INTEGRATED SIGNATURES OF DISEASE
USING NETWORK METHODS

By

David L Gibbs

A DISSERTATION

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

November 2012

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the PhD Dissertation of

David L. Gibbs

"*INTEGRATED SIGNATURES OF DISEASE USING NETWORK METHODS*"

Has been approved

_____

Shannon McWeeney, PhD

_____

Beth Wilmot, PhD

_____

Larry David, PhD

_____

David Maier, PhD

_____

Richard Smith, PhD

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Acknowledgments

This dissertation would not be possible without a titanic, monumental, *tremendous*, universe-sized amount of help and support from my advisor and mentor, Shannon McWeeney. I am truly in her debt.

I'm grateful for the counsel of my committee who never failed to steer me in the right direction. Untold, vast amounts of grammatical and editing help were provided by David Maier, who also always had great suggestions about figures and better ways to explain things, as well as several key insights that strengthened this work tremendously. Also, thanks goes to Dick Smith who was always enthusiastic regarding the utility of the work. Thanks to Beth Wilmot for taking on the exam chair with such short notice and Larry David for jumping in with no hesitation.

Institutionally, I've been fortunate in receiving a great education, from an understanding and smart faculty, as well as an organized and responsive staff. I am thankful for the support I have received with a fellowship from the National Library of Medicine. Thanks also to the Biodev group and the PhD meetings where I was able to test out presentation materials and get useful feedback.

Of course, finishing graduate school would not be possible if it were not for the unwavering support of my family. Especially my parents who always encouraged dreaming and doing.

Last, and most importantly, I am so thankful for the love of my wonderful wife, Tara, and darling daughter Beatrix. *They* are the light of my life. Thank you!

Abstract

This work develops the methods necessary for deriving integrated network signatures of disease with an application in systems biology of infectious disease.

Co-expression network models help identify important biochemical pathways, biomarkers and targets for research, but they typically focus on gene expression. In this work, co-expression network methodologies are extended to proteomics and applied to data derived from mice infected with either influenza or SARS-CoV.

Although, peptide-level co-expression networks are promising, the determination of parent proteins is difficult, especially due to degenerate peptides mapping to multiple proteins. Protein inference attempts to solve this problem. While there are a handful of models, none have been purposed towards high-throughput tag-based proteomics. As such, a new model for protein inference is developed, representing a new approach to the problem.

Lastly, two methods of data integration are explored. First, using a new application of correlated factor analysis, and second, by joining independently constructed co-expression networks. These integration methods allow the discovery of new integrated network signatures of disease and suggest new paths for biomedical research.

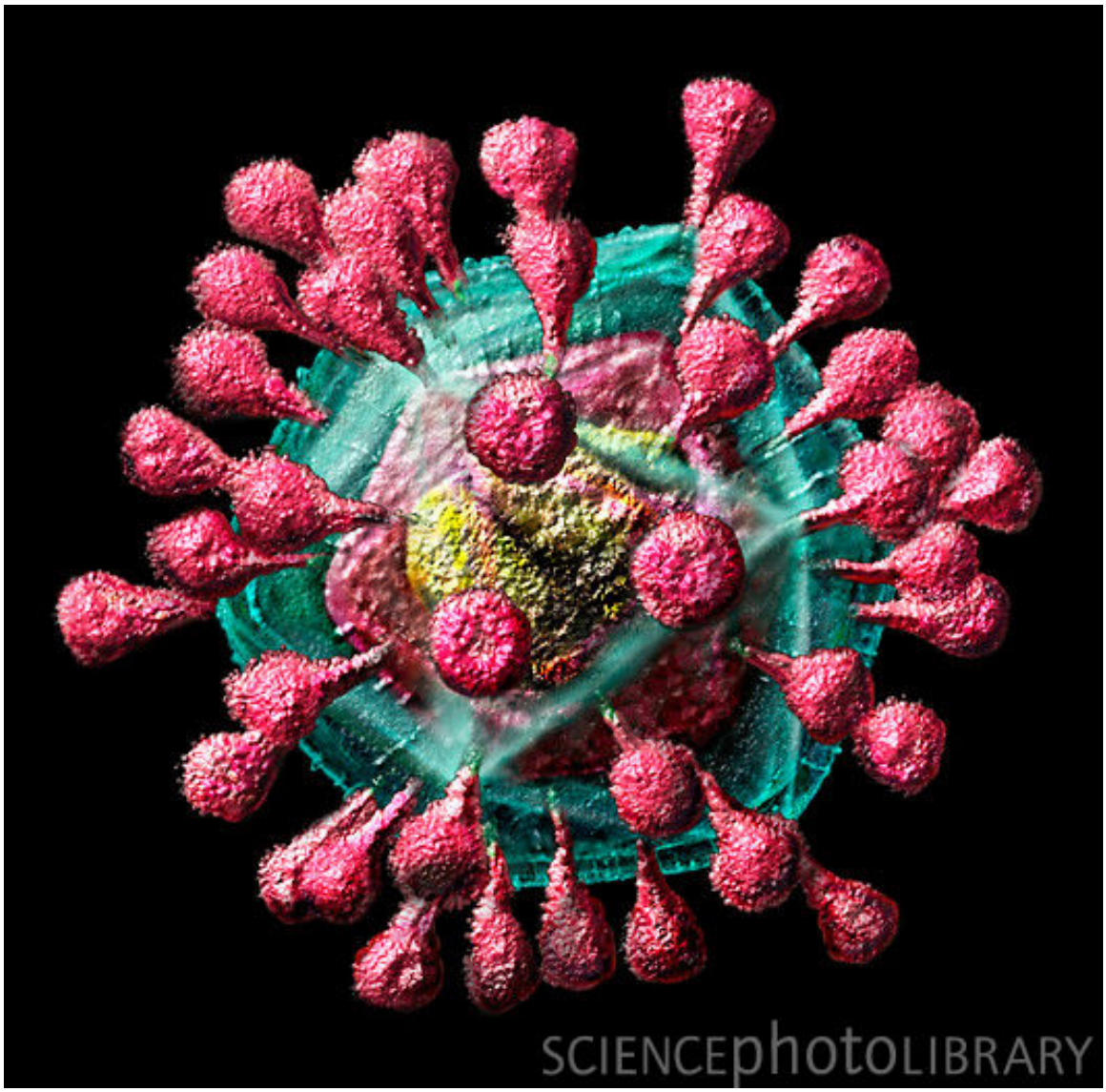

Figure 1. The SARS-CoV corona virus (SciencePhotoLibrary).

1. Introduction

Disease signatures, or biomarkers, are sets of biological components such as genes or proteins, that together can be used for prediction of clinically relevant phenotypes. The uses are diverse. For example, phenotypes can describe disease subtypes or pathological severity.

In the past, most signatures were derived using single data types, such as gene expression microarrays, and were subsequently found to be poor predictors. However, it is becoming apparent that the integration of multiple data types and the use of network models will vastly increase the accuracy and robustness of signatures (Sung, Wang, Chandrasekaran, Witten & Price, 2012).

Weighted Gene Co-expression Network Analysis (B. Zhang & Horvath, 2005) uses gene expression data to discover disease signatures. These methods are extended to proteomics and it is shown that the resulting peptide networks are scale-free, modular, and associate with clinically relevant phenotypes.

In order to integrate peptide networks to gene expression data, protein inference must be performed. A new method of protein inference is developed for high-throughput tag-based proteomics (Smith, et al., 2002a). This model is compared to an established method and shown to make reliable predictions on both simulated and real data.

With annotated peptide networks, relationships to independently constructed transcript networks are discovered, uncovering integrated signatures of disease. Joint co-expression network integration is compared to another method of integration called Correlated Factor Analysis (CFA). This work supports the idea that multi-omic signatures are feasible, robust, and biologically informative.

This work is applied to studying the host response after infection by either SARS-CoV and influenza viruses (Ksiazek et al., 2003; Shortridge et al., 1998) (Figure 1).

Aim1

A novel method for de novo protein co-expression network analysis is developed.

Hypothesis: Peptide data is useful for co-expression analysis, and inferred sub-networks have focused functional profiles.

1. Protein co-expression networks are constructed using proteomics data from influenza and SARS-CoV infected mouse studies, and a human cohort study.
2. Network significance is evaluated using permutation testing.
3. The network is evaluated by identifying strongly connected central "hub" peptides, and functional profiling on subnetworks.
4. Protein co-expression networks are also evaluated by comparison to known protein-protein interaction networks using the Mouse Protein-Protein Interaction database.

Aim 2

A novel method of protein inference for tag-based proteomics is developed.

Hypothesis: a method based on ideas from network flow provide a solution.

1. Simulated data sets are generated using parameters reflecting real data. The simulated data is compared to real data in terms of topological properties on graphs connecting peptides to proteins.
2. A method of protein inference is developed tailored to high-throughput tag-based proteomics.

3. The method is compared to an established method for validation on both simulated and real data.

Aim 3

Integration of gene and protein data is demonstrated using two methods representing “early” and “late” methodologies. Hypothesis: Integrated analysis allows the discovery of relationships between previously unrelated genes and proteins relevant to disease.

1. “Early” data integration is accomplished by extending correlated factor analysis resulting in immediately integrated gene and protein data.
2. The most variable genes and genes that match the protein data are used for gene co-expression network construction.
3. “Late” data integration is accomplished by joining independently constructed peptide and transcript networks. The subnetworks are selected for integration by a combination of member overlap, eigenvector correlation, and phenotype correlation.
4. Integrated networks obtained using both methods are evaluated by comparing functional profiles.
5. An integrated signature for SARS-CoV infection is found using joined co-expression modules.

2. Background

Importance

The determination of biological components and their relative importance is important for understanding the disease process. The systems involved in the progression of disease are complex, and have proven difficult to analyze. Network methods have shown promise in decomposing complex problems, and have been applied to a broad variety of biological studies. In particular, systems biology is focused on learning about complex systems using network methods as a primary tool.

However, networks composed of a single data type do not tend to reflect the complex biological networks observed in living things. Although difficult, in order to better model and represent the disease state, it is important to incorporate a variety of data types into our models. A way forward is to focus on technologies that allow high-throughput “Omic” level measurements, such as transcriptomic and proteomic measurements. The combination of network methods and ‘omic data integration will lead to more clinically relevant models of disease.

Globally, millions of individuals are affected by respiratory viral infection every year (Dixon, 1985). However, the complete mechanism of viral pathology is still not clearly understood (Peiris J, n.d.; Safronetz, et al., 2011b). For SARS, the most common form of damage to the host is lung pathology, and in particular diffuse alveolar damage. Influenza pathology is thought to result from a

combination of a disregulation of immune response pathways (de Jong et al., 2006) and high viral replication rates (Hatta et al., 2010). It is quite likely that different respiratory viruses share some of the causes of pathology.

The result of this work could lead to new targets for further investigation, potentially leading to new therapies.

Critical Barriers

Proteomics provides valuable information. Direct measurements of protein levels are important since this information cannot be inferred reliably from microarray data (Nie, Wu, Culley, Scholten, & Zhang, 2007). However, for a number of reasons, there is a large amount of missing data and significant difficulty in the inference of proteins in a biological mixture. In addition, there remains a distinct lack of de novo network methods applicable to proteomics.

We should be able to improve our understanding of biology by integrating multiple sources of information such as gene expression and proteomics. However, data integration has proven difficult (Joyce & Palsson, 2006), and there are a limited number of successes with virtually no examples in virology. To date, no integrated models exist for host response in viral infection.

Understanding the range of pathogenicity among viruses as a function of host response is important for the improvement of public health (Mauad et al., 2009). The mechanisms behind tissue damage are not clear. Simple measures such as viral titer do not predict the level of damage (Safronetz, et al., 2011a). It is therefore crucial to have a clear understanding of how the host system,

including the innate and adaptive immune systems, responds to viral infection and how it relates to physiological damage (Hatta et al., 2010).

Application to systems biology of infectious disease

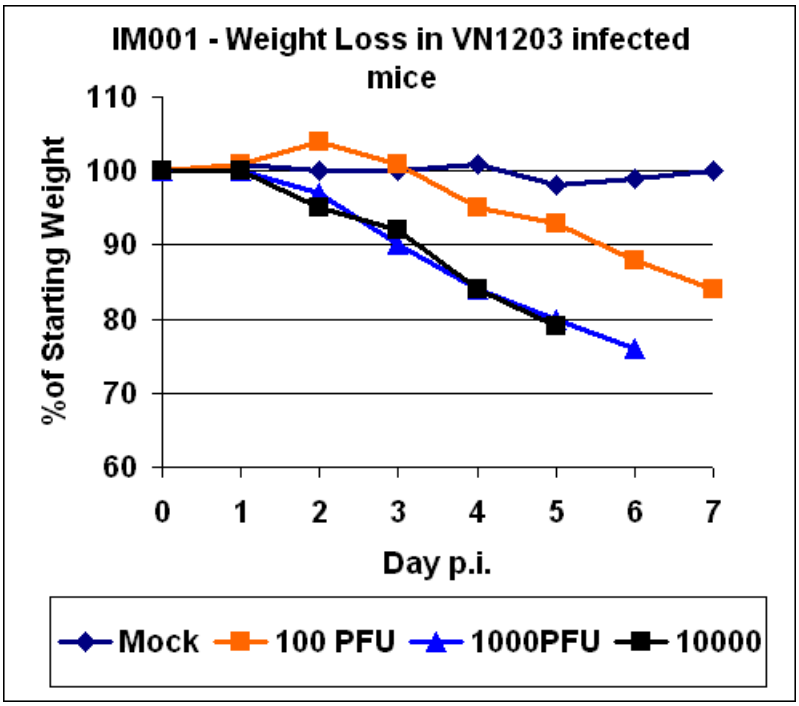

Figure 2. Weight loss of mice infected by influenza. Mice were given varying dosages of influenza virus and host response time series data was generated including both transcript and proteomics data. Significant amounts of weight loss was observed for infected mice. The effect of larger doses is easily seen.

The work in this proposal is applied to publicly available data from an NIAID Systems Biology project (Systems Virology Center, NIAID Contract No. HHSN272200800060C) where mice are inoculated with differing dosages of SARS or influenza and measurements are taken while the infection progresses (Figure 2). Both viruses are enveloped RNA respiratory viruses that are harbored in animal reservoirs (Wendong Li et al., 2005b). The viruses infect epithelial tissue in the respiratory track (Fang Li, Li, Farzan, & Harrison, 2005a).

The SARS virus is the mouse adapted MA15 strain (Roberts et al., 2007). The animals used in the study were 20 weeks old C57bl/6 mice. SARS doses were given at 10^2, 10^3, 10^4, 10^5 PFU. Each day, for 7 days, weight is measured using combined groups of five mice. On days 1, 2, 4, and 7 RNA and protein levels were measured, and pathology samples taken. The Baric Laboratory at the University of North Carolina, Chapel Hill, conducted all SARS infections and the accompanying phenotypic characterizations. The Katze Laboratory at the University of Washington produced the gene expression data using the Agilent 4x44 mouse microarray platform. Protein data was produced by the Pacific Northwest National Lab Proteomic Research Resource for Integrated Biology using LC IMS/MS (Baker et al., 2010) for the tag database creation and LC-MS for the sample analysis.

Proteomic data comes in the form of peptide observations stemming from fragmented proteins. A mass-and-time tag database is used to identify peptides contained in sample mixtures (Ksiazek et al., 2003; Shortridge et al., 1998; Smith, et al., 2002a). Technical replicates are averaged. Missing values are recorded when no data was present for any technical replicate.

Many transcripts in the microarray annotation do not have a matching protein in the tag database. To further characterize the differences between the microarray data and the tag database, the goProfiles package from Bioconductor is used. All GO term nodes on the biological process tree levels 2, 3, and 4 were examined. Associated Entrez gene identifiers for the microarray and proteins in the tag database were mapped using Bioconductor annotation databases.

For each GO term, the associated Entrez IDs were compared to the microarray and proteomics database. The two comparisons resulted in a set of GO terms unique to the microarray, a set of GO terms for the intersection between array and proteomics database, and a set unique to the proteomics.

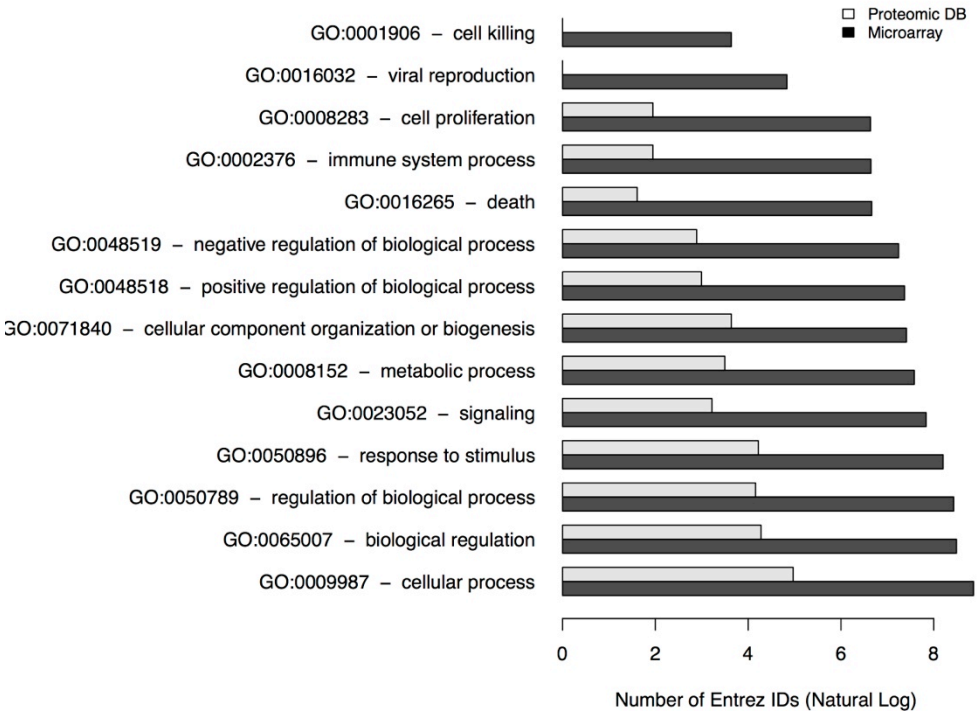

Figure 3. Comparison of microarray and tag database GO BP term coverage. For both the microarray annotation and the mass tag database, which describes all possible peptide identifications, the number of entities mapping to high-level GO terms is compared on the log scale. In some categories, including cell killing, and viral reproduction, the proteomics database is noticeably absent. In all other categories, proteomic coverage is lacking to some degree.

As shown in Figure 3, the proteomic data has reduced functional category coverage. However, with respect to the microarray, a large portion of the GO term coverage comes from genes unique to the array. For example, considering level 2 BP GO terms, the coverage provided by the genes unique to the microarray is often almost twice as great as those genes in the intersection between microarray and proteomics database. Therefore, using the intersection

of EGs between microarray and proteomics database greatly diminishes the GO term coverage.

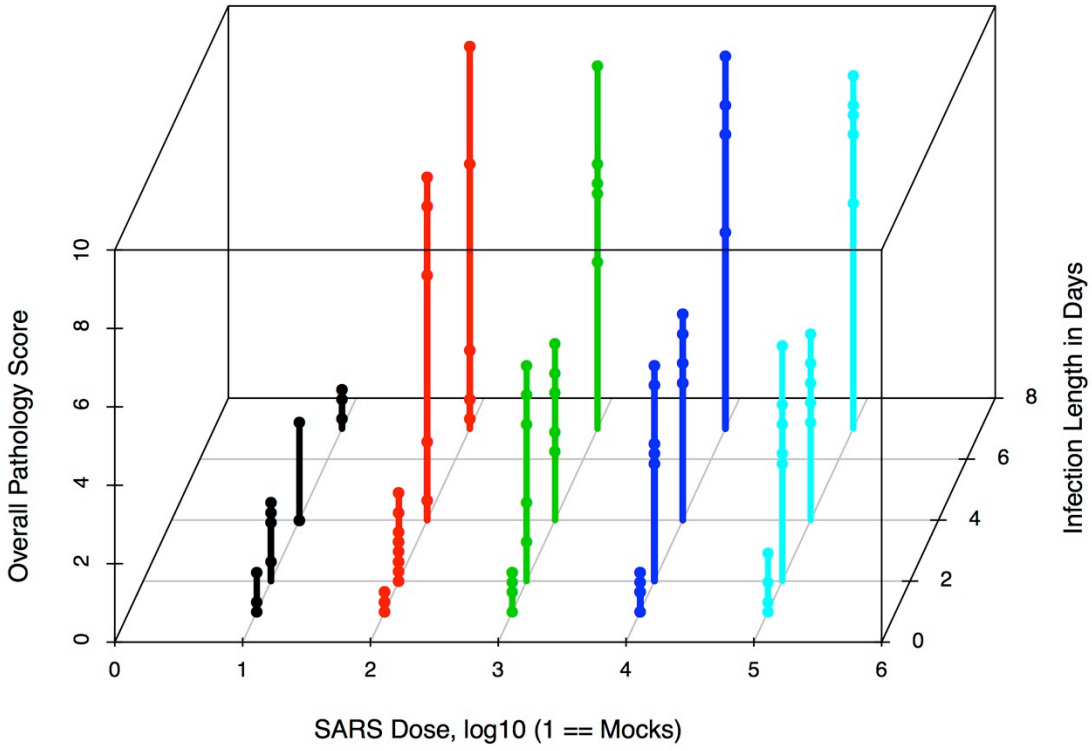

Figure 4. Overall pathology score by dose and day. Comparing the overall pathology score for each mouse infected with the SARS-CoV virus shows increasing amounts of lung damage with time (moving backwards into the box). While the three higher dosages follow a similar trend, the 10^2 profile appears quite different.

The phenotypic information associated with this project is provided by a pathologist who has evaluated the lungs of each subject. Measurements for a wide range of variables are provided including features such as airway constriction, inflammation, airway inflammation, debris, denudation, the state of the vasculature, and whether signs of pneumonia are observed.

Conveniently all of the variables are combined into a single measure entitled “Overall Total Score”, which correlates very well with most other variables as shown in Figure 5.

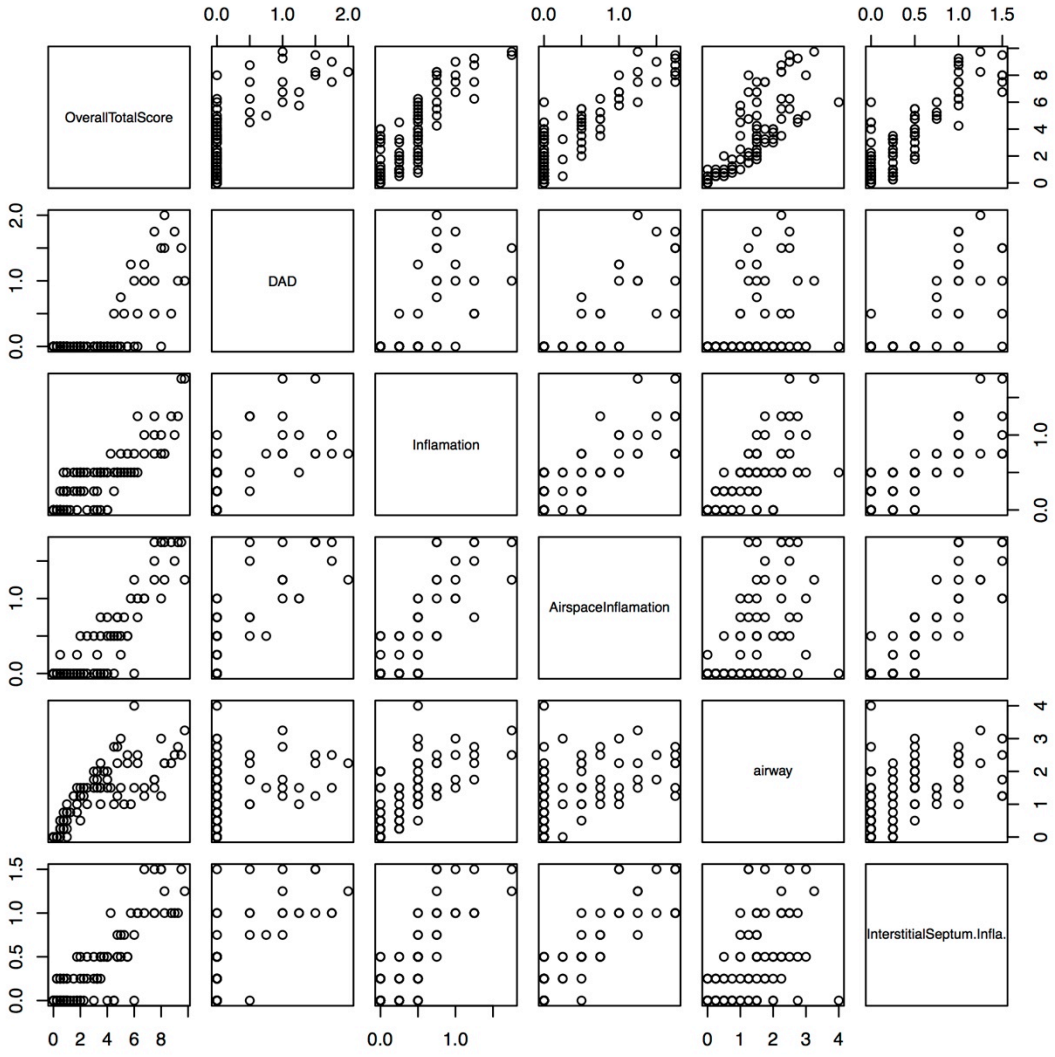

Figure 5. Comparison of the overall pathology score to inflammation related phenotypes. Although Diffuse aveolar damage (DAD) does not start until the overall score is above 4, the other variables correlate strongly implying that the overall score is a good surrogate for mouse lung health.

When the lung pathology is observed as the Overall Total Score by time and by dosage, we see that the trend is strongest by time, and less difference is seen by dosage (see Figure 4). All mice that received a dose of SARS ended up

with observable lung pathology. Strangely, there are two points from mice that received a low (10^2 PFU) dose, but by day four had an extremely strong response to virus, which is seen in Figure 4.

Aim 1 Background

It has been observed that many naturally occurring networks can be described as having a "scale-free" topology (Albert & Barabási, 2002; B. Zhang & Horvath, 2005) (Ravasz & Barabási, 2003; Smith, et al., 2002a). The Internet, social networks, and gene regulatory networks all show this fundamental organization. A scale free topology describes the node connectivity within the network. If we consider a network to consist of a set of nodes connected by edges, then in a scale-free network, a very small number nodes have the highest number of connections whereas most nodes show very modest connectivity. In fact the distribution of connectivity is linear on the log scale (Barabási & Oltvai, 2004; Dixon, 1985).

Researchers began to notice similar patterns in gene expression profiles (Bergmann, Ihmels, & Barkai, 2003; Wendong Li et al., 2005b; Segal, Friedman, Kaminski, Regev, & Koller, 2005; Shai et al., 2003). Given some perturbation, such as viral infection, certain groups of genes are co-regulated and thus co-expressed. When co-expression patterns are conserved across species, it implies a possible functional relationship (Fang Li et al., 2005a; Stuart, Segal, Koller, & Kim, 2003). A connection between co-expression networks and scale-free topologies has been observed (Carter, 2004; Fraser & Marcotte, 2004; Satija & Lal, 2007; B. Zhang & Horvath, 2005). The weighted gene co-expression

network analysis (WGCNA) technique has been thoroughly developed (Langfelder & Horvath, 2008; Langfelder, Luo, Oldham, & Horvath, 2011; Ai Li & Horvath, 2009; Peiris J, n.d.; Safronetz, Rockx, Feldmann, Belisle, Palermo, Brining, Gardner, Proll, Marzi, Tsuda, et al., 2011b; Yip & Horvath, 2007) and has been applied to studying cancer (de Jong et al., 2006; H. L. Kim, 2004; Shai et al., 2003), evolution (Carlson et al., 2006; Hatta et al., 2010; Oldham, Horvath, & Geschwind, 2006; Oldham et al., 2008), cardiac disease (Dewey et al., 2011; Mauad et al., 2009), and applied to mouse systems (MacLennan et al., 2009; Safronetz, et al., 2011a).

However, de novo network methods have seen very little application in Proteomics. There have been attempts to validate yeast protein-protein interaction networks using gene co-expression, but they have not been completely successful (Bhardwaj & Lu, 2005; Hatta et al., 2010; Tirosh & Barkai, 2005).

Bing Zhang has previously written about protein co-expression analysis (Nie et al., 2007; Bing Zhang et al., 2006). The work focused on the yeast proteome, and was centered on computing abundance correlations among a sizable total of 1,119 proteins. These correlations were clustered and it was found that distinct, biologically significant clusters formed in response to cell perturbations. This work is very important in showing the potential of quantitative proteomics co-expression studies, but significantly leaves out the advent of scale-free topologies and lacks the use of more modern network building and clustering methods.

Since the 2006 Zhang study, gene co-expression networks have matured. This work aims to bring recent de novo network methods to proteomics, allowing for functional subnetwork discovery, potentially assigning function and relationships to proteins with unknown characteristics and identifying new relations between proteins.

Aim 2 Background

The protein inference problem attempts a prediction as to what proteins are likely be responsible for generating an observed set of peptides during the course of a mass spectroscopy experiment (Joyce & Palsson, 2006; Nesvizhskii & Aebersold, 2005, Casadevall & Pirofski, 1999; Huang, Wang, Yu, & He, 2012; Serang & Noble, 2012a). Typically when people discuss this problem, they are referring specifically to *liquid chromatography tandem mass spectroscopy* (Aebersold, 2003; Roberts et al., 2007). This type of experiment takes a biological mixture containing some variety of proteins, and digests (fragments) the proteins, using an enzyme such as trypsin, into a mixture of peptides, which are short sequences of amino acids (Baker et al., 2010; Mihályi & Szent-Györgyi, 1953; Northrop, 1922). This peptide mixture is injected into the instrument where liquid chromatography separates the peptides by hydrophobicity (Yoshida, 2004). Tandem mass spectroscopy involves two stages. The first stage of mass spectroscopy produces spectra for the intact peptide mixture most recently exiting (eluting from) the chromatographic column. Then a sampling of spectra is taken, and the peptides linked to said spectra are sent to the second stage that involves further fragmentation resulting in what is termed the *fragmentation*

spectra. Peptide sequences can be inferred from fragmentation spectra (Noble & MacCoss, 2012) using special software such as SEQUEST (Eng, McCormack, & Yates, 1994).

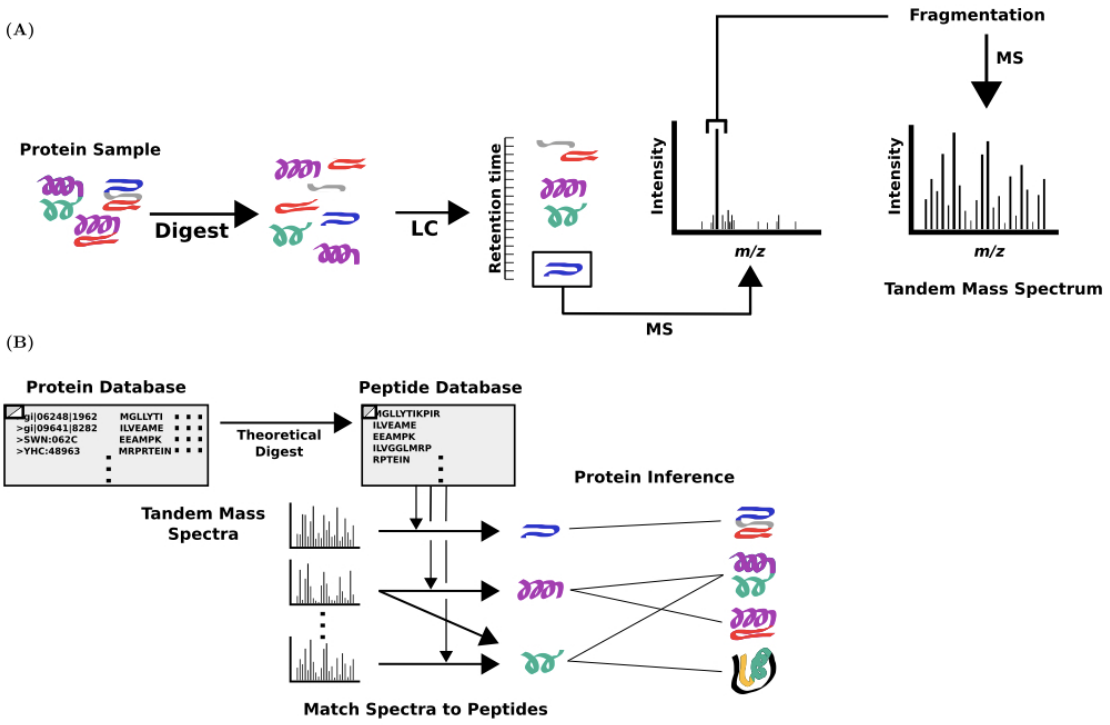

Figure 6. Process of tandem mass spectroscopy. Figure taken from Serang 2012. In (A), proteins are digested into peptide fragments, and separated chromatographically, resulting in peptide spectra, and fragmentation spectra. Then in (B) using a protein database, and an in silico protein digest, spectra from (A) are matched against simulated spectra from the peptide database. A tripartite graph is constructed mapping spectra to peptide sequences which are then mapped to proteins.

The expected spectrum for a given peptide sequence can be predicted computationally. The search algorithm attempts to match the observed spectra to what is expected given a particular sequence.

Currently, all protein inference models are based upon the output of tandem mass spectroscopy. A tripartite graph connects the spectra to peptide sequences based upon a protein database (prior knowledge) and a theoretical protein digest (see Figure 6). The peptide sequences are then connected to proteins based on sequence matching. The goal in protein inference is to select a

set of proteins such that the number of false positives is minimized while maximizing the number of true positives (sensitivity). This problem is extremely difficult due to false edges between spectra and peptide sequences, and peptide sequences that map to multiple proteins. To illustrate the depth of the problem, some proteins are indistinguishable from alternate proteins since all the constituent peptides are degenerate.

There are primarily three methods of protein inference. The first involves *set-cover* methods. The simplest method would be to remove any degenerate peptides, and to take proteins that have at least one or two unique supporting peptides. This method throws out a large portion of data and greatly limits the diversity of proteins identified, making it less than satisfactory.

Another set cover method, IDPicker, attempts to find the minimum set of proteins necessary to explain the observed peptides (Bing Zhang, Chambers, & Tabb, 2007). The results are considered to be a conservative estimate. This computationally intense algorithm is solved using a number of heuristics. One such method for improving the run time of these algorithms is through protein bundling, where proteins with similar sets of peptides are combined into a single meta-protein, reducing the graph size considerably. Another method to improve tractability is to partition the graph of peptides mapping to proteins into smaller, individual graphs, requiring less work each.

The second class of protein inference methods are termed *iterative* methods, of which ProteinProphet is one. Again, this algorithm is operating on the graph of peptides mapping to proteins. In this case, an expectation-

maximization-like problem is solved where iteratively peptide identification scores (the quality of the match from spectra to peptide sequence) are used to update posterior probabilities on proteins. The posteriors are then used to update weights on the peptide-protein graph edges, and the algorithm continues until convergence. The problem with ProteinProphet is that, as it has been observed, very high posteriors may be assigned to a protein with a single, high scoring peptide, or with several poor scoring peptides (Serang, MacCoss, & Noble, 2010). Other protein-inference methods such as Scaffold, PANORAMICS, and EBP also fall into this category (Feng, Naiman, & Cooper, 2007; Price et al., 2007; Searle, 2010).

Lastly, Bayesian models make up the third type of protein inference. These probabilistic models attempt to model the physical process of mass spectroscopy. I will focus on two methods, MSBayes, and Fido (Serang et al., 2010; Serang & Noble, 2012b). MSBayes is interesting in that peptide detectability was incorporated into the model. Detectability models attempt to predict which peptides are easily measured, and which are not (Yong Fuga Li, Arnold, Tang, & Radivojac, 2010b). This contribution is novel and interesting, although the model used for prediction involved hundreds of parameters, and was not necessarily useful for other mass spectroscopy platforms. More practical and useful for this work, Webb-Robertson et al. showed that the mass tag database could be used to train the detectability classifier (Webb-Robertson et al., 2008).

Fido, on the other hand, appears as a relatively simple probabilistic method with just three parameters. Although it appears simple, it is equivalent to "computing every possible set of present proteins and evaluating their net contribution." As the number of proteins grows, the collection of all possible protein sets grow exponentially, which requires some inventive mathematics to remain tractable. In deriving the graphical model, seven assumptions are made to aid model clarity. Describing the three model parameters, first there is a probability, alpha, that a peptide is emitted from some given protein. Second, there is a probability, beta, that a peptide (matched from spectra) is in fact noise. Finally, each protein has an identical prior probability, gamma. This model is shown to perform very well compared to other methods, including ProteinProphet, which is in the Serang review on protein inference methods. Other methods here include the Hierarchical statistical model (Shen, Wang, Shankar, Zhang, & Li, 2008), the nested mixture model (Q. Li, MacCoss, & Stephens, 2010a).

Aim 3 Background

Recently, a large amount of work has been produced focused on the integration of transcript and proteomic data (B. Cox, Kislinger, & Emili, 2005; Daemen et al., 2008; Fagan, Culhane, & Higgins, 2007; Joyce & Palsson, 2006; Kellam, 2001; Torres-Garc i a, Zhang, Runger, Johnson, & Meldrum, 2009; Waters, Pounds, & Thrall, 2006).

Some researchers have been working towards creating models that allow the prediction of protein levels based on the transcript level (Fagan et al., 2007)

which allows researchers to generate values for missing proteins, producing more complete data sets. However, these models have not performed well, partly because of the complexity of the transformation of transcript to protein.

In this project, two methods of integration will be explored: Correlated Factor Analysis (CFA) (C. S. Tan et al., 2009) and a novel approach: the integration of gene and protein co-expression networks.

The “early” method of integration will be based on Correlated Factor Analysis a new method with deep roots in statistics (Browne, 1980; Ihara & Kano, 1986), psychometrics (Tucker, 1958) and climatology (Salim & Pawitan, 2007). CFA has recently been extended to integrate transcriptomic and proteomic data (C. S. Tan et al., 2009). However, CFA has not been widely accepted and its feasibility on larger data sets is not known. In its only such use, 15,918 genes were integrated with 89 proteins. This project uses data that contains over 2,000 proteins, so it remains to be seen if further development and validation is needed for application to larger protein sets.

Summary

Integrated network methods are important for the future of systems biology, and provide a way forward when dealing with increasingly complex biology. Using networks, we are no longer are looking at single gene or protein effects, but instead look at the behavior of the group, which is more in line with how biological systems actually work. When groups of genes are collected in a network module, it relaxes the multiple-testing problem faced when working on

very large sets of single genes. In addition, it is more biologically accurate when our networks contain multiple data types.

Protein co-expression network analysis is a novel development, and is a significant addition to the field. Also, correlated factor analysis provides a valuable tool for data integration. These de novo, derived networks can show us potential biomarkers and novel, important targets for further research. Using these techniques to further the understanding of the host response to viral infection is key in understanding mechanisms of pathology.

3. Protein co-expression Network Analysis

Introduction

In the cell, biological networks are populated with active, working, proteins. Systems biology aims to take a holistic view of activities in the cell, while embracing complexity (Ideker, Galitski, & Hood, 2001) (Tisoncik & Katze, 2010). As systems biology moves forward, models making use of quantitative proteomic data will become increasingly necessary (Kellam, 2001) (Ideker & Sharan, 2008).

However, large-scale quantitative proteomics is still developing and can be challenging and complex in practice (Domon & Aebersold, 2006a; 2006b). Briefly, proteins are digested enzymatically, producing a multitude of peptide fragments. Using liquid chromatography coupled to mass spectroscopy (referred to as LC-MS) the digested (fragmented) mixture is separated and quantified, resulting in a set of peptide identifications with abundance measurements. Peptide identifications are made either by spectral searching or by mapping features to an accurate mass and time (AMT) tag database. Tag databases are previously constructed using pooled samples processed on a tandem MS/MS platform (Zimmer, Monroe, Qian, & Smith, 2006).

Currently, virtually all of protein-interaction networks are constructed using protein-protein interaction (PPI) databases. However, manually curated PPI databases are regularly revised as our understanding of biology grows. PPI databases are typically quite heterogeneous, containing different experiment types and model organisms leading to sparse annotation and a lack of

experimental concordance. In addition, interaction temporality and contextual information is lacking. Coverage, selection bias, and detection bias all remain problems (Bonetta, 2010; Figeys, 2008).

De novo approaches offer an alternative under which prior knowledge of protein interaction is eliminated and replaced by direct measurements of abundance. In this paper, we introduce a novel approach to proteomic network analysis that is applicable to peptide and protein level data. By using methods derived from weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008; B. Zhang & Horvath, 2005), we show that unbiased de novo co-expression peptide networks can be constructed and used for determining potential biomarkers, functional module prediction, and the discovery of important elements of human disease.

Methods

Protein co-expression network construction

Peptide networks are described using a graph with nodes representing peptides and edge weights representing similarity of abundance profiles (Figure 8). Edge weights are calculated using peptide intensities. Although not always representative of absolute abundance, intensity is frequently used to track relative peptide abundance and to infer protein abundance (J. Cox & Mann, 2011). In this work, we do not attempt to rectify situations where proteins are represented by a single peptide or where degenerate peptides map to multiple proteins.

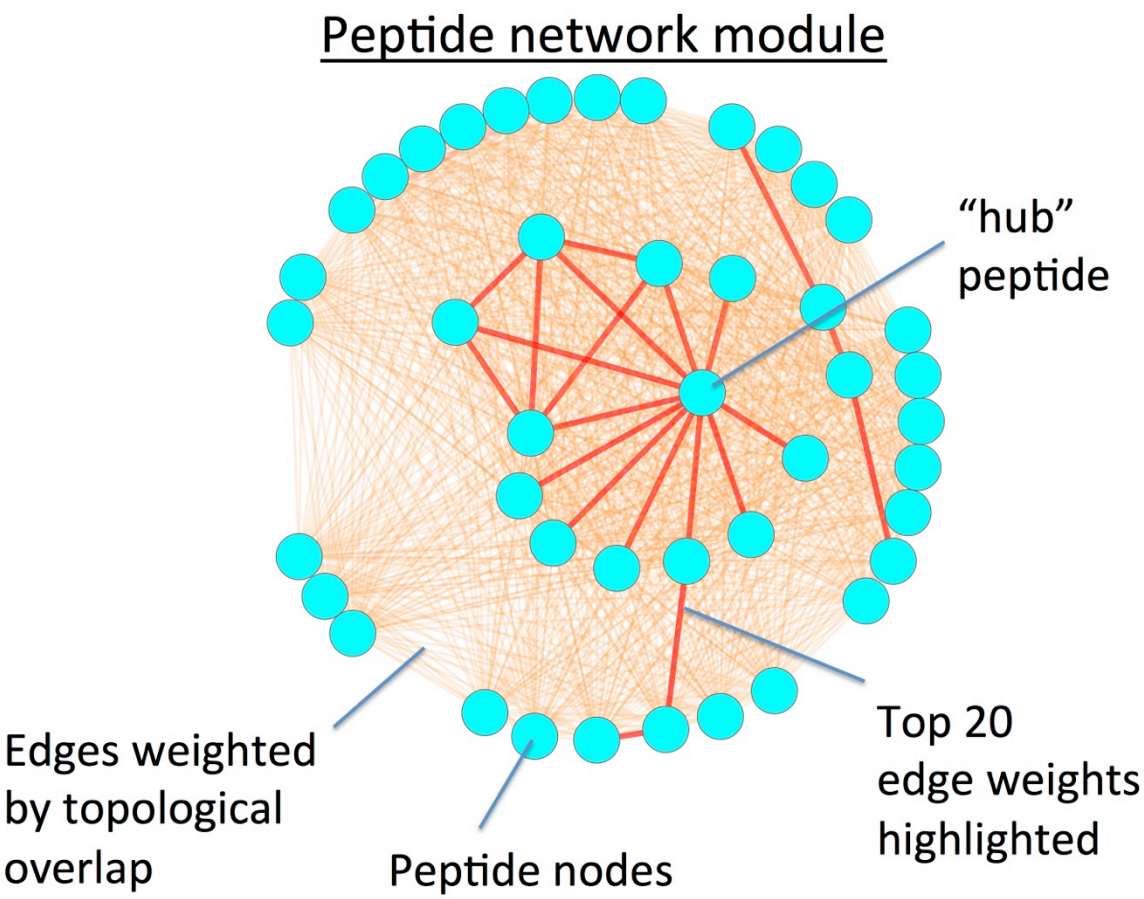

Figure 7. A peptide network module. Peptide networks are decomposed into sub-networks, or modules. Peptide nodes in the network are connected by weighted edges representing similarity in abundance profiles. Hub peptides are highly connected, and can be described as central and important. The most weighted edges are highlighted showing a two sub-graphs.

Construction of the network follows the WGCNA method (Langfelder, 2010; Langfelder et al., 2011; Langfelder & Horvath, 2012; Langfelder, Zhang, & Horvath, 2008). Signed Pearson's correlations are computed pairwise for all peptides in the filtered data set resulting in an adjacency matrix (Mason, Fan, Plath, Zhou, & Horvath, 2009). According to the scale-free criterion, a power (beta) is selected that transforms the distribution of node degrees to log-linear. The scaled adjacency matrix is used to compute the topological overlap matrix (TOM).

Topological overlap is a similarity metric that incorporates information from neighboring nodes, making it more robust to noisy correlations. The TOM is computed as $TOM_{ij} = (l_{ij} + a_{ij}) / [\min(k_i, k_j) + 1 - a_{ij}]$ where l_{ij} is defined as the sum of pairwise products of row i and column j in adjacency matrix a and k_i and k_j are the summation of row i in adjacency matrix a.

Modules, or subnetworks, are composed of strongly connected peptides. Modules are discovered by hierarchical clustering of (1-TOM) using the "average" agglomeration method, followed by branch cutting with the dynamic hybrid treecut algorithm (Figure 11).

Module significance

Similar to previous work (Bankhead, 2010; Iancu, Kawane, et al., 2012b; Oldham et al., 2006; 2008), module significance was examined using permutation testing. Empirical p-values are computed by comparing the mean topological overlap of peptides within a module to random samplings. For a given module with size n, mean edge weight is computed. For a number of trials, t, a sample of peptides is drawn with size equal to n, and the mean edge weight computed. If this value is equal to, or higher than the observed module mean, a count in incremented. The p-value is equal to (counts/t). In this work 10,000 random samples were drawn.

Module Summaries

After assigning peptides to modules, an aggregate module signature is computed. The first principal component, after singular value decomposition on the subset of peptide abundance data by module, is a vector with length equal to the number of samples (an "eigenvector"). This vector acts as an overall

summary of the module. Modules can be sorted according to correlations between eigenvector summaries and biological phenotypes. Additionally, the relative “importance” of any given peptide within a module is found by computing a correlation to the eigenvector summary (called the Kme) (Oldham et al., 2008). Peptides with a large correlation to the eigenvector are said to be more central, and important within the module.

Module Concordance

Concordance among a set of peptides relates to the shared sign of the slope when regressed against a given vector such as time. Our approach to investigating concordance involves constructing protein sub-networks, initially as “all-to-all” networks. After applying a topological overlap threshold, edges start to fall away. This pruning results in a set of disjoint connected components. To examine whether concordant peptides are connected in the network, a linear model is constructed for each peptide using a reference variable such as time or a phenotypic trait. Peptides are classified as increasing (+1), decreasing (-1), or no-slope (0) depending on the adjusted p-value. If a connected sub-graph contains both increasing and decreasing peptides, it is considered a discordant component.

Peptide connectivity by protein

Similar to testing for module significance, the connectivity among peptides mapping to a given protein can be tested by permutation. For each protein with greater than two peptides, the pairwise edge weights are averaged. Then for a number of trials, the same number of peptides are randomly sampled, and the

mean pairwise topological overlap recorded. The empirical p-value is taken as the number of times the random sample has values equal or greater than the observed case divided by the number of trials. This test can also be applied using correlations between the peptides.

Protein-protein interaction enrichment

Permutation testing is used to determine whether a significant amount of PPI edges exist within a module. Co-expression modules are thought to reflect, to some degree, true protein interactions. To examine this premise, we compare the contents of modules with known PPIs. Within each module, peptides were filtered for centrality, and then mapped to proteins. Proteins with any number of mapping peptides are included. Degenerate mappings were allowed. The number of observed PPIs within a module is counted and compared to the number of PPIs in a random module for a number of trials. P-values are computed as before. The PPI databases HPRD (Keshava Prasad et al., 2009) and MPPI (Yellaboina, Dudekula, & Ko, 2008) were used for human and mouse data respectively.

Pathway enrichment

After PPI enrichment tests, significant sets of proteins were collected by module. Querying KEGG with these proteins (Kanehisa, 2004; Kawashima, Katayama, & Sato, 2003), using the R package KEGGSOAP (J Zhang & Gentleman, n.d.), provided a list of potential pathways to investigate by module. For each pathway returned, a hypergeometric test was performed using significant PPIs from the module and other proteins taking part in the pathway. The universe is defined as

the subset of proteins in the mass tag database with roles in known KEGG pathways. P-values are adjusted using the Benjamini and Yekutieli method (Benjamini & Yekutieli, 2001).

Biological functional enrichment

Functional enrichment on modules was computed using the R package GOstats (Falcon & Gentleman, 2007). Similar to computing PPI significance, peptides are mapped to proteins. Proteins are counted once within any module. Proteins mapped to from degenerate peptides are allowed. The universe is defined as all proteins found in the AMT mass tag database, similar to microarray studies. Annotation databases in Bioconductor (2.8) are used for mouse and human annotations. The conditional test is used which tests leaves of the Gene Ontology tree first, removing those mapped entities from the gene list. Then parent nodes are counted if the remainder of list members are significant, providing the most detailed GO terms with the least amount of correlation between terms. P-values were Bonferroni adjusted according to the number of GO terms tested.

Data sources

Quantitative LC-MS data, including two mouse disease studies and one human study, is used. The Thermo Electron Exactive platform was used to generate data. Accurate Mass and Time tag (AMT) databases were developed at PNNL. VIPER (v3.48) was used to align individual samples with the AMT database and identify peptides (Monroe et al., 2007). Identifications have confidence metrics:

the probability for a correct match, the STAC score, and the probability for a unique database match; the *uniqueness probability* (UP) (Stanley et al., 2011).

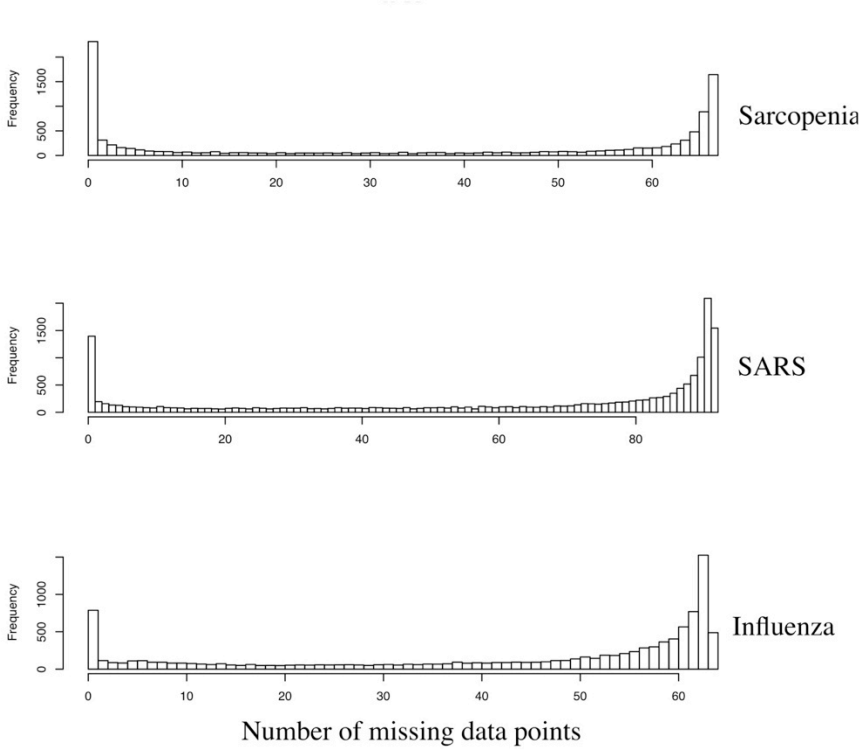

Figure 8. Missingness in proteomics data. Peptides may be observed in all samples, or only a subset. Above, a histogram shows the frequency of missing data points (NAs) per peptide. A peptide that is observed in all samples is represented in the tall bar (has few or no NAs) on the left while peptides present in few or no samples are shown on the right (mostly missing data).

Peptides with STAC scores > 0 and UP > 0 are used. Peptide abundances are normalized by total ion count per sample and log10 scaled. Many peptides are identified in some samples and missing in others; peptides with less than 10% missing data across samples are used (Figure 7). See *http://omics.pnl.gov* for more information.

The infectious disease data came from the publically available data (Systems Virology Center, NIAID Contract No. HHSN272200800060C). We utilize both longitudinal SARS-CoV and influenza mouse studies. This data is generated using C57BL/6J mice exposed to either a mouse adapted SARS-CoV

(MA-15, from the Baric Laboratory at the University of North Carolina, Chapel Hill) or avian influenza virus (A/Vietnam/1203/2004 (H5N1, VN1203), from the Kawaoka Laboratory at the University of Wisconsin at Madison) (Barnard, 2009; Roberts et al., 2007). Measurements took place on post infection days 1, 2, 4 and 7.

SARS control samples include three technical replicates per day. Infected samples are five technical replicates with viral dosages of 10^2, 10^3, 10^4, and 10^5 plaque forming units (PFU) per day. Abundance measurements for 16,890 peptides mapping to 3,277 proteins were recorded. After missingness filtration, 2,008 peptides mapping to 707 proteins remained with 352 proteins associated with a single peptide, and 355 proteins with two or more associated peptides.

Influenza control samples include three technical replicates per day. Infected samples include five technical replicates with dosages of 10^2, 10^3, and 10^4 PFU per day. Abundances for 10,285 peptides mapping to 2,661 proteins were recorded. After missingness filtration, 989 peptides associated with 493 proteins remained with 274 proteins associated with a single peptide and 219 proteins associated with at least 2 peptides.

The human proteomics data (currently unpublished) comes from a sub-cohort of participants selected from a large (N = 6000) longitudinal observational study of musculoskeletal health in older (≥ 65 years) men (MrOS) (Orwoll et al., 2005) (Cawthon et al., 2007). The data used here focuses on the sarcopenia

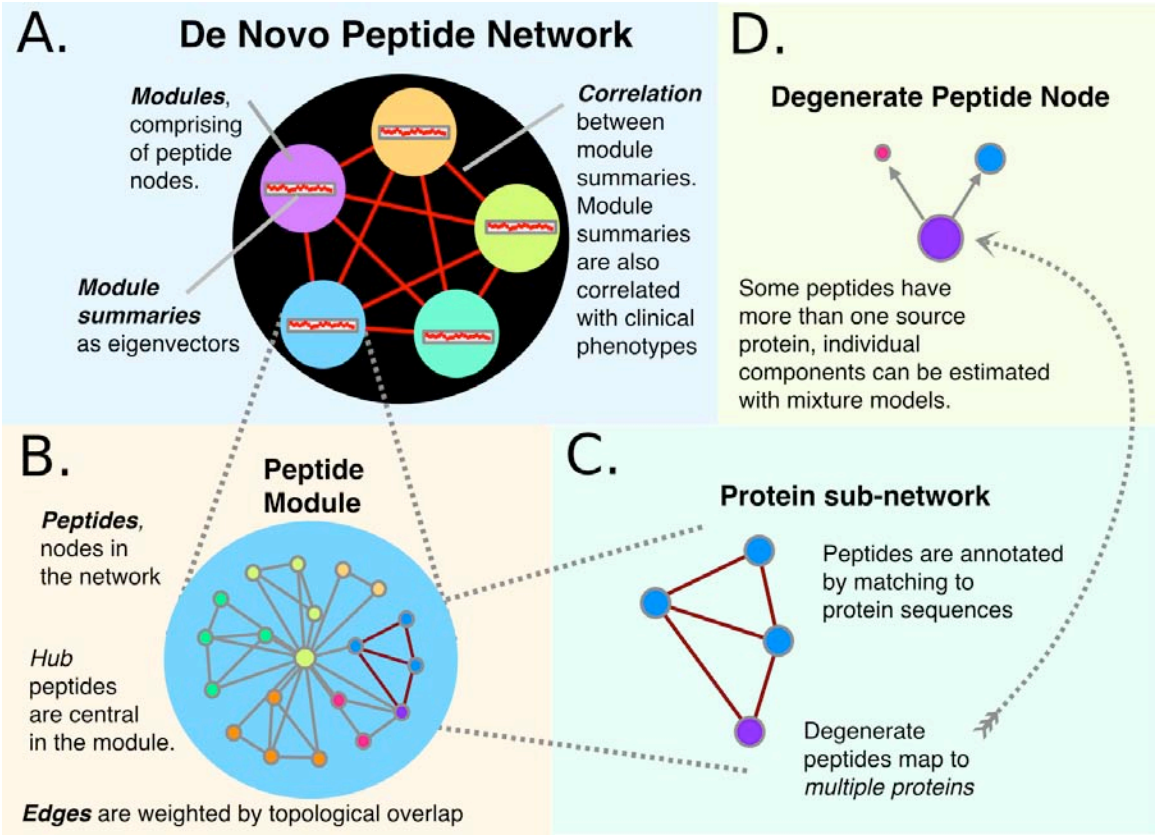

Figure 9. Detail of the peptide network. (A) The peptide network is decomposed into a set of subnetworks, or modules. Module members - peptide nodes - are more connected to other peptides within the module than to those in other modules. (B) Modules can be summarized by taking the first eigenvector from data mapping to these peptides. Taking the correlation of peptides to the module summary gives the relative importance, or centrality of peptides. (C) Taking peptides that map to a given protein defines a protein subnetwork. (D) Difficulties arise when a peptide node maps to multiple proteins, rendering it degenerate.

phenotype (related to loss of lean mass and muscle performance) (Morley, Baumgartner, Roubenoff, Mayer, & Nair, 2001). We used an initial proteomics study that included 68 samples from two carefully phenotyped groups (sarcopenic (N=38) and non-sarcopenic (N=30)) based on lean mass and leg power. Abundances for 10,679 peptides mapping to 1,868 proteins were recorded. After missingness filtration, 2,845 peptides mapping to 685 proteins remained with 505 proteins associated with a single peptide and 180 proteins with at least two peptides.

Results

Peptide networks are approximately scale-free.

Scale-free network topologies have node degree distributions following the power law. There is a continuous range of node degrees, with the fewest nodes having the greatest number of connections (Albert, 2005; Ravasz & Barabási, 2003). We have found that peptide networks share this topology and have biologically informative graph properties similar to those found in gene co-expression networks (Figure 9, Table 1). We have found that, within limits, missing data does not negatively affect the model fit (Figure 10).

Data	Peptides	Proteins	Power	R^2	Slope	MeanK	Modules
Influenza	989	493	15	0.82	-1.31	7.00	6
SARS	2008	707	16	0.76	-1.67	10.8	14
Sarcopenia	2845	685	15	0.81	-1.55	25.22	19

Table 1. LC-MS data is applicable to co-expression network construction methods. R^2 describes the scale-free topology fit. Definitions of mean K: network connectivity using the adjacency matrix.

Considering only significant modules, the SARS network contains 14 modules spanning from 65 to 369 peptides with a mean size of 133.9 peptides. The Influenza network contains 6 modules spanning from 56 to 327 peptides with a mean size of 141.3 peptides. The network dendrogram is shown in Figure 11. The sarcopenia network contains 18 modules spanning 477 to 36 peptides with a mean size of 142.25 peptides.

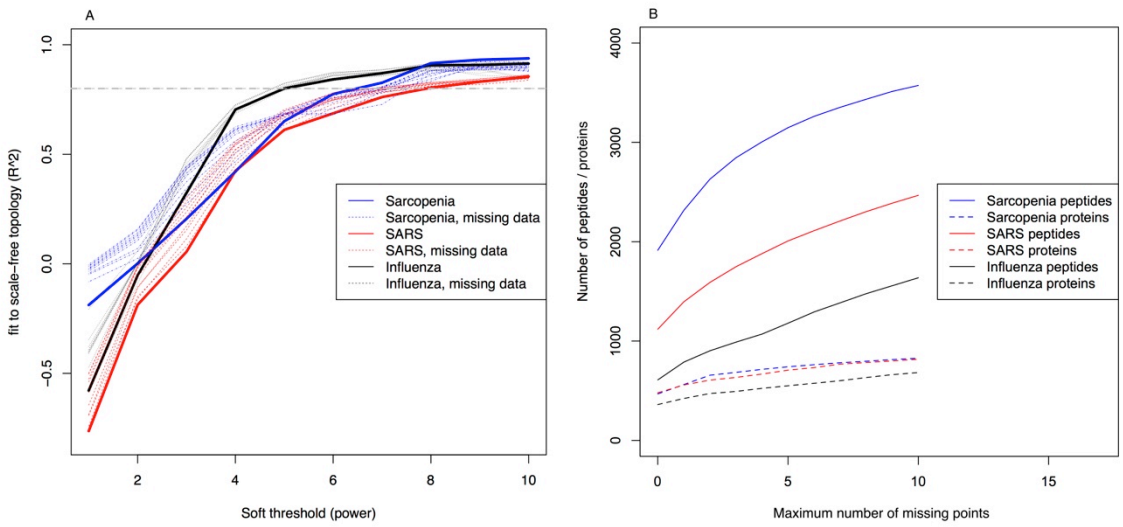

Figure 10. Protein co-expression networks are shown to be "approximately scale-free". As the soft thresholding power, β, grows, the resulting adjacency matrix increasingly fits the scale-free model (A). This trend is robust to missing data. Data sets were generated with varying amounts of missing data ranging from a maximum of one to ten missing data points per peptide, shown here with lighter, broken lines. With more missing data, the total number of proteins and their constituent peptides increases (B). The rate of increase is higher for peptides compared to proteins, improving the peptide-to-protein ratio. For network analysis, it is strongly in our interest to incorporate peptides with missing data, making imputation an attractive option. The return on the number of proteins diminishes with increased peptides.

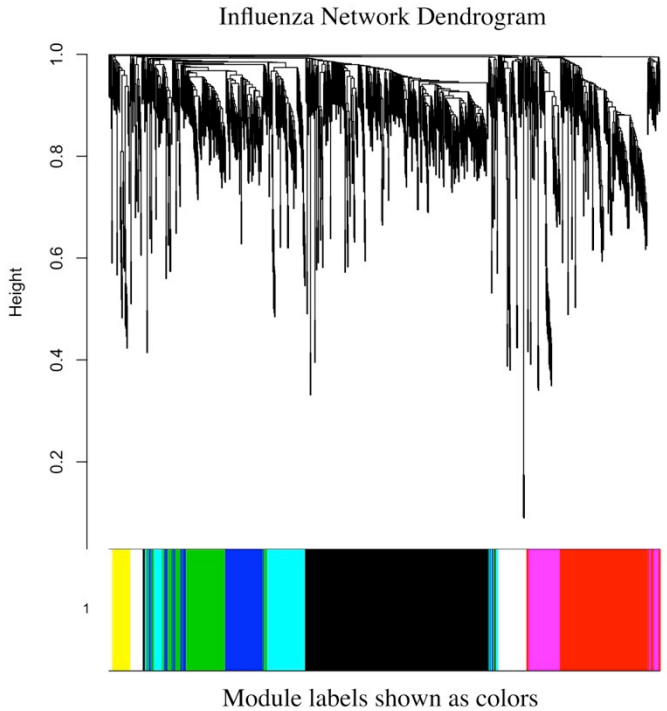

Figure 11. Dendrogram and recovered modules for influenza network. The dendrogram is generated from the distance matrix (1 - TOM), and modules are recovered from branch cutting using the dynamic hybrid treecut algorithm.

Significant modules correlate with phenotypes

Using module summaries, correlation to biological phenotype can guide the discovery of biomarkers (Figure 12). In this work, all modules were found significant with the exception of the sarcopenia network, which had one insignificant module (p-value 0.33).

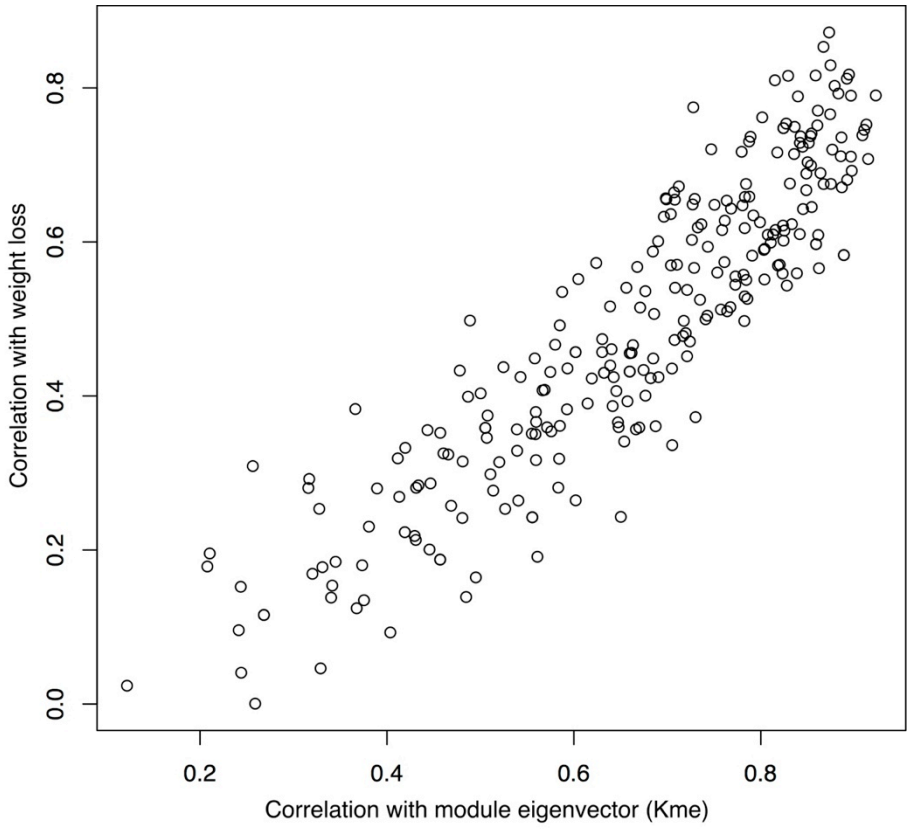

Figure 12. Correlation structure within a module. An illustration from the Influenza data is shown where each point represents a single peptide within module 1. In modules where the eigenvector is strongly correlated with a biological phenotype, an upward trend is observed between the K_{me} of a peptide and the correlation with the given phenotype, demonstrating structural order within the module. Prioritization along these dimensions suggest further experiments.

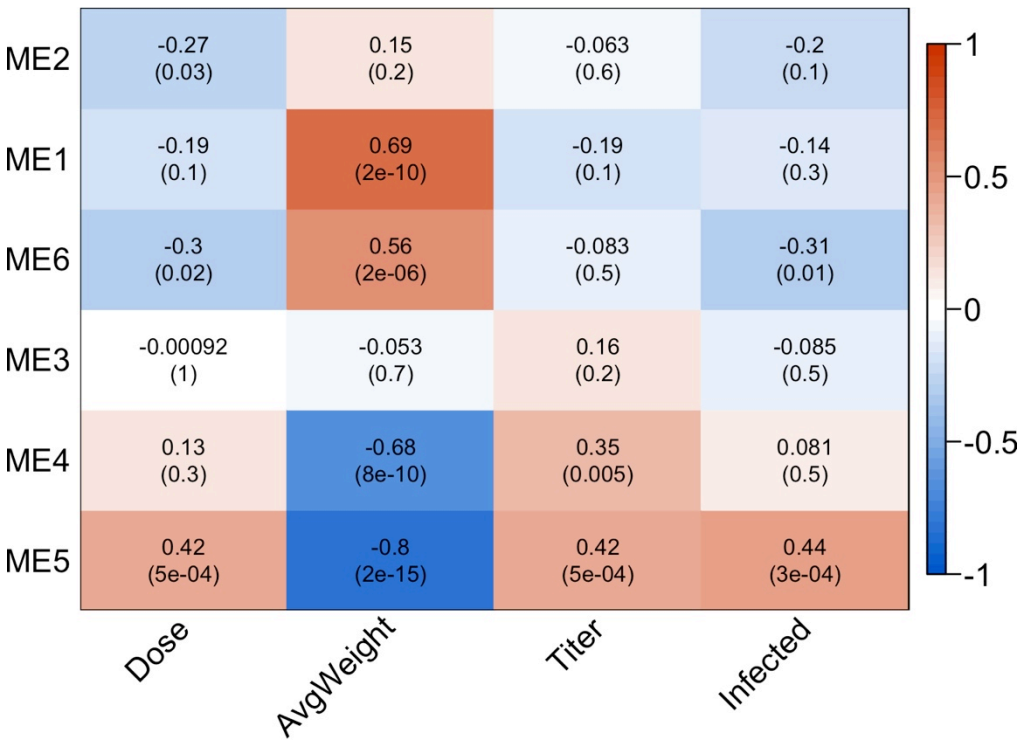

Figure 13. Influenza module-phenotype heatmap. The de novo modules (represented by module eigenvectors, ME), are highly correlated to clinically relevant phenotypes. An illustration from the influenza dataset is shown. In each cell, Pearson correlations are on top, with p-values below. Clear patterns emerge showing positive and negative correlation clusters. Average weight is an important phenotype in this experiment as it clearly showed the infection severity. Above we see modules 1 and 6 correlate positively with each other while modules 4 and 5 correlate negatively.

The influenza network showed strong correlations with average weight loss, an important indicator of severe infection (Figure 13). Two modules showed positive correlation (p-values 2e-10 and 2e-6) and two modules showed negative correlation (p-values 8e-10 and 2e-15), again showing the dichotomous split in phenotype correlations.

In the SARS network, strong correlations with pathological features were observed including diffuse alveolar damage, tissue inflammation, and alveoli parenchyma pneumonia (see Figure 14).

The strongest correlations found were found with time (module 3, Pearson correlation 0.8, p-value 1e-22) potentially relating to infection progress.

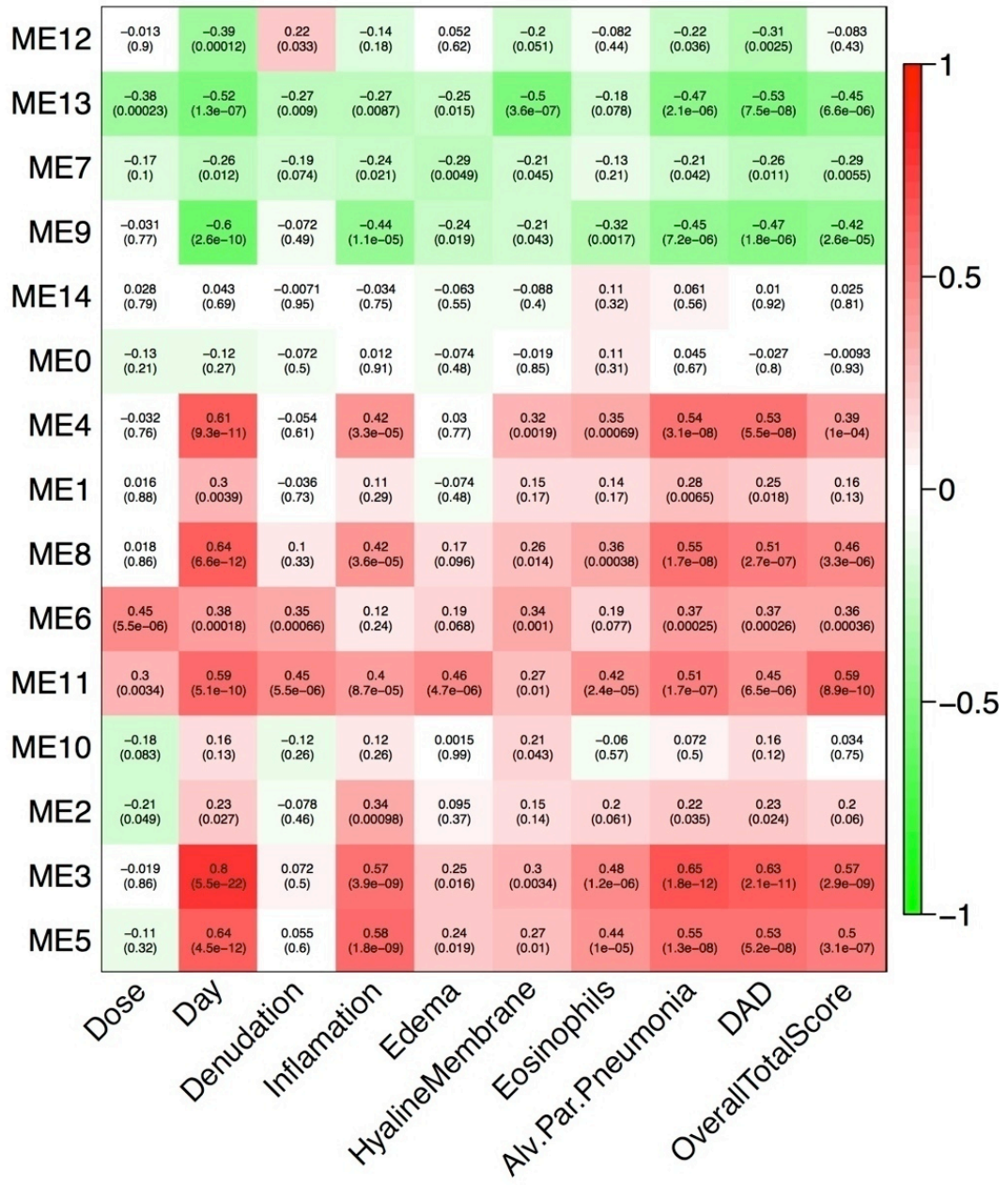

Figure 14. The SARS de novo modules (represented by module eigenvectors, ME) are highly correlated to clinically relevant phenotypes. An illustration from the SARS dataset is shown. In each cell, Pearson correlations are on top, with p-values below. Clear patterns emerge showing positive and negative correlation clusters. As expected, related phenotypes such as airspace inflammation, interstitial septum inflammation, and diffuse alveolar damage (DAD) tend to be correlated in the same direction showing an overarching biological process at work. *Label Key: Alv.Par.Pneumonia: alveolar parenchyma pneumonia, DAD: diffuse alveolar damage, OverallTotalScore: cumulative score calculated by a pathologist.*

It is possible that the relationship of modules correlating in opposing directions is biologically derived. A closer look with eigenvector networks might reveal this.

The sarcopenia network showed weak correlations with sample phenotypes. Several modules correlate with technical variables, indicating that the normalization method did not completely remove systematic effects. To avoid such problems, strong normalization techniques and potentially surrogate variable analysis should be explored.

Peptides are connected by protein

Given a complex biological mixture, a significant problem remains in confidently identifying the protein component. This problem is made worse by the existence of degenerate peptides. We have found that the connectivity of a protein's constituent peptides is far from random, the peptides have strong edge weights within the protein sub-graph. This feature is potentially useful for resolving cases of degenerate peptide mapping, increasing confidence in protein identification.

Topological overlap thresholds eliminate edges with weights below a given point. This filtering tends to fragment the network. However, at a topological overlap threshold of 80% (keeping only edges in the top 20% of all weights), most proteins remain connected (Sarcopenia 84%, SARS 72%, influenza 63%).

To test for significant protein connectivity, we compared the mean topological overlap between constituent peptides and similar numbers of randomly selected peptides. A t-test between the mean protein connectivities and random connectivities shows significant connections between constituent

peptides (Table 2). This result suggests that the network structure should be helpful in resolving degenerate mappings by comparing the connection to alternative proteins. A given peptide, mapping to multiple proteins, could be assigned to a single protein based on a strong connection to it.

Data	Peptides	Proteins	Mean TO	RandomTO	p-value
Sarcopenia	2845	685	0.089	0.004	2.09e-14
SARS	2008	707	0.025	0.004	2.2e-16
Influenza	989	493	0.028	0.005	8.99e-16

Table 2. Protein subnetworks are strongly connected. Given a complex biological mixture, a significant problem remains in confidently identifying the protein component. This problem is made worse by the existence of degenerate peptides. This result statistically shows that the connectivity for a protein's constituent peptides is far from random. Network topology may be useful for resolving cases of degenerate peptide mapping, increasing confidence in protein identification. Results from a two sample t-test between topological overlap (TO) of peptides derived from the same protein versus peptides selected at random.

Connected peptides are concordant

When considering peptide abundance trends, it is desirable to have modules and proteins trending in the same direction (Figure 15). This aspect is important for inferring protein abundance based upon the observed peptides. For example, if we have four peptides mapping to a given protein, where two of the peptides have increasing abundance over time, and two of the peptides have decreasing abundance over time, then this protein shows discordance among its constituent peptides. Quantitative estimates of protein abundance should take account of this phenomenon. Also considering pathway dynamics, peptides mapping to members of a pathway should together reflect the shifts in protein levels that result from biological events. However, an intriguing idea is the possibility that

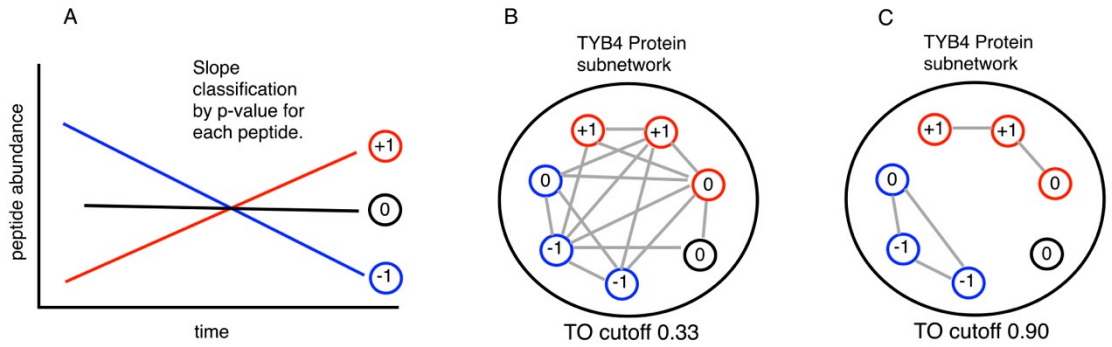

Figure 15. Utility of de novo network inference in resolving peptide level discordance. Protein sub-networks constructed using peptide topological overlap show correlated clusters. Taken together, some proteins show conflicts between constituent peptides, where given a variable such as time, some peptides are increasing in abundance and other decreasing (A). To examine these trends, only proteins with uniquely mapping peptides were used. A protein sub-network was constructed by taking associated peptides, and keeping only edges with topological overlaps in a specified upper quantile (e.g. the upper twenty percent of all topological overlaps). In all three data sources, as the edge threshold is raised, the number of connected components with discordant peptides dramatically decreases (B, C), suggesting that inference of protein abundance can be guided by network topology.

discordance observed among peptides might reflect the activities of different protein isoforms.

Using proteins with unique peptide mappings, 48 of 218 proteins in the SARS data were discordant (Bonferroni adjusted p-value cutoff of 0.1). After applying a topological overlap threshold of 80% (as above), the number of discordant components dropped to 24, and with a threshold of 0.9 dropped to 12. In the sarcopenia network, peptides that were modeled against leg strength provided the most discordant proteins. At the same p-value threshold, 11 of 139 proteins had discordant peptides. After applying the topological overlap thresholds of 0.8 and 0.9, these dropped to 4 and 3 respectively. In influenza, 15 of 115 discordant proteins dropped to 7 and 3. It appears that edge strength is predictive of concordance among constituent peptides for any given protein. These results show potential utility for both protein inference and quantification.

Co-expression modules are thought to be useful for predicting new PPIs and pathway members. Previously, it has been observed that gene expression and protein expression are loosely coupled. It follows that protein co-expression networks should be more useful for making predictions about proteins. Using the HURD and MPPI protein-protein interaction databases, significant interactions were identified in all three experiments. After adjusting for multiple testing, the influenza network had 5/6 modules with PPI enrichment, the SARS network had 12/14 modules enriched and the sarcopenia network had 10/19. It is interesting to note that some modules overlap in terms of mapped proteins. This effect is typically the result of highly similar protein sequences, for example the set of histones, which produce many degenerate peptides, pointing to the essential problem of protein inference in tag-based proteomics that presently remains an open question.

To illustrate the potential utility of PPI enrichment in protein co-expression modules, the sets of proteins involved in influenza PPIs are associated with KEGG pathways. When defining the universe as all proteins contained in the mass tag database (5,521 proteins), a range of significant pathways were found including “regulation of actin cytoskeleton” (mmu:04810), the “tight junctions” pathway (mmu:04530), and the “antigen presentation and processing” pathway (mmu:04612). When the universe is restricted to proteins with known roles in KEGG pathways (2,539 proteins), the antigen presentation pathway alone remained significant in two modules. These pathways are important in the pathological progression of influenza, confirming the relation of network structure

to known biology, and showing the utility of highlighting PPIs in protein co-expression network analysis. Looking ahead, it is possible that by examining the local network structure around proteins taking part in specific pathways, that new members or interactions may be discovered, although this remains an open question.

Gene ontology term enrichment

Modules appear to have overarching functional organization. To further examine this idea, associations with GO terms are tested. For each module, the set of unique genes forms the test set. Gene Ontology enrichment, by module, was evaluated using the GOstats package. All three data sources showed gene enrichment with highly significant adjusted p-values (10^{-3} to 10^{-12}). In the SARS and influenza networks, enrichment for biological processes such as DNA packaging, cellular component assembly, and cellular complex assembly is observed. The sarcopenia network modules also showed significant functional enrichment including immune response and blood processes, lending evidence to the claim of biologically relevant modular organization. Protein co-expression networks should be useful in guiding downstream experiments.

Conclusions

We have demonstrated the feasibility of constructing de novo protein co-expression networks. These networks have a biologically meaningful and approximately scale-free topology (like many other validated biological networks) and contain statistically significant modules. The module summaries significantly correlate with clinically relevant phenotypes. Modules show significant

enrichment for known biological function. The network structure is potentially useful for resolution of degenerate peptides and inference of protein abundance. Peptides can be sorted according to their module centrality and relationship to phenotypic traits, allowing researchers to prioritize targets for further research. Finally, modules can provide a natural aggregate representation for composite biomarker discovery. These results suggest a significant advancement for proteomic network analysis.

4. Protein Inference For Tag-Based Proteomics

Introduction

Systems biology is an experimental approach to learning about the complex systems present in living organisms (Tisoncik & Katze, 2010). To date, many systems models are focused on gene expression. However, it is becoming clear that models will benefit from the inclusion of proteomics data (Vogel & Marcotte, 2012, Malmström & Lee, 2007; Weston & Hood, 2004). In order to get a more meaningful sense of how cells respond to given conditions, direct measurements on protein products are necessary (J. Cox & Mann, 2007).

Biological systems are known to be noisy and contain a lot of cross-talk (Donaldson & Calder, 2010; Koh, Teong, Clement, Hsu, & Thiagarajan, 2006; Waltermann & Klipp, 2011). Great numbers of replicates generally improve the modeling of complex systems and the power of statistical tests. Consider the task of learning a complex statistical model (Needham, Bradford, Bulpitt, & Westhead, 2007). A good model fit requires hundreds of samples. Therefore, to be truly useful for systems modeling, high-throughput experiments are necessary.

One approach that has found success in high-throughput proteomics is that of tag-based LC-MS (Smith, et al., 2002a; Smith, et al., 2002b; Zimmer et al., 2006). For these measurements, first, tandem mass spectroscopy (MS) is used to create a database of peptide tags using pooled samples. Then, subsequently, in LC-MS runs, features (the resulting data) are matched to peptides in the tag database by elution time and mass. To judge the quality of a match to the tag

database, several metrics are computed, including the STAC score, the probability of a correct match to the database, and the uniqueness probability (UP), which indicates the uniqueness of the match (Stanley et al., 2011). The result is a list of attributed features (peptides) across a number of samples, each with a STAC and UP score, in addition to information about the quality of the tag itself.

Although interesting work can be done at the peptide level, ultimately researchers wish to know the contents of the original biological mixture, and how its abundance was affected by biological or environmental events (Rappsilber & Mann, 2002). The protein inference problem is defined as the challenge of finding the most likely set of proteins that would generate the observed peptide data (Huang et al., 2012; Nesvizhskii, 2010; Nesvizhskii & Aebersold, 2005; Serang & Noble, 2012a). This problem is made especially challenging due to the degenerate nature of many peptides as well as the large amount of missing data and the appearance of low confidence identifications.

Work has been done on this topic, however, currently all methods are based on LC-MS/MS proteomics, requiring peptide identifications and statistics from either MASCOT, SEQUEST, or X!Tandem coupled with PeptideProphet probabilities (Craig & Beavis, 2004; Eng et al., 1994; Nesvizhskii, Keller, Kolker, & Aebersold, 2003; Perkins, Pappin, Creasy, & Cottrell, 1999; Keller, Nesvizhskii, Kolker, & Aebersold, 2002). Results from tandem MS are significantly different compared to tag-based proteomics, and as such, it is helpful to have a protein inference method tailored accordingly. The assumptions made, and the statistics

derived for tandem MS are quite different than tag based proteomics due to the nature of the problem. In tandem MS, the range of peptides possibly identified is orders of magnitude greater compared to the closed universe of the tag database suggesting the Poisson distribution. Another difference is that we are not making peptide-spectrum matches (PSMs), but instead make peptide-tag matches (PTMs if you will). The mechanics of peptide identification is performed differently.

There are three primary ways of performing protein inference. The first, as implemented by IDpicker, is a set cover solution (Ma et al., 2009). The result reflects the minimal set of proteins needed to explain the observed peptides. This method has the advantage of great specificity, but sensitivity can suffer (Serang et al., 2010). Secondly, iterative methods, such as ProteinProphet, use methods related to expectation-maximization to produce posterior probabilities on proteins (Nesvizhskii et al., 2003). A nice feature is that all proteins receive a score, which makes it flexible in terms of choosing cutoffs. Lastly, statistically motivated Bayesian models such as Fido make clear assumptions and attempt to find the maximum a posteriori protein set (Serang & Noble, 2012b). These methods try to model the physical process of proteomics. Probabilistic methods are desirable due to the easily interpretable outcomes.

In this chapter, a novel solution for the protein inference problem aimed at tag based proteomics is presented. The model takes inspiration from network-flow methods, where in this case, quality information attached to each identified peptide flows towards the protein. The flow is either helped or hindered by the

degeneracy of the peptide. The protein takes the sum of information, resulting in an identification score. The model also incorporates protein detectability, trained using the mass tag database, making it specific to the problem at hand.

For validation purposes, and because there are no available gold standards in tag-based proteomics, I developed a method for generating large numbers of simulated LC-MS data sets, reflecting real data. The method is compared to Fido, which in the literature has been compared to the most popular protein inference methods, giving us an idea of the accuracy of protein inference.

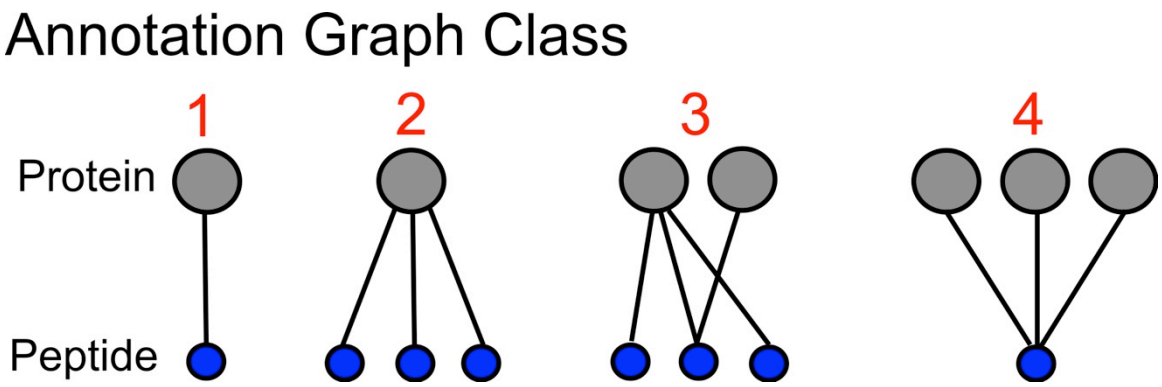

Figure 16. Annotation graphs. The mapping of peptides to protein by sequence can be represented as a bipartite graph termed an "annotation graph" since it is constructed using our current knowledge of the proteome. There are four classes of graph. The first class contains single peptides that uniquely map to single proteins. The second contains multiple peptides that uniquely support a single protein. The third, and where most problems lie, consists of a mixture of degenerate and non-degenerate peptides. Degenerate peptides are defined as peptides that do not unique map to proteins. The fourth class consists of only degenerate peptides, and proteins each with a single peptide.

Annotation graphs

Common to all methods of protein inference is the notion of viewing the problem graphically, connecting proteins to constituent peptides, which then map to observed spectra (Bing Zhang et al., 2007). These complexes are termed annotation graphs because they are built from and describe our current knowledge about the proteome (Figure 16).

The mapping of peptides to protein by sequence can be represented as a bipartite graph termed an “annotation graph” since it is constructed using our current knowledge of the proteome. There are four classes of graph. The first class contains single peptides that uniquely map to single proteins. Although this class is technically contained in either Class 2 or Class 4, we keep it separate since it has been discussed frequently in the context of protein inference (Gupta & Pevzner, 2009). The second contains multiple peptides that uniquely support a single protein. The third, and where most problems lie, consists of a mixture of degenerate and non-degenerate peptides. Degenerate peptides are defined as peptides that do not unique map to proteins. The fourth class consists of only degenerate peptides, and proteins each with a single peptide.

Thus, we can see that we are annotating data with prior knowledge. These graphs represent a map of peptides to proteins, clearly illustrating degeneracies, and cases where one protein is eclipsed by the peptides mapping to another protein (Slotta, McFarland, & Markey, 2010).

Methods

Data

Two quantitative LC-MS data sets involving mouse disease studies are used. The Thermo Electron Exactive platform generated the data. Accurate Mass and Time tag (AMT) databases were developed in-house at PNNL. Peptide identifications are made with VIPER (v3.48) (Monroe et al., 2007). Identifications have confidence metrics: the probability for a correct match, the STAC score, and the probability for a unique database match, the uniqueness probability

(UP)(Stanley et al., 2011). Peptides with STAC scores > 0 and UP > 0 are used providing the largest data set possible. See *http://omics.pnl.gov* for more information.

The infectious disease data came from publically (Systems Virology Center, NIAID Contract No. HHSN272200800060C). We utilize both longitudinal SARS-CoV and influenza mouse studies. This data is generated using C57BL/6J mice exposed to either a mouse adapted SARS-CoV (MA-15, from the Baric Laboratory at the University of North Carolina at Chapel Hill) or avian influenza virus (A/Vietnam/1203/2004 (H5N1, VN1203), from the Kawaoka Laboratory at the University of Wisconsin at Madison). Measurements took place on post infection days 1, 2, 4 and 7.

SARS control samples include three technical replicates per day. Infected samples are five technical replicates with viral dosages of 10^2, 10^3, 10^4, and 10^5 PFU per day. Abundance measurements for 16,890 peptides mapping to 3,277 proteins were recorded.

Influenza control samples include three technical replicates per day. Infected samples include five technical replicates with dosages of 10^2, 10^3, and 10^4 PFU per day. Abundances for 10,285 peptides mapping to 2,661 proteins were recorded.

Simulation Framework

Starting with LC-MSSim (Schulz-Trieglaff, Pfeifer, Gröpl, Kohlbacher, & Reinert, 2008) and later evolving into MSSimulator (Bielow, Aiche, Andreotti, & Reinert, 2011), simulations of mass spectroscopy have advanced steadily. The latter is

now part of the TOPP proteomics software project based at the Max Plank Institute (Bertsch, Gröpl, Reinert, & Kohlbacher, 2011). In my work, a simulation framework was constructed around MSSimulator using the R scripting language (Ihaka & Gentleman, 1996) automating the process, enabling large amounts of simulated data to be produced with a great degree of flexibility (Figure 17).

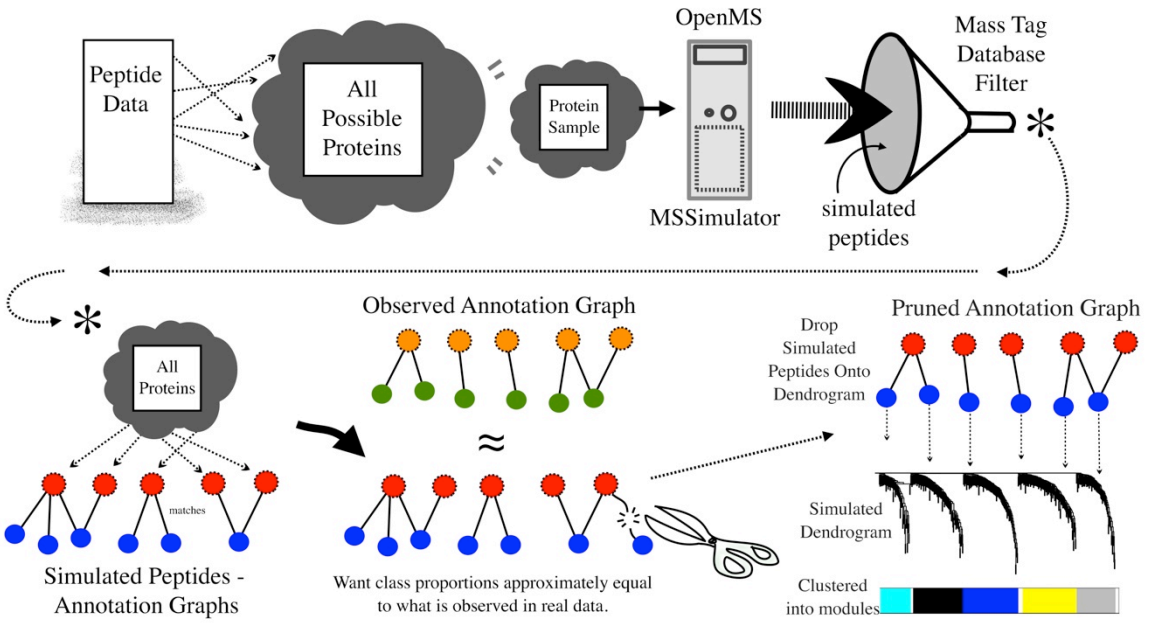

Figure 17. The peptide data simulation pipeline. Given an LC-MS data set and mass tag database, peptides are mapped to all possible proteins, making up the sample pool. A random sample of proteins is processed by MSSimulation which simulates the entire LC-MS pathway from enzymatic digestion to chromatographic elution to charge state prediction and detection. The resulting set of simulated peptides is matched against the mass tag database, resulting in a set of peptides reflecting a realistic observation. Using a protein database, annotation graphs are constructed for the simulated data, which are pruned in a greedy manner so the proportion of each annotation graph classes match proportions found in the real data.

The goal of the simulation is to produce simulated data sets that accurately reflect observed data both in terms of the peptides identified and annotation graphs produced. To this end, the simulation pipeline requires observed data consisting of identified peptides with quality scores and the mass tag database used to identify the peptides.

The pipeline begins by specifying the desired number of samples and proteins per sample. The pool of proteins available for sampling is formed by matching all possible proteins using the observed data. In this way, the simulated peptides result from a similar background. After sampling proteins from the pool, a FASTA file is generated for use with MSSimulator, essentially giving us a known set of proteins for method evaluation. When run, MSSimulator simulates the entire process of mass spectroscopy consisting of in-silico digestion, chromatographic elution, ionization, charge state estimation, feature generation, and detectability prediction. MSSimulator takes an extensive set of option parameters that allow one to tune the output accordingly. The result is a list of simulated peptides for each sample. To better match our data, the set of peptides is filtered by what would be identified using the mass tag database. Using quality scores from real observed data, we generate sampling distributions that are applied to the simulated data, giving each peptide a STAC score, uniqueness probability, and proportion of observations across samples.

After the peptide simulation, annotation graphs are constructed connecting the simulated peptides to the set of proteins in the mass tag database. It was observed that the proportion of each class of annotation graph did not match observed proportions from real data. Commonly, there were too many annotation graphs of classes two and three, and not enough cases of one peptide mapping to one protein (Class 1) making the simulation potentially biased and necessitating the need for pruning.

To transform the set of annotation graphs to better match the observed input data, graph pruning was used to greatly improve the realism by shaping the proportions of each class of annotation graph and the numbers of peptides mapping to each protein. Graphs are pruned by first separating all connected graphs into four classes as stated previously. The proportion of each class is compared to the proportions found in the real observed data for comparison. If the simulated annotation graphs have a class proportion greater than what is observed in real data, then a random graph of that category is selected, and a random edge in the graph is cut. This greedy approach to graph pruning gradually cuts out peptides, and brings the category proportions into alignment, although at a cost of lost simulated peptides.

Simulated Data Sets

To validate the methods, four simulated data sets were generated using the SARS-CoV data. Each simulated data set is generated using a subset of the observed data. Peptide subsets were selected according to the level of missingness, using maximum levels of 10%, 25%, 50%, or 75%. For example, if our goal is to retain all peptides with a maximum amount of missing data of 10%, then a peptide identified in 91% of the samples, missing in 9%, would be accepted. Subsetting the data has an effect on the distributions of quality scores and proportions of annotation graph classes. The quality scores were sampled from distributions constructed from each subset. The tag quality score from the initial construction of the mass tag database is used.

The four simulated datasets were generated using different numbers of protein inputs as well. With respect to the missingness levels, the numbers of proteins sampled were 300, 600, 900, and 1200. After the simulation, the number of proteins mapped to by simulated peptides is often different from the starting number. This change is due to some proteins subject to poor ionization or elution times that are too long or short, and effectively lost during the process. The annotation graphs were pruned to match the full SARS-CoV data set. Table 3 shows the features of the simulated data sets.

DataSet	MaxMissing	DataPoints	Samples	Peptides	MappedProteins	TrueProteins
Sim10 Unpruned	10%	127860	60	2131	374	300
Sim25 Unpruned	25%	210300	60	3505	637	600
Sim50 Unpruned	50%	309060	60	5151	918	900
Sim75 Unpruned	75%	361980	60	6033	1167	1200
Sim10 Pruned	10%	84600	60	1410	247	300
Sim25 Pruned	25%	138540	60	2309	501	600
Sim50 Pruned	50%	206940	60	3449	748	900
Sim75 Pruned	75%	262260	60	4371	986	1200

DataSet	Graphs	Class1	Class2	Class3	Class4
Sim10 Unpruned	222	0.086	0.613	0.279	0.022
Sim25 Unpruned	433	0.122	0.656	0.203	0.018
Sim50 Unpruned	650	0.108	0.688	0.183	0.021
Sim75 Unpruned	845	0.129	0.692	0.166	0.013
Sim10 Pruned	202	0.406	0.495	0.089	0.010
Sim25 Pruned	410	0.040	0.495	0.093	0.010
Sim50 Pruned	622	0.405	0.495	0.088	0.011
Sim75 Pruned	819	0.403	0.496	0.090	0.011

Table 3. Simulated data sets. Four simulated data sets were generated using SARS-CoV data subset by missingness. For example, the Sim10 data set is simulated using peptides that are present in at least 90% of samples, or missing in a maximum of 10%. In each case, 60 samples were simulated. Above the proportions of annotation graph class are shown before and after the pruning step. Class 1 annotation graphs are produced by pruning Classes 2, 3 and 4.

Virology subsets for comparison

For each data source, influenza and SARS-CoV, three subsets were taken by missingness with levels of 10%, 50%, and 100% producing a variable number of observed peptides and proteins mapped. At the 100% subset, all peptides were

used even if observed in a single sample providing the greatest range of proteins possible. Table 4 shows the contents of each dataset.

DataSet	MaxMissing	DataPoints	Samples	Peptides	MappedProteins
SARS10	10%	345910	128	1759	714
SARS50	50%	727117	128	4519	1376
SARS100	100%	1009000	128	16890	3277
FLU10	10%	134938	188	1019	500
FLU50	50%	278945	188	2551	972
FLU100	100%	391064	188	10285	2661
	Annotation				
DataSet	Graphs	Class1	Class2	Class3	Class4
SARS10	480	0.548	0.279	0.1375	0.035
SARS50	1026	0.52	0.347	0.107	0.025
SARS100	2643	0.412	0.487	0.089	0.012
FLU10	342	0.614	0.213	0.137	0.035
FLU50	718	0.599	0.255	0.112	0.033
FLU100	2128	0.49	0.409	0.085	0.016

Table 4. Description of real data subsets. Three subsets of each data source were generated according to levels of missingness. SARS10 and FLU10 each contained peptides with a maximum of 10% missing data. SARS50 and FLU50 contained peptides with a maximum of 50% missing data across samples, and SARS100 and FLU100 contained all identified peptides. Depending on the level of missingness the proportions of annotation graph change, typically with the proportion of class 2 graphs increasing with increasing amounts of missing data.

Network flow model for protein inference

This protein inference model is based on the idea of information flow on a bipartite graph, which represents a new approach to the problem. Within a given annotation graph, we have peptides connected to proteins. Each of the peptides is observed across some number of samples, and for each sample a set of quality scores (STAC, UP, tag quality score) is kept. The model begins with the product of quality scores, and "flows" them towards the protein targets. The "flow" of quality information is limited by degeneracy and missingness of the peptides, but rewarded for uniqueness. The protein targets accumulate information, but the total is again limited by a prior estimate of detectability. A diagram of the model is shown in Figure 18.

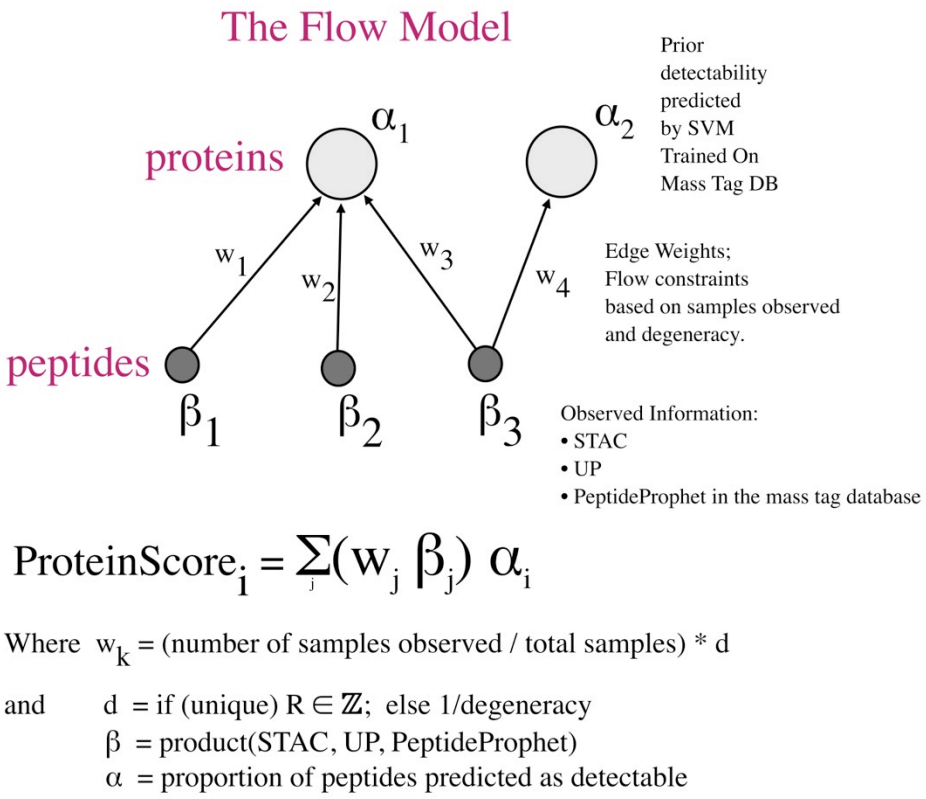

Figure 18. The protein inference model. Schematic illustrating the calculation of the protein score in the Flow Model for protein inference.

The prior probability, α_i, describes the predicted detectability for any given protein before any data has been collected (Yong Fuga Li et al., 2010b; Tang et al., 2006; Webb-Robertson et al., 2008). A uniform distribution might be useful except that it would ignore the fact that we have a resource available, namely the mass tag database. Naively, we might define the prior as the proportion of tags for the given protein compared the set of all tags in the mass tag database. However, this assignment would punish small proteins that have few proteotypic peptides.

Peptide detectability describes the variation in observations given the physical-chemical properties of peptide sequences interacting with MS

instrument architectures. It has been observed that some peptides are detected more often, and depending on the MS platform used, the set of detectable peptides change. Overall, the current sentiment in the literature is that detectability is primarily a function of the sequence of a given peptide. Depending on the sequence, overall peptide properties become more or less prominent, such as hydrophobicity or the flexibility of the peptide backbone.

Webb-Robertson et al. proposed an SVM classifier for peptide detectability trained using the mass tag database. If we consider the in silico digestion of a protein sequence, producing a finite number of peptides, some proportion of those can be found in the mass tag database, essentially meaning that for this experiment these particular peptides are detectable. Peptides not found in the mass tag database are considered undetected and labeled negative examples, while peptides found in the database are labeled as positive examples. Then, for each peptide resulting from the in silico digestion, a feature vector is generated using previously determined constants of hydrophobicity, flexibility, disorder, and other modeled characteristics of peptides as described in the previous work on detectability.

Fortunately, great quantities of amino acid characteristics are found in the R package “Seqinr” (Charif, 2007wf). By counting the specific residues found in any given peptide, we can easily calculate a mean characteristic such as Grantham polarity or average histidine composition. A combination of the best performing features from the previous three papers was used here. In this case, the in silico protein digestion algorithm “dig2” was used (Palmblad, 2000) and the

SVM implementation found in the R package “e1071” was used for prediction. All proteins found in the mass tag database are used for prediction. Using protein sequences from Uniprot (downloaded July 2012), an in silico trypsin digestion (allowing up to 2 missing cleavages) is performed and 47 features are computed for each predicted peptide. Those predicted peptides that are found in the mass tag database are labeled detected and all others are labeled undetected. This dataset is used to tune and train an SVM (radial basis, polynomial = 3, gamma = 0.1, cost = 1). For any given protein, an in silico digest produces a set of predicted peptides, each of which can be scored for detectability and the proportion of detectable peptides used for prior probability. The final detectability score for a given protein is the proportion of constituent peptides that are classified as detectable. A peptide is determined to be detectable if prediction results in a probability greater than 0.5. The model is then defined as

$$\underline{\text{ProteinScore}}_i = \alpha_i \sum_j (w_j * \beta_j)$$

β_j = quality information for peptide j
(product of max STAC, UP, PeptideProphet)

w_j = Edge_j = the information limiting edge
= (B or P) * Presence

α_i = Predicted detectability for protein i.

B = bonus given to unique edges (2 & 3 are considered)

P = punishment given to degenerate edges (1/degeneracy)

Presence = proportion of samples where a peptide is identified.

This model rewards proteins that have multiple, non-degenerate, supporting peptides with high STAC scores, uniquely mapping to identifications in the mass-tag database, having repeatable measurements across samples,

and are predictably detectable from prior information. This model reflects a belief that a given protein is present. Proteins with low scores lack enough evidence to be believable, which is different than saying we have evidence against a given protein. This model is sufficiently transparent and tractable, easy to understand and modify making it accessible for modification in the future.

Features Used In Peptide Detectability Prediction

These features are taken from and described in the seqinr package.

(Charif & Lobry, 2007).

1. Mass
2. Length
3. Mass To Length
4. Average Positive Charge
5. Total Positive Charge
6. Average Negative Charge
7. Total Negative Charge
8. Number Of Nonpolar Hydrophobic Residues
9. Polar Hydrophobic Residues
10. Uncharged Polar Hydrophilic Residues
11. Charged Polar Hydrophilic Residues
12. Total Positively Charged Polar Hydrophilic
13. Total Negatively Charged Polar Hydrophilic
14. Eisenberg Scale Hydrophobicity
15. Hopp-Woods Hydrophilicity
16. Kyte-Doolittle Hydrophobicity
17. Roseman Hydropathy
18. Grantham Polarity
19. Vihinen Flexibility
20. Grantham Polarity
21. Fauchere Normalized Van Der Waals Volume
22. Weber-Lacey Rf Value In High Salt Chromatography
23. Zimmerman Bulkiness
24. Zimmerman Polarity
25. Zimmerman Isoelectric Point
26. Eisenberg-Mclachlan Atom Based Hydrophobic Moment
27. Shannon's Entropy On Sequence
28. Count Of Each Amino Acid (20 Features)
29. Detected (In Mass Tag Database)

Results

Simulated Data Sets Compared to Real Data

The four simulated data sets are compared to the full set of SARS-CoV data revealing quite similar distributions of quality scores and annotation graphs. The comparisons are made to show that the simulated data is useful for validation. In each case, the proportion of annotation graph classes was held essentially constant, while the number of input proteins was varied (Figure 19). In addition, the distribution of quality scores, which are sampled from during the simulation, were constrained so that simulations with smaller numbers of proteins had higher scoring quality metrics. As the size of the sample pool increased, the scores became more similar to the full SARS data set.

For each simulated set, the proportion of peptides to mapped proteins is quite similar. The full SARS set averages five peptides per proteins. The sim10 set averages slightly above 5.5, and the other simulated sets average slightly more than 4.5. More importantly, the topology of the annotation graphs makes the largest impact on the protein inference. More complicated graphs make for more difficult inference.

All simulated data sets have an annotation-graph-to-protein ratio (number of graphs compared to number of proteins) of about 0.8, which agrees with the full SARS set. We do find some disagreement when examining the average node degree for proteins by annotation graph class.

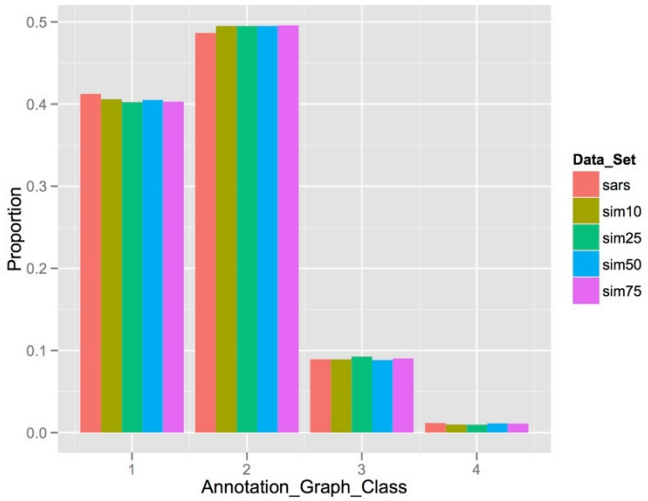

Figure 19. Proportions of annotation graphs are similar between simulated data and SARS-CoV data.

Class 1 will always have a degree of 1, and similarly with Class 4. In Class 2, where we have some number of uniquely mapping peptides connected to a single protein, we find the sim10 data set has considerably more peptides connected to each protein than other simulations and the full SARS set. The other three simulations agree with the SARS set (see Figure 20).

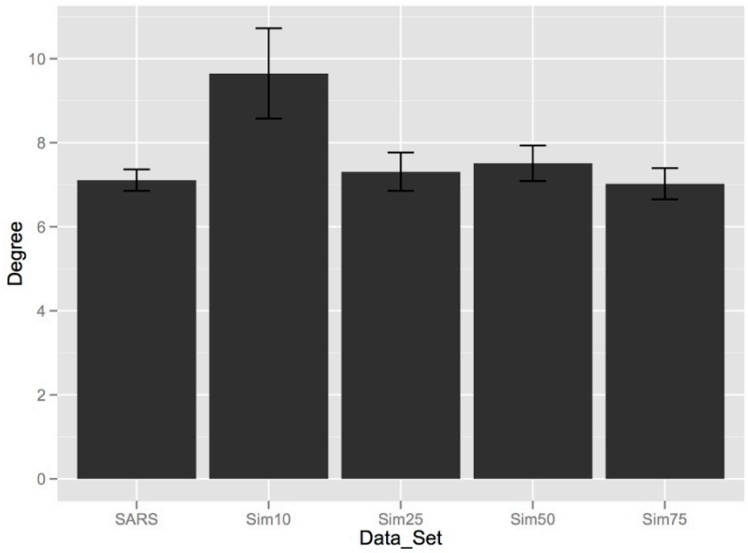

Figure 20. Annotation graph Class 2 protein node degree. The degree of annotation class 2 proteins shows the number of supporting peptides. Simulations Sim25, Sim50, and Sim75 show similar amounts of peptides in each graph, while Sim10 has a higher degree.

In Class 3, where we have mixtures of multi-mapping peptides and proteins, the SARS full set has an average degree of close to 12, whereas the simulations have average degrees of slightly less than 8 (Figure 21), likely a result of graph pruning.

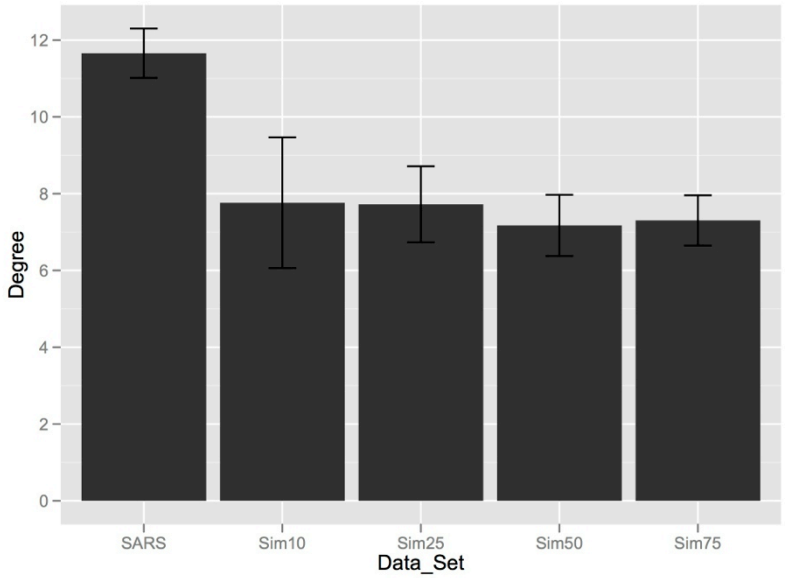

Figure 21. Annotation graph Class 3 protein node degree. The degree of proteins in annotation graph class 3. SARS-CoV data shows a much higher degree, meaning that proteins in those graphs have more supporting peptides.

Class 3 graphs have the most potential for variation due to the mixture of multi-mapping peptides and proteins, and due to the way that proteins are sampled, graphs are pruned and quality scores sampled. Some definite trends in the topology of Class 3 graphs are observed. For the full SARS set and simulation sets other than Sim10, the number of protein targets in each class 3 graph are largely similar. The Sim10 set has a smaller number of proteins, so appears to have a larger boxplot whisker, but it is not appreciably different (Figure 22).

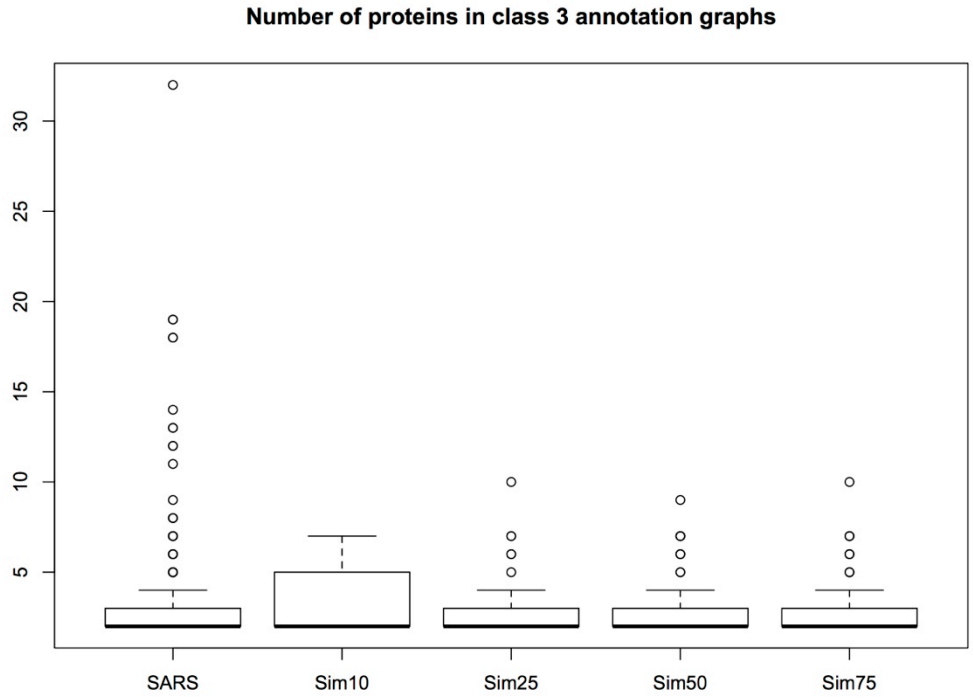

Figure 22. The number of proteins in each Class 3 annotation sub-graph. Although Sim10 appears to be an outlier, the boxplot is affected by the small number of graphs. In other cases, the mean and third quartile line up with real data. SARS does show a larger number of possible outliers, where some annotation graphs contain a large number of proteins, potentially making them difficult targets for protein inference.

There is some variation of degeneracy within Class 3 graphs (Figure 23). A *degeneracy ratio* is the number of edges in a graph divided by the number of peptides. So a graph with all uniquely mapping peptides would score 1, and graphs with degeneracy, or more edges than peptide nodes, would score > 1. Overall, Class 3 graphs are dominated by graphs with a small amount of degeneracy, and rarely does the degeneracy-ratio rise above 5. Without considering outliers, the mean degeneracy starts low, and increases across the simulated sets until reaching parity with the observed SARS data.

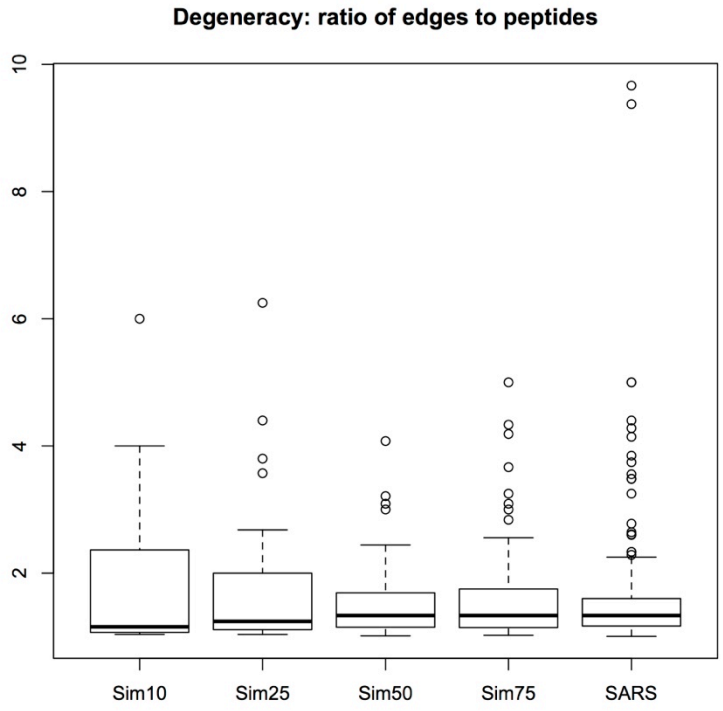

Figure 23. The ratio of edges in class 3 annotation graphs to the number of peptides, showing the relative amount of degeneracy contained in each graph. Here, Sim10 and Sim25 have slightly smaller amounts of degeneracy in these graphs, while Sim50 and Sim75 have very similar means.

The quality scores, which directly influence the protein inference results, show good agreement. The STAC scores tend to be higher in the Sim10 simulations, but gradually return to parity with the SARS data. Conversely, with the uniqueness probability scores, the Sim10 set is lower and then across the simulations, Sim25, Sim50, and Sim75, these scores return upward to the SARS data.

In summary, the proportion of peptides and annotation graphs to proteins in simulated annotation graphs is in agreement with observed data. The average degree for protein nodes in the simulations is generally similar, although in Class 3 graphs, the simulations have distinctly fewer peptides forming sparser graphs. The mean level of degeneracy in Class 3 graphs is somewhat lower in the

simulations with fewer proteins, but becomes similar to the observed data in Sim50 and Sim75 data sets. Additionally the number of proteins in Class 3 graphs are very similar across all sets. The quality scores follow the same distributions, with the smaller simulated sets having quality metrics skewed towards higher values. Overall, the Sim75 data set is qualitatively most similar to the full SARS set.

Protein detectability

Protein sequence and structure variation leads to variation in peptide detectability, but is also affected by laboratory protocols and the instrumentation used. In this work, protein detectability was defined as the proportion of detectable constituent peptides. An SVM classifier was trained to predict peptide detectability, and on cross-validation tests using the tag database, achieved 73% accuracy. The task was to predict if a given peptide produced from in silico digestion would have been detected when building the mass tag database. The predictive model allows that some peptides produced after digestion, and not appearing in the mass tag databased are still able to have detectability scores as high as other peptides found in the database. When setting the detectability threshold at 0.5, most proteins had detectability scores around 0.3. For example, in digesting proteins BAG2_MOUSE, APOA4_MOUSE, and LCAT_MOUSE, 16, 29, and 24 peptides were produced respectively, and the proportion of detectable peptides was 0.313, 0.310, and 0.4 (See Figure 24).

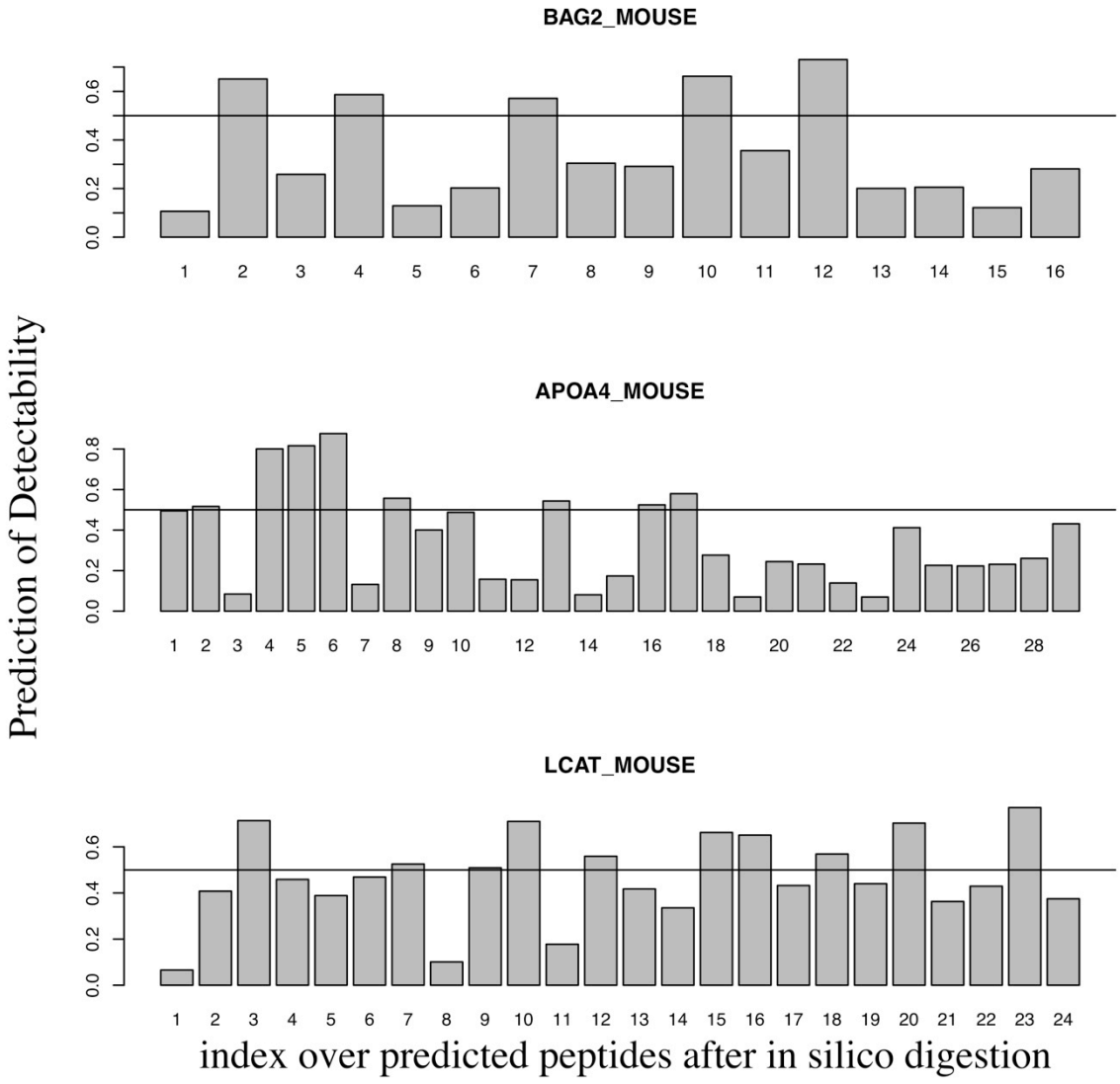

Figure 24. Peptide detectability. The Y axis shows probability for detection and the X axis indexes the predicted peptides. Prediction of peptide and protein detectability using an SVM classifier trained using the mass tag database. The classifier makes a prediction for each peptide after an in silico digestion. The proportion of detectable peptides is used as the detectability score for a given protein.

Simulated protein inference

For simulated data, known proteins allow exact predictive performance to be measured. Protein inference using both Fido and the flow model give each protein a score. Fido results in a posterior probability, while the flow model results in a numeric score. By thresholding the scores, we partition the list of

proteins into present and not-present. The partitioned list can be compared to the set of proteins used in the simulation giving true and false positive rates. By varying the threshold, we produce receiver operating characteristic (ROC) plots. A perfect predictor would have an area under the curve (AUC) of one, while a random predictor would have an AUC of 0.5.

For simulated data sets, Fido was used with default parameters given by application website (hosted by the Nobel lab) which are described as being very robust parameters. To be fair, a very limited amount of tuning was performed for either model. For the flow model, one parameter, the bonus given to unique edges in parameter W, was tried with two values, 2 and 3. With parameter B = 2, the results appeared very similar to Fido, and with parameter B = 3, the results surpassed Fido. It is possible that with proper parameter optimization, Fido would perform better, but in limited testing, parameters were not found that improved the predictions.

			% Agreement in Top Ranked Proteins		
Data Set	Flow ROC	Fido ROC	Cut @ 90%	Cut @ 80%	Cut @ 60%
Sim10	0.949	0.938	0.981	0.964	0.858
Sim25	0.909	0.893	0.989	0.932	0.867
Sim50	0.901	0.871	0.993	0.925	0.875
Sim75	0.913	0.883	0.993	0.926	0.885

Table 5. Prediction results for each simulated data set using both the Flow and Fido models. The Flow ROC and Fido ROC columns show the area under the curve while the Cut columns show similarity between Flow and Fido when comparing the top X%.

The ROCs for the flow model show consistent AUCs above 0.9 which is considered very good. Fido also performs very well with AUCs over 0.9 in half of the trials indicating that scores do a good job separating true and false proteins. As the threshold changes, there is an abrupt transition between true and false

predictions giving the predictor the ability of high sensitivity and simultaneously high specificity (Figure 25).

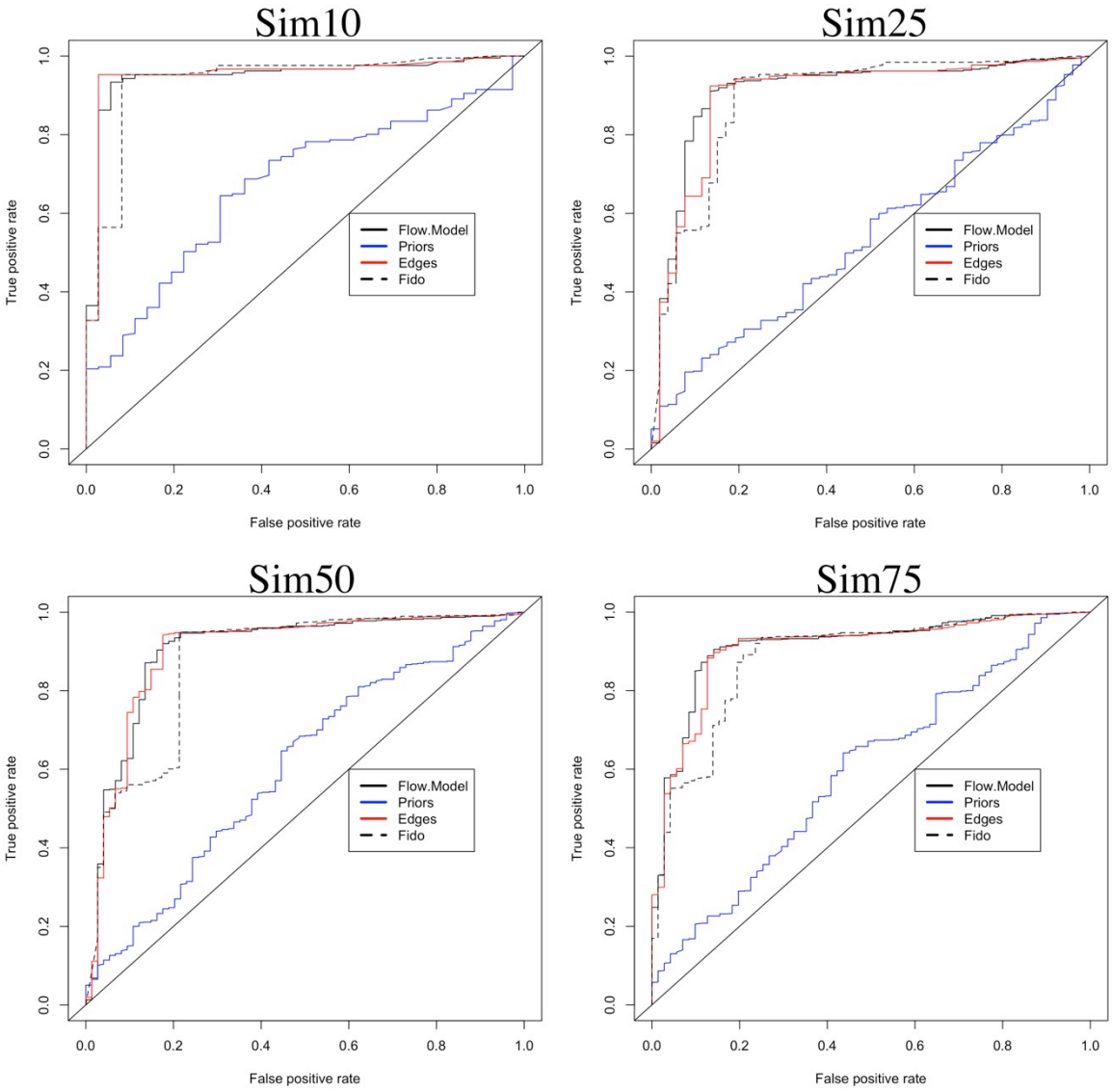

Figure 25. ROC curves for prediction on simulated data sets using both Flow and Fido models. The dotted black line represents the prediction made by Fido. The blue line represents prediction made using the detectability prior alone. The red line represents the quality flow information before the prior is applied. The black line represents the final Flow prediction.

When a threshold is selected, as already mentioned, the list of proteins is apportioned into a list of proteins we believe are in the original biological mixture, and those we do not. With a given threshold we can compare the predictive result of each algorithm. One comparison method is to count the number of

proteins in agreement divided by the number possible. In this work, I looked at cut points of 0.9, 0.8 and 0.6. For example, the proteins are ranked, and the top 90% are taken for comparison. The flow model and Fido compare favorably where at 80 and 90%, the two lists are nearly identical, which also implies the bottoms of the lists are identical as well. When we compare the top 60% of each list, there is less agreement ranging from 0.858 to 0.885 percent. It appears there is a differing order in the middle of the ranked list of proteins (see Discussion).

SARS and influenza protein inference

In protein inference on real observed data, the proteins responsible for the observed peptides are unknown. To compare the algorithms, protein inference is performed, and proteins are ranked according to the scores they receive. The same proteins are used in the results of each algorithm. Then taking the top X% of each list (from Fido and Flow), the intersection is used to get the percent agreement.

The two inference models produce different score curves, which is shown in Figure 26. Notably the Flow model makes a smooth continuous range of scores. In the full SARS data, which maps to 3185 proteins, the lowest 1000 scores show approximately 600 with scores close to zero, which afterwards increases at a steep rate. Fido, on the other hand, tends to give somewhat high scores to many proteins. Considering that these are posterior probabilities, we observe that only slightly more than 200 of 3185 proteins receive zero probabilities. After this point, there is a large discontinuous jump of posterior

probabilities to 0.4, which quickly increase so that more than 2300 protein have posterior probabilities of greater than 0.6.

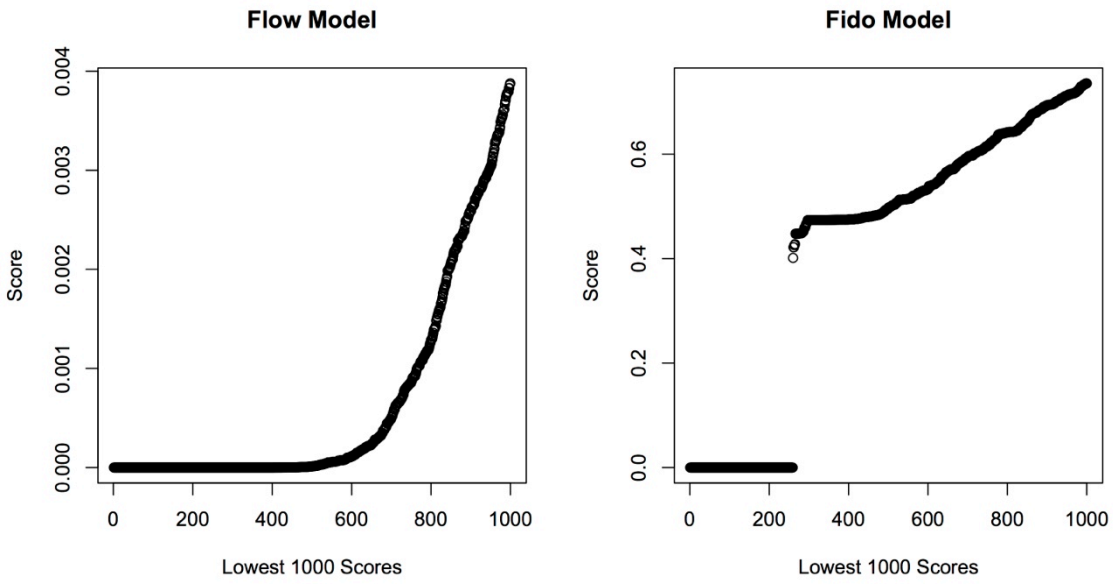

Figure 26. The lowest 1000 protein inference scores produced by the Flow and Fido models.

For all datasets tested, a similar trend in agreement is observed (Figure 27). The agreement starts very low in comparing the top 1-5% of proteins, and then quickly increases upwards peaking around 20% of the top scores. A drop is observed between 40 to 60%, finally rising to 100% as the lists converge. It is not surprising that the top of the lists do not agree since Fido ranks many proteins with a score of one, becoming randomly sorted as the list is ranked. If we are concerned with taking a large portion of the proteins, then we find that the two algorithms have a very good agreement starting at approximately the 80% level.

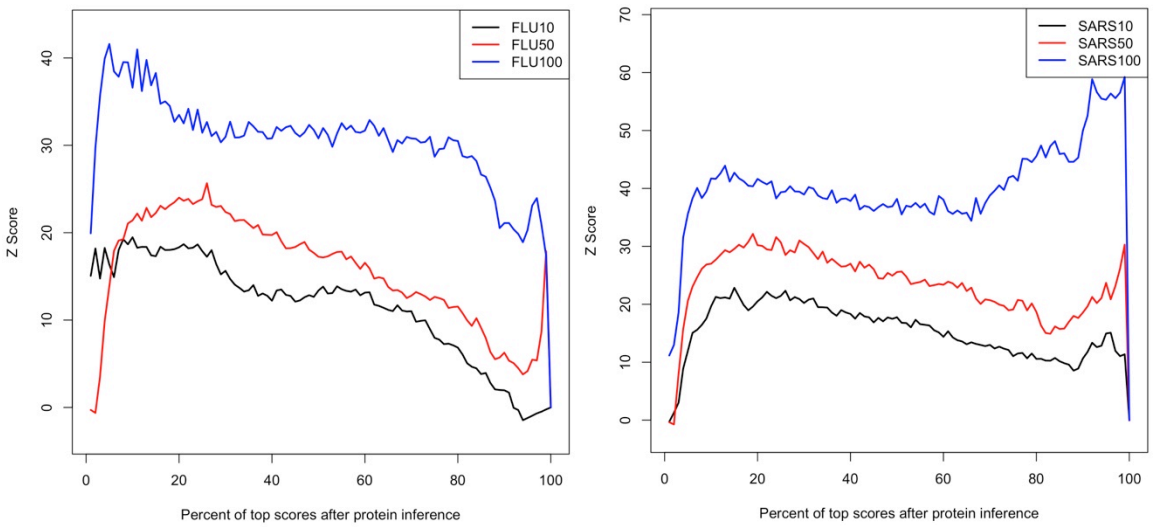

Figure 27. Comparisons of protein inference using Flow and Fido with real data. Three data sets were generated by taking only peptides with a maximum of 10% missing data (SARS10, FLU10), a maximum of 50% missing data (SARS50, FLU50), or the largest data set including all peptides (SARS100, FLU100). In each case, after protein inference, proteins are compared after selection by scores in the top x% (which is along the x axis). The observed overlap in the selection is normalized compared to what is expected by chance.

If we consider that the Flow model makes no use of latent variables, and instead represents a transformation of the observed data, then with some assumptions we are able to define minimum scores. Suppose that we are interested in proteins that have at least one peptide with a STAC score of 0.6, a UP score of 0.8, and a tag Peptide Prophet score of 0.9. In addition, we want this peptide to be observed in at least half the samples, and non-degenerate, with a prior probability of 0.5. Taking the product of these scores gives 0.108. However, when the SARS full dataset is examined, only 1,228 of the proteins have scores above this threshold. In Fido, this threshold is 0.84, which might be a good cut point for confidence.

By examining low scoring proteins, a few different aspects become clear. For one, many of the low-scoring proteins are those seen in a small fraction of the samples. From the SARS full data set, 4272/16890 peptides are present in at

most 2% of the samples. Fido does not take the number of times a peptide is observed into account. It is simply operating with the assumption that if we saw it once, then it should be in the annotation graph.

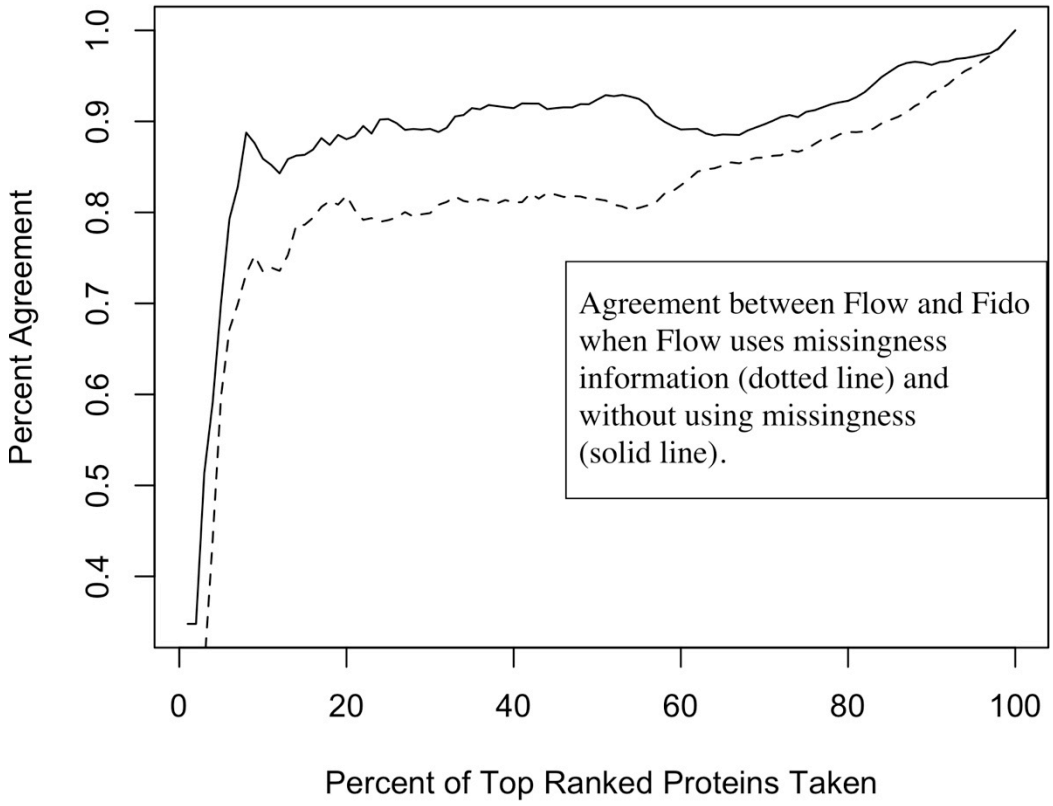

Figure 28. Percent agreement between protein inference of the Flow and Fido model using the SARS100 data. When the Flow model does not consider missing data information, the results from the two algorithms becomes more similar, as seen in the SARS10 comparisons.

To demonstrate the effect of accounting for the number of observations, protein inference was performed on the full SARS data, but setting the Presence parameter to one, as if every peptide was observed in every sample. The result is much improved agreement with Fido (Figure 28).

Discussion

Differences from Fido – Use Case Example

Clearly, since Fido does not take the number of times a peptide is observed into account, the data should be manually cleaned before protein inference. However, it appears that even without data cleaning the Flow model performs fairly well. Proteins supported by peptides either observed very rarely or with very low quality scores are accordingly ranked low.

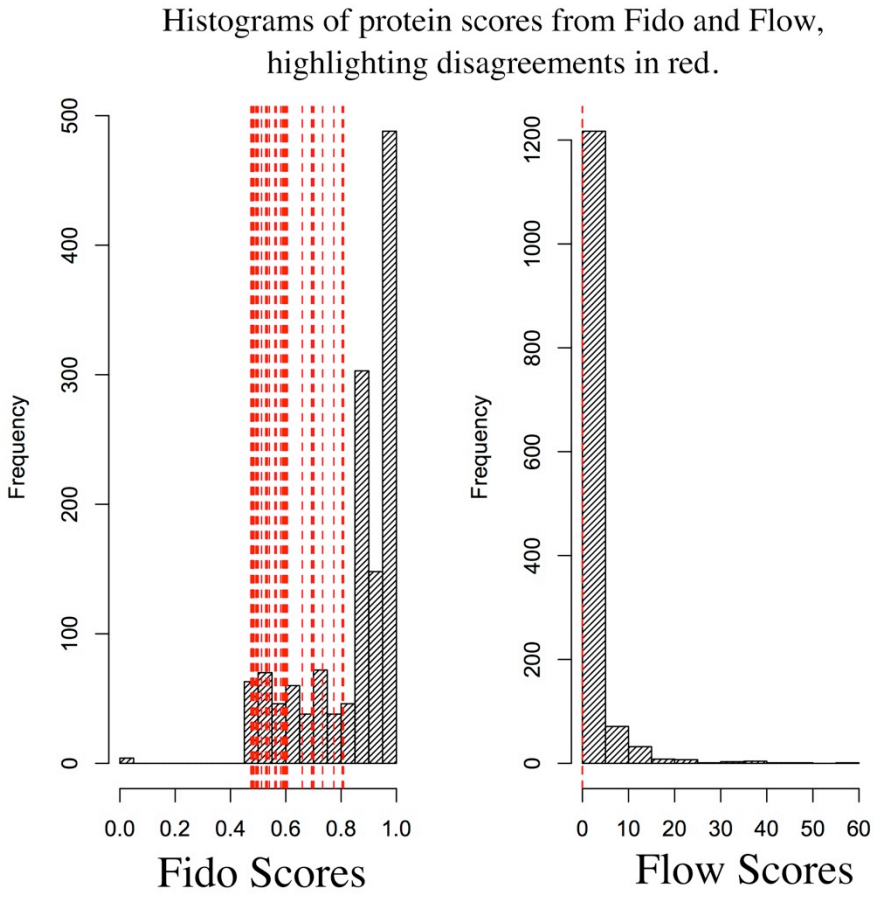

Figure 29. Disagreements (shown in red) in the top 95% of protein inference. Fido is giving a broad range of scores compared to the Flow model. These are proteins that are accepted by Fido and not by the Flow model. A majority of disagreements come from proteins with a single supporting peptide, often with very low STAC scores.

However, there remains some disagreement between the algorithms. Using the SARS50 dataset and taking the top 95% of the ranked protein list, proteins are examined that are accepted by Fido and not by the Flow model. Some discordance is shown below in Figure 29. With this subset, in Flow, the scores are concentrated towards the bottom of the ranked list. In Fido, scores are spread out, including some proteins with posterior probabilities above 0.75. In total, 52 proteins did not agree. These proteins had Flow model prior probabilities ranging from 0.44 to 0.76, but extremely low flow-model-edge scores. An example is shown below in Table 6.

Protein	Fido Score	Flow Score	STAC	Prior	Edge
ACBE1	0.81	5.50E-07	0.096	0.5	9.00E-07
ATX10	0.73	5.07E-08	0.009	0.49	1.02E-07
SUCB2	0.77	0	0.19	0.46	0
DHPR	0.7	5.90E-07	0.006	0.58	1.01E-06

Table 6. Score discrepancies between the Fido and Flow models. Puzzlingly, these proteins have high posterior probabilities from Fido when considering the extremely low STAC scores. In this case, these proteins were supported by a single peptide.

In this example, all of the proteins are supported by single peptides. It is surprising then that Fido would give a posterior probability of 0.81 to protein ABCE1 when it has a STAC score (which is used by Fido) of 0.096.

An approach unique to protein inference

The Flow method represents a new approach to protein inference that tracks very well with a more theoretical, but still well tested algorithm. The Flow model is useful for a first pass examination of the data since it clearly delineates which proteins have suitable information supporting them, and takes advantage of multiple samples by down weighting peptides that are rarely observed.

The prior probabilities are based directly on the mass tag database, which should take account of the protocols used, and the mass spectroscopy platform used. There is certainly room for improvement in protein detectability, and changes to this model could improve the predictive power.

Conclusions

An automated method for producing simulated LC-MS peptide data was produced that harnessed the MSSimulator. Simulated annotation graphs were made to match annotation graph topology distributions present in real data using a greedy graph pruning method. Simulations were validated by comparison of the annotation graph topology and features to the full SARS data. Both protein inference algorithms Fido and the flow model performed very well on simulated data, with the flow model performing slightly better. When applied to real data, the results are in good agreement. Some disagreement is observed for proteins with single supporting peptides. In those cases, the flow model produces more believable scorings.

This is the first protein inference method designed with high-throughput tag-based proteomics in mind. The method is simple in computation and requires no latent variables. Other methods do not take missingness or the particulars of tag-based proteomics into account. This method is also the first network flow based model, which demonstrates a new class of methods in protein inference.

6. An Integrated Systems Signature of SARS-CoV Infection

Introduction

Cells contain an incredible number of molecular entities forming a dense interconnected network of interactions. In order to take a more global systems view in our analysis, multiple data types, representing the range of biological entities, must be integrated. Considerable attention has been given to the problem of data integration in large scale systems approaches (Adourian et al., 2008; Fagan et al., 2007; Gat-Viks, Tanay, Raijman, & Shamir, 2006; McAteer & Skerrett, 2009; Mcgarvey et al., 2009; Troyanskaya, Dolinski, Owen, Altman, & Botstein, 2003; Vaske et al., 2009). Given the many types of "omic" data, the most general problems involve annotation and interpretation (Palsson & Zengler, 2010). Currently, the best and most informative way to integrate biological data remains an open question.

More recently the study of host-pathogen systems has become increasingly important for understanding viral pathology (Aderem et al., 2011; Peng et al., 2009; S.L. Tan, Ganji, Paeper, Proll, & Katze, 2007; Zak & Aderem, 2009; Forst, 2006; Joyce & Palsson, 2006). In these studies, we learn about the relationships among components associated with the host's response to infection. This work is focused on the first line of defense: the innate system (Katze & He, 2002; Takaoka & Yanai, 2006; Zak & Aderem, 2009).

SARS-CoV is an upper repertory virus that caused a global pandemic in 2002 (Ksiazek et al., 2003; Wendong Li et al., 2005b; Low & McGeer, 2003;

Roberts et al., 2007; Rockx et al., 2011; Stadler et al., 2003). Commonly encountered, the corona family of viruses is responsible for a large portion of reported “colds” (Masters, 2006). Understanding the host response to this infection could lead to new therapeutic actions (Dykxhoorn & Lieberman, 2006).

In this work, two methods of data integration are explored and applied to the problem of integrating the transcriptome and proteome. This problem is significant since these two “omes” are a great distance apart in terms of biological processes. In addition to the biology, the technologies used for measuring transcripts and peptides are incredibly different and require different treatments. Using SARS-CoV infected mouse-lung microarray data and tag-based proteomic data, data integration should work to improve our systems understanding of SARS infection.

In this chapter, a new method for integrating transcription and proteomic data is developed using co-expression networks (B. Zhang & Horvath, 2005). Another previously published method, correlated factor analysis (CFA) (C. S. Tan et al., 2009) is explored. The two different methods represent “early” (CFA) and “late” (co-expression networks) approaches. CFA is considered “early” in the sense that data is integrated immediately, whereas integrating pre-built networks is deemed “late”.

Interpreting integrated results might possibly bring new understanding of the response to SARS-CoV, which could lead to biomarkers for pathogenicity, and the ability to predict host response. Ideally, the identification of individuals

requiring more intensive care due to an over-reactive host response could improve triage and reduce the chance of a fatal infection.

Background

In this chapter, we are working with tag-based proteomics and microarray based transcriptomics, which are two experimental domains. Each of them has distinct methods for generating data, learning about organisms, and investigating disease. Tag-based proteomics is described previously in Chapter 4. Following is a very brief description of transcriptomics.

Within any given organism, DNA is transcribed and mRNA molecules (transcripts) are actively transported in the cell. The transcripts ultimately encode protein sequences, but only after a long series of biological transformations and editing events (Vogel & Marcotte, 2012). Measuring transcript abundance of a given organism is thought to give some insight into the state of the cell at a discrete time point. One way expression measurements are taken comes by lysing cells and purifying mRNA. The mRNA is processed according to specific microarray protocols involving the addition of florescent tags (Wolber, Collins, Lucas, De Witte, & Shannon, 2006). The solution is washed over the microarray, at which point the RNA fragments hybridize with (bind to) oligonucleotide probes on the surface of the array. The probes, bound with florescent RNA, are measured by imaging. The brighter the spot, the more RNA has hybridized.

The probes are designed to limit degeneracy; great effort was spent making the probes as unique to a given gene as possible. In this work, the Agilent 4x44 microarray is used. This microarray carries 43,803 probes that map

to 20,835 Entrez gene IDs. We see that genes are often mapped to by multiple probes.

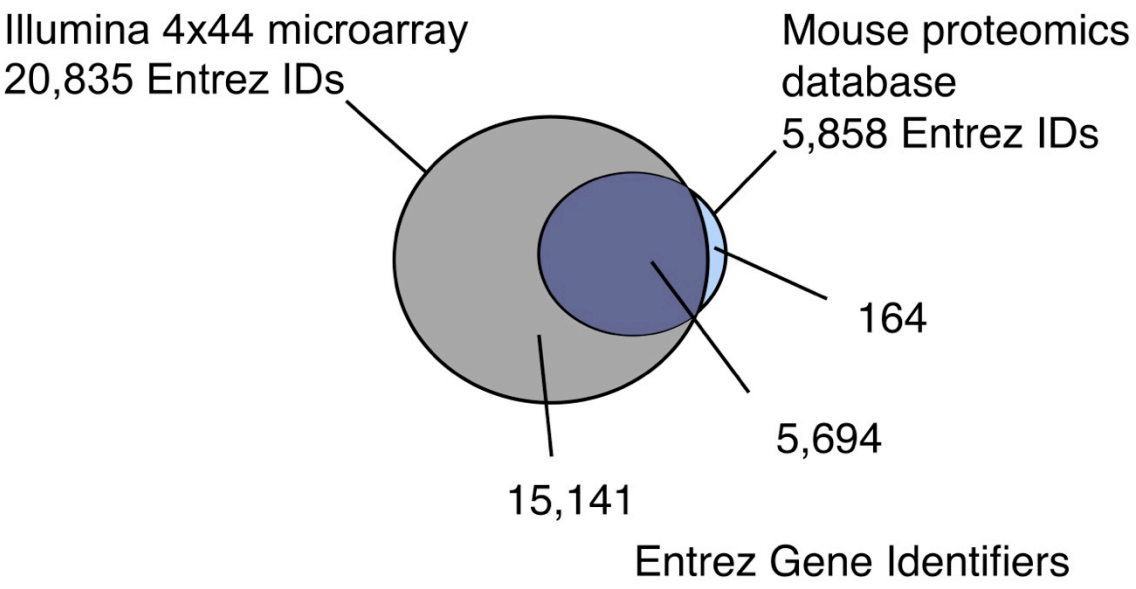

Figure 30. Overlap of entrez gene IDs between the Illumina expression microarray and the mass tag database used to identify peptides in the infectious disease experiments utilized.

Although microarrays are able to cover almost the entirety of the mouse transcriptome, there is lesser coverage for the proteome (Figure 30). The number of proteins measured is expected to improve, but for now remains a limiting factor. Here, the mass tag database used to identify peptides, contains entries for 5,858 Entrez IDs (after mapping Uniprot protein IDs to Entrez). The overlap between microarray and tag database consists of 5,694 IDs, leaving 164 proteins unique to the proteomics side, and 15,141 Entrez IDs unique to the transcript side.

The GO hierarchy is a tree of nodes, connected by edges, and can be viewed in terms of levels (Ashburner et al., 2000). The three roots of the tree are “Biological Processes” (BP), “Molecular Function” (MF), or “Cellular Component” (CC). A non-unique set of genes are mapped to each node in the tree. Here we

are focusing on the BP tree. The root can be considered level 1. Level 2 is the set of GO terms that can be reached by traversing not more than 1 edge. Level 3 are all GO terms that are one additional edge from Level 2 nodes. As we traverse down into the tree, the terms change from more general to more specific (Figure 31).

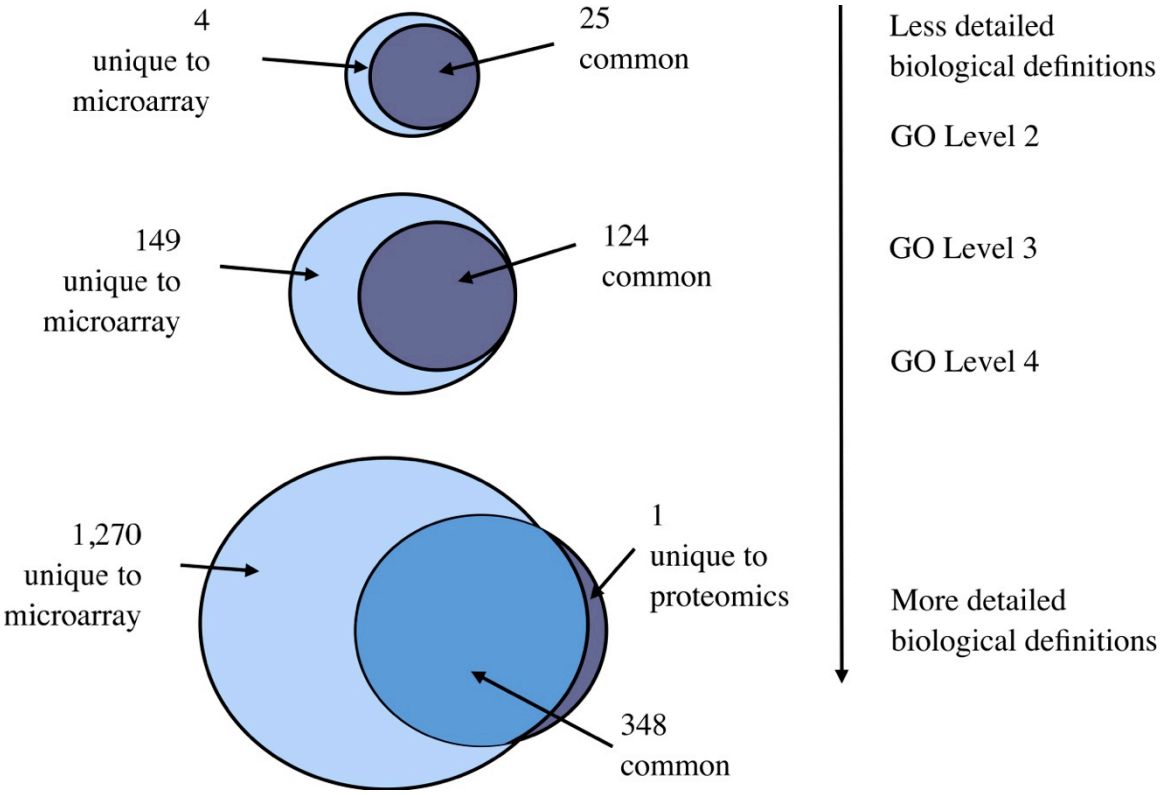

Figure 31. Comparison of GO IDs for members of the Illumina expression microarray and the mass tag database. At "upper" levels of the GO hierarchy, the GO terms are more general, while at "lower" levels the terms become more specialized. In the comparison, while the array has more GO terms unique to its members, it is not until GO level 4 that the proteomics side has a single unique GO term.

At level 2 in the BP tree, 25 GO terms are common between the tag database and the array, with only 4 GO terms unique to the array. At level 3, the difference becomes more distinct where 124 GO terms are common between the data types, but 149 are unique to the array. Finally at level 4 (as deep as this

analysis would dare go), 348 GO terms were common, 1,270 GO terms were unique to the array, and a single GO term was unique to the tag database.

The GO term unique to the virology tag database was positive regulation of leukotriene production involved in inflammatory response (GO:0035491), a specialized term and informative given the topic. In addition, among GO terms unique to the array, immunity related terms are found. These terms are shown in Table 7.

GOID	Description	Array	Proteomics
GO:0070664	negative regulation of leukocyte proliferation	405	0
GO:0002269	leukocyte activation involved in inflammatory response	243	0
GO:0019079	viral genome replication	238	0
GO:0046788	egress of virus within host cell	154	0
GO:0046725	negative regulation of viral protein levels in host cell	59	0
GO:0042089	cytokine biosynthetic process	41	0
GO:0019072	viral genome packaging	39	0

Table 7. GO terms unique to the Illumina expression microarray. The array column shows the number of genes mapping to that particular GO term using Bioconductor annotation packages. These GO terms were selected because of their association with immunity relation functions.

When looking at KEGG pathways mapped to Entrez IDs within either the tag database or the microarray probeset, the difference is smaller than what was seen in the GO term analysis. 199 KEGG pathways mapped to the union of Entrez IDs. Of these only 2 were unique to the array including lipoic acid metabolism (mmu00472) and D-arginine and D-ornithine metabolism (mmu00785). While its beneficial to our analysis to have this bountiful overlap, it

likely demonstrates the sparsity of our pathway knowledge. Many proteins and genes lack functional annotation, and there is no doubt that many pathways remain to be discovered.

Methods

Data

Proteomic Data: Abundance measurements for 16,890 peptides mapping to 3,277 proteins were recorded. Taking all observed peptides, protein inference was performed using the Flow and the Fido models (see Chapter 4). Proteins were accepted with scores above 0.95 in both the Flow and Fido models resulting in 691 proteins.

Peptide data was filtered by protein inference mapping, STAC (> 0.6), UP (>0.5) and Peptide Prophet tag score (>0.9) which resulted in 611,812 observations, or a matrix of 188 sample rows by 9,326 peptides. Sample replicates were combined by taking the mean over peptides, increasing the number of peptides with observations for each sample. Peptides were then filtered by missingness, taking peptides with not more than 20% missing data, resulting in a matrix of 92 samples by 2,273 peptides that mapped to 467 Uniprot protein IDs.

Transcript Data: This microarray data, matched to the proteomic data, became public on Nov 01, 2011. Processing details were taken from the GEO repository:

“Twenty-week-old C57BL/6 mice were infected by intranasal instillation of 10^2,10^3, 10^4 or 10^5 PFU of SARS CoV MA15 in 50 µl of PBS or mock-infected with PBS alone. At days 1, 2, 4 and 7 days post-infection, lungs were harvested.

Specific lobs of the lung from each animal were harvested and briefly rinse tissue in cold (4ºC) PBS. Following the RNAlater (Ambion) protocol, tissue was cut into small chunks (<0.5cm in any single dimension) and place immediately into a 10-20 volumes (w/v) (e.g. 100mg/ml) RNAlater. After a 4ºC incubation for overnight, samples were stored at -80ºC further processing. Lung tissue was removed from RNAlater, washed in a small volume of Trizol, homogenized in 10-20 volumes (w/v) TRIzol and stored at -80°C until RNA isolation.

All TRIzol lysates were processed simultaneously: they were phase-separated, and RNA was isolated from the aqueous phase (diluted 2 fold with RLT buffer) using Qiagen RNeasy Mini columns and the manufacturer’s recommended protocol (Qiagen Inc., Valencia, CA). RNA quality was assessed on an Agilent 2100 Bioanalyzer using the nanochip format, and only intact RNA was used for microarray analyses.

The Agilent One-Color Microarray-Based Gene Expression Analysis Protocol was followed for the Cy3-cDNA probe preparation. The Agilent One-Color Microarray-Based Gene Expression Analysis Protocol was followed for hybridization and array washing. Two hundred fifty ng of each RNA sample was hybridized to one Agilent 4X44K human HG (Design ID 014850) array.

Dry slides were scanned on an Agilent DNA microarray scanner (Model G2505B) using the XDR setting. Raw images were analyzed using the Agilent Feature Extraction software (version 9.5.3.1) and the GE1-v5_95_Feb07 extraction protocol. All arrays were required to pass Agilent QC flags. Extracted raw data were background corrected using the norm-exp method and quantile normalized using Agi4x44PreProcess and RMA Bioconductor packages."

With the procedures above, 92 expression profiles were collected from the four time points (1, 2, 4, 7 days) and four dosage levels (10^2, 10^3, 10^4, 10^5 SARS pfu), and included 3 mock samples per day. 31,416 probes passed probe QC flags for all replicates of at least one infected time point.

After assessing biological replicates, the following replicates were removed: Mock at Day 7, replicate 2 and PFU 10^2 on Day 4, replicate 3, since the mock clustered with infected samples, and clustered with early infected and mock samples, rather than with samples where infections are more progressed.

Pathology

Measurements for a wide range of pathological variables are provided including features such as airway constriction, inflammation, airway inflammation, debris, denudation, the state of the vasculature, and whether signs of pneumonia are observed. Conveniently, all of the variables are combined into a single measure entitled “Overall Total Score”, which is highly correlated with the other variables.

When the lung pathology is observed as the Overall Total Score by time and dosage, we see that the trend is strongest by time, and less difference is seen by dosage. All mice receiving a dose of SARS have observable lung

pathology. The response of individuals receiving dosages of 10^2 PFU appear somewhat different than individuals receiving higher dosages (see Figure 32).

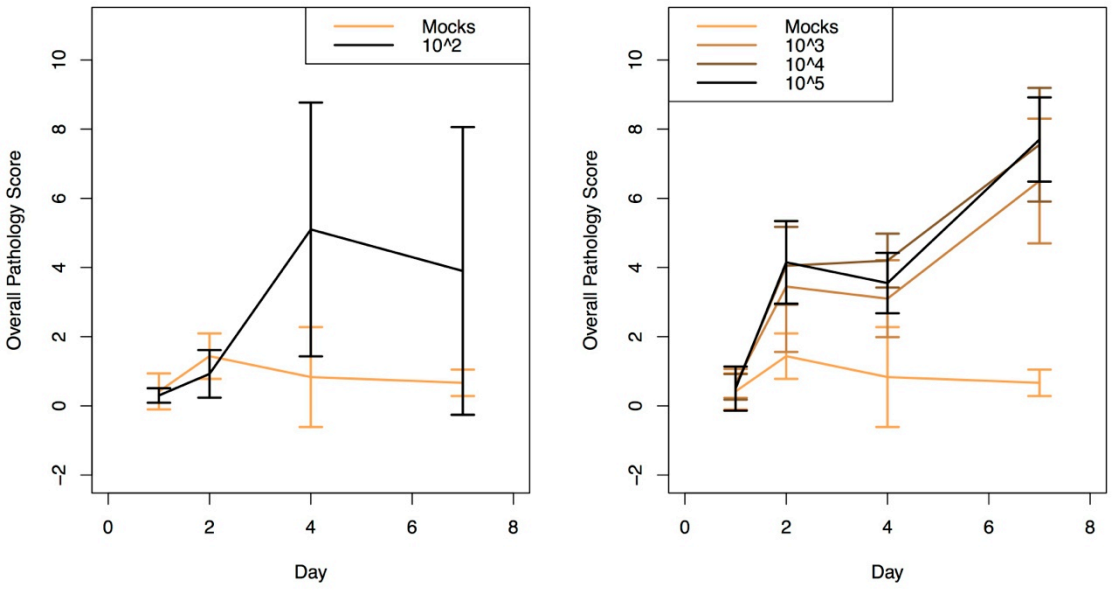

Figure 32. Overall pathology scores by dosage. The figure on the left shows the mocks and the 10^2 dosage level, while the figure on the right shows the 10^3, 10^4, 10^5 dosage levels. The host response patterns are clearly different.

Methods Used

Correlated factor analysis

Correlated factor analysis (CFA) (Tan et al., 2009), a method for data integration, uses ideas from Maximum Covariance Analysis (Storch & Zwiers, 2002) and essentially consists of singular value decomposition on the covariance matrix of gene and protein data. CFA is an interesting approach to the integration problem in that it requires no free parameters and results in immediate joint analysis. Using CFA, the hope is to discover patterns of correlation, shared between the two data sources, that are associated with the biology of disease. In this case,

we are extending the method to transcript and peptide data, which is a novel use of data.

First, the cross-covariance matrix is computed from the matrices of transcript and peptide data by

$$\mathbf{\Sigma} = \frac{XY'}{(n-1)}$$

where *X* and *Y* represent the transcript and peptide data matrices respectively, and n is the number of samples. The covariance matrix, $\mathbf{\Sigma}$, has dimension *p* by *q* where *p* is the number of transcripts and *q* is the number of peptides. Singular value decomposition applied to the covariance matrix results in three matrices:

$$\mathbf{\Sigma_{p\,x\,q}} = \boldsymbol{U_{p\,x\,r} D_{r\,x\,r} V'_{q\,x\,r}}$$

where the left singular values *U* consist of the eigenvectors of $\mathbf{\Sigma\Sigma'}$, the right singular values consist of the eigenvectors of $\mathbf{\Sigma'\Sigma}$, and *D* holds eigenvalues. The k^{th} left and right singular vectors and the singular value together are called a factor or a *"pattern-pair"*. Within *D*, the eigenvalues on the diagonal decrease in value, and describe the relative amount of covariance explained by pattern-pair *k*.

$$\frac{\lambda_k^2}{\Sigma\,\lambda_t^2}$$

which is simply the square of the eigenvalue over the sum of eigenvalues squared.

CFA analysis is focused on the comparison of pattern-pairs, or the pairing of column vectors from U_k and V_k where *k* takes a value from 1 to *r*. In order to determine the number of pattern-pairs to use, a permutation testing approach is

used. Essentially, rows are permuted on the peptide matrix, and CFA analysis is performed, the maximum eigenvalue is stored from each permutation. Significance in pattern-pairs can then be judged by taking only those with eigenvalues greater than all permuted eigenvalues. Here, 1000 permutations are performed.

Once significant pattern-pairs have been identified, the question becomes which members of the pattern-pair should be considered. In this case, each pattern-pair contains 828 transcripts and 2246 peptides. Certainly not all of the members contribute equally to the variance. We are most interested in members that have large contributions to the covariance. Therefore, we can extract members that have larger vector loadings that directly correspond to variance contributions. Transcript and peptide vectors should be considered separately. The method here involves taking the full vector a_k, one of the columns of *U* or *V*, and computing

$$u_k = X' a_k$$

which results in a vector with length equal to the number of samples. This computation is equivalent to taking the dot product by sample between the observed data and the pattern. A subset of a_k is taken based on the top *W*% of absolute values. This subset is again used to compute the score u_k^*. The correlation is computed between u_k and u_k^*. As the top *W*% increases, the similarity between the two score vectors becomes increasingly similar. Our goal is to take the smallest number of members that explains most of the covariance. I

have developed an automated method for this procedure. Epsilon can be defined as the error allowed in finding the "top" of the slope. The algorithm is listed below.

1. Compute the curve generated from correlation of u_k and u_k^* at varying Ws.
2. Fit this curve to a Loess equation.
3. At various points along the Loess curve, x, compute the slope of the tangent.
4. When the slope – epsilon < 0, return x.

Performing this method on each pattern-pair results in a set of transcripts and another set of peptides that explain a great deal of the variance observed.

However, as this is longitudinal data, we are concerned with finding sets of transcripts and peptides that vary together. The results of CFA show a mixture of entities trending up and down over time. By focusing on transcripts or peptides with large absolute values, we do not pay attention to important information contained in the sign of the value. Therefore, by taking the top *W*% of values, and then separating by sign, we have broken the set into those peptides or transcripts that trend upwards or downwards over time.

To judge whether similar genes are represented by the important peptides and transcripts in each pattern-pair, a Jaccard-index was computed between pattern-pairs.

Network Integration

Network based integration of transcript and peptide data is performed using independent co-expression networks built using each data type individually. A great deal has already been written about methods behind WGCNA, see Chapter 3 and (Iancu, et al., 2012a; Langfelder & Horvath, 2008; Langfelder et al., 2012;

Mason et al., 2009) for a small sampling of the literature. Here, signed, robust correlations were used, the scaling term, beta, was chosen so that the R^2 for peptide networks was greater than 0.8 and for transcript networks greater than 0.9, and *Partitioning Around Medoids* (PAM) was respected in branch cutting. The test involved 1000 permutations performed to assess statistical significance on connectivity in modules.

Each network is partitioned into a number of modules, each of which contains either transcript or peptide nodes. After mapping transcripts and peptides to Entrez gene IDs using the protein inference results, the overlap is measured between pairs of modules. To test the significance of each overlap, random modules are constructed by keeping the module sizes fixed and varying the contents. Ten thousand permutations are performed, and the number of overlaps greater than each module pairing is recorded. In this way, we have an empirical p-value for each overlap. Overlaps were taken as significant if less than 1% of permutations had overlaps greater in magnitude.

Once significant overlaps are found, the correlation between module eigenvectors is computed. In addition, the module eigenvectors are correlated with phenotype data to determine if overlapping modules similarly correlate with phenotype information. The combination of these three measures of correlation between modules describes an integrated signature, or more explicitly, the signature is a multi-omic collection of modules that share mapped IDs, correlation of summary eigenvectors, and correlation with sample phenotypes.

Functional enrichment

Using KEGG (Kanehisa, 2004) using the R package KEGGSOAP (J Zhang & Gentleman, n.d.), proteins found in each module provided a list of potential pathways to investigate. For each pathway returned, a hypergeometric test was performed using proteins from the module and other proteins taking part in the pathway. The universe is defined as the subset of proteins in the mass tag database with roles in known KEGG pathways. P-values are adjusted using the Benjamini and Yekutieli method (Benjamini & Yekutieli, 2001).

GO term enrichment was performed as in Chapter 3 using GOstats (Falcon & Gentleman, 2007).

Results

CFA results

Only the first pattern-pair was found to be significant after permutation testing. It is noted that this style of statistical test, in this context, may be too conservative. It might be more reasonable to consider all singular values, rather than only the highest value for each permutation. In these results, while the first singular value was quite high, the drop in singular values was extremely sharp, and quickly became close to zero (Figure 33), indicating that the first pattern-pair explains the greatest amount of variation in the covariance between transcript and peptides. The singular value for the first pattern-pair is 37.16 corresponding to 79.05% of covariance explained. In Tan's work, the first three pattern-pairs were significant, and cumulatively explained 74.8% of covariance.

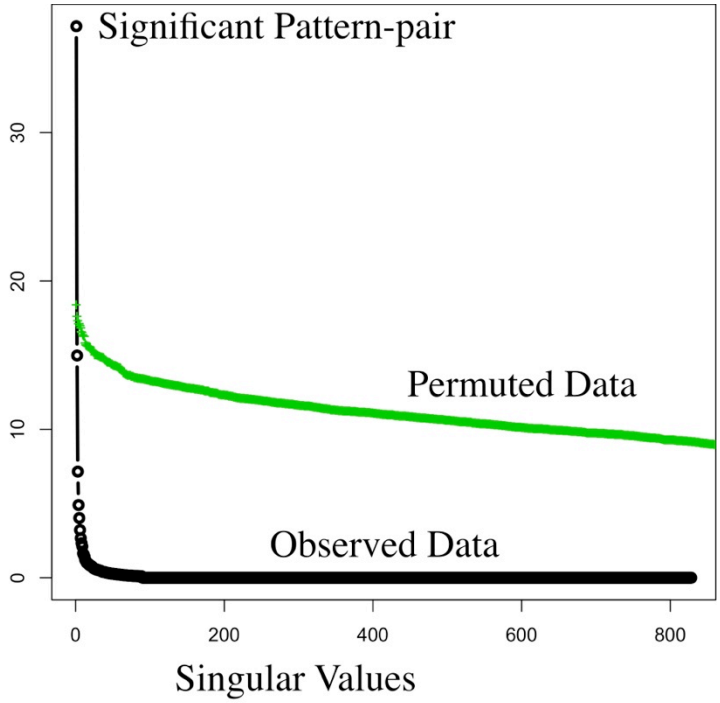

Figure 33. Significance of pattern pairs by permutation testing. The black line shows the singular values associated with each pattern pair, from left to right. Only the first pattern pair showed a singular value greater than what was generated by permutations, making it the only significant pattern pair. The green line was generated by taking 1000 permutations on the data, and each time taking only the highest singular value.

By taking the top *W%* of absolute values in the pattern-pair, and comparing this subset to the full set, separately for transcripts and peptides, we can find the smallest number of entities that explain most of the covariance (Figure 34) allowing us to focus on a much smaller number of “active” entities. Here, using the previously described algorithm, *W* was chosen that induced the tangent slope closest to 0.1 with a Loess span of 0.25 capturing most of the variance while limiting the number of entities to analyze.

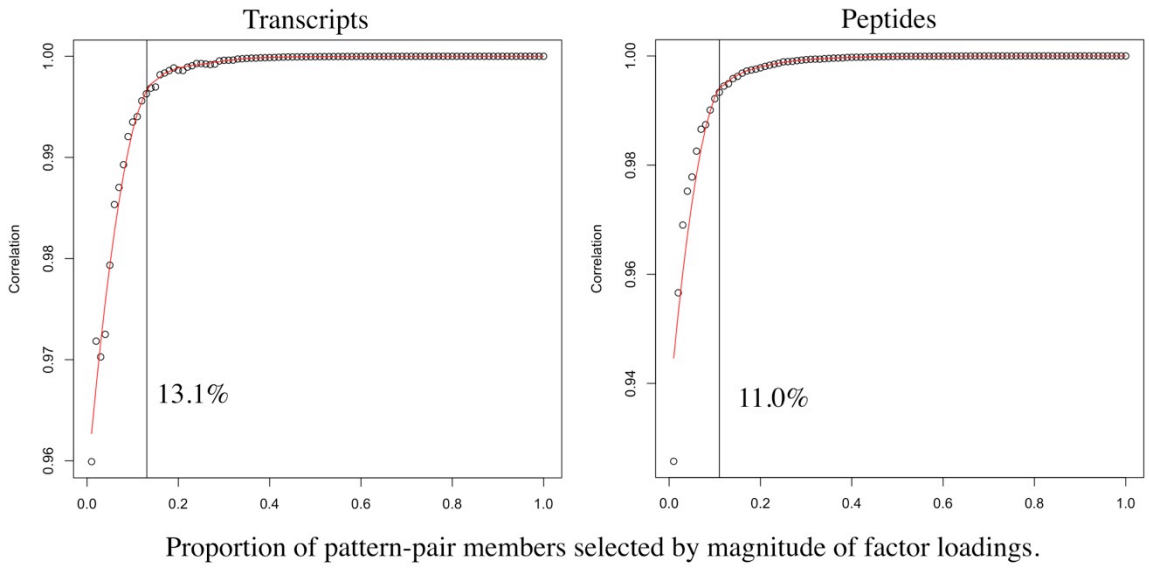

Figure 34. Selection of cutoff for important members in the pattern pair. For the left and right singular vectors separately, the transcript and peptide members are selected that explained most of the variation. The y-axis shows the correlation between scores computed using the entire pattern and a selection of members. As the selection becomes larger, the variance explained starts to become equivalent to taking all members. The selection is made by taking those that fall before the shoulder of the curve.

The top W% found were 13.1% for transcripts, and 11.0% of peptides, resulting in a set of 108 transcripts and 247 peptides, which mapped to 108 and 150 entrez gene IDs. It is encouraging that we have proteins supported by multiple peptides in the most “interesting” peptides.

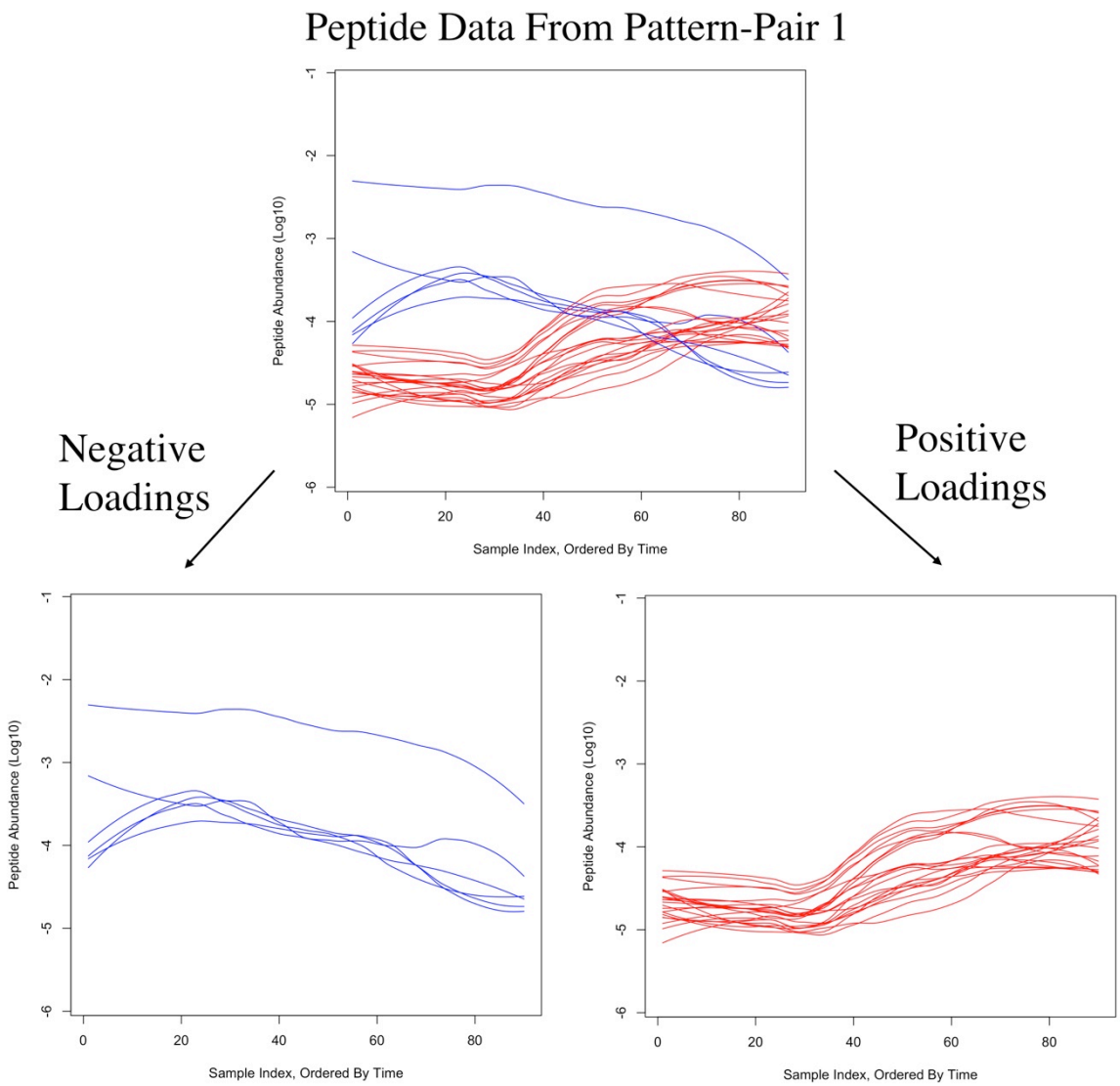

Figure 35. Decoupling pattern pair signals by abundance trends over time. When the peptides from pattern pair 1 are considered, we find members that trend both upwards and downwards over time. Since our experiment is highly concerned with temporality, it is desired to be able to separate important entities by abundance trends. By filtering significant pattern pair members by the sign of the vector loading (in the singular vectors U and V), we can easily separate the members by abundance trend over time.

Peptides and transcripts that were part of the top W% of the pattern-pair were found to trend in both directions which is quite significant for a study that is longitudinal in nature. Therefore, it is desirable to separate entities by direction over time. Temporality separation was accomplished by separating values in

eigenvectors of U and V by sign. Of the 108 genes, 63 had upward trend while 45 had a downward trend.

In terms of the gene IDs represented by peptides, 71 had downwards trends while 176 had upwards trends. An example of the separation is shown in Figure 35.

Transcript Data Results				Peptide Data Results		
Loading	Entrez	Gene Name		Loading	Entrez	Gene Name
0.216	56702	Hist1h1b		0.108	15458	Hpx
0.174	80838	Hist1h1a		0.101	15458	Hpx
0.173	14957	Hist1h1d		-0.088	12409	Cbr2
-0.156	17189	Mb		0.084	15458	Hpx
0.146	12842	Col1a1		0.082	15458	Hpx
0.130	360198	Hist1h3a		0.079	15439	Hp
0.128	16644	Kng1		0.078	15458	Hpx
-0.117	14859	Gsta3		0.076	15439	Hp
0.116	12842	Col1a1		0.075	16644	Kng1
0.113	16906	Lmnb1		0.074	15439	Hp
-0.112	14859	Gsta3		0.074	18405	Orm1
-0.111	55990	Fmo2		-0.072	100503605	Beta-s
-0.109	11657	Alb		-0.072	15129	Hbb-b1
0.103	12842	Col1a1		0.071	16644	Kng1
-0.101	14261	Fmo1		0.070	22041	Trf
-0.099	21743	Inmt		0.066	12266	C3
0.098	97165	Hmgb2		0.066	12266	C3
-0.097	55990	Fmo2		0.066	14473	Gc
0.097	97165	Hmgb2		0.064	12266	C3
0.095	319158	Hist1h4i		0.063	15458	Hpx

Table 8. Table of CFA results from pattern-pair 1. These results represent members of pattern-pair 1 that have the largest role in explaining the observed variation. The transcripts and peptides have been mapped to appropriate entrez gene IDs.

Between the extracted members of the transcripts and peptides, and after mapping to entrez gene IDs, 39 were found to be similar from a union with size 200 (Jaccard index 0.195). Permutation testing showed this overlap to be insignificant. The overlap in the pattern-pair was made worse when considering the transcripts and peptides separately by direction of trend over time. When considering only transcripts and peptides that trended downwards, the

intersection was 4. When considering transcripts and peptides that trended upwards, the intersection was 16, which is fewer than half of the intersection taking everything together, pointing out that a significant portion of the overlap was comprised of discordant transcript-peptide pairs. The discordance here might be due to entities with negligible slopes that are counted in one direction or the other. However, these partitions are made by classifying the loadings in the eigenvectors of U and V, not by slope alone.

CFA GO enrichment analysis results

Using the mass tag database as the universe, GO enrichment analysis was performed using the R package GO stats. For the transcript data, due to the very large table of results, only the lowest 10 p-values are shown along with a selection of other, interesting, GO terms (Table 9). The GO term with the greatest significance, cellular component assembly, is three levels from the root of the GO hierarchy (tree).

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0022607	2.9E-11	5.228	9.545	33	580	cellular component assembly
GO:0034622	1.3E-09	7.041	3.587	19	218	cellular macromolecular complex assembly
GO:0043933	7.7E-09	5.272	5.955	23	385	macromolecular complex subunit organization
GO:0051258	9.4E-07	24.167	0.356	6	23	protein polymerization
GO:0003013	6.9E-06	6.988	1.728	10	105	circulatory system process
GO:0006334	2.2E-05	12.775	0.592	6	36	nucleosome assembly
GO:0051259	2.6E-05	5.356	2.436	11	148	protein oligomerization
GO:0007517	2.8E-05	6.633	1.613	9	98	muscle organ development
GO:0006461	3.1E-05	5.270	2.496	11	169	protein complex assembly
GO:0007015	4.1E-05	6.274	1.695	9	103	actin filament organization
GO:0030048	1.8E-04	17.861	0.296	4	18	actin filament-based movement
GO:0006333	2.5E-04	7.792	0.905	6	55	chromatin assembly or disassembly
GO:0070836	2.7E-04	Inf	0.033	2	2	caveola assembly
GO:0006323	2.8E-04	7.635	0.922	6	56	DNA packaging
GO:0007018	2.8E-04	7.635	0.922	6	56	microtubule-based movement
GO:0030837	4.0E-04	13.881	0.362	4	22	negative regulation of actin filament polymerization
GO:0071841	4.3E-04	2.227	19.023	33	1156	cellular component organization or biogenesis at cellular level
GO:0002376	6.2E-04	2.769	7.191	17	437	immune system process

Table 9. Significant GO terms for the CFA transcript pattern. Here the most significant GO terms are associated with component and complex assembly.

While it is not extremely specialized, it is somewhat descriptive, coinciding with other members of the list including terms involving complex assembly, nucleosome assembly, chromatin assembly, caveola assembly.

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0010951	9.20E-08	7.98E+00	2.06E+00	13	83	negative regulation of endopeptidase activity
GO:0052547	1.98E-07	5.77E+00	3.41E+00	16	137	regulation of peptidase activity
GO:0051346	2.22E-07	6.14E+00	3.01E+00	15	121	negative regulation of hydrolase activity
GO:0071824	6.06E-07	9.57E+00	1.34E+00	10	54	protein-DNA complex subunit organization
GO:0030162	1.23E-06	7.63E+00	1.79E+00	11	72	regulation of proteolysis
GO:0006334	1.99E-06	1.19E+01	8.96E-01	8	36	nucleosome assembly
GO:0034097	3.32E-05	4.74E+00	2.96E+00	12	119	response to cytokine stimulus
GO:0006333	5.38E-05	7.04E+00	1.37E+00	8	55	chromatin assembly or disassembly
GO:0009611	5.43E-05	3.40E+00	5.80E+00	17	233	response to wounding
GO:0044092	6.05E-05	3.37E+00	5.85E+00	17	235	negative regulation of molecular function
GO:0006323	6.15E-05	6.90E+00	1.39E+00	8	56	DNA packaging
GO:0034622	8.54E-05	3.40E+00	5.42E+00	16	218	cellular macromolecular complex assembly
GO:0042221	1.24E-04	2.26E+00	1.80E+01	34	724	response to chemical stimulus
GO:0043434	1.44E-04	5.26E+00	1.99E+00	9	80	response to peptide hormone stimulus
GO:0016584	1.45E-04	6.02E+01	1.24E-01	3	5	nucleosome positioning
GO:0030099	3.27E-04	4.66E+00	2.21E+00	9	89	myeloid cell differentiation
GO:0022607	3.65E-04	2.26E+00	1.44E+01	28	580	cellular component assembly
GO:0019882	3.81E-04	9.67E+00	6.47E-01	5	26	antigen processing and presentation
GO:0030239	5.28E-04	1.34E+01	3.98E-01	4	16	myofibril assembly
GO:0002526	5.31E-04	6.80E+00	1.04E+00	6	42	acute inflammatory response
GO:0006414	5.47E-04	8.82E+00	6.97E-01	5	28	translational elongation
GO:0045780	7.69E-04	2.41E+01	1.99E-01	3	8	positive regulation of bone resorption
GO:0030048	8.53E-04	1.15E+01	4.48E-01	4	18	actin filament-based movement
GO:0051099	8.91E-04	7.80E+00	7.71E-01	5	31	positive regulation of binding

Table 10. Significant GO terms for the CFA peptide pattern. The top of the list shows an emphasis on regulatory GO terms.

Essentially this group of GO terms seems associated with large complex assembly, which makes sense given the context of viral infection.

When the enriched GO terms for the protein side are considered, while we do see some assembly related terms, there are many more terms involving enzymatic activity including endopeptidases, hydrolases, and responses to stimulus of various sorts (Table 10).

CFA PPI Networks

A small amount of member overlap was observed when considering the top hits from the significant pattern pairs. However, when pattern pair members were

mapped into the protein-protein interaction space, a highly connected network was found. The PPI network was constructed using the BioNetBuilder2 (Konieczka et al., 2009) plugin for the software package Cytoscape (Smoot, Ono, Ruscheinski, Wang, & Ideker, 2011). Based on work involving mapping similar genes across species (known as orthologs), a database was constructed containing what are called interologs, or orthologs with known interactions (Yellaboina et al., 2008). The database is called the Interologger database, and pulls data from other public databases such as HPRD, MINT, Bind, Biogrid, KEGG, MPPI, and others. Using a list of the genes and proteins found in the CFA results, PPI edges are placed between entities from the pattern pair. Neighboring nodes were allowed if the Interologger score (indicating sequence homology for the orthologs) was above 0.5 and the neighbor node was connected to both a transcript and peptide node. The resulting PPI network is show in Figure 36.

In the network representation, larger nodes are more responsible for the covariance observed. We see that a collection of histones and ribosomal proteins were active as well as actin associated proteins. The largest nodes tend to be those that are present in both data types. PPI edges are weighted by confidence, so that high confidence edges pull associated nodes together closely, giving us a clear way to link the pattern pair in a way that is (more) understandable by biologists, and presents a way forward in analysis, namely with graph analysis.

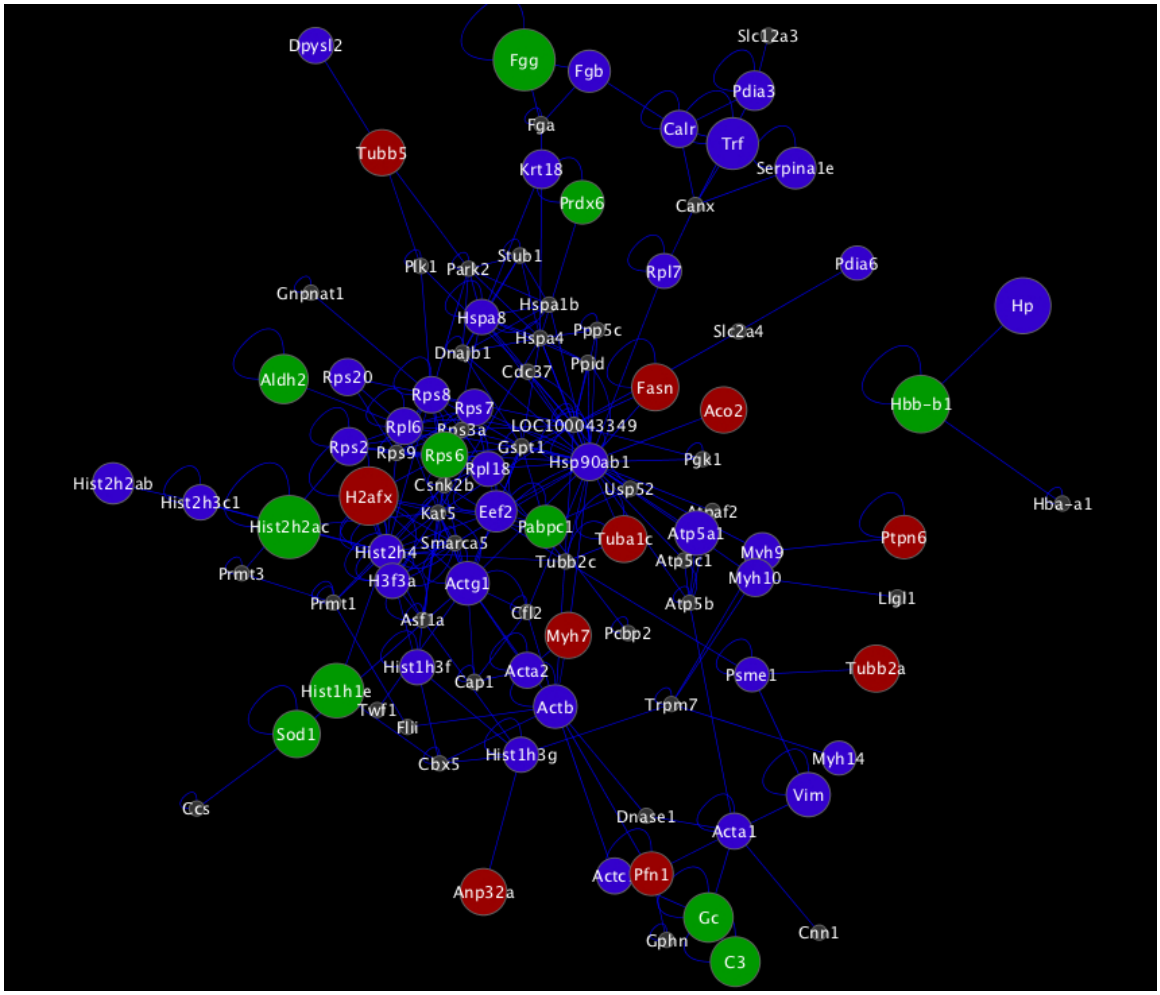

Figure 36. Protein-protein interaction network for CFA pattern pair 1. Known interactions are shown between members of the CFA pattern pair. The network shows transcripts in red, peptides in blue, and members that were shared between the transcript and peptide patterns in green. Extra nodes were allowed if the PPI score was above 0.5 that was used as a cutoff for significant edges using the Interologger database via the BioNetBuilder2 plugin for Cytoscape 2.8.

Network Integration: the most variable transcripts.

It is common when building transcript co-expression networks to use the most variable data. In fact, it is within the variance that biological signals are found. Therefore, co-expression transcript networks were constructed selecting the most variable 7000 transcripts. Of these, 5,531 had Entrez gene IDs, which produced a network containing 19 modules and was compared to a protein co-expression network containing 14 modules. Comparing these networks did not

result in any significant overlaps between modules, and further analysis was not continued.

Intersection networks

Networks were constructed by taking the entrez gene ID and sample ID (since two samples had been removed from the transcript data) intersection between quality filtered peptide and transcript data. This corresponded to 2246 peptides mapping to 445 Uniprot IDs (in the mass tag database) and 490 Entrez gene IDs. For the transcript network, 828 probes were used corresponding to 439 Uniprot IDs and 447 Entrez gene IDs. Between the two data sets, 429 Uniprot IDs and 437 Entrez IDs were shared. Some discrepancy is observed, since peptides are often degenerate, mapping to multiple proteins, which can then map from a given Uniprot ID to multiple Entrez IDs. The transcript network consists of 7 modules while the peptide network contains 14.

The intersection networks were found to be significantly overlapping. Significance here is defined empirically as less than one percent of permuted overlaps being larger than observed overlaps. By that definition nine distinct member overlaps were observed, forming three distinct subgraphs.

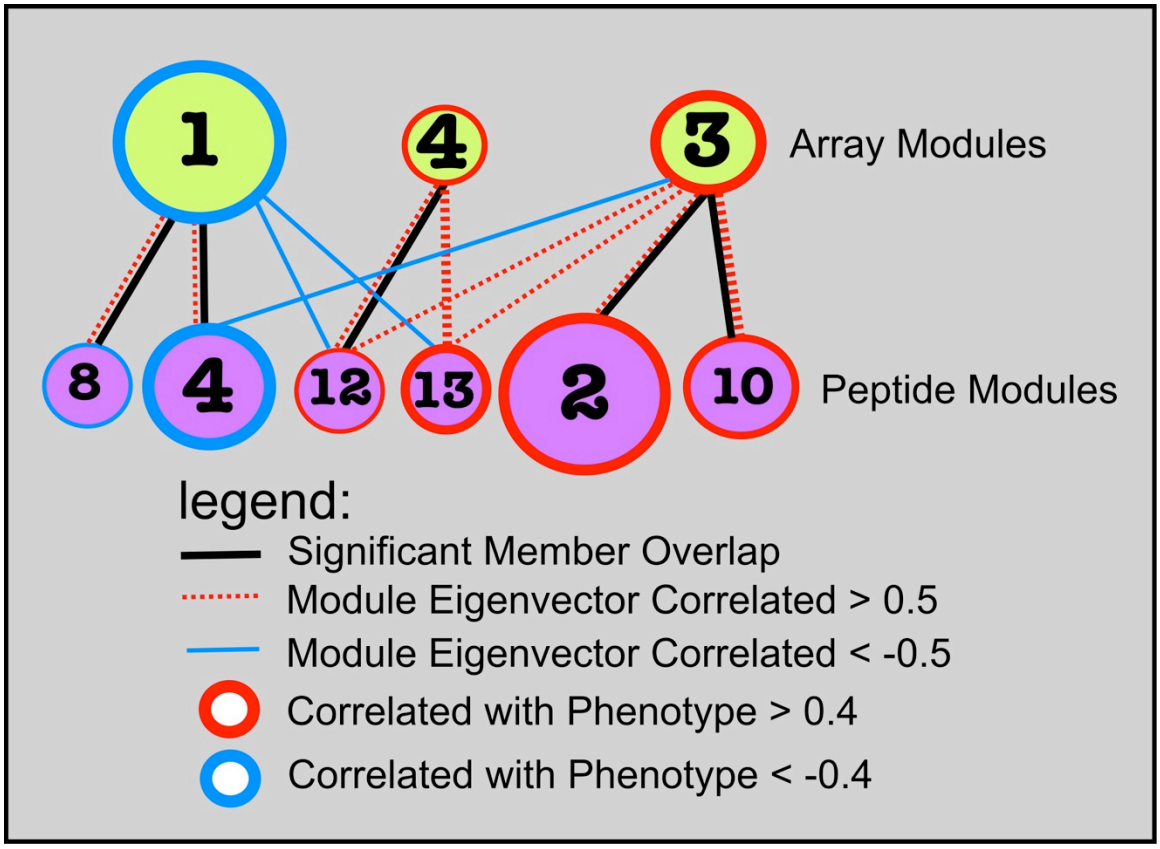

Figure 37. Multi-module, multi-omic integrated co-expression signature for SARS-CoV infection. Three kinds of edges are shown. Black edges indicate a significant overlap in module membership. Red edges indicate significant positive correlation between module eigenvectors. Blue edges indicate a significant, but negative, correlation between module eigenvectors. Node sizes correspond to module size, outlines show correlation with the overall pathology score. Although two modules might have a significant module eigenvector correlation, at times the correlation to the same phenotype can be different. The integrated graph above shows three subgraphs, (1,4,8), (4,12,13), and (3,2,10).

The overlap of module members is highlighted by the correlation between module eigenvectors (Figure 37). From the nine edges in the overlap graph, eight showed significant eigenvector correlation (p-value < 0.0006). The correlation between transcript module 1 and peptide modules 4 and 8 was 0.523 and 0.434 respectfully (p-values 1.16e-07 and 1.908e-05). The correlation between transcript module 4 and peptide modules 12 and 13 was 0.696 and 0.683 (p-values 1.159e-13 and 2.554e-14. The correlation between eigenvectors of transcript module 3 and peptide modules 2 and 10 was 0.755 and 0.801 (p-value

2.2e-16 for both). It should be explicitly mentioned that module membership overlap is not necessary for module eigenvector correlation, as seen with array module 3 and peptide modules 12 and 13 or array module 4 and peptide module 13. These are potentially very interesting sets where, in response to viral infection, a set of transcripts responds, mirrored by the response of a set of proteins that are not directly encoded by these genes. The eigenvector overlap plots in Figure 39 show joint trends over time and dosage. Array module 1 with peptide modules 4 and 8 show abundance peaks on day 1, which quickly drop to valleys by day 7. Conversely, array module 3 with peptide modules 2 and 10 show low points on day 1 and peaks at day 7.

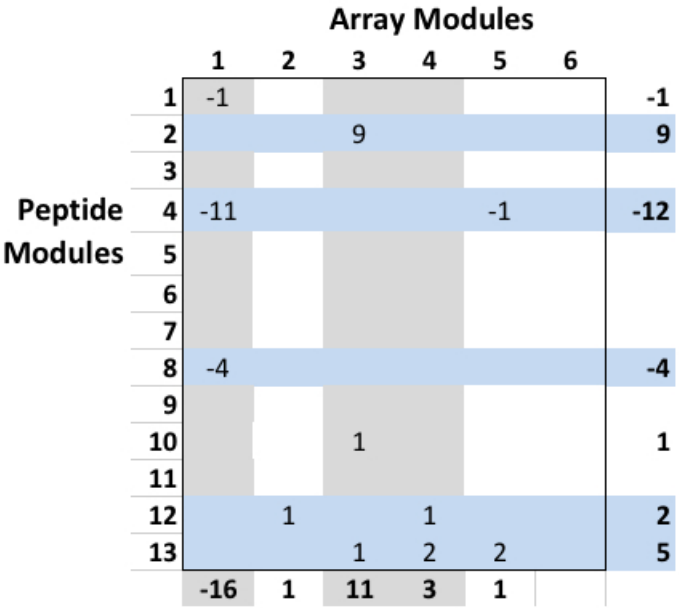

Figure 38. Relationship of modules to phenotypes. For each phenotype including diffuse aveolar damage, inflammation, edema, and the overall pathology score, both peptide and transcript module eigenvectors were correlated with phenotype, and the pair (transcript and peptide) with the highest correlations are awarded +1, while the pair of modules with the most negative correlations were awarded a -1. All other entries are 0. Clear patterns show array module 1 with the bulk of maximum negative correlations and array module 3 with the bulk of maximum positive correlations. A larger number of peptide modules appear due to multiple peptide modules associating with single array modules.

Lastly, and differently, array module 4 with peptide modules 12 and 13 seem to respond more to dosage than time. In summary, we have two different patterns of module response, one by time and the other by dosage.

The correlation in eigenvectors implicitly brings a shared correlation to sample phenotypes due to the similar vector structures between the modules. Although some variation is observed, and at times correlated eigenvectors do not share correlation with a given phenotype. However, shared correlation with phenotypes is readily observed in the overlap graph above. In the case of transcript module 3 and peptide module 2, we have strong correlations to the overall pathology score, with similar magnitudes (0.652 and 0.571 respectively with p-values 3.344e-12 and 4.158e-09). With transcript module 4 and peptide modules 12 and 13, we have similar correlation magnitudes (0.449, 0.451, and 0.552 respectively (p-values 8.969e-06, 8.088e-06, and 1.648e-08). Moving in the opposite direction, array module 1 and peptide modules 4 and 8 correlate with the overall pathology score -0.633, -0.505, and -0.352 (p-values 2.244e-11, 3.762e-07, and 0.0007).

However, given the rich set of phenotypes available, it is more telling to approach each individually, and to judge which modules are more correlated to each of the phenotypes. We sketch a brief algorithm thus: for each phenotype, the maximum and minimum correlating pair was found. A matrix is constructed where, when regarding each phenotype, if a pair of modules is maximum, +1 is added to the matrix element corresponding to this pair, and if the module pair has

the minimum correlation (i.e. negative correlation), a -1 is added to the matrix position corresponding to the pair.

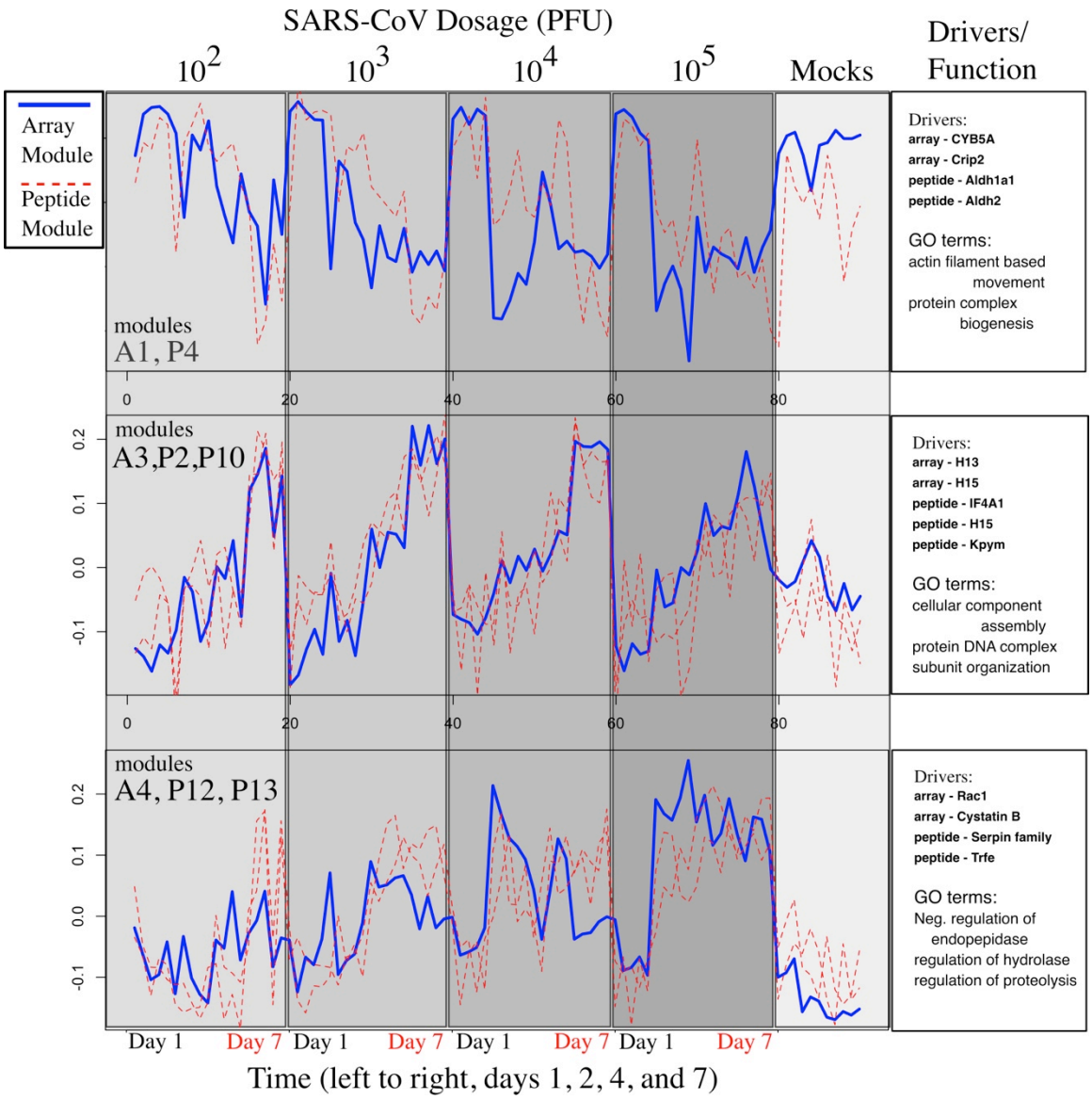

Figure 39. Overlapping eigenvectors. Similar response patterns are shown between array and peptide modules. The blue lines show array module eigenvectors plotted across an index over 92 samples. Red lines show the peptide modules. The grey background shows viral dosage arranged by time. The left side of each grey rectangle is day 1 and the right side is day 7. The top most figure shows array module 1 and peptide module 4. The middle figure shows array module 4 with peptide modules 12 and 13. The bottom figure shows array module 3 with peptide modules 2 and 10. The bottom plot is the best example of shared signal between array and peptide modules, and exhibits clear evidence for a multi-omic signature.

After integration a clear set of integrated modules becomes apparent. Figure 38 describes this set. We see that in terms of array modules, modules 1, 3, and 4 take nearly all of the possible maximum and minimum correlations. Peptide modules 2, 4, 13, 8, and 12 show the most correlations to phenotypes respectively. The phenotypes ascribed to array module 4 include OverallTotalScore, Vasculature, DAD, Eosinophils, HyalineMembrane, Exudates, Day, and Alveoli Parenchyma Pneumonia. The phenotypes ascribed to peptide module 8 include Denudation, Debris, and Edema.

Array and peptide modules, constructed after taking correlation signs into account, are separated (imperfectly) by abundance trend over time. Array modules 1 and peptide modules 4 and 8 show decreasing abundance over time, while the other modules show increasing abundance trends (Figure 40).

Annotation of Integrated Modules

Annotation mapping to transcripts and peptides were acquired by using the Uniprot web service (Apweiler, 2004; Magrane & Consortium, 2011; C. H. Wu, 2006). Within each module, each entity can be associated with a number of keywords and a protein family.

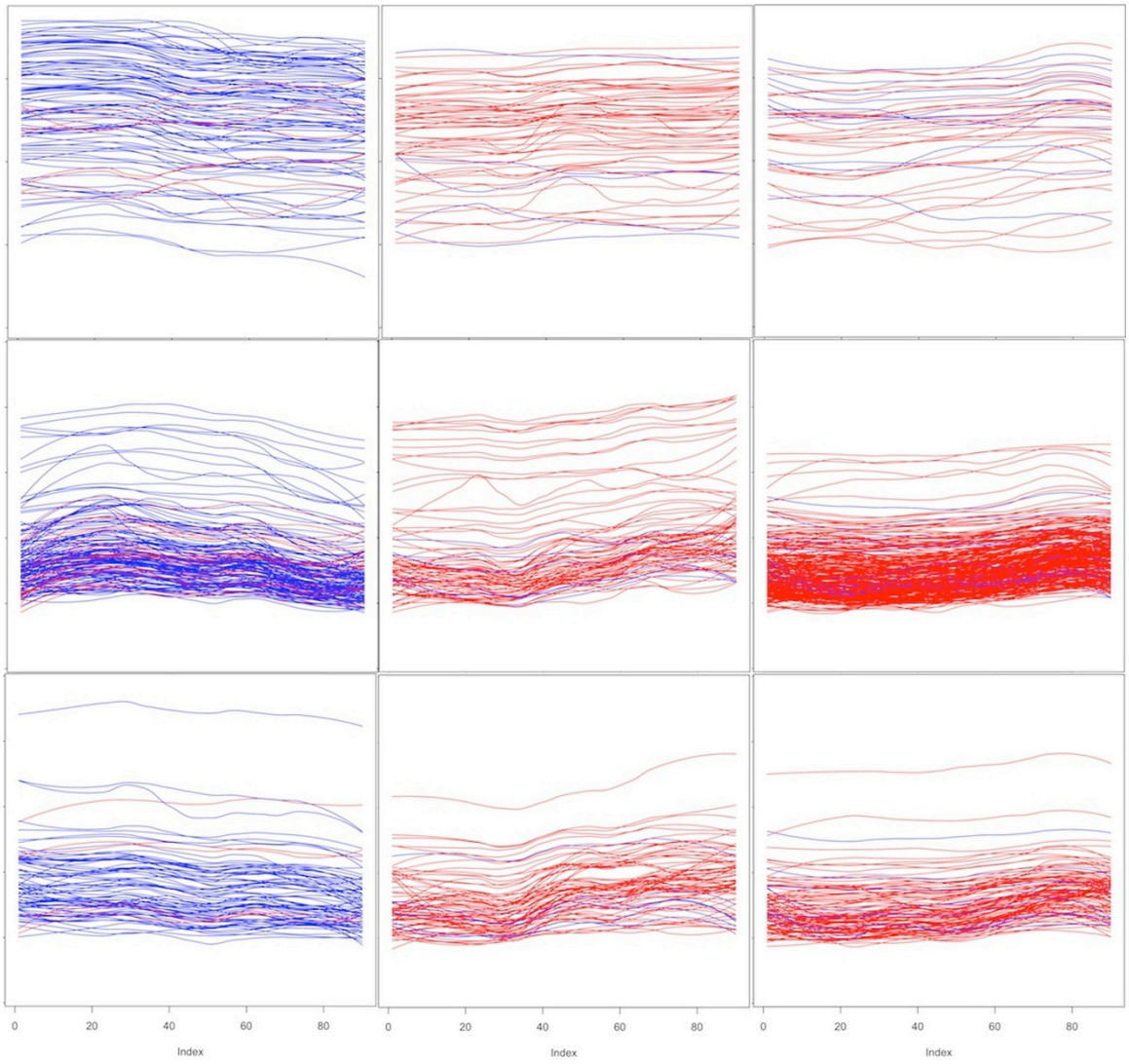

Figure 40. Abundance trends for array and peptide modules show joined modules share similar trends over time. The sample index is ordered by day, and then by viral dosage. The plots show smoothed Loess lines, that in some cases mask very noisy signals. The left most column from top to bottom shows array module 1 and peptide modules 4 and 8, in that order. The middle column shows array module 4 with peptide modules 12 and 13, and the right column shows array module 3 with peptide modules 2 and 10.

By filtering module members by correlation with the module eigenvector to which they belong, also called the centrality of a node, we can rank the members by importance. The more central a node, the more similar the expression or abundance profile is, compared to the module eigenvector. In cases where the module eigenvector correlates very strongly with a given phenotype, we are

interested is what biological entities relate most closely. Tables of the top module members are shown below in Tables 11, 12, and 13.

ME1	Array Entrez	Protein	Protein Full Name	Protein Family
0.960	109672	CYB5_MOUSE	Cytochrome b5	Cytochrome b5 family
0.955	68337	CRIP2_MOUSE	Cysteine-rich protein 2 (CRP-2) (Heart LIM protein)	
0.954	12389	CAV1_MOUSE	Caveolin-1	Caveolin family
0.948	56431	DEST_MOUSE	Destrin (Actin-depolymerizing factor) (ADF) (Sid 23)	Actin-binding proteins ADF family
0.947	14199	FHL1_MOUSE	Four and a half LIM domains protein 1 (FHL-1) (KyoT) (RBP)	
0.937	20742	SPTB2_MOUSE	Spectrin beta chain, non-erythrocytic 1 (Beta-II spectrin)	Spectrin family
0.936	20655	SODC_MOUSE	Superoxide dismutase [Cu-Zn] (EC 1.15.1.1)	Cu-Zn superoxide dismutase family
0.935	20324	SDPR_MOUSE	Serum deprivation-response protein (Cavin-2)	PTRF/SDPR family
0.927	71960	MYH14_MOUSE	Myosin-14 (Myosin heavy chain 14) (Myosin heavy chain)	
0.925	14862	GSTM1_MOUSE	Glutathione S-transferase Mu 1 (EC 2.5.1.18) (GST 1-1)	GST superfamily, Mu family
ME4	Peptide ID	Protein	Protein Full Name	Protein Family
0.917	7647003	AL1A1_MOUSE	Retinal dehydrogenase 1 (RALDH 1) (RalDH1) (EC 1.2.1.36)	Aldehyde dehydrogenase family
0.910	21465157	ALDH2_MOUSE	Aldehyde dehydrogenase, mitochondrial (EC 1.2.1.3) (AHD-M1)	Aldehyde dehydrogenase family
0.904	22221232	ALDH2_MOUSE	Aldehyde dehydrogenase, mitochondrial (EC 1.2.1.3) (AHD-M1)	Aldehyde dehydrogenase family
0.897	20975837	GSTM1_MOUSE	Glutathione S-transferase Mu 1 (EC 2.5.1.18) (GST 1-1)	GST superfamily, Mu family
0.896	26406968	ALDH2_MOUSE	Aldehyde dehydrogenase, mitochondrial (EC 1.2.1.3)	Aldehyde dehydrogenase family
0.890	21563097	FHL1_MOUSE	Four and a half LIM domains protein 1 (FHL-1) (KyoT) (RBP)	
0.881	7454569	CAV1_MOUSE	Caveolin-1	Caveolin family
0.880	20745210	INMT_MOUSE	Indolethylamine N-methyltransferase (EC 2.1.1.49) (EC 2.1.1.96)	NNMT/PNMT/TEMT family
0.876	20747700	INMT_MOUSE	Indolethylamine N-methyltransferase (EC 2.1.1.49) (EC 2.1.1.96)	NNMT/PNMT/TEMT family
0.867	20746372	SODC_MOUSE	Superoxide dismutase [Cu-Zn] (EC 1.15.1.1)	Cu-Zn superoxide dismutase family
ME8	Peptide ID	Protein	Protein Full Name	
0.882	20750889	SBP1_MOUSE	Selenium-binding protein 1 (56 kDa selenium-binding protein)	Selenium-binding protein family
0.849	20750952	PRDX6_MOUSE	Peroxiredoxin-6 (EC 1.11.1.15) (1-Cys peroxiredoxin)	AhpC/TSA family, Rehydrin subfamily
0.849	47465602	BCAM_MOUSE	Basal cell adhesion molecule (B-CAM cell surface glycoprotein)	
0.841	20750707	SBP1; SBP2_MOUS	Selenium-binding protein 1 (56 kDa selenium-binding protein)	Selenium-binding protein family
0.830	20768320	SBP1; SBP2_MOUS	Selenium-binding protein 1 (56 kDa selenium-binding protein)	Selenium-binding protein family
0.819	20746226	PRDX6_MOUSE	Peroxiredoxin-6 (EC 1.11.1.15) (1-Cys peroxiredoxin)	AhpC/TSA family, Rehydrin subfamily
0.810	7598040	ALDH2_MOUSE	Aldehyde dehydrogenase, mitochondrial (EC 1.2.1.3)	Aldehyde dehydrogenase family
0.805	20758902	SBP1; SBP2_MOUS	Selenium-binding protein 1 (56 kDa selenium-binding protein)	Selenium-binding protein family

Table 11. The ten most central module members for array module 1, and peptide modules 4, and 8. The first column shows the correlation with the module eigenvector, describing its centrality in the module. In the array module we find a cytochrome family member and an assortment of other proteins including caveolin, crip2, and destrin. On the peptide module side, we find an emphasis on aldehyde dehydrogenases and selenium binding proteins. It is thought that selenium binding protein might be involved in the sensing of xenobiotics, which could be protein imported with the virus.

In these cases, the top ten members in each module are examined. In Table 11, between module array module 1 and peptide module 4, similar keywords include Zinc, Metal-binding, Oxidoreductase, DNA-binding, Acetylation. Between array module 3 and peptide module 2 the most common shared keywords are Acetylation, DNA-binding, Actin-binding, Ubl conjugation. For module 4, the most common keywords include Protease, Proteasome, Threonine protease, ATP-binding, and Acetylation, while for peptide modules 12 and 13 the

most common keywords include Protease inhibitor, Serine protease inhibitor, Signal, and Secreted. In Table 12, for array module 4 protein families include Cystatin family, DEAD box helicase family, eIF4A subfamily, Pyruvate kinase family, Peptidase T1A family, Small GTPase superfamily, Rho family, Serpin family, Tubulin family, while for peptide modules 12 and 13 shared protein families include many from the Serpin family, ALB/AFP/VDB family, Transferrin family, and additionally for module 13 only, the Peptidase S1 family, Fetuin family and Hemopexin family.

ME4	Array Entrez	Protein	Protein Full Name	Protein Family
0.939	19353	RAC1_MOUSE	Ras-related C3 botulinum toxin substrate 1 (p21-Rac1)	Small GTPase superfamily, Rho family
0.929	13014	CYTB_MOUSE	Cystatin-B (Stefin-B)	Cystatin family
0.926	13014	CYTB_MOUSE	Cystatin-B (Stefin-B)	Cystatin family
0.888	13681	IF4A1_MOUSE	Eukaryotic initiation factor 4A-I (eIF-4A-I) (eIF4A-I) (EC 3.6.4.13)	DEADbox helicase family, eIF4A
0.875	20717	SPA3M_MOUSE	Serine protease inhibitor A3M (Serpin A3M)	Serpin family
0.871	13681	IF4A1_MOUSE	Eukaryotic initiation factor 4A-I (eIF-4A-I) (eIF4A-I) (EC 3.6.4.13)	DEAD box helicase family, eIF4A
0.869	26441	PSA4_MOUSE	Proteasome subunit alpha type-4 (EC 3.4.25.1)	Peptidase T1A family
0.864	26442	PSA5_MOUSE	Proteasome subunit alpha type-5 (EC 3.4.25.1)	Peptidase T1A family
0.863	67951	TBB6_MOUSE	Tubulin beta-6 chain	Tubulin family
0.855	18746	KPYM_MOUSE	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40)	Pyruvate kinase family
ME12	Peptide ID	Protein	Protein Full Name	Protein Family
0.950	26384724	ALBU_MOUSE	Serum albumin	ALB/AFP/VDB family
0.927	20745314	SPA3C / SPA3G / ..	Serine protease inhibitor A3K (Serpin A3K) (Contrapsin) (SPI-2)	Serpin family
0.921	20752234	CO3_MOUSE	Complement C3 (HSE-MSF)	
0.909	20778355	SPA3K_MOUSE	Serine protease inhibitor A3K (Serpin A3K) (Contrapsin) (SPI-2)	Serpin family
0.909	20751228	A1AT1/A1AT3	Alpha-1-antitrypsin 1-1 (AAT) (Alpha-1 protease inhibitor 1)	Serpin family
0.902	20748135	CO3_MOUSE	Complement C3 (HSE-MSF)	
0.902	26389720	ALBU_MOUSE	Serum albumin	ALB/AFP/VDB family
0.900	20746192	A1AT1/A1AT3	Alpha-1-antitrypsin 1-1 (AAT) (Alpha-1 protease inhibitor 1)	Serpin family
0.892	20751066	TRFE_MOUSE	Serotransferrin (Transferrin) (Beta-1 metal-binding globulin)	Transferrin family
0.888	20767488	CO3_MOUSE	Complement C3 (HSE-MSF)	
ME13	Peptide ID	Protein	Protein Full Name	Protein Family
0.952	20750867	TRFE_MOUSE	Serotransferrin (Transferrin) (Beta-1 metal-binding globulin)	Transferrin family
0.951	20758893	VTDB_MOUSE	Vitamin D-binding protein (DBP) (VDB) (Gc-globulin)	ALB/AFP/VDB family
0.931	20746152	TRFE_MOUSE	Serotransferrin (Transferrin) (Beta-1 metal-binding globulin)	Transferrin family
0.928	21759809	VTDB_MOUSE	Vitamin D-binding protein (DBP) (VDB) (Gc-globulin)	ALB/AFP/VDB family
0.924	20758798	HEMO_MOUSE	Hemopexin	Hemopexin family
0.923	20776242	CO3_MOUSE	Complement C3 (HSE-MSF)	
0.923	20758841	HEMO_MOUSE	Hemopexin	Hemopexin family
0.918	21468145	FIBG_MOUSE	Fibrinogen gamma chain	
0.912	20744484	TRFE_MOUSE	Serotransferrin (Transferrin) (Beta-1 metal-binding globulin)	Transferrin family
0.908	20750875	TRFE_MOUSE	Serotransferrin (Transferrin) (Beta-1 metal-binding globulin)	Transferrin family

Table 12. The ten most central module members for array module 4 and peptide modules 12 and 13. The first column shows the correlation with the module eigenvector, describing its centrality in the module. In this case we find in the array module, Rac1 which has a role in a wide assortment of pathways, and Cstb which is a intracellular thiol proteinase inhibitor. On the peptide module side we find other protease inhibitors, including members of the Serpin family. Serotransferrins are typically associated with the task of transporting iron atoms, but have also been associated with cell proliferation, although that latter role is less clear.

When the most central module elements are considered, we can see what protein families are most associated with clinical phenotypes. For array module 1 and peptide module 4, overlapping families include the Caveolin family, GST superfamily, Mu family, Cu-Zn superoxide dismutase family and module 4 contains members of the Aldehyde dehydrogenase family. In array module 3, central members include proteins from families of CAP family, Histone H1/H5 family, Histone H2A family, Intermediate filament family while for peptide module 2 protein families include GTP-binding elongation factor family, EF-Tu/EF-1A subfamily, Heat shock protein 90 family, Histone H1/H5 family, and the DEAD box helicase family, eIF4A subfamily (Table 13).

Overall, it appears that each set of integrated modules is somewhat discrete in its functionality. Array module 1 and peptide module 4 are related to metal binding proteins interacting with DNA and engaging in changes to Acetylation patterns. Array module 3 and peptide module 2 seems to be associated with chromatin modifications, and indeed are enriched with histones and actin binding proteins, and elongation factors. Array module 4 and peptide modules 12 and 13 are enzyme driven and have quite a few Serpin family members, which are protease inhibitors.

ME3	Array Entrez	Protein	Protein Full Name	Protein Family
0.970	14957	H13_MOUSE	Histone H1.3 (H1 VAR.4) (H1d)	Histone H1/H5 family
0.952	56702	H15_MOUSE	Histone H1.5 (H1 VAR.5) (H1b)	Histone H1/H5 family
0.937	16906	LMNB1_MOUSE	Lamin-B1	Intermediate filament family
0.933	97165	HMGB2_MOUSE	High mobility group protein B2 (High mobility group protein 2)	HMGB family
0.927	80838	H11_MOUSE	Histone H1.1 (H1 VAR.3) (Histone H1a) (H1a)	Histone H1/H5 family
0.912	97165	HMGB2_MOUSE	High mobility group protein B2 (High mobility group protein 2)	HMGB family
0.903	320332	H4_MOUSE	Histone H4	Histone H4 family
0.895	12331	CAP1_MOUSE	Adenylyl cyclase-associated protein 1 (CAP 1)	CAP family
0.892	319176	H2A2C_MOUSE	Histone H2A type 2-C (H2a-613B)	Histone H2A family
0.886	319156	H4_MOUSE	Histone H4	Histone H4 family
ME2	Peptide ID	Protein	Protein Full Name	Protein Family
0.939	7470155	MYH9_MOUSE	MYH9_MOUSE	Myh9
0.937	6964781	IF4A1; IF4A2; ..A3	Eukaryotic initiation factor 4A-I (eIF-4A-I) (eIF4A-I) (EC 3.6.4.13)	DEAD box helicase family, eIF4A
0.915	6991655	H15_MOUSE	Histone H1.5 (H1 VAR.5) (H1b)	Histone H1/H5 family
0.913	7101861	MYH9_MOUSE	Myosin-9 (Cellular myosin heavy chain, type A)	
0.912	6949196	ENPL_MOUSE	Endoplasmin(Tumor rejection antigen gp96)	Heat shock protein 90 family
0.904	21071453	MYH9_MOUSE	Myosin-9 (Cellular myosin heavy chain, type A)	
0.894	8559873	H15_MOUSE	Histone H1.5 (H1 VAR.5) (H1b)	Histone H1/H5 family
0.894	147423023	MYH9_MOUSE	Myosin-9 (Cellular myosin heavy chain, type A)	
0.891	6958762	EF1A1; EF1A2;	Eukaryotic initiation factor 4A-I (eIF-4A-I) (eIF4A-I) (EC 3.6.4.13)	DEAD box helicase family, eIF4A
0.890	26752702	MYH9_MOUSE	Myosin-9 (Cellular myosin heavy chain, type A)	
ME10	Peptide ID	Protein	Protein Full Name	Protein Family
0.919	20925153	KPYM_MOUSE	Pyruvate kinase isozymes M1/M2 (EC 2.7.1.40)	Pyruvate kinase family
0.918	6955105	COF1_MOUSE	Cofilin-1 (Cofilin, non-muscle isoform)	Actin-binding proteins ADF family
0.907	20961596	PDIA3_MOUSE	Protein disulfide-isomerase A3 (EC 5.3.4.1)	Protein disulfide isomerase family
0.897	20765979	PROF1_MOUSE	Profilin-1 (Profilin I)	Profilin family
0.895	7148581	ENPL_MOUSE	Endoplasmin (Tumor rejection antigen gp96)	Heat shock protein 90 family
0.892	8557322	PDIA3_MOUSE	Protein disulfide-isomerase A3 (EC 5.3.4.1)	Protein disulfide isomerase family
0.889	6948889	GRP78_MOUSE	78 kDa glucose-regulated protein (GRP-78)	Heat shock protein 70 family
0.886	7004018	ROA3_MOUSE	Heterogeneous nuclear ribonucleoprotein A3 (hnRNP A3)	
0.871	6964344	ROA3_MOUSE	Heterogeneous nuclear ribonucleoprotein A3 (hnRNP A3)	
0.865	6959492	HS71A; HS71B;..	Heat shock cognate 71 kDa protein	Heat shock protein 70 family

Table 13. The ten most central module members for array module 3 and peptide modules 2 and 10. The first column shows the correlation with the module eigenvector, describing its centrality in the module. Array module three comprises of histones mostly. This concentration is interesting since it shows very strong module eigenvector correlation with peptide module 10, which comprises of pyruvate kinase isozymes and an actin binding protein. Pyruvate kinase isozyme is associated with apoptosis when transported to the nucleus.

GO enrichment for intersection networks

Tabulation of GO terms shows largely similar trends as keyword and protein family analysis. In the first sub-graph including array module 1, peptide module 4 and peptide module 8, the most significant GO terms include processes involving actin filament processes, component assembly, and oxidation reduction processes (Table 14). Although there are some common themes running across the modules, the direct overlap of GO terms is very sparse.

Array Module 1		Bonnferoni Adj.				
GOBPID	Pvalue	Pvalue	ExpCount	Count	Size	Term
GO:0030048	2.04E-08	5.51E-05	0.802372	9	18	actin filament-based movement
GO:0042542	1.77E-07	4.76E-04	1.560168	11	35	response to hydrogen peroxide
GO:0043933	3.28E-07	8.86E-04	16.71448	39	386	macromolecular complex subunit organization
GO:0055114	5.29E-07	1.43E-03	16.34708	38	377	oxidation-reduction process
GO:0010035	7.22E-07	1.95E-03	5.21542	19	117	response to inorganic substance
GO:0042744	8.40E-07	2.27E-03	0.624067	7	14	hydrogen peroxide catabolic process
GO:0055002	2.27E-06	6.12E-03	1.961355	11	44	striated muscle cell development
GO:0006461	4.03E-06	1.09E-02	12.35972	30	284	protein complex assembly
GO:0006084	6.80E-06	1.83E-02	1.42644	9	32	acetyl-CoA metabolic process
GO:0006979	6.86E-06	1.85E-02	4.457624	16	100	response to oxidative stress
Peptide Module 4						
GO:0006334	1.20E-12	2.13E-09	0.902238	13	36	nucleosome assembly
GO:0022607	2.20E-10	3.93E-07	14.56112	41	581	cellular component assembly
GO:0071824	3.91E-10	6.97E-07	1.353358	13	54	protein-DNA complex subunit organization
GO:0006333	5.02E-10	8.93E-07	1.37842	13	55	chromatin assembly or disassembly
GO:0006323	6.39E-10	1.14E-06	1.403482	13	56	DNA packaging
GO:0034622	1.44E-08	2.57E-05	5.463555	22	218	cellular macromolecular complex assembly
GO:0030036	3.28E-05	5.84E-02	5.012435	16	200	actin cytoskeleton organization
GO:0071841	4.17E-05	7.42E-02	28.97188	49	1156	cellular component organization or biogenesis at cell level
GO:0045214	7.04E-05	1.25E-01	0.250622	4	10	sarcomere organization
GO:0016584	0.0001483	2.64E-01	0.125311	3	5	nucleosome positioning
Peptide Module 8						
GO:0030036	3.41E-06	4.11E-03	2.831452	13	200	actin cytoskeleton organization
GO:0006749	4.52E-05	5.45E-02	0.410561	5	29	glutathione metabolic process
GO:0051146	6.61E-05	7.96E-02	1.03348	7	73	striated muscle cell differentiation
GO:0055001	7.13E-05	8.60E-02	0.72202	6	51	muscle cell development
GO:0055114	0.0001702	2.05E-01	5.875263	16	415	oxidation-reduction process
GO:0045214	0.0003044	3.67E-01	0.141573	3	10	sarcomere organization
GO:0022607	0.0003291	3.97E-01	8.225368	19	581	cellular component assembly
GO:0046700	0.0004138	4.99E-01	5.705376	15	403	heterocycle catabolic process
GO:0015986	0.0004143	4.99E-01	0.15573	3	11	ATP synthesis coupled proton transport
GO:0030240	0.0005878	7.08E-01	0.042472	2	3	skeletal muscle thin filament assembly

Table 14. The ten most significant enriched GO terms for array module 1 and peptide modules 4 and 8. An emphasis is found on actin processes (internal cellular movement) and complex and component assembly.

In array module 4, peptide module 12 and peptide module 13, enriched GO terms tend to associate with regulation of processes. In particular, the regulation of endopeptidase, regulation of hydrolases, and response to immune signaling like cytokines (Table 15). Here the GO term overlap is strong.

Array Module 4		Bonnferoni Adj.				
GOBPID	Pvalue	Pvalue	ExpCount	Count	Size	Term
GO:0010951	4.31E-05	5.59E-02	0.968624	7	83	negative regulation of endopeptidase activity
GO:0051346	6.90E-05	8.95E-02	1.412091	8	121	negative regulation of hydrolase activity
GO:0009259	0.0001509	1.96E-01	4.66807	14	400	ribonucleotide metabolic process
GO:0052547	0.0001656	2.15E-01	1.598814	8	137	regulation of peptidase activity
GO:0009154	0.0001941	2.52E-01	4.189593	13	359	purine ribonucleotide catabolic process
GO:0009205	0.0002915	3.78E-01	4.364645	13	374	purine ribonucleoside triphosphate metabolic process
GO:0072523	0.0003409	4.42E-01	4.434666	13	380	purine-containing compound catabolic process
GO:0006163	0.0003556	4.61E-01	5.064856	14	434	purine nucleotide metabolic process
GO:0009141	0.0003588	4.65E-01	4.458007	13	382	nucleoside triphosphate metabolic process
GO:0048532	0.0003989	5.17E-01	0.035011	2	3	anatomical structure arrangement
Peptide Module 12						
GO:0010951	7.78E-18	8.36E-15	0.508131	14	83	negative regulation of endopeptidase activity
GO:0030162	5.86E-17	6.29E-14	0.440788	13	72	regulation of proteolysis
GO:0043086	1.39E-15	1.49E-12	1.144825	16	187	negative regulation of catalytic activity
GO:0051336	7.41E-14	7.95E-11	2.11211	18	345	regulation of hydrolase activity
GO:0043434	5.30E-13	5.70E-10	0.489765	11	80	response to peptide hormone stimulus
GO:0034097	1.70E-12	1.83E-09	0.728525	12	119	response to cytokine stimulus
GO:0065009	1.57E-11	1.68E-08	4.37115	21	714	regulation of molecular function
GO:0009719	8.58E-09	9.21E-06	1.181557	11	193	response to endogenous stimulus
GO:0042221	2.47E-08	2.65E-05	3.088454	15	607	response to chemical stimulus
GO:0080090	9.34E-08	1.00E-04	7.652573	22	1250	regulation of primary metabolic process
Peptide Module 13						
GO:0043086	2.62E-09	3.42E-06	1.323704	12	187	negative regulation of catalytic activity
GO:0009611	2.70E-09	3.53E-06	1.649321	13	233	response to wounding
GO:0010951	3.58E-09	4.67E-06	0.587526	9	83	negative regulation of endopeptidase activity
GO:0051336	3.59E-09	4.69E-06	2.442127	15	345	regulation of hydrolase activity
GO:0006952	7.42E-09	9.68E-06	1.790893	13	253	defense response
GO:0065009	4.70E-08	6.13E-05	5.054142	19	714	regulation of molecular function
GO:0006953	3.24E-07	4.23E-04	0.15573	5	22	acute-phase response
GO:0030162	5.18E-07	6.76E-04	0.509661	7	72	regulation of proteolysis
GO:0080134	9.53E-07	1.24E-03	1.451119	10	205	regulation of response to stress
GO:0050727	1.66E-06	2.17E-03	0.382246	6	54	regulation of inflammatory response

Table 15. The ten most significant enriched GO terms for arrays 4 and peptide modules 12 and 13. Here we find an emphasis on regulation, involving enzymes such as endopepidases, metabolic processes, and signaling involving cytokines an inflammation.

For array module 3, peptide module 2 and peptide module 10 have enriched GO terms that are associated with nucleosome assembly, and DNA-protein complex organization. Also enriched are terms with actin cytoskeleton organization and processes (Table 16). Again, there is strong GO term overlap across modules.

Overall, we see that the module set including array module 1 with peptide modules 4 and 8 and the module set including array module 3 with peptide

modules 2 and 10, both include cellular complex assembly, but where the abundance trends over time are in opposing directions.

Array Module 3		Bonnferoni Adj.				
GOBPID	Pvalue	Pvalue	ExpCount	Count	Size	Term
GO:0022607	3.78E-08	5.48E-05	9.337	28	581	cellular component assembly
GO:0034622	4.00E-08	5.80E-05	3.503	17	218	cellular macromolecular complex assembly
GO:0006334	6.66E-08	9.66E-05	0.579	8	36	nucleosome assembly
GO:0030036	7.45E-08	1.08E-04	3.214	16	200	actin cytoskeleton organization
GO:0071824	1.81E-06	2.63E-03	0.868	8	54	protein-DNA complex subunit organization
GO:0006333	2.09E-06	3.04E-03	0.884	8	55	chromatin assembly or disassembly
GO:0006323	2.41E-06	3.50E-03	0.900	8	56	DNA packaging
GO:0006996	1.02E-05	1.48E-02	11.597	27	762	organelle organization
GO:0071841	4.05E-05	5.87E-02	18.577	35	1156	cellular component organization or biogenesis at cell level
GO:0051494	0.0001299	1.88E-01	0.804	6	50	negative regulation of cytoskeleton organization
Peptide Module 2						
GO:0034622	2.35E-12	4.71E-09	7.257	31	218	cellular macromolecular complex assembly
GO:0006334	1.18E-08	2.37E-05	1.198	11	36	nucleosome assembly
GO:0051258	1.21E-08	2.43E-05	2.830	16	85	protein polymerization
GO:0009259	1.95E-08	3.90E-05	13.315	36	400	ribonucleotide metabolic process
GO:0022607	2.50E-08	5.00E-05	19.341	45	581	cellular component assembly
GO:0006163	5.09E-08	1.02E-04	14.447	37	434	purine nucleotide metabolic process
GO:0006753	5.80E-08	1.16E-04	16.478	40	495	nucleoside phosphate metabolic process
GO:0009203	1.26E-07	2.52E-04	11.784	32	354	ribonucleoside triphosphate catabolic process
GO:0071824	1.29E-07	2.58E-04	1.798	12	54	protein-DNA complex subunit organization
GO:0009146	1.44E-07	2.88E-04	11.851	32	356	purine nucleoside triphosphate catabolic process
Peptide Module 10						
GO:0006334	1.93E-10	2.93E-07	0.606	10	36	nucleosome assembly
GO:0071824	8.64E-10	1.31E-06	0.909	11	54	protein-DNA complex subunit organization
GO:0022607	6.70E-09	1.02E-05	9.782	30	581	cellular component assembly
GO:0034622	1.31E-08	1.99E-05	3.670	18	218	cellular macromolecular complex assembly
GO:0006333	1.71E-08	2.60E-05	0.926	10	55	chromatin assembly or disassembly
GO:0006323	2.06E-08	3.12E-05	0.943	10	56	DNA packaging
GO:0006259	1.05E-07	1.59E-04	4.199	18	255	DNA metabolic process
GO:0071841	2.58E-07	3.92E-04	19.462	41	1156	cellular component organization or biogenesis at cell level
GO:0030036	4.83E-06	7.32E-03	3.367	14	200	actin cytoskeleton organization
GO:0006996	5.32E-06	8.06E-03	8.578	23	572	organelle organization

Table 16. The ten most significant enriched GO terms for array module 3 and peptide modules 2 and 10. This set of terms shows an emphasis on protein-DNA complex subunit organization, DNA packaging, and nucleosome assembly. Cellular macromolecular complex assembly, also present in each list, is actually 5 levels from the root of the biological process GO hierarchy and involves the assembly of very large protein structures.

The first sub-graph declines in abundance over time, the latter increasing in abundance. The sub-graph of modules 3, 2, and 10 show a tendency towards DNA-protein interactions. The module set including array module 4 and peptide

modules 12 and 13 stand apart in the amount of enrichment associated with regulation and response to biological events. This module set seems more related to dosage rather than time.

Discussion

Using two different methods, a connected biological signal is observed across different data types. The data types, transcriptomic and proteomic, represent very different views in the cell. First, microarray measurements are fairly close to the DNA. In this case, we are observing the host response, changes in the genomic program over time, as infection progresses. On the other hand, the proteomic data represents the biological machinery actively doing work in the cell. To my knowledge, this study is the first evidence showing clear modularization of the proteome, connected to modularization in the transcriptome, in response to viral infection. Very little is known about how proteome modules change in response to biological events, or even what is typical in normal healthy cells.

CFA analysis returned a set of transcripts and peptides that were together co-varying in a significant way, and explained a large portion of the variance. Although a high degree of member overlap was not observed across the data types, by using protein-protein interaction databases, the individuals were found to be connected, resulting in a very interesting, and novel, sort of biological network, where edges represent observed PPIs, but the nodes are mixed representing genes and proteins. Inference on networks such as this might reveal new pathways important for understanding infection. In addition to the

connected network, enriched GO terms were observed to be shared across the pattern pair. GO terms including cellular macromolecular complex assembly, DNA packaging, chromatin assembly and disassembly, actin filament based movement, nucleosome assembly, and immune process related terms are shared. Although the results from CFA contain individuals with abundance trends moving in both directions, with a simple extension it was shown that the signal could be separated into groups that trend together over time.

Joined co-expression modules bring another approach to integrating data types. In this case, networks are constructed separately, and then, along with phenotypic data, modules of each type are compared to determine what combination best agrees. Agreement is defined in three ways: module membership overlaps, module eigenvector correlation, and similar correlation to phenotype. A brief remark: it is quite remarkable that eigenvectors from these distinct data types would be as similar as we observed, and that modules with very little overlap in terms of membership would have such similar responses. Clearly strong biological organization is observed which should lead to insights into viral infection.

When considering the enriched GO terms and biological entities found to be most important across CFA and joint network analysis, we find similarities, showing that by using either method results in shared entities responding to infection. Similarly, each method gives credence to the other, making it much more likely that the observed biological signals are real.

It appears that we are observing very fast acting pathways, interactions of DNA and proteins, transcription and translation, that likely evolved as fast responses to viral infection. Effectively we have discovered the beginnings of a true multi-omic signature of SARS-CoV viral infection that may have relation to other viral respiratory infections as well.

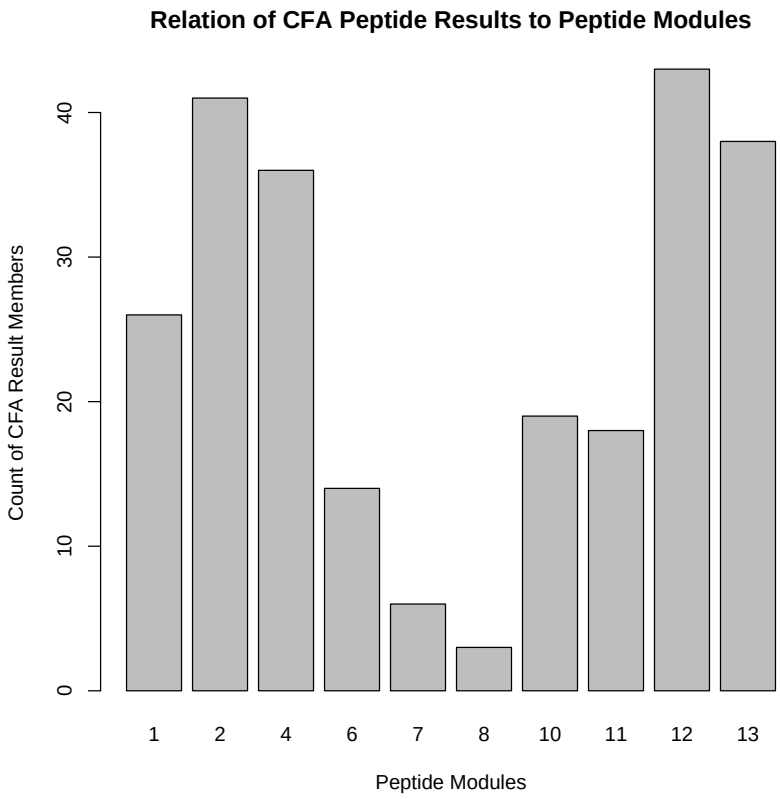

Figure 41. Mapping CFA peptide results to peptide co-expression modules shows a lack of definitive clustering.

Comparison of early and late analysis

To get a sense of how the results of CFA and joint co-expression analysis relate to one another, I compare the peptide modules to the peptide portion of the significant pattern pair. First, taking the pattern-pair subset of 247 peptides, it is observed that these peptides fall within 10 of 14 total modules.

Similar quantities of CFA peptide result members exist in the largest modules (modules 1 and 2) and the much smaller modules, such as modules 12 and 13 (Figure 41). Certainly, it is interesting to note that in modules most important to the joint co-expression module set, those peptide modules joined to gene expression modules are also those that have the largest quantity of CFA results overlapping, with the exceptions of modules 8 and 10.

When looking within the modules, the K_{me} of peptides that are part of the pattern-pair 1 subset do not have proportional loadings. Put another way, there is not a linear relationship between CFA loadings and module centrality (K_{me}) (Figure 42). This suggests that one method is not a surrogate for the other.

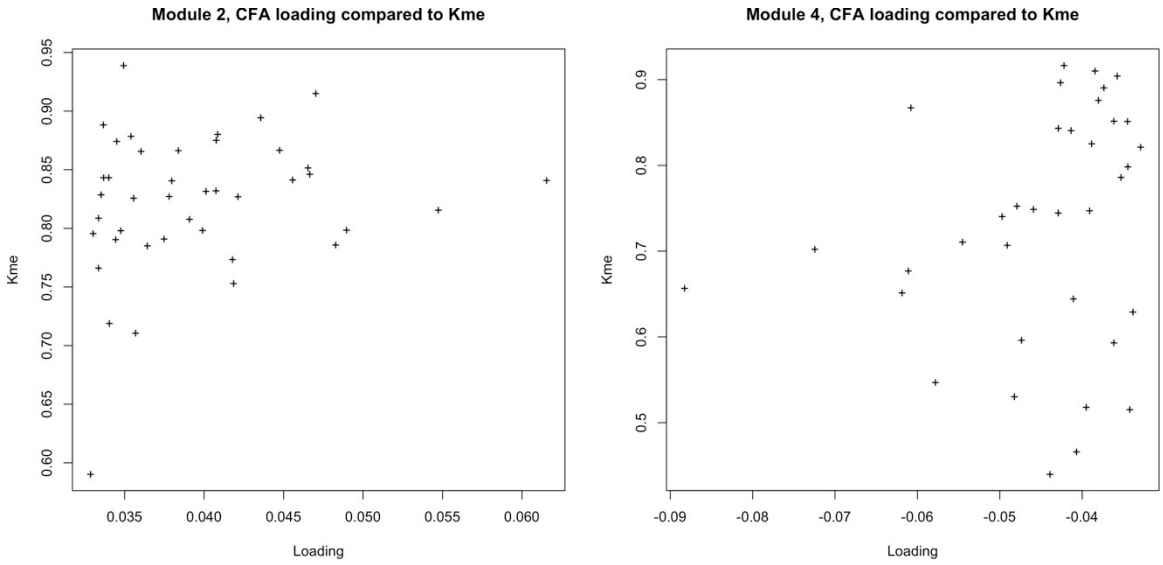

Figure 42. Relationship between CFA pattern-pair loading and K_{me} in peptide results. There is not a strong linear relationship between the module structure and the CFA loadings. However, there is a strong relationship in peptide abundance over time and the sign of the loading and module structure, showing the results of one method are not surrogates for the other.

In terms of comparing the mechanics of the two methods, CFA with few parameters to adjust, remains a much simpler process compared to the construction of co-expression modules, which require many parameters and sub-methods such as clustering and branch cutting, making the process complex and

variable. The flexibility comes with a reward however. The structure of the network is rich, and we find a much greater proportion of the data is useful for analysis. Even just considering the ability to relate phenotypes to modules, which is currently not present in CFA analysis, this richness is seen easily.

Correlated factor analysis resulted in one significant pattern pair and this is likely reasonable. It is thought that each pattern pair might correspond to distinct cellular activities and pathways, and would thereby allow us to decompose cellular activity into manageable pieces. The intersection is taken between gene and protein data in creating the integrated signature, removing a large portion of the transcript variance. It is common in network construction to take the most variable transcripts, and to use those in the network. But when transcript data was subset by peptides present in at least 80% of samples, the variance ranking on transcripts spanned the entire range meaning that our transcripts represented the entire range of variance possible. Many of the transcripts were of low variance. In this light maybe the result found with CFA is reasonable, that only a smaller portion of transcripts accounted for much of the variance, because there was little to start with.

However, the integrated networks were constructed and were found to be quite significant when considering mean topological overlap among the nodes in the graph. In addition, the significant GO terms strongly overlapped with those found in CFA analysis.

Co-expression modules tend to separate the data by abundance trend over time, while CFA does not. I was able to show that with a small change in the

CFA method members could be separated by trend. Also, the modules appear to apply some degree of functional sorting as evidenced by the GO term analysis, as well as by protein family.

The discovery of eigenvectors shared between peptide and transcript data, strongly and significantly correlating to clinical phenotypes shows a true multi-omic signature for viral infection.

Conclusions

Looking at integrated data is important in the future of systems biology. Here we have an example of how integrated analysis can shed light on the problem of pathology. With these methods it is surely possible to learn new and interesting things about the relationship between the host and pathogen, but also more generally between the transcriptome and the proteome. Systems biology is about relationships, and there is no better way to learn than by taking the larger view provided by data integration.

6. Discussion of Aims

Completed Work

This work carried though a unified arc connecting one cellular dimension to another. First, I showed that deriving protein co-expression networks using peptide-level data are feasible and contained interesting biological information. Then a novel method of protein inference was developed with a focus on tag-based proteomic data. Finally, with an ability to confidently map peptides to proteins, it is shown that integration is possible in two ways using either Correlated Factor Analysis or by the joining of co-expression modules. It was shown that the joined modules shared biological enrichments, potentially showing biological pathways in the cell that respond to cellular events in parallel, a biologically important finding.

Peptide Networks

Using tag-based proteomic data, the first de novo peptide networks were constructed using methods primarily developed for investigations of gene regulation. It was shown that these networks have a similar scale-free topological structure, and that peptides naturally grouped into modules, or sub-networks. Peptides within a given module are more connected to one another than to members of other modules. Modules were summarized taking the first eigenvector from data corresponding to the module, and were found to correlate strongly with clinical phenotypes. Within certain modules, strongly correlating with phenotypes, we see trends where higher centrality within the module (by

correlation with the module eigenvector) leads to higher correlation with the phenotype. This observation has been made in gene and transcript networks, and is considered the mark of a valuable signature describing disease. It is informative that we would also see the effect in peptide networks.

These results are significant since the method allows one to define an aggregate signature. The signature does not consist of a single protein, or even a single peptide, but an group of peptides that are highly related to each other in terms of similar abundance profiles over samples, and potentially through protein-protein interactions. The aggregate signature may contain multiple proteins, supported by different numbers of peptides. In gene and transcript co-expression networks, it is thought that highly connected members may in fact be reflective of protein-protein interactions in the cell. However, it seems much more likely that these types of interaction inferences may be better realized using proteomic data for obvious reasons. Similarly, results show highly significant pathway enrichment, and large numbers of known PPIs within modules. Again, we would expect that if pathways are being modulated, it would be observable in proteomic data. These ideas are supported by the observations that peptides are more connected when considering them grouped by protein (compared to random groups). Also, strongly connected peptides are concordant considering trends of abundance over time.

Although we observe proteins that split between different modules, it is possible that these disconnected groups of peptides are in fact reflective of different protein isoforms, interactions, or contexts. These methods give us the

potential for disambiguating between protein isoforms or pathway roles, that were previously unattainable.

Protein inference

A new protein inference model, tailored to tag-based high throughput proteomics, was developed and validated using both simulated data and two real data sets by comparing results to what I consider is the state of the art in protein inference. This new model, which I have termed “the Flow model”, represents a new approach to the protein inference problem by using methods similar to those used in network flow problems.

Simulated data was produced by building a harness around the MSSimulator, part of the TOPP proteomics processing pipeline. The harness allowed an easy way to generate a large number of samples using a known set of input proteins providing a way to formally test protein inference methods with known true positives.

However, the vast number of parameters available to the MSSimulator makes it difficult to tune. To get around this challenge, it was observed that the annotation graphs from simulated data were in different proportions that what is observed using real data. Therefore, a simple graph pruning algorithm was applied to the simulated annotation graphs, bringing the proportion of each graph class in line with the observed data. It remains unclear whether other aspects of the data are disturbed by the graph pruning. In this way, by starting with a set of proteins similar to what we find in real data, we are able to produce a set of

simulated peptides, likely be observed, and appearing in realistic proportions of annotation graph classes.

Most protein inference implementations are focused on tandem mass spectroscopy, and as such, require the output of peptide search algorithms such as Peptide Prophet or MASCOT making them unusable for tag-based proteomics, since the assumptions and the software used are quite different from tag-based proteomic work. However, a recently implemented model, Fido, is based on a probabilistic model and does not require tandem MS output files, though it does require PeptideProphet probabilities. In a review article written by the model designer, Fido is compared to other “main stream” protein inference methods, potentially giving us an indirect way to compare our model to others.

The Flow model, which “flows” quality information from supporting peptides to proteins, describes a simple computation that compares very favorably to the Fido method, which depends of the inference of latent variables requiring sophisticated computation methods. The sum of information is rewarded for uniqueness and punished for degeneracy and missingness, and scaled by a SVM model trained using the mass tag database. This SVM is used for prediction of detectability. A protein is considered detectable if, after digestion using trypsin, it contains peptides that are themselves detectable by the mass spectroscopy platform. The mass tag database provides a record of what peptides were in fact found to be detectable and therefore embodies an excellent source of information for training such a classifier.

The Flow model performed very well using simulated data, and generally compares very well to Fido. One difference lies in the fact that the Flow model takes missingness into account whereas Fido does not. In the Flow model, if a peptide is observed in only 10% of the samples, reflecting the amount information that is passed to the protein. Fido on the other hand simply takes the fact that the peptide was observed, and regards it no further. So, in summary, the Flow model is appropriate for use with tag-based proteomics, compares well with other methods but additionally takes missingness and a prior detectability into account.

Data Integration

Integration of transcript and proteomic data was shown both feasible and useful using two different methods. First, representing more of an “early” approach to integration, the CFA method was utilized, which essentially takes the singular value decomposition of the covariance matrix for the two types of data. The covariance matrix shows what transcripts and peptides vary together during infection over time. The decomposition of the covariance allows us to find groups of peptides and transcripts that, taken together, explain most of the variance observed. The decomposition of the cross-covariance matrix results in “pattern-pairs”. In this work, using permutation testing, only one pattern-pair was found to be significant. The members of the pattern-pair are taken if they have eigenvector loadings that are above a given threshold. The pattern-pairs might be considered a joint aggregate signature for infection by SARS-CoV.

However, longitudinal data was not considered in the original work describing CFA. In this work, I was able to show that with a simple change, it was possible to separate the resulting sets of peptides and transcripts into groups that had similar abundance trends over time, important for describing dynamic systems in response to biological perturbations and events. The patterns resulting from CFA did not overlap in terms of gene identities, but did overlap in terms of enriched GO terms. The transcripts and peptides were found to take part in a connected protein-protein interaction network that contained shared members as well as transcript nodes and peptide nodes. This biological network is highly interesting and possibly holds novel information describing biological events.

The other, “late” integration approach involved taking previously constructed co-expression networks and joining modules by rules of relation. The modules were joined by observing their relationships in three ways. First, simply a count of overlapping members after mapping to a common Entrez identifier. This overlap count was confirmed as statistically significant using permutation testing. Then overlapping modules were compared by correlation between the module eigenvectors. Lastly, it was observed that the joined modules also similarly correlated with phenotypes in magnitude and direction. Therefore, we have modules that overlap by member, correlate by eigenvector, and correlate similarly to phenotypes. By using these metrics, three module sets were found that represented a multi-omic signature of SARS-CoV infection. The joined modules were found to have similar GO term enrichment, since they were

composed of similar protein families and highly similar eigenvectors that literally overlap one another. The joined modules also share distinct functional signatures.

Links were found between integrated results found by the CFA method and those joined modules from integrated co-expression analysis in terms of the enriched functional signatures on the integrated sets. These results show sets of biological entities that responds in similar ways to viral infection (by SARS-CoV). The result of transcript and peptide modules that co-vary so similarly in a temporal context potentially shows which biological pathways exist in such a manner as to rapidly respond in situations of host defense, such as the innate immune response.

Potential Avenues for Exploration

This work has opened many avenues for further research, and in this portion I will discuss the potentially fruitful topics for further investigation. First I will discuss aspects of peptide network construction that could be improved, then I'll discuss protein inference and data integration.

Peptide Networks

In building peptide networks, the first step involves transforming the data into normalized abundance measures. In my work, normalization involved taking the output from VIPER, dividing each peptide abundance by the total sum of abundances for the particular sample, and log transforming. In comparison to transcript microarray data, which typically uses quantile normalization, this form of normalization is weaker (not necessarily a bad thing). It would be interesting to

DATA	IMPUTED	POWER	R^2	MEAN K	MED K	SD K
Sarcopenia	None	7	0.83	21.99	13.26	23.05
Sarcopenia	KNN	7	0.81	24.40	15.10	25.07
Sarcopenia	Linear	7	0.84	18.90	10.70	20.63
SARS	None	8	0.78	8.31	5.84	7.95
SARS	KNN	8	0.80	8.52	5.98	8.06
SARS	Linear	8	0.82	7.42	5.01	7.39
Influenza	None	5	0.83	17.49	12.86	14.89
Influenza	KNN	5	0.80	18.10	13.70	15.33
Influenza	Linear	5	0.81	15.90	11.60	13.80

Table 17. Imputation was carried out in two ways, using KNN[29] and estimation with linear models. A general trend is observed where linear estimation results in higher scale-free topology fit (R^2) but lower mean connectivity and variance. In all examples, KNN resulted in the highest mean connectivity and standard deviation. R^2 describes the scale-free topology fit. Definitions of mean K, median K, sd K: network connectivity using the adjacency matrix. mean, median, and standard deviation on network connectivity using the adjacency matrix.

test different methods of normalization, and potentially the removal of systematic artifacts, and to judge how the resultant networks change. I did do some preliminary work on performing robust normalization, taking the difference between each peptide abundance and the sample median, and dividing by the IQR (inter quartile ratio) which lines up the medians across all samples as well as the 1st and 3rd quartile whiskers. The effect was lesser connectivity in the network as well as a smaller R^2 fit to scale-free topology. But whether the network would align better to the more normalized transcript data is not known.

After data normalization, the next thing to consider would be the approach to missingness. Correlation matrices require that most of the data be present. So that presents us with two choices, we can use nearly complete data at the cost of leaving out many identified peptides, or we can attempt to perform imputation on missing data.

Imputation was evaluated using both K nearest neighbors (KNN) and estimation with a linear model (LM) (Hastie, Tibshirani, Narasimhan, & Chu, n.d.). The linear model simply uses group as the predictor variable and peptide intensity as the response. Data was imputed for peptides with less than 5% missing data and subsequently used for network construction (Table 17). Although network statistics were generally comparable, networks built with KNN imputed data showed a slightly reduced fit to the scale-free model, but higher mean and variance in terms of connectivity. The accuracy of imputation was estimated using simulated missing data. For each dataset, peptides with complete data are selected. Then 1%, 5%, 10%, and 20% of data points are randomly removed. After imputation, the mean, median, and standard deviation of the error distribution (imputed – real) is examined.

Both the KNN and LM error distributions are centered at zero, but skewed in towards underestimation (see Supplementary Figure 3). QQ plots of real and imputed data show that both methods strongly underestimate abundance at low ranges and overestimate at the high range.

In all cases, there is no clear trend in mean error, or the standard deviation of error, with increasing amounts of missing data. Overall, KNN error distributions tend to be much narrower than LM-produced error distributions. The maximum observed standard deviation in the SARS KNN imputed error distribution is 0.1485 while in the LM imputed data it is 0.2292 on the log10 scale (see supplementary Table 1 for all methods and data).

To quantify the effect of imputation on network construction, modules from two networks are pairwise compared using Fisher's exact test for the overlap of module members. A 2x2 table is shown below, where each variable is a count of members.

Net1 / Net 2	In Module A	Out Module A
In Module B	X	Y
Out Module B	Z	W

For each pair of modules, the Fisher's exact test result and the number of overlapping peptides is recorded. This information is used to calculate a comparison score, defining the quality of module alignment between networks. The comparison score, $c/min(n,m)$, is defined where c is the number of significant module overlaps, n and m are the number of modules in the networks. An overlap is counted if the intersection contains greater than 10 peptides and the Fisher's test p-value is less than 0.05 after a Bonferroni correction over all pairs.

A perfect module-wise one-to-one network alignment produces a score of 1, whereas randomized modules, on average, produce a score close to zero ($c << min(m,n)$). When module overlaps become non-specific, one module overlapping with several others, the score rises above one ($c > min(n,m)$).

As the amount of missing data increased, comparison scores increased, indicating diminished network alignments. Without imputation, SARS networks showed mean comparison scores in the range of 1.19 to 1.39, increasing with the amount of missing data. Influenza scores increased from 0.98 to 1.10 and Sarcopenia comparison scores increased from 1.41 to 1.91.

The KNN imputed data produced network alignments that generally tracked well with the non-imputed comparisons. For the sarcopenia dataset, by far the largest set of complete data, imputing with KNN led to better network alignments (0.29 mean comparison points lower), while in the other datasets the alignments were very similar (mean comparison point different 0.04 and -0.04 for SARS and Influenza respectively). The LM-imputed networks show similar results except in the case of 20% missing SARS data, where the comparison score was 2.08, indicating a large degree of non-specific overlaps.

A more detailed assessment of imputation is needed to determine what degree and method will lead to the most robust and confident networks. By allowing more missing data, a more diverse population of peptides enters the model, increasing the information content. At the same time, with more missing data, we have greater uncertainty in correlations and calculations involved in network construction. With larger numbers of samples, it becomes more likely that the computed correlations are closer to the true correlations, but the magnitude of error is still unknown. Using more robust correlation methods such as bootstrapping, Tukey's bi-weight, or multiple imputation techniques might mitigate this problem.

Two other topics that might have a strong effect on the interpretation of protein co-expression networks include feature selection on peptides that go into the network, and correlation variance "carry-through". First, in the co-expression network, the number of peptides supporting any given protein is widely variable. Many proteins have only a single peptide, while others have tens. It is possible

that proteins with large numbers of peptides could be biasing the network structure. To combat this bias, after protein inference when the set of identified proteins is decided on, the best one to three peptides for each protein could be selected according to the criterion used in Flow model protein inference, putting all proteins on “even ground” and allowing masked peptides to become more important in the network. Secondly, it would be interesting and useful to estimate the variance of correlations using bootstrap methods, and to determine what connections in the network are more variable and less confident. If these confidence measures could be passed to the branch cutting methods that derive the modules, then a great deal more confidence could be given to modules acting as signature of disease.

Protein Inference

With more data sets, a greater variety of simulated data could be generated spanning a larger portion of the proteome. The simulated sets should have more varied annotation graph class proportions, meaning cases with fewer singlets and more complex graph structures. More data would allow better comparisons between Fido and the Flow model as well.

Currently, the score that results from the Flow model, for each protein, is at minimum zero, but has no upper limit. It would be more intuitive if the score resulting from the graph portion of the model, the quality information flow, was probabilistically framed. This would match up with the prior, and would also be more comparable with other probabilistic methods. One way of making these comparisons might be to simply divide the Flow score (before the prior is applied)

by the possible maximum sum if all quality information was perfect and the peptide had been measured for all samples, giving the ratio of the observed score to the potential maximum score. It would then be enlightening to compare the probabilities attained with those from Fido, a probabilistic method.

Lastly, the Flow model would benefit from a better prior model. Currently, the prior uses a collection of features attained from three different studies on predicting protein detectability. However, it was found in those studies that changes in platform and experiment led to different sets of features that performed best. It is therefore uncertain whether there exists a best subset of features that works “well enough” in all cases, or whether each protein inference must perform a feature selection step when training the SVM. Training of the prior is the most time consuming step, so improvement here would carry on to improve the usability of the method.

Data Integration

Annotation, connecting the transcript and peptide to their correct source gene, can be one of the most difficult aspects of data integration,. For transcripts, this annotation is straightforward since microarray probes have been designed specially for avoiding degeneracy among genes and have good documentation for each. On the other hand, given a peptide, it can be quite difficult to determine what gene it resulted from. When considering theoretical proteins found in databases such as TREMBL, it quickly becomes daunting. Our knowledge of the proteome is still rapidly expanding, directly affecting our peptide-transcript integration solution.

After we have produced a set of joined modules, the next step would be to aggregate the summary eigenvectors into a single signature. This meta-vector would correlate, and thereby connect, each of the modules, both transcript and peptide. This single meta-vector would also show the correlation strength, and connectivity, of the joined modules to phenotype, simultaneously. The search for a meta-vector would involve finding a vector so that the angle between each module eigenvector and the meta-vector is minimized. This meta-vector would reflect cellular activity that is taking place across the fundamental biological molecular entities, on a systems level.

Lastly, I must state the importance of increasing the peptide data coverage (with respect to the proteome), which lets us build intersection co-expression networks that include more members, containing more variability and allowing us to find increasingly interesting and important biological signatures.

7. Summary

In this work, a new strategy for data integration has been developed and shown to hold promise for advancing systems biology. The work was applied to a data from a large-scale infectious disease study that aimed to compare influenza and SARS infections in mice. The data was generated by dosing mice with varying concentrations of virus, and measuring gene expression and protein abundance in lung tissue at four time points.

First, using methods normally reserved for gene expression studies, the first de novo protein co-expression networks were constructed using peptide

level data. The networks are shown to be feasible, novel, and composed of functional modules. The networks appear useful for protein biomarker discovery.

Second, a novel method for protein inference was developed using ideas from network flow that scored each protein with summarized peptide information, producing the first method of protein inference aimed at high-throughput tag-based proteomics and also represents a new approach to the problem. The method was tested on simulated data, and compared to a more established model using real data.

Finally, using the previous results, two methods of data integration were explored: correlated factor analysis and joint co-expression network analysis. Both methods showed potential in discovering integrated network signatures of disease. The joint co-expression analysis shows potentially the first evidence for modularization in the proteome, mirroring what is known about the transcriptome.

References

Aderem, A., Adkins, J. N., Ansong, C., Galagan, J., Kaiser, S., Korth, M. J., Law, G. L., et al. (2011). A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *mBio*, *2*(1), e00325–10. doi:10.1128/mBio.00325-10

Adourian, A., Jennings, E., Balasubramanian, R., Hines, W. M., Damian, D., Plasterer, T. N., Clish, C. B., et al. (2008). Correlation network analysis for data integration and biomarker selection. *Molecular BioSystems*, *4*(3), 249. doi:10.1039/b708489g

Aebersold, R. (2003). Mass spectrometry-based proteomics. *Nature*.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of cell science*, *118*(Pt 21), 4947–4957. doi:10.1242/jcs.02714

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. doi:10.1103/RevModPhys.74.47

Apweiler, R. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, *32*(90001), 115D–119. doi:10.1093/nar/gkh131

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Baker, E. S., Livesay, E. A., Orton, D. J., Moore, R. J., Danielson, W. F., Prior, D. C., Ibrahim, Y. M., et al. (2010). An LC-IMS-MS Platform Providing Increased Dynamic Range for High-Throughput Proteomic Studies. *Journal of proteome research*, *9*(2), 997–1006. doi:10.1021/pr900888b

Bankhead, I. M. (2010). Network Guided Disease Classifiers. *Bimolecular Interaction and Disease: Function and Disease, Keystone Symposia on Molecular and Cellular Biology.*

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. doi:10.1038/nrg1272

Barnard, D. L. (2009). Animal models for the study of influenza pathogenesis and therapy. *Antiviral research*, *82*(2), A110–22. doi:10.1016/j.antiviral.2008.12.014

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bergmann, S., Ihmels, J., & Barkai, N. (2003). Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Biology*, *2*(1), e9.

Bertsch, A., Gröpl, C., Reinert, K., & Kohlbacher, O. (2011). OpenMS and TOPP: open source software for LC-MS data analysis. *Methods in molecular biology (Clifton, N.J.)*, *696*, 353–367. doi:10.1007/978-1-60761-987-1_23

Bhardwaj, N., & Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, *21*(11), 2730–2738. doi:10.1093/bioinformatics/bti398

Bielow, C., Aiche, S., Andreotti, S., & Reinert, K. (2011). MSSimulator: Simulation of mass spectrometry data. *Journal of proteome research*, *10*(7), 2922–2929. doi:10.1021/pr200155f

Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. *Nature*, *468*(7325), 851–854. doi:10.1038/468851a

Browne, M. W. (1980). Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, *33*, 184–199.

Carlson, M., Zhang, B., Fang, Z., Mischel, P., Horvath, S., & Nelson, S. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, *7*(1), 40. doi:10.1186/1471-2164-7-40

Carter, S. L. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, *20*(14), 2242–2250. doi:10.1093/bioinformatics/bth234

Casadevall, A., & Pirofski, L.-A. (1999). Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infection and immunity*, *67*(8), 3703–3713.

Cawthon, P. M., Marshall, L. M., Michael, Y., Dam, T.-T., Ensrud, K. E., Barrett-Connor, E., Orwoll, E. S., et al. (2007). Frailty in Older Men: Prevalence, Progression, and Relationship with Mortality. *Journal of the American Geriatrics Society*, *55*(8), 1216–1223. doi:10.1111/j.1532-5415.2007.01259.x

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *Structural Approaches to Sequence Evolution*, 207–232.

Cox, B., Kislinger, T., & Emili, A. (2005). Integrating gene and protein expression data: pattern analysis and profile mining. *Methods (San Diego, Calif.)*, *35*(3), 303–314. doi:10.1016/j.ymeth.2004.08.021

Cox, J., & Mann, M. (2007). Is Proteomics the New Genomics? *Cell*, *130*(3), 395–398. doi:10.1016/j.cell.2007.07.032

Cox, J., & Mann, M. (2011). Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. *Annual Review of Biochemistry*, *80*(1), 273–299. doi:10.1146/annurev-biochem-061308-093216

Craig, R., & Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*, *20*(9), 1466–1467. doi:10.1093/bioinformatics/bth092

Daemen, A., Gevaert, O., De Bie, T., Debucquoy, A., Machiels, J.-P., De Moor, B., & Haustermans, K. (2008). Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 166–177.

de Jong, M. D., Simmons, C. P., Thanh, T. T., Hien, V. M., Smith, G. J. D., Chau, T. N. B., Hoang, D. M., et al. (2006). Fatal outcome of human influenza A (H5N1) is associated with high viral load and hypercytokinemia. *Nature medicine*, *12*(10), 1203–1207. doi:10.1038/nm1477

Dewey, F. E., Perez, M. V., Wheeler, M. T., Watt, C., Spin, J., Langfelder, P., Horvath, S., et al. (2011). Gene coexpression network topology of cardiac development, hypertrophy, and failure. *Circulation. Cardiovascular genetics*, *4*(1), 26–35. doi:10.1161/CIRCGENETICS.110.941757

Dixon, R. E. (1985). Economic costs of respiratory tract infections in the United States. *The American journal of medicine*, *78*(6), 45–51.

Domon, B., & Aebersold, R. (2006a). Mass spectrometry and protein analysis. *Science Signalling*, *312*(5771), 212. doi:10.1126/science.312.5771.211

Domon, B., & Aebersold, R. (2006b). Challenges and Opportunities in Proteomics Data Analysis. *Molecular & Cellular Proteomics*, *5*(10), 1921–1926. doi:10.1074/mcp.R600012-MCP200

Donaldson, R., & Calder, M. (2010). Modelling and Analysis of Biochemical Signalling Pathway Cross-talk. (E. Merelli & P. Quaglia, Eds.)*Electronic Proceedings in Theoretical Computer Science*, *19*, 40–54. doi:10.4204/EPTCS.19.3

Dykxhoorn, D. M., & Lieberman, J. (2006). Silencing viral infection. *PLoS Medicine*, *3*(7), e242. doi:10.1371/journal.pmed.0030242

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, *5*(11), 976–989. doi:10.1016/1044-0305(94)80016-2

Fagan, A., Culhane, A. C., & Higgins, D. G. (2007). A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, *7*(13), 2162–2171. doi:10.1002/pmic.200600898

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, *23*(2), 257–258. doi:10.1093/bioinformatics/btl567

Feng, J., Naiman, D. Q., & Cooper, B. (2007). Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics*, *23*(17), 2210–2217. doi:10.1093/bioinformatics/btm267

Figeys, D. (2008). Mapping the human protein interactome. *Cell Research*, *18*(7), 716–724. doi:10.1038/cr.2008.72

Forst, C. V. (2006). Host-pathogen systems biology. *Drug Discovery Today*, *11*(5/6), 220–227. doi:1359-6446/06/$

Fraser, A. G., & Marcotte, E. M. (2004). A probabilistic view of gene function. *Nature Genetics*, *36*(6), 559–564. doi:10.1038/ng1370

Gat-Viks, I., Tanay, A., Raijman, D., & Shamir, R. (2006). A probabilistic methodology for integrating knowledge and experiments on biological networks. *Journal of computational biology : a journal of computational molecular cell biology*, *13*(2), 165–181. doi:10.1089/cmb.2006.13.165

Gupta, N., & Pevzner, P. A. (2009). False discovery rates of protein identifications: a strike against the two-peptide rule. Journal of Proteome Research, 8(9), 4173.

Hastie, T., Tibshirani, R., Narasimhan, B., & Chu, G. (n.d.). Impute: Imputation for microarray data. Bioconductor.

Hatta, Y., Hershberger, K., Shinya, K., Proll, S. C., Dubielzig, R. R., Hatta, M., Katze, M. G., et al. (2010). Viral replication rate regulates clinical outcome and CD8 T cell responses during highly pathogenic H5N1 influenza virus infection in mice. *PLoS Pathogens*, *6*(10). doi:10.1371/journal.ppat.1001139

Huang, T., Wang, J., Yu, W., & He, Z. (2012). Protein inference: a review. *Briefings in bioinformatics*. doi:10.1093/bib/bbs004

Iancu, O. D., Darakjian, P., Malmanger, B., Walter, N. A. R., McWeeney, S., & Hitzemann, R. (2012a). Gene networks and haloperidol-induced catalepsy. *Genes, brain, and behavior*, *11*(1), 29–37. doi:10.1111/j.1601-183X.2011.00736.x

Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., & McWeeney, S. (2012b). Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics*, *28*(12), 1592–1597. doi:10.1093/bioinformatics/bts245

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome research*, *18*(4), 644–652. doi:10.1101/gr.071852.107

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, *2*, 343–372. doi:10.1146/annurev.genom.2.1.343

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299–314.

Ihara, M., & Kano, Y. (1986). A new estimator of the uniqueness in factor analysis. *Psychometrika*, *51*, 563–566. doi:10.1007/BF02295595

Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating “omics” data sets. *Nat Rev Mol Cell Biol*, *7*(3), 198–210. doi:10.1038/nrm1857

Kanehisa, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, *32*(90001), 277D–280. doi:10.1093/nar/gkh063

Katze, M., & He, Y. (2002). Viruses and interferon: a fight for supremacy. *Nature Reviews Immunology*.

Kawashima, S., Katayama, T., & Sato, Y. (2003). KEGG API: A web service using SOAP/WSDL to access the KEGG system. *GENOME ….*

Kellam, P. (2001). Post-genomic virology: the impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. *Reviews in medical virology*, *11*(5), 313–329.

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, *74*(20), 5383–5392.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., et al. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Research*, *37*(Database), D767–D772. doi:10.1093/nar/gkn892

Kim, H. L. (2004). Using Protein Expressions to Predict Survival in Clear Cell Renal Carcinoma. *Clinical Cancer Research*, *10*(16), 5464–5471. doi:10.1158/1078-0432.CCR-04-0488

Koh, G., Teong, H. F. C., Clement, M. V., Hsu, D., & Thiagarajan, P. S. (2006). A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk. *Bioinformatics*, *22*(14), e271–e280. doi:10.1093/bioinformatics/btl264

Konieczka, J. H., Drew, K., Pine, A., Belasco, K., Davey, S., Yatskievych, T. A., Bonneau, R., et al. (2009). BioNetBuilder2.0: bringing systems biology to chicken and other model organisms. *BMC Genomics*, *10 Suppl 2*, S6. doi:10.1186/1471-2164-10-S2-S6

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., et al. (2003). A novel coronavirus associated with severe acute respiratory syndrome. *The New England journal of medicine*, *348*(20), 1953–1966. doi:10.1056/NEJMoa030781

Langfelder, P. (2010). Tutorial for the WGCNA package for R.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. doi:10.1186/1471-2105-9-559

Langfelder, P., & Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*, *46*(i11).

Langfelder, P., Castellani, L. W., Zhou, Z., Paul, E., Davis, R., Schadt, E. E., Lusis, A. J., et al. (2012). A systems genetic analysis of high density lipoprotein metabolism and network preservation across mouse models. *Biochimica et biophysica acta*, *1821*(3), 435–447. doi:10.1016/j.bbalip.2011.07.014

Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Computational Biology*, *7*(1), e1001057. doi:10.1371/journal.pcbi.1001057

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720. doi:10.1093/bioinformatics/btm563

Li, Ai, & Horvath, S. (2009). Network module detection: Affinity search technique with the multi-node topological overlap measure. *BMC research notes*, *2*, 142. doi:10.1186/1756-0500-2-142

Li, Fang, Li, W., Farzan, M., & Harrison, S. C. (2005a). Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science (New York, NY)*, *309*(5742), 1864–1868. doi:10.1126/science.1116480

Li, Q., MacCoss, M. J., & Stephens, M. (2010a). A nested mixture model for protein identification using mass spectrometry. *Annals of Applied Statistics*, *4*(2), 962–987.

Li, Wendong, Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., Wang, H., et al. (2005b). Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science (New York, NY)*, *310*(5748), 676–679. doi:10.1126/science.1118391

Li, Yong Fuga, Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., & Tang, H. (2009). A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. *Journal of Computational Biology*, *16*(8), 1183–1193. doi:10.1089/cmb.2009.0018

Li, Yong Fuga, Arnold, R. J., Tang, H., & Radivojac, P. (2010b). The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *Journal of proteome research*, *9*(12), 6288–6297. doi:10.1021/pr1005586

Low, D. E., & McGeer, A. (2003). SARS--one year later. *The New England journal of medicine*, *349*(25), 2381–2382. doi:10.1056/NEJMp038203

Ma, Z.-Q., Dasari, S., Chambers, M. C., Litton, M. D., Sobecki, S. M., Zimmerman, L. J., Halvey, P. J., et al. (2009). IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *Journal of proteome research*, *8*(8), 3872–3881. doi:10.1021/pr900360j

MacLennan, N. K., Dong, J., Aten, J. E., Horvath, S., Rahib, L., Ornelas, L., Dipple, K. M., et al. (2009). Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice. *Molecular genetics and metabolism*, *98*(1-2), 203–214. doi:10.1016/j.ymgme.2009.05.004

Magrane, M., & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, *2011*(0), bar009–bar009. doi:10.1093/database/bar009

Malmström, J., & Lee, H. (2007). Advances in proteomic workflows for systems biology. *Current opinion in biotechnology*.

Mason, M. J., Fan, G., Plath, K., Zhou, Q., & Horvath, S. (2009). Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, *10*, 327. doi:10.1186/1471-2164-10-327

Masters, P. S. (2006). The molecular biology of coronaviruses. *Advances in virus research*, *66*, 193–292. doi:10.1016/S0065-3527(06)66005-3

Mauad, T., Hajjar, L. A., Callegari, G. D., da Silva, L. F. F., Schout, D., Galas, F. R. B. G., Alves, V. A. F., et al. (2009). Lung Pathology in Fatal Novel Human Influenza A (H1N1) Infection. *American Journal of Respiratory and Critical Care Medicine*, *181*(1), 72–79. doi:10.1164/rccm.200909-1420OC

McAteer, J., & Skerrett, S. (2009). A Bayesian Integration Model Of High-Throughput Proteomics And Metabolomics Data For Improved Early Detection Of Microbial. *Pacific Symposium on Biocomputing*, *463*, 451–463.

Mcgarvey, P. B., Huang, H., Mazumder, R., Zhang, J. A., Chen, Y., Zhang, C., Cammer, S., et al. (2009). Systems Integration of Biodefense Omics Data for Analysis of Pathogen-Host Interactions and Identification of Potential Targets. *PLoS One*, *4*(9), e7162. doi:10.1371/journal.pone.0007162.t001

Mihályi, E., & Szent-Györgyi, A. G. (1953). TRYPSIN DIGESTION OF MUSCLE PROTEINS. *Journal of Biological Chemistry*.

Monroe, M. E., Tolić, N., Jaitly, N., Shaw, J. L., Adkins, J. N., & Smith, R. D. (2007). VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics*, *23*(15), 2021–2023. doi:10.1093/bioinformatics/btm281

Morley, J. E., Baumgartner, R. N., Roubenoff, R., Mayer, J., & Nair, K. S. (2001). Sarcopenia. *Journal of Laboratory and Clinical Medicine*, *137*(4), 231–243. doi:10.1067/mlc.2001.113504

Needham, C. J., Bradford, J. R., Bulpitt, A. J., & Westhead, D. R. (2007). A primer on learning in Bayesian networks for computational biology. (F. Lewitter, Ed.)*PLoS Computational Biology*, *3*(8), e129. doi:10.1371/journal.pcbi.0030129

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, *73*(11), 2092–2123. doi:10.1016/j.jprot.2010.08.009

Nesvizhskii, A. I., & Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP*, *4*(10), 1419–1440. doi:10.1074/mcp.R500012-MCP200

Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, *75*(17), 4646–4658.

Nie, L., Wu, G., Culley, D., Scholten, J., & Zhang, W. (2007). Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit Rev Biotechnol*, *27*(2), 63–75. doi:10.1080/07388550701334212

Noble, W. S., & MacCoss, M. J. (2012). Computational and Statistical Analysis of Protein Mass Spectrometry Data. *PLoS Computational Biology*, *8*(1), e1002296.

Northrop, J. H. (1922). The mechanism of the influence of acids and alkalies on the digestion of proteins by pepsin or trypsin. *The Journal of General Physiology*, *5*(2), 263–274. doi:10.1085/jgp.5.2.263

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, *103*(47), 17973–17978. doi:10.1073/pnas.0605938103

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nature Neuroscience*, *11*(11), 1271–1282. doi:10.1038/nn.2207

Orwoll, E., Blank, J. B., Barrett-Connor, E., Cauley, J., Cummings, S., Ensrud, K., Lewis, C., et al. (2005). Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study--a large observational study of the determinants of fracture in older men. *Contemporary clinical trials*, *26*(5), 569–585. doi:10.1016/j.cct.2005.05.006

Palsson, B., & Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nature Chemical Biology*, *6*(11), 787–789. doi:10.1038/nchembio.462

Peiris J, C. C. Y. L. C. Y. N. J. (n.d.). Innate immune responses to influenza A H5N1: friend or foe? *Trends in Immunology*, *30*(12), 574–584. doi:doi:10.1016/j.it.2009.09.004

Peng, X., Chan, E. Y., Li, Y., Diamond, D. L., Korth, M. J., & Katze, M. G. (2009). Virus-host interactions: from systems biology to translational research. *Current opinion in microbiology*, *12*(4), 432–438. doi:10.1016/j.mib.2009.06.003

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18), 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Price, T. S., Lucitt, M. B., Wu, W., Austin, D. J., Pizarro, A., Yocum, A. K., Blair, I. A., et al. (2007). EBP, a program for protein identification using multiple tandem mass spectrometry datasets. *Molecular & cellular proteomics : MCP*, *6*(3), 527–536. doi:10.1074/mcp.T600049-MCP200

Rappsilber, J., & Mann, M. (2002). What does it mean to identify a protein in proteomics? *Trends in Biochemical Sciences*, *27*(2), 74–78. doi:10.1016/S0968-0004(01)02021-7

Ravasz, E., & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *67*(2 Pt 2), 026112.

Roberts, A., Deming, D., Paddock, C. D., Cheng, A., Yount, B., Vogel, L., Herman, B. D., et al. (2007). A Mouse-Adapted SARS-Coronavirus Causes Disease and Mortality in BALB/c Mice. *PLoS Pathogens*, *3*(1), e5. doi:10.1371/journal.ppat.0030005

Rockx, B., Feldmann, F., Brining, D., Gardner, D., LaCasse, R., Kercher, L., Long, D., et al. (2011). Comparative pathogenesis of three human and zoonotic SARS-CoV strains in cynomolgus macaques. *PLoS One*, *6*(4), e18558. doi:10.1371/journal.pone.0018558

Safronetz, D., Rockx, B., Feldmann, F., Belisle, S. E., Palermo, R. E., Brining, D., Gardner, D., Proll, S. C., Marzi, A., Lacasse, R. A., et al. (2011a). Pandemic swine-origin H1N1 influenza A virus isolates show heterogeneous virulence in macaques. *Journal of virology*, *85*(3), 1214–1223. doi:10.1128/JVI.01848-10

Safronetz, D., Rockx, B., Feldmann, F., Belisle, S. E., Palermo, R. E., Brining, D., Gardner, D., Proll, S. C., Marzi, A., Tsuda, Y., et al. (2011b). Pandemic Swine-Origin H1N1 Influenza A Virus Isolates Show Heterogeneous Virulence in Macaques. *Journal of virology*, *85*(3), 1214–1223. doi:10.1128/JVI.01848-10

Salim, A., & Pawitan, Y. (2007). Model-based maximum covariance analysis for irregularly observed climatological data. *Journal of Agricultural, Biological, and Environmental Statistics*, *12*, 1–24. doi:10.1198/108571107X177078

Satija, N., & Lal, S. K. (2007). The Molecular Biology of SARS Coronavirus. *Annals of the New York Academy of Sciences*, *1102*(1), 26–38. doi:10.1196/annals.1408.002

Schulz-Trieglaff, O., Pfeifer, N., Gröpl, C., Kohlbacher, O., & Reinert, K. (2008). LC-MSsim--a simulation software for liquid chromatography mass spectrometry data. *BMC Bioinformatics*, *9*, 423. doi:10.1186/1471-2105-9-423

Searle, B. C. (2010). Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, *10*(6), 1265–1269. doi:10.1002/pmic.200900437

Segal, E., Friedman, N., Kaminski, N., Regev, A., & Koller, D. (2005). From signatures to models: understanding cancer using microarrays. *Nature Genetics*, *37*(6s), S38–S45. doi:10.1038/ng1561

Serang, O., & Noble, W. (2012a). A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface*, *5*(1), 3–20.

Serang, O., & Noble, W. S. (2012b). Faster mass spectrometry-based protein inference: junction trees are more efficient than sampling and marginalization by enumeration. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, *9*(3), 809–817. doi:10.1109/TCBB.2012.26

Serang, O., MacCoss, M. J., & Noble, W. S. (2010). Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research*, *9*(10), 5346–5357. doi:10.1021/pr100594k

Shai, R., Shi, T., Kremen, T. J., Horvath, S., Liau, L. M., Cloughesy, T. F., Mischel, P. S., et al. (2003). Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, *22*(31), 4918–4923. doi:10.1038/sj.onc.1206753

Shen, C., Wang, Z., Shankar, G., Zhang, X., & Li, L. (2008). A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. *Bioinformatics*, *24*(2), 202–208. doi:10.1093/bioinformatics/btm555

Shortridge, K. F., Zhou, N. N., Guan, Y., Gao, P., Ito, T., Kawaoka, Y., Kodihalli, S., et al. (1998). Characterization of avian H5N1 influenza viruses from poultry in Hong Kong. *Virology*, *252*(2), 331–342. doi:10.1006/viro.1998.9488

Slotta, D. J., McFarland, M. A., & Markey, S. P. (2010). MassSieve: Panning MS/MS peptide data for proteins. *Proteomics*, *10*(16), 3035–3039. doi:10.1002/pmic.200900370

Smith, R. D., Anderson, G. A., Lipton, M. S., Masselon, C., Pasa-Tolic, L., Shen, Y., & Udseth, H. R. (2002a). The use of accurate mass tags for high throughput microbial proteomics. *OMICS*, *6*, 61–90. doi:10.1089/15362310252780843

Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., et al. (2002b). An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, *2*(5), 513–523. doi:10.1002/1615-9861(200205)2:5<513::AID-PROT513>3.0.CO;2-W

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, *27*(3), 431–432. doi:10.1093/bioinformatics/btq675

Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.-D., & Rappuoli, R. (2003). SARS — beginning to understand a new virus. *Nature Reviews Microbiology*, *1*(3), 209–218. doi:10.1038/nrmicro775

Stanley, J. R., Adkins, J. N., Slysz, G. W., Monroe, M. E., Purvine, S. O., Karpievitch, Y. V., Anderson, G. A., et al. (2011). A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics. *Analytical chemistry*, *83*(16), 6135–6140. doi:10.1021/ac2009806

Storch, von, H., & Zwiers, F. W. (2002). Statistical analysis in climate research.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, NY)*, *302*(5643), 249–255. doi:10.1126/science.1087447

Sung, J., Wang, Y., Chandrasekaran, S., Witten, D. M., & Price, N. D. (2012). Molecular signatures from omics data: From chaos to consensus. Biotechnology Journal, 7(8), 946–957. doi:10.1002/biot.201100305

Takaoka, A., & Yanai, H. (2006). Interferon signalling network in innate defence. *Cellular Microbiology*, *8*(6), 907–922. doi:10.1111/j.1462-5822.2006.00716.x

Tan, C. S., Salim, A., Ploner, A., Lehtiö, J., Chia, K. S., & Pawitan, Y. (2009). Correlating gene and protein expression data using Correlated Factor Analysis. *BMC Bioinformatics*, *10*(1), 272. doi:10.1186/1471-2105-10-272

Tan, S.-L., Ganji, G., Paeper, B., Proll, S., & Katze, M. G. (2007). Systems biology and the host response to viral infection. *Nature biotechnology*, *25*(12), 1383–1389. doi:10.1038/nbt1207-1383

Tang, H., Arnold, R. J., Alves, P., Xun, Z., Clemmer, D. E., Novotny, M. V., Reilly, J. P., et al. (2006). A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, *22*(14), e481–e488. doi:10.1093/bioinformatics/btl237

Tirosh, I., & Barkai, N. (2005). Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinformatics*, *6*, 40. doi:10.1186/1471-2105-6-40

Tisoncik, J. R., & Katze, M. G. (2010). What is systems biology? *Future Microbiology*, *5*(2), 139. doi:10.1128/JVI.05605-11

Torres-Garc i a, W., Zhang, W., Runger, G. C., Johnson, R. H., & Meldrum, D. R. (2009). Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics*, *25*(15), 1905–1914. doi:10.1093/bioinformatics/btp325

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., & Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8348–8353. doi:10.1073/pnas.0832373100

Tucker, L. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*, 111–136. doi:10.1007/BF02289009

Vaske, C. J., House, C., Luu, T., Frank, B., Yeang, C.-H., Lee, N. H., & Stuart, J. M. (2009). A Factor Graph Nested Effects Model To Identify Networks from Genetic Perturbations. *PLoS Computational Biology*, *5*(1), e1000274. doi:10.1371/journal.pcbi.1000274.t001

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, *13*(4), 227–232. doi:10.1038/nrg3185

Waltermann, C., & Klipp, E. (2011). Information theory based approaches to cellular signaling. *Biochimica et biophysica acta*, *1810*(10), 924–932. doi:10.1016/j.bbagen.2011.07.009

Waters, K. M., Pounds, J. G., & Thrall, B. D. (2006). Data merging for integrated microarray and proteomic analysis. *Briefings in functional genomics & proteomics*, *5*(4), 261–272. doi:10.1093/bfgp/ell019

Webb-Robertson, B. J. M., Cannon, W. R., Oehmen, C. S., Shah, A. R., Gurumoorthi, V., Lipton, M. S., & Waters, K. M. (2008). A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, *24*(13), 1503–1509. doi:10.1093/bioinformatics/btn218

Weston, A. D., & Hood, L. (2004). Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. *Journal of proteome research*, *3*(2), 179–196. doi:10.1021/pr0499693

Wolber, P. K., Collins, P. J., Lucas, A. B., De Witte, A., & Shannon, K. W. (2006). The Agilent in situ-synthesized microarray platform. *Methods in enzymology*, *410*, 28–57. doi:10.1016/S0076-6879(06)10002-6

Wu, C. H. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, *34*(90001), D187–D191. doi:10.1093/nar/gkj161

Yellaboina, S., Dudekula, D. B., & Ko, M. S. (2008). Prediction of evolutionarily conserved interologs in Mus musculus. *BMC Genomics*, *9*, 465. doi:10.1186/1471-2164-9-465

Yip, A. M., & Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, *8*, 22. doi:10.1186/1471-2105-8-22

Yoshida, T. (2004). Peptide separation by Hydrophilic-Interaction Chromatography: a review. *Journal of Biochemical and Biophysical Methods*, *60*(3), 265–280. doi:10.1016/j.jbbm.2004.01.006

Zak, D. E., & Aderem, A. (2009). Systems biology of innate immunity. *Immunological Reviews*, *227*(1), 264–282. doi:10.1111/j.1600-065X.2008.00721.x

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*, Article17. doi:10.2202/1544-6115.1128

Zhang, Bing, Chambers, M. C., & Tabb, D. L. (2007). Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of proteome research*, *6*(9), 3549–3557. doi:10.1021/pr070230d

Zhang, Bing, VerBerkmoes, N. C., Langston, M. A., Uberbacher, E., Hettich, R. L., & Samatova, N. F. (2006). Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of proteome research*, *5*(11), 2909–2918. doi:10.1021/pr0600273

Zhang, J, & Gentleman, R. (n.d.). KEGGSOAP. R Package.

Zimmer, J. S. D., Monroe, M. E., Qian, W.-J., & Smith, R. D. (2006). Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrometry Reviews*, *25*(3), 450–482. doi:10.1002/mas.20071