UNIVERSITY OF CALIFORNA, SAN DIEGO

Advances in Computational Mass Spectrometry:

Phosphoproteomics and Proteogenomics

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Bioinformatics

by

Samuel Harris Payne

Committee in charge:

Professor Vineet Bafna, Chair
Professor William F. Loomis, Co-Chair
Professor Steven P. Briggs
Professor Elizabeth A. Komives
Professor Terrence Hwa

2008

The Dissertation of Samuel Harris Payne is approved, and it is acceptable in quality and for publication on microfilm:

____

____

____

____

Co-Chair

____

Chair

University of California, San Diego

2008

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to sincerely thank my advisor Vineet Bafna, for his steady and encouraging guidance. He was instrumental in helping me find my niche in graduate school, and finding research that I could flourish with. Vineet was always available for me and is a generous man. Vineet always focused on getting the correct result, regardless of its affect on the research.

I would also like to thank the members of my committee. I have been blessed with significant scientific interaction with every committee member; each was open, generous, and kind. With each, I felt elevated beyond a graduate student to peer. I thank Bill Loomis for his thoughtful and careful mentoring of a young computer science student. Bill spent many hours helping me understand and appreciate the beauty of molecular biology. His tutoring was invaluable in helping me to be a competent bioinformaticist. He always stressed to me that computation alone would not solve any significant or interesting problems. I thank Steve Briggs, for his guidance and encouragement. I benefited greatly from regularly attending his lab meetings and watching his discerning eye. I thank Betsy Komives for her friendly and informative tutorage in the experimental aspects of mass spectrometry. Finally, I thank Terry Hwa for his respect and solicitation of my ideas.

My work would not have been possible without funding from the National Institutes of Health and the National Science Foundation, both of which supplied training grants that allowed me to research here at UCSD. The

NIH bioinformatics training grant gave me the opportunity to rotate through many labs and find the research that interested me. The NSF plant systems biology IGERT helped me find exciting and productive collaborations and allowed me to present my research at conferences. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622.

I would like to thank the many friends and family who supported and influenced my education. First, my wife Mollie has been the standard of unwavering support. She always helped me to focus on what is most important for both my degree and my life. Without her love and support, none of this would have been possible. My two boys, Harrison and Noah, whose energy fills me. I thank my sister, Sara Anderson, for introducing me to bioinformatics. I would like to thank my parents for their support and encouragement. They have also been an example for me in balancing life and work. Finally, I would like to thank my grandfather, John W Payne, who made the educational leap from shepherd to college professor, for his diligence and determination in pursuing education.

I would also like to thank past scientists who had a significant impact on my career. Bryan Morse, at Brigham Young University, and Bernard Heymann, at the National Institutes of Health, both provided research opportunities for me as an undergraduate in biomedical imaging. I would also

like to thank Steve Wasserman, at UCSD, for letting me rotate in his lab and learn basic experimental techniques.

I would like to thank past and present members of the Bafna, Briggs and Zhou lab for helping me with my research. Stephen Tanner for mentoring in computational mass spectrometry, and for writing a wonderful software tool, Inspect, which I was able to extend and use for all my research needs. I would like to thank Natalie Castellana for her diligent work with me in the Arabidopsis proteogenomic annotation. Zhouxin Shen, Marcus Smolka, and Claudio Albuquerque all generated the data that I used for my research. I am indebted to them for their gifted knowledge of analytical chemistry and mass spectrometry.

Chapter 2, in full, was published as "Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis". Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, and Banfa V. 2008. Journal of Proteome Research, in press. The dissertation author was the primary author of this paper.

Chapter 3 is in preparation for publication as "Proteogenomic Discovery, Correction and Confirmation of Arabidopsis Gene Models". Castellana NC, Payne SH, Shen Z, Stanke M, Bafna V, and Briggs SP. 2008, submitted. The dissertation author, Natalie Castellana and Zhouxin Shen were the primary authors of this paper.

Chapter 5, in full, was published as “Retention and Loss of Amino Acid Biosynthetic Pathways Based on Whole Genome Sequences.” Payne SH and Loomis WF (2006) Eukaryot Cell 5(2):272-6. The dissertation author was the primary author of this paper.

Vita

2002	Bachelor of Science, Brigham Young University
2008	Doctor of Philosophy, University of California, San Diego

Publications

Castellana NE, Payne SH, Shen Z, Stanke M, Bafan V, Briggs SP. Proteogenomic discovery, correction and confirmation of Arabidopsis gene models. In review.

Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, Bafna V. Phosphorylation specific MS/MS scoring for rapid and accurate phospho-proteome analysis. Poster presented at 2008 ASMS in Denver.

Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, Bafna V. Phosphorylation specific MS/MS scoring for rapid and accurate phospho-proteome analysis. Poster presented at Ninth International Symposium on Mass Spectrometry in the Health and Life Sciences, 2007 in San Francisco.

Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, Bafna V. Phosphorylation specific MS/MS scoring for rapid and accurate phospho-proteome analysis. In press, J Proteome Res

Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H. High-coverage phosphoproteome analysis using HILIC based mult-dimensional chromatography. in press Mol Cell Proteomics

Tanner S, Payne SH, Dasari S, Shen Z, Wilmarth P, David L, Loomis WF, Briggs SP, Bafna V. Accurate Annotation of Peptide Modifications through Unrestrictive Database Search. Journal of Proteome Research 2008, 7:170-181 PubMed

Rana BK, Insel PA, Payne SH, Abel K, Beutler E, Ziegler MG, Schork NJ, O'Connor DT. Population-based sample reveals gene-gender interactions in blood pressure in White Americans. Hypertension. 2007 Jan;49(1):96-106. PubMed

Payne SH and Loomis WF (2006) Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. Eukaryot Cell 5(2):272-6. PubMed

Hirose S, Payne SH, Loomis WF (2006) cis-Acting site controlling bidirectional transcription at the growth-differentiation transition in Dictyostelium discoideum. Eukaryot Cell. 2006 Jul;5(7):1104-10. PubMed

Payne SH (2005) Metabolic pathways, p. 41-57. In W. F. Loomis and A. Kuspa (ed.), Dictyostelium genomics. Horizon Press, Far Hills, N.J.

Fields of Study

Major Field: Bioinformatics and Proteomics

Studies in Bioinformatics

Professor Vineet Bafna

Studies in Proteomics

Professors Vineet Bafna, Steven Briggs and Elizabeth Komives

Studies in Plant Biology

Professor Steven Briggs

Studies in Molecular Biology

Professor William Loomis

ABSTRACT OF THE DISSERTATION

Advances in Computational Mass Spectrometry:

Phosphoproteomics and Proteogenomics

by

Samuel Harris Payne

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2008

Professor Vineet Bafna, Chair

Professor William Loomis, Co-Chair

The proteome is a dynamic group of proteins, interacting with and modifying each other in response to the environment. Tandem mass spectrometry has become the most convenient and high-throughput means of assaying the proteome. Modern instruments are capable of generating data for tens of thousands of peptides from thousands of proteins in a single experiment. In this work we present two important applications on proteomics: phosphoproteomics and proteogenomics.

Protein signaling is dominated by reversible phosphorylation. Understanding which proteins are phosphorylated, when, where, and by whom is key to understanding most cellular signaling. A variety of obstacles make

assaying phosphopeptides with tandem mass spectrometry a difficult task. First, phosphorylation is reversible and transitory. Therefore, although many proteins can be phosphorylated, very few are phosphorylated at any given time. Moreover, the phosphorylation event may be sub-stoichiometric. Thus a small fraction of peptides in a proteomic sample are phosphorylated. Experimental mass spectrometrists have overcome this with the adoption of phosphopeptide enrichment protocols. A sample containing perhaps 1% phosphopeptides can be purified to over 90% phosphopeptides. However, even with a high concentration of phosphorylated peptides, phosphoproteomics suffers from a second challenge, poor spectral quality. Spectra generated by phosphopeptides have low information content and are difficult to interpret. We present an approach for learning the features of phosphopeptide spectra, and model these features in a Bayesian network. This probability model, when applied to the scoring function of Inspect, achieves a dramatic increase in sensitivity versus other peptide identification software.

The second field of study presented in proteogenomics. The task of annotating the genome for protein coding genes is difficult, and requires substantial effort. Yet this is the arguably the most important outcome of the genomic era. Most annotation pipelines utilize nucleotide centric information, such as cDNA or homology to known genes, to refine their computational predictions. Unfortunately error rates are still suspected to be high, both in

terms of genes which are mispredicted and genes which are wholly missing from the annotation. We present our work on utilizing peptides obtained from mass spectrometry to reannotate the genome. We collect a large corpus of MS/MS spectra from Arabidopsis thaliana and annotate spectra from 18,024 peptides which are not currently in the proteome. Using these peptides we present gene models for 778 genes missing from the current annotation, and refine or correct an additional 695 loci, showing that proteogenomics can dramatically improve the quality of a genome annotation.

Chapter 1: Introduction: Proteomics and Mass Spectrometry

As genomics continues to receive significant media attention and funding as the new vehicle of biological research, Erin O'Shea and Jonathan Weissman remind us that "biological systems ultimately need to be explained in terms of the activity, regulation and modification of proteins" [Ghaemmaghami 2003]. Proteomics then is the large-scale study of proteins, their structure, function and interactions.

In one of the first proteome-scale experiments, Ghaemmaghami and colleagues quantified protein abundance of 4251 yeast proteins, or 80% of the proteome, during log-phase growth [Ghaemmaghami 2003]. In this landmark effort, all yeast open reading frames (ORFs) were tagged with a dual specificity epitope to facilitate efficient purification from cell lysates. To classify protein localization, the same group created a separate library of GFP tagged ORFs [Huh 2003]. Twenty-two distinct sub-cellular localizations were used to classify 75% of the proteome, including 70% of previously unlocalized proteins.

Many proteins perform their function in a complex, or in conjunction, with other proteins. Therefore, protein interactions are another important characteristic of the proteome. Direct physical interactions can be assayed with co-immunoprecipitation, or with two-hybrid screening [Krogan 2006, Fields 1989]. From this set of binary interactions, large interaction networks can be constructed and compared [Suthram 2005]. Another kind of interaction is functional interaction, such as proteins working together in a metabolic pathway. Although they may not physically touch, they form a logical interaction. These types of interaction for metabolic and signaling pathways are curated at KEGG [Kanehisa 2000].

The most significant drawback to these projects is the substantial time investment required to create materials. Clone libraries created for the Weissman and O'Shea papers consisted of thousands of genetically modified yeast strains. These libraries are a valuable resource to yeast researchers, and the results from these papers are of tremendous impact on the scientific community in general. However, similar libraries have yet to be created for other organisms, a testament to the technical difficulty involved in whole proteome library creation. Similarly, the effort required to make yeast-two-hybrid libraries is significant. For this reason, a different high-throughput technology must be utilized to carry forth proteomic research.

Mass spectrometry-based proteomics is able to offer insights into protein content, quantitation, modification, and interaction, without requiring extensive clone library creation [Aebersold 2003]. Mass spectrometers take as input an ionized biological sample. The output is a spectrum, or mass to charge ratios (m/z) of the constituent species of a sample. Protein mass spectrometry primarily uses short polypeptides, proteolyzed proteins, as the input sample as these are more amenable to the mass range of the instrument. Additionally, protein mass spectrometry utilizes tandem mass spectrometry, where a single m/z region is isolated and fragmented to produce more information about the species. From this output, protein sequence can be inferred. Commonly, liquid chromatography is coupled to the front end of the mass spectrometry instrument, separating a complex peptide sample and concentrating its constituent peptides into a small elution window. In this manner, information on thousands of proteins can be gained from a single experiment.

Klaus Biemann discussed the use of electron collision to produce a series of peptide fragments which could be used to infer protein sequence over 40 years ago [Biemann 1966]. In this work, he presents a *de novo* algorithm for peptide annotation and discusses the difficulties arising from non-protein peptides, unexpected amino acids, or unexpected linkages, and neutral losses which are not part of the primary fragment ladder.

In 1981 Fred McLafferty discussed the novel technique of tandem mass spectrometry, which uses two stages of MS. In the first stage, all ions present in the sample are measured. Then a single species is isolated, and all others are ejected. In the second stage of MS, the isolated species is collided with gas atoms to produce peptide fragment ions [McLafferty 1981]. He also discusses advances in chromatography, and the superiority of using two orthogonal separation criteria. It is much easier to isolate a single species by distinguishing on two characteristics than on only one. Both of these ideas (two stage mass spectrometry and two dimensional chromatography) have become standards in the proteomics field.

1.1 Computational Mass Spectrometry

Initially, most spectra were annotated by hand. Often many of them were confirmed by chemically synthesizing the peptide and confirming the analysis. Two issues initiated the movement towards computational proteomics: speed and speed. With the invention and wide-spread adoption of ion-trap mass spectrometers, tens and hundreds of thousands of spectra began to be generated in a single experiment. This massive increase in data volume overwhelmed the ability of manual annotation. Secondly, as more protein sequence was being discovered and cataloged, the idea of using a database to speed up annotation software became realistic. Although still incomplete today, protein repositories like the NCBI and swiss-prot/trembl had a reasonably large subset of the human proteome, and using this to limit the search space of MS/MS interpretation algorithms could significantly increase the speed of annotation.

Computational mass spectrometry came into its own in the 1990s, starting with the publication of SEQUEST, the first sequence aided algorithm (database search) to automatically identify peptide sequences from tandem mass spectra [Eng 1994]. In this paper, formalisms currently used by most database search engines were set forth: data preprocessing, database filtering, and scoring. An MS/MS spectrum corresponds to (hopefully) only one peptide species. Thus only a small portion of the database is relevant. Database filtering is the idea of quickly eliminating much of the obviously unrelated sequences, so that scoring can focus its efforts on the most likely candidates. SEQUEST starts filtering with an *in silico* proteolytic digestion of the protein database, to produce peptides from proteins. For each spectrum, peptides with parent masses inconsistent with the spectrum were filtered out and removed from consideration. This parent mass filtering paradigm has remained the dominant filtering technique for most spectral identification programs. Candidate peptides which pass the filter are then considered for scoring. The task of scoring is to give some numeric measure of how well a peptide matches the spectrum. The best candidate, the one with the highest score, is considered the winner. SEQUEST's scoring attempts to count the number of *b/y* fragment ions predicted from the candidate peptide which are observed in the spectrum. The assumption here is that the correct peptide will have the most peaks in common with the observed spectrum. Formally, each peptide creates a theoretical spectrum consisting of *b/y* ions, ^{13}C isotopes and neutral loss of water and ammonia and *a* ions. A simple dot product between the peak lists counts the matching peaks. Spectra with more matching peaks score better. Then, the top 500 scoring peptides are re-scored with a more complex scoring function, the X-corr. Finally, the authors note that the difference between the best scoring peptide and the runner up, denoted the “delta score”, is a valuable metric as false annotations often have multiple peptides which score equally well. This realization proved to be the basis for subsequent post-processing validation algorithms.

Later algorithms, including Mascot and X!Tandem were also released. Mascot uses a probabilistic scoring function, although the details are not publically available [Perkins 1999]. X!Tandem’s major innovation was twofold: first, it was released free for public use, secondly, it had a more advanced database filter than either Mascot or SEQUEST which made is faster, although less sensitive [Craig 2004].

In the Bafna lab, Stephen Tanner created Inspect, a database search engine with several improvements over the then available tools [Tanner 2005]. First, Inspect was created with a highly efficient database filter, which makes it orders of magnitude faster than SEQUEST or X!Tandem. The filtering strategy comes from sequence tags within the spectrum. Consecutive peaks of the b/y ladder can be interpreted to annotate consecutive amino acids in the peptide. The concept of sequence tags was presented in 1994, and is commonly used to judge the quality of a spectrum annotation during manual validation [Mann 1994]. By filtering peptides except those with a tag match, only ~5 peptides per megabase of database are scored, as opposed to tens of thousands with a parent mass filter. Sensitivity losses for this are typically minimal, and acceptable given the dramatic speed improvement. The second major advance of Inspect was the Bayesian network based scoring function.

In brief, this uses information from other ions in the spectrum to evaluate the probability that a given ion assignment is correct. This idea is discussed in full in Chapter 2, where the dissertation author presents his work adapting this framework for phosphorylated peptides.

Methods for results validation evolved separately from annotation algorithms. Initially researches used an *ad hoc* validation process and rigorous probability values were not often enforced. Commonly researches trusted all results with a given final score and perhaps manually validated a subset of the results. This was facilitated by the publication of score cutoffs which the software authors thought were sufficient for a "significant" assignment, e.g. Mascot originally reported "a significant match is typically a score of the order of 70" [Perkins 1999]. However, as the number of spectra per experiment grew, it became increasingly important to quantify the likelihood of an annotation being correct.

Work by Keller and colleagues explored how to assign a probability value to each spectrum assignment [Keller 2002]. A crucial piece of information in the calculation of a p-value is to determine the probability that a false hit receives a given score. To this end, they generated spectra from a set of 18 purified proteins. These spectra were searched with SEQUEST against a special database: the 18 true sequences and the fly proteome, which consisted of over 10,000 decoy sequences. With a database of over 99% decoy sequences, the likelihood of a false-positive match to a target sequence is unlikely. To a first approximation, any annotation to a fly sequence was false, and any annotation to their 18 was correct. They created a new multi-feature score, which contained the X-corr score and the delta score, and plotted the distribution of scores for the true and false annotations. Then, for any given score, they could empirically determine the probability that the annotation was false.

Most MS/MS experiments do not contain such a highly purified and known set of true positive proteins. Thus, to be applicable, this method had to be generalized. They showed that the scores of any search could be modeled as a mixture of distributions. A gamma distribution was chosen to model the false annotations, with its tail protruding to the right and slightly overlapping with the true distribution, which was modeled with a normal distribution. Many labs now have advanced this idea and run all their samples against a target-decoy database, and directly measure the false-positives at a given score.

1.2 Post-translational Modifications

Proteins are frequently modified to initiate or regulate their function. These modifications can be sequence processing, such as cleavage to activate neural signaling peptides, or in terms of chemical additions to amino acid sequences like the addition of a methyl group to a lysine. Most commonly the later is referred to as a post-translational modification (PTM), although this may technically apply to both cases. For this dissertation, the term post-translational modification refers to a chemical, covalent modification on an amino acid. It has long been known that a vast diversity of modifications decorate protein sequences [Uy 1977]. Uy and Wold cataloged 140 amino acids and amino acid derivatives (PTM) present in proteins. Recent work by the Zubarev group has shown that, in MS/MS samples, there is nearly one modification per amino acid [Nielsen 2006]. Here I highlight just two well-studied examples of modifications that change protein function: the histone code and phosphorylation mediated NF-KB activiation.

Four histone proteins make up the core components of chromatin fibers that organize DNA within the nucleus. The packing state of chromatin has a great influence on the ability of transcription factors to find their DNA targets and mediate gene transcription. Condensed chromatin, or heterochromatin, is tightly associated and affords little access to DNA. Hence, genes in these regions are generally less active transcriptionally. Decondensed chromatin, or euchromatin, allows for greater access to the DNA and generally correlates with increased gene transcription. Post-translational modifications on the histone proteins (methylation, acetylation, phosphorylation, and ubiquitylation) are thought to be one of the factors that influence and regulate the state of chromatin. Some modifications, e.g. methylation of H3 lysine residue number 9, have been shown to affect the state of chromatin. This modification leads to dimerization of H3 with heterochromatin protein HP1 and stabilizes higher order heterochromatin [Cosgrove 2005]. The Histone Code is a generalization of this phenomena, and states that post-translation modification of the histone proteins regulates the chromatin state, and thus leads to changes in gene expression.

Protien phosphorylation is the most widely studied and most widely utilized PTM for signal transduction [Ubersax 2007]. Phosphorylation is a reversible addition of PO_4 to (most commonly) serine, threonine, or tyrosine. Protein phosphorylation is an ideal means of signaling, because the cell does not need to transcribe or translate any new proteins. The constituent members of the signal are already present, and only need to be activated, or relay the signal. In NF-KB signaling, a transcription factor (NF-KB) is held inactive outside the nucleus by its primary inhibitor iKB. Extracellular signals activate iKB kinase, which phosphorylates the inhibitor iKB leading to its degradation. Once the inhibitor is degraded, NF-KB is free to enter the nucleus and promote gene transcription. Thus phosphorylation of iKB serves as a signal for its own degradation, and also a switch for the general control mechanism of the response pathway.

Detecting post-translational modifications is an especially attractive application for tandem mass spectrometry, as the chemical addition changes the mass of the peptide. Early work by Carr and colleagues used mass spectrometry to identify n-Tetradecanoyl as a post-translational modification on the n-terminus of cAMP-dependent protein kinase [Carr 1982]. In particular, the identification of n-terminal modifications was especially exciting because modifications to the n-terminus caused complications with Edman sequencing, the standard protein sequencing technique at the time. Indeed, protein mass spectrometry is the only technique capable of identifying PTMs at a proteomic scale.

Some modifications, like methylation, are tightly bound to the peptide and do not alter the fragmentation. Thus identifying peptides with these modifications is straightforward. Database search algorithms accept as input a list of potential modification masses and their respective amino acid. Programs then alter their amino acid dictionary to include the new masses (e.g. 142 for methylated lysine). However, some modifications like phosphorylation are weakly bound to the peptide and alter the fragmentation characteristics [DeGnore 1998]. The phosphate bond is so weak that collision inside the mass spectrometer frequently breaks only the phosphate bond, and nothing else. This leaves the peptide intact and depletes the spectrum of sequence informative backbone breaks. Unfortunately, the changes to the spectrum are so dramatic that standard search algorithms perform poorly. To overcome this, many researchers resort to manual validation of large datasets searching for phosphorylated peptides. In Chapter 2, work from the dissertation author presents the computational formalism for solving this problem. By learning the characteristic fragmentation of phosphorylated peptides, weaker spectra can be accurately annotated.

1.3 Proteogenomics

The identification of a complete protein-coding catalog is a fundamental goal of genome projects. As *ab initio* gene prediction algorithms remain inaccurate, additional evidences are often incorporated into genome annotation pipelines, most commonly cDNA and EST libraries [Brent 2008]. Although extremely beneficial, cDNA and EST libraries have both a theoretical and practical drawback. First, ESTs and cDNAs are evidence of transcription and could represent non protein-coding sequence. Work by Clamp and colleagues reject 20% of the human protein-coding genes as chance RNA transcripts [Clamp 2007]. More practically, transcript evidence is often erroneous. In the latest Arabidopsis genome release, curators rejected 25% of the proposed gene model updates due to poor-quality sequences, ambiguous orientation, misalignment and other problems [Swarbreck 2008].

An alternate technique to improve genome annotation is comparative genomics. Originally used in an *ad hoc* manner to find known genes, several genomes have now been systematically aligned to genomes of closely related organisms [Kellis 2003, Lin 2007, Clamp 2007]. The alignment reveals regions of genome conservation which are then analyzed for functional signatures which can distinguish between protein-coding and non-coding elements. This approach is both labor and cost intensive, in that it requires the completion of additional genome projects before the target genome benefits.

Large scale tandem mass spectrometry experiments can identify tens of thousands of peptides from thousands of proteins. In addition to the original biological context, the observed peptide sequences are direct evidence of gene expression and translation. Peptide sequences can be mapped back to the genome to validate current gene models and also discover novel protein sequences [Tanner 2007, Savidor 2006, Brunner 2007]. Proteomics-based genome annotation, or proteogenomics, obviates several problems inherent in transcript-based protein prediction. Specifically, proteogenomics can determine reading frame, translational start and stop sites, alternative splicing, and the validity of short ORFs. By complementing current annotation pipelines with proteogenomics, a more complete and accurate protein coding catalog can be achieved.

One of the fundamental drawbacks of proteogenomics is the inherent sampling limitations of a mass spectrometer [Stasky 2004]. Unlike nucleotide based assays which work on hybridization, MS/MS data is collected in sync with chromatography. Thus, when more peptides elute from the column than can be sampled by the machine - which frequently happens in complex samples - not every peptide can be observed. Observed peptides are also limited by ionizability, mass range of the instrument and the length of the peptide after proteolysis. Although this may change with newer technology, the efficiency of a single experiment to capture the diversity of the protein sample is currently not as complete as microarrays.

In Chapter 3, we demonstrate the benefit of proteogenomics in both validating and discovering protein-coding genes of Arabidopsis thaliana. We extend previous proteogenomic research in two ways. First, we utilize alternative fractionation techniques to extend coverage of the proteome, specifically low-abundance proteins. Second, we explicitly and thoroughly search the genome for novel protein coding sequences. We search 21 million tandem mass spectra against the TAIR7 proteome and find 126,055 unique peptides from 12,702 proteins. Additionally we identify 18,024 peptides which are not in the current annotation. These represent almost 900 new genes, and corrections and additions to another nearly 700 current genes.

1.4 Open Problems

As the proteomics community continues to grow and adventure into new applications, I believe that computational modeling will play an increasingly important role in research. Some may claim that technological advances (such as the zero mass error instruments discussed at ASMS 2008) will make computational tools obsolete. However, I argue that instrumentation and experimental improvements will allow or often require new computational tools. I cite an example from the history of MS/MS instrumentation. Hyper accurate instrumentation is a “new movement” in tandem mass spectrometry easing many computational burdens. However, in the 1960s, Klaus Biemann was reporting 0.003 mass unit deviations [Biemann 1966], the equivalent of the new Orbitrap. So what have we gained in the last 40 years? Speed. The LTQ instrument records orders of magnitude more data than its predecessors. The sacrifice for this technological advancement was accuracy. So as instrumentation changed, it opened new doors for computational research. Large data sets were the driving force for efficient algorithms like Inspect. Also, less instrument accuracy drove the need for computational error correction, like parent mass correction. Now with the Orbitrap, label free quantitation is possible, but requires rigorous statistical and computational modeling. Thus, I believe that future advances will create, not obviate, computational proteomics research. Among a myriad of possible topics, I discuss here on two potential research topics: the reapplication of an old computational tool – the self-convolution, and the integration of systems biology into proteomics.

As instrumentation becomes more accurate, the need for parent mass correction may become obsolete. The underlying algorithm, however, may take on a new application. Spectrum self-convolution serves to find the regularly occurring ion pairs with a spectrum. This is accomplished by multiplying the intensity of a peak and its cognate (parent mass – peak mass). Its original intent was to find the parent mass value (parent mass +/- epsilon) which maximized such pairs. However, now that parent mass is not an issue the utility changes. Using this to judge the number of b/y matching pairs, Alexey Nesvizhskii used this algorithm to filter spectra which are unlikely to be true peptide spectra. I however, wish to explore another application, that of discovering novel fragmentation.

As shown in chapter 2, the definition of peak pair can be expanded beyond the simple b/y ladder. By including a numeric offset, we can also find pairs with a neutral loss, e.g. b-water and y. In the context of phosphopeptides this idea is necessary for proper parent mass correction, as the spectra have weak b/y ladders and cannot be corrected without the inclusion of the neutral loss offsets. In collaboration with Pieter Dorrenstein at UCSD, we are now attempting to do the same for phosphopantetheinyalted peptides. They too undergo a frequent neutral loss that removes intensity from the b/y ladder. Using the spectrum self-convolution on accurate parent mass spectra, we can discover regular neutral loss products in unannotated spectra. We hope this work will make significant headway into one of MS/MS's most intriguing questions, "what are the other 80% of spectra?"

Another area of interest for me is the mix of proteomics and systems biology. As datasets continue to increase in size, the proteomics community has sometimes devolved into an arms race. Who can make the most spectra? Who can find the most phosphorylation events? As an active participant in some races, I cannot be too critical. Such preliminary studies are essential, but limited in application. In the end, discovery and not description is the goal.

Phosphoproteomics, for example, is in need of computation tools to advance their questions. For phosphorproteomics research, finding the site of modification is really only the beginning. This single site is not a pathway. We must determine the responsible kinase, and most importantly the intent of modification. The MAP kinase cascade is a classic example of phosphorylation-mediated cell signaling. Here, multiple kinases phosphorylate each other to relay the signal down to its final effectors. A MAP kinase kinase kinase (or Map3K) is activated by some extra cellular stimuli and starts the relay by phosphorylating a MAP kinase kinase, Map2K. This protein in turn phosphorylates a MAP kinase, which then relays the signal downstream again through phosphorylation. Absent context, a large phosphoproteomics survey would only note three proteins with phosphorylation sites, but miss the coordination and structure of the signaling cascade.

Systems biology works to integrate information at a systems level into interaction maps and pathways. This type of infrastructure is what is needed for MS/MS applications to uncover biological impact. Currently this is sometimes done in an *ad hoc* manner, and sometimes it is left undone. In the future, mapping to known pathways will also lose impact, as the purpose is to gain new insight.

Chapter 2: Phosphorylation-specific MS/MS scoring.

The promise of mass spectrometry as a tool for probing signal-transduction is predicated on reliable identification of post-translational modifications. Phosphorylations are key mediators of cellular signaling, yet are hard to detect, partly because of unusual fragmentation patterns of phospho-peptides. In addition to being accurate, MS/MS identification software must be robust and efficient to deal with increasingly large spectral data-sets. Here we present a new scoring function for the Inspect software for phosphorylated peptide tandem mass spectra for ion-trap instruments, without the need for manual validation. The scoring function was modeled by learning fragmentation patterns from 7677 validated phospho-peptide spectra. We compare our algorithm against SEQUEST and X!Tandem on testing and training datasets. At a 1% false positive rate, Inspect identified the greatest total number of phosphorylated spectra, 13% more than SEQUEST and 39% more than X!Tandem. Spectra identified by Inspect tended to score better in several spectral quality measures. Furthermore, Inspect runs much faster than either SEQUEST or X!Tandem, making desktop phosphoproteomics feasible. Finally, we used our new models to reanalyze a corpus of 423,000 LTQ spectra acquired for a phospho-proteome analysis of S. cerevisiae DNA damage and repair pathways and discover 43% more phospho-peptides than the previous study.

2.1: Introduction

Finding sites of protein modification has been of great interest in proteomics [Jensen 2006]. Protein phosphorylation, which regulates many cellular processes [Hunter 2000], has been a prime target of research. To enable the large-scale discovery of protein phosphorylation sites, a variety of experimental techniques have been developed for phospho-peptide enrichment [Zhou 2001, Andersson 1986, Pinske 2004]. As a result, tandem mass spectrometry has been widely used to annotate the phosphoproteome of both whole cells [Macek 2007, Chi 2007, Olsen 2006, Villén 2007, Molina 2007, Chitteti 2007, Shu 2004], and sub-cellular fractions [Nousiainen 2006, Trinidad 2006, Lee 2007]. As the protocols for isolating phospho-peptides improve, the bottleneck for phospho-peptide identification has shifted to data interpretation of the MS/MS spectra. Most search algorithms are not optimized specifically for phospho-peptide spectra which could have very different characteristics.

Phospho-peptide fragmentation under collision induced disassociation (CID) is perceptibly different from unmodified peptides. Cleavage is highly biased to the phosphoester bond [DeGnore 1998]. Phosphate loss from the precursor typically dominates the MS/MS spectrum, averaging 20-30% of the total ion current. Moreover, b/y ions also frequently lose the phosphate, further weakening the signal of the b/y ladder, complicating peptide identification. Highlighting the difficulty of accurate phospho-peptide identification are studies which set a weak score cutoff followed by either exhaustive manual validation [Nousiainen 2006, Lee 2007], or substantial post-processing techniques to obtain a low false-discovery rate [Macek 2007, Villén 2007]. Such attempts to recover misscored false-negatives require subjective intervention to ensure quality identifications. Although manual validation is invaluable for gaining an overall confidence in the results, its application to phospho-proteome scale searches (tens of thousands of spectra) is neither realistic nor prudent [Beausoleil 2006].

Current algorithmic improvements for phospho-peptide identification focus on post-processing instead of the original scoring function. Lu et al. developed criteria for automated validation which judges annotations based on characteristics of phospho-peptide spectra [Lu 2007]. Due to the potential ambiguity in the placement of the phosphate group within the peptide, Beausoleil and colleagues have developed a confidence metric for phosphate localization [Beausoleil 2006]. While validation and localization are important, they help primarily in reducing false-positives, but not false-negatives. To overcome poor scoring of false-negatives, a scoring function must be trained to discriminate annotations based on the unique fragmentation probabilities of phospho-peptide spectra. Moreover, the development of an improved scoring function does not preclude application of post-processing techniques.

Our strategy for scoring phospho-peptides is based on well-established principles, specifically, that fragmentation of the peptide backbone is not uniform [Loo 1993, Hunt 1986, Havilio 2003, Breci 2003]; all ion types are not equally likely to appear in the spectrum with uniform intensity. Classic examples include proline directed fragmentation and the isotopic envelope: fragmentation N-terminal to a proline produces more intense b/y ions than fragmentation C-terminal to the proline; isotopic peaks (e.g., y+1) rarely occur without the monoisotopic peak. More generally, the expected intensity of an ion can change based on flanking residue, related peak presence or other factors. If we consider annotations in context, we obtain a more discerning scoring function. The main contribution of our paper is an automated system that learns the fragmentation propensities and peak dependencies of phospho-peptides using a large training corpus of annotated spectra. We use this knowledge to devise a Bayesian network [Jensen 2001] based scoring function for the Inspect software [Tanner 2005].

Our new algorithm outperforms current algorithms (SEQUEST, COMET, X!Tandem) in both speed and accuracy on large training and testing datasets of spectra acquired on ion-trap instruments. On a small test set of 6410 spectra, at a fixed false-discovery rate for each program (1%), Inspect had the highest true-positive rate, annotating 13% more spectra than SEQUEST and 39% more than X!Tandem. Additionally, when we reanalyze a previously published dataset of 423,000 spectra, we recover 43% more phospho-peptides than the original work [Smolka 2007]. A better recovery of phospho-peptides from the spectra provides a more complete view of the phospho-proteome, enabling researchers to better understand the dynamic signaling processes of the cell. Furthermore, the run time was one or two orders of magnitude faster than current algorithms, making desktop phospho-proteome analysis possible. The new models have been incorporated into the Inspect software package, which is freely available for download from our webserver, http://peptide.ucsd.edu/. In addition to the strong performance of the new models, we discuss the distinct characteristics of phospho-peptide fragmentation probabilities and the use of Bayesian networks for probabilistic scoring, both of which are of independent interest.

2.2 Materials and Methods

Overview: MS/MS peptide identification programs typically have four major stages: spectral-preprocessing, database filtering (searching), scoring and validation [Eng 1994]. Each stage functions as a distinct module within Inspect. Ion-trap instruments, like the LTQ, are the workhorse instrument of proteomics. However, the accuracy of these instruments necessitates the preprocessing steps of parent mass correction and charge state determination. The experimental parent mass is often off by 2-3 Da. Moreover, the charge state of an LTQ spectrum is ambiguous because the isotopic envelope of the precursor cannot be established. High accuracy instruments such as a QTOF or Orbitrap may not require these corrections.

After parent mass and charge state are determined, the spectrum is searched against a protein database to produce a list of candidate annotations. Database filtering is used to rapidly eliminate many of the peptides from the database without explicitly scoring them. Parent-mass based filters are common but not as effective when dealing with post-translational modifications. Inspect uses a tag-based search for filtering by performing a partial de novo interpretation during pre-processing [Tanner 2005]. Tag-based filtering is orders of magnitude more efficient than other filters, but requires accurate tagging.

The filtered peptides are rank-ordered based on scoring against the spectrum. This score represents how well the annotation agrees with the spectrum's peak list. The best candidate peptide should get the highest score, followed by the next best candidate, and so on. Even with an accurate scoring function, the top-scoring peptide might still not be the correct one. It could be, for example, that the correct peptide is not in the database, or that there isn't enough information in the spectrum to distinguish between the top two peptides. A final validation step is used to determine the probability that the top scoring peptide is the correct one. In this work we focus on the preprocessing and scoring steps in the context of phosphorylated peptides. The filtering and validation steps remain unchanged. The new models are incorporated into the Inspect software version 2007.07.12 and later.

Parent Mass Correction: Correcting the observed parent mass is a crucial preprocessing step for any de novo MS/MS program [Dancík 1999]. Inspect's tag generation utilizes a partial de novo interpretation of the spectrum, and is therefore sensitive to erroneous parent mass values. Peptide fragmentation creates matching b/y ion pairs, whose mass sums to the parent mass of the precursor ion. Thus given a spectrum, we can determine the parent mass by finding matching b/y ion pairs. Our parent mass correction routine is based on spectrum self-convolution introduced by Dancík et al. [Dancík 1999]. Define M as the measured mass of the charge 1 precursor ion; Pi as the m/z of the ith peak in a spectrum; *I(v)* is the intensity of the peak at mass value *v* (binned to 0.3 Da). Dancík corrected the parent mass of a spectrum within the range [M – ε, M + ε] as

$$M^* = \underset{M-\varepsilon \leq m \leq M+\varepsilon}{\operatorname{argmax}} \sum_i I(P_i) I(m - P_i)$$

The intuition here is that at the correct parent mass, M^*, we will see a large number of high intensity cognate pairs corresponding to the b/y ladder [Dancík 1999, Venable 2006].

We extend this algorithm in two ways, exploiting the neutral losses from phospho-peptides. Here it is necessary to define the two types of phosphate neutral losses from peptides. The loss of phosphate from a b or y ion is called a fragment neutral loss. The second ion type is neutral loss from the precursor ion, or $M - p$. These two distinct ion types are used in different ways in the models.

Our first extension to the Dancík algorithm is the inclusion of a mass offset into the convolution. In addition to the cognate pair at (P_i, $M^* - P_i$), we also expect to see a pair at (P_i, $M^* - P_i + 1$), corresponding to matching a +1 isotope, e.g. b/y+1 or b+1/y. Similarly we would expect to see cognate pairs from neutral losses. We modify the original convolution equation to take as input an arbitrary offset, O:

$$f_{m,O} = \sum_i I(P_i)I(m - P_i + O)$$

Using the training data as input, we plot $\square_{m,O}$ for all values of O between -101 and +4 (Figure 2.1). The highest $\square_{m,O}$ values represent offsets for which intense pairs (P_i, $M^* - P_i + O$) were found. For example, when no offset is applied (Figure 2.1, x=0), equation 2 sums the intensity of all b/y peak pairs in the training set. The strong peaks at -18 and -17 (water and ammonia loss) and +1 (isotopic peak) all correspond to known biological events. A strong peak was observed at -98, phosphate loss. Unfortunately, this feature did not add discriminatory power to the parent mass correction model; see Results for a possible explanation. Based on these observations, we define a feature vector

$\overline{\mathbf{F}}(m)$ = [f(m,0) f(m,1) f(m,–17) f(m,–18)]. At the correct parent mass, we expect to see strong values in $\overline{\mathrm{F}}(m)$.

Our second extension of the Dancík algorithm is the explicit use of the precursor neutral loss. For phosphopeptides, we expect to see an intense neutral loss from the precursor, M–p. We model this by the intensity and skew of the peak from the expected position, m/z – 98/z. The most intense peak at this location (±0.5 Da) is assigned the M –p identity. Its intensity, Ip, and skew from expected location, Sp, are added to the feature set. We use the feature set $(\overline{\mathrm{F}}(m), I_p, S_p)$ as input to a Linear Discriminant Analysis model for distinguishing the correct mass from a range. Formally,

$$M^* = \underset{M-\varepsilon \leq m \leq M+\varepsilon}{\operatorname{argmax}} \quad \mathrm{LDA}(\overline{\mathbf{F}}(m), I_p, S_p)$$

The model was trained to find the optimal linear combination of features by comparing correct and incorrect parent mass values for spectra in the training set. We show in Results that this model vastly outperforms models for unmodified spectra. For charge state correction, we closely follow the features and methods of Klammer et al., [Klammer 2005], but include the M – p peak intensity as an additional feature.

Scoring: Inspect’s scoring function is comprised of six features: percent of total ion current explained by the annotation, fraction of b ions observed, fraction of y ions observed, length of the peptide, number of enzymatically digested endpoints, and the cut-score (described below). Values for each of these features are used as input into a Support Vector Machine [Noble 2006], which returns the final score of Inspect, the MQScore. A new set of fragmentation probabilities impacts only the cut score, as explained below.

Note that a peptide (with parent residue mass PM) can be described by a set of cuts, or prefix residue masses P1 < P2 < · · · < PM. Note that if a certain cut Pj is indeed a true cut for the spectrum, we will see many peaks corresponding to the fragment ions that support this cut (b, y, b-H2O. . .) . Figure 2.2 illustrates this for the peptide RGSphosDVEDASNAK. CID fragmentation between the 7th and 8th residue (cut P7) predominantly produces b7 and y5. However, we also see other related ions. Following Frank and Pevzner [Frank 2005], let $P_{CID}(\bar{I}\,|P_j,S)$ denote the probability of detecting a set of ions, $\bar{I}$, given that P_j is a valid cut of the spectrum S. From Figure 2.2, Pj is P7; $\bar{I}$ is [b7, b7 + 1, b7-H3PO4, y5, y5 + 1, y5-NH3, and y5-H2O]. As the null hypothesis, let $P_{\varphi}(\bar{I}\,|P_j,S)$ denote the probability of observing $\bar{I}$ by chance. The cut-score of a peptide is given by $\sum_j Score(P_j,S)$where,

$$Score(P_j,S) = \log\frac{P_{CID}(\bar{I}\,|P_j,S)}{P_{\varphi}(\bar{I}\,|P_j,S)}$$

The critical part of this is the determination of $P_{CID}(\bar{I} = [I_0, I_1, \dots]|P_j,S)$ given that the occurrence of fragment ions are not independent. It is usually not possible to estimate all dependencies due to lack of sufficient training samples. We approximate this with a Bayesian network [Jensen 2001] described by a directed acyclic graph on the ion-types with limited outgoing edges (dependencies) for each ion-type. Let $I_{\pi(i)}$ denote the set of ions that I_i depends upon. Then,

$$P_{CID}\left(\vec{I} = [I_0, I_1, \ldots] \middle| P_j, S\right) \cong \prod_i P_{CID}(I_i \mid P_j, I_{\pi(i)}, S)$$

The set of dependencies $I_{\pi(i)}$ is not well-understood for phospho-peptides. Therefore, we computed a minimum entropy architecture based on observed fragmentation in our training data-set.

To get robust estimates of conditional probabilities, each possible combination of values should have a potentially large number of observed instances. To prevent the network from being too large and to ensure that the calculated statistics are well-formed, we only include as nodes the ion types which are regularly observed in phospho-peptide CID fragmentation. We required an observed frequency of 1 instance per spectrum. Observed frequency was calculated by making an offset frequency histogram [Dancík 1999] of all spectra in the training set, Figure 2.3. Also, we generalize this framework slightly. Nodes in the network include ion type and also associated meta data. The ion-types are listed in Table 2.1. A variety of meta data was investigated. Only those with a high information content were kept: amino acid flanking the break, spectrum region (divide m/z range into 5 equally sized bins), and whether the phosphate group is on this fragment of the peptide (ContainPhos).

To estimate $P_{CID}(I_i | P_j, I_{\pi(i)}, S)$, we tabulated the values for these nodes for each cut of each spectrum in the training set. From this large table we calculated both entropy (Shannon information entropy) and conditional entropy. Bayesian networks require a topological ordering for the directed acyclic graph. As many network reconstructions are possible, to algorithmically compute the optimal ordering would require a much larger training data set. Therefore, we use an ordering based on ion prevelance, or the fraction of possible ions (for a given ion type) observed in the training data. When including the non-ion type nodes, our final order was: spectrum region, flanking amino acid, ContainPhos, and the ion list as ordered above, Table 2.1. Let Pred(X) denote the set of nodes that precede node X in the node order. To construct the network, we choose at most 2 parent nodes (i.e. $I_{\pi(j)}$) for each node (i.e. I_j). Thus,

$$I_{\pi(j)} = \underset{P_1, P_2 \in Pred(I_j)}{\operatorname{argmin}} \quad H(X | P_1, P_2)$$

An example probability table is shown in Table 2.2. Peak binning into strong, medium, weak, and absent are based off the median peak intensity, and is a learned parameter. The set of I_(j) comprises the Bayesian network as shown in

Figure 2.4. After the network structure is finalized, conditional probability tables representing the ion profiles are stored and this becomes our Bayesian model.

Generating the Training Set: 62,000 LTQ MS/MS spectra were generated from S. cerevisiae and an additional 109,000 LTQ MS/MS spectra were generated from S. pombe. These spectra came from whole cell lysates, purified by IMAC as described in Sample Preparation. To obtain a corpus of highly confident phosphorylated spectra, we relied on the overlap in annotation from four independent programs: Inspect [Tanner 2005], SEQUEST [Eng 1994], COMET [Keller 2005], and X!Tandem [Craig 2004]. SEQUEST was run on a SageN Sorcerer system; other programs were downloaded and installed on a local linux cluster. Each program searched the dataset allowing up to two phosphorylations on serine, threonine, or tyrosine as a variable modification; parent mass tolerance of 3 Da, fragment mass tolerance of 0.5 Da. SEQUEST, COMET, and X!Tandem set a semi-tryptic cleavage specificity with 2 missed cleavages; Inspect has no such parameter due to the tag, not parent mass, filter. The database for S. cerevisiae was downloaded from http://www.yeastgenome.org on January 12, 2007; the database for S. pombe was downloaded from http://pombe.nci.nih.gov/genome on July 28, 2006. Each database was concatenated with decoy protein sequences. To create the decoy database, we shuffled each protein record once. Results of Inspect and X!Tandem were ranked based on the provided p-values. SEQUEST results were processed with the trans proteomic pipeline, and the Peptide Prophet p-value was used for ranking. COMET results were ranked by using both the ΔN and Z-score. Each program's results were filtered to 2% false-discovery rate, as measured by hits to decoy sequences [Elias 2005]. We compiled the training set from these filtered results by requiring that a spectrum be identified by at least three of the four programs. We observed that the overlap between any two programs was typically 70%. The final set consisted of 7677 spectra (5218 charge 2 and 2459 charge 3). The total number of distinct peptides was 2293 charge 2 and 1087 charge 3. This training set was used for all model building. There are two test sets. The first is 6410 LTQ MS/MS spectra from S. cerevisiae whole cell lysate enriched for phospho-peptides by IMAC. Time trials for this test set were performed on a single processor of the linux cluster (including a local installation of SEQUEST). The second is 423,000 LTQ MS/MS spectra from S. cerevisiae as described [Smolka 2007]. All datasets are available from the authors on request.

Sample Preparation: Cell growth. For Saccharomyces cerevisiae, fifty milliliters of budding yeast cells (BY4741) were grown in YPD medium to an OD600 of 0.5 and cells were treated with 0.05% MMS for 3 hours. For Schizosaccharomyces pombe, fifty milliliters of fission yeast cells (FY259) were grown in YES medium to an OD600 of 0.5 and cells were treated with 0.01% MMS for 3 hours.

Protein extraction and trypsin digestion. Cells were broken in an ice-cooled bead-beater with 2 ml lysis buffer containing 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.2% NP40, 0.5 mM DTT, 5 mM NaF, 10 mM β-glycerolphosphate, 1mM sodium vanadate, 5mM EDTA, 1mM PMSF, 0.2mM Benzamidine, 1 μM Leupeptin,1.5 μM Pepstatin. Cell debris was removed by centrifugation at 30,000xG for 30 minutes. Approximately 10 mg of proteins were then denatured by boiling in the presence of 2% SDS and 10 mM DTT for 5 minutes. Proteins were alkylated with 50 mM iodoacetamide, precipitated with 3 volumes of cold ethanol:acetone (1:1, v/v) and then resuspended with buffer containing 2 M urea and 50 mM Tris-HCl, pH 8.0. Twenty micrograms of trypsin (Worthington, Lakewood, NJ) was added for overnight digestion, and then the tryptic peptides were desalted using a 200-milligram C18 column (Waters).

Phosphopeptide purification and mass spectrometry. Desalted peptides were dried in speed-vac, resuspended in 150 μL of 1 % acetic acid and loaded to a gel loading tip column containing 25 μL of immobilized metal affinity column (IMAC) resin. IMAC resin was prepared from silica Ni-NTA (Qiagen), where the nickel was substituted by iron as the bound metal. After loading of the peptides, the IMAC resin was washed twice with 25 μL of wash buffer containing 25 % acetonitrile, 100 mM NaCl and 0.1 % acetic acid.

Bound phosphopeptides were successively eluted by four different eluting solutions containing increasing concentrations of phosphoric acid (0.01%, 0.05%, 0.1%, 1%) to yield four distinct eluted fractions. Each of the four elutions was performed with 100 µL of solution and processed independently. Each fraction was transferred to a silanized glass insert (National Scientific, Rockwood, TN), dried under reduced pressure, resuspended in 10 µL of 0.1 % TFA and subjected to mass spectrometry analysis. Mass spectrometry experiments were performed using the 1100 QuadPump HPLC system (Agilent, Santa Clara, CA), the Ultimate 3000 autosampler (Dionex, Sunnyvale, CA), and the LTQ tandem mass spectrometer (Thermo Fischer Scientific, San Jose, CA). Four microliters of each eluted fraction were loaded using the autosampler via a 5 µL sample loop directly to an in-house packed 125 µm (inner diameter) x 20 cm microcapillary RP-HPLC column, packed with 3 µm C18 resin (Magic beads; Michrom Bioresources, Auburn, CA). For RP-HPLC-MS/MS analysis, Buffer I consisted of 0.1 % formic acid and 2 % acetonitrile. Buffer II consisted of 0.1 % formic acid and 80 % acetonitrile. A 120 min gradient from 15 % to 35 % Buffer II was used. Xcalibur 2.2 software (Thermo Fischer Scientific, San Jose, CA) was used for the data acquisition, and mass spectrometer was set to perform one full MS scan followed by 6 consecutive MS/MS scans according to the ion intensities detected in the full MS scan. The minimal threshold for the dependant scans was set to 6500 counts, and a dynamic exclusion list was used with the following settings:

repeat count of 1, repeat duration of 2 seconds, exclusion list size of 150, exclusion duration of 60 seconds, and exclusion mass width of 0.2 % relative to the reference mass. Raw data files were converted to mzXML with ReAdW 2006Nov01, http://tools.proteomecenter.org/ReAdW.php.

2.3 Results and Discussion

To generate the models we first obtained a highly confident training set of 7677 phospho-peptide spectra. These spectra were identified by at least three independent algorithms (see Methods).

Correcting the observed parent mass is a crucial preprocessing step for any de novo MS/MS program [Dancík 1999]. Inspect uses a partial *de novo* for tag generation and database filtering, and is therefore sensitive to erroneous parent mass values. As explained in Methods, we create a new model that explicitly uses neutral loss of phosphate. The trained models produce a significant improvement over the uncorrected and generic models (Table 2.3). For charge 3 spectra, the observed parent mass is only accurate (within 0.5 Da) 5% of the time. After parent mass correction, the accuracy is 90%. The new phosphorylation specific model has nearly twice as many spectra accurately predicted to 0.3 Da. We explored efficacy of using -98 as an offset for this model (Methods). However, the presence of amino acid masses close to 98 Da was confounding (i.e. 97 Da for proline and 99 Da for valine). Given the inaccuracy of the instrument, an offset of -98 Da could be a fragment neutral loss, or merely the next peak in the b/y ladder (compare the broad peak surrounding -98 with the narrow peak at -18 in Figure 2.1). Thus, when -98 was added to the feature set, the model gained no extra discriminatory power.

The scoring function of Inspect uses six spectrum features as input into an Support Vector Machine [Noble 2006] to get the final MQScore. As described in Methods, we use a Bayesian network to model the probability that each assigned peak is correct. Phospho-peptide fragmentation characteristics lead to a Bayesian network that is significantly different from the one for unmodified peptides. For example, our model clearly shows that b ions are twice as likely to be accompanied by a phosphate neutral loss than y ions (Table 2.2). Indeed the probability of observing y-H3PO4 given a strong y ion is very similar to the probability of observing b-H3PO4 given an absent b ion.

After training, the new Inspect program was run on the test dataset of 6410 MS/MS spectra and filtered to a 1% false-discovery rate. These results were compared to the results of SEQUEST and X!Tandem (Figure 2.5). We first note that Inspect is orders of magnitude faster than SEQUEST and X!Tandem. Inspect ran in 30 minutes on a desktop PC (1.6 GHz, 2GB RAM). X!Tandem took 6 hours and SEQUEST took 36 hours. As for identifications, Inspect identified a total of 1089 phospho-peptide spectra. This is 13% more than SEQUEST and 39% more than X!Tandem at the same false-discovery rate. Moreover, Inspect also had the strongest overlap with other confidently identified spectra.

When looking at the overlap in Figure 2.5 we see 15-20% of any program's annotations were unique. As it is possible that some of these could be false-positive identifications, we attempted to objectively compare the quality of these single program identifications. First, we plotted several features of phosphorylation spectra as discussed by Lu et al. [Lu 2007]. For each feature, we compare the unique annotations to the 501 consensus spectra (Figure 2.6). The most distinguishing feature of a phospho-peptide is the fragment neutral loss, e.g. b-H3PO4 [Lu 2007]. As a labile modification, the phosphate is frequently lost during CID, thus a true phospho-peptide spectra will contain many fragment neutral loss peaks. Figure 2.6a plots a histogram of the fragment neutral loss count per spectrum. Each line in the graph represent the distribution of fragment neutral losses in the identified spectra. The blue line is the distribution of the 501 consensus spectra; green is for the 203 spectra uniquely identified by Inspect; red is for the 116 spectra uniquely identified by SEQUEST; and grey is for the 92 spectra uniquely identified by X!Tandem. Here Inspect is the most similar to the consensus spectra, averaging more fragment neutral loss peaks per spectrum than SEQUEST or X!Tandem. A second highly characteristic feature of phospho-peptides is the intense M – p peak [Lu 2007]. This peak is typically the base peak of the spectrum and contains 20-30% of the total ion current. Figure 2.6b plots the intensity of this peak compared to be base peak of the spectrum. Here both Inspect and SEQUEST are very similar to the consensus spectra, each having a high percentage of spectra where the base peak is the M–p peak. Next we look at a common spectral quality metric, the fraction of b and y ions observed (Figure 2.6c, d). Again, Inspect more closely resembles the distribution of the consensus spectra, having on average a higher percentage of b/y ions observed. Another common quality assurance check, the intensity of proline directed fragmentation, shows no difference between the program's annotations, Figure 2.7.

A close look of the false-negatives of Inspect (the 74 spectra identified by SEQUEST and X!Tandem but not Inspect) shows the current deficiencies of the program. A total of 53 spectra were missed due to tagging errors. Of these, 39 are charge 3 spectra which are notoriously harder to tag. However, even though Inspect mis-tagged these spectra, it still identified nearly 20% more charge 3 spectra than SEQUEST (Table 2.4), a true-positive gain more than covering the false-negative loss. In ongoing research, we plan to improve the tagging accuracy of higher charge peptides. Among the remaining false-negatives, 10 represent spectra that Inspect identified but at a less significant p-value than 0.01. Another 9 of the false-negatives are charge determination errors; remaining spectra score poorly in the phosphorylation specific scoring function. When considering the 1165 spectra that Inspect identifies in total, losses for tagging (< 5%) and charge correction (< 1%) are minimal. Moreover, the number of false-negatives for Inspect is smaller than either SEQUEST or X!Tandem. It is worth reiterating that as mass-spectrometers become more accurate, charge detection, and tag identifications will improve dramatically. This potentially enables longer tags, further improving the speed of search.

After training and testing our models, we compared their performance against some of our previous work. A subset of the authors recently published a phospho-proteome analysis of DNA damage and repair pathways in S. cerevisiae [Smolka 2007]. This study identified 2457 non-redundant phospho-peptides found in both wild-type and kinase-null cells, using the COMET software. The dataset of 423,000 LTQ MS/MS spectra ran for ~40 days on a 22 processor Linux cluster, a total of ~ 21, 000 CPU hours. We re-ran these spectra with Inspect and annotated 41,077 spectra (8118 distinct peptides) at a false-discovery rate of 1%. When we restricted the results to peptides found in both wild-type and kinase-null samples we found 3518 non-redundant phospho-peptides, an increase of 43% from the original results. Additionally, the speed of Inspect was evident, running in less than 3 days on a single-processor desktop PC (66 CPU hours). A grid compiled version of Inspect finished the computation in 2 hours.

2.4 Conclusions

Recent studies have shown the importance of post-translational modifications (phosphorylations in particular) in mediating cellular signals.

While identification of phosphorylated peptides is key to these analyses, manual validation remains a standard of sorts in MS/MS phosphorylation studies. One reason for this standard is that a phospho-peptides's characteristic fragmentation pattern is easily picked out by eye. However, the other reason is simply that existing software are not trained to take advantage of the unique fragmentation patterns. We close this gap by training Inspect on a corpus of 7677 validated phospho-peptide spectra (3380 peptides).

In both training and multiple testing datasets, the new program discovers more phospho-peptides at a given false-discovery rate than any of the other programs considered. No program annotated all spectra; each algorithm has a measurable false-negative rate. Here we show that the learned scoring function of Inspect out performs the other algorithms and has the lowest false-negative rate. Second, examination of the quality of the identifications using a variety of objective criteria show that the Inspect identifications are of uniform high quality. Moreover, the tag-based filtering approach of Inspect allows it to be somewhere between 10 and 100 times faster than X!Tandem, SEQUEST and COMET. Our methodology is quite general and will be applied to other important modifications and instrumentation as data becomes available. With an increase in the quality and throughput of mass spectrometry data, our methods will find broad applicability.

Acknowledgments: S.H.P. is supported by NSF IGERT training grant DGE0504645. V.B. acknowledges support from NIH grant R01-RR16522. M.B.S. and H.Z. acknowledge support from NIH grant GM080469-01 and the Ludwig Institute for Cancer Research. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. The authors wish to thank Claudio Albuquerque, Ari Frank, and Nuno Bandeira for their many helpful conversations. The authors declare no financial conflict of interest. This chapter initially published as “Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome anlaysis.” Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, and Banfa V. 2008. Journal of Proteome Research, in press. The dissertation author was the primary author of this paper.

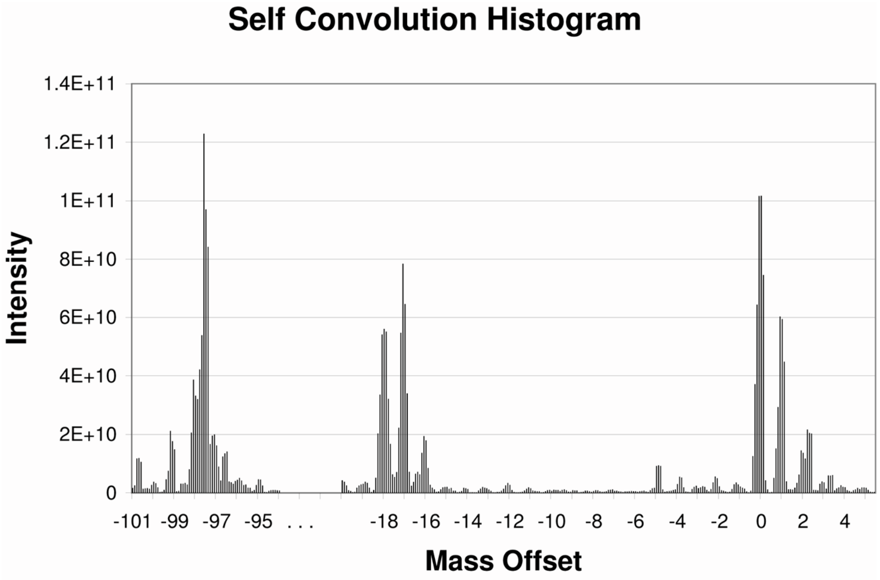


Figure 2.1 - Self Convolution of spectra in the training set. A spectrum self-convolution, as described in [Dancik 1999] is the product of a spectrum and it is reflection. Formally, eq 1 describes it as the product of intensity of a peak and it is cognate. Eq 2 introduces a mass offset, O, applied to the cognate peak. In this figure, O is plotted along the x-axis. The y-axis represents the value of the convolution in intensity units. As the self-convolution in eq 2 is applied to many spectra (all spectra in the training set), frequently observed offsets stand out. x = 0 represents the matching of b and y ions. The peak at x = 1 represents matching of an isotope to the b/y ladder: b + 1 and y, or b and y + 1. The peaks at -18, -17 and -98 correspond to the neutral losses of water and ammonia and phosphate. The peak for phosphate loss is not used in the final model. See Results and Discussion for possible explanation.

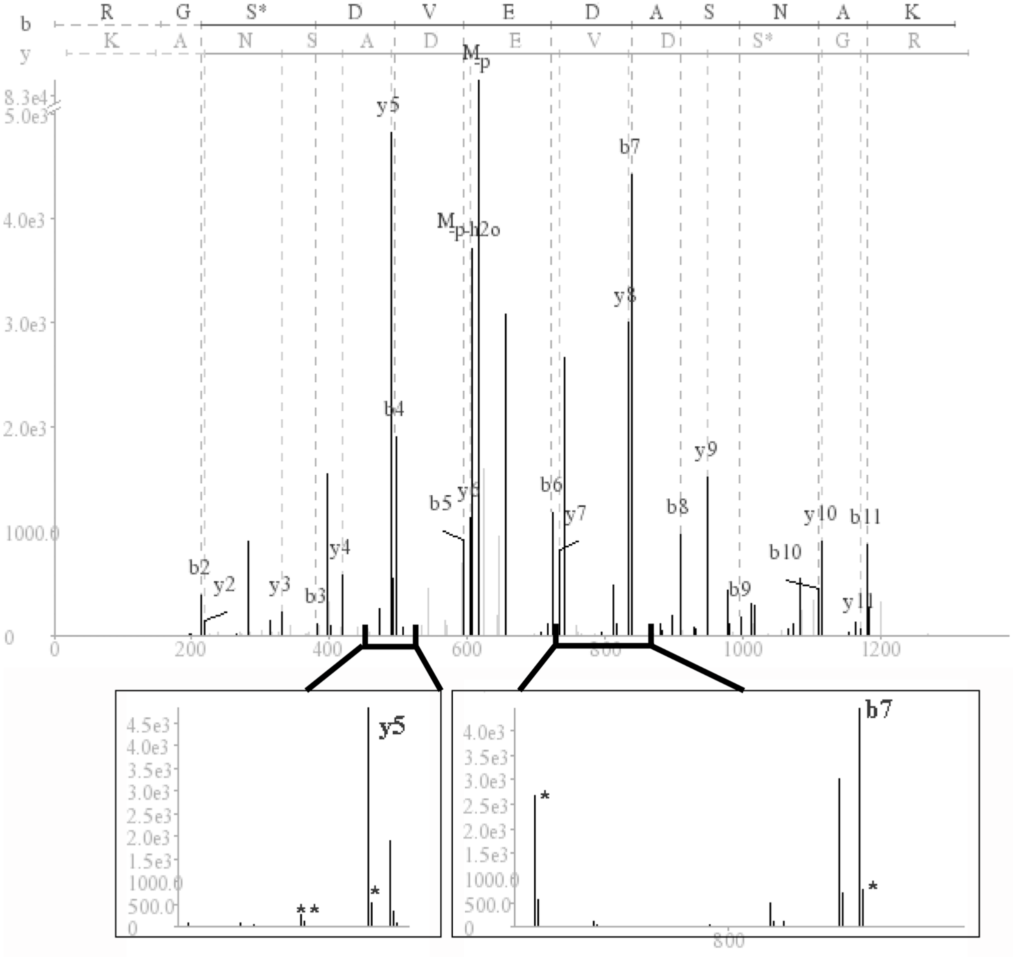


Figure 2.2 - A cut of the peptide. When the peptide RGSphosDVEDASNAK is fragmented between the seventh and eighth residue, the predominant resulting species are b7 (RGSphosDEVD) and y5 (ASNAK). Seven peaks support the this cut of the peptide, each adding to the confidence in the assignment. Zoom-in images around b7 and y5 show the related ions (/) present in this cut. In the b7 image, related ions include b7 + 1 and b7 - 98. In the y5 image, related ions include y5 + 1, y5 - 17 and y5 - 18. Note the break in the y-axis scale.

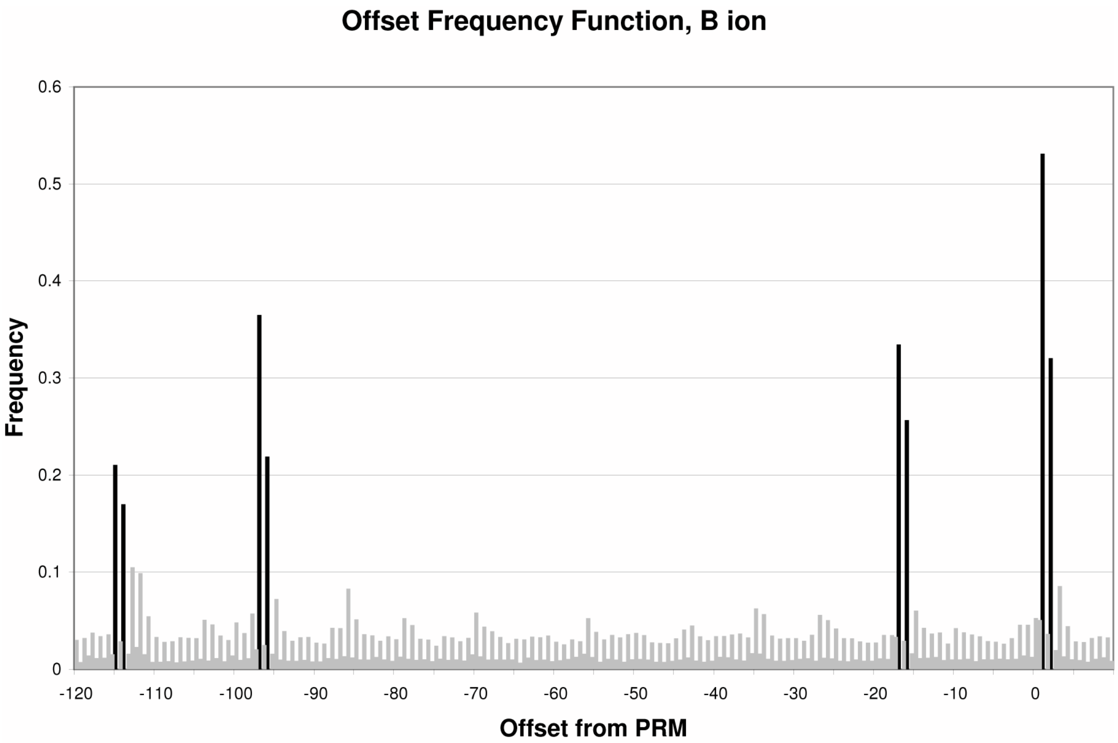


Figure 2.3 - The offset frequency function of b ions. Offsets from the prefix residue mass are plotted [Dancik 1999]. Offsets in black are the regularly occurring ions included in the model. Offsets in gray are not included. The strong gray offsets (e.g., -113) were discovered to be parts of the b/y ladder and not a novel neutral loss. Differentiating offsets caused by regularly occurring neutral losses from offsets caused by neighboring b peaks was done by iteratively removing the strongest offset from the spectra and repeating the analysis.

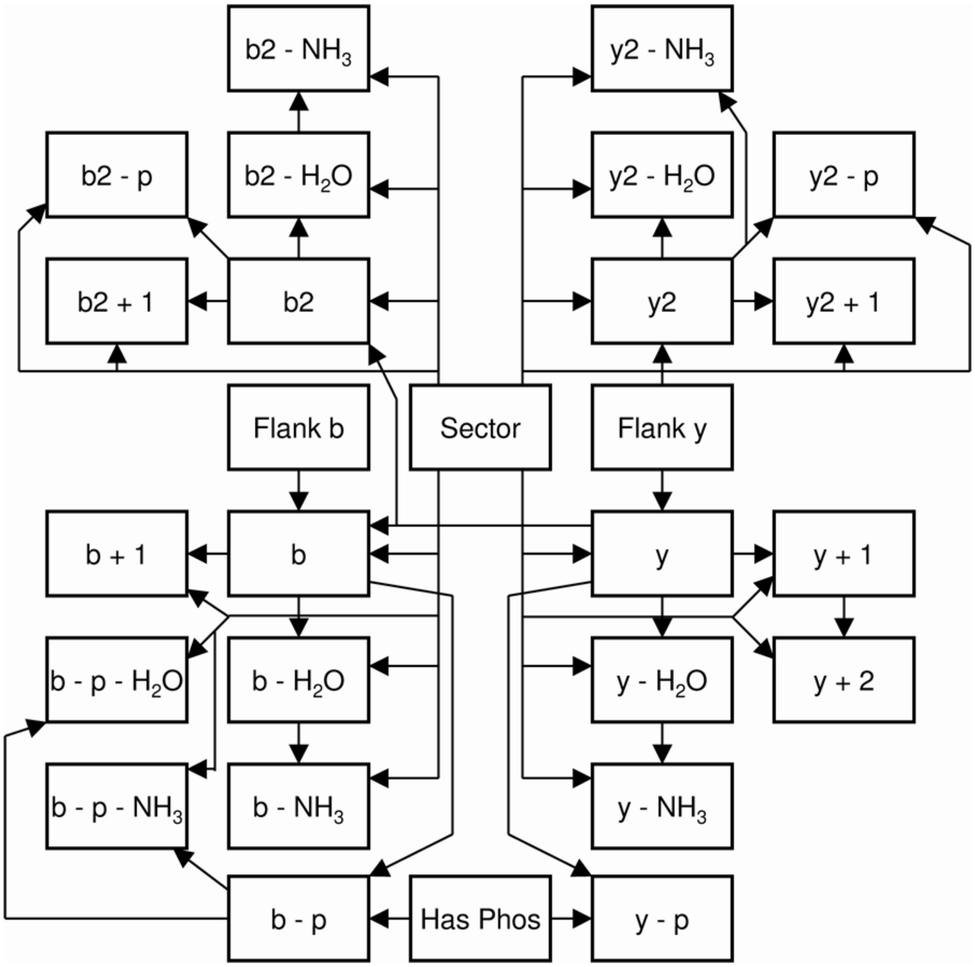


Figure 2.4 - Bayesian Network architecture. The nodes and connections of the Bayesian network.

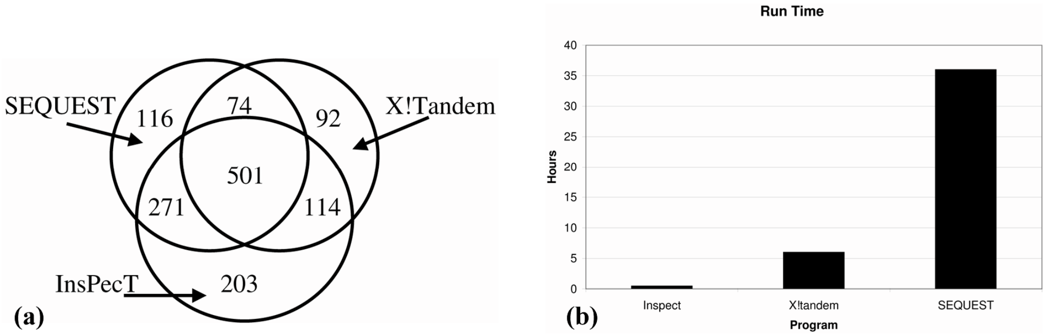


Figure 2.5 – Benchmarking Inspect's new scoring function on test data set 1. The test data set of 6410 MS/MS spectra was searched with the new Inspect models, SEQUEST and X!Tandem. Each program used as input the same mzXML spectrum file and the same database. Search parameters allowed up to 2 phosphorylations per peptide. Results of each program were filtered to 1% false-discovery by using the hits to the decoy database. (a) Overlap between annotations is plotted in a Venn diagram. Numbers represent individual spectra identified by an algorithm(s). (b) Run times are plotted for each program (single processor desktop PC).

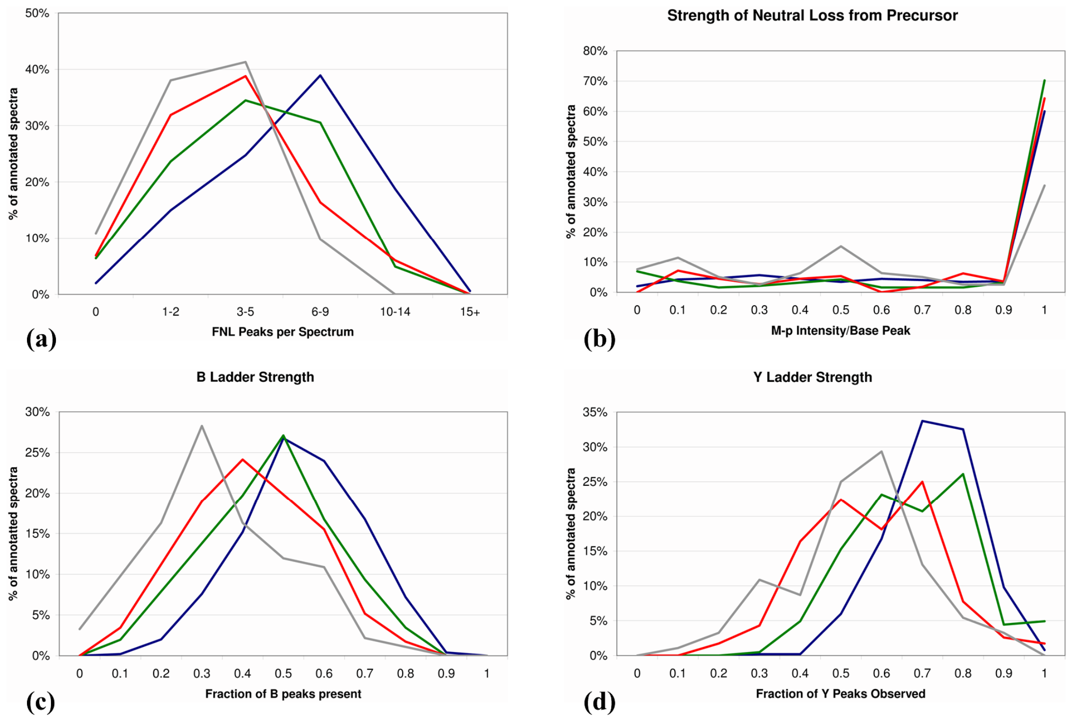


Figure 2.6 – Phosphopeptide spectral qualities. For each plot, the spectral quality feature was tabulated for all spectra in the testing data set. The blue line represents spectra annotated by all programs (consensus spectra). The green line represents spectra uniquely annotated by Inspect, red for unique SEQUEST annotations, and gray for unique X!Tandem annotations. Each figure shows the overlaid histogram of the results. (a) Number of fragment neutral loss peaks per spectrum. (b) Intensity of the M - p peak compared to the base peak of the spectrum. (c) Strength of the b ion ladder. (d) Strength of the y ion ladder.

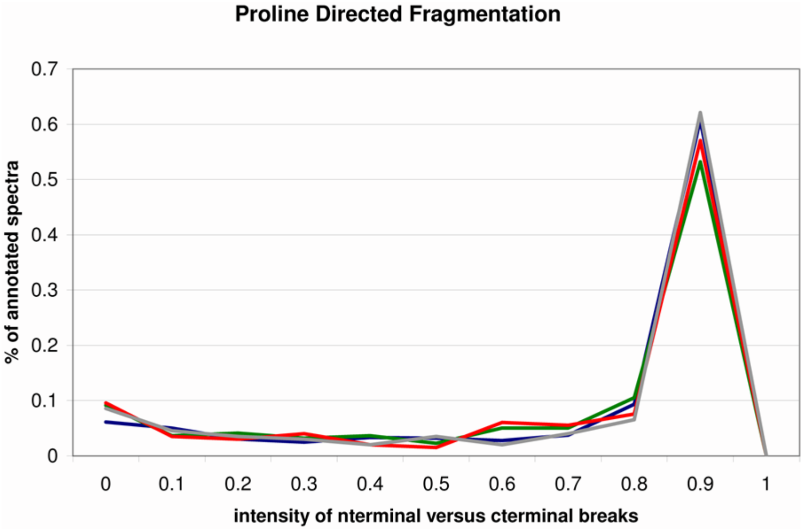


Figure 2.7 - Proline directed fragmentation. Here, we compare the intensity of peaks produced by breaks either nterminal or cterminal to proline. As in Figure 6, the blue line is consensus spectra; green line is unique Inspect identifications; red line is unique SEQUEST identifications; gray line is unique X!Tandem identifications. The plotted function is X) N/(N + C) where N and C represent the intensity of the N-terminal and C-terminal breaks, respectively. It is expected that N-terminal ions are much more intense than C-terminal ions, due to proline directed fragmentation. As seen, all programs have identical profiles, having a large majority of breaks where the N-terminal ion is an order of magnitude stronger than the C-terminal ion.

Table 2.1 - Ion Set. Ions included in the Bayesian Network.

C-terminal ions	y	y + 1	y + 2	y - H2O	y - NH3	y - H3PO4
N-terminal ions	b	b + 1	b - H2O	b - NH3	b - H3PO4	b - H2O - H3PO4
fragments with a +2 charge	y2+	y2+ + 1	y2+ - H2O	y2+ - NH3	y2+ - H3PO4	
	b2+	b2+ + 1	b2+ - H2O	b2+ - NH3	b2+ - H3PO4	

Table 2.2 - Conditional Probability Table. Using the observed intensity for Ij and Ip(j), we look up the learned conditional probability and score the peak assignment Ij accordingly. This table shows two conditional probabilities: the third column for b - H3PO4 given b, the fourth column for y - H3PO4 given y. Notice the distinct propensities for fragment neutral loss of a b ion compared to a y ion. For example, a strong b peak produces a medium/strong b - H3PO4 43% of the time. Remembering that on average only 50% of b ions contain the phosphate moiety, almost all phosphorylated b peaks are accompanied by a neutral phosphate loss. In contrast, a strong y peak produces a medium/strong y - H3PO4 only 25% of the time, or roughly half of the fragments containing a phosphate.

		P_{CID} (I_j | $I_{\prod(j)}$, S, P_j)	
$I_{\prod(j)}$ Intensity	I_j Intensity	I_j = b-H_3PO_4 $I_{\prod(j)}$ =b	I_j = y-H_3PO_4 $I_{\prod(j)}$ =y
strong	strong	22.50%	9.50%
strong	medium	20.40%	15.20%
strong	weak	3.80%	4.80%
strong	absent	53.30%	70.50%
medium	strong	8.10%	1.70%
medium	medium	25.10%	11.40%
medium	weak	10.80%	8.60%
medium	absent	55.90%	78.40%
weak	strong	3.10%	0.40%
weak	medium	12.90%	4.00%
weak	weak	14.50%	9.40%
weak	absent	69.50%	86.30%
absent	strong	5.40%	1.50%
absent	medium	15.90%	4.10%
absent	weak	9.30%	4.40%
absent	absent	69.40%	90.00%

Table 2.3 - Performance of Parent Mass Correction. Data represent the percent of spectra that are correct to a given accuracy. The general model is the default Inspect model and was trained on unmodified peptides. The phosphorylation specific model was trained on phosphopeptides and includes phosphorylation-specific features, see Materials and Methods.

accuracy	no correction	general model	phosphorylation specific model
Charge 2			
0.1 Da	4.3	10.8	22.1
0.3 Da	15.6	47.8	76.4
0.5 Da	29.6	83.6	93.8
Charge 3			
0.1 Da	0.5	9.7	26.8
0.3 Da	2	44.4	70.7
0.5 Da	4.9	77.9	90.2

Table 2.4 Comparison of Inspect and SEQUEST. The test data set of 6410 MS/MS spectra from S. cerevisiae was run permitting up to two phosphorylations per peptide. All results were filtered to 1% false positives and then compared between programs. Shown are only the spectra with a phosphorylated peptide annotation. The rows for peptides are counts of nonredundant phosphopeptide species.

	Inspect	SEQUEST
Total Spectra	1089	962
2+ Spectra	700	645
3+ Spectra	389	317
2+ Peptides	619	589
3+ Peptides	333	287
Run Time	0.5 h	36 h

Chapter 3: Proteogenomic discovery, correction and confirmation of Arabidopsis gene models.

Gene annotation underpins genome science but gene model error rates are suspected to be high. Most often protein sequence is obtained from the genome using transcript evidence or computational predictions. However, to directly identify coding DNA, we obtained 144,079 distinct peptide amino acid sequences by tandem mass spectrometry from *Arabidopsis thaliana*. The peptides were derived from three different interpretations of the genome: a six-frame translation, an exon splice-graph, and the currently annotated proteome. 18,024 peptides were found that do not correspond to the currently annotated proteome. By integrating these peptides into the gene finder AUGUSTUS we predict 778 new genes and to refine or expand the annotation of 695 current gene models.

3.1 Introduction

A complete protein-coding catalog is a fundamental goal of genome projects. Much of modern biological research, from micro-array chips to protein family and functional classification, depends on a complete and accurate proteome. Extensive proteomic catalogs have been developed through the integration of gene prediction algorithms, cDNA and EST sequences, and comparative genomics (Kellis 2003, Lin 2007). As emerging research is incorporated into annotation pipelines and manual curation efforts, gene models continue to improve and expand. High throughput gene annotation pipelines utilize a variety of information sources, and benefit most significantly when new data contains orthogonal information to what is currently available (Brent 2008).

Recent advances in both experimental and computational peptide mass spectrometry have enabled the production of large proteomics datasets with broad coverage of the proteome (Baerenfaller 2008, Brunner 2007, Tanner 2007). Proteogenomics, using proteomic information to reannotate the genome, supplements nucleotide-based annotation in that it directly determines reading frame, translation start and stop sites, exact splice boundaries, and the validity of short open reading frames. By complementing DNA-based annotation with proteogenomics, a more complete and accurate protein-coding catalog can be obtained (Tanner 2007, Gupta 2007, Savidor 2006, Fermin 2006, Desiere 2005). In spite of its potential for improving gene annotation, proteogenomic analysis has not integrated into large-scale genome annotation pipelines or public protein sequence repositories.

Mass spectrometry is an ideal tool for the high-throughput identification of proteins in a sample. However, there are some considerations which must be addressed in a genome-wide proteogenomic effort. First, tandem mass spectrometry samples are biased towards the more abundant proteins in the cell. To more deeply sample the proteome, a diversity of samples must be assayed (see Methods). Secondly, as peptides are sampled in line with chromatography, the depth of peptide sampling relies heavily on the ability of chromatography to evenly space peptide elution. Finally, proteins in the cell may be post-translationally modified (e.g. phosphorylated), which can be readily identified by MS/MS software. Moreover, the identification of post-translationally modified proteins is essential for understanding the dynamic proteome.

In this project, we demonstrate the benefit of proteogenomics in discovering and validating protein coding genes of *Arabidopsis thaliana*. Our goal is both to achieve broad coverage of the proteome, but more importantly to discover new protein coding sequences within the arabidopsis genome. We searched 21 million tandem mass spectra against two greatly expanded databases: a six frame translation of the genome, and a splice-exon graph that compactly encodes putative splicing events (Tanner 2007). We introduce a new method for false-discovery rate (FDR) estimation for novel events and novel genes. From 2.7 million total spectrum annotations (144,079 peptides), we identify 18,024 peptides not in the current genome annotation (4018 spliced peptides). These peptides are classified according to their genomic location, either internal to an annotated gene or in the intergenic region between protein-coding genes. By integrating peptide sequences with other genome annotation hints, we produce 778 new genes (930 transcripts) and refine 695 current gene models (964 transcripts).

A recent publication by Baerenfaller et al., (Baerenfaller 2008) attempts a similar proteogenomic annotation of Arabidopsis. In both our and their study, a significant portion of the proteome was validated. However, we annotate dramatically more novel protein coding sequences. Our ability to annotate more spectra is in part due to the computational advancements of the database search tool, Inspect. Inspect's Bayesian scoring function is more sensitive than that of SEQUEST, annotating more spectra at a given false-discovery rate. This allows us to detect peptides which span splice boundaries as well efficiently determine presence and location of post translational modifications. Further, our experimental techniques enabled a diverse sampling of the proteome, beyond plant organs, to enrich for phosphopeptides. Our analysis extends beyond the Baerenfaller study to integrate our peptide information into an automated gene prediction pipeline.

3.2 Results

Novel Genes: To achieve broad proteomic coverage, we acquired 21 million spectra from four Arabidopsis organs and a variety of phosphopeptide enrichment techniques. Spectra were searched with Inspect against three databases and filtered to a 1% cumulative false-discovery rate at the spectrum level (Figure 3.1). From the expanded databases (see Methods), we annotate 18,024 peptides not in the current genome annotation, termed "novel peptides”. 16,348 peptides mapped to a single locus in the genome including 4018 peptides (22%) that spanned novel splice junctions. Both the six frame translation and the spliced-exon database contributed equally to the novel peptides and largely did not overlap with each other, with only 5% coming from both. This indicates that both databases are valuable for proteogenomic studies because they provide different possibilities for discovering novelty. Using the common protein reporting standard of two peptides per protein, we focused on 1,765 novel peptide clusters comprising 4,575 novel peptides (see Methods). We classify novel peptide clusters according to their position relative to annotated protein coding models. We define *intragenic* clusters as those falling within the boundaries of a protein coding gene and *intergenic* clusters as those falling in the intergenic space.

Using our novel intergenic peptides, we predict 778 novel genes (with 930 transcripts) with the gene finder AUGUSTUS (Stanke 2006). Evidence from peptides as well as from EST alignments, and genomic conservation with rice, poplar, medicago, were given as 'hints' to AUGUSTUS, which tries to find gene models that are in agreement with the hints and that have high likelihood in an *ab initio* probabilistic gene structure model. Resulting gene models included alternative splice variants, if suggested by the evidence. Of the 778 novel genes, 55 have both EST and homology support in addition to our peptides. 455 genes have support by the peptides and ESTs; 70 genes are supported by the peptides and conservation only. The remaining 198 genes have no other support than the peptides. As an independent validation of our discoveries, 52 of the 778 loci have now been incorporated in the newest Arabidopsis genome release (TAIR8).

To discover homology with the novel genes, we excised the surrounding nucleotide sequence and searched against the non-redundant database of proteins (NCBI nr version 03/26/08). For 539 of the loci, the underlying sequence revealed a close homolog (e-value < 1e-10), providing additional validation, and functional assignments for the new genes. Although many of the novel genes we discover are homologous to genes of unknown function, we highlight a novel gene involved in photosynthesis, Figure 3.2. Our predicted protein, supported by 13 novel and uniquely located peptides, aligns with proteins targeted to the chloroplast thylakoid lumen (e-value 1e-75). It also contains the PsbP pfam domain characteristic of photosystem II. A second novel locus, containing 4 uniquely located peptides on chromosome 4 shows strong similarity (e-value 1e-85) with a heat-shock protein (AT4G12770) in Arabidopsis.

We also note several interesting structural features of the intergenic clusters. First, a significant fraction (64%) of intergenic clusters overlap annotated pseudogenes or transposons (Figure 3.3). An example of a translated pseudogene is at locus AT2G15040, ATRLP18: Receptor like protein 18, which has high homology to disease resistance proteins in both Arabidopsis and other plants. We identify 5 peptides, 3 of which are uniquely located at this locus, confirming translation. While most pseudogenes are believed to be non-translated artifacts, transposons (which like pseudogenes are not typically included in the proteome) contain active protein-coding genes. We find evidence of translated proteins in transposons which are unrelated to transposon activity. For example, we identify 3 peptides within the locus AT4G07947. Although annotated as a pseudogene in TAIR7 it has been reclassified as a transposable element gene in TAIR8. The genomic region containing these peptides has high similarity to the ubiquitin-like protease (Ulp1) family in Arabidopsis, suggesting this may be a gene traveling as 'cargo' with the transposable element (Jiang 2004).

Refining gene models: In addition to the novel genes, we discovered peptide clusters overlapping annotated gene models, suggesting refinement of the existing annotation, e.g. new exon, exon boundary change, exon skipping or modified translation boundaries. Using AUGUSTUS to refine existing models with the new peptide evidence, we predict 964 new or altered transcripts in 695 genes. The refinement events can be classified according to their type, location, and the transcript being modified. A majority (521) of the events are novel exons, of which 314 are located within introns and 207 are in untranslated regions (UTR) of TAIR7 gene models. Exon boundary changes were also prevalent, with typical instances including 5' extension of the first exon and alternative donor/acceptor splice sites. We find evidence for 180 instances of exon extension, and 191 instances of exon shortening. In five transcripts, peptide evidence supports an exon skipping event. Some intragenic loci indicate gene extension beyond the borders of the annotated gene model. 323 of these gene extension events were also discovered.

To provide additional support to our gene refinement events, we again excised the nucleotide sequence surrounding a novel cluster and searched against the non-redundant database of proteins (NCBI nr version 03/26/08). For 348 loci, the underlying sequence containing the novel events revealed a close homolog (e-value < 1e-10). Several genes that have been extensively studied are included among the refined gene models. For example, we found an additional 200 amino acid exon in the 5’ UTR of MAPK phosphatase (AT3G55270; Figure 3.4). Also, we identified eight peptides corresponding to four missing or mispredicted exons at locus AT1G79920 (heat shock protein 70). The new sequence completes the canonical HSP70 pFam domain. A final example is the gene PMI1 (AT1G42550) which, when mutated, results in impaired plastid movement and localization (DeBlasio 2005). We find 6 peptides upstream of the annotated start codon, providing at least 130 amino acids of new sequence.

We identified 70 cases in which the reading frame of the annotation is different from the observed peptides, Table 3.1. Assignment of reading frame is particularly difficult for nucleotide-based genome annotation (e.g. cDNA). However, proteomic evidence unambiguously defines the frame of translation. The 70 frame correction events are supported by multiple peptides and extensive homology to other proteins (see Methods)**.** We present here two examples: first a whole gene frame correction, and second a partial gene correction. Locus AT3G22240 is a short 51 amino acid protein with no discernable homology. Four peptides suggest translation in another frame. Translation in the new reading frame yields a protein high sequence identity to PCC1, pathogen and clock controlled protein. The second example is AT1G63500, a protein kinase, which has four novel peptides in the annotated 5' UTR. These peptides point to a large expansion of the gene and a misprediction of the current first exon (Figure 3.5).

Validation: In addition to gene discovery, we identified 126,055 distinct peptides (1.72 million amino acids) and confirmed gene models for 12,702 proteins (40% of the TAIR7 genes), Figure 3.6. Our claims of coverage are conservative. We count only proteins covered by at least two peptides, one of which must uniquely map to the designated locus. Of the sequenced peptides, 87% map to a unique genomic location, unambiguously identifying 10,690 proteins. In addition, we observed proteins from highly homologous gene groups that could not be attributed to a single locus (see Methods). The arabidopsis genome has high rates of tandem and segmental duplication and many loci contain multiple gene predictions that differ only in the non-translated regions (Cannon 2004). We observed peptides from 883 groups of indistinguishable proteins (2,012 proteins), bringing the total confirmed gene models to 12,702.

3.3 Conclusions

Historically, the proteomic and genomic communities have operated independently, with the genomic community in charge of annotation efforts. The predicted proteome is then passed over to the proteomics community for validation, and identification of post-translational events. We assert that much is to be gained by joining forces, and incorporating proteomic evidence upfront into the genomics pipelines. Proteogenomics provides an orthogonal data source to predict gene models, with levels of sensitivity that are complementary to cDNA sequencing. By investing in proteogenomics to complement more traditional cDNA and EST data at the onset of genome annotation, a more complete and accurate proteome can be achieved even in the early releases. Here we provide proteomic evidence for 778 new genes and refine 695 current gene models, using the reference annotation from TAIR7. Recently, TAIR has release the next revision of the genome/proteome, TAIR8. Only a small number of our novel peptides (3%) appear in the TAIR8 release suggesting that the proteogenomic approach still has much to offer the arabidopsis community.

3.4 Materials and Methods

Sample Preparation: In total 21,170,989 MS/MS spectra were collected from 45 LC-MS/MS experiments. For Arabidopsis organ samples (leaf, root, flower, and silique), ~ 2g of fresh organs were cut form wild type Arabidopsis (Col-0) using a sharp razor blade and transferred into a 50ml conical tube filled with liquid nitrogen immediately. Frozen organs were ground in a ceramic mortar and pestle with liquid nitrogen for 15 minutes to fine powders, and then transferred to a 50ml conical tube. 50ml cold (-20 ^{o}C) methanol containing 0.2 mM Na_3VO_4 was added to the conical tube. Samples were incubated at -20 ^{o}C for 15 minutes and then spun down in a refrigerated centrifuge at 4,000g for 5 minutes. Supernatant was discarded. Two more methanol washes were performed and followed by three acetone washes using the same procedure. After final acetone wash, sample pellets were dried in an Eppendorf Vacufuge Concentrator at 4 ^{o}C. Proteins were extracted by adding 1ml of 0.2% RapiGest (Waters) with 0.2 mM Na_3VO_4 to the dry pellet and incubated on ice for 15 minutes. Samples were spun down at 16,000g in a refrigerated centrifuge for 15 minutes. Pellets were discarded and the supernatants were ready for protein digestion.

For MM2d cells, cell pellets (~ 100uL pellet volume) were washed by 1ml HEPES saline buffer (10mM HEPES, 150 mM NaCl) three times. 250 ul of protein extraction buffer (2% RapiGest from Waters plus 0.2 mM Na_3VO_4) was added to the cell pellet. Samples were sonicated in a Branson Sonifier 450 sonicator equipped with a high intensity cup horn (Branson Part No.101-

147-046) at 40% output power for 2 minutes with circulating cooling water. Cell lysates were centrifuged at 16,100g, 4 ^{o}C, for 15 minutes. Pellets were discarded and the supernatants were diluted 10 times in 50 mM HEPES buffer (pH 7.2) and ready for protein digestion.

Cysteines were reduced and alkylated using 1 mM Tris(2-carboxyethyl)phosphine (Fisher, AC36383) at 95 °C for 5 minutes then 2.5 mM iodoacetamide (Fisher, AC12227) at 37 °C in dark for 15 minutes. Proteins were digested with trypsin (Roche, 03 708 969 001) overnight then 1% TFA (pH 1.4) was added to precipitate RapiGest. Samples were incubated at 4 ^{o}C overnight and then centrifuged at 16,100g for 15 minutes. Supernatant was collected and centrifuged through a 0.22 uM filter.

An Agilent 1100 HPLC system (Agilent Technologies, Wilmington, DE) delivered a flow rate of 300 nL min^{-1} to a 3-phase capillary chromatography column through a splitter. Using a custom pressure cell, 5 µm Zorbax SB-C18 (Agilent) was packed into fused silica capillary tubing (200 µm ID, 360 µm OD, 20 cm long) to form the first dimension reverse phase column (RP1). A 5 cm long strong cation exchange (SCX) column packed with 5 µm PolySulfoethyl (PolyLC) was connected to RP1 using a zero dead volume 1 µm filter (Upchurch, M548) attached to the exit of the RP1 column. A fused silica capillary (100 µm ID, 360 µm OD, 20 cm long) packed with 5 µm Zorbax SB-C18 (Agilent) was connected to SCX as the analytical column (RP2). The electro-spray tip of the fused silica tubing was pulled to a sharp tip with the inner diameter smaller than 1 µm using a laser puller (Sutter P-2000). The peptide mixtures were loaded onto the RP1 column using the custom pressure cell. Columns were not re-used. Peptides were first eluted from the RP1 column to the SCX column using a 0 to 80% acetonitrile gradient for 150 minutes. The peptides were fractionated by the SCX column using a series of salt gradients (from 10 mM to 1 M ammonium acetate for 20 minutes), followed by high resolution reverse phase separation using an acetonitrile gradient of 0 to 80% for 120 minutes.

Spectra were acquired on LTQ linear ion trap tandem mass spectrometers (Thermo Electron Corporation, San Jose, CA) employing automated, data-dependent acquisition. The mass spectrometer was operated in positive ion mode with a source temperature of 150 ^{o}C. As a final fractionation step, gas phase separation in the ion trap was employed to separate the peptides into 3 mass classes prior to scanning; the full MS scan range was divided into 3 smaller scan ranges (300-800, 800-1100, and 1100-2000 Da) to improve dynamic range. Each MS scan was followed by 4 MS/MS scans of the most intense ions from the parent MS scan. A dynamic exclusion of 1 minute was used to improve the duty cycle.

Final totals for spectrum count were: 6,336,450 spectra from roots, 1,415,293 spectra from *M. incognita* infected roots, 2,660,544 from leaves, 1,284,713 from flowers, 1,206,222 from siliques, and 8,267,767 from phospho-peptide enriched MM2D cell lysates. All data is uploaded to the Tranche repository (http://tranche.proteomecommons.org/); a hash key for download is available from the authors upon request.

Database Construction: Proper database construction is crucial for novel peptide recovery. For gene model confirmation, we use the TAIR7 release of the Arabidopsis proteome (www.arabidopsis.org). For proteomic discovery, we constructed two greatly expanded databases. The first database was the six frame translation of the Arabidopsis genome, containing 210 M amino acids. The second database was a spliced-exon graph containing *ab initio* gene predictions from the AUGUSTUS software (Tanner 2007, Stanke 2006, Stanke 2008). AUGUSTUS reported multiple transcripts per locus with sampling parameter = 100. Additionally, we edited the exon length distribution to make short exons (< 100 base pairs) 3x more likely. All resulting exon and intron predictions were incorporated into the graph where each node is a putative exon and each directed edge indicates a putative splice junction. The resulting graph contained about 16 M amino acids. For the MS/MS searches, all three databases were combined with decoy sequences formed by shuffling each target sequence. To ensure minimal overlap between target and decoy sequences, any 8mer appearing in the decoy sequences which also appears in the target database was re-shuffled.

Mass Spectrometry, Peptide Identification and Location. All spectra above were converted from the vendor formatted RAW files to mzXML using the ReAdW software in centroid mode (Nov1, 2006 version). Spectra were searched against the three databases with the Inspect software, release 2007.09.05. All datasets, excepting the phospho-peptide enriched samples, were searched without allowing any post-translational modifications (PTMs). Parameters for this search were: PM tolerance 3.0 Da, 0.5 Da fragment ion tolerance, 25 tags/spectrum, +57 Da fixed modification on cysteine. For phospho-enriched samples, we allowed a variable modification of +80 on STY, max of 2 PTMs/peptide and searched with Inspect's phosphopeptide specific scoring function (Payne 2008). All results are filtered to 1% spectrum-level false discovery rate using the decoy database strategy (Elias 2007). In this strategy, a scrambled database of the same size is concurrently searched with the target sequences against the spectra. A score cutoff is chosen such that no more than 1% of the spectra are annotated with a peptide from decoy sequences. To count proteins validated by our TAIR7 database search, we map peptides back to their protein(s). We report proteins with two or more peptides, and at least 1 uniquely mapped peptide. For proteins groups which have exactly identical coding sequences we report the group of proteins, as they share all peptides and do not have any uniquely mapped peptides. As we require multiple peptides per protein identification, our 1% spectrum-level FDR translated to an empirical 0.6% protein-level FDR. The source code for Inspect is available at our lab website, http://peptide.ucsd.edu/.

Clustering and Homology Search: Novel peptides (all of which have a genomic location) were clustered. Peptides within 1000 nucleotides were linked; clusters were aggregated by single linkage. We find that the vast majority of peptides within current genes fit this clustering (98%). Any fixed width cluster has the potential to misgroup peptides from multiple genes into a single group. This is overcome by the gene finding algorithm which creates the best gene model given the evidences, including splitting clusters. Clusters were classified as intragenic or intergenic depending on whether they overlapped a TAIR7 protein coding gene model. We extracted the DNA sequence of each cluster with 500 nucleotides abutting the first and last amino acid of the predicted peptides and searched versus the NCBI non-redundant protein database (NCBI nr) using blast with default parameters. Supplemental table 1 lists the top five results of each search.

Frame Correction: We found evidence of many novel peptides out of frame with the current gene models. From these we picked a subset to highlight and present in Supplemental Table 2. From the list of all novel peptides which overlap a known gene locus, we generated a list of peptides which overlap the coding region of the locus but in a different frame. Our reported results require at least two out-of-frame novel peptides. To increase our confidence in the assertion that the gene is (at least partially) mis-predicted, we also tabulated several features for these novel peptides. As splicing sometimes causes that only a portion of the peptide is out of frame with the reference annotation, we filtered out peptides which had fewer than three amino acids out of frame. In these cases we do not doubt the accuracy of the MS/MS annotation. However, with only one or two amino acid(s), there are likely several close genomic regions with an appropriate nucleotide sequence. Additionally, we determined whether the novel peptides conflict with observed MS/MS peptides which support the current gene model and frame. On a few occasions there was peptide and homology support for the both the annotated frame and a new frame, possibly suggesting alternative splicing. (There are instances within the current annotation of the same DNA sequence being translated in multiple frames.) 70 proteins had novel peptides which met these three requirements: multiple peptides out of frame, sequence out of frame is at least 3 amino acids, and no conflicting TAIR peptides. There were also instances of novel peptides present in the 5' and 3' untranslated regions of genes. In some instances these are likely to merely be expansion of the current sequence. However, some of these also appear to be frame mis-predictions.

False Discovery Rate Calculations: **Local False Discovery Rate (lFDR):** The most commonly reported statistic for false-discovery rate is the cumulative false-discovery rate, cFDR, or the fraction of false-positive spectra with a score greater than *t*. This number is meant as an estimate of error for a data set and is often misinterpreted as a confidence in a single spectrum annotation. For example, consider a data set with 1000 spectra annotated at a 1% cFDR at score t_0. At this cutoff, 10 of the 1000 spectra are estimated to be false-positives. As the score cutoff is relaxed and more spectra are accumulated, we set the next cutoff t_1 for a 5% cFDR and note 1200 spectra are annotated. 60 of the 1200 are estimated false positives. cFDR assignments to these new spectra would be between 1% and 5%. However, this is not accurate. The entire data set has a 5% false discovery rate, but for the 200 newly included spectra, the false discovery rate is much higher. Of the 60 total false-positive spectra, 50 came from these 200 new annotations, or 25% false discovery. Thus a cumulative FDR calculation fails to deliver an accurate quality metric. Unfortunately, this point has not been previously addressed in proteomics studies. When considering that we annotate over 2.7 million spectra at a cFDR of 1%, a more lax 5% cFDR could have included significant number of spectra which in reality had an unacceptable false-discovery rate.

To more accurately measure the quality of our assignments, we define a local false discovery rate (Efron 2001). For score t, and bin-size δ, define *local-FDR* (lFDR_δ(t)), as the fraction of incorrect identifications with score in [t,t+δ).

$$lFDR_\delta(t) = f_0(t) / f_0(t) + f_1(t)$$

Where $f_0(t)$ is the number of false annotations with score in [t,t+δ) and $f_1(t)$ is the number of true annotations with score in [t,t+δ). While local FDR is a continuous function, we empirically measure it over a discretized range. Unlike microarray experiments where the number of false data points must be estimated, by using the decoy database search strategy, we can directly count this value; f_0 is simply the distribution of matches to the decoy database and f_1 is the distribution of matches to the true database. We compute a local FDR for each spectrum using δ=0.1. Our dataset is large, therefore, a significant number of spectrum-peptide matches fall in each bin and we can achieve a more accurate local FDR. For higher score regions with fewer spectrum-peptide matches, bins were expanded to include at least 1000 annotations. As spectra of different charge states have distinct score distributions (data not shown), the FDR should be separately calculated. Inspect identifies spectra of charge <= 3, and we compute lFDR separately for charge 3 spectra. A change in FDR for peptides of different lengths has also been reported. As Inspect explicitly takes peptide-length into account while scoring, this bias is not observed in our identifications (data not shown). The minimum peptide length for Inspect is 7 amino acids; 0.8% of all reported spectral identifications are of length 7.

Spectral FDR versus Peptide FDR (pFDR): The redundancy introduced by repeated observation (multiple spectra) of peptides changes the false discovery rate for peptides. If 100 spectra identified the same peptide, the peptide identification is incorrect only if all spectral identifications are incorrect. At the same time, spectra identifying the same peptide cannot be treated as independent. A systematic error might lead to similar spectra to all be mis-annotated. Therefore, we conservatively assign the FDR for a peptide to be the minimum local FDR of all spectra identifying that peptide.

Event level FDR (eFDR): The identification of distinct peptides can be reasonably assumed to be independent. Even peptides that overlap in sequence have completely different spectra, as prefix/suffix masses are all changed by the distinct terminal residues. In identifying an *event*, (e.g. a novel exon), we estimate the FDR of the event as the product of the local FDR of the peptides supporting that event, and use an eFDR cut-off of 5%.

Acknowledgments: We would like to acknowledge Anand Patel for his efforts in homology searches and alignment visualization. SHP was supported by National Science Foundation IGERT Plant Systems Biology training grant DGE-0504645. ZS and SPB were supported by National Science Foundation Grant IBN 0619411. VB and NC were supported in part by the grant NIH R01-RR16522. MS was supported by DFG grant STA 1009/4-1. This chapter is in preparation for publication as “Proteogenomic Discovery, Correction and Confirmation of Arabidopsis Gene Models.” Castellana NC, Payne SH, Shen Z, Stanke M, Bafna V, and Briggs SP. 2008, submitted. The dissertation author, Natalie Castellana and Zhouxin Shen were the primary authors of this paper.

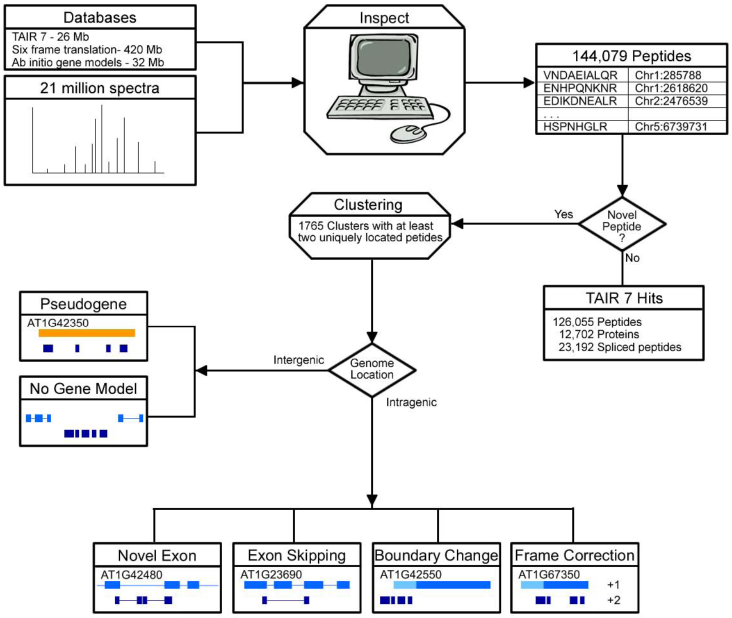
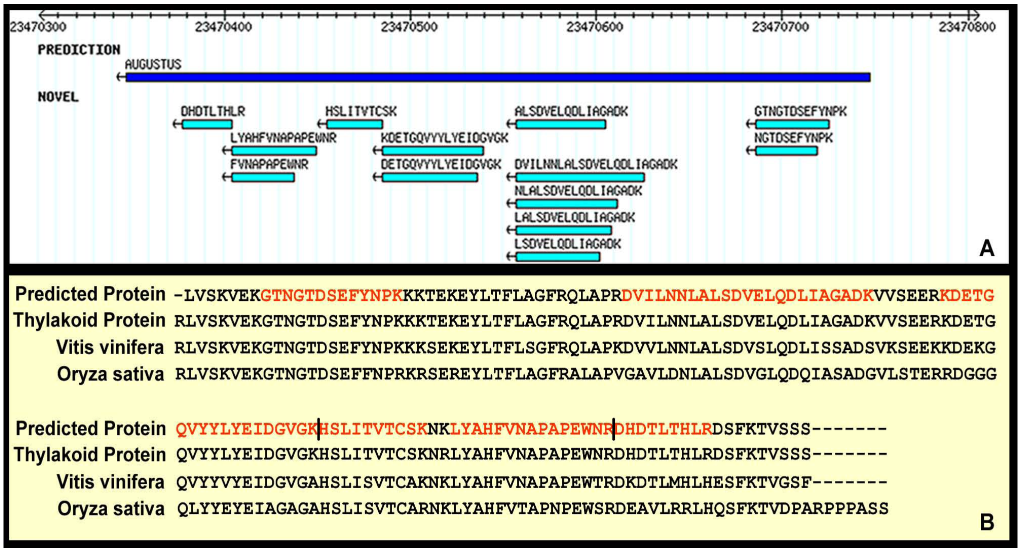


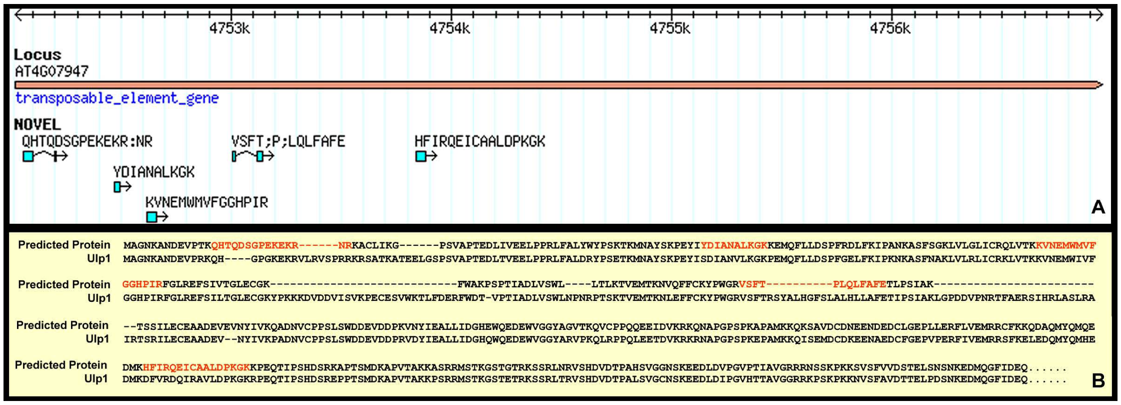
Figure 3.1 - Workflow. All mass spectra are compared to three databases using Inspect. Spectra are filtered to a 1% false discovery rate and grouped into peptides. Novel peptides are separated from those that appear in TAIR7. Novel peptides are then segregated based on genome location. Those that overlap a current gene model (intragenic) are further classified by how they refine the model. Peptides that do not overlap a gene model (intergenic) are classified by whether they overlap a pseudogene.



```
Predicted Protein -LVSKVEKGTNGTDSEFYNPKKKTEKEYLTFLAGFRQLAPRDVILNNLALSDVELQDLIAGADKVVSEERKDETG
Thylakoid Protein RLVSKVEKGTNGTDSEFYNPKKKTEKEYLTFLAGFRQLAPRDVILNNLALSDVELQDLIAGADKVVSEERKDETG
   Vitis vinifera RLVSKVEKGTNGTDSEFYNPKKKSEKEYLTFLSGFRQLAPKDVVLNNLALSDVSLQDLISSADSVKSEEKKDEKG
    Oryza sativa RLVSKVEKGTNGTDSEFFNPRKRSEREYLTFLAGFRALAPVGAVLDNLALSDVGLQDQIASADGVLSTERRDGGG

Predicted Protein QVYYLYEIDGVGK|HSLITVTCSKNKLYAHFVNAPAPEWNR|DHDTLTHLRDSFKTVSSS-------
Thylakoid Protein QVYYLYEIDGVGKHSLITVTCSKNRLYAHFVNAPAPEWNRDHDTLTHLRDSFKTVSSS-------
   Vitis vinifera QVYYVYEIDGVGAHSLISVTCAKNKLYAHFVNAPAPEWTRDKDTLMHLHESFKTVGSF-------
    Oryza sativa QLYYEYEIAGAGAHSLISVTCARNKLYAHFVTAPNPEWSRDEAVLRRLHQSFKTVDPARPPPASS
```

Figure 3.2 - Novel gene supported discovered by proteogenomics. (A) A cluster of 13 uniquely located peptides which do not overlap a current gene model (Chr3). The prediction track shows the single exon gene model produced by AUGUSTUS. (B) The predicted sequence shows strong homology to a Thylakoid lumen family protein (sp|P82658|TL19_ARATH). It also shows strong similarity to proteins in both grapevine (emb|CA040861.1) and rice (Os08g0504500).



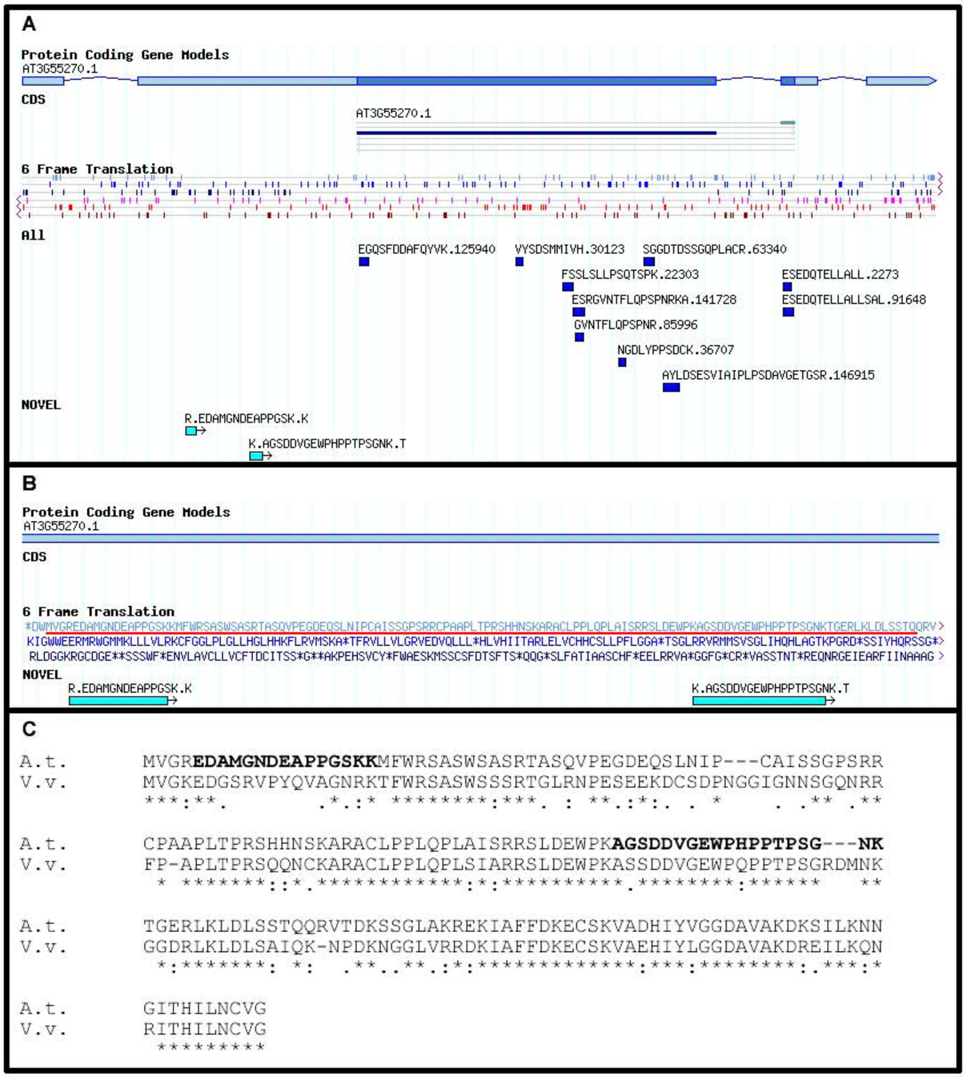
```
Predicted Protein  MAGNKANDEVPTKQHTQDSGPEKEKR------NRKACLIKG------PSVAPTEDLIVEELPPRLFALYWYPSKTKMNAYSKPEYIYDIANALKGKKEMQFLLDSPFRDLFKIPANKASFSGKLVLGLICRQLVTKKVNEMWMVF
             Ulp1  MAGNKANDEVPRKQH----GPGKEKRVLRVSPRRKRSATKATEELGSPSVAPTEDLTVEELPPRLFALDRYPSETKMNAYSKPEYISDIANVLKGKPEMQFLLDSPFGELFKIPKNKASFNAKLVLRLICRKLVTKKVNEMWIVF

Predicted Protein  GGHPIRFGLREFSIVTGLECGK----------------------------FWAKPSPTIADLVSWL----LTLKTVEMTKNVQFFCKYPWGRVSFT----------PLQLFAFETLPSIAK------------------------
             Ulp1  GGHPIRFGLREFSILTGLECGKYPKKKDVDDVISVKPECESVWKTLFDERFWDT-VPTIADLVSWLNPNRPTSKTVEMTKNLEFFCKYPWGRVSFTRSYALHGFSLALHLLAFETIPSIAKLGPDDVPNRTFAERSIHRLASLRA

Predicted Protein  --TSSILECEAADEVEVNYIVKQADNVCPPSLSWDDEVDDPKVNYIEALLIDGHEWQEDEWVGGYAGVTKQVCPPQQEEIDVKRKQNAPGPSPKAPAMKKQKSAVDCDNEENDEDCLGEPLLERFLVEMRRCFKKQDAQMYQMQE
             Ulp1  IRTSRILECEAADEV--NYIVKPADNVCPPSLSWDDEVDDPRVDYIEALLIDGHQWQEDEWVGGYARVPKQLRPPQLEETDVKRKRNAPGPSPKEPAMKKQISEMDCDKEENAEDCFGEPVPERFIVEMRRSFKELEDQMYQMHE

Predicted Protein  DMKHFIRQEICAALDPKGKKPEQTIPSHDSRKAPTSMDKAPVTAKKASRRMSTKGSTGTRKSSRLNRVSHDVDTPAHSVGGNSKEEDLDVPGVPTIAVGRRRNSSKPKKSVSFVVDSTELSNSNKEDMQGFIDEQ......
             Ulp1  DMKDFVRDQIRAVLDPKGKRPEQTIPSHDSREPPTSMDKAPVTAKKPSRRMSTKGSTETRKSSRLTRVSHDVDTPALSVGCNSKEEDLDIPGVHTTAVGGRRKPSKPKKNVSFAVDTTELPDSNKEDMQGFIDEQ......
```

Figure 3.3 – Peptides overlapping a predicted transposable element gene. (A) 5 peptides, 4 unique, overlap locus AT4G07947 which is annotated as a transposable element gene. (B) Sequence alignment of an Arbaidopsis Ulp1 (ubiquitin like protease) showing strong conservation.



```
*DWMVGREDAMGNDEAPPGSKKMFWRSASWSASRTASQVPEGDEQSLNIPCAISSGPSRRCPAAPLTPRSHHNSKARACLPPLQPLAISRRSLDEWPKAGSDDVGEWPHPPTPSGNKTGERLKLDLSSTQQRV>
 KIGWWEERMRWGMMKLLLVLRKCFGGLPLGLLHGLHHKFLRVMSKA*TFRVLLVLGRVEDVQLLL*HLVHIITARLELVCHHCSLLPFLGGA*TSGLRRVRMMSVSGLIHQHLAGTKPGRD*SSIYHQRSSG*>
  RLDGGKRGCDGE**SSSWF*ENVLAVCLLVCFTDCITSS*G**AKPEHSVCY*FWAESKMSSCSFDTSFTS*QQG*SLFATIAASCHF*EELRRVA*GGFG*CR*VASSTNT*REQNRGEIEARFIINAAAG >
```

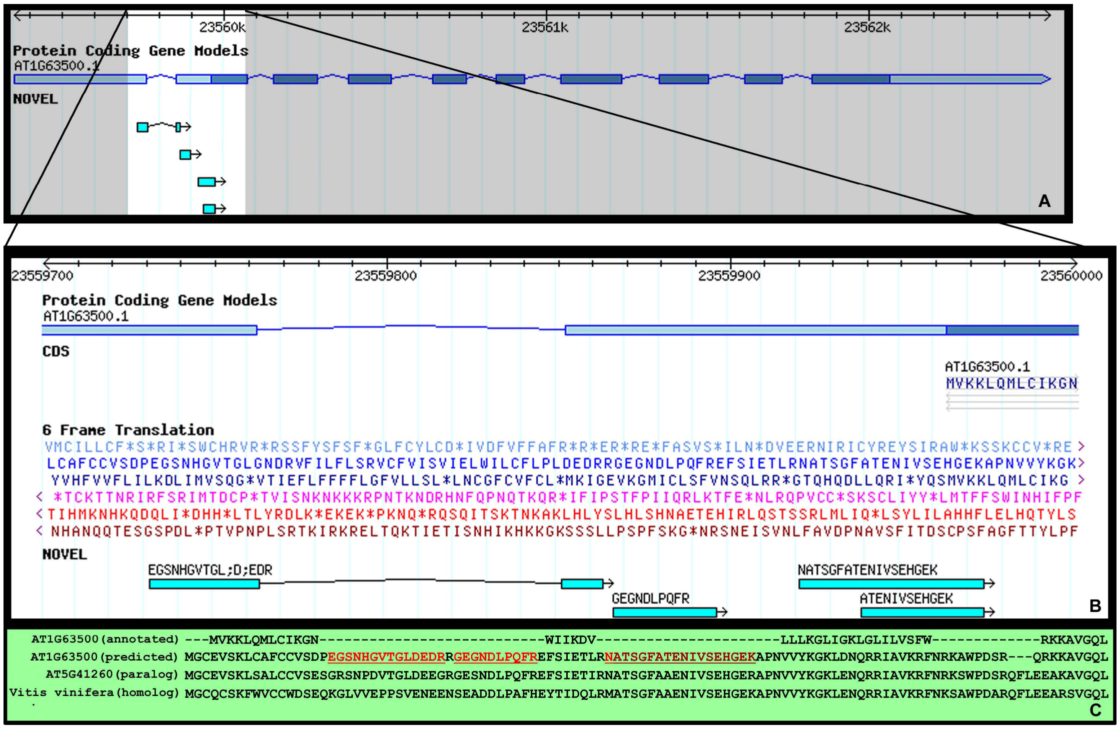
```
A.t.      MVGREDAMGNDEAPPGSKKMFWRSASWSASRTASQVPEGDEQSLNIP---CAISSGPSRR
V.v.      MVGKEDGSRVPYQVAGNRKTFWRSASWSSSRTGLRNPESEEKDCSDPNGGIGNNSGQNRR
          ***:**.       .*.:* ********:***. : **.:*:. . *    . .** .**

A.t.      CPAAPLTPRSHHNSKARACLPPLQPLAISRRSLDEWPKAGSDDVGEWPHPPTPSG---NK
V.v.      FP-APLTPRSQQNCKARACLPPLQPLSIARRSLDEWPKASSDDVGEWPQPPTPSGRDMNK
           * *******::*.************:*:**********.********:******   **

A.t.      TGERLKLDLSSTQQRVTDKSSGLAKREKIAFFDKECSKVADHIYVGGDAVAKDKSILKNN
V.v.      GGDRLKLDLSAIQK-NPDKNGGLVRRDKIAFFDKECSKVAEHIYLGGDAVAKDREILKQN
           *:*******: *:  .**..**.:*:*************:***:********:.***:*

A.t.      GITHILNCVG
V.v.      RITHILNCVG
           *********
```

Figure 3.4 – Addition of a 5' exon to MAP Kinase Phosphatase, AT3G55270. A). The predicted protein coding sequence is well covered by peptides (All track), but we observe two additional peptides in the 5' UTR. B). The two novel peptides lie in a single ORF, frame 1. C). A portion of the translated UTR sequence is aligned to grapevine (V.v.). Peptides from proteogenomics are bolded.



```
VMCILLCF*S*RI*SWCHRVR*RSSFYSFSF*GLFCYLCD*IVDFVFFAFR*R*ER*RE*FASVS*ILN*DVEERNIRICYREYSIRAW*KSSKCCV*RE >
LCAFCCVSDPEGSNHGVTGLGNDRVFILFLSRVCFVISVIELWILCFLPLDEDRRGEGNDLPQFREFSIETLRNATSGFATENIVSEHGEKAPNVVYKGK>
YVHFVVFLILKDLIMVSQG*VTIEFLFFFFLGFVLLSL*LNCGFCVFCL*MKIGEVKGMICLSFVNSQLRR*GTQHQDLLQRI*YQSMVKKLQMLCIKG >
< *TCKTTNRIRFSRIMTDCP*TVISNKNKKKRPNTKNDRHNFQPNQTKQR*IFIPSTFPIIQRLKTFE*NLRQPVCC*SKSCLIYY*LMTFFSWINHIFPF
<TIHMKNHKQDQLI*DHH*LTLYRDLK*EKEK*PKNQ*RQSQITSKTNKAKLHLYSLHLSHNAETEHIRLQSTSSRLMLIQ*LSYLILAHHFLELHQTYLS
< NHANQQTESGSPDL*PTVPNPLSRTKIRKRELTQKTIETISNHIKHKKGKSSSLLPSPFSKG*NRSNEISVNLFAVDPNAVSFITDSCPSFAGFTTYLPF
```

```
   AT1G63500(annotated) ---MVKKLQMLCIKGN---------------------------WIIKDV----------------------LLLKGLIGKLGLILVSFW-------------RKKAVGQL
   AT1G63500(predicted) MGCEVSKLCAFCCVSDPEGSNHGVTGLDEDRRGEGNDLPQFREFSIETLRNATSGFATENIVSEHGEKAPNVVYKGKLDNQRRIAVKRFNRKAWPDSR---QRKKAVGQL
     AT5G41260(paralog) MGCEVSKLSALCCVSESGRSNPDVTGLDEEGRGESNDLPQFREFSIETIRNATSGFAAENIVSEHGERAPNVVYKGKLENQRRIAVKRFNRKSWPDSRQFLEEAKAVGQL
Vitis vinifera(homolog) MGCQCSKFWVCCWDSEQKGLVVEPPSVENEENSEADDLPAFHEYTIDQLRMATSGFAAENIVSEHGEKAPNVVYKGKLENQRRIAVKRFNKSAWPDARQFLEEARSVGQL
```

Figure 3.5 – Refined Gene Model. TAIR locus AT1G63500 encodes a protein kinase. (A) Four novel peptides map within the 5' UTR and the first exon. (B) Zoom of the region shows that the current first exon (frame 3) is out of frame with the peptides (frame 2). (C) Sequence alignment with Arabidopsis and grapevine proteins supports translation in the frame supported by peptides (observed peptides highlighted in alignment).

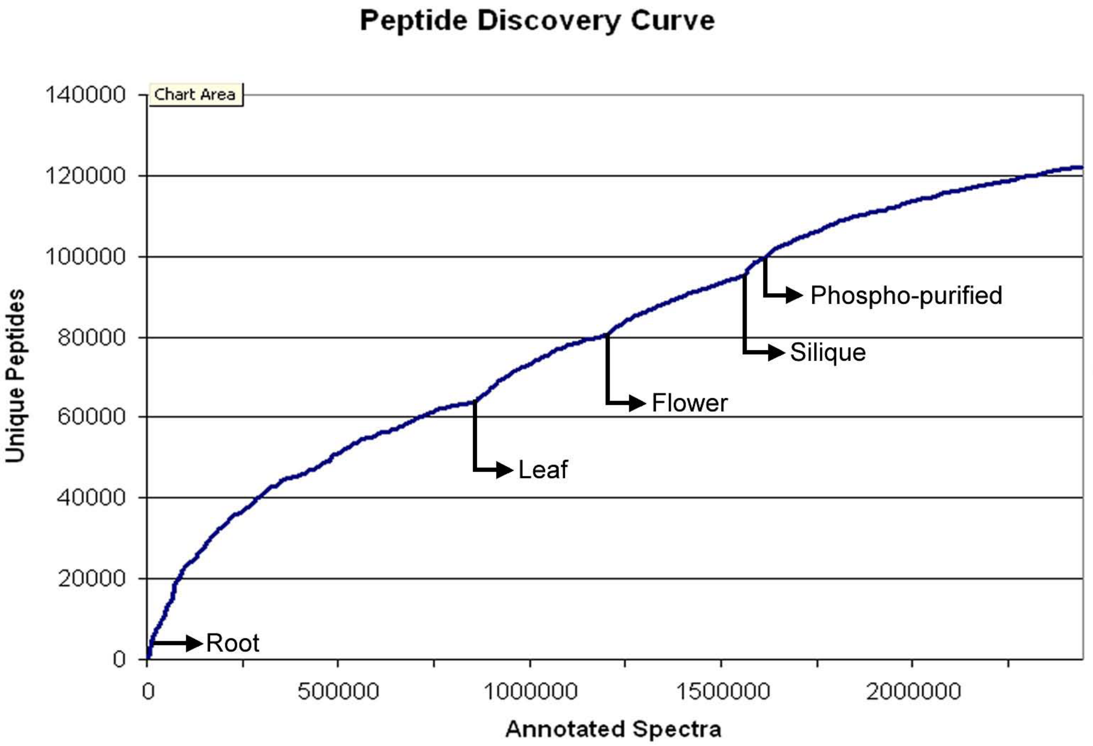


Figure 3.6 - Discovery curve showing the number of distinct peptides matching to TAIR7 recovered as a function of the number of annotated spectra. The discovery curve is separated to show the contribution of each individual dataset.

Table 3.1 - Frame Correction Loci. Each locus is proposed to be annotated (at least partially) out of frame. There are multiple lines of evidence. Shown are the number of proteomic peptides identified in a frame different from the TAIR annotation, whether each peptide has at least three amino acids out of frame, and whether there are proteomic peptides which confirm the TAIR annotation, and thus conflict with the proposed new frame annotatino.

Locus	Peps Out of frame	Min length met	Conflicts with TAIR peps
AT1G11905	3	Y	N
AT1G15120	3	Y	N
AT1G16840	2	Y	N
AT1G18580	2	y	n
AT1G19130	10	y	n
AT1G21695	2	Y	Y
AT1G29560	3	y	n
AT1G29600	2	Y	N
AT1G53705	3	y	n
AT1G56100	11	y	n
AT1G63500	2	y	n
AT1G63820	2	y	y
AT1G65090	2	y	n
AT1G67350	9	y	n
AT1G71360	2	Y	Y
AT1G71530	3	y	n
AT1G72840	2	y	n
AT1G76820	14	y	n
AT1G79920	7	y	n
AT2G02730	3	y	n
AT2G02810	2	y	n
AT2G03600	2	y	n
AT2G07280	2	y	n
AT2G09865	2	y	n
AT2G24680	2	y	n
AT2G30120	3	y	n
AT2G34200	2	y	n
AT3G06290	2	y	n
AT3G08770	3	y	y
AT3G09660	2	y	n
AT3G19290	2	y	n

AT3G22240	5	y	n

Table 3.1 (cont.)

Locus	Peps Out of frame	Min length met	Conflicts with TAIR peps
AT3G47836	3	y	n
AT3G55970	2	y	n
AT4G00340	0	y	n
AT4G01925	2	y	n
AT4G02260	2	y	n
AT4G02700	2	y	n
AT4G05631	2	y	n
AT4G14240	2	y	n
AT4G15020	0	y	n
AT4G15770	0	y	n
AT4G15930	2	y	y
AT4G16260	4	y	y
AT4G16380	0	y	n
AT4G17120	2	y	n
AT4G17245	3	y	n
AT4G19460	2	y	2:1 novel:TAIR
AT4G20310	2	y	n
AT4G22850	2	y	n
AT4G23760	3	y	n
AT4G36600	2	y	n
AT4G36780	2	y	n
AT4G36800	4	y	n
AT4G39420	2	y	y
AT5G02160	5	y	n
AT5G06350	2	y	n
AT5G07670	2	y	n
AT5G13010	2	y	n
AT5G13850	16	y	y
AT5G26680	3	y	n
AT5G32070	2	y	n
AT5G39570	5	y	n
AT5G43530	2	y	n
AT5G43580	2	y	n
AT5G44280	2	y	n
AT5G46540	2	y	n

AT5G58790	2	y	n
AT5G59980	3	y	n

Chapter 4: Usability and Software Development

A conflict of interest underlies the creation of academic software. The primary focus of graduate students and their professors is to publish. The most important qualities for publication are novelty of the algorithm or application. Improving the usability of a current idea is generally not considered publishable. Thus given a limited amount of time, researchers typically opt to invest in producing new ideas at the expense of refining or improving existing software. The command line often is the only interface; a tool unknown to many biologists.

Although this decision is best for the individual researchers, it may not be optimal for the scientific field as a whole. The result is "research grade" software, meaning that it works well for a limited set of inputs. Atypical or unanticipated inputs are generally poorly dealt with. Two classic examples of this in the proteomics community are spectrum file formats and ever changing quantitation protocols.

The effort required to parse a new spectrum format is not overly demanding. Depending on the complexity of the format and the flexibility of the software, this task could take up to two weeks of focused software engineering. It is a poor assumption that researchers are good software engineers; software is never as flexible as it could be. Regardless, most programs only allow a subset of formats.

The proteomics community is becoming more quantitative. More often researchers are looking to quantify the difference between samples, both in the relative and absolute. At this leading edge of research, there is no consensus on the best way to establish a quantitative experiment. There is a tremendous diversity of setups: labeled and label free, amino acid labeling (SILAC) or n/c terminus labeling (e.g. O16/O18 digestion), iTRAQ or other covalent reagents, using MS or MS/MS. From a computational perspective, there are too many possible set-ups to be reasonably encompassed into a single algorithm or software tool. The result is software targeted towards a few (or even just one) experimental protocol.

4.1 Helping the End User

The UCSD Computational Mass Spectrometry group, members of the Bafna and Pevzer lab, has created a large variety of tools and algorithms, each with a slightly different purpose. Inspect is a general purpose database search engine [Tanner 2005]; PepNovo is a *de novo* interpretation tool that does not require a database. SpectralNetworks utilizes the sequence redundancy to create a network of related spectra, which can then be interpreted; MSGenerationFunction finds the true pvalue of a spectrum by scoring all possible interpretations of the spectrum. Each has a separate interface design.

Even when tools are theoretically similar, their approach, design and code base are distinct. For example, Inspect and PepNovo are both tools for peptide identification. They take spectra as input and output a scored list of candidate peptides; they both score candidate peptides using a Bayesian network model of peptide fragmentation. For the end user, the most important difference however, is the interface. The input format (e.g. command line parameters) and output format are distinct. This specific example holds for each of the 5-10 software programs written in the lab. Each is independently implemented, shares little code with other projects, and has unique I/O styles. Thus even though the desired output is often similar, and end user must learn the nuances of each program.

This problem is especially significant for bioinformatics, a field attempting to bridge the gap between experimental and computational biologists. The end user is often a biologist with little computer experience, who wants to test the software. To emphasize the beginning level of the average user, I found that the task of locating a command prompt and navigating the directory structure of the user's own computer was difficult. In my personal experience, the lack of investment in a more familiar interface environment prohibits adoption of our software.

I developed a set of tutorials aimed to help the first time user learn how to use Inspect. The four tutorials step through both basic and advanced applications: basic set up and use, advanced input options, blind searches, and grid applications. The first tutorial, as is evident by its name, is the most basic. It walks users through installation of the software and their first run of Inspect. I found that a purely text based explanation of the steps was not enough. Highlighting the conceptual gaps, users often asked, “How do I open Inspect?” In response, I added a multitude of screen shots, appendices and a glossary. The screen shots were essential in helping novice users understand the concept of command line interfaces. The second tutorial, advanced Inspect use, introduced options and workflows that I routinely use: decoy database, post-translational modifications. From my perspective as a developer, these are not necessarily an advanced concept, but they were confusing when included in the first tutorial. The same is true for the third and fourth tutorial. As I tested these tutorials with the BENG 208 students and each new collaborator, I noticed that even with the tutorials and demos, most biology users were unsure of themselves and anxious as to whether they were running the program correctly. This uncertainty impacted their desire to adopt our tools as part of their standard platform, or experimental protocols. Even after labs had seen an improvement in their results versus their current software they were prone to stick with the status quo.

4.2 Center for Computational Mass Spectrometry

Increased spectral acquisition rate in modern MS/MS instrumentation, and the ensuing deluge of data, have created a severe bottleneck in proteomics workflows. Quite simply, instruments can create spectra faster than traditional algorithms can interpret them. With researchers producing more and larger datasets, this basic problem has become more significant. Even for Inspect, which is orders of magnitude faster than algorithms like SEQUEST or X!Tandem, processing large amounts of data is only realistic with grid computing. A large computational resource available to the community would be truly beneficial.

The emerging diversity of computational algorithms presents a different challenge. There is no single data analysis pipeline. Simple database searches can be augmented with more targeted searches for post-translational modifications, spectral networks, quantification, or searched de novo. Each algorithm, each viewpoint, adds unique and valuable information. In addition to spectrum annotation, data pre-processing and results post-processing is a fruitful area of algorithmic research. Given the variety of options at each step, there are myriads of possible workflows. Routing data from program to program is a major time loss for research. A computational architecture built to facilitate this interconnection is a pressing need for the community.

The Bafna and Pevzer lab had attempted two previous occasions to make their tools available online. But still the tools remained under utilized. Ingolf Krueger was brought in for his expertise in service oriented software and another attempt was made to bring the tools to the end user. The new web tool had two major goals: to provide a simple and unified interface to all lab software, to port the computation to a super computer or grid system. Although I was not involved in the initial stage of the tool development, I soon became an integral part of the team, working extensively with both of these two goals.

Creating computational resources to process large data and to connect new tools in new ways creates two distinct problems. First, the architecture must be agile and flexible to allow the addition of new algorithms and new workflows. Secondly, it must be accessible to new users. As the lab continues its research, many new tools will be developed. The task of integrating them into a single web service is being addressed by members of the Krueger lab. This work encompasses large data transfer, grid interfacing, process control, intuitive interface design, and running a very heterogeneous set of tools. To promote accessibility, all tools need a common interface, which is provided by the web service. Moreover, algorithms cannot be an entirely black box. Proper use of the algorithm requires knowing the meaning and impact of search options, and is essential for users to make coherent workflows. We attempt to address all of these problems: providing the computational resources necessary to perform large proteomics searches, providing workflow tools to promote repeated analysis of the data with a variety of algorithms, educate the user through open source software, and finally to make the interface simple and intuitive.

Another input that I performed was extending our set of collaborators. Having a set of users was essential for good interface design. Members of the Bafna and Pevzner lab are developers, not end users. Although they set out the initial use cases and design, it was imperative to include others in the process. I worked with our collaborators at UCSD to organize a demo and interface feedback session. This was the first time that our target audience was able to talk directly with the developers (members of the Krueger lab). We had members of 3 mass spectrometry labs attend, and the feedback was very useful in determining the basic usability features that they expected. Specific examples like searching and sorting the output, and group data sharing came out of this discussion.

4.3 Hardware Acceleration of MS-Alignment

Another key focus for increasing the use of Inspect is to improve the speed of MS-Alignment, the unrestrictive search algorithm within Inspect. Many biologists are keen to discover post-translational modifications (PTMs) in an unrestricted format. Most algorithms search for PTMs only when requested by the user. Even at this point, they search only the list of modifications that the user inputs. Typically this list includes phosphorylation, methionine oxidation, lysine or arginine methylation, etc. This paradigm, the enumerated PTM search, requires the user to know beforehand the modifications of interest. This search paradigm is very useful in targeted research, for example phospho-proteome studies where the sample has been purified for phosphopeptides. However, more and more researchers want to look for any and all modifications.

Proteins undergo significant post-translational modification in order to modulate their structure, regulate their function, and as part of signaling networks. The functions of many post-translational modifications are well characterized. Phosphorylation is used to signal or activate proteins. Methylation and acetylation are used to modify the state of chromatin and influence gene regulation. These are fairly common modifications that can be found in any proteomics sample. However, there are also several well-characterized rare modifications, e.g. actin arginylation and dipthamide [Karakozova 2006, Van Ness 1980]. Hundreds of other modifications are known; recent research points to newly discovered in vivo modifications, as well as many unknown chemical adducts [Tanner 2008, Chen 2007]. In light of these and future discoveries, the ability to identify all modifications in a sample is particularly attractive. An unrestrictive search can confidently identify rare and uncharacterized modifications.

The central challenge of unrestrictive search is speed. In the unmodified search, database search programs filter the database to only consider a subset of the potential peptides as candidates for scoring. Inspect uses sequence tags to filter the database; other programs use a parent mass filter. Regardless, each of these relies on a limited alphabet composing the peptide sequences. There are 20 amino acids used in protein sequences. When users enumerate a modification (e.g. oxidized methionine), a new mass can be added to the 20 standard amino acids. This slightly expanded alphabet can be easily incorporated into filtering techniques. Unrestricted search, however, have a massively expanded alphabet, and cannot readily adopt standard filtering algorithms. Instead of a 20 character alphabet, the unrestricted search considers 20 amino acids, and up to 500 modified versions of each (up to 250 Daltons addition or subtraction to the mass of the amino acid). By considering all possible modifications, and without any filtering MS-Alignment is at least two orders of magnitude slower than Inspect.

A slower run time for MS-Alignment is at odds with the trend for increasingly large data sets in proteomics. Large data sets, containing millions or tens of millions of spectra, are now routine. Attempting to perform an unrestrictive search on such datasets is a major time investment, and cannot be undertaken without significant computational resources. However, these large datasets are very valuable for the unrestrictive search. Work by Tanner and colleagues have utilized the redundant information in large datasets to accurately provide probability values to sites of post-translational modifications [Tanner 2008].

To make MS-alignment more practical, and extend the benefit of unrestricted modification search to the general proteomics community, the software must be accelerated. Some algorithmic work has pursued novel filtering techniques. Here, however, we have chosen to accelerate the program in hardware. In collaboration with Convey Computers, we are implementing MS-Alignment on a FPGA-based computation engine. Initial profiling of the code revealed that 99.4% of run time was spent inside a single function, making MS-Alignment an ideal candidate for hardware based acceleration. The function, FreeMod.c::SeekMatch1PTM, utilizes a dynamic programming table to place post-translational modifications within a candidate peptide. This is essentially the search function for MS-Alignment, although it is a complete search of the database, unaided by filtering.

To make the function amenable to FPGA hardware specifications, a rewrite was required. In collaboration with Mark Kelly and Glen Edwards at Convey, this function has been transformed. Based on their analysis they estimate a 300-1000 fold speed up in the function when run on the Convey FPGA platform. Overall, this translates to a 100 – 150 fold speed up for the program as a whole. This is a significant milestone, not only for the dramatic speed up, but also because this amount of acceleration obviates the need for grid computing. The current pipeline for MS-Alignment is to distribute jobs on to the FWGrid (fwgrid.ucsd.edu). This large computing resource has allocated users up to 128 nodes at a time. Typically I run MS-Alignment at full capacity on my account. Thus, if Convey is able to achieve the projected speed up, they will have equaled the resources of the grid. This is a substantial savings.

By using a single processor with FPGAs, we save the electricity and system administration overhead of using a supercomputer.

Chapter 5: Retention and Loss of Amino Acid Biosynthetic Pathways based on Analysis of Whole Genome Sequences

Plants and fungi can synthesize each of the 20 amino acids by using biosynthetic pathways inherited from their bacterial ancestors. However, the ability to synthesize nine amino acids (Phe, Trp, Ile, Leu, Val, Lys, His, Thr, and Met) was lost in a wide variety of eukaryotes that evolved the ability to feed on other organisms. Since the biosynthetic pathways and their respective enzymes are well characterized, orthologs can be recognized in whole genomes to understand when in evolution pathways were lost. The pattern of pathway loss and retention was analyzed in the complete genomes of three early-diverging protist parasites, the amoeba Dictyostelium, and six animals. The nine pathways were lost independently in animals, Dictyostelium, Leishmania, Plasmodium, and Cryptosporidium. Seven additional pathways appear to have been lost in one or another parasite, demonstrating that they are dispensable in a nutrition-rich environment. Our predictions of pathways retained and pathways lost based on computational analyses of whole genomes are validated by minimal-medium studies with mammals, fish, worms, and Dictyostelium. The apparent selective advantages of retaining biosynthetic capabilities for amino acids available in the diet are considered.

5.1 Introduction

Before the genomic era, minimal-medium studies offered essential information about the metabolic potential of an organism. Comparative genomics can now discover the same information about an organism even when the life cycle or environment is so complex as to preclude the defining of a minimal medium. The amino acid requirements of a variety of organisms were the subject of considerable interest in the last century. The first successful synthetic diet using purified amino acids was reported in 1935 for rats from the laboratory of W. C. Rose (McCoy 1935). Subsequent work showed that rats, mice, or salmon fed on diets lacking any one of nine amino acids (Phe, Trp, Ile, Leu, Val, Lys, His, Thr, or Met) would waste away and die (Greenstein 1961, Rose 1948). These are known as the essential amino acids. The other 11 amino acids found in proteins could be omitted from the diet with no deleterious effects and so were considered nonessential. Yeasts such as Saccharomyces cerevisiae, as well as plants such as Arabidopsis thaliana, are able to grow in media devoid of amino acids, demonstrating that they can synthesize all of the amino acids from sugars and fats in the media or generated photosynthetically. Clearly, the common progenitor of plants, fungi, and animals carried genes for all of the enzymes in the 20 amino acid biosynthetic pathways, but almost half of them have been lost in consumers. Now that the sequences of a considerable number of eukaryotic genomes have been completed, we can inspect them to determine when in evolution the pertinent genes were lost.

Since the biosynthetic pathways and their respective enzymes are well characterized in mammals and fungi, orthologs can be recognized in whole genomes. When key enzymes in a pathway are missing, it can be concluded that the respective amino acid is not synthesized. We analyzed the genomes of two alveolates, Cryptosporidium hominis and Plasmodium falciparum; one euglenozoid, Leishmania major; and six animals, Homo sapiens, Tetraodon nigroviridis, Ciona intestinalis, Drosophila melanogaster, Anopheles gambiae, and Caenorhabditis elegans. Previously, we used this approach to predict the metabolic capabilities of a free-living soil amoeba, Dictyostelium discoideum, for which a defined medium had been developed (Franke 1977). We confirmed that the pathways leading to the five amino acids not added to the medium were intact but also found genes for the pathways leading to four other amino acids that were included in the medium (Payne 2005). We verified that Dictyostelium cells could synthesize these four amino acids by successfully growing cells in media lacking all of them. Omission of any of the 11 remaining amino acids in the new minimal medium precluded growth of the cells, confirming the bioinformatic analyses which showed that genes for the pathways to these 11 amino acids were missing in the genome (Payne 2005). This approach seems sufficiently robust to apply to other organisms with fully sequenced genomes so as to follow the pattern of gene retention and loss during evolution.

5.2 Materials and Methods

Databases. Complete genome sequences for humans (H. sapiens), the fly D. melanogaster, the mosquito A. gambiae, the worm C. elegans, the yeast S. cerevisiae, and the plant A. thaliana were downloaded from the National Center for Biotechnology Information. The genome for the fish T. nigroviridis was downloaded from Genoscope (genoscope.cns.fr), and that for the chordate C. intestinalis was downloaded from the Department of Energy Joint Genome Institute. The genomes for D. discoideum, P. falciparum, L. major, and C. hominis were recovered from their respective websites: dictybase.org, plasmodb.org, sanger.ac.uk/Projects/L_major/, and hominis.mic.vcu.edu.

Orthology. The amino acid sequences of enzymes in the amino acid biosynthetic pathways of S. cerevisiae and H. sapiens were downloaded from the KEGG website (www.kegg.com; Kanehisa 1997, Kanehisa 2000). Pfam domains in these enzymes were used to collect potential orthologs from other organisms [http://hmmer.wustl.edu/; Bateman 2004]. The yeast and human enzymes were also compared to gene products in the other complete genomes by using the BlastP program. Genes with the pertinent Pfam domain(s) and a BLAST score of e-80 or better were considered functional orthologs. Genes below this BLAST threshold that still had the pertinent Pfam domains were checked by mutual best BLAST hit. A BlastP cutoff of e-20 was used for the early-diverging eukaryotes. Genes were considered missing if there were no hits at better than e-1. The smallest of the enzymes used to

BLAST was 221 amino acids and so should be easily recognized. In addition to the yeast and human enzymes, bacterial enzymes from KEGG were also used to query the genomes of the early-diverging eukaryotes. No additional putative amino acid biosynthetic genes were found.

Pathways. After orthologous genes were collected, the biosynthetic pathway to each amino acid in each organism was analyzed. If every enzyme in the pathway had an ortholog in a genome, the pathway was considered functional in that organism. If one or more enzymes in a pathway were missing in a genome, the pathway was considered nonfunctional in that organism. When a biosynthetic pathway appeared to be nonfunctional in an organism that was phylogenetically close to organisms with an intact pathway, we manually inspected the genome.

5.3 Results

By searching the genomes of the 10 eukaryotic organisms for protein sequences and Pfam domain structures, we were able to find probable functional orthologs to amino acid metabolic enzymes. These orthologs were organized into the biosynthesis pathways, which were then predicted to be either functional or nonfunctional. Almost always, nonfunctional pathways lacked pertinent genes even when the BLAST threshold was set at e-1. When a biosynthetic gene was present, the BLAST hit score was always better than e-60 (the great majority better than e-80). This resulted in a very clear distinction between present and missing genes. There were a few cases where genes with intermediate BLAST scores were found but were clearly not functional orthologs. For example, yeast isopropylmalate dehydrogenase, which is encoded by a leucine biosynthesis gene, recognized a human gene (IDH3A) with a BLAST score of e-17. This gene, however, encodes isocitrate dehydrogenase (National Center for Biotechnology Information gi,18314368).

Multistep biosynthetic pathways in which an essential gene is missing were almost always found to have lost all of the genes dedicated to that pathway (Table 1). In only two cases was a single gene missing from the pathway. The Dictyostelium serine pathway lacks a functional phosphoserine phosphatase. As expected, Dictyostelium requires serine for growth (Payne 2005). The Leishmania methionine pathway lacks cystathionine gamma-synthase. Although it is possible that a gene unrelated to the traditional cystathionine gamma-synthase could have acquired this function and remained unrecognized, such cases are rare. Alternative pathways not including the missing genes would have been considered in Dictyostelium, worms, fish, or mammals when their minimal media were defined. However, all of the amino acids expected to be required were directly observed to be essential.

The parasites Leishmania, Cryptosporidium, and Plasmodium grow within mammalian cells, an environment rich in amino acids. They diverged before the separation of plants and animals and must have inherited genes for the biosynthesis of all amino acids (Fig. 1). However, they subsequently lost the ability to make most amino acids (Tables 2 and 3). Cryptosporidium, whose metabolic capabilities are sparse (Xu 2004), has retained the biosynthetic pathways for only four amino acids: asparagine, glutamine, glycine, and proline. Plasmodium has retained the ability to synthesize these four plus aspartate and glutamate. The euglenozoid Leishmania diverged separately from Cryptosporidium and Plasmodium. Although they are all intracellular parasites, Leishmania has a more expansive set of amino acid biosynthetic pathways: alanine, asparagine, aspartate, cysteine, glutamate, glutamine, glycine, proline, and tyrosine.

Dictyostelium, a phagocytic amoeba, possesses amino acid biosynthetic capabilities very similar to those of metazoans (Tables 2 and 3). In addition to losing all of the genes in pathways for the human-essential amino acids, they have lost genes required for serine and arginine biosynthesis. These bioinformatic predictions have been experimentally verified by testing the ability of Dictyostelium to grow in media lacking these amino acids (Payne 2005). Since the yeast S. cerevisiae diverged from the line leading to vertebrates after Dictyostelium and has retained the ability to synthesize the 20 amino acids, loss of these pathways in the amoebae must have occurred independently of the subsequent loss in the animal branch [Bapteste 2002].

All of the enzymes for the biosynthesis of the 11 amino acids known to be nonessential for rodents can be easily recognized in the human and fish genomes (Table 2). Ten of the 11 pathways are complete in the genomes of all metazoans, but enzymes for the synthesis of arginine are missing in C. intestinalis, D. melanogaster, A. gambiae, and C. elegans. Arginine has previously been shown to be an essential amino acid for C. elegans (Vanfleteren 1980). The loss of the arginine biosynthetic enzymes appears to have occurred independently in these organisms, after each diverged from the line leading to vertebrates, which can still make arginine. All animals lack the biosynthetic pathways for isoleucine, leucine, valine, phenylalanine, tryptophan, lysine, methionine, threonine, and histidine (Table 3). For genes dedicated to these pathways, no putative homolog could be found in any animal genome.

Only four pathways are universally conserved in the 12 eukaryotes examined: those leading to asparagine, glutamine, glycine, and proline. The enzymes for glycine and glutamine synthesis, serine hydroxymethyltransferase and glutamine synthase, are highly conserved in all organisms. Although the ability to synthesize asparagine and proline is universally conserved, analysis of the enzymes in the pathways suggests that the means for their synthesis is not. Asparagine synthase (glutamine hydrolyzing) transaminates aspartate from glutamine and is present in animals, Dictyostelium, and Plasmodium. Alternately, aspartate can be amidated with free ammonia by the aspartate ammonia-ligase present in Cryptosporidium and Leishmania. Likewise, proline is synthesized in Dictyostelium and Plasmodium from arginine via ornithine and glutamate semialdehyde, while animals, Leishmania, and Cryptosporidium encode enzymes to convert glutamate to proline via phospho-glutamate and glutamate semialdehyde. The final steps in these pathways to proline are both catalyzed by pyrroline-5-carboxylate reductase.

5.3 Discussion

When an organism becomes a consumer by eating other organisms, all of the amino acids are available in the diet and no longer need to be synthesized. Unless amino acid biosynthetic pathways serve other essential functions besides providing an amino acid, they are unnecessary and dispensable. Genes in the dispensable pathways accumulate deleterious mutations, lose the ability to encode functional enzymes, and are eventually deleted from the genome. Deletion was the fate for almost all of the genes specific to the pathways that were lost (Table 1). It is important to note that this common set of pathways was lost in at least four independent evolutionary instances (Fig. 1). This process of complete purging of genes in a pathway is not without precedent (Hittinger 2004). Our predictions of pathways retained and pathways lost based on computational analyses of whole genomes are validated by the observed minimal requirements of mammals, fish, worms, and

Dictyostelium (Greenstein 1961, Payne 2005, Vanfleteren 1980). Therefore, our predictions for insects, Ciona, and the parasites are likely to be substantiated when minimal media are defined for these organisms.

Selective conservation of a pathway over time indicates that it is indispensable for the metabolic needs of the organism. An example of this is the arginine synthesis pathway, which is part of the urea cycle used in mammals and embryonic fish to remove excess nitrogen (Walsh 1998, Wright 1995). Dictyostelium and invertebrates utilize alternate nitrogen excretion metabolites (Craig 1960, Payne 2005, Wright 1998). Therefore, the urea cycle and arginine synthesis are not essential and were lost in these organisms.

Only four pathways (Asn, Gln, Gly, and Pro) are universally conserved. The importance of these pathways is emphasized by their retention in Cryptosporidium. Its tiny genome lacks genes for the tricarboxylic acid cycle and the biosynthetic pathways to sugars and nucleotides (Xu 2004), yet it has retained these capabilities. The synthesis of glycine also produces 5,10-methylenetetrahydrofolate. This is the major source of one-carbon units used in dTMP and purine ring biosynthesis, making this pathway indispensable. Glutamine synthesis is essential for nitrogen assimilation, detoxification, and general nitrogen metabolism (e.g., transamination), and the same may be true for asparagine. The utility of the pathway to proline is not immediately obvious. However, yeast strains in which the gene encoding pyrroline-5-carboxylate reductase is deleted have been found to grow slowly even in rich media with a plentiful supply of proline (Barbara Dunn, personal communications).

Dictyostelium feeds on bacteria and yeast. They have lost the pathways to all of the amino acids essential for humans. Surprisingly, they have retained all of the biosynthetic pathways found in metazoans, except that for serine. Loss of the ability to synthesize serine appears to result from the fairly recent inactivation of phosphoserine phosphatase, since a pseudogene can be recognized in the genome. Our bioinformatically predicted pathway loss and retention are validated by minimal-medium studies (Payne 2005).

Ten pathways (Ala, Asp, Asn, Gly, Ser, Cys, Tyr, Pro, Glu, and Gln) were uniformly conserved in the animal lineage. Discerning the selective advantage they might provide is aided in some cases by symptoms of human metabolic disorders. The tyrosine synthesis pathway is also part of the phenylalanine catabolic pathway. Mutations in phenylalanine hydroxylase, which makes tyrosine, are the cause of phenylketonuria. The resulting buildup of phenylalanine and relative depletion of tyrosine cause clinical symptoms of seizures and mental retardation (Kahler 2003). Likewise, loss of cystathionine beta-synthase, a component of the cysteine pathway, causes a buildup of homocysteine, which contributes to ocular, skeletal, nervous system, and vascular problems (Townsend 2004). Defects in enzymes of the serine biosynthesis pathway lead to congenital microcephaly, severe psychomotor retardation, and intractable seizures [de Koning 2004]. When any of these three pathways fails, the resulting metabolic imbalance has severe consequences.

The importance of the glycine, asparagine, and glutamine pathways has been previously discussed. The alanine pathway, like proline, does not have an obvious reason for conservation. In the same yeast experiment, knocking out alanine transaminase was detrimental to cells, even in rich media with a plentiful supply of alanine (Barbara Dunn, personal communications). The importance of aspartate and glutamate is likely to result from their nitrogen handling. It is likely that free-living organisms require a more responsive nitrogen-handling capability than parasites and therefore require the ability to synthesize glutamate and aspartate, so as not to rely only on the simple transaminase reactions involving glutamine and asparagine.

The metabolic capabilities of parasites living within other cells offer unique insights into the loss and retention of pathways. Parasites lack several pathways that are conserved in all of the free-living organisms we studied. The ability to thrive without these pathways may be confined to the obligate intracellular parasites that rely on their host for additional metabolic functions. For instance, the lack of phenylalanine hydroxylase in Cryptosporidium may not result in a phenylalanine-tyrosine imbalance if these amino acids are rapidly exchanged with the host where excess phenylalanine can be metabolized. Thus, by identifying conspicuous voids in metabolic capabilities, we can learn what critical functions the host provides for its parasite with the potential of intervention.

Acknowledgments This work was supported by an NSF Biocomplexity Grant and NIH grant GM62350 to W.F.L. S.H.P. is supported by an NIH training grant (GM08806). We thank Rolf Olsen and Christophe Anjard for useful discussions. Barbara Dunn, Stanford University, Stanford, CA, kindly shared un-published results. We have no financial conflict of interest. This chapter, in full, was published as “Retention and Loss of Amino Acid Biosynthetic Pathways Based on Whole Genome Sequences.” Payne SH and Loomis WF (2006) Eukaryot Cell 5(2):272-6. The dissertation author was the primary author of this paper.

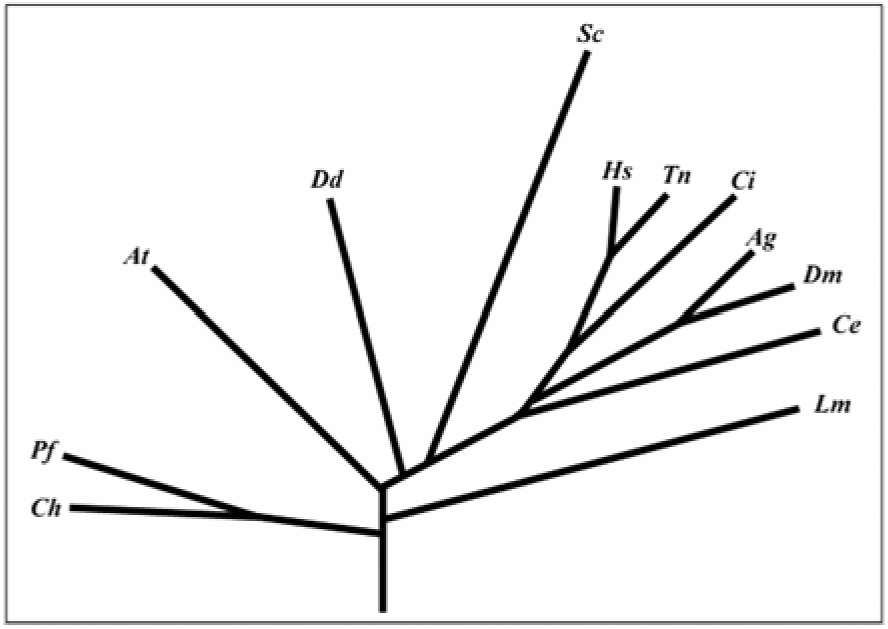


Figure 5.1 - Evolutionary tree depicting the branching order of the organisms studied (adapted from reference 15). The tree is rooted on seven archaebacterial genomes. Abbreviations: *Ch*, *C. hominis*; *Pf, P. falciparum*; *At, A. thaliana*; *Dd, D. discoideum*; *Sc, S. cerevisiae*; *Hs, H. sapiens*; *Tn, T. nigroviridis*; *Ci, C. intestinalis*; *Ag, A. gambiae*; *Dm, D. melanogaster*; *Ce, C. elegans*; *Lm, L. major*.

Table 5.1 - The biosynthetic pathways leading to these 11 amino acids involve multiple steps. The number of genes in each pathway in *S. cerevisiae* was taken from KEGG. The number of genes for each pathway that are missing in the other organisms was determined from their complete genomes. The amino acids are given in the single-letter code. (a) *C. elegans* is missing all three genes of the arginine (R) biosynthetic pathway, but vertebrates have maintained them for use in the urea cycle (see text).

Pathway	No. of dedicated genes	No. of enzymes missing in:				
		Animals	D.d	P.f	C.h	L.m
ILV	5	5	5	5	5	5
H	7	7	7	7	7	7
K	3	3	3	3	3	3
FW	11	11	11	10	10	11
R	3	3a	3	3	3	2
T	5	5	5	5	5	5
M	2	2	2	2	2	1
S	3	0	1	3	3	2

Table 5.2 - The synthesis capabilities of the following eukaryotic organisms are shown: *H. sapiens* (*H.s.*), *T. nigroviridis* (*T.n.*), *C. intestinalis* (*C.i.*), *D. melanogaster* (*D.m.*), *A. gambiae* (*A.g.*), *C. elegans* (*C.e.*), *S. cerevisiae* (*S.c.*), *D. discoideum* (*D.d.*), *A. thaliana* (*A.t.*), *C. hominis* (*C.h.*), *L. major* (*L.m.*), and *P. falciparum* (*P.f.*). A plus sign indicates that the genes for all of the enzymes of the biosynthetic pathway are present in the genome. A minus sign indicates that one or more of the genes encoding an enzyme in the pathway are missing in the genome. (*a) L. major* and *C. hominis* amidate aspartate with ammonia.

Amino acid	Vertebrates		Chordate	Invertebrates			Unicellular organisms		Plant	Intracellular parasites		
	H.s.	T.n.	C.i.	D.m.	A.g.	C.e.	S.c.	D.d.	A.t.	L.m.	P.f.	C.h.
Alanine	+	+	+	+	+	+	+	+	+	+	–	–
Asparagine	+	+	+	+	+	+	+	+	+	+[a]	+	+[a]
Aspartate	+	+	+	+	+	+	+	+	+	+	+	–
Arginine	+	+	–	–	–	–	+	–	+	–	–	–
Cysteine	+	+	+	+	+	+	+	+	+	+	–	–
Glutamate	+	+	+	+	+	+	+	+	+	+	+	–
Glutamine	+	+	+	+	+	+	+	+	+	+	+	+
Glycine	+	+	+	+	+	+	+	+	+	+	+	+
Proline	+	+	+	+	+	+	+	+	+	+	+	+
Serine	+	+	+	+	+	+	+	–	+	–	–	–
Tyrosine	+	+	+	+	+	+	+	+	+	+	–	–

Table 5.3 - The synthesis capabilities of the following eukaryotic organisms are shown: *H. sapiens* (*H.s.*), *T. nigroviridis* (*T.n.*), *C. intestinalis* (*C.i.*), *D. melanogaster* (*D.m.*), *A. gambiae* (*A.g.*), *C. elegans* (*C.e.*), *S. cerevisiae* (*S.c.*), *D. discoideum* (*D.d.*), *A. thaliana* (*A.t.*), *C. hominis* (*C.h.*), *L. major* (*L.m.*), and *P. falciparum* (*P.f.*). A plus sign indicates that the genes for all of the enzymes of the biosynthetic pathway are present in the genome. A minus sign indicates that one or more of the genes encoding an enzyme in the pathway are missing in the genome.

Amino acid	Vertebrates		Chordate	Invertebrates			Unicellular organisms		Plant	Intracellular parasites		
	H.s.	T.n.	C.i.	D.m.	A.g.	C.e.	S.c.	D.d.	A.t.	L.m.	P.f.	C.h.
Histidine	–	–	–	–	–	–	+	–	+	–	–	–
Isoleucine	–	–	–	–	–	–	+	–	+	–	–	–
Leucine	–	–	–	–	–	–	+	–	+	–	–	–
Lysine	–	–	–	–	–	–	+	–	+	–	–	–
Methionine	–	–	–	–	–	–	+	–	+	–	–	–
Phenylalanine	–	–	–	–	–	–	+	–	+	–	–	–
Threonine	–	–	–	–	–	–	+	–	+	–	–	–
Tryptophan	–	–	–	–	–	–	+	–	+	–	–	–
Valine	–	–	–	–	–	–	+	–	+	–	–	–

References

Andersson L, Porath J. 1986. Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. Anal Biochem, 154:250–254.

Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller M, Philippe H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc. Natl. Acad. Sci. USA 99:1414–1419.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. Nucleic Acids Res. 32:D138–D141.

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S. 2008. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science, 320:938-41

Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for highthroughput protein phosphorylation analysis and site localization. Nat Biotechnol, 24:1285–1292.

Biemann K, Cone C, Webster BR, Arsenault GP. 1966. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. J Am Chem Soc., 88:5598-606.

Breci LA, Tabb DL, Yates JR 3rd, Wysocki VH. 2003. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. Anal. Chem, 75:1963–71.

Brent MR. 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nature Reviews Genetics, 9:62-73

Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O, Lee H, Pedrioli PG, Malmstrom J, Koehler K, Schrimpf S, Krijgsveld J, Kregenow F, Heck AJ, Hafen E, Schlapbach R, Aebersold R. 2007. A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol, 25:576-83.

Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol, 4:10.

Carr SA, Biemann K, Shoji S, Parmelee DC, Titani K. 1982. n-Tetradecanoyl is the NH2-terminal blocking group of the catalytic subunit of cyclic AMP-dependent protein kinase from bovine cardiac muscle. Proc Natl Acad Sci U S A, 79:6128-31.

Chen Y, Sprung R, Tang Y, Ball H, Sangras B, Kim SC, Falck JR, Peng J, Gu W, Zhao Y. 2007. Lysine propionylation and butyrylation are novel post-translational modifications in histones.Mol Cell Proteomics, 6:812-9.

Chi A, Huttenhower C, Geer LY, Coon JJ, Syka JE, Bai DL, Shabanowitz J, Burke DJ, Troyanskaya OG, Hunt DF. 2007. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc Natl Acad Sci U S A, 104:2193–2198.

Chitteti BR, Peng Z. 2007. Proteome and phosphoproteome dynamic change during cell dedifferentiation in Arabidopsis. Proteomics, 7:1473–1500.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A, 104:19428-33.

Cosgrove MS, Wolberger C. 2005. How does the histone code work? Biochem Cell Biol, 83:468-76.

Craig R. 1960. The physiology of excretion in the insect. Annu. Rev. Entomol. 5:53–68.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 20:1466–1467.

Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. 1999. De novo peptide sequencing via tandem mass spectrometry. J Comput Biol, 1999, 6:327–342.

DeBlasio SL, Leusse DL, Hangarter RP (2005) A plant-specific protein essential for blue-light-induced chloroplast movements. Plant Physiol 139:101-14.

DeGnore JP, Qin J. 1998. Fragmentation of phosphopeptides in an ion trap mass spectrometer. J Am Soc Mass Spectrom, 9:1175–1188.

Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R. 2005. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol, 6:R9.

de Koning TJ, Klomp LW. 2004. Serine-deficiency syndromes. Curr. Opin. Neurol. 17:197–204.

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* 96**:**1151–1160.

Elias JE, Haas W, Faherty BK, Gygi SP. 2005. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods, 2:667–675.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 4:207-14.

Eng JK, McCormack AL, Yates JR. 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, 5:976–989.

Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ. 2006. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol, 7:R35.

Fields S, Song O. 1989. A novel genetic system to detect protein-protein interactions. Nature, 340:245-6.

Frank A, Pevzner P. 2005. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem, 77:964–973.

Franke J, Kessin R. 1977. A defined minimal medium for axenic strains of Dictyostelium discoideum. Proc. Natl. Acad. Sci. USA 74:2157–2161.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature, 425:737-41.

Greenstein JP, Winitz M. 1961. Chemistry of the amino acids. John Wiley & Sons, Inc., New York, N.Y.

Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD, Pevzner PA. 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res, 17:1362-77.

Havilio M, Haddad Y, Smilansky Z. 2003. Intensity-based statistical scorer for tandem mass spectrometry. Anal Chem, 75:435–444.

Hittinger CT, Rokas A, Carroll SB. 2004. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc. Natl. Acad. Sci. USA 101:14144–14149.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. Nature,425:686-91

Hunt DF, Yates JR 3rd, Shabanowitz J, Winston S, Hauer CR. 1986. Protein sequencing by tandem mass spectrometry. Proc Natl Acad Sci U S A, 83:6233–6237.

Hunter T. 2000. Signaling-2000 and beyond. Cell, 100:113–127.

Jensen FV. 2001. Bayesian Networks and Decision Graphs. Springer.

Jensen ON. 2006. Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol, 7:391–403.

Jiang B, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:163-7.

Kahler SG, Fahey MC. 2003. Metabolic disorders and mental retardation. Am. J. Med. Genet. C Semin. Med. Genet. 117:31–41.

Kanehisa M. 1997. A database for post-genome analysis. Trends Genet. 13:375–376.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.

Karakozova M, Kozak M, Wong CC, Bailey AO, Yates JR 3rd, Mogilner A, Zebroski H, Kashina A. 2006. Arginylation of beta-actin regulates actin cytoskeleton and cell motility. Science, 313:192-6.

Keller A, Eng J, Zhang N, Li XJ, Aebersold R. 2005. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol, 1:2005.0017.

Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem, 74:5383-92

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423:241-54.

Klammer AA, Wu CC, MacCoss MJ, Noble WS. 2005. Peptide charge state determination for low-resolution tandem mass spectra. Proc IEEE Comput Syst Bioinform Conf, pages 175–185.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440:637-43

Lee J, Xu Y, Chen Y, Sprung R, Kim SC, Xie S, Zhao Y. 2007. Mitochondrial phosphoproteome revealed by an improved IMAC method and MS/MS/MS. Mol Cell Proteomics, 6:669–676.

Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE, Roark M, Wiley KL Jr, Kulathinal RJ, Zhang P, Myrick KV, Antone JV, Celniker SE, Gelbart WM, Kellis M. 2007. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res, 17:1823-36.

Loo JA, Edmonds CG, Smith RD. 1993. Tandem mass spectrometry of very large molecules. 2. Dissociation of multiply charged proline-containing proteins from electrospray ionization. Anal Chem, 65:425–438.

Lu B, Ruse C, Xu T, Park SK, Yates J 3rd. 2007. Automatic validation of phosphopeptide identifications from tandem mass spectra. Anal Chem, 79:1301–1310.

Macek B, Mijakovic I, Olsen JV, Gnad F, Kumar C, Jensen PR, Mann M. 2007. The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. Mol Cell Proteomics, 2007, 6:697–707.

Mann M, Wilm M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem,66:4390-9.

McCoy RH, Meyer CD, Rose WC. 1935. Feeding experiments with highly purified amino acids. VIII. Isolation and identification of a new essential amino acid. J. Biol. Chem. 112:283.

McLafferty FW. 1981. Tandem Mass Spectrometry. Science, 214:280-287.

Molina H, Horn DM, Tang N, Mathivanan S, Pandey A. 2007. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci U S A, 104:2199–2204.

Nielsen ML, Savitski MM, Zubarev RA. 2006. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. Mol Cell Proteomics, 5:2384-91.

Noble WS. 2006. What is a support vector machine? Nature Biotechnology, 24:1565–1567.

Nousiainen M, Silljé HH, Sauer G, Nigg EA, Körner R. 2006. Phosphoproteome analysis of the human mitotic spindle. Proc Natl Acad Sci U S A, 103:5391–5396.

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. 2006. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell, 127:635–648.

Olsen RM. 2005. How many protein encoding genes does Dictyostelium discoideum have? p. 265–278. In W. F. Loomis and A. Kuspa (ed.), Dictyostelium genomics. Horizon Press, Far Hills, N.J.

Payne SH. 2005. Metabolic pathways, p. 41–57. In W. F. Loomis and A. Kuspa (ed.), Dictyostelium genomics. Horizon Press, Far Hills, N.J.

Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, Bafna V. 2008. Phosphorylation specific scoring of MS/MS spectra for rapid and accurate phosphoproteome anlaysis. J Proteome Res, in press.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20:3551-67.

Pinkse MW, Uitto PM, Hilhorst MJ, Ooms B, Heck AJ. 2004. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. Anal Chem, 76:3935–3943.

Rose WC, Oesterling MJ, Womack M. 1948. Comparative growth on diets containing ten and nineteen amino acids, with further observations upon the role of glutamic and aspartic acids. J. Biol. Chem. 176:753–762.

Savidor A, Donahoo RS, Hurtado-Gonzales O, Verberkmoes NC, Shah MB, Lamour KH, McDonald WH. 2006. Expressed peptide tags: an additional layer of data for genome annotation. J Proteome Res, 5:3048-58.

Shu H, Chen S, Bi Q, Mumby M, Brekken DL. 2004. Identification of phosphoproteins and their phosphorylation sites in the WEHI-231 B lymphoma cell line. Mol Cell Proteomics, 3:279–286.

Smolka MB, Albuquerque CP, Chen SH, Zhou H. 2007. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. Proc Natl Acad Sci U S A, 104:10364–10369.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24:637-44.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res, 34:W435-9.

Stasyk T, Huber LA. Zooming in: fractionation strategies in proteomics. 2004. Proteomics, 4:3704-16.

Suthram S, Sittler T, Ideker T. 2005. The Plasmodium protein network diverges from those of other eukaryotes. Nature, 438:108-12.

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res, 36:D1009-14.

Tanner S, Payne SH, Dasari S, Shen Z, Wilmarth PA, David LL, Loomis WF, Briggs SP, Bafna V. 2008. Accurate annotation of peptide modifications through unrestrictive database search.J Proteome Res, 7:170–81.

Tanner S, Shen Z, Ng J, Florea L, Guigó R, Briggs SP, Bafna V. 2007. Improving gene annotation using peptide mass spectrometry. Genome Res, 17:231-9.

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. 2005. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem, 77:4626–4639.

Trinidad JC, Specht CG, Thalhammer A, Schoepfer R, Burlingame AL. 2006. Comprehensive identification of phosphorylation sites in postsynaptic density preparations. Mol Cell Proteomics, 5:914–922.

Townsend DM, Tew KD, Tapiero H. 2004. Sulfur containing amino acids and human disease. Biomed. Pharmacother. 58:47–55.

Ubersax JA, Ferrell JE Jr. Mechanisms of specificity in protein phosphorylation. 2007. Nat Rev Mol Cell Biol, 8:530-41.

Uy R, Wold F. 1977. Posttranslational covalent modification of proteins. Science, 198:890-6.

Van Ness BG, Howard JB, Bodley JW. ADP-ribosylation of elongation factor 2 by diphtheria toxin. 1980. Isolation and properties of the novel ribosyl-amino acid and its hydrolysis products. J Biol Chem, 255:10717-20.

Vanfleteren JR. 1980. Nematodes as nutritional models, p. 47–79. In B. M. Zuckerman (ed.), Nematodes as biological models, vol. 2. Academic Press, Inc., New York, N.Y.

Venable JD, Xu T, Cociorva D, Yates JR 3rd. 2006. Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra. Anal Chem, 78:1921–1929.

Villén J, Beausoleil SA, Gerber SA, Gygi SP. 2007. Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A, 104:1488–1493.

Walsh PJ. 1998. Nitrogen excretion and metabolism, p. 199–214. In D. H. Evans (ed.), The physiology of fishes, 2nd ed. CRC Press, Inc., New York, N.Y.

Wright DJ. 1998. Respiratory physiology, nitrogen excretion and osmotic and ionic regulation, p. 103–131. In R. N. Perry and D. J. Wright (ed.), The physiology and biochemistry of free-living and plant-parasitic nematodes. CABI Publishing, New York, N.Y.

Wright P, Felskie A, Anderson P. 1995. Induction of ornithine-urea cycle enzymes and nitrogen metabolism and excretion in rainbow trout (Oncorhynchus mykiss) during early life stages. J. Exp. Biol 198:127–135.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA. 2004. The genome of Cryptosporidium hominis. Nature 431:1107–1112.

Zhou H, Watts JD, Aebersold R. 2001. A systematic approach to the analysis of protein phosphorylation. Nat Biotechnol, 19:375–378.