The pipeline for processing user input data and computing interaction scores was developed using Python. The visualization module that generates summary reports of scored interactions was developed using the google visualization API. The system is deployed on a server managed by the University of Michigan Medical School Information services (MSIS) using Apache, an open source web server. The UI facilitates uploading the user data, selecting various analysis options and computing scores. User analyses (referred to here as ‘user jobs’) are managed using TORQUE, an open source computing resource manager, and are processed on a first-come, first-served basis. Processing user jobs via TORQUE enables streamlined utilization of server resources and enhances the stability of the system. SAINT analysis is computationally intensive and is hence executed on FLUX, a shared high-performance computing service at the University of Michigan. Robust system integration between the web server and FLUX was designed in a fire-and-forget approach that minimizes connection failures (Appendix A). In this approach, the web server ‘fires’ a processing request to FLUX and ‘forgets’. FLUX picks up a request, processes it on a first-come, first-served basis and sends the results back to the web server. A daemon on the web server receives the results and presents them to the user. In addition to professional data backup and system management, both FLUX and MSIS provide on-demand scalability of

computing resources. The modular design of the software system facilitates easy expansion of functional capabilities of SPrint, like incorporating new scoring and visualization modules. Details of the software architecture are provided in Appendix A.

Scoring modules

SPrint comprises two scoring modules: enrichment- and specific- based. The enrichment scoring module implements an empirical scoring function (FC score [1]) and a probabilistic scoring model (SAINT [20]). The specificity scoring module implements two empirical scoring functions: EScore (unpublished) and CompPASS [21]. Each of the functions is described in detail below, with an emphasis on their relevance and applicability.

Interaction scoring: FC score

The primary FC score (FC-A, or just FC) can be considered an alternative to SAINT scoring described in [20]. It is computed for each bait-prey interaction pair (initially separately for each biological replicate of the bait). It is defined as the ratio of the normalized spectral count of protein i in purification with bait j , $T_{i,j}$, to the average normalized spectral count of that protein across the negative controls (user controls or selected CRAPome controls), C_i , calculated as follows.

$$FC_{ij} = \frac{T_{i,j} + \alpha}{C_i + \alpha}, \text{ where } i = \text{prey}, j = \text{bait}$$

The normalized spectral counts are computed as $T_{i,j} = \frac{SC_{i,j}}{N_j}$, where the normalization factor N_j is the sum over all proteins identified in the experiment with bait j , $N_j = \sum_i SC_{i,j}$. Similarly, the counts are normalized in each negative-control experiment $x = 1$ through n , $C_{i,x} = \frac{SC_{i,x}}{N_x}$, before the averaged normalized count across all n controls, $C_i = \frac{1}{n} \sum C_{i,x}$, is computed. A small background factor α is added to prevent division by 0, calculated as $\alpha = \frac{\beta}{ave(N_x)}$, where $ave(N_x)$ is the average normalization factor across all n negative controls. The parameter β is by default set to 1. When the bait protein is analyzed in multiple biological replicates, the FC

scores computed independently for each bait replicate are averaged to arrive at the final FC score.

The secondary, more conservative FC score (FC-B) can be used in addition to SAINT or the primary FC-A score for improved detection of several classes of challenging contaminants. It is computed as described above, except that C_i is computed by averaging the three highest normalized spectral counts across all controls (by default, using the combined set of selected CRAPome controls and the user controls, when available). Furthermore, in the case of biological replicates for the bait protein, the final FC-B score is computed by default as the geometric mean of the FC scores for each replicate.

Interaction scoring: SAINT

SAINT was described in [20]. Here the data were analyzed using SAINT options 'LowMode = 0, MinFold = 0, Normalize = 1.' In general, SAINT performance varies depending on the choice of options, especially MinFold (requiring a certain minimum fold change as a part of probability calculation) and Normalize (normalization to the total spectral count in each experiment). SAINT runs with the options specified above slightly outperform SAINT results with other options applied to these data sets (data not shown). When the bait protein is analyzed in multiple biological replicates, SAINT probabilities computed independently for each bait replicate are averaged, and the average probability (AveP) is reported as the final SAINT score. For in-depth discussion of these options, see ref. [25]. SPInt also allows alternative specifications for combining biological replicates (for example, geometric mean as a more conservative approach).

SAINT has been shown to perform well when using a sufficient number of matching negative controls (ideally, at least 3–5 controls) showing a high degree of reproducibility. At the same time, SAINT can be sensitive to changes in the spectral count distributions of a given protein in either the controls or the bait samples, and its performance may thus be affected if the bait sample quality is poor or the negative controls are heterogeneous. SAINT is also computationally intensive.

Interaction scoring: EScore

The EScore can be considered as an alternative to CompPASS score described in [21, 23]. It is computed for each interaction in medium/large scale data sets. EScore of a bait-prey pair is defined as the ratio of its log transformed FC score to the average, log transformed FC score of the same prey across all baits in the data set. Missing values are represented by zero fold change. A small background factor (μ) is added to both the numerator and denominator, to serve as a smoothing factor [26]. It is calculated as the average log transformed FC for the entire data set. The log transformation of the FC values is computed as $\log_2(1 + x)$ as opposed to $\log_2(x)$ in order to achieve a positive, monotonic transformation.

$$EScore(i, j) = \frac{\log_2(1 + FC_{i,j}) + \mu}{\sum_j \log_2(1 + FC_{i,j}) + \mu} ; i = prey, j = bait$$

$$where \mu = \frac{1}{N} \sum_{i,j} \log_2(1 + FC_{i,j}) , N = total\ number\ of\ interactions$$

The EScore has similarities to the FC score in the sense that the protein abundance of true purification ($T_{i,j}$) is replaced with the fold change score ($FC_{i,j}$), and the estimated abundance across negative controls (C_i) is replaced by the average fold change of a prey across all the baits. In other words, we attempt to estimate the extent to which the fold change of a bait-prey pair is higher than the average fold change of the prey across all baits in the data set. Both FC-A and FC-B can be used for computing the EScore, the choice of which depends on the data set. In general, the difference between the two is minimal owing to the log transformation applied to the fold change values. In the case of the test data sets, the EScore was generated using FC-B.

Interaction scoring: CompPASS

CompPASS was originally described in [23]. An improved version was subsequently described in [21]. The un-normalized WD score described in [21] is implemented in SPrint, with one minor modification; the SPrint implementation uses the average spectral abundance of an interaction as opposed to total spectral abundance (see below).

$$WD_{i,j} = \sqrt{(\lambda\omega)^p * x_{i,j}}$$

$$\lambda = \frac{N}{f_i}; \quad \omega = \frac{\sigma_j}{\bar{x}_i}$$

$x_{i,j}$ = ave. spectral count of prey i in purification with bait j

f_i = frequency of prey i across all bait purifications

p = number of replicate purifications for bait j

ω = coefficient of variation of prey i

The WD-CompPASS score can be interpreted as follows. The score for a bait-prey pair is (a) directly proportional to the average prey abundance ($x_{i,j}$), (b) inversely proportional to the frequency of identification of the prey across the data set (f_i) and (c) directly proportional to the number of times it is identified reproducibly in biological replicates (p). The weighting factor (ω) is the coefficient of variation of the prey across the data set and can be interpreted as a metric of specificity. In summary, CompPASS combines metrics of abundance ($x_{i,j}$), specificity (λ, ω) and reproducibility (p) to score an interaction. The final score includes a square root transformation, to reduce its range.

Comparison to literature data

To rapidly benchmark scoring performance and to provide users with a view of the new data in the context of previously published results, a mapping of the interactions to those deposited in the iRefIndex repository [27] (V 9.0) is provided. iRefIndex was selected because of its comprehensiveness in the number of interactions annotated and the relative ease of download and data mapping. The database is created by aggregating protein interactions from several primary interaction databases, including “BIND, BioGRID, CORUM, DIP, HPRD, InnateDB, IntAct, MatrixDB, MINT, MPact, MPIDB, MPPI and OPHID”³. Aggregated interactions are de-duplicated and systematically mapped to unique identifiers using the Secure Hash Algorithm. Each entry from the database is then mapped to a pair of genes (interacting proteins) using an in-house

³ <http://irefindex.org/wiki/index.php?title=iRefIndex>

mapping tool. Entries identified as ‘complex’ are excluded from this mapping. Owing to uncertain quality of previously reported interactions involving ribosomal proteins, which are among the most common contaminating proteins in AP-MS experiments, we excluded all RPL and RPS proteins from the computation of ROC curves shown in Figure 2-3.

Preparation of test data

Two data sets, generated by the Gingras Lab (University of Toronto, CA), were used to illustrate the ‘enrichment’ scoring module of the SPInt pipeline. The first data set (referred to here as **DS1**) has two biological replicates for each of the following four baits. RAF1 is a serine/threonine kinase that binds to Ras, several chaperones and 14-3-3 proteins [28, 29]. EIF4A2 is a translation initiation factor that is part of the EIF4F complex, which bridges the mRNA cap structure to the ribosome via the EIF3 complex [30]. WASL (also known as N-WASP) belongs to the Wiskott-Aldrich syndrome (WAS) family of proteins, involved in transduction of signals from receptors on the cell surface to the actin cytoskeleton[31]. Finally, MEPCE, the 7SK snRNA methylphosphate capping enzyme, interacts with numerous transcriptional and RNA-processing proteins[32]. The second data set (referred to here as **DS2**) has two biological replicates using ORC2L as the bait protein. Both data sets used a common set of negative controls.

Data for DS1 and DS2 was generated as follows. Cloning and expression of EIF4A2, RAF1 and MEPCE has been previously described[33]. WASL and ORC2L were amplified by PCR from Mammalian Gene Collection constructs BC052955 and BC014834, respectively, and were cloned into pcDNA5-FRT-Flag (using EcoRI/NotI for WASL and AscI/NotI for ORC2L), and the junctions were sequenced. The primers used were WASL_5'EcoRI, GATCGAATTCATGAGCTCCGTCCAGCAGC; WASL_3'NotI, GATCGGGCCGCTCAGTCTTCCCACTCATCATCATC; ORC2L_5'AscI, GATCGGGCCGCAATGAGTAAACCAGAATTAAGGAAGAC; ORC2L_3'NotI, GATCGGGCCGCTCAAGCCTCTTCTTCC. The resulting vectors were stably co-transfected with the Flp-recombinase-expressing vector pOG44 into Flp-In T-REx 293 cells (Invitrogen). Selection of stable transformants (single clones), clonal expansion, induction of protein expression and AP-MS were performed essentially as described in ref. [33] using Flag M2 agarose beads (Sigma). Two biological replicate analyses of each bait

were performed, alongside six negative controls (cells expressing the tag alone). All samples were analyzed on an LTQ mass spectrometer coupled to an online C18 reversed-phase column. The detailed protocol is no. 48 in the CRAPome database.

The data generated for DS1 and DS2 (see above) were processed separately as follows. The mass spectrometry data were searched using the X! Tandem/TPP/ABACUS pipeline [34]. MS/MS spectra were searched against RefSeq protein sequence database version 47 (ref. [35]; H. sapiens), appended with an equal number of decoy sequences, using X! Tandem [13] with k-score plug-in. MS/MS spectra were searched using a precursor-ion mass tolerance of -1 to $+4$ Da (average mass) window. Cysteine carbamylation (C + 57.0215) and methionine oxidation (M + 15.9949) were specified as variable modifications. The search results were processed using PeptideProphet [14] and then further processed using ProteinProphet [36] to create protein summary files. All the PeptideProphet results generated from individual experiments were processed together using ProteinProphet to generate a single protein summary file (protXML file). This combined protXML file, as well as the pepXML and protXML files from each individual experiment, were then processed using ABACUS [37] to generate a combined spectral count matrix using default parameters (accepting proteins with at least one peptide having PeptideProphet probability of 0.99 or greater and protein probability as computed by ProteinProphet of 0.9 or greater). Each row in the filtered ABACUS file represented a protein group from the combined protXML file, with a single accession number selected among indistinguishable protein entries forming that group. Spectral counts for the representative proteins were extracted from pepXML files for each individual experiment. The FDR for the combined protein list was less than 1% as estimated using decoy sequences. The filtered ABACUS file was manually modified to generate a matrix formatted SPrInt input file (see 'Data formats'). Data were uploaded to the SPrInt pipeline and analyzed using the enrichment scoring module. The resulting input data matrices for both DS1 and DS2 can be downloaded from <http://www.crapome.org/?q=suppdata> .

Two medium/large scale data sets were used to illustrate the 'Specificity/Network' scoring module of SPrInt. The first large scale data (referred to here as **DS3**) was from 'Interlaboratory

reproducibility of large-scale human protein-complex analysis by standardized AP-MS' by Varjosalo et al., Nature Methods (2013) [38]. The data was downloaded from peptide atlas (PASS00117) and processed using X! Tandem/TPP/ABACUS as described above (for DS1 and DS2). Since the data was collected on a LTQ Orbitrap XL mass spectrometer, precursor-ion mass tolerance of 100 p.p.m. was used for X! Tandem search. The second large scale data set (referred to as **DS4**) is from 'A quantitative chaperone interaction network for exploring the wiring of cellular protein folding pathways' [39]. For DS4, processed data was directly obtained from the authors, who generated the SPrInt input file (list format) using ProHits [40], their laboratory information management system. The details of data processing can be found in their original manuscript.

Data Formats

User data can be uploaded to SPrInt in either the list or the matrix format. The 'list' formatted input file (Figure 2-1 a) needs to contain fields for each interaction: 'Bait Name (BAIT column)', 'Experiment Name (AP Name column)', 'Prey Name (Prey column)' and 'Spectral Count (SPC column)'. Each row in this file lists the spectral count (SPC column) for each protein (referenced in Prey column) in purification with a particular bait protein (bait protein/gene identifier is referenced in the BAIT column). When multiple biological replicates for the same bait are available, they are distinguished using different text strings in the AP Name column (for example, 'R1', 'R2', etc.). Negative control runs are specified by the text string 'CONTROL' or 'CTRL' in the BAIT column (and named differently in the AP Name column: for example, 'UC1', 'UC2', etc.). The 'matrix' formatted input file (Figure 2-1 b) must have a column for each affinity purification (AP). The AP names (i.e. column names) must be unique. The first column of the matrix is reserved for prey names (PROTID column). Each prey is represented in a separate row. Every cell in the matrix contains the spectral count for the corresponding bait-prey pair in each AP experiment. Missing data is represented with zero counts. The second row of the matrix is reserved for annotating the AP names (i.e., column names specified in the first row). In this row, the bait name is specified against the corresponding AP experiment. Negative controls are annotated as 'CONTROL' or 'CTRL'.

Access to SPrint

SPrint and other tools (CRAPome, Plnt and RePlnt) were implemented as an integrated system, on a common software platform. All the tools can be accessed at <http://www.crapome.org/>. SPrint requires user registration to maintain data integrity and privacy. Users can access the results of previously performed analyses, until they are purged by the system administrator.

a	BAIT	AP Name	Prey	SPC
	bait-1	AP-1	prey-1	10
	bait-1	AP-2	prey-1	21
	bait-2	AP-3	prey-9	10
	bait-2	AP-4	prey-10	80
	***	***	***	***
	CTRL	UC-1	prey-11	11
	CTRL	UC-2	prey-2	30
	***	***	***	***

b	PROTID	AP-1	AP-2	AP-3	AP-4	***	UC-1	UC-2
	N/A	bait-1	bait-1	bait-2	bait-2	*	CTRL	CTRL
	prey-1	10	7	22	17	*	9	15
	prey-2	92	71	17	32	*	5	0
	prey-3	7	6	45	52	*	17	45
	***	*	*	*	*	*	*	*
	***	*	*	*	*	*	*	*
	***	*	*	*	*	*	*	*
	prey-n	*	*	*	*	*	*	*

Figure 2-1: Input file formats for SPrint.

Input file formats for SPrint. a) List format. In the list format, each interaction is specified in a separate row using the Bait Name, AP Name, Prey Name and Spectral Counts (SPC). b) Matrix format. In the matrix format, spectral count values from each experiment are specified in a separate column. Each row is represented by a single prey protein. The second row is used to annotate the AP names (main text).

Results and Discussion

Creation of the SPrint processing pipeline

The analysis pipeline of SPrint is shown (Figure 2-2 a). Users upload their data and score interactions using the 'enrichment' scoring module and/or the 'specificity/network' scoring module. Optionally, users can also include negative controls from the CRAPome repository (chapter 4) to supplement their own negative controls. Relevant CRAPome controls can be selected with the help of controlled vocabulary used to annotate experiments (Figure 2-2b). Each analysis, referred to as a 'user job', can be visualized and downloaded. An integrated network reconstruction and visualization tool, Plnt, is used for network generation and analysis. Plnt is described in chapter 3 of this thesis.

Uploading user data and quality control: User data can be formatted in the list or the matrix format (Methods) and uploaded for analysis through the graphical user interface (Figure 2-2 c). The input data consists of one or multiple AP-MS experiments, ideally including biological replicates. The validity of any analysis is contingent on the quality of input data; hence SPInt generates three important quality metrics to identify and weed out low quality AP-MS experiments (Figure 2-2 d). First, data (spectral counts) from each experiment are visualized using box plots; with the bait counts highlighted using a red asterisk symbol. A good AP experiment is expected to recover the bait protein with reasonably high counts. This box plot provides a visual cue on the success of an AP experiment. Second, common contaminants (e.g. keratins, tubulins) must not dominate the total yield in an experiment. If data is collected in the data dependent acquisition mode, the dominance of such contaminants indicates that the real interactors were not sampled and sequenced adequately. Venn diagrams indicating the fraction of data that can be attributed to common contaminants in each experiment assist in determining the quality of the sample. Finally, the similarity among replicates for each bait purification helps in choosing the appropriate strategy for combining scores (Methods). Scatter plot generated using replicates assists in determining the replicate similarity. While we recommend that the users ensure the availability of adequate number of negative controls in each data set, we have shown that standardized negative controls from the CRAPome database can assist in discriminating true interactors and contaminants (chapter 2).

Scoring interactions and visualizing results. Enrichment analysis is performed using the significance analysis of interactome (SAINT) score [20, 25, 41] and/or a simpler fold-change (FC) calculation (detailed below). The specificity analysis is performed using CompPASS [21, 23] and EScore (detailed below). The options for all the scoring functions can be specified through the GUI (Figure 2-2 e, f). User jobs are queued and processed on a first-come, first-served basis. The results of completed jobs are presented in a tabular format and can be downloaded as a tab-delimited file. Previously reported interactions documented in the interaction database aggregator iRefIndex (V 9.0; ref. [27]) are also mapped onto user data. Summary graphical views of the data are provided for each bait protein (Figure 2-2 g) or for all baits combined, enabling the user to view their data at a glance. These visualizations, especially the ROC curves

generated using comparisons with literature data (Methods), can assist the users to determine appropriate cutoffs to filter the data. The network reconstruction tool (PInt) can be launched from this results page ('View Network in PInt' hyperlink) and interaction maps can be generated (Figure 2-2 h).

Analysis of small scale data sets

Eliminating background contaminants in small scale data sets is based on enrichment scores that compare true purifications with negative controls. The enrichment scoring module of SPrInt implements two complementary scoring strategies, both based on quantitative comparisons of abundance levels (estimated using spectral counts) of co-precipitating proteins (prey) in purifications with bait proteins against the distribution of prey abundances across a set of negative controls (Methods). SAINT, described previously [25, 41-43], performs advanced statistical modeling of the input bait-prey spectral count data and reports a posterior probability of true interaction. A simpler FC calculation is based on the ratio of average normalized spectral counts in bait purifications to negative controls. FC scoring is customizable and, in addition to the calculation of the standard FC score (referred to as primary score, or FC-A), involves the computation of a secondary, more stringent score (FC-B; see below). Both FC and SAINT calculations are run in parallel, allowing specification of key model parameters via the user interface[25]. A comparison of their relative performances for each of the tested baits can be assessed by a receiver operating characteristic (ROC) analysis. These curves are generated automatically by the SPrInt pipeline for every analysis.

The analysis of a small scale data set using SPrInt is illustrated here with the help of two data sets, DS1 and DS2 (see 'preparation of test data'). DS1 consists of two biological replicates of each of the following four baits: RAF1, EIF4A2, WASL and MEPCE. MEPCE and EIF4A2 have many documented interactors [27], whereas WASL and RAF1 have fewer known interactors; all proteins provide challenges for background definition because of their association with polypeptides with contaminant-like behavior (chaperones, cytoskeletal proteins, RNA-binding proteins and so on; Table 4-3). DS2 consists of two biological replicates for ORC2L bait. In

addition, six matching controls (user controls) were generated that are applicable to both DS1 and DS2. Data was processed as described (Methods) and interactions were scored using SPrint.

The results generated by DS1 were evaluated by plotting ROC curves based on the information extracted from iRefIndex [27]. The protein interaction list (all four baits combined) was sorted on the basis of either the SAINT probability or the FC-A score computed using the six user controls (Figure 2-3 a). Although SAINT outperformed the FC-A score on this data set, both scoring schemes were able to efficiently recapitulate known interactions from the literature. Both scores also tracked very similarly for most of the proteins analyzed (Figure 2-3 b), with SAINT essentially providing a statistical conversion of the FC score onto the probability scale via the mixture-model analysis of the underlying spectral count distributions. We further visualized the performance of the interaction scores by plotting the distribution of scores (histograms) separately on the basis of iRefIndex annotation, which showed that high-scoring interactions (SAINT probability above 0.9, FC score above 4) are clearly enriched for previously reported interactions (Figure 2-3 c, d). The SPrint interface provides (both separately for each analyzed bait and for all baits combined) an ROC and a histogram view (with mouse-over function), which enables the user to explore the reported interactions at different scores for SAINT or FC and assists in establishing appropriate thresholds.

One issue affecting the scoring of AP-MS data is the existence of contaminants (for example, myosin and the proteins that co-purify with it) that are usually present in small amounts across most controls but that can spike to high abundance in some controls (or across batches of purifications), making detection of the true interactors much more difficult. Such contaminants are normally 'diluted out' when multiple experiments are used for FC calculation, or even SAINT analysis (Figure 2-4 a). To assist in the identification of these 'rare' contaminants, we implemented a more conservative FC score, FC-B, that is automatically calculated to supplement normal scoring using SAINT or the FC-A score (Methods and Figure 2-4 a). We applied this more stringent scoring scheme to DS2, which, through visual inspection of the results, was found to contain large quantities of myosin contamination. Although SAINT is capable of identifying true interactors in successful experiments, as exemplified by EIF4A2 of

DS1 (Figure 2-4 b; note the relatively good agreement between the SAINT score and FC-B score), it assigned a high probability to myosins and associated proteins in the ORC2L samples (Figure 2-4 c). By contrast, the FC-B scores readily distinguished between these contaminants and true interaction partners (ORC3, ORC4 and ORC5 are in iRefIndex [27], and LRWD1 is reported in PubMed [44]). Notably, the SPrint interface enables rapid visualization of the samples likely affected by this type of low-frequency contaminants by providing comparisons between FC-B and SAINT or FC-A (Figure 2-2 g).

Analysis of medium/large scale data sets

The background contaminant profile is largely dependent on the experimental protocol and less on the bait protein itself. Accordingly, when several bait purifications are performed under similar experimental conditions, *bona fide* interactions are more specific than contaminants. Hence metrics of specificity of a prey to one or few baits can be used to determine whether the prey is promiscuous interactor or not. Several scores have been developed in the past including HGScore [45], MiST [46] and CompPASS [21, 23]. In addition to such scores, we developed a simple yet robust enrichment specificity score (EScore) that estimates specificity based on a primary enrichment score (Methods). The ‘Specificity/network’ scoring module of SPrint implements EScore and CompPASS. While EScore is based on the FC scores (generated by the ‘enrichment scoring’ module of SPrint), CompPASS can be computed directly using the spectral count values. In that context, these scores can be considered complementary to each other and are generated in parallel by SPrint.

The utility of specificity scores is illustrated here with the help of two medium/large scale data sets (DS3 and DS4, see ‘Preparation of test data’). DS3 is from an inter-laboratory study comprising of 32 human kinases [38]. Only the subset of data generated by the group at ETH Zurich (ETHZ) is used here. DS4 is a non-typical data set that profiles the chaperone network (HSP-90 interactome) [39]. Only the subset of data generated using FLAG-AP purification (VS1) is used here. Being ‘sticky’ proteins themselves, chaperones have a high propensity to manifest as non-specific interactions in typical AP-MS experiments. As scavenging proteins, they also interact with a several proteins in the cell, including common contaminants such as tubulins.

While the ETH samples were purified using a two-step (tandem) protocol, the HSP-90 samples were generated using single step purification. Both data sets have at least two biological replicates for each bait purification and more than six negative controls.

Both DS3 and DS4 were first subjected to enrichment scoring. Even after stringent thresholds (SAINT ≥ 0.9 and empirical FC-A ≥ 4) were applied to filter the data, several prey that have a high propensity for non-specific interaction (such as tubulins, keratins and ribosomal processing proteins) were found in the filtered list. While it is possible that the negative controls in the analysis were not able to accurately model the background profile of such proteins, they may well be genuine interactors. Typically, they are manually excluded from the analysis on a case by case basis. Fortunately, medium/large scale data sets lend themselves for additional scoring using specificity metrics. A combination of EScore and CompPASS scores was able to distinguish such promiscuous and non-specific interactions from specific and potentially *bona fide* interactions (Figure 2-5). As expected, common contaminants (tubulins, keratins and ribosomal proteins) scored low in both DS3 and DS4 (black triangles in Figure 2-5 a, b respectively). DS4 is a non-typical data set that characterizes the HSP-90 interactome. As scavenging proteins, they interact with several proteins in the cell, including common contaminants in AP-MS experiments (Table 4-3). As expected, the same set of common contaminants that scored low in DS3 were relatively high scoring in DS4. Taken together, these observations indicate that a combination of EScore and CompPASS, which employ metrics of specificity to score interactions, can be used to differentiate *bona fide* interactions from promiscuous and non-specific background contaminants in large/medium scale data sets.

Additional confirmation of the performance of specificity scores is provided by the fact that trypsin, which was found in high quantities across several bait purifications but not in negative controls in the HSP-90 data set (DS4), scored low on both EScore and CompPASS (Figure 2-5 b). Trypsin is used in sample preparation to digest proteins and is not a part of the cell lysate. It is typically removed manually before scoring interactions, but is used here for validating the performance of the scores.

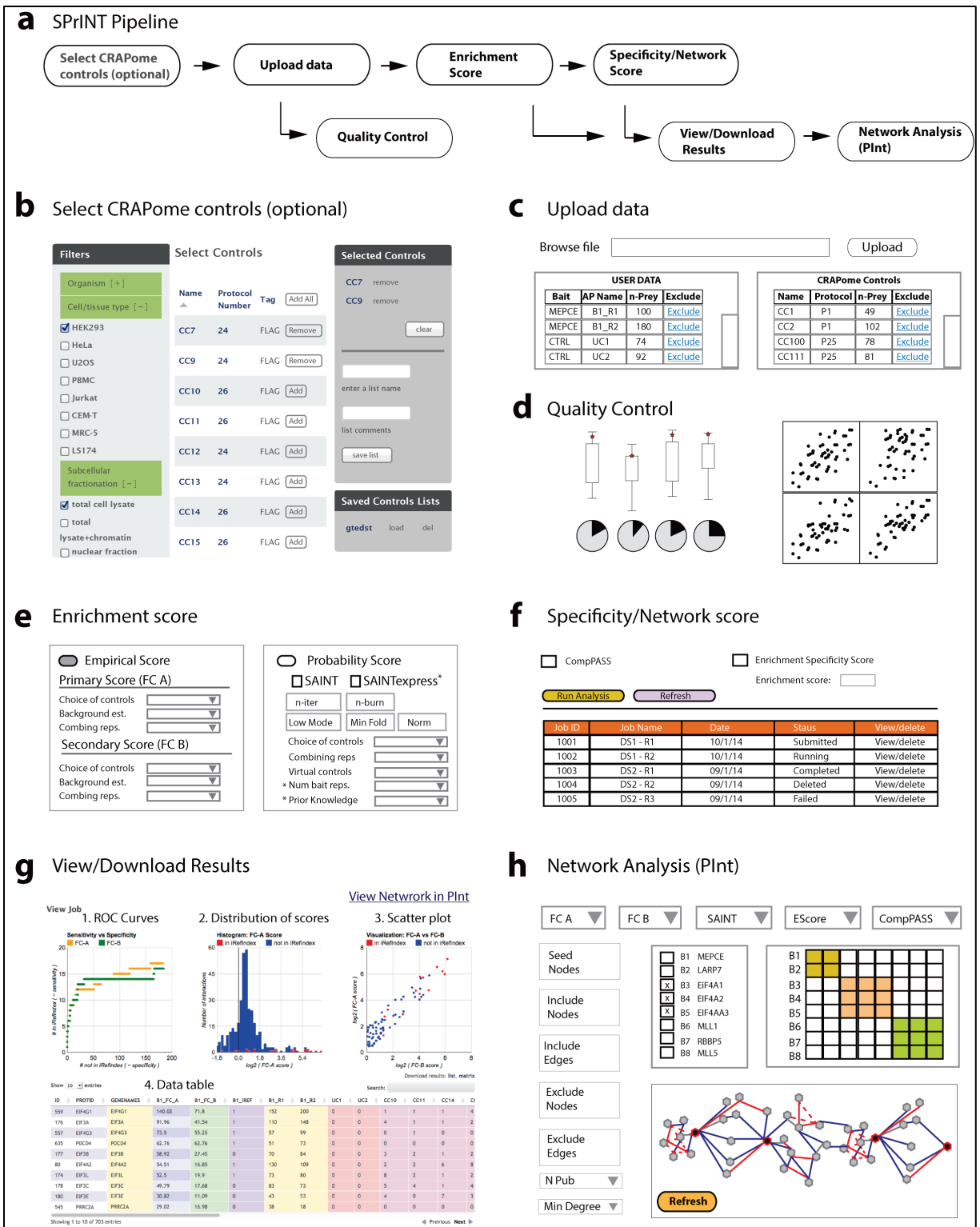


Figure 2-2: SPRINT pipeline and the graphical user interface.

(a) Overview of the SPRINT pipeline. CRAPome controls are selected (optional), data is uploaded and interactions are scored online using the ‘enrichment’ and/or ‘specificity/network’ scoring modules. Results are visualized and

filtered interactions are used for network reconstruction using PInt (main text). (b) The graphical interface for selecting CRAPome controls is shown. Desired controls are selected with the help of CVs. (c) Data is uploaded in either list or matrix format (see 'Data formats'). (d) Several visualizations are generated to assess the quality of input data (main text). (e, f) The options for enrichment and specificity scoring functions are specified through the graphical user interface. User requests are processed on a first-come, first-served basis. The status of a submitted job is displayed on the user interface ('Status'). (g) Scored interactions are visualized online. ROC curves (panel 1), and distribution of scores (panels 2) assist in determining appropriate cut-offs. Each interaction is also plotted against two selected scores to easily identify high scoring interactions (panel 3). Users can mouse over the points on the plot to view the details. The actual data is also presented in a tabular format (panel 4). PInt can be launched from the results screen ('View Network in PInt' hyperlink in the top right corner). (h) PInt interface (see chapter 3).

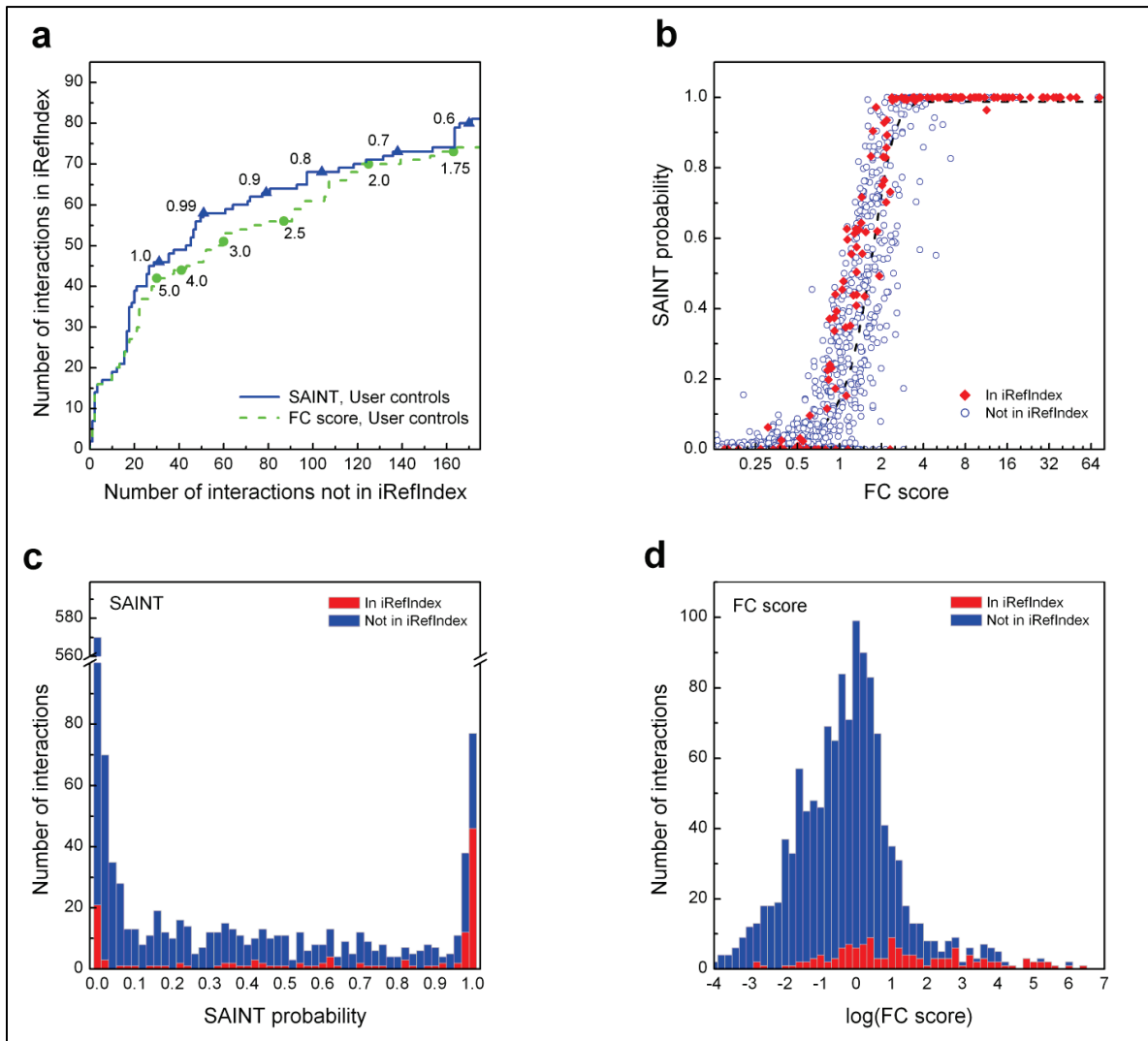


Figure 2-3: Enrichment scoring functions of SPInt illustrated on a four-bait data set.

(a) Comparison between the FC-A and SAINT for interactions scored using negative-control runs ($n = 6$) provided by the user; the receiver operating characteristic is based on the interactions in iRefIndex. Note that when SAINT scores are identical, ties are broken by the FC-A score. Selected SAINT probability or FC-A score thresholds are represented by triangles and circles, respectively. (b) The relationship between SAINT probability and FC score is

well represented by a sigmoid function (dashed curve). (c,d) Histogram visualization of the data presented in b can help with data exploration and threshold selection.

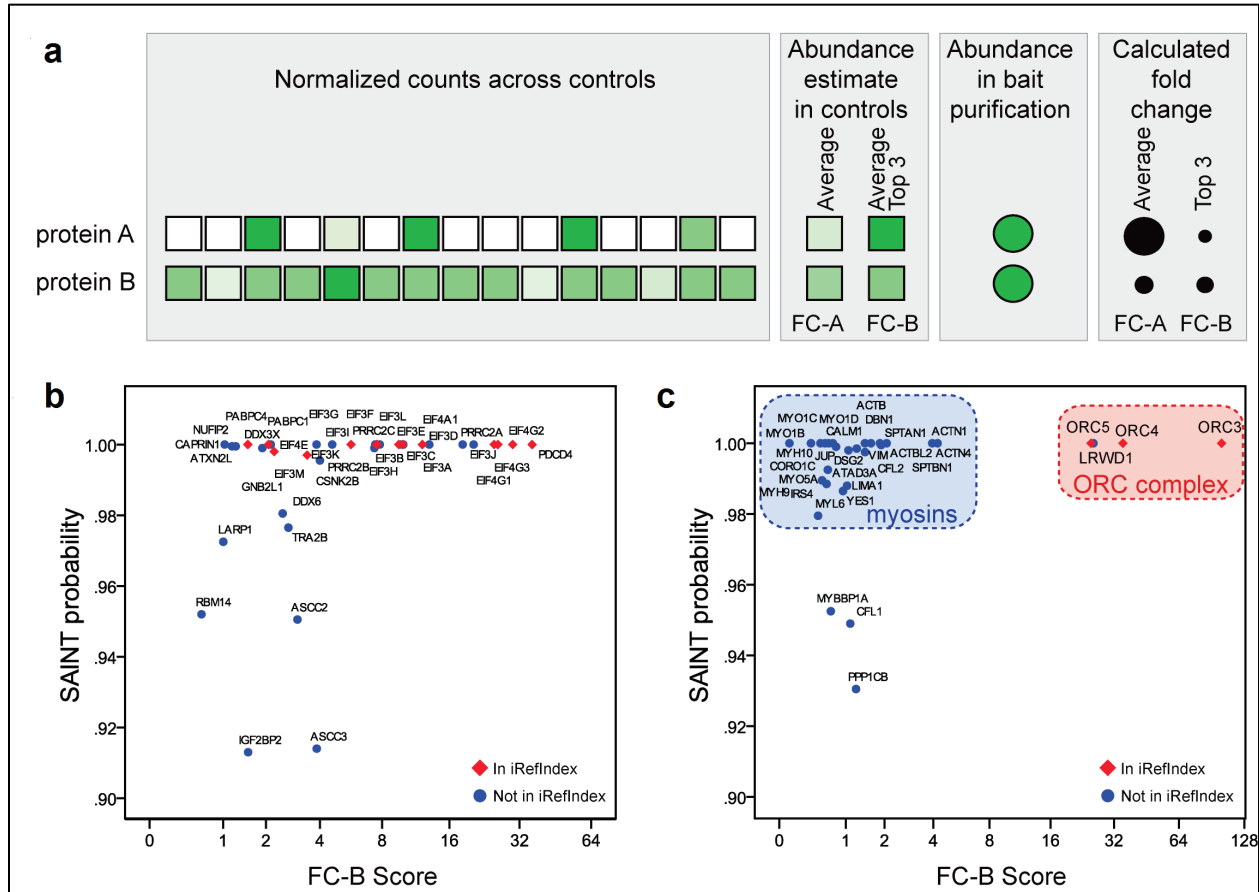


Figure 2-4: Utility of FC B in filtering sporadic contaminants.

(a) Illustration of the consequences of averaging all spectral counts instead of selecting the top three maximal values for scoring protein-protein interactions. Protein A represents a contaminant in the purification scheme that is detected with variable counts across the 15 selected controls (the intensity of shading is proportional to the spectral counts). By contrast, protein B is a contaminant detected with similar counts across all selected controls. The FC-A calculation averages the counts across all controls, whereas the more stringent FC-B score takes the average of the top three highest spectral counts for the abundance estimate. The resulting FC-A and FC-B scores are represented schematically, where a larger circle indicates a higher fold change, with FC-A and FC-B assigning a similar score to protein B but not to protein A. (b) Comparison of SAINT & FC-B scoring with good bait samples. Note that only the top of the map (the interactions with SAINT probability ≥ 0.9) are displayed. (c) Same as b for bait samples (ORC2L) contaminated with myosin: the more stringent fold-change score FC-B helps in discriminating between true interaction partners (labeled “ORC complex”) and contaminants (labeled “myosins”).

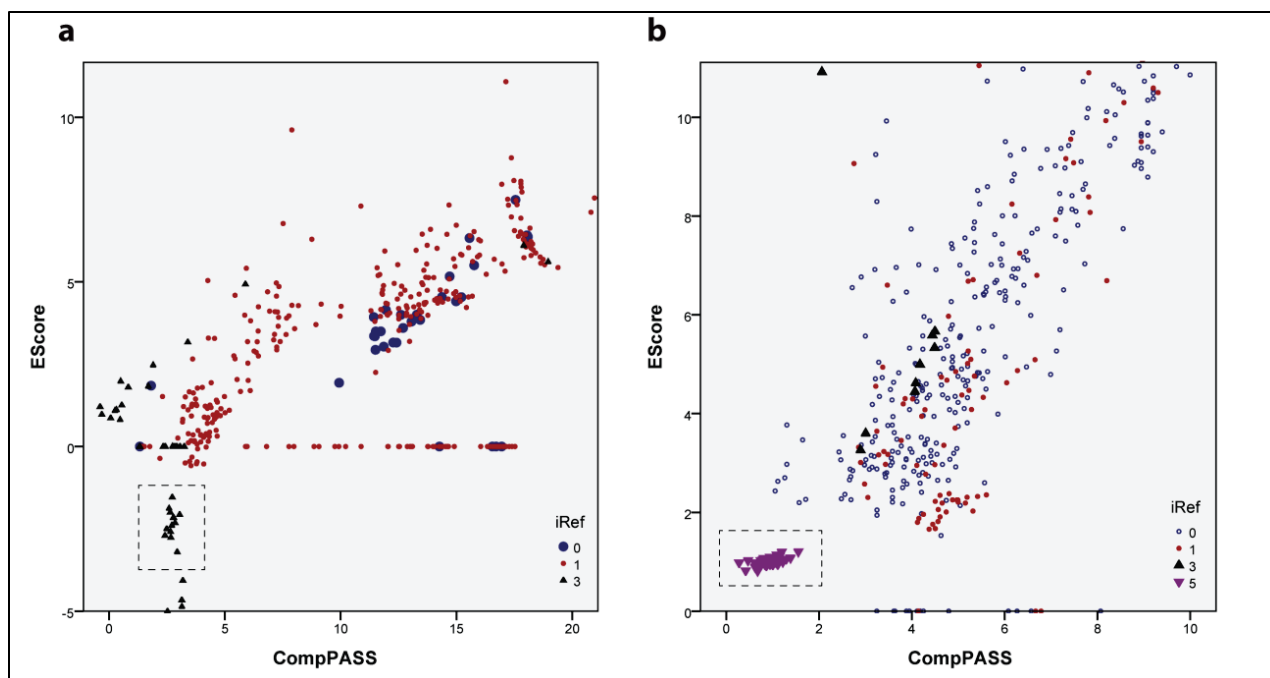


Figure 2-5: Specificity scoring functions of SPInt illustrated using two medium scale data sets.

(a) Using specificity scores to identify promiscuous, non-specific interactors illustrated using DS3 (Methods). The EScore and CompPASS scores of highly enriched interactions (SAINT 0.9 or higher and FC-A 4 or higher) are shown here. Red dots indicate that the interaction was documented in iRefIndex database. The blue dots correspond to high scoring interactions that were not documented in iRefIndex (Methods). The black triangles are common contaminants (tubulins, keratins and ribosomal proteins). (b) Utility of specificity scores illustrated using DS4 (Methods). The inverted purple triangles correspond to trypsin, a common contaminant in AP-MS experiments. It is used here as a positive control, to demonstrate the performance of specificity scores (main text). In both data sets, common contaminants (represented by black triangles) have high enrichment scores, but relatively low specificity scores. The same set of common contaminants are relatively high scoring in DS4 (b) compared to DS3 (a). This observation is attributed to the non-typical nature of DS4, which profiles the HSP-90 interactome (main text).

Concluding Remarks

SPInt is a freely available versatile tool for scoring interactions and can handle a wide variety of AP-MS data sets. An integrated network reconstruction and visualization tool (PInt) helps generate and visualize interaction maps. The pipeline is also integrated with the CRAPome repository, which makes available standardized negative controls. The performance of enrichment scoring models can potentially be improved by using CRAPome controls. An easy-to-use web interface facilitates rapid analysis and visualization, even for those who may be new to mass spectrometry.

SPrInt currently uses spectral counts as the sole quantitation. While spectral count data has its own advantages, there is a growing interest in using peptide ion intensity to measure protein abundance. Accordingly, scoring models that can optimally process intensity data, such as SAINT-MS1 [47], need to be incorporated into SPrInt. In addition, strategies to use both spectral counts and peptide ion intensity in parallel to score interactions need to be developed.

iRefIndex database is currently used to rapidly benchmark scored interactions. Additional databases, including the RePrInt database described in chapter 5, need to be incorporated to expand the scope of prior knowledge. The pipeline also needs to be improved to accept user provided list of interactions to enable the processing of special data sets, such as host-viral interactomes.

Contributions

This work is the result of collaboration between the Nesvizhskii Lab (Univ. of Michigan, Ann Arbor) and the Gingras Lab (Univ. of Toronto, CA). Bioinformatics pipelines/methods were developed by Dattatreya Mellacheruvu, under the guidance of Dr. Alexey Nesvizhskii. Dattatreya Mellacheruvu and Zachary Charles Wright (Application programmer (sr), Univ. of Michigan, Ann Arbor) implemented the system.

CHAPTER 3

PInt: A tool for analysis of Protein Interaction Networks

Introduction

Protein-interaction networks (PINs) provide a blue print of the underlying biological mechanisms of the cell. They also assist in functional interpretation of differentially expressed genes/proteins from other studies that compare normal cells with disease/mutant variants. Reconstruction of protein interaction networks essentially involves piecing together *bona fide* protein interactions. SPInt (presented in chapter 2) is a versatile tool for scoring interactions; high scoring interactions (HCIs) are potentially *bona fide*. We present here **PInt**, a tool for reconstructing and visualizing PINs using HCIs generated by SPInt. Both tools are tightly integrated; PInt is launched from the user interface of SPInt.

Small scale data sets generate incomplete interaction maps. Hence, integration of prior knowledge is critical for the analysis of such small scale networks. However, extracting relevant prior knowledge is not straightforward, owing to the fact that biological molecules (here proteins and core protein complexes) often have multiple roles/functions within the cell and are hence shared among several sub networks. PInt provides two models for extracting relevant prior knowledge (which we call the 'network context') from publicly available protein interaction databases.

PINs generated from medium/large scale data sets comprehensively profile the landscape of the interactome. The task of interpreting such complex networks can be simplified by narrowing down to the sub-network of interest. PInt provides a mechanism to identify and zoom into tightly connected sub-networks. Further, all networks generated by PInt can be downloaded in the portable 'graphml' format, for analysis using more sophisticated tools.

Reconstruction and analysis of biological networks (in this case PINs) is often a semi-manual and iterative process. Accordingly, computationally intensive (and hence time consuming) network queries of PInt are optimized. A simple to use web interface, streamlined network analysis framework and a software design that minimizes latency enable PInt to serve as a powerful tool for the generation, rapid visualization and analysis of protein interaction networks from AP-MS data. Additionally, PInt can also create quantitative networks by qualifying the edges with enrichment/specificity scores generated by SPInt.

PInt is an important component of a suite of computational tools (SPInt, PInt, CRAPome and RePInt) for analyzing AP-MS data. Post publication, it will be publicly available at www.crapome.org.

Methods

Design and implementation of PInt

The design principles of PInt are similar to that of SPInt described in chapter 2. The user interface was developed using Drupal and MySQL and SQLite relational databases. The processing pipeline is developed using Python and SQLite. Network generation is performed using the 'networkx' library of Python and a 'graphml' file is generated. This output file can be downloaded and imported to advanced network analysis tools like GraphViz, Cytoscape [48], etc. In order to facilitate rapid visualization, PInt also generates browser embedded network visualization using Cytoscape web [4]. Prior knowledge represented in iRefIndex database is stored in neo4j⁴, an open source persistent graph database, which is designed to optimize the performance of network queries. PInt is hosted on a virtual server managed by the Medical School information Services (MSIS) of the University of Michigan. Apart from providing professional maintenance and backup services, the computing infrastructure managed by MSIS can easily be upgraded and scaled.

⁴ <http://neo4j.com/>

Generating a database of interactions from literature

iRefIndex [27] is an aggregator of protein interactions from various primary sources and was used to represent the set of interactions from literature (prior knowledge). The interaction file (V 9.0) was downloaded and parsed using an in-house Python script. Protein IDs are mapped to their gene name(s) using ID-to-gene name mappings downloaded from the ensemble 'BioMart' database and a list of interactions was generated, where the proteins are referenced by their gene names. When multiple gene names map to one (or both) of the members of an interaction, the entry is duplicated corresponding to each (mapped) gene symbol. The process generates a redundant list and inflates the size of the database, but is helpful to resolve issues with mapping 'user' data to prior knowledge. Common contaminants (tubulins, keratins and ribosomal proteins; Table 4-3) are excluded during the process. While this helps in reducing a large number of false positives, it may also exclude a few genuine interactions. Protein complexes represented in iRefIndex were also omitted, pending analysis on whether the spoke model or the matrix model is better suited for representing such complexes. Finally, the list of interactions is stored in neo4j. The nodes are indexed to accelerate queries against the database. Data is refreshed periodically (approximately on a half yearly basis) using updated iRefIndex files and ID mappings.

Generating the network context

Two approaches to generate a network context were implemented in Plnt. The first approach (which we refer to as the 'simple' approach) involves querying the neo4j database (see 'Generating a database of interactions from literature') using a list of proteins relevant to the network of interest. This list is generated manually by the user and provided as an input through the user interface ('include nodes' field, Figure 3-1 b). Interactions in the database that connect any two proteins in the list are used to generate the network context. In other words, this approach attempts to connect proteins in the input list with direct links (edges). A more comprehensive approach (which we call the 'scaffolding' approach) is to use a set of nodes (referred to here as the 'seed' nodes) to generate a network by connecting each seed node with every other seed node with either a direct link or the shortest path between them. In

other words, the scaffolding approach attempts to connect seed nodes either directly or by introducing new nodes. Such new nodes are also referred to as ‘white’ nodes in STRING-db.

Using bait-bait cluster gram to identify sub-networks

For our purposes, sub-networks comprise of closely related baits. We define relatedness of two baits as the degree of shared interactions between them. For every data set analyzed using Plnt, cluster analysis is used to identify sub-networks. First, scored interactions are filtered to generate a list of high confidence interactions (HCIs). Interactions with a high SAINT probability (≥ 0.9) are defined to be HCIs. A bait-prey matrix is generated using this list of interactions. Each bait-prey pair in this matrix is represented by its square root transformed average spectral count value. The square root transformation helps reduce the range of spectral count values. In addition to implicitly normalizing the data, it also reduces the variance in the data set. A bait-bait matrix is then generated, where each cell ($C_{i,j}$) in this matrix is represented by the correlation between columns i and j in the bait-prey matrix. In other words, the correlation between two baits is used to generate the bait-bait matrix. This bait-bait matrix is clustered using Gene Cluster 3.0 [49] with correlation as the similarity metric and average linkage as the clustering method. The clustered matrix is visualized as a heat map, generated using Java Treeview [50].

Preparation of test data

Two data sets were used to illustrate the reconstruction of protein interaction networks using Plnt. The first data set (referred to here as **DS5**) was from “Nuclear import of histone deacetylase 5 by requisite nuclear localization signal phosphorylation”, by Greco et al., *Molecular & Cellular Proteomics* (2011) [51]. Only a fraction of the entire data was taken to represent a typical small scale data set. Briefly, two biological replicates of eGFP tagged HDAC5 protein, stably expressed in HEK293 cells, was affinity purified and analyzed on an LTQ Orbitrap XL mass spectrometer over a 90 minute gradient. Two negative controls (tag-only purifications) were also included.

The second data set (referred to here as **DS6**) was from “The functional interactome landscape of the human histone deacetylase family” by Joshi et al., *Molecular Systems Biology* (2013) [4].

This data set comprises of affinity purification experiments of eleven Histone deacetylases (HDACs) and represents a typical medium/large scale data set. Briefly, stable CEM-T cells were used to express eGFP tagged bait proteins. Affinity purification is carried out using anti bodies conjugated to magnetic (dyna) beads. Samples were analyzed on an LTQ-Orbitrap Velos mass spectrometer over a 90 minute gradient. At least two biological replicates were included for each bait purification. The data set also included six negative controls (tag-only purifications).

Each data set was processed separately using the X! Tandem/TPP/ABACUS pipeline as described earlier ('Preparation of test data (DS1)', Chapter 2), except for the following differences. The UniProt sequence database (*H. sapiens*) was used for searching MS/MS spectra and the precursor mass tolerance was specified as 100 p.p.m. The ABACUS output file was manually edited to generate the matrix formatted SPrInt input file (see 'Data formats', Chapter 2) and scored online.

Results and Discussion

Creation of PInt

The pipeline for reconstruction of interaction networks using PInt is shown in Figure 3-1 a. The graphical user interface for the pipeline is shown in Figure 3-1 b. First, interactions are scored using SPrInt (as described in chapter 2). A list of high confidence interactions is generated by filtering the output generated by SPrInt (step 1). The filtering options can be specified through the GUI (Figure 3-1 b, 1). In step 2, a subset of baits is selected (Figure 3-1 b, 2) to narrow down the analysis to a sub-network. The bait-bait cluster gram, displayed as a heat map, can be used to select closely related baits and hence the corresponding sub-network (see 'Using bait-bait cluster gram to identify sub-networks'). By default, all baits are selected. This is an optional step and is not relevant for small scale data sets. In step 3, a 'network context' is generated using prior knowledge (see 'Generating network context'). This is also an optional step and may not be necessary for medium/large scale data sets as detailed below. Currently, iRefIndex database is used to represent prior knowledge, but the system can easily include other databases. The parameters for generating the network context (Figure 3-1 b, 3) include 'seed nodes', 'include nodes', 'include edges', 'exclude nodes' and 'exclude edges'. 'Seed nodes' is used to generate a

network context using the ‘scaffolding’ approach (see ‘Generating network context’). ‘Include nodes’ is used to generate a network context using the ‘simple’ approach (see ‘Generating network context’). Users can specify nodes that need to be excluded from an analysis using the ‘exclude nodes’ option. This feature is useful when it is necessary to manually exclude certain proteins (such as ribosomal proteins) from an analysis. Similarly, a list of interactions can be (manually) appended to the prior knowledge generated from iRefIndex using the ‘include edges’ option. This feature is useful when iRefIndex does not document interactions that are relevant to an analysis. It is also possible that our parsing scripts may omit a few interactions documented in iRefIndex (see ‘Generating a database of interactions from literature’). In step 3, high confidence interactions generated from the user data are merged with the network generated from prior knowledge. Optionally, the resulting network can be pruned. The options for pruning the network (such as excluding nodes with a small degree and limiting the prior knowledge by the number of supported publications) are specified through the GUI (Figure 3-1 b, 4). Pruning allows narrowing down the network to its core-components. The output (interaction map) is displayed using Cytoscape web, an embedded plugin for viewing networks on a browser (Figure 3-1 b, 5).

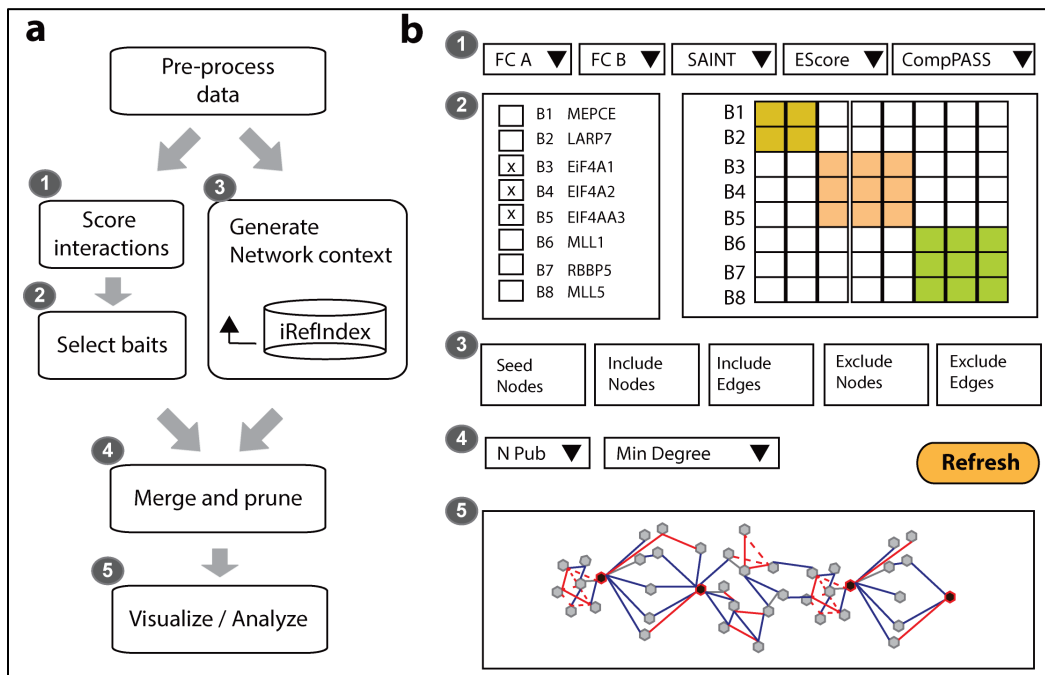


Figure 3-1: Plnt pipeline and the graphical user interface.

(a) PInt processing pipeline. (b) Graphical user interface. **Step 1:** Interactions are scored using SPrint (chapter 2) and filtered to identify high confidence interactions. The options for filtering are shown in (b-1). **Step 2:** Baits are (optionally) selected using the heat map shown in (b-2). Sub networks can be created by selecting closely related baits. **Step 3:** A network context is generated. The options for generating network context are shown in (b-3) and explained in the main text. **Step 4:** User generated network is merged with literature derived network (prior knowledge). **Step 5:** Networks are visualized online using Cytoscape web.

Analysis of small scale networks

Networks generated using small scale data sets present an incomplete (biological) picture. In the case of our small-scale test data set (DS5), the network generated from high confidence interactions in user data ($\text{SAINT} \geq 0.9$) was too simplistic to elucidate the mechanistic role of HDAC5 in the cell. This is a good example when integration with prior data significantly enhances the interpretability of user data. It turns out that reconstruction of protein interaction networks in higher mammals is a distributed enterprise, involving several small-scale analyses. Hence, there is a strong case for a tool that provides a framework for systematic integration of prior knowledge into a small scale analysis. A ‘simple’ network context (see ‘Generation of network context’) derived by specifying a list of relevant proteins (in this case, HDACs 1-11) is shown (Figure 3-2 b). The network context extracted using the scaffolding approach (see ‘Generation of network context’) generates a more comprehensive picture (Figure 3-2 c). The combined network generated by integrating high confidence interactions from user data with a network context generated using the scaffolding approach is shown (Figure 3-2 d). As expected, this combined network is more interpretable due to the inclusion of several important members of the HDAC complex, the RNA processing complex and the PP2A system (Figure 3-2 c).

It is important to note that the network context generated in step 3 is biased towards the input proteins provided by the user. Such a bias is not necessarily un-warranted. Small scale analyses are usually performed to explore a specific biological question, so it may be advantageous to interpret the collected data in the same context.

Analysis of medium/large scale networks

Baits in large/medium-scale networks are typically selected to explore the landscape of a certain biological phenomenon (e.g. Kinase signaling). Accordingly, a network context exists

implicitly for such networks and deriving one from prior knowledge (using ‘simple’ or ‘scaffolding’ approach) is redundant. Scale, however, introduces complexity into any network analysis. Hence, the ability to dissect the network and zoom into smaller sub-networks is of prime importance for the analysis of medium/large scale networks. Sub-networks can be derived based on the bait-bait similarity (see ‘Generation of network context’). In the case of our medium/large scale test data set (DS6), the full network (Figure 3-3 a) projects a relatively complex picture. The bait-bait similarity analysis suggests two groups in the data (Figure 3-3 b). These groups correspond to class I and class II HDAC networks respectively. The corresponding sub-networks (Figure 3-3 d, e) present a simpler picture compared to the full network (Figure 3-3 a). A pruned version of the full network was generated by retaining only those nodes with at least two connections (Figure 3-3 c). As expected, nodes in this pruned network are highly connected. An initial analysis suggests that important players in the landscape of HDAC interactome are highlighted by pruning the network. A more detailed analysis is needed to examine the effects of pruning from a biological point-of-view, which is beyond the scope of current work.

Contributions

This work is the result of collaboration between the Nesvizhskii Lab (Univ. of Michigan, Ann Arbor) and the Gingras Lab (Univ. of Toronto, CA). Bioinformatics pipelines/methods were developed by Dattatreya Mellacheruvu, under the guidance of Dr. Alexey Nesvizhskii. Dattatreya Mellacheruvu and Zachary Charles Wright (Application programmer (sr), Univ. of Michigan, Ann Arbor) implemented the system.

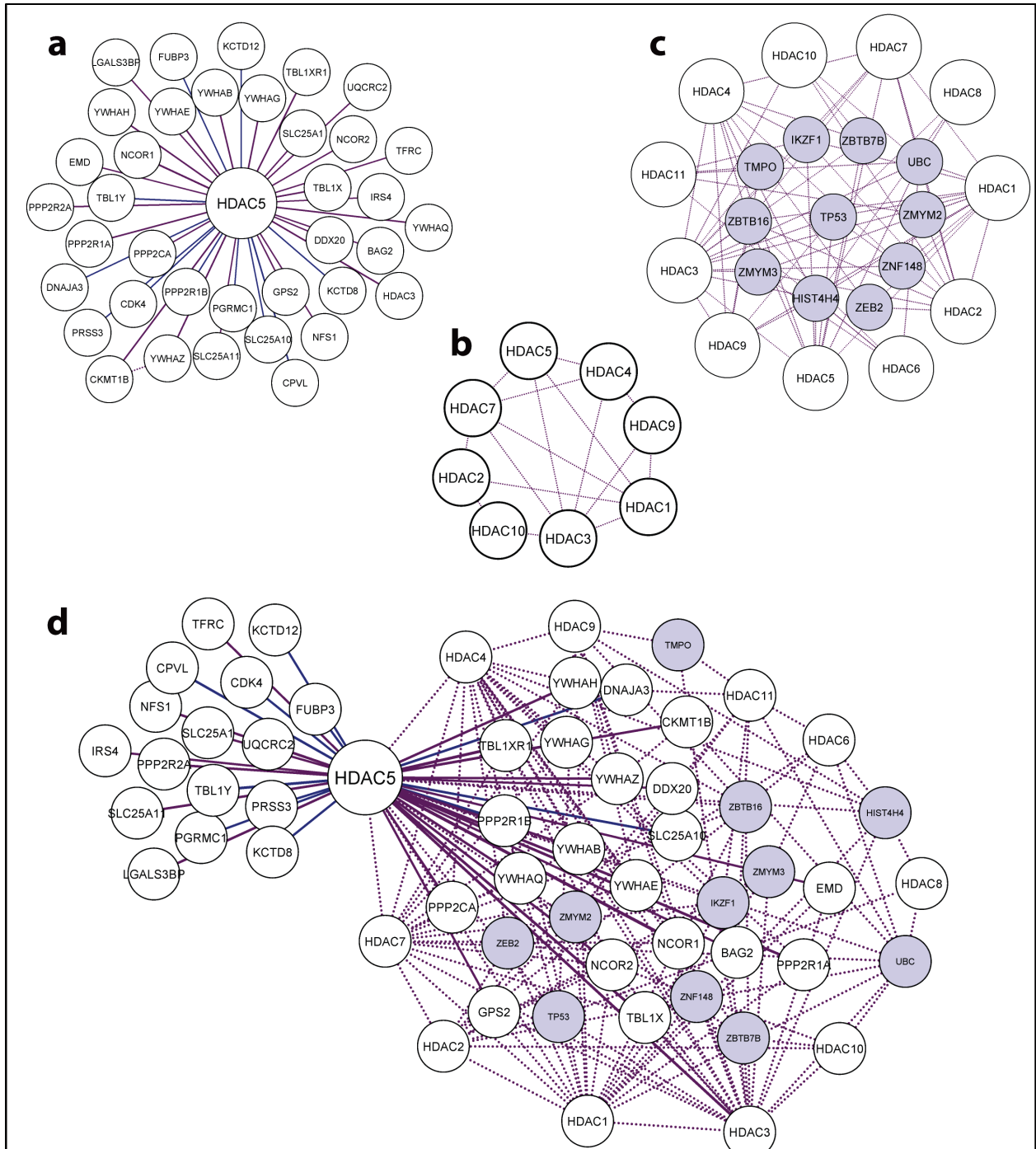


Figure 3-2: Analysis of a small scale network using Plnt.

(a) Interaction network of HDAC5 generated from a small scale data set (DS5). (b) Network context derived using 'simple' approach (Methods). HDACs 1-11 were provided as the input. (c) Network context derived using the 'scaffolding' approach (Methods). HDACs 1-11 were specified as the 'seed' nodes (Methods). Scaffolding approach generates a more comprehensive picture by introducing new nodes (shown in blue). (d) Combined network generated from user data and prior knowledge.

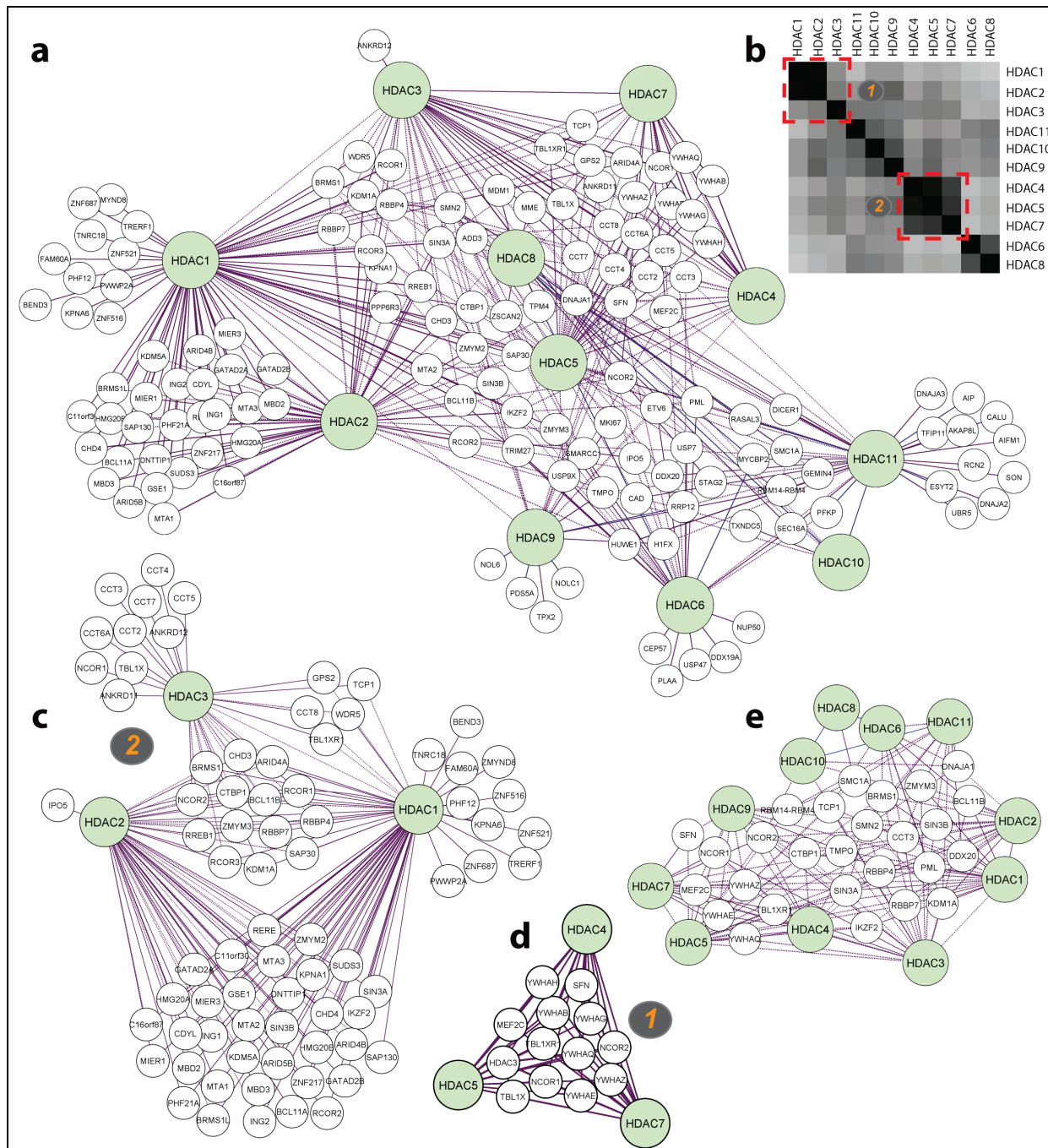


Figure 3-3: Analysis of a medium/large scale network using PInt.

(a) Network generated using high confidence interactions (SAINT ≥ 0.9) from a medium scale data set (DS6) comprising eleven baits (HDACs 1-11). (b) Bait-bait similarity assessed using cluster grams. The data is visualized as a heat map. Two clusters are clearly visible. (c, d) Sub-networks corresponding to clusters 1 and 2 shown in b). (e) A pruned network created using the full data set, but by excluding nodes with degree < 2 . Pruning highlights critical nodes in the network (Methods).

CHAPTER 4

CRAPome: The Contaminant Repository for Affinity Purification Mass Spectrometry Data

Contents of this chapter have been published by the author in Nature Methods [1].

Introduction

AP-MS has become a widely used approach for the identification of protein-protein interactions [22]. In most cases, however, a large number of nonspecific interactors (here referred to as 'background contaminants', or 'contaminants') are co-purified with bait proteins and identified by mass spectrometry. Methods to discern bona fide interacting partners from background contaminants are thus essential. In the case of affinity purification using epitope-tagged proteins, this distinction is often aided by the inclusion of negative-control purifications, typically consisting of one or more mock purifications using the same support resin and cell line but without expression of the polypeptide(s) of interest ('bait' proteins). These controls (when isotope labeling [52-55] is not used) can be considered universal, meaning that they are useful for filtering the background from any bait protein subjected to the same purification scheme [9, 42, 53, 56-58].

A question arises when designing and performing AP-MS experiments as to how to use previous knowledge regarding background contaminants to best score interaction data. Small variations in the sample or sample preparation may influence the recovery of proteins, including contaminants. It is therefore not uncommon for a negative-control experiment to fail to capture a complete set of contaminants owing to undetected variations at one or more experimental steps. This issue is compounded by the fact that low-abundance peptides (and hence proteins) may not be reliably detected in a given mass spectrometry analysis. Analyzing

one or a few negative-control samples will thus generally not allow for a comprehensive characterization of background contaminants for a given purification regime.

Here we present the CRAPome, a web-accessible resource that stores and annotates negative controls generated by the proteomics research community and enables their use for scoring AP-MS data. Users employ an intuitive graphical user interface to explore the database by either querying one protein at a time or downloading background contaminant lists for selected experimental conditions. The repository is tightly integrated with SPrint (chapter 2), a versatile tool for scoring protein interactions. Users can select relevant CRAPome controls to augment their own negative controls for scoring interactions. We also describe here, the database structure and composition and provide examples of the use of this resource to filter contaminants with properly chosen controls. The CRAPome accommodates a variety of purification schemes. Though it currently contains only *H. sapiens*, *S. cerevisiae* and *E. coli* data, it will be expanded to include other species.

Methods

Design and architecture of the CRAPome repository

The CRAPome interface was developed using Drupal, an open-source PHP-based web framework, and MySQL and SQLite databases. The processing pipeline for adding data to the database and querying/extracting data from the database was developed using Python and SQLite. The actual data for each experiment ('data'; Figure 4-1), such as the protein/gene accession numbers, the sequences of the identified peptides, peptide probabilities and the spectral counts, are stored in a SQLite database. The attributes used to annotate the experimental conditions (metadata) are stored in a separate MySQL database. The separation of data and metadata is performed for the convenience of developing the web interface, which allows annotation of experiments (management of metadata) directly by data contributors, whereas the processing and management of the data themselves is performed by the database administrator. Further details of the software design are provided in Appendix A.

In order to keep the annotation of data consistent, the attributes and values that describe the experimental conditions are predefined. The corpus of these attributes (and their values) is referred to as the 'controlled vocabularies', or CVs (Table 4-1). In addition to the CVs, each experiment deposited in the CRAPome repository is also annotated with a detailed description of the experimental protocol that enables users to obtain additional details about the experiments.

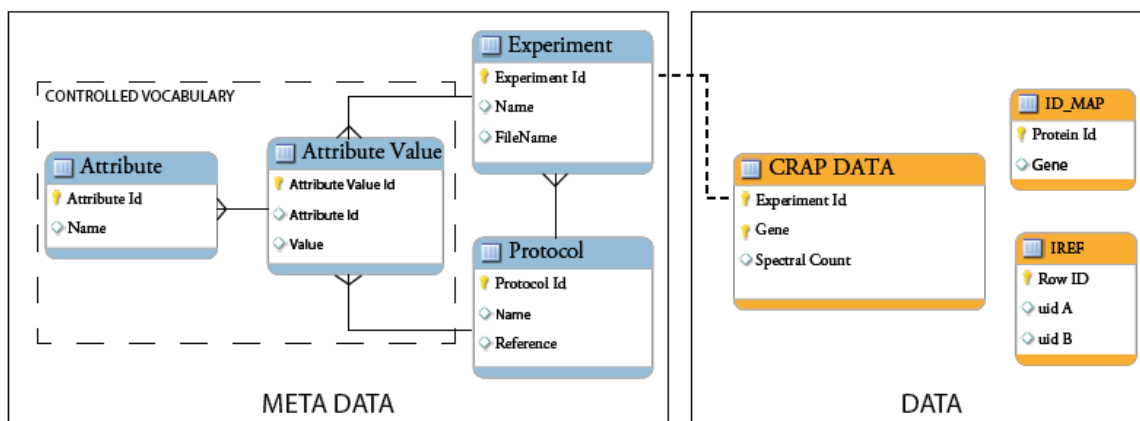


Figure 4-1: Schema of the CRAPome database.

Meta data is stored in MySQL database, where as the data is stored in SQLite database. The separation allows streamed processing and annotation of experiments (main text).

Processing of mass spectrometry data and population of the CRAPome database

Data sets were obtained from the contributing laboratories in the .raw or .mgf file formats. The files were converted to the open mzXML file format and further processed using the X! Tandem/Trans-Proteomic Pipeline (TPP) suite of tools [14, 15, 34]. For the initial release (V 1.0), MS/MS spectra were searched against RefSeq protein sequence database version 47 (ref. [35]; H. sapiens) or SGD ORF protein sequence database orf_trans.20100105.fasta (S. cerevisiae), appended with an equal number of decoy sequences, using X! Tandem[13] with k-score plug-in. For the purposes of simplicity and uniformity, we developed two standard parameter templates for processing using X! Tandem and TPP, which were applied to data generated on low- or high-mass accuracy instruments, respectively. MS/MS spectra were searched using a precursor-ion mass tolerance of 100 p.p.m. (monoisotopic mass) or using -1 to +4 Da (average mass) windows for high- and low-mass accuracy instruments, respectively. All other database

search parameters were identical: cysteine carbamylation (C + 57.0215) and methionine oxidation (M + 15.9949) were specified as variable modifications. The search results were processed using PeptideProphet (high-mass accuracy data were analyzed using the high-mass accuracy binning option) and then further processed using ProteinProphet to create protein summary files. For each experiment, all contributing data (multiple gel-band fractions, technical replicates, etc.) were combined to generate a single set of PeptideProphet and ProteinProphet output files (pepXML and protXML files, respectively).

One of the submitted data sets [59] consisted of a very large number (300) of negative controls in which proteins were separated using 1D SDS-PAGE. In a fraction of these experiments, only selected bands were analyzed using MS. To avoid the problem of data inconsistency due to missing MS data for a subset of gel fractions, and to reduce the total number of entries in the CRAPome representing this data set, we combined the individual experiments from this data set to generate ten composite experiments (protocol no. 66; experiments CC185–CC194).

To build the CRAPome database, we extracted spectral counts from protXML files using an in-house software tool. For each protein in the protXML file, peptide-to-spectrum matches with a probability ≥ 0.9 were extracted. The cumulative sum of the spectral assignments for these peptides constituted the spectral count for the corresponding protein. The spectral count was computed for each protein in the output file regardless of whether peptides mapping to a given protein could also map to other proteins. We note that this represents a deviation from the conventional approach of performing stringent false discovery rate (FDR) filtering and removing redundant or inconclusive, i.e., not supported by unique peptides, protein identifications [36]. (The results of such stringent filtering are described below; see ‘Global analysis and reduced gene counts.’) The liberal approach for creating protein summaries for each experiment taken here in fact enables a conservative approach for scoring protein interactions. As discussed in ref. [57], it ensures that the spectral counts of proteins from homologous families such as keratins, tubulins and actins are not underestimated owing to the ambiguities related to the identification of shared peptides. Finally, RefSeq protein accession numbers were mapped to official gene identifiers using Ensembl BioMart [60] tools and were displayed as corresponding

gene symbols (entries with NP accession numbers only; proteins with XP numbers and those with NP accession numbers that cannot be mapped to gene symbols are presently not visible in the database). When multiple proteins mapped to the same gene entry, the maximum spectral count among these proteins was selected as the spectral count for that gene. These data provided the basis of the CRAPome accessible online and were used to calculate redundant gene counts shown in Table 4-2.

Quality control

As part of the process of creating the database, the CRAPome administrator performs a quality-control check of the database search results. Experiments containing only a few identifications (fewer than 10 gene symbols with nonzero counts) are removed automatically, and experiments with 10–50 gene symbols are inspected in more detail. Furthermore, all negative-control experiments generated using the same protocol are inspected for consistency, and inconsistent samples are removed. Last, possible carryover issues are identified and referred to the data depositors for further inspection. From the 402 experiments submitted to the CRAPome, 42 experiments were excluded on the basis of these quality-control steps.

Integrated scoring tool (SPrint)

The pipeline for scoring interactions (SPrint) and CRAPome are built on a common software platform and are tightly integrated. SPrint is described in chapter 2. Relevant CRAPome controls, selected using CVs, are passed as an input to SPrint. Data is extracted from the CRAPome database and used for scoring interactions.

Global analysis of CRAPome and reduced gene counts

To allow a more informative analysis of the contaminant profiles and comparison with other data, we processed all pepXML and protXML files generated as described above using a more conventional set of filtering thresholds. All pepXML files used to generate the CRAPome repository (human data subset in version 1.0, 343 files) were processed together using ProteinProphet to generate a single protein summary file (protXML file). This combined protXML file, as well as the pepXML and protXML files for each individual experiment, were then processed using ABACUS [37] to generate a combined spectral count matrix using default

parameters (accepting proteins with at least one peptide having PeptideProphet probability of 0.99 or greater and protein probability as computed by ProteinProphet of 0.9 or greater). Each row in the filtered ABACUS file represented a protein group from the combined protXML file, with a single accession number selected among indistinguishable protein entries forming that group. Spectral counts for the representative proteins were extracted from pepXML files for each individual experiment. The FDR for the combined protein list was less than 1% as estimated using decoy counts. The resulting spectral count matrix was used to compute similarity scores to generate the cluster gram (see below) and to analyze the global properties of the data such as frequency of identification across the entire data set (Table 4-2, 'reduced gene count').

Gene Ontology (GO) enrichment analysis was performed on the reduced list, and only the top 25% most abundant proteins in each experiment were considered (1,427 genes in total). The analysis was done using the online DAVID tool [61], with the analysis restricted to level 3 biological process (BP), molecular function (MF), or cellular component (CC).

To generate the cluster gram (Figure 4-4), we first computed experiment-experiment similarity scores using the cosine function from square root-transformed spectral counts (data from protocol no. 66 (ref. [59]) were excluded from this analysis; see above). For computing the final cluster gram, we required that each experiment had at least two additional experiments with a similarity score of 0.7 or higher. The final cluster gram was generated using Cluster 3.0 software [49], with single-linkage clustering using Pearson correlation (un-centered) as the similarity measure. The cluster gram was visualized using Tree View software [50].

Contaminant propensity as a function of protein abundance

To generate the list of proteins and protein abundances in the HEK293 whole-cell lysate, we used publicly available data taken from ref. [62]. Raw mass spectrometry data for this cell line were downloaded from the original publication and processed as described above (see “Global analysis and reduced gene counts”). For each identified protein (representative protein per group; see above) in the filtered ABACUS file, the summed spectral count across the four biological replicates was taken as a measure of the protein abundance in the cell line. A global

histogram of protein abundances was then generated by binning (Figure 4-4 a). The background contaminant propensity was then calculated as a fraction of HEK293 cell line-identified proteins in each spectral count bin that were also detected in at least one HEK293 experiment in the CRAPome. For this comparison, we selected CRAPome (V 1.0) experiments having the 'Cell Line' attribute value 'HEK293' only and queried protein accession numbers identified in the HEK293 whole-cell lysate against the CRAPome HEK293 identified proteins. We then plotted the 'fraction in CRAPome' as a function of protein abundance (binned spectral counts).

Preparation of test data sets

The utility of CRAPome database is illustrated using a typical small scale data (DS1, described in chapter 2). Briefly, the data set comprises two biological replicates of each of the following four baits: MEPCE, RAF1, WASL and RAF1. Six negative controls, i.e., tag-only purifications, were also included. The data was processed as described earlier (see "Preparation of test data", chapter 2).

Access to CRAPome

The CRAPome can be accessed at <http://www.crapome.org/>. No registration is required to access the database. However, registered users can save selected lists of controls (Figure 4-2 c). The database and experimental protocols can be downloaded as text files from the website (<http://crapome.org/?q=Download>).

Results and Discussion

Creation of the CRAPome Repository

The CRAPome database is a web-accessible (<http://www.crapome.org/>) repository of negative control AP-MS experiments (both published [33, 42, 51, 56, 58, 59, 63-76] and unpublished) associated with detailed protocols and controlled vocabularies (CVs; Table 4-1) used to organize the data. Data contributors first submit raw mass spectrometry files (Figure 4-2 a), which are processed using a uniform data analysis pipeline and by several quality-control checks (Methods) before the association of metadata (CVs and text-based protocols). These annotated negative-control runs form the core of the repository. The initial version (V 1.0, March 2013)

comprised 360 experiments contributed by 12 laboratories, of which the bulk of the data (343 experiments) were generated using human cell lines. This large data set covers many of the most commonly used AP-MS protocols. For each experiment, mapping of the protein identifiers to the HUGO gene nomenclature committee (HGNC) gene symbol is performed, and spectral count information is parsed to the relational database (Methods). The database is expandable, and new data are added to the CRAPome using the same deposition and annotation process. New protocols and CVs will adapt the database to new experimental workflows.

Attribute Name	Attribute Values
Organism	human
Cell/tissue type	HEK293, HeLa, U2OS, PBMC, Jurkat, CEM-T, MRC-5, LS174
Cell/tissue subtype	none, HEK293T, HEK293 Flp-In T-REx, Jurkat-Flp-In
Drug treatment	aphidicolin, rapamycin, nocodazole, MG132, none, IFN-beta, DMSO, okadaic acid, doxycycline+thymidine, tetracycline+thymidine, thymidine+nocodazole
Subcellular fractionation	total cell lysate, total lysate+chromatin, nuclear fraction, cytosolic fraction
Epitope tag	FLAG, HA, GFP, TAP, HaloTag, Strep-HA
Control protein	RFP, GFP, FLAG, mCherry, tag alone, untransfected, uninduced, NLS-RFP
AP steps	single, tandem
Affinity approach 1	M2 anti-FLAG, anti-GFP camel, anti-GFP rabbit, HA-7 anti-HA, HaloLink, IgG, Streptactin, 2xFLAG, SBP, anti-GFP mouse
Affinity support 1	agarose, magnetic (dynabead), magnetic (agarose coated), nano-magnetic, microMACS
Affinity approach 2	none, M2 anti-FLAG, anti-GFP camel, anti-GFP rabbit, calmodulin, HA, 2xHA, HA-7 anti-HA, anti-GFP mouse
Affinity support 2	none, agarose, magnetic bead (dynabead), magnetic beads, agarose coated, nano-magnetic beads, microMACS
Fractionation	SDS-PAGE, 1D LC-MS, MudPIT, RP-RP, GeLC
Instrument type	Velos-Orbitrap, LTQ-Orbitrap, LTQ, LCQ, LTQ-FT, 5600 TripleTOF

Table 4-1: Controlled vocabulary for annotating experiments (V 1.0).

Graphical user Interface

End users access the database via a web interface (Figure 4-2 c, d). After selecting the organism of interest (currently *H. sapiens*, *S. cerevisiae* or *E. coli*), the database can be queried in two ways (called “user workflows”).

1. Query selected proteins. In workflow 1, users submit queries consisting of protein or gene identifiers and retrieve summaries of the occurrence of queried entries. An expanded view summarizes the conditions and protocols in which the protein has been identified, associated with spectral count information (Figure 4-2 c).

2. Create contaminant lists. Workflow 2 generates background lists from a subset of the CRAPome controls. In this case, the user simply selects the list of desired controls (filtered using CVs and protocol details; Figure 4-2 d) and downloads the resulting tables of contaminants. Quantitative parameters, including the occurrence of identification across selected controls and the average spectral counts across selected controls in which the protein was detected, are included (a maximum of 30 experiments can be viewed online; the entire data set can be downloaded as a tab-delimited file from <http://www.crapome.org/>). Registered users can also save the selected list of controls for future use.

Using CRAPome to score interactions

We tested whether the controls deposited in the CRAPome could be used for scoring interactions in the absence of user controls. Although we recommend always using at least some user controls for scoring interactions, there are certainly cases in which such controls do not appropriately model the background. Controls from the repository were thus selected on the basis of the CVs and protocols. We identified two relevant control sets from two different laboratories that fulfilled our criteria (HEK293 cells, Flag tag, single-step purification on M2 agarose) which contained 10 (set 1; CRAPome protocol no. 56) and 11 (set 2; CRAPome protocol no. 26) experiments, respectively. Using ROC analysis, we showed that each of these sets of controls performed very similarly to the user controls in both SAINT (Figure 4-3 a) and FC (Figure 4-3 b) calculations.

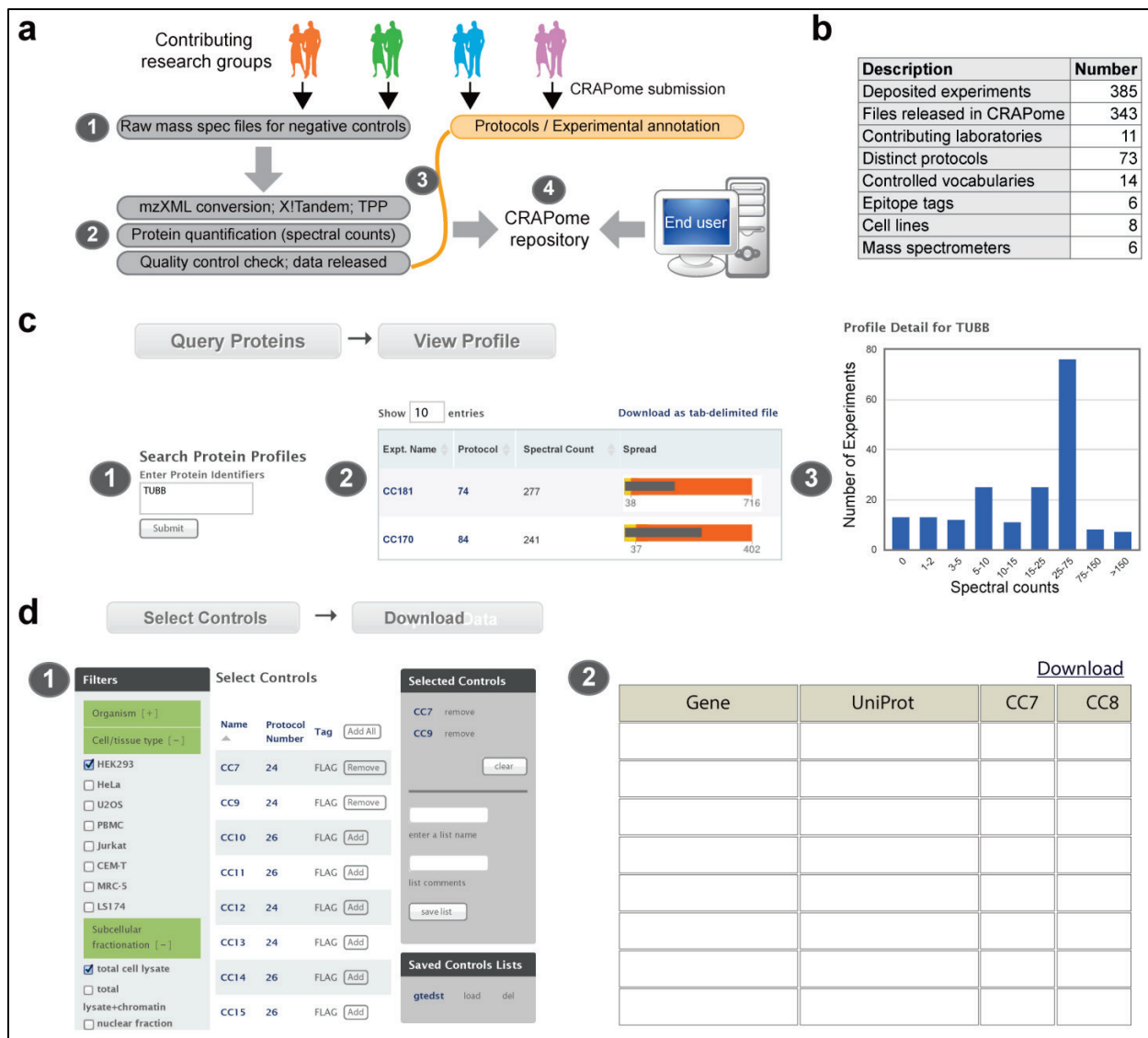


Figure 4-2: Creation of CRAPome database and the graphical user interface.

(a) Creation of the CRAPome. (1) Contributors to the CRAPome submit raw mass spectrometry files for negative-control runs as well as detailed experimental protocols and mapping information. (2) Raw mass spectrometry files are first converted to mzXML and analyzed by X! Tandem and the Trans-Proteomic Pipeline (TPP); counts are extracted for protein quantification, and the CRAPome administrator performs a quality-control check (Methods). (3) Released high-quality runs (data) are associated with experimental descriptions and protocols (metadata) by the CRAPome administrator in consultation with the data provider. (4) The CRAPome database is queried by external users via the web interface. (b) Overview of the data in CRAPome (V 1.0). (c) Overview of the CRAPome workflow 1. (1) Proteins are queried against the CRAPome by inputting one of several identifiers, which are mapped to corresponding gene symbols. Different views enable exploration of the contaminant profile of each queried protein, either as a summary table (2) or in graphical formats (3). (d) Overview of the CRAPome workflow 2 that is used to generate lists of contaminant proteins. Desired controls are selected, with the help of CVs. Data are displayed in a tabular format and can be downloaded as a text file.

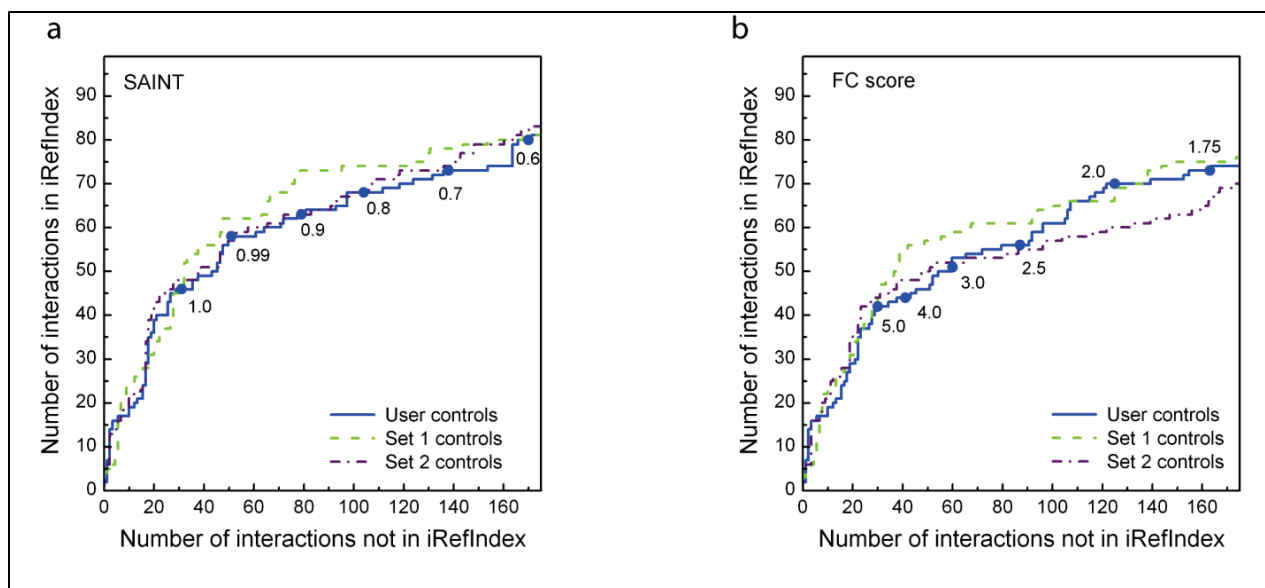


Figure 4-3: Scoring protein interactions using controls from the CRAPome (V 1.0) with SAINT.

Scoring protein interactions using controls from the CRAPome with SAINT (a) and FC-A (b): user controls ($n = 6$) are compared to two sets of controls from the CRAPome, selected according to the CVs (set 1 = 10 controls; set 2 = 11 controls).

Characterization of CRAPome

We mined the database to determine (i) which proteins have a higher propensity to be contaminants and (ii) how background proteins differ according to experimental conditions. First, to understand whether the abundance level of a protein in a sample increases the propensity of the protein to be a contaminant, we plotted the proteins reported in the CRAPome repository (restricting the analysis to HEK293 cells, by far the most common human cell line in the CRAPome) against a list of proteins ranked by their abundance estimates on the basis of whole-proteome analysis of HEK293 cell lysate [62]. We observed a clear relationship between the abundance of a protein in HEK293 and its detection in at least one of the HEK293 experiments in the CRAPome database (Figure 4-4 a). We next analyzed the frequency of detection of individual proteins in the CRAPome (mapped to gene names, as throughout this manuscript). Using stringent filtering (protein false discovery rate <1%), 4,449 non redundant protein groups (or 7,782 gene names without compression of the data; Methods) were identified. Of these, 14 proteins were detected in >90% of all experiments, and 89 were detected in >50% of the experiments: percentages that qualified these proteins as ubiquitous

contaminants (Table 4-2). These include keratins, cytoskeletal proteins such as tubulins and actins, and high-abundance proteins including translation elongation factors and histones (Table 4-3). Other proteins were not detected consistently across all purifications but were abundant (in terms of total spectral counts) across the database: these were notably enriched for several functional categories, predominantly those associated with RNA functions. However, a large fraction of the proteins present in the CRAPome were detected in only a small fraction of experiments: 3,571, or 80% of the proteins in the CRAPome, were found in $\leq 10\%$ of the experiments.

Frequency in CRAPome (%)	Redundant gene counts	Reduced gene counts
>90	15	14
>75	37	30
>50	110	89
>20	504	463
>10	898	878
≤ 10	6,884	3,571
Total	7,782	4,449

Table 4-2: Frequency of detection across CRAPome database (V 1.0).

Data are for *H. sapiens*. The two counts are computed at different frequencies. (i) “Redundant” gene counts are based on a generous estimation of shared peptides: in this case, each protein or gene to which a given peptide is matched is counted as a contaminant. (ii) “Reduced” gene counts are based on a more stringent definition of protein/gene parsimony, as described in Methods.

To further explore the contaminant propensity of the proteins in the CRAPome, we computed the similarity of all experiments (restricting the analysis to human data only), generating a heat map (Figure 4-4b and Methods). The data clustered primarily according to experimental conditions (though there was a bias in the type of background detected across different laboratories). Several of the clusters could be further separated into sub-clusters, as exemplified by the “Flag HeLa agarose” cluster, which showed a clear separation based on subcellular fractionation (cytoplasmic or nuclear) performed before AP-MS (Figure 4-4 c). Using our analysis of the most important determinants of background behavior as a basis, we annotated all experiments in the CRAPome (V 1.0) using 14 categories of CVs (Table 4-1), which can be used to select experiments that are most similar to those in a query set. Complete

protocol descriptions of the experiments are also provided by selecting the desired protocol number.

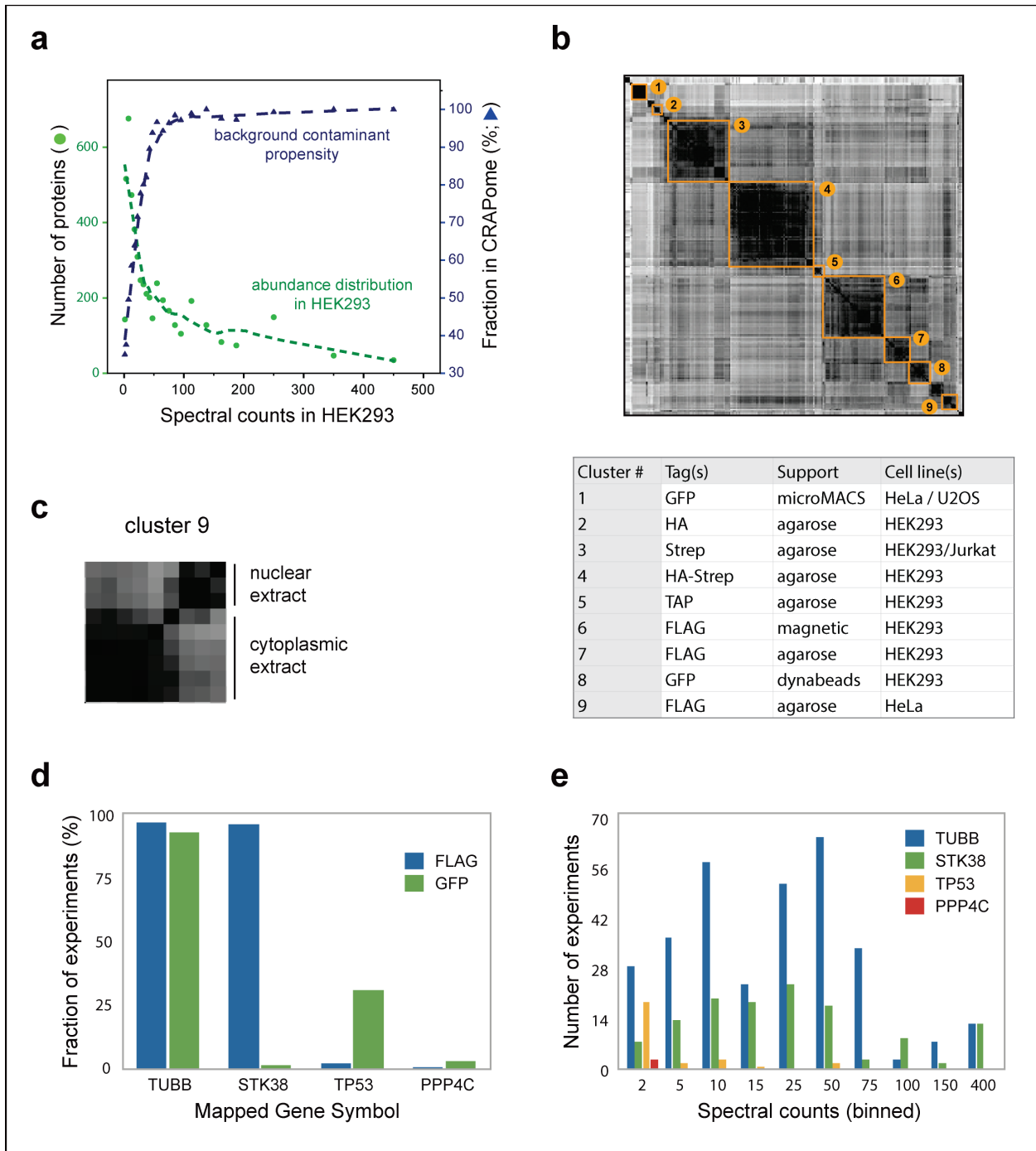


Figure 4-4: Characterization of CRAPome (V 1.0).

Relationship between the detection of a given protein in the CRAPome and its protein abundance. The abundance distribution in HEK293 cells was calculated from shotgun mass spectrometry data (Methods). The left axis indicates

the number of proteins identified at each of the spectral count abundances (green circles; green dashed line shows fit to data); the right axis indicates the fraction of the proteins at a given binned abundance in the CRAPome database (blue triangles). (b) Similarity clusters of all experiments. All experiments in the CRAPome were scored for similarity in their contaminant profiles according to a cosine function: the size of each cluster represents the number of experiments with strong similarity. Selected similarity clusters are indicated alongside their composition. (c) Cluster 9, described in b as Flag-tagged agarose in HeLa cells, can be further defined as two sub-clusters on the basis of subcellular fractionation performed before the affinity purification (cytoplasmic and nuclear fractions); other clusters can also be refined. (d) Example of epitope-tag specificity for selected proteins and genes. (e) Spectral count distribution of the proteins in d across the entire data set. Spectral count bins are shown for all nonzero experiments. The highest spectral count boundary for each bin is shown.

To illustrate the different contaminant propensities of individual proteins, and the need to account for not only the overall frequency of detection in the data set but also the experimental conditions, we analyzed the frequency distribution of four proteins with two types of epitope tags, Flag and GFP (Figure 4-4 d). Tubulin- β (TUBB) was detected across nearly all of the experiments, irrespective of the epitope tag. By contrast, the serine/threonine kinase STK38 co-purified in nearly all Flag experiments but not in GFP experiments, whereas the tumor suppressor protein p53 (TP53) was detected predominantly in GFP-based affinity purification protocols. The serine/threonine phosphatase PPP4C was not detected at a high frequency in experiments performed with either of these epitope tags (it was identified in 3 of 343 experiments across the entire database). Frequency and experimental conditions are also clearly insufficient to describe contaminant propensity: abundance measures are also critical. For instance, if a protein is detected at a high frequency but low abundance (that is, a low number of spectral counts in a high number of mass spectrometry runs) in the CRAPome but is detected with a high spectral count in bait purifications performed by a user, it is more likely to be a true interactor than if it were always detected with high abundance in the CRAPome. To illustrate this concept, we compared the nonzero values in Figure 4-4 d for the four proteins, but we specifically examined spectral count distributions (binned values). This analysis revealed that whereas TUBB and STK38 were often present in very high counts in the CRAPome, TP53 was usually detected with much lower spectral counts (Figure 4-4 e). Such comparisons can be easily accessed via the CRAPome user interface (workflow 1). They also provide the basis for statistical or empirical scoring of interactions described in SPInt (chapter 2).

Gene family	Example gene symbols
Heat-shock proteins	<i>HSPA1A, HSPA8, HSPA2</i>
Keratins	<i>KRT1, KRT10, KRT2</i>
Tubulins	<i>TUBA1B, TUBA3C, TUBB</i>
Actins	<i>ACTB, ACTA2, ACTBL2</i>
Elongation factors	<i>EEF1A, EEF1A2</i>
Histones	<i>HIST1H1C, H2AFX, HIST2H2BE</i>
Ribonucleoproteins	<i>HNRNPK, HNRNPU, HNRNPH1</i>
Ribosomal proteins	<i>RPS3, RPS18, RPL23</i>

Table 4-3: Most Frequently detected protein families across the CRAPome (V 1.0).

Shown are the most frequently detected protein families, alongside some of the most frequently detected representative genes (*H. sapiens* data, V 1.0).

Concluding Remarks

Although lists of contaminating proteins have been reported in the past [53, 77, 78] there has been no central repository for this type of data or freely available software tools for using these lists. The CRAPome facilitates access to a standardized (in terms of protein identification pipeline, ID mapping, abundance measures and so on) set of negative-control experiments, organized via CVs based on experimental considerations. The freely accessible user interface is intuitive and informative, even for those who may be new to mass spectrometry.

We are currently using spectral counts as the sole quantification tool in the repository, but extension of the system to other types of quantification (especially that based on peptide ion intensity, which is becoming possible as high-mass resolution instruments are increasingly being used for AP-MS experiments) may help to further discriminate between background contaminants and true interactors. We expect a constant stream of negative-control data to be deposited in the CRAPome. As contributors continue depositing their data in the repository, robustness in scoring will increase, and in-depth characterization of contaminant behavior will be possible. The CRAPome can be used as a retrospective tool to analyze AP-MS data, and it will be instrumental to curators of protein-protein interaction databases. It should also assist with establishing guidelines regarding the scoring and annotation of such data. Widespread adoption of the CRAPome (by experimentalists, computational biologists, database curators and

reviewers alike) will improve the overall quality of AP-MS protein interaction data, addressing one of the key challenges in proteomics research.

Contributions

This work is the result of collaboration between the Nesvizhskii Lab (Univ. of Michigan, Ann Arbor) and the Gingras Lab (Univ. of Toronto, CA). Bioinformatics pipelines/methods were developed by Dattatreya Mellacheruvu, under the guidance of Dr. Alexey Nesvizhskii. Dattatreya Mellacheruvu and Zachary Charles Wright (Application programmer (sr), Univ. of Michigan, Ann Arbor) implemented the system.

CHAPTER 5

RePrint: A Repository of Protein Interactions Generated from Affinity Purification Mass Spectrometry Data

Introduction

Interaction databases are central for creating comprehensive protein interaction networks. These databases typically store lists of interactions that are aggregated from various sources, including those derived from literature and processed high-throughput data sets. While some databases like HPRD [79] solely depend on manual curation methodologies, others like DIP [80] employ automated procedures such as text mining. Meta databases such as iRefIndex [27, 81] and IntAct [81] aggregate data from several other databases and facilitate the consolidation of available information. In addition to providing scores that indicate the confidence of an interaction, these databases also maintain the data provenance. Standardized data formats have facilitated easy exchange of information between databases.

In spite of serious efforts to develop and maintain well curated protein interaction databases, high false positive rates have been reported among several of them. On the other hand, there is also an ever increasing demand to expand the coverage of interactome captured by these databases, notwithstanding several limiting factors such as under sampling of the interactome itself. While the false positive rates are generally low in manually curated databases, they are labor-intensive and hence limited in their scope. Computational curation methodologies help in expanding the scope of a database, but they also inflate error rates due to inherent algorithmic limitations. Prediction based approaches are limited by the availability of training data and accuracy of prediction models.

Contemporary interactome analyses are increasingly being performed using high-throughput approaches. In particular, AP-MS has emerged as an efficient, sensitive and high throughput

approach to survey protein interactions[82]. With several AP-MS analyses being performed regularly, a data-driven strategy can now be adapted for creating protein interaction databases. Such a strategy circumvents the need for manual/computer curation and associated problems. We present here RePrint, a **repository of protein interactions** created using systematically aggregated spectral data from several AP-MS studies. All data sets are systematically annotated using standardized controlled vocabularies and processed uniformly to identify and quantitate proteins in the sample. Interactions are then scored using SPrint (chapter 2) and saved in a database. The compendium of such scored interactions forms the core of RePrint, which is publicly accessible through a graphical user interface.

RePrint implements a new score (RScore) to 'merge' evidence from multiple data sets, when available. It also provides a novel pipeline to create comprehensive interaction networks. Post publication, the repository and associated tools will be publicly available at www.reprint-apms.org.

Methods

Design and architecture of RePrint

RePrint was implemented by leveraging the infrastructure developed for CRAPome (chapter 4) and SPrint (chapter 2). Briefly, the user interface was built using Drupal, MySQL and SQLite. The pipeline to populate the RePrint database was created using Python and SQLite. The workflow for creating interaction maps using RePrint data was created using the 'networkx' library of Python. The user interface is deployed using Apache, an open source web server, on virtual servers managed by the Medical School Information Services of the University of Michigan (MSIS). Details of the software design and implementation are presented in Appendix A.

Processing of mass spectrometry data and population of RePrint database

Each data set considered for RePrint is processed separately as follows. The raw files are converted to the open mzXML file format and further processed using the X! Tandem/TPP/ABACUS suite of tools [14, 15, 34]. *H. sapiens* data is searched against RefSeq protein sequence database version 56 (ref. [35]) appended with an equal number of decoy sequences,

using X! Tandem [13] with k-score plug-in. For the purposes of simplicity and uniformity, two standard parameter templates are used for processing data using X! Tandem and TPP depending on the mass accuracy of the instrument (low or high). MS/MS spectra are searched using a precursor-ion mass tolerance of 100 p.p.m. (monoisotopic mass) or using -1 to $+4$ Da (average mass) window for high- and low- mass accuracy instruments, respectively. All other database search parameters are kept identical: cysteine carbamylation (C + 57.0215) and methionine oxidation (M + 15.9949) are specified as variable modifications. The search results are processed using PeptideProphet (high-mass accuracy data are analyzed using the high-mass accuracy binning option) and then further processed using ProteinProphet to create protein summary files. All pepXML files in the data set are processed together using ProteinProphet to generate a single protein summary file (protXML file). This combined protXML file, as well as the pepXML and protXML files for each individual experiment, are then processed using ABACUS [37] to generate a combined spectral count matrix using default parameters (accepting proteins with at least one peptide having PeptideProphet probability of 0.99 or greater and protein probability as computed by ProteinProphet of 0.9 or greater). Each row in the filtered ABACUS file represents a protein group from the combined protXML file, with a single accession number selected among indistinguishable protein entries forming that group. Spectral counts for the representative proteins are extracted from pepXML files for each individual experiment. It is ensured that the FDR for the combined protein list is low (less than 5%) as estimated using decoy counts. The ABACUS output file is then (manually) edited to generate the corresponding SPrInt input file in the matrix format (Methods, Chapter 2). This input file generated for each data set is uploaded to SPrInt and interactions are scored online. The results for each data set are manually inspected by an expert analyst. Standardized metrics for quality control are still under development. When a data set produces reasonable results, it is annotated using the admin interface (Appendix B), assigned a data set ID and marked for inclusion in RePrInt.

An in-house Python script is used to aggregate scored interactions of all data sets marked for RePrInt. The (SPrInt) results file for each data set is parsed and a master list of interactions is generated. All necessary information such as the confidence scores, spectral abundance values,

and meta-data are also included for each interaction. RePrint assigns a unique ID for each interaction and references both bait and prey proteins by their corresponding gene names. The mappings from protein ID to its gene name(s) are downloaded from Ensembl BioMart [60]. When multiple gene names map to either the prey or the bait protein, the interaction is duplicated to generate a redundant list. All data is finally stored in a SQLite database. The bait and prey columns are indexed to enable faster queries on the database.

Interaction scoring: RePrint score

RePrint score (RScore) is defined as the probability that the observed interaction is a true interaction. When multiple data sets profile the same interaction, each can be considered as an independent evidence for the interaction. Accordingly, the overall probability that the observed interaction is true is computed using the following formula.

$$RScore_i = 1 - \prod_{i=1}^n (1 - p_i)$$

where p_i is the probability that the observed interaction is true in dataset i

Currently, SAINT is the only scoring model that generates a probability score. Accordingly, RScore is computed using the SAINT score.

Preparation of test data

The process of generating comprehensive interaction maps using the network reconstruction algorithm (workflow 3) of RePrint is illustrated with the help of three data sets. All three data sets targeted the mammalian Hippo pathway in *H. sapiens*. The first data set (DS7) was from “Protein Interaction Network of the Mammalian Hippo Pathway Reveals Mechanisms of Kinase-Phosphatase Interactions” by Couzens et al., Science Signaling (2013)[5]. Data generated from FLAG AP-MS experiments were considered for this analysis. Briefly, 21 FLAG tagged bait proteins were stably expressed in Ip-In 293 T-REx or Flp-In HeLa T-REx cell lines, affinity purified and analyzed on an LTQ mass spectrometer under two conditions - treated with okadaic acid and otherwise. Two biological replicates were generated for each bait purification. 15 negative controls were also included. Okadaic acid (OA), a potent inhibitor of serine and threonine phosphatases, resulted in an increased phosphorylation of some of the targeted baits. Four of

the negative controls generated included GFP as the control protein, while the rest were flag-only purifications. In summary, DS7 is a comprehensive data set that attempts to profile the interactions of several important proteins related to the Hippo pathway.

The second data set (DS8) was taken from “modular control of the co-activator YAP1” by Hauri et al., *Molecular systems biology* (2013) [83]. Here, 34 Strep-HA tagged bait proteins were stably expressed in HEK Flp-In 293 T-Rex cells, affinity purified and analyzed on an LTQ Orbitrap XL mass spectrometer. Two biological replicates were included for each bait purification. More than sixty negative controls (Strep-HA-GFP or Strep-HA-RFP-NLS) generated in their laboratory were used to score interactions.

The third data set (DS9) was taken from “Defining the Protein–Protein Interaction Network of the Human Hippo Pathway” by Wang et al., *Molecular & Cellular Proteomics* (2014) [84]. In this data set, 32 clonally modified bait proteins expressing an SBP triple tag construct in 293T cells were tandem affinity purified and analyzed on an LTQ velos instrument. 31 unrelated bait purifications were used as negative controls.

All three data sets were downloaded from public repositories: DS7 from massive (MSV000078450), DS8 from peptide atlas (PASS00281) and DS9 from proteome exchange (PXD000415). Data was processed as per the procedure described above (‘Processing of mass spectrometry data and population of RePrInt database’). The FDR in the filtered results generated by ABACUS for each of the data sets was <1%. Each data set was scored independently using SPrInt with default parameters.

Access to RePrInt

As described in earlier chapters, SPrInt, Plnt, CRAPome and RePrInt were implemented as an integrated system, on a common software platform. Post publication, the repository and tools will be available at <http://www.reprint-apms.org>. While querying and downloading data (workflows 1 and 2) are available anonymously, workflow 3 requires user registration. Users can access previous analyses until they are purged by the system administrator (after an advanced notification).

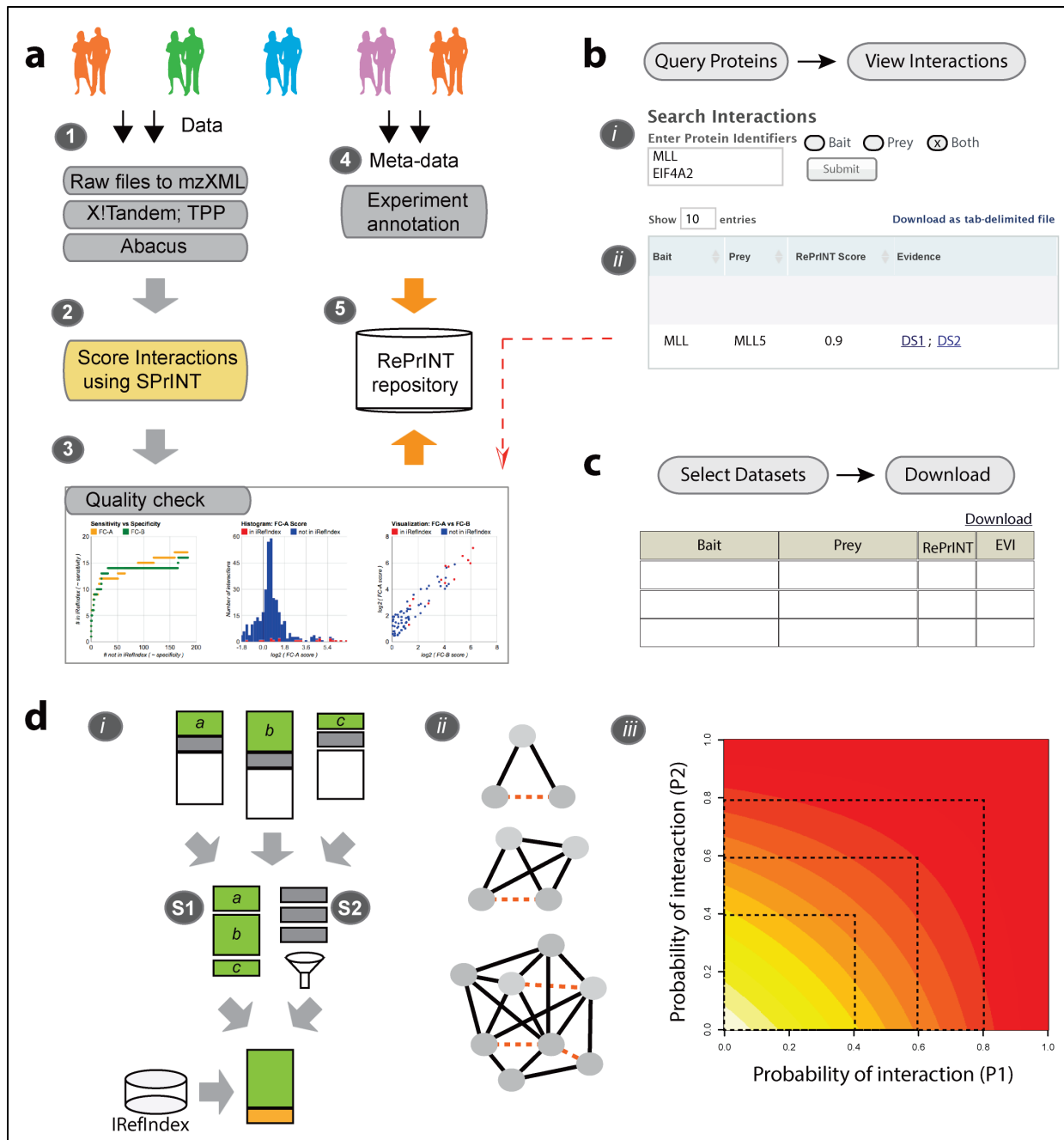


Figure 5-1: Creation of RePrint and the graphical user interface.

(a) RePrint processing pipeline. 1) Raw data is collected and processed uniformly using the X! Tandem/TPP/ABACUS pipeline (Methods). 2) Interactions are scored using SPriNT. 3) The quality of interactions is assessed manually. 4) Data sets that pass quality control check are annotated using the admin interface. 5) Processed results are submitted to RePrint database. (b) Querying interactions or workflow 1 (main text). (c) Downloading data or workflow 2 (main text). (d) Pipeline for generating protein interaction networks (workflow 3). (i) Illustration of the algorithm for network reconstruction. First, high confidence interactions are selected using various scores generated by SPriNT (S1). The cut-offs are relaxed to generate ‘borderline interactions’ (S2, main text). Borderline interactions are subjected to topological filtering before they are merged with the high scoring

interactions. (ii) Topological filtering criterion. Interactions that complete a network motif are more likely to be true than other borderline interactions. The triangle, quadrilateral motifs are shown. Similarly, interactions that interconnect members of a complex are also more likely to be true than other borderline interactions. (iii) Simulation showing the RePrint score (Methods) as a function of SAINT probabilities from two hypothetical data sets. RePrint score is shown on the color scale (red=1, white=0). The gradual curving of bands illustrates that RePrint can be used as a 'soft' cut-off, by merging evidence from multiple data sets (main text).

Results and Discussion

Creation of RePrint

The RePrint database is intended to be a web-accessible (<http://www.reprint-apms.org/>) repository of scored interactions from AP-MS experiments (both published and unpublished). The pipeline for populating the RePrint database is shown in Figure 5-1 a. **In step 1**, data is aggregated in the form of raw mass spectrometry files, which is processed uniformly using the X! Tandem/TPP/ABACUS pipeline (Methods). **In step 2**, each data set is scored online using SPrint with default parameters. Standardized negative controls from the CRAPome database are included in the analysis of a data set, when necessary. **In step 3**, the results generated by SPrint are manually inspected for quality control. When a data set meets the quality check criterion, it is marked for annotation. **In step 4**, experiments are annotated as described in Appendix B. Both CRAPome (chapter 4) and RePrint follow the same approach for annotating experiments. Briefly, the meta-data consists of three components: a) provenance of the data, b) bait description and c) the experimental protocol. The details of where and when the data was generated constitute the data provenance. The amino acid sequence, UniProt/RefSeq accession IDs and associated gene symbols are used to describe the bait protein. The experimental protocol is described using a standardized set of attributes, also known as 'controlled vocabulary (CV)'. The cell line, epitope tag, affinity approach and support are some of the important attributes in the CV (Table 4-1). Free text annotation is also included to accommodate any descriptions that the CV may not capture adequately. Once annotated, the data set is submitted to the RePrint database. **In step 5**, processed results (i.e., scored interactions generated by SPrint) of submitted data sets are parsed and stored in a SQLite database. The repository is currently under development and includes at least 6 large/medium scale data sets.

Graphical user interface

End users access the RePrInt database through a web interface (Figure 5-1 b, c). After selecting the organism of interest (currently *H. sapiens*), the database can be queried/accessed in two ways (called 'user workflows'). A third workflow helps in creating interaction maps using RePrInt data. The user interface of workflow 3 is under development.

Workflow 1: Query interactions. Here, users query the database using a list of protein or gene identifiers and retrieve scored interactions (Figure 5-1 b). As mentioned earlier, interactions in RePrInt are stored as pairs of bait-prey proteins. Accordingly, users can specify (through the GUI) whether the query should consider the input list of proteins as prey and/or bait. The results are displayed as a list of interactions. Protein abundances in bait purifications and the corresponding negative controls are shown along with the SPrInt and RePrInt scores. A hyperlink (orange arrow in Figure 5-1 b) to the SPrInt results page of the source data assists the user in interpreting/evaluating the confidence score of an interaction in the context of the original data. In other words, a fine level of detail is available to the user in addition to a brief summary.

Workflow 2: Download interaction lists. This workflow allows users to select subsets of experiments and download lists of scored interactions (Figure 5-1 c). The GUI allows users to select data sets of interest by filtering on the controlled vocabulary.

Workflow 3: Generate interaction networks. Here, users can generate interaction maps using scored interactions available in the RePrInt database. After selecting data sets of interest, users specify parameters for network reconstruction and generate interaction maps. The pipeline for network generation is detailed below. Networks can be downloaded as lists of interactions or in the portable 'graphml' file format.

Using RePrInt to generate interaction networks

In a typical setting, scored interactions are filtered to generate a list of 'high confidence interactions' (HCIs), which are then used to generate interaction maps. Stringent filtering of scored interactions reduces the false positive rate among HCIs, however it limits the inclusion of weak and transient interactions. By their very nature, weak and transient interactions result

in correspondingly low prey abundances. Accordingly, the confidence scores for such weak interactions are low. This makes the task of identifying *bona fide* weak/transient interactions quite challenging. As a first step in the direction for creating comprehensive protein interaction networks, we devised a novel strategy to ‘rescue’ *bona fide* interactions that barely miss the filtering criterion (referred to here as ‘borderline interactions’). Our pipeline uses a two step procedure for network reconstruction (Figure 5-1 d, (i)). First, a high-confidence network is generated using stringently filtered HCIs. The network is then expanded to include borderline interactions. It is common knowledge that relaxing the filtering criterion can potentially inflate the error rate (i.e., false positive identifications). Hence, we devised a ‘topological’ filter to prune the list of borderline interactions and preferentially include those that are more likely to be *bona fide*. We reasoned that interactions which complete a network motif (such as triads, tetrads or protein complexes (Figure 5-1 d, (ii)) are more likely to be *bona fide* than others. Filtered borderline interactions are appended to the high confidence interactions to generate the final network. Optionally, prior knowledge from the iRefIndex database is mapped to the final network.

A single experimental protocol may not capture all the interacting partners of a bait in an AP-MS experiment. Hence, it is a common practice to repeat an AP-MS experiment in slightly altered experimental conditions, in order to increase the coverage. Also, signaling pathways and networks are often studied by multiple groups due to their relevance to drug discovery. Accordingly, several protein interactions are often surveyed by multiple data sets. When the goal is generate a comprehensive interaction network, a question arises as to how to ‘merge’ evidence from multiple sources. The simple approach is to assign the maximum confidence score across data sets to each interaction. However, weak and transient interactions, which score low across the board, do not get selected in this approach. Hence we developed RScore (Methods), which computes the confidence of an interaction by treating each data set as an independent experiment/trial. It can be proved analytically that RScore is greater than the maximum SAINT score. Its ability to serve an ‘aggregator’ of evidence is illustrated below, using a hypothetical example.

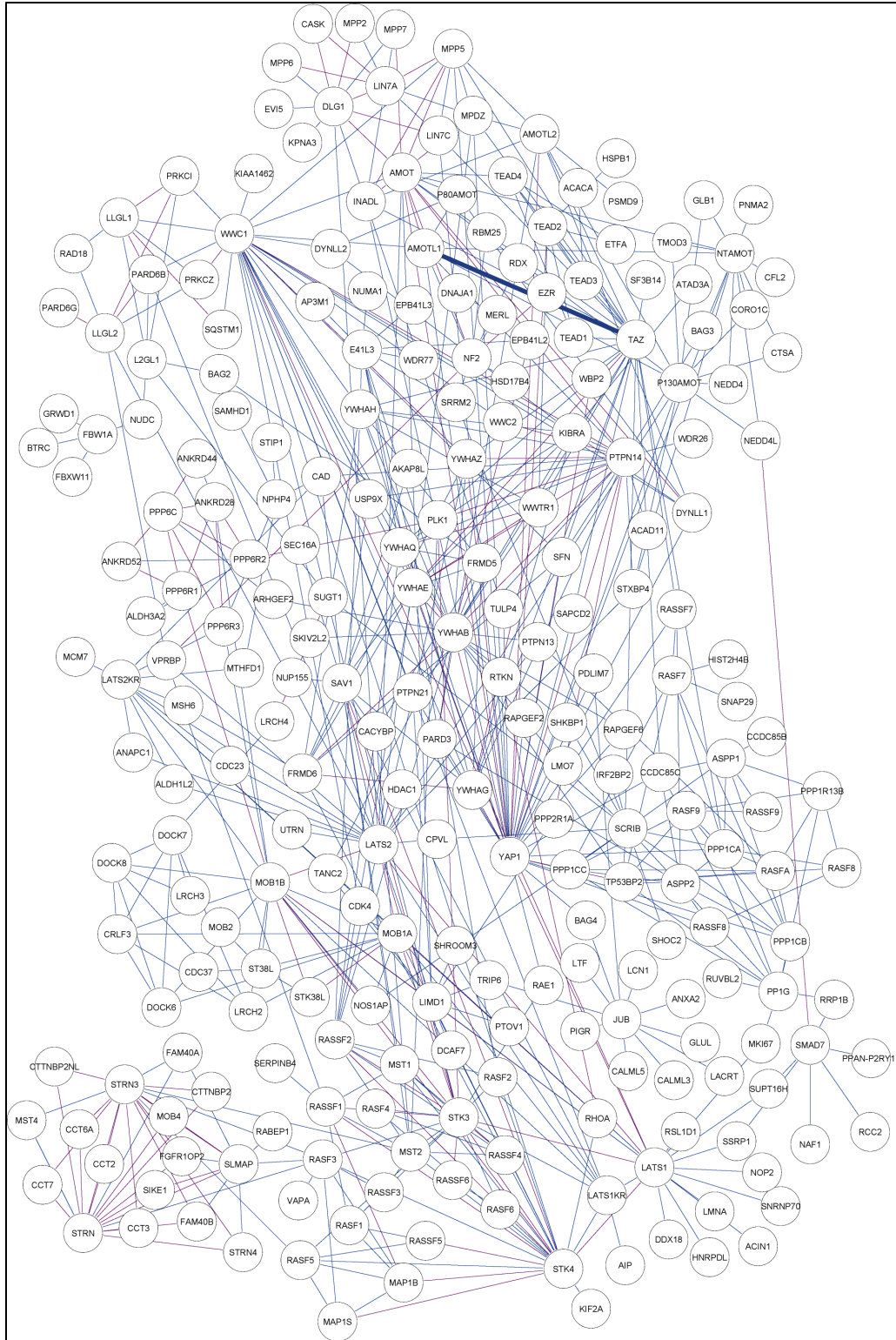


Figure 5-2: Hippo pathway generated from three data sources.

Hippo network generated by merging high scoring interactions from three data sets (DS7, DS8 and DS9). Manual inspection indicates that network recapitulates important aspects of Hippo network.

Consider two interactions profiled across two data sets. The first interaction has SAINT probabilities 0.78 and 0.79 from each data set respectively. The second interaction has SAINT probabilities 0.80 and 0.75 respectively. When a list of high confidence interactions is generated by applying a 'hard' cut-off of $\text{SAINT} \geq 0.8$, the second interaction would pass the cut-off, whereas the first would not. Intuitively, the difference between them seems to be a numerical artifact (owing to several reasons such as technical and biological variation among samples, the discrete nature of spectral count values and the granularity of scores themselves). Combining evidence using RScore generates comparable confidence scores for both the interactions (0.953 and 0.950 respectively). Hence a filtering approach that uses RScore is more likely to capture both the interactions.

We further evaluated the behavior of RScore using simulated data sets (Figure 5-1 d (iii)). RScore is plotted as a function of two hypothetical data sets that generate SAINT scores P1 and P2 for each interaction. The gradual curving of the threshold (colored bands in the heat map) indicates that RScore behaves like a 'soft' cut-off, by aggregating evidence from the two constituent data sets. The simulation suggests that RScore is a better metric to 'rescue' weak and transient interactions than the traditional approach of using maximum SAINT probability.

The utility of the RePrint database and network reconstruction strategy is illustrated using three data sets targeting the HIPPO pathway, namely DS7, DS8 and DS9. The 'high quality' network was generated using stringently filtered interactions ($\text{SAINT} \geq 0.9$, $\text{FC-A} \geq 4$ and $\text{FC-B} \geq 2$). A less stringent filtering criterion was applied ($\text{RScore} \geq 0.85$) to generate a list of 'borderline' interactions. The topological filtering criterion was defined as the ability of a borderline interaction to generate a triangle subgraph in the high-quality network (generated in step 1). Borderline interactions that passed the topological filter were then merged with the high confidence network. The iRefIndex database was queried with the nodes (proteins) in the final network and prior knowledge was mapped to the final network. If an interaction was reported in iRefIndex database, the edge was colored red (otherwise blue). The network generated using this algorithm resulted in a comprehensive interaction map (Figure 5-2) that includes several important members of the HIPPO pathway. The topological filter prunes a vast

percentage of interactions (Figure 5-3). 20% of interactions in the pruned network were reported in iRefIndex database, compared to 15% in the network that is not pruned. Taken together, these results indicate that the two-step approach for network generation creates comprehensive networks while controlling the error rates. The nature of interactions excluded by the pruning operation needs to be evaluated further.

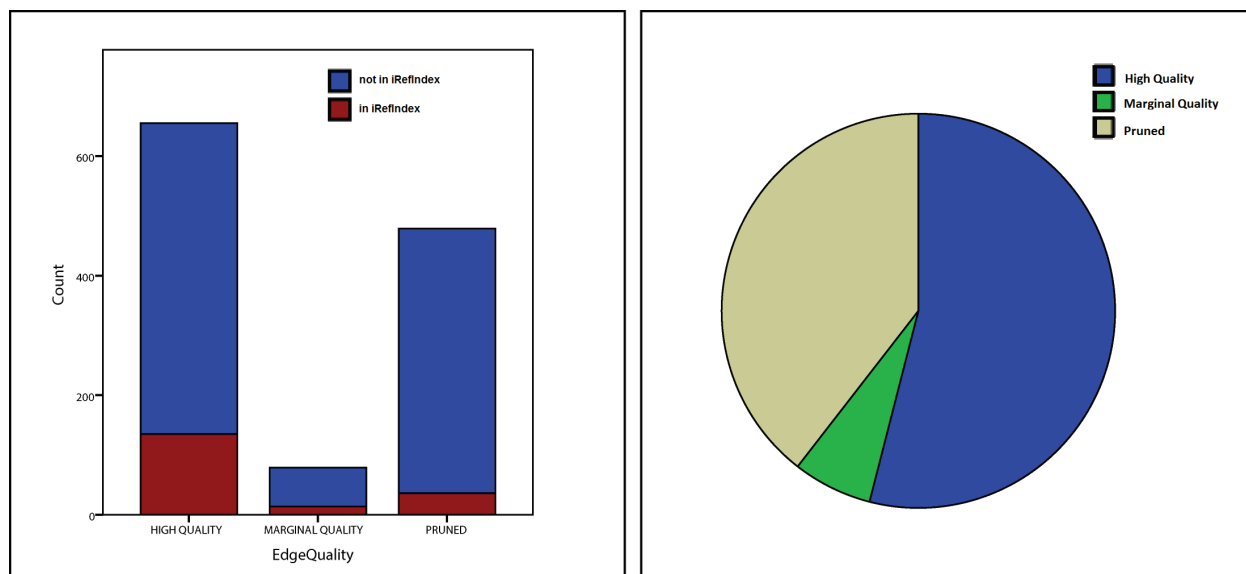


Figure 5-3: Utility of topological filtering.

Concluding Remarks

Although several protein interaction databases exist, there is no central repository for storing uniformly processed AP-MS data. The wide spread adaptation of AP-MS as the technology of choice for surveying protein interactions and the availability of a central repository of systematically annotated and uniformly processed AP-MS data provides a platform for in-depth interactome analyses. Periodic reprocessing of data using updated sequence databases and scoring functions facilitates greater utilization of raw (spectral) data. The availability of standardized negative controls in the integrated CRAPome repository can potentially improve the scoring of challenging AP-MS data sets.

At a basic level, the network reconstruction workflow is similar to other publicly available tools such as GeneMania [85] and STRINB Db [86]. However, there are two key differences. First, the availability of uniformly scored protein interactions provides greater control over filtering the

data in order to identify HCIs. Second, quantitative interaction networks can be generated by assigning a weight (such as the FC Score) to each edge in the network.

As with the CRAPome database, RePrint also uses spectral counts as the sole quantification tool, but extension of the system to other types of quantitation (especially that based on peptide ion intensity) can help identify those interactions that weren't identified using spectral count data.

The RePrint database is easily scalable. With large volumes of AP-MS data being generated regularly, the repository can play a central role in comprehensively profiling the landscape of protein interactions.

Contributions

This work is the result of collaboration between the Nesvizhskii Lab (Univ. of Michigan, Ann Arbor) and the Gingras Lab (Univ. of Toronto, CA). Bioinformatics pipelines/methods were developed by Dattatreya Mellacheruvu, under the guidance of Dr. Alexey Nesvizhskii. Dattatreya Mellacheruvu and Zachary Charles Wright (Application programmer (sr), Univ. of Michigan, Ann Arbor) implemented the system.

CHAPTER 6

Conclusions and Future Directions

Summary of the thesis

The introduction (**chapter 1**) for the thesis provided a brief overview of the significance of protein interactions in biology and discussed several important aspects of affinity purification mass spectrometry (AP-MS), a powerful technology to survey native protein interactions in the cell. A brief description of the sample preparation and different flavors of APMS were described. The processing of mass spectrometry data was detailed, with a focus on open source tools such as X! Tandem, PeptideProphet, ProteinProphet and ABACUS. Finally, the ubiquitous presence of non-specific background interactions in AP-MS data and the need for an informatics solution to identify *bona fide* interactions were discussed.

Chapter 2 of the thesis presented a standardized pipeline (SPrint) for identifying *bona fide* interactions from AP-MS data. Two basic strategies for scoring interactions were described; enrichment and specificity scoring. While specificity scoring can only be performed on medium/large scale data sets comprising several bait purifications, enrichment scoring requires negative controls generated in parallel. Each module implements two scoring functions that are, in general, complementary to each other. While one of the scoring functions in each module was described earlier in literature, the other is a novel implementation by the author. The tool also generates several visualizations of the data that help in interpreting the results. In summary, the author and colleagues have created a versatile tool that is capable of processing a wide variety of AP-MS data sets. Its simple-to-use graphical user interface enables experimentalists to rapidly analyze their data. Greater acceptance and adaptation of the tool by the research community will implicitly facilitate standardized data processing practices.

Chapter 3 of the thesis presented PInt, an integrated tool for network reconstruction using high scoring, *bona fide* protein interactions. Small scale AP-MS data sets can not capture the landscape of an interactome; hence systematic integration of prior knowledge is essential for the analysis of such networks. PInt provides two strategies for extracting prior knowledge, referred to as the ‘network context’. On the other hand, it is often the case that networks generated from medium/large scale data sets look like hairballs and are difficult to interpret. PInt provides a strategy to dissect such networks into constituent modules, thus enhancing their interpretability. Further, options for network pruning help identify core network modules. Biological network analysis is often an iterative and semi-manual process; hence integrated tools for scoring interactions and network reconstruction are invaluable for systematic, end-to-end network analysis.

Chapter 4 of the thesis presented CRAPome, a repository of standardized negative controls generated from several AP-MS data sets. High quality negative controls that profile the non-specific background in AP-MS data are not always readily available. Fortunately, negative controls in epitope-tag based AP-MS experiments are bait-independent; hence they can potentially be re-used as long as the experimental conditions remain the same. Systematic annotation using controlled vocabularies enables the selection of suitable CRAPome controls that can be included in a SPInt analysis. The utility of such standardized negative controls was demonstrated using a benchmark data set. The graphical user interface of CRAPome allows querying (using protein/gene lists) the database to generate ‘contaminant profiles’ on-the-fly. While several large laboratories have the practice of aggregating negative controls, CRAPome is the first international effort to create a central repository of standardized negative controls. The database is easily expandable and has been steadily growing ever since its original publication.

Chapter 5 of the thesis presented RePInt, a repository of protein interactions from AP-MS data. While existing protein interaction databases aggregate ‘curated’ lists of high confidence interactions, RePInt takes a data driven approach by aggregating raw spectral data and making available uniformly scored protein interactions. Two other important contributions are the

creation of a novel scoring scheme (RScore) to ‘merge’ evidence from multiple data sets and a novel network reconstruction algorithm. By their very nature, several weak and transient interactions are not ‘high scoring’. Accordingly, they often are discarded as noise. RScore and the network reconstruction algorithm can help rescue such *bona fide* interactions without inflating the error rates. RePrInt also facilitates retrospective and periodic re-processing of AP-MS data with updated sequence databases and scoring schemes. The repository is easily scalable and has the potential to develop as a prime source for interactome data.

Impact on research community

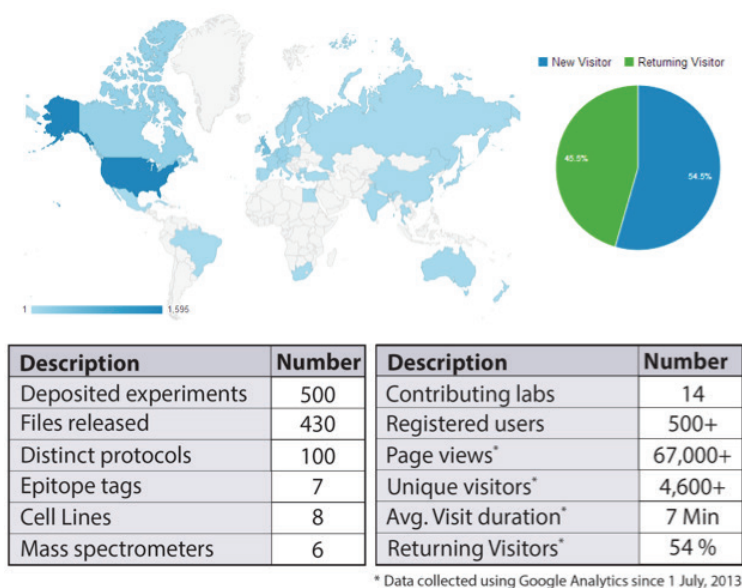


Figure 6-1: Impact on research community.

The map shows the geographic locations from where CRAPome was accessed. Areas in dark blue (e.g. the USA) correspond to geographic locations with relatively high usage (data collected using Google Analytics). The pie chart shows the fraction of returning visitors (approx. 55%). The tables describe the current status of the database (V 1.1).

Since its publication, CRAPome and SPPrInt (originally, workflow #3 of CRAPome) have been used extensively by several research groups (Figure 6-1). The repository is actively supported by several laboratories that have regularly contributed data and assisted in annotating experiments (Table 6-1). The website has had more than 67000 page views and more than 500 registered users as on October, 2014; (data collected using Google Analytics). The pipeline for

scoring interactions has been used more than 2000 times, with some researchers using the results directly in their publications (e.g. Figure 1A in ref. [87]). The original publication has been cited approximately 60 times (Google Citation count retrieved in December, 2014). A new release of the repository and associated tools is forthcoming in the first quarter of 2015. In summary, CRAPome has generated significant interest and is being widely adapted by the research community.

Contributor	Lab
Abdallah al-Hakim	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Albert J. R. Heck	Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
Amber L. Couzens	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Anne-Claude Gingras	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Annie Bouchard	Institut de recherches cliniques de Montreal (IRCM), Montreal, QC, Canada
Beatriz Gonzalez Badillo	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Benoit Coulombe	Institut de recherches cliniques de Montreal (IRCM), Montreal, QC, Canada
Brian Raught	Ontario Cancer Institute, Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
Daniel Durocher	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Denis Faubert	Institut de recherches cliniques de Montreal (IRCM), Montreal, QC, Canada
Giulio Superti-Furga	CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
Ileana M. Cristea	Department of Molecular Biology, Princeton University, Princeton, NJ, USA
Jacques Colinge	CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
Jean-Philippe Lambert	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Keiryn L. Bennett	CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
Marilyn Goudreault	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Matthias Gstaiger	Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
Mihaela E. Sardi	Stowers Institute for Medical Research, Kansas City, MO, USA
Mike P. Washburn	Stowers Institute for Medical Research, Kansas City, MO, USA

Nicole St-Denis	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Nina C. Hubner	Department of Molecular Biology; Faculty of Science; Nijmegen Centre for Molecular Life Sciences; Radboud University; Nijmegen, The Netherlands
Richard D. Bagshaw	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Rob M. Ewing	Center for Proteomics and Bioinformatics, and Department of Genetics and Genome Science, Case Western Reserve University School of Medicine, Cleveland, OH, USA
Ruedi Aebersold	Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
Shabaz Mohammed	Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
Vincentius A. Halim	Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
Wade H. Dunham	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Yana V. Miteva	Department of Molecular Biology, Princeton University, Princeton, NJ, USA
Zhen-Yuan Lin	Centre for Systems Biology, Samuel Lunenfeld Research Institute at Mount Sinai Hospital, Toronto, ON, Canada
Tuo Li	Department of Molecular Biology, Princeton University, Princeton, NJ, USA

Table 6-1: List of contributors for the CRAPome repository (V 1.0).

Utility of interaction networks in biological data analysis

As described in chapter 1, biological networks are blue prints of the underlying biological processes. While networks generated from AP-MS data represent the landscape of physical interactions, those generated using approaches such as genetic screens [88] represent functional interactions. Other low throughput methods are also typically used to reconstruct what are commonly referred to as the ‘pathways’. In addition to providing an insight into the underlying mechanisms, especially cell signaling, these networks are indispensable for the interpretation of high-throughput biological data sets. The utility of interaction networks in biological data analysis is illustrated using two specific analyses performed by the author and colleagues. Both studies were related to elucidating the regulation of pseudohyphal/invasive growth in yeast [89]. *S. cerevisiae*, when subjected to environmental stress such as nitrogen deprivation, undergoes dramatic and reversible morphological transformation to form multicellular, elongated cells that resemble hyphae. Such a transformation may lead to invasive

growth and formation of bio-films. This dimorphic behavior is also observed in other virulent organisms like *Candida albicans*, and is hence an interesting phenomenon to study.

Signaling cascades regulating invasive growth in yeast: In this study, a genome wide analysis was performed by over-expressing 4909 genes (one at a time) and measuring the resulting morphological changes under normal vegetative growth conditions [90]. 551 genes responsible for invasive/pseudohyphal growth were identified in this screen. Following analysis was performed for the functional interpretation of this resulting gene set. First, enriched KEGG pathways[91] were identified using the DAVID functional analysis tool [61]. The resulting pathways (MAPK signaling, Cell progression and Meiosis) had a high degree of overlapping gene sets, hence they were parsed using an in-house tool and visualized as a single network. Genes responsible for pseudohyphal growth were highlighted. Network visualization indicated that signaling cascades involving of Kss1 and Hog1 genes link the MAPK (signaling) pathway to cell cycle and meiosis pathways. These cues lead to further (biological) analyses, which helped elucidate the role of Hog1 in the regulation of pseudohyphal growth in yeast.

Role of SKS1 in the regulation of pseudohyphal growth in yeast. Here, the role of serine/threonine-protein kinase (sks1) as a link between pseudohyphal growth and glucose signaling pathways was investigated using quantitative phosphoproteomics [92]. Phosphopeptides from wild type cells were compared to those from sks1-K39R (kinase dead) mutant cells using SILAC (stable isotope labeling by/with amino acids in cell culture), grown under nitrogen/glucose limitation. The set of differentially expressed phosphoproteins (identified using differentially expressed phosphopeptides) were functionally interpreted using network analysis as follows. First, physical and genetic interactions among the members of the (KEGG) glycolysis, MAPK signaling and cell cycle pathways (reported in KEGG[91], BioGrid [93] and GeneMania [85]) and sks1 were used to generate a network scaffold. Both MAPK signaling and cell cycle pathways were known to be required for pseudohyphal growth in wild type cells and were hence selected for network generation. The resulting network indicated that sks1 is strongly interconnected with the glycolysis pathway.

Future directions

Projects discussed in this thesis created a computational and informatics frame work for the analysis of AP-MS data. Following are some potential improvements in the near future.

Using peptide ion intensity based quantitation. The tools and repositories presented here use spectral counts as the sole quantitation metric for scoring interactions. While spectral count data is robust, peptide ion quantitation is more sensitive [17, 94]. Both spectral count based quantitation and peptide ion intensity based quantitation can be generated in parallel. The models and databases presented here need to be extended to incorporate peptide ion intensity based quantitation. Further study is needed to understand (a) the similarities between the two approaches and (b) the advantages of each approach to identify *bona fide* protein interactions.

Analyzing the dynamic interactome. A majority of current interactome studies focus on profiling the landscape of protein interactions. These studies generate a static picture of the interactome. In reality, the interactome is dynamic. Data generated using the typical data dependent acquisition (DDA) strategy may not have the necessary sensitivity to study the dynamic interactome. However, newer approaches such as AP-SWATH are beginning to gain popularity for such analyses [95, 96]. Models for scoring interactions presented in this thesis need to be extended to handle such data. Accurate methods for data normalization that are critical for the analysis of differential and dynamic interactions also need to be developed.

Expanding the scope of RePrInt. The computational framework for analysis of protein interactions using tag-based AP-MS can be extended to similar approaches, such as immunoprecipitation mass spectrometry (IP-MS) and the analysis of RNA-protein interactions using AP-MS [9, 97, 98]. While the experimental principle in all these approaches is the same, i.e. purifying a bio-molecule complex and analyzing the constituent proteins using mass spectrometry, the data generated varies significantly. The sources of non-specific background interactions also vary significantly based on the experimental approach. For example, a major source of non-specificity in IP-MS protocols is the anti-body cross reactivity [99]. Similarly, sequence and structural homology of RNA molecules is a source of non-specificity in the

purification of RNA-protein complexes using RNA as the bait. In summary, data processing pipelines need to be tailored for each experimental approach.

Integrating data from complementary approaches. In addition to AP-MS, several complementary approaches for profiling protein-protein interactions are being developed actively. There is significant interest in prediction based methods [100] and structural mass spectrometry [101]. Integration of data from such technologies into the models for scoring interactions and network reconstruction can significantly improve the results. Also, proteogenomic approaches for the processing of spectral data can facilitate deep profiling of the interactome [102].

Creating disease specific interactomes. Most of the interactome data is generated using cell lines, which are derived from various disease states. While the current effort is to generate a composite interactome, future studies need to focus on disease specific interactomes. A comparison of several disease specific interactomes can lead to novel drug targets.

Creating a composite bio-molecule interaction database. Current interactome analyses are largely unidimensional, in the sense that profile a single type of interaction (protein-protein, RNA-protein, etc.) in any given analysis. The cell however functions as a single unit, which includes intricate interactions among multiple (types of) bio-molecules. Accordingly, all interactome data needs to be integrated to generate a composite bio-molecule interaction database.

Appendix A

Software Manual

1. Introduction

All the software tools discussed in this thesis have been developed on a common software platform and are tightly integrated. We outline here the important elements of the software design and development. All software tools used here are ‘open source’ and hosted on virtual servers managed by the University of Michigan Medical School Information Services (MSIS) and FLUX, the university wide high performance compute cluster.

Three software environments, namely sandbox, development and production, enable streamlined project execution. The sandbox is used for developing proof-of-concept applications and testing new features. Once the design is finalized, it is implemented in the development environment. There is no separate ‘test’ environment, as is usually the case in enterprise systems. Instead, testing happens in the development environment before it is readied for a release. A ready-for-release version is re-christened as the new production version by pointing the production URL to the ready-for-release version. The release management plan includes syncing user-data and meta-data (see below).

2. Design paradigm

All software is designed in the model-view-controller architectural framework (Figure A-1). This framework facilitates modular software development.

The database layer supports (a) the repo-data, i.e. protein lists with spectral abundance values extracted from raw spectral files, (b) the meta-data, i.e. experimental conditions described using controlled vocabularies and free text, and (c) the user-data, i.e. user uploaded input files and scored interactions. Repo-data and user-data are stored in SQLite databases. The meta-data is stored in a MySQL database. The database schema of CRAPome database is shown in

Figure 4-1. The RePrint schema is essentially the same. Meta-data is populated through the graphical user-interface (see below). User-data is stored in SQLite databases by controller scripts. CRAPome data is populated using in-house Python scripts. RePrint data is first generated using SPrint, which is then used to populate the RePrint database (see chapter 5).

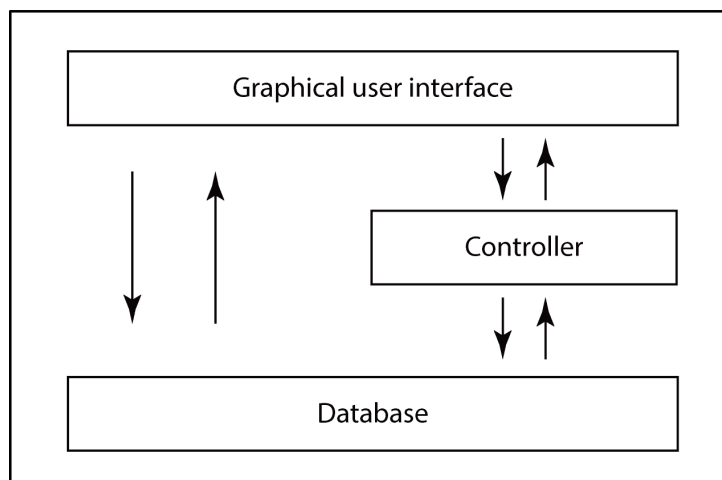


Figure A-1: Software design paradigm.

The controller scripts are the link between the graphical user interface and the database layers. Its major functions include: (a) generating formatted results and summary statistics directed against the CRAPome and RePrint databases (chapters 4 and 5 respectively), (b) computing SPrint scores (chapter 2) and (c) generating and displaying PInt networks (chapter 3). All these scripts were developed in the Python programming language.

The graphical user-interface (GUI) is developed using Drupal, a PHP-based web development framework. Drupal was originally intended to be a content management system, but current versions have all the essential features of a web framework. Drupal provides all the basic features of a web-enabled software tool, such as the login module, access control, etc. The GUI provides an interface to: (a) populate the meta-data, (b) access the repo- and meta- data and (c) access the SPrint and PInt pipelines. Graphical visualizations generated by SPrint are rendered using the Google Visualization API. Networks generated by PInt are rendered using Cytoscape web.

3. Software architecture diagram

The software architecture diagram indicating various components and their interconnections is shown (Figure A-2).

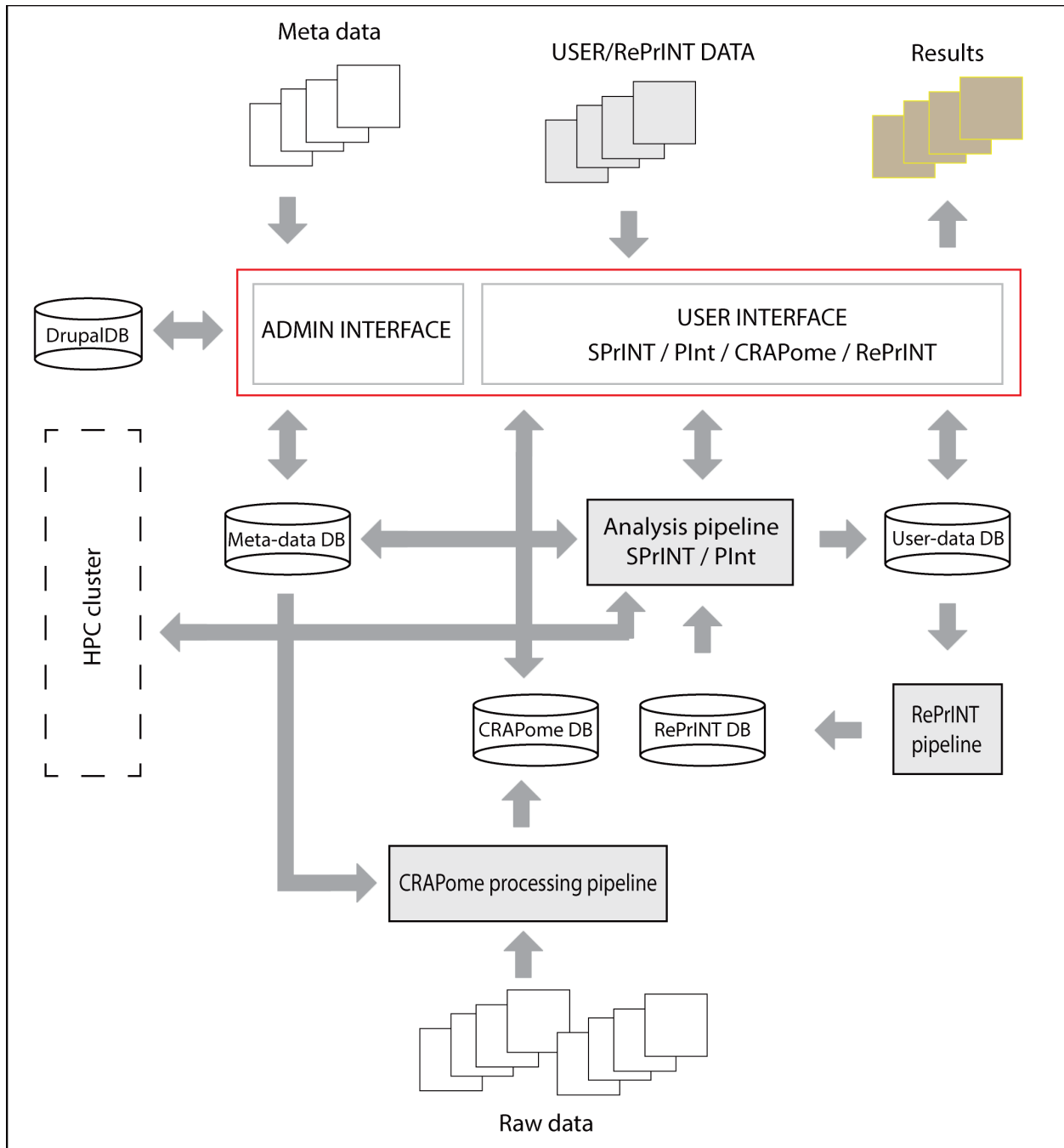


Figure A-2: Software architecture diagram.

The GUI, developed in Drupal, is highlighted in red. Controller scripts, developed in Python, are shown in a box with a grey background. Databases are represented with cylinders and interfaces are indicated with double headed arrows. Data files are indicated using cascaded squares. The 'ADMIN INTERFACE' accepts meta-data and populates

the Meta-data DB. The 'USER INTERFACE' provides access to the CRAPome and RePrInt databases and SPrint and Plnt pipelines. 'Analysis pipeline' generates scored interactions and networks. The 'CRAPome processing pipeline', developed in Python extracts protein and peptide lists from processed spectral data and populates the 'CRAPome DB'. The 'RePrInt pipeline' extracts scored interactions corresponding to RePrInt data sets and populates the 'RePrInt DB'. The 'Analysis pipeline' interacts with 'HPC cluster', i.e. FLUX, and facilitates processing of computationally intensive jobs (see 'System integration').

4. System integration

A fire-and-forget approach is used to design the system integration components, between the web server and the university wide high performance compute cluster (FLUX) (Figure A-3). An NFS mountable disk drive that is visible to both the web server and FLUX serves as a conduit between the systems. The web server has an 'inbox' (IN) and an 'outbox' (OUT) on the conduit. A data file that needs to be processed on FLUX is placed in the outbox. A daemon, i.e. a program that runs every second, on FLUX monitors the outbox for new input files, picks them up and processes the data. The results are placed in the inbox. A daemon on the web server monitors the inbox, picks up new results and presents them to the user.

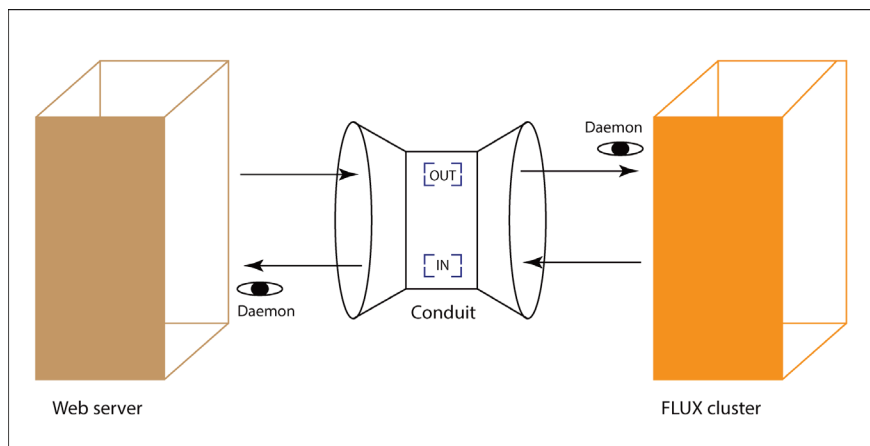


Figure A-3: Overview of system integration.

5. Processing user requests (jobs)

User data is processed using TORQUE, an open source computing resource management system. The job execution pipeline is shown in Figure A-4. The GUI facilitates uploading user data, selecting analysis options and submitting a request to process their data (job). Each request is stored in a 'userJob' table. A daemon, i.e. a program that runs every one second, monitors the table for new requests, and submits the job to TORQUE. The queuing system of TORQUE processes user requests one at a time, on a first-come, first-served basis. Processed results are stored in a SQLite database (see 'Design paradigm' and 'Software architecture

diagram'). The queue also updates the user job table when a job is completed. The results of the completed jobs (data tables and visualization) can be viewed on the GUI.

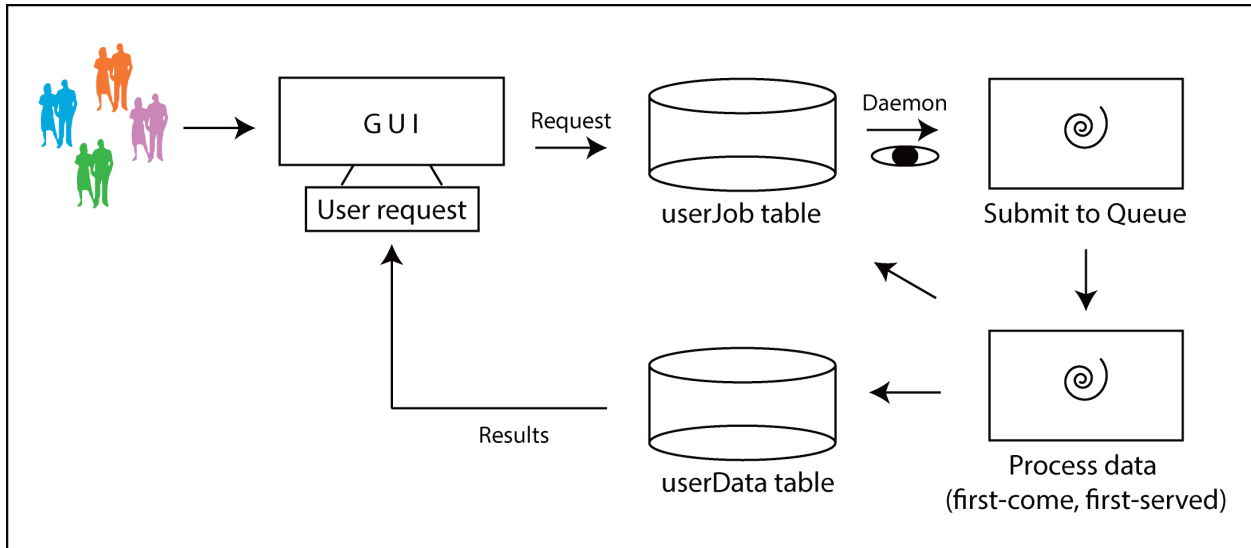


Figure A-4: Overview of user job execution.

Appendix B

Annotator Manual

1. Overview

CRAPome and RePrInt are repositories of affinity purification coupled with mass spectrometry (AP-MS) experiments. An annotator is usually the contributor of mass spectrometry data. Contributors first submit raw mass spectrometry files to the CRAPome/RePrInt administrator. The administrator processes the data through X! Tandem/TPP/ABACUS/SPrInt pipeline and performs a basic quality check (see Chapters 4, 5). Experiments that pass the quality control checks are annotated as follows.

2. Annotation procedure

Annotation is a four step process (Figure B-1). First a 'data set' entry is created, which captures the provenance of the data. Second, an entry for each bait in the data set is created, as needed. Third, a protocol is defined using controlled vocabularies. Finally, an experiment entry is created for each experiment. The corresponding data set, bait and protocol records are associated to each experiment. Each of these steps is illustrated in the following sections.

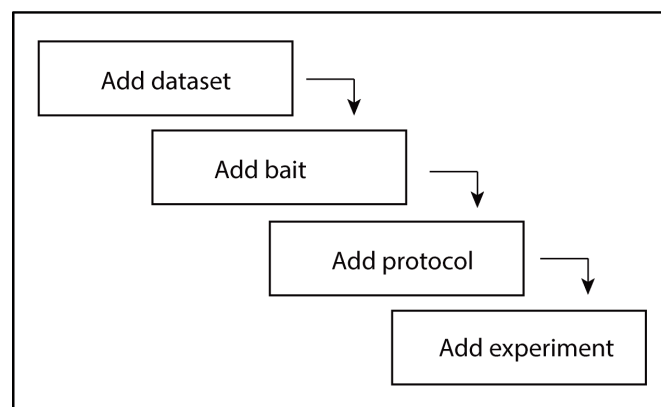


Figure B-1: Overview of annotation procedure.

3. Accessing the system as an annotator

Annotators are assigned a higher level of privileges than regular registered users. They can create data sets, baits, protocols and link protocols to experiments. Annotator-level login access can be requested by emailing the CRAPome administrator. The screens that allow an annotator to add/edit experiments and protocols is shown (Figure B-2, Figure B-3).

Name	Prey Count	File Name	Protocol(s)	Tag	Fractionation	Cell Line	Action
CC40	486	ACG_FLAG_Trex293_MAG_9560	293 Flip-In FLAG mag LTQ	FLAG	total cell lysate	HEK293 magnetic (agarose coated)	edit
		HeLa_chAP1_Orbi_11JUN2011				HeLa magnetic (agarose coated)	edit
CC22	343	ACG_2276_MK_1_Flag_control_20090427	293 stables FLAG agarose LTQ -...	FLAG	total cell lysate	HEK293 agarose	edit

Figure B-2: Experiment view (annotator login)

Prot. ID	Prot. Name	Comments	Epitope Tag	Fractionation	Cell Line	Action*
14	293 Flip-In FLAG mChip LTQ	Gingras lab - version 1.0	FLAG	1D LC-MS	HEK293	edit
15	293 Flip-In FLAG mChip Orbitrap	Gingras lab - version 1.0	FLAG	1D LC-MS	HEK293	edit
16	293 Flip-In FLAG mChip Yeast	Gingras lab - version 1.0	FLAG	1D LC-MS	HEK293	edit
22	293T transient FLAG LTQ	Pawson lab protocol	FLAG	1D LC-MS	HEK293	-
23	293T transient GFP LTQ	Pawson lab protocol	GFP	1D LC-MS	HEK293	-
24	293 stables FLAG agarose LTQ - GC	Gingras lab - version 1.0; Ginny Chen	FLAG	1D LC-MS	HEK293	edit
25	293 Flip-In FLAG agarose LTQ - MM	Gingras lab - version 1.0; Michael Mullin	FLAG	1D LC-MS	HEK293	edit
26	293 Flip-In FLAG agarose LTQ - AA	Durocher lab	FLAG	1D LC-MS	HEK293	-
27	HeLa Flip-In FLAG mChip LTQ		FLAG	1D LC-MS	HeLa	edit
28	HeLa Flip-In FLAG mChip Orbitrap	Gingras lab - version 1.0	FLAG	1D LC-MS	HeLa	edit
29	293 Flip-In pools FLAG magnetic LTQ - AC	Gingras lab - version 2.0 - Amber Couzens	FLAG	1D LC-MS	HEK293	edit
32	example protocol	This is an example Protocol.	FLAG	1D LC-MS	HEK293	edit

*NOTE: Deletion is only permitted for administrators on unassociated protocols.

Figure B-3: Protocol view (annotator login)

4. Managing controlled vocabularies

Each experiment is annotated using a pre-defined set of controlled vocabularies(Figure B-4). Annotators can only use pre-defined CVs; new CVs are added by the system administrator.

Current Attributes	
Attribute Name	Attribute Values
Organism	human
Cell/tissue type	HEK293, HeLa, U2OS, PBMC, Jurkat, CEM-T, MRC-5, LS174
Cell/tissue subtype	-, HEK293T, HEK293 Flp-In T-REx, Jurkat-Flp-In
Drug treatment	aphidicolin, rapamycin, nocodazole, MG132, none, IFN-beta, DMSO, okadaic acid, doxycycline+thymidine, tetracycline+thymidine, thymidine+nocodazole
Subcellular fractionation	total cell lysate, total lysate+chromatin, nuclear fraction, cytosolic fraction
Epitope tag	FLAG, HA, GFP, TAP, HaloTag, Strep-HA
Control protein	RFP, GFP, FLAG, mCherry, tag alone, untransfected, uninduced, RFP
AP steps	single, tandem

Figure B-4: Controlled vocabularies (annotator login).

5. Adding data sets

A data set record is intended to capture the data provenance. The screen for adding a dataset is shown (Figure B-5 a). Primary information, such as who generated the data and when is captured here.

6. Adding baits

A bait record is intended to capture the full details of the bait protein. Primary information, such as the sequence, accession number and gene name are captured (Figure B-5 b). When modified baits are used for affinity purification, the reference sequence is indicated in the 'Original Sequence' field.

Figure B-5: Adding data sets and baits (annotator login).

7. Adding protocols

A protocol record is intended to capture the experimental conditions. Annotators can add/edit protocols according to the guidelines shown in Figure B-6.

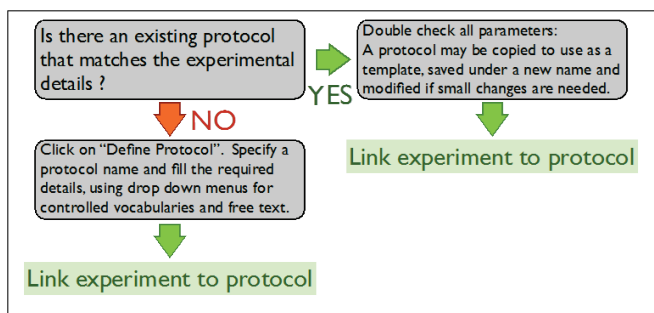


Figure B-6: Guidelines for adding new protocols (annotator login).

A protocol description has two parts: A) defining the set of controlled vocabularies, i.e., attributes and values (Figure B-7) and B) providing free form text to capture additional details (Figure B-8). The system warns the annotator when a protocol is duplicated.

ORAPome
Contaminant Repository for Affinity Purification
An online database of common contaminants in AP-MS experiments

Home Experiments Protocols Define Experiment Define Protocol About Tutorial Logout

Add Protocol

NOTICE: Greek or other non-Latin characters/symbols

Protocol Name *

enter an protocol name

Protocol Comments

enter protocol comments

Select Lab
ACG

Attributes

Cell Line

Cell Line Subtype

Epitope Tag

Subcellular fractionation

Fractionation

Affinity support 1

Figure B-7: Creating new protocol; part A: define controlled vocabulary (annotator login).

Biological Material
Stable pools, HEK293. Transfection of low passage HEK293 (CRL1573) with vector using lipofectamine PLUS; selection for ~14 days with 750ug/ml active G418. Amplification off cells in 5-8 x 150mm plates; harvesting at 80-95% confluence. Harvest through scraping, followed by 3 washes with PBS. Cell pellet either frozen on dry ice and stored dry at -80C, or processed immediately.

Affinity Purification
Cells were lysed by [passive lysis assisted by freeze-thaw]. Briefly, to the frozen cell pellet, 1:4 or 1:5 pellet weight:volume ratio of lysis buffer was added. Lysis buffer was 50 mM Hepes-NaOH pH 8.0, [100 mM KCl], 2 mM EDTA, [0.1% NP40], 10% glycerol, 1 mM PMSF, 1 mM DTT and Sigma protease inhibitor cocktail, P8340, 1:500. No phosphatase inhibitors were added. Resuspended pellets were incubated on ice (or on a rotator at 4°C) for 10 min to assist lysis, then pipetted up and down to break up pellet. Tubes were frozen and thawed once (liquid nitrogen or dry ice ~5min, 37°C with agitation, then put on ice, and the lysate transferred to 2 ml Eppendorf tubes. An aliquot (20ul) was taken to monitor solubility (This aliquot was spun down, the supernatant transferred to a fresh tube, and 6 µl 4X Laemmli sample buffer added to the supernatant. The pellet was resuspended in 26 µl 2X Laemmli sample buffer). The 2 ml tubes were centrifuged at 14000 rpm for 20 min at 4°C, and the supernatant transferred to fresh 15 ml conical tubes. The protein concentration was measured (using BSA as a control). To the rest of the lysate, 25-30 µl packed [FLAG M2 agarose beads] pre-washed 4X in lysis buffer were added, and the mixture incubated 2 hours at

Peptide Preparation
Trypsin (1 µg Sigma Trypsin Singles, T7575) dissolved in 70 µl of 50 mM ammonium bicarbonate pH 8 was added each sample. Tubes were vortexed, briefly centrifuged, and incubated at 37°C overnight. After quickly centrifuging the samples, an additional amount of trypsin (0.25 µg) was added, and the samples incubated for another 3-4 hours. The samples were acidified by adding 2 µl of 50% formic acid, and lyophilized in the speed-vac. The samples were stored at -40°C. When ready for mass spectrometry, 20 µl 5% formic acid was added to the samples and the samples were centrifuged at max speed for 10 min.

LC-MS
The ammonium bicarbonate was evaporated, and the samples were resuspended in HPLC buffer A (2% acetonitrile, 0.1% formic acid), then directly loaded onto capillary columns packed in-house with Magic 5 µm, 100A, C18AQ. MS/MS data was acquired in data-dependent mode (over a 65min - 2 hr acetonitrile 2-40% gradient) on a ThermoFinnigan LTQ, equipped with a Proxeon NanoSource and an Agilent 1100 capillary pump.

Publication Reference
Chen et al., J Biol Chem, 2008. PMID:18715871; Chen and Gingras, Methods, 2007. PMID: 17532517; Goudreau et al., Mol Cell Proteomics, 2009. PMID: 18782753; Kean et al., J Biol Chem., PMID: 21561862.

Figure B-8: Creating new protocol; part B: adding protocol details (annotator login).

Add information details pertaining to the biological material (How were the cells grown and harvested? How was the recombinant protein expressed? Has a subcellular fractionation been performed?), the affinity purification step, the procedure for preparing the peptides (including fractionation at the protein or peptide level when applicable), and details of the LC-MS/MS analysis. If the Method has been published, add citations in the "Publication reference" box.

8. Adding experiments

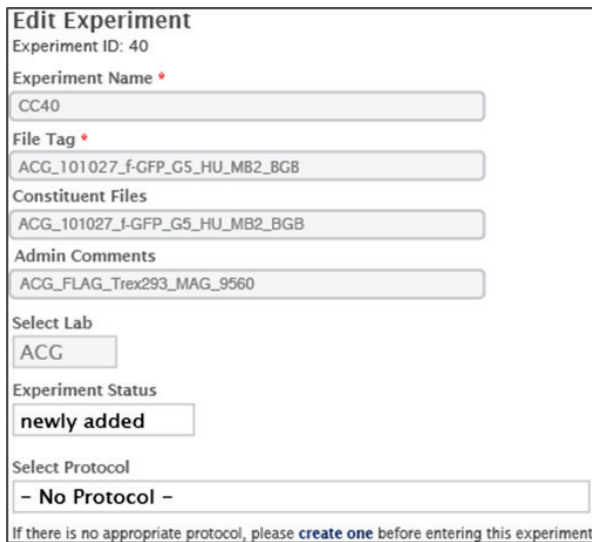
An experiment record primarily captures the file name and associated data set, bait and protocol. It is often the case that the system administrator defines the experiment record with dummy protocol information, while the annotator ‘edits’ a record to update the annotation (Figure B-9, Figure B-10). Data set, bait and protocol records that are defined separately (see above) and linked to an experiment (Figure B-11).



Name	Num Preys	File Tag	Protocol Number	Protocol	Tag	Fractionation	Cell Line	Affinity Supp.	
CC40	486	ACG_101027_f-GFP_G5_HU_MB2_BCB							edit
CC5	339	ACG_11303_untagged_HeLa_chAP1_Orbi_11JUN2011	28	HeLa Flp-In FLAG mChip Orbitra...	FLAG	total lysate+chromatin	HeLa	magnetic (agarose coated)	edit

Figure B-9: Experiment view (annotator login).

Clicking on “edit” on the right enable linking a protocol to the experiment.



Edit Experiment
Experiment ID: 40

Experiment Name *
CC40

File Tag *
ACG_101027_f-GFP_G5_HU_MB2_BGB

Constituent Files
ACG_101027_f-GFP_G5_HU_MB2_BGB

Admin Comments
ACG_FLAG_Trex293_MAG_9560

Select Lab
ACG

Experiment Status
newly added

Select Protocol
- No Protocol -

If there is no appropriate protocol, please [create one](#) before entering this experiment.

Figure B-10: Editing experiments to associate protocol (annotator login).

Data entered by the administrator is greyed out. Select a protocol to link to the experiment. Create new protocols as needed, as described above.

Experiment Status
show

Select Protocol
20-293 Flp-In FLAG mag LTQ
If there is no appropriate protocol, please **create one** before entering this experiment.

Controlled Vocabulary

Organism: human

Cell/tissue type: HEK293

Cell/tissue subtype: HEK293 Flp-In T-REx

Drug treatment: none

Subcellular fractionation: total cell lysate

Epitope tag: FLAG

Control protein: GFP

AP steps: single

Figure B-11: Linking protocol to experiments using the ‘Select Protocol’ drop down menu (annotator login).

9. Deleting meta data

To avoid accidental deletion of data, annotators are restricted to adding and editing meta-data that are assigned to their lab. A lab is assigned to the annotator at the time of login generation. The status (e.g. newly defined, ready-for-release, obsolete, etc.) of every meta-data (i.e., data set, bait, protocol and experiment) is indicated by the annotator. Obsolete records are purged by the system administrator periodically.

References

1. Mellacheruvu, D., et al., *The CRAPome: a contaminant repository for affinity purification-mass spectrometry data*. Nat Meth, 2013. **10**(8): p. 730-736.
2. Grunberg, S. and S. Hahn, *Structural insights into transcription initiation by RNA polymerase II*. Trends Biochem Sci, 2013. **38**(12): p. 603-11.
3. Korobeinikova, A.V., M.B. Garber, and G.M. Gongadze, *Ribosomal proteins: structure, function, and evolution*. Biochemistry (Mosc), 2012. **77**(6): p. 562-74.
4. Joshi, P., et al., *The functional interactome landscape of the human histone deacetylase family*. Mol Syst Biol, 2013. **9**: p. 672.
5. Couzens, A.L., et al., *Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions*. Sci Signal, 2013. **6**(302): p. rs15.
6. Nusse, R., *Wnt signaling*. Cold Spring Harb Perspect Biol, 2012. **4**(5).
7. Cohen, P., *Protein kinases--the major drug targets of the twenty-first century?* Nat Rev Drug Discov, 2002. **1**(4): p. 309-15.
8. Walhout, A.J.M. and M. Vidal, *High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping*. Methods, 2001. **24**(3): p. 297-306.
9. Dunham, W.H., M. Mullin, and A.C. Gingras, *Affinity-purification coupled to mass spectrometry: basic principles and strategies*. Proteomics, 2012. **12**: p. 1576-1590.
10. Hunt, D.F., et al., *Protein sequencing by tandem mass spectrometry*. Proc Natl Acad Sci U S A, 1986. **83**(17): p. 6233-7.
11. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry*. Nat Methods, 2007. **4**(10): p. 787-97.
12. Eng, J.K., A.L. McCormack, and J.R. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database*. J Am Soc Mass Spectrom, 1994. **5**(11): p. 976-89.
13. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra*. Bioinformatics, 2004. **20**: p. 1466-1467.
14. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search*. Anal. Chem., 2002. **74**: p. 5383-5392.
15. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry*. Anal. Chem., 2003. **75**: p. 4646-4658.
16. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics*. J Proteomics, 2010. **73**(11): p. 2092-123.
17. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present*. Anal Bioanal Chem, 2012. **404**(4): p. 939-65.
18. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
19. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.

20. Choi, H., *SAINT: probabilistic scoring of affinity purification-mass spectrometry data*. Nat. Methods, 2011. **8**: p. 70-73.
21. Behrends, C., et al., *Network organization of the human autophagy system*. Nature, 2010. **466**(7302): p. 68-76.
22. Gingras, A.C., et al., *Analysis of protein complexes using mass spectrometry*. Nat. Rev. Mol. Cell Biol., 2007. **8**: p. 645-654.
23. Sowa, M.E., et al., *Defining the human deubiquitinating enzyme interaction landscape*. Cell, 2009. **138**(2): p. 389-403.
24. Krasner, G. and S. Pope, *A Description of the Model-View-Controller User Interface Paradigm in the Smalltalk-80 System*. Journal of Object Oriented Programming, 1988. **1**(3): p. 26--49.
25. Choi, H., *Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT*. Curr. Protoc. Bioinformatics, 2012. **39**: p. 8.15.
26. Manning, C.D., et al., *Introduction to Information Retrieval*. 2008: Cambridge University Press. 496.
27. Razick, S., G. Magklaras, and I.M. Donaldson, *iRefIndex: a consolidated protein interaction database with provenance*. BMC Bioinformatics, 2008. **9**: p. 405.
28. Tzivion, G., Z. Luo, and J. Avruch, *A dimeric 14-3-3 protein is an essential cofactor for Raf kinase activity*. Nature, 1998. **394**: p. 88-92.
29. Wartmann, M. and R.J. Davis, *The native structure of the activated Raf protein kinase is a membrane-bound multi-subunit complex*. J. Biol. Chem., 1994. **269**: p. 6695-6701.
30. Gingras, A.C., B. Raught, and N. Sonenberg, *eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation*. Annu. Rev. Biochem., 1999. **68**: p. 913-963.
31. Miki, H., K. Miura, and T. Takenawa, *N-WASP, a novel actin-depolymerizing protein, regulates the cortical cytoskeletal rearrangement in a PIP2-dependent manner downstream of tyrosine kinases*. EMBO J., 1996. **15**: p. 5326-5335.
32. Jeronimo, C., *Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme*. Mol. Cell, 2007. **27**: p. 262-274.
33. Dunham, W.H., *A cost-benefit analysis of multidimensional fractionation of affinity purification-mass spectrometry samples*. Proteomics, 2011. **11**: p. 2603-2612.
34. Deutsch, E.W., *A guided tour of the Trans-Proteomic Pipeline*. Proteomics, 2010. **10**: p. 1150-1159.
35. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy*. Nucleic Acids Res., 2012. **40**: p. D130-D135.
36. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem*. Mol. Cell. Proteomics, 2005. **4**: p. 1419-1440.
37. Fermin, D., et al., *Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis*. Proteomics, 2011. **11**: p. 1340-1345.
38. Varjosalo, M., et al., *Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS*. Nat Meth, 2013. **10**(4): p. 307-314.
39. Taipale, M., et al., *A Quantitative Chaperone Interaction Network Reveals the Architecture of Cellular Protein Homeostasis Pathways*. Cell, 2014. **158**(2): p. 434-448.
40. Liu, G., et al., *ProHits: integrated software for mass spectrometry-based interaction proteomics*. Nat Biotechnol, 2010. **28**(10): p. 1015-7.
41. Breitkreutz, A., *A global protein kinase and phosphatase interaction network in yeast*. Science, 2010. **328**: p. 1043-1046.
42. Skarra, D.V., *Label-free quantitative proteomics and SAINT analysis enable interactome mapping for the human Ser/Thr protein phosphatase 5*. Proteomics, 2011. **11**: p. 1508-1516.

43. Choi, H., et al., *SAINT: probabilistic scoring of affinity purification-mass spectrometry data*. Nat Methods, 2011. **8**(1): p. 70-3.
44. Shen, Z., *A WD-repeat protein stabilizes ORC binding to chromatin*. Mol. Cell, 2010. **40**: p. 99-111.
45. Guruharsha, K.G., et al., *A protein complex network of Drosophila melanogaster*. Cell, 2011. **147**(3): p. 690-703.
46. Jager, S., et al., *Global landscape of HIV-human protein complexes*. Nature, 2012. **481**(7381): p. 365-70.
47. Choi, H., et al., *SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments*. J Proteome Res, 2012. **11**(4): p. 2619-24.
48. Saito, R., et al., *A travel guide to Cytoscape plugins*. Nat Methods, 2012. **9**(11): p. 1069-76.
49. de Hoon, M.J., et al., *Open source clustering software*. Bioinformatics, 2004. **20**: p. 1453-1454.
50. Page, R.D., *TreeView: an application to display phylogenetic trees on personal computers*. Comput. Appl. Biosci., 1996. **12**: p. 357-358.
51. Greco, T.M., et al., *Nuclear import of histone deacetylase 5 by requisite nuclear localization signal phosphorylation*. Mol. Cell. Proteomics, 2011. **10**: p. M110.004317.
52. Selbach, M. and M. Mann, *Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK)*. Nat. Methods, 2006. **3**: p. 981-983.
53. Trinkle-Mulcahy, L., *Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes*. J. Cell Biol., 2008. **183**: p. 223-239.
54. Trinkle-Mulcahy, L., *Resolving protein interactions and complexes by affinity purification followed by label-based quantitative mass spectrometry*. Proteomics, 2012. **12**: p. 1623-1638.
55. Tackett, A.J., *I-DIRT, a general method for distinguishing between specific and nonspecific protein interactions*. J. Proteome Res., 2005. **4**: p. 1752-1756.
56. Hubner, N.C., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions*. J. Cell Biol., 2010. **189**: p. 739-754.
57. Nesvizhskii, A.I., *Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments*. Proteomics, 2012. **12**: p. 1639-1655.
58. Sardiou, M.E., *Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics*. Proc. Natl. Acad. Sci. USA, 2008. **105**: p. 1454-1459.
59. Ewing, R.M., *Large-scale mapping of human protein-protein interactions by mass spectrometry*. Mol. Syst. Biol., 2007. **3**: p. 89.
60. Kasprzyk, A., *BioMart: driving a paradigm change in biological data management*. Database (Oxford), 2011. **2011**: p. bar049.
61. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nat. Protoc., 2009. **4**: p. 44-57.
62. Thakur, S.S., *Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation*. Mol. Cell Proteomics, 2011. **10**: p. M110.003699.
63. Al-Hakim, A.K., et al., *Interaction proteomics identify NEURL4 and the HECT E3 ligase HERC2 as novel modulators of centrosome architecture*. Mol. Cell. Proteomics, 2012. **11**: p. M111.014233.
64. Chen, G.I., *PP4R4/KIAA1622 forms a novel stable cytosolic complex with phosphoprotein phosphatase 4*. J. Biol. Chem., 2008. **283**: p. 29273-29284.
65. Cristea, I.M., et al., *Fluorescent proteins as proteomic probes*. Mol. Cell. Proteomics, 2005. **4**: p. 1933-1941.
66. Daniels, D.L., *Examining the complexity of human RNA polymerase complexes using HaloTag technology coupled to label free quantitative proteomics*. J. Proteome Res., 2012. **11**: p. 564-575.

67. Forget, D., *The protein interaction network of the human transcription machinery reveals a role for the conserved GTPase RPAP4/GPN1 and microtubule assembly in nuclear import and biogenesis of RNA polymerase II*. Mol. Cell. Proteomics, 2010. **9**: p. 2827-2839.
68. Goudreault, M., *A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein*. Mol. Cell. Proteomics, 2009. **8**: p. 157-171.
69. Kean, M.J., *Structure-function analysis of core STRIPAK proteins: a signaling complex implicated in Golgi polarization*. J. Biol. Chem., 2011. **286**: p. 25065-25075.
70. Kruiswijk, F., *Coupled activation and degradation of eEF2K regulates protein synthesis in response to genotoxic stress*. Sci. Signal., 2012. **5**: p. ra40.
71. Sato, S., *A set of consensus mammalian mediator subunits identified by multidimensional protein identification technology*. Mol. Cell, 2004. **14**: p. 685-691.
72. de Lau, W., *Lgr5 homologues associate with Wnt receptors and mediate R-spondin signalling*. Nature, 2011. **476**: p. 293-297.
73. Tsai, Y.C., et al., *Functional proteomics establishes the interaction of SIRT7 with chromatin remodeling complexes and expands its role in regulation of RNA polymerase I transcription*. Mol. Cell. Proteomics, 2012. **11**: p. M111.015156.
74. Rudashevskaya, E.L., *A method to resolve the composition of heterogeneous affinity-purified protein complexes assembled around a common protein by chemical cross-linking, gel electrophoresis and mass spectrometry*. Nat. Protoc., 2013. **8**: p. 75-97.
75. Pichlmair, A., *Viral immune modulators perturb the human molecular network by common and unique strategies*. Nature, 2012. **487**: p. 486-490.
76. Varjosalo, M., *Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS*. Nat. Methods, 2013. **10**: p. 307-314.
77. Chen, G.I. and A.C. Gingras, *Affinity-purification mass spectrometry (AP-MS) of serine/threonine phosphatases*. Methods, 2007. **42**: p. 298-305.
78. Gingras, A.C., *A novel, evolutionarily conserved protein phosphatase complex involved in cisplatin sensitivity*. Mol. Cell Proteomics, 2005. **4**: p. 1725-1740.
79. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update*. Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
80. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update*. Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
81. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases*. Nucleic acids research, 2014. **42**(Database issue): p. D358-63.
82. Gingras, A.C., et al., *Analysis of protein complexes using mass spectrometry*. Nat Rev Mol Cell Biol, 2007. **8**(8): p. 645-54.
83. Hauri, S., et al., *Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1*. Mol Syst Biol, 2013. **9**: p. 713.
84. Wang, W., et al., *Defining the protein-protein interaction network of the human hippo pathway*. Mol Cell Proteomics, 2014. **13**(1): p. 119-31.
85. Montojo, J., et al., *GeneMANIA: Fast gene network construction and function prediction for Cytoscape*. F1000Res, 2014. **3**: p. 153.
86. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life*. Nucleic Acids Res, 2014.
87. Smith-Kinnaman, W.R., et al., *The interactome of the atypical phosphatase Rtr1 in Saccharomyces cerevisiae*. Mol Biosyst, 2014. **10**(7): p. 1730-41.
88. Forsburg, S.L., *The art and design of genetic screens: yeast*. Nat Rev Genet, 2001. **2**(9): p. 659-668.

89. Ryan, O., et al., *Global Gene Deletion Analysis Exploring Yeast Filamentous Growth*. Science, 2012. **337**(6100): p. 1353-1356.
90. Shively, C.A., et al., *Genetic Networks Inducing Invasive Growth in Saccharomyces cerevisiae Identified Through Systematic Genome-Wide Overexpression*. Genetics, 2013. **193**(4): p. 1297-1310.
91. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets*. Nucleic Acids Research, 2011.
92. Johnson, C., et al., *The Yeast Sks1p Kinase Signaling Network Regulates Pseudohyphal Growth and Glucose Response*. PLoS Genet, 2014. **10**(3): p. e1004183.
93. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2015 update*. Nucleic Acids Res, 2015. **43**(Database issue): p. D470-8.
94. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review*. Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.
95. Lambert, J.P., et al., *Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition*. Nat Methods, 2013. **10**(12): p. 1239-45.
96. Collins, B.C., et al., *Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system*. 2013. **10**(12): p. 1246-53.
97. Tsai, B.P., et al., *Quantitative profiling of in vivo-assembled RNA-protein complexes using a novel integrated proteomic approach*. Mol Cell Proteomics, 2011. **10**(4): p. M110.007385.
98. Oeffinger, M., *Two steps forward--one step back: advances in affinity purification mass spectrometry of macromolecular complexes*. Proteomics, 2012. **12**(10): p. 1591-608.
99. Malovannaya, A., et al., *Analysis of the human endogenous coregulator complexome*. Cell, 2011. **145**(5): p. 787-99.
100. Zhang, Q.C., et al., *PrePPI: a structure-informed database of protein-protein interactions*. Nucleic Acids Res, 2013. **41**(Database issue): p. D828-33.
101. Petrotchenko, E.V., et al., *Analysis of protein structure by cross-linking combined with mass spectrometry*. Methods Mol Biol, 2014. **1156**: p. 447-63.
102. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies*. Nat Meth, 2014. **11**(11): p. 1114-1125.