UNIVERSITY OF CALIFORNA, SAN DIEGO

Efficient and Accurate Bioinformatics Algorithms for Peptide Mass Spectrometry

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

Doctor of Philosophy

in

Bioinformatics

by

Stephen Will Tanner

Committee in charge:

Professor Vineet Bafna, Chair
Professor Julian Schroeder, Co-Chair
Professor Steve Briggs
Professor Trey Ideker
Professor Pavel Pevzner

2007

The Dissertation of Stephen Will Tanner is approved, and it is acceptable in quality and for publication on microfilm:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2007

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to gratefully acknowledge the support and advice of my advisor, Vineet Bafna, and to the other fine scientists on my committee: Julian Schroeder, Pavel Pevzner, Steve Briggs, and Trey Ideker. I give special thanks to fellow students Ari Frank, Nuno Bandeira, Samuel Payne, Julio Ng, Natalie Castellana, Nitin Gupta, Qian Peng, and Vagisha Sharma for discussions, collaborations, and help improving InsPecT. Thanks are also due to Helge Weissig of ActivX and Pankaj Agarwal of GlaxoSmithKline for two eye-opening visits to industry. I am grateful to all our collaborating labs, without whose data our work would not have been possible.

This project was supported by US National Institute of Health grant NIGMS 1-R01-RR16522. Stephen Tanner is supported by a NSF IGERT training grant DGE0504645. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. Part of this investigation was supported using the computing facility made possible by the Research Facilities Improvement Program Grant Number C06 RR017588 awarded to the Whitaker Biomedical Engineering Institute, and the Biomedical Technology Resource Centers Program Grant Number P41 RR08605 awarded to the National Biomedical Computation Resource, UCSD, from the National Center for Research Resources, National Institutes of Health.

Chapter 5, in full, was published as "Identification of Post-translational Modifications via Blind Search of Mass-Spectra". Tsur, D. and Tanner, S. and Zandi, E. and Bafna, V. and Pevzner, P.A. 2005. Nature Biotechnology **23**: 1562-1567. The dissertation author and Dr. Tsur were primary authors of this paper.

Chapter 6 is in preparation for publication as "Accurate Annotation of Peptide Modifications through Unrestrictive Database Search". Tanner, Stephen and Payne, Samuel H. and Dasari, Surendra and Shen, Zhouxin and Wilmarth, Philip and David, Larry and Loomis, William F. and Briggs, Steven P. and Bafna, Vineet 2007, in preparation. The dissertation author was the primary author of this paper.

Chapter 7, in full, was published as "Improving gene annotation with mass spectrometry". Tanner, Stephen and Shen, Zhouxin and Ng, Julio and Florea, Liliana and Guigo, Roderic and Briggs, Steven P and Bafna, Vineet 2007. Genome Research 17(2), 231-239. The dissertation author was the primary author of this paper.

Chapter 8 is in preparation for publication as "Generalized gene set queries for microarray analysis". Tanner, S. and Agarwal, P. 2007, in preparation. The dissertation author was the primary author of this paper.

Vita

1992	Bachelor of Arts, Brigham Young University
1995	Master of Arts, University of Wisconsin, Madison
1996-2001	TenFold Corporation
2002	Symantec Corporation
2003	Bachelor of Science, University of Utah
2007	Doctor of Philosophy, University of California, San Diego

Publications

Generalized gene set queries for microarray analysis. Stephen Tanner, Pankaj Agarwal, 2007. In preparation.

Clustering Tandem Mass Spectra: From Spectral Libraries to Spectral Archives. Ari M. Frank, Nuno Bandeira, Zhouxin Shen, Stephen Tanner, Steven P. Briggs, Richard D. Smith, Pavel A. Pevzner, 2007. In preparation.

Accurate Annotation of Peptide Modifications through Unrestrictive Database Search. Stephen Tanner, Samuel H. Payne, Surendra Dasari, Zhouxin Shen, Phillip Wilmarth, Larry David, William F. Loomis, Steven P. Briggs and Vineet Bafna, 2007. In preparation.

Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Nitin Gupta, Stephen Tanner, Navdeep Jaitly, Joshua Adkins, Mary Lipton, Robert Edwards, Margaret Romine, Andrei Osterman, Vineet Bafna, Richard D. Smith, Pavel Pevzner, 2007. In preparation.

Patricelli, M.P., Szardenings, A.K., Liyanage, M., Nomanbhoy, T.K., Wu, M., Weissig, H., Aban, A., Chun, D., Tanner, S., Kozarich, J.W., 2007. Functional Interrogation of the Kinome Using Nucleotide Acyl Phosphates. Biochemistry 46: 350-358.

Improving gene annotation using peptide mass spectrometry. Stephen Tanner, Pavel A. Pevzner, Liliana Florea, Steve Briggs, Zhouxin Chen, Julio Ng, Roderic Guigo, and Vineet Bafna, 2007. Genome Research 17(2):231-239.

Annotation of peptide mass spectra with a genomic exon graph. Stephen Tanner, Vineet Bafna. Poster presented at the 2006 ASMS conference in Seattle.

Unrestrictive identification of post-translational modifications through peptide mass spectrometry. Stephen Tanner, Pavel A. Pevzner, Vineet Bafna, 2006. Nature Protocols 1(1),67-72.

Age-Related Changes in Human Crystallins Determined from Comparative Analysis of Post-Translational Modifications in Young and Aged Lens: Does Deamidation Contribute to Crystallin Insolubility? P.A. Wilmarth1, S. Tanner, S. Dasari, M.A. Riviere, V. Bafna, P.A. Pevzner, and L.L David, 2006. Journal of Proteome Research 5(10), 2554--2566.

Identification of Post-translational Modifications via Blind Search of Mass-Spectra. Dekel Tsur, Stephen Tanner, Ebrahim Zandi, Vineet Bafna, Pavel A. Pevzner. Nature Biotechnology 23, 1562-2567 (01 Dec 2005).

Identification of Post-translational Modifications via Blind Search of Mass-Spectra. Dekel Tsur, Stephen Tanner, Ebrahim Zandi, Vineet Bafna, and Pavel A. Pevzner IEEE Computer Society Bioinformatics Conference (CSB) 2005.

S. Tanner, H. Shu, A. Frank, L.. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna. Inspect: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. Anal. Chem., 77(14):4626–4639, 2005.

Peptide Sequence Tags for Fast Database Search in Mass Spectrometry. Ari Frank, Stephen Tanner, and Pavel Pevzner. Conference on Research in Computational Molecular Biology (RECOMB) 2005.

Frank A, Tanner S, Bafna V, Pevzner P. Peptide sequence tags for fast database search in mass-spectrometry. J Proteome Res. 2005 Jul-Aug;4(4):1287-95.

Fields of Study

Major Field: Bioinformatics and Systems Biology

Studies in Bioinformatics

Professors Vineet Bafna and Pavel Pevzner

Studies in Plant Biology

Professors Julian Schroeder and Steven Briggs

Studies in Systems Biology

Professor Trey Ideker

ABSTRACT OF THE DISSERTATION

Efficient and Accurate Bioinformatics Algorithms for Peptide Mass Spectrometry

by

Stephen Will Tanner

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2007

Professor Vineet Bafna, Chair

Professor Julian Schroeder, Co-Chair

Peptide tandem mass spectrometry has emerged as a key technology to detect and measure proteins in biological systems. A core problem is the interpretation of tandem mass spectra. These spectrum annotations are then used to study post-translational modifications, disease biomarkers, protein-protein interactions, and subcellular localization. Technological breakthroughs have led to the generation of ever-increasing volumes of data. Experiments generating tens of millions of spectra are routine, and require efficient algorithms to be effectively analyzed. Filters using sequence tags, as implemented in the InsPecT software toolkit, allow spectra to be rapidly searched against a large proteomics database. The MS-Alignment algorithm

addresses the still more challenging problem of interpreting mass spectra in the presence of unanticipated modifications. A key consideration is the efficient handling of large data volumes without the need for manual intervention.

In any high-throughput biological experiment, calculation of a false discovery rate is essential. The use of a decoy database of shuffled proteins is emerging as a key method for measuring false discovery rates. In addition, decoy database allows a direct comparison of the quality of results from different search parameters, instrument settings, or software tools. We adopt a principled approach to correcting or filtering spurious annotations and experimental artifacts. A key idea is the focus on error rate, not at the level of individual spectra, but at the level of distinct peptides, or modification sites. Results at this higher level can be made more accurate by integrating data across mass spectra.

Additional research is presented on the analysis of RNA microarrays. Here the goal is the identification of gene sets - such as members of a pathway - which are up- or down-regulated. Here the fundamental data is differential transcription levels, as measured by a t statistic. As with mass spectra, leveraging separate measurements (expression levels across many genes) improves accuracy. And computing the false positive rate in a principled way, with an appropriate null model, is vital for computing valid p-values.

Chapter 1: Introduction: High-throughput analysis of mass spectra

Mass spectrometry is the tool of choice for protein identification, fueled by ongoing improvements in instrumentation, as well as improved availability of software for interpreting mass spectra [Aebersold 2003]. This dissertation focuses primarily on the interpretation of tandem mass spectra. Such spectra are formed by proteolytically digesting proteins (typically with trypsin), isolating short peptides (through 1D or 2D liquid chromatography followed by a first round of mass spectrometry), then fragmenting the short peptides to form N- and C-terminal ions. The mass spectrometer separates these ions by their mass-to-charge ratio, and measures their masses to within half a Dalton or less. Each distinct molecular species produces one peak in the resulting mass spectrum.

Interpreting tandem mass spectra is a central problem in peptide mass spectrometry. A pure *de novo* approach as implemented by PepNovo [Frank 2005] or SHERENGA [Dancik 1999], attempts to reconstruct the source peptide from these peaks alone. Database search, as implemented in early tools Sequest [Eng 1994] and Mascot [Perkins 1999], compares the spectrum with each candidate peptide from a protein sequence database such as Swiss-Prot [Boeckmann 2003]. Both approaches have advantages - a database is necessary to distinguish between possible peptides with similar scores, but considering every database peptide is extremely expensive. I developed the InsPecT tool [Tanner 2005], implementing a hybrid strategy which combines

local de novo reconstruction (tag generation) with database search. Tag-based filtering allows InsPecT to run much faster than Sequest, even when allowing non-tryptic peptides and/or considering a list of user-specified modifications. Another algorithm from our group, MS-Alignment [Tsur 2006], performs an unrestrictive ("blind") search of modifications up to a given size.

The sequencing of the human genome [Venter 2001, IHGSC 2004] marked the transition of biology into a new high-throughput era. New technologies such as microarrays and high-throughput mass spectrometry generate unprecedented volumes of biological data. Analysis of this data is a daunting task. New computational methods are often required simply to cope with the volumes of data available. And accurate analysis of data which overwhelms manual curation requires careful statistical modeling. The field of mass spectrometry has suffered acutely from both of these problems. My dissertation will pay particular attention to these two problems - the removal of computational bottlenecks, and maintaining accuracy in high-throughput studies - in the context of peptide mass spectrometry.

The problem of computational bottlenecks becomes obvious whenever one confronts a data-set much too large to be analyzed by hand. Data volumes have reached the scale which not only defies manual analysis, but also overwhelms naive algorithms. For example: In the past, mass spectra were typically analyzed by a trained operator. Early search tools such as Sequest greatly accelerated the process of interpreting peptide mass spectra.

However, newer mass spectrometers can generate a million mass spectra in a day. These instruments generate data much faster than legacy software tools can analyze it. Newer algorithms, such as tag-based filtering, and spectrum clustering - are necessary in order for data analysis to keep up with data generation.

Some examples will illustrate the problems of scale. When I started work in mass spectrometry, I began with a "large" corpus of 40,000 tandem mass spectra from 22 runs on an LCQ instrument [Keller 2002]. The newer LTQ model generates as many as 200,000 spectra from a single run - and often dozens or hundreds of runs must be analyzed together. Several of our recent studies included over ten million spectra. Simply storing the data required the purchase of additional hard drives which were quickly filled to capacity. All this analysis (particularly unrestrictive search) requires computational power: I have used well over 20 CPU-years during my research, and other users of my software have used many more. Parallel computing raises its own set of logistical challenges. Far from being a transparent "grid", compute clusters require careful oversight to coordinate transfers of large files, rescue crashed or hung jobs, etc. The biggest payoff of the proteomics website currently in development may be in saving the time spent herding jobs through the clusters.

Large-scale analyses also exacerbate software engineering problems such as memory management and error reporting. For instance, suppose we

find that one sub-job of a large search generated no output. This could result from a core dump, a crash due to a memory leak, loss of power to a compute node, or a large scan range that simply has no tandem spectra in it. All of these exigencies, and many more, have arisen in our lab while analyzing large data-sets. Troubleshooting each failure by hand quickly becomes unfeasible, and high-level error reporting is necessary.

The problem of accuracy stems from multiple hypothesis testing. Suppose we were to look for up-regulated genes in a microarray that measures 30,000 genes. If we take a gene-level false positive rate of 0.05, we expect to see 1,500 false positives. Similar problems arise in peptide mass spectrometry. Simply applying the Bonferroni correction to lower the p-value cutoff may wipe out all the significant results. In addition, in mass spectrometry, a low false discovery rate at the spectrum level may not be sufficient (or necessary) to ensure a low false discovery rate in the list of peptides, proteins, or post-translational modifications identified. By applying more sophisticated statistical models, or making use of machine learning methods, it is possible to improve sensitivity while retaining high selectivity. A key point here is that large data volumes are a blessing in disguise, since redundant measurements (e.g. spectra from the same peptide) can be used to improve accuracy. The need to carefully judge the accuracy of high-throughput experiments is becoming widely acknowledged: Just as graphs of

experimental measurements should include error bars, so too should every list of high-throughput discoveries quantify the false discovery rate.

Chapter 2 covers InsPecT's tag-based search in detail. We note that peptides are much more likely to break toward the middle of their sequence than on the edges. The theoretical peaks corresponding to breaks near the edges of a peptide are not typically distinguishable from noise. Thus, a full *de novo* reconstruction of a peptide is often unfeasible. The cleanest, strongest peaks of spectra tend to lie in the central mass range. The central mass range of a spectrum generally features several clear "rungs" of the b and y peak ladders. This makes tag generation feasible for medium-quality spectra - particularly when several candidate tags are allowed for each spectrum. There is some overlap between chapters 2 through 4 and the original InsPecT paper [Tanner 2005]. However, InsPecT's tag generation has been through multiple rounds of revisions since the initial paper, and now uses a more sophisticated scoring model.

Chapter 3 discusses spectrum scoring. This can be seen as several problems in statistical modeling at ever-increasing scales. First, we ask whether a particular mass represents a break in our peptide. Moving up one level, we ask whether a peptide annotation adequately explains a complete spectrum. Moving up still further, we consider the validity of peptide annotations (involving many spectra) or protein annotations (involving many peptides). Like most problems related to machine learning, obtaining the right

collection of features is a crucial step. Obtaining "cheap" features is also important, since we must score hundreds of masses and hundreds to thousands of peptide candidates while still processing spectra in under one second.

Chapter 4 covers several important steps in spectrum pre-processing. These steps handle unusual file formats, clean up dirty data, and lay the groundwork for the core search algorithms. These are unglamorous steps that are often glossed over, but they have a dramatic impact. Some algorithms are O(n) in the number of spectrum peaks, so filtering out noise peaks can more than double the speed of tag generation. Parent masses as reported by ion trap instruments are notoriously inaccurate, and correcting the masses allows us to search with a narrower mass window, which speeds things up greatly. And charge determination approximately halves the time needed to search spectra, since only one charge state needs to be considered.

Chapter 5 presents the MS-Alignment algorithm for identifying post-translational modifications. Hundreds of different chemical modifications, with roles ranging from structure stabilization to signal transduction, are known to occur in nature. In the past, tools were provided which searched for one modification at once, requiring the researcher to guess modification types in advance. InsPecT supports several modifications at once, but speed and accuracy suffer visibly when over a dozen modifications are considered. MS-Alignment is one of the first tools to provide an unrestrictive (or "blind") search

for all modification types at once. This search problem is much more challenging than ordinary search, since the "virtual database" of singly- and doubly-modified peptides is orders of magnitude larger than the sequence database. For efficiency reasons, enumerating all such peptides becomes impossible, and we employ a dynamic programming algorithm to obtain an optimal match. In addition, note that in a typical database search, one peptide candidate is clearly superior to all others - this is the key advantage of database search over de novo interpretation. When modifications are considered, we encounter many "δ-correct" annotations which assign modifications the wrong mass or attachment site, but explain the spectrum peaks as well as (or better than) the true peptide.

Chapter 6 builds upon the initial development of MS-Alignment by presenting an analysis procedure, PTMFinder, which better distinguishes correct from spurious modification sites. In addition, an empirical false discovery rate is computed, through the use of decoy (shuffled) protein records. In addition, PTMFinder is able to integrate known biochemical knowledge into its analysis: All things being equal, a known chemical adduct (such as methylation of lysine) is more likely to be correct than a novel, unexplained mass shift. This allows users to compromise between "unrestrictive" search (with few δ-correct" annotations) and truly "blind" search (which enables discovery of novel chemical events). A key idea is that this procedure integrates data across multiple spectra, across multiple charge

states, and even across multiple overlapping peptides. Techniques (such as PeptideProphet) which go beyond analysis of isolated spectra can wield significantly more statistical power.

Chapter 7 applies peptide mass spectrometry to the problem of genome annotation. Identification of protein-coding regions remains a challenging undertaking: Coding signals are difficult to distinguish from pseudogenes or other noise. Short exons and alternative splicing is particularly challenging. EST data is very valuable in this context, particularly for finding intron boundaries, but the data is incomplete, biased, and noisy. Peptide sequences identified through mass spectrometry provide an orthogonal line of evidence for particular exons and particular splicing patterns. I generated an exon graph which represents putative exons and introns from the human genome, identified both through experimental evidence and prediction algorithms. A vast collection of spectra were then searched against this exon graph, identifying novel exons and splice boundaries within the human genome. Variant alleles for hundreds of coding SNPs were also observed. It is clear from this work that an exon graph represents the proteome much more faithfully and efficiently than a simplistic table of strings. Additional work along these lines is being pursued by Nitin Gupta in prokaryotic genomes, Natalie Castellana in plant genomes, and Melissa Key in the Drosophila genome.

Chapter 8 presents work on analysis of microarray experiments. Early microarray experiments focused on the identification of individual genes that

are up- and down-regulated in response to stimuli, or to unsupervised clustering. Here we bring the analysis up a level to searching for biologically-connected sets of genes that are differentially expressed (gene set query), or by directly comparing entire experiments for relationships (gene vector query). These holistic searches help deal with the high level of noise in microarray data: Often no one gene is differentially regulated to a statistically significant extent, while the up-regulation of a family of genes may be quite clear. Constructing a reasonable null model for these queries is challenging, since related genes are co-regulated (and hence their expression correlates tightly). However, once this is done, even simple algorithms can generate exciting biological findings. This chapter strays somewhat from my focus on mass spectrometry data. However, the two key themes of this dissertation - the challenges of scaling up to very high-throughput analysis, and the need to compute accurate false discovery rates - are stressed here as well.

What does the future hold for peptide mass spectrometry? In the past, pioneering work in algorithms has come very early in the lifetime of each generation of instruments. As early as 1995, papers were published discussing post-translational modifications, and whole-genome searches were attempted [Kuster 2001] even before the underlying genome sequences were complete. However, building robust software tools which reliably solve mass spectrometry problems is an ongoing and iterative process. Some challenging problems - such as how to handle alternative splicing in whole-genome

searches - were not fully articulated until quite recently. Other problems, such as the quest for quantitative biomarkers, or the analysis of cross-linked peptides, have proven to vastly more difficult than they initially appeared.

1.1 Open Problems

Several open problems in peptide mass spectrometry suggest themselves. One is the routine incorporation of tandem mass spectra into genome annotation. Chapter 7 demonstrates the feasibility of the approach. The gene scoring model we employed is simplistic, and a truly first-class gene finder would incorporate MS/MS identifications alongside all the valuable data sources, including ESTs and homology information. In addition, incorporating MS/MS data into a genomic database involves significant effort and curation, which often goes beyond the scope of a single publication. There is often a gap, in both time and completeness, between the making of high-throughput discoveries and their incorporation into databases such as The Arabidopsis Information Resource (TAIR).

Quantification of relative abundance is an important area in mass spectrometry. In particular, comparing post-translational modifications across samples could potentially discover modifications - such as phosphorylation involved in cellular signaling, or histone deacetylation - that are involved in for cellular processes, or serve as disease biomarkers. Several quantitation methods have been introduced. Some incorporate isotopic labels while others do not, and some use MS/MS identifications exclusively while others rely

heavily upon consistent elution times. I had the opportunity to do some work in this area - including a space-efficient implementation of dynamic time warping (DTW) - while working at ActivX. A rigorous comparison of different quantification methods would be a valuable step toward the next generation of quantification technologies.

Environmental proteomics is another potential source of exciting analysis challenges. Peptide spectra acquired from a community of bacterial species (living in sea water, dirt, or an infection) must be analyzed carefully, since in addition to identifying the peptide sequences and their proteins, the list of species (and their relative abundance) must be considered.

Other open problems are being dealt with by other UCSD students. One open problem is that of filtering and clustering spectra, for faster (and potentially more sensitive) search. Another is shotgun protein sequencing through the use of multiple proteases (or a non-specific protease).

1.2 Future Directions

The most significant impacts for peptide mass spectrometry will likely come from two sources. First is improvements in instrumentation, which have already greatly improved the quality of incoming data, and promise to make routine those techniques (such as top-down proteomics) which are barely tractable with today's technology. Second is the integration of analysis software into coherent workflows. The core problem of peptide identification has been addressed many, many times, by methods ranging from linear

programming to finite state automata. Leveraging these spectrum annotations to make biological discoveries is an untidy but extremely important problem. InsPecT represents one tool in a complex toolbox which will continue to evolve. Constructing a set of tools - or a "pipeline" is a challenging problem which typically involves the linking together of data and software from multiple labs and vendors.

Chapter 2: Tag-based search

A peptide sequence tag [Mann 1994] is a short peptide (typically 1-5 residues), together with a prefix mass and suffix mass. In tag-based search, one or more tags are first constructed by performing a partial de novo reconstruction of a peptide from its fragmentation spectrum. The tag sequences are added to a trie [Aho 1976], which can then scan a sequence database in linear time. Once a tag's sequence is matched, tag extension checks that the masses of flanking residues from the database match the tag's flanking masses. If tag extension is successful, the resulting peptide candidate is scored. Filtering a large sequence database with tags greatly decreases the number of peptides which must be scored, possibly at some cost in sensitivity [Tabb 2003] [Sunyaev 2005] [Day 2004] [Bern 2007].

Database filtration is particularly important in the context of post-translational modifications, or of multiple splice isoforms, where enumerating all candidates becomes a very time-consuming problem. In this case the sequence tags are analogous to the seeds used by BLAST [Altschul 1997], as perfect matches within a longer inexact match.

2.1: Tag Generation

InsPecT constructs two PRM nodes for each peak, by treating the peak as both a b fragment and a y fragment. It then adds directed edges between any two PRMs whose masses differ by approximately the mass of an amino acid. The difference between the theoretical difference and the actual edge

length (at most 0.5Da) is termed the skew, K. Tags are then generated by finding paths of length 3 through the resulting directed acyclic graph. The score of a tag is computed as follows:

$$Score = a\sum_i PRM(M_i) + bS(\sum_i |Ki|) + cS'(|\sum_i Ki|)$$

Here PRM(M_i) is the PRM score (as described in Chapter 3) of mass M_i. The functions S and S' penalize tags whose edges have a large skew. Valid tags tend to have significantly smaller skew than spurious tags (Figure 2.1). The edge scoring function is derived from the likelihood-ratio for a given skew cutoff. The weights (*a, b, c*) were trained through multiple runs of the tag generator on the training corpus.

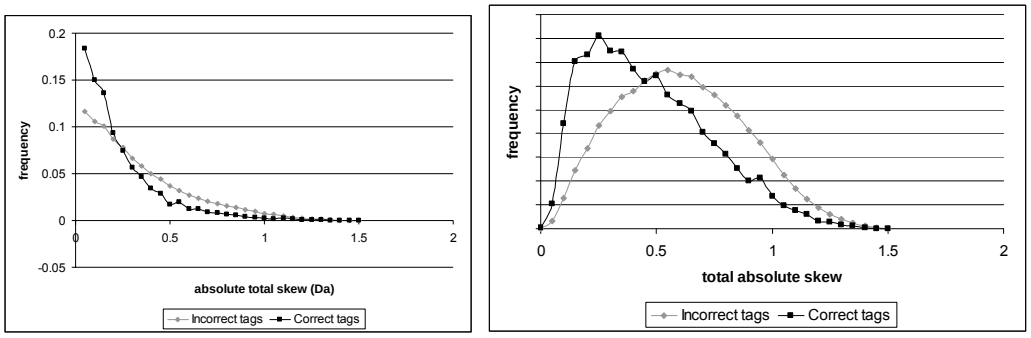


Figure 2.1: Histograms of tag edge skew, for correct and incorrect tags across the training corpus. Valid tags have lower total absolute skew on average (right), and the skews for individual edges tend to cancel out (left).

Several other minor issues arise during tag generation. If a spectrum contains both b and y peaks for a PRM, then two PRM nodes will be generated, generating multiple near-identical tags; these tags must be merged. PRM nodes for the "goalpost" masses, mass 0 and for the parent residue mass, are always added. Amino acids I and L (and, in ion trap data, Q

and K) are indistinguishable by mass, so these amino acids are treated synonymously. The top tags (by default, 25) for each spectrum are retained, in order to accommodate most *de novo* sequencing errors.

2.2 Trie Search

Peptide sequence tags are added to a tree, which allows the search of an arbitrary number of strings against a database in linear time. Because iterating over a large database has a non-trivial cost (particularly if it must be read from the disk), InsPecT constructs one trie for each block of 100 spectra, amortizing the cost of scanning the database. Each leaf of the trie contains a list of one or more associated tags (with associated spectra, and prefix and suffix masses); internal nodes have no tags.

Once a tag's sequence is matched, we consider whether the flanking amino acids can (possibly with modifications added) match the tag's prefix and suffix masses. This **tag extension** step provides much of the filtration efficiency: Assuming an average amino acid mass of 100Da, an arbitrary match's flanking mass has only a 1% chance to match (to the nearest Dalton) the mass from the peptide sequence tag. In order to handle tag extension efficiently, we define a *decoration* to be an allowed collection of post-translational modifications (including the empty decoration). We sort all possible decorations by mass. Tag extension begins by considering the candidate mass provided by the largest possible decoration, and an empty suffix. Suffix extension then proceeds iteratively:

- If the candidate mass is too large, the next smallest decoration is considered. If no more decorations remain, the procedure terminates
- If the candidate mass is too small, the suffix is extended by one letter. If the end of a protein is encountered, the procedure terminates.
- If the candidate mass is correct (within a pre-specified tolerance), extension succeeds.

Prefix extension proceeds similarly. When searching a splice-tolerant database (.ms2db format), extension proceeds by a depth-first search in the exon graph. When both prefix and suffix extension succeed, a candidate peptide is generated. Each candidate peptide is then scored (Chapter 3.3), and the best such candidates are retained and reported.

2.3 UCSD Proteomics Web Server

InsPecT and MS-Alignment are available to researchers through a user-friendly online interface. Dr. Ingolf Krueger of the Service-Oriented Software and Systems Engineering Laboratory (S3EL) directs a group of computer scientists in ongoing development of the web interface. The same flexible, service-oriented infrastructure [Krueger 2006] also supports other proteomics tools, such as PepNovo [Frank 2005] and the Shotgun Protein Sequencer (SPS) [Bandeira 2007]. I drew up the initial requirements documents for the website, and helped guide development over several iterations. This interface will allow UCSD bioinformatics students to pursue many more collaborations with other labs, without having to spend time on the

logistics of transferring files and running routine searches. Figure 2.2 shows the front page of the current website.

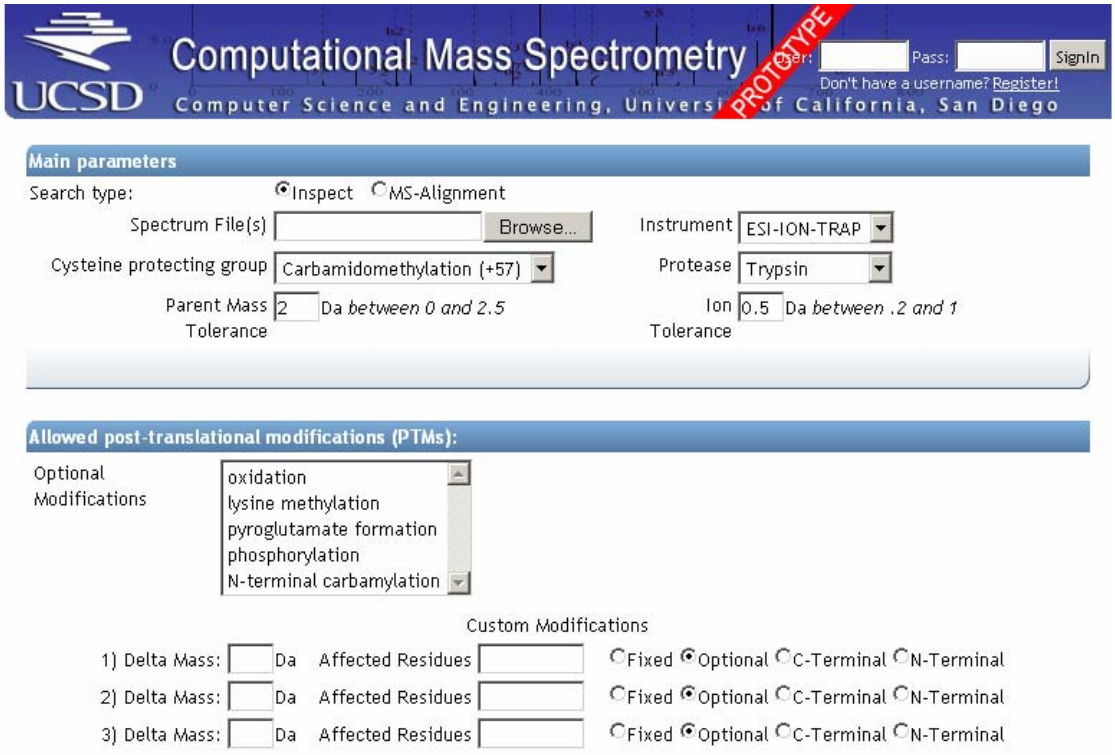


Figure 2.2: Screenshot of the UCSD Proteomics website. Users can upload spectra for analysis, along with protein sequence databases to be searched. The website provides an interface to InsPecT search, MS-Alignment (Chapter 5), and other tools such as PepNovo.

2.4 Comparison with Sequest

InsPecT was both used to search a collection of 770,000 peptide spectra derived from *Arabidopsis thaliana* cell extract against a database of corpus (Chapter 3.1) against a database of 34,892 proteins derived from the IPI database, together with shuffled protein records. Similarly, Huilin Zhao's lab searched the same sample against the same database using the cluster version of Sequest [Keller 2005]. The false discovery rates for each search

were quantified (Chapter 3.4), in order to compare results between tools. At a false discovery rate of 5%, Sequest annotated 81,912 spectra, and 8,051 distinct peptides. At the same 5% false discovery rate, InsPecT successfully annotated 120,903 spectra, and 11,753 distinct peptides. This represents an improvement of 48% more spectra and 46% more peptides. These improvements in accuracy are important for many applications, including relative protein quantitation.

Chapter 3: Spectrum Scoring

Spectrum scoring measures the probability that a partial spectrum interpretation (tag), spectrum annotation, or peptide is correct. There are several subproblems of scoring:

- Scoring a prefix residue mass (PRM), in the context of tagging or unrestrictive search.
- Scoring a cut point. This is the same process as scoring a PRM, except in the context of a tag or peptide, so that one or both flanking amino acids are known.
- Scoring a spectrum annotation. This score considers the scores of each individual cut point, as well as additional features such as the percentage of the top N peaks which are b or y fragments.
- Scoring a peptide. This attaches a p-value to a peptide in the context of a whole mass spectrometry experiment, and may bring together information from many spectra.

One consideration in scoring is that the model should be kept simple, so that we can score hundreds of masses, and thousands of candidate peptides, while processing each spectrum in less than a second. (This is in addition to the pressure from the "curse of dimensionality" to remove unnecessary features). Another challenge is that the output of one scoring model may be input for the next: The match quality scores (MQScores) of spectra are used in validating peptides annotating the spectra. Because of the

interactions between models, and because of anticipated generalization to new instruments (FT, QTOF), an infrastructure to re-train scoring models is essential. Inspect's scoring models can be re-trained programmatically, and since the models are stored in separate files on disk, Inspect need not be recompiled in order to do so.

3.1 Training Corpus

I developed a large corpus of annotated spectra for use in training the scoring functions. To avoid over-fitting to one particular instrument or sample type, these samples come from three sources:

- OMICS2002- A collection of 19,000 spectra derived from a mixture of known proteins, as described in [Keller 2002].
- OMICS2004 - A collection of 7,147 spectra derived from a known mixture of proteins and synthetic peptides, as described in [Purvine 2004].
- HEK293 - A collection of 790,000 spectra derived from whole-cell extract from human embryonic kidney cell line 293, as described in [Tanner 2007].

The OMICS data-sets have the advantage that they come from a known protein mixture, so that filtering of both false positives and false negatives is possible [Hogan 2006]. In each case, annotations were obtained by searching spectra against a large database including decoy proteins, as described in 3.4. Spectrum matches were retained at a false discovery rate of

5% from the OMICS data-sets, and 1% from the (much larger) HEK293 data-set. A total of 2,965 spectra from the OMICS 2002 data-set, 2,078 spectra from the OMICS 2004 data-set, and 196,970 spectra from the HEK293 dataset were obtained. This training corpus is stored in our lab's data repository, so that improvements in algorithm performance can be tracked over time. Because peptide abundance and detectability varies greatly, training scripts process the same peptide species at most five times; this helps avoid over-training on fragmentation properties specific to abundant proteins.

3.2 PRM and Cut Point Scoring

A potential prefix residue mass (PRM) is scored based upon the intensities observed at the theoretical fragment masses for various fragment types. InsPecT scores PRMs for each mass value (with bin size 0.1Da), using a Bayesian network [Finn 2001]. The Scorpion tools are used to build the Bayesian network and train the probabilities. Most of the network nodes correspond to "witness peaks" for a given PRM. There are intensity levels: High, present, and absent. The highest intensity level corresponds to the top N peaks in the spectrum, where N is equal to the parent mass divided by 50 Da. (Intuitively, we expect to see up to two strong peaks per amino acid, where amino acids have average mass ~100Da).

The fragment masses of interest are learned from a corpus of annotated spectra, by constructing an offset frequency histogram (Figure 3.1) [Dancik 1999]. This histogram has peaks corresponding to singly- and doubly-

charged prefix and suffix fragments. However, the histogram contains significant levels of noise. For instance, a y peak preceding an Alanine residue will correctly contribute to the suffix peak count at offset 0, but will also contribute to the counts for offset 71, as well as to counts for prefix peaks. Therefore, we adopt an iterative approach: Once an offset has been selected, we re-generate the histogram, this time disregarding any peaks matching a known offset, re-generating the histogram. In this way, subtle peaks such as b2-H_2O can still be detected. An advantage of this approach is that it can learn new fragment types (e.g. from phosphopeptides). Table 3.1 summarizes the peaks identified in this way.

Table 3.1: Peak types in doubly-charged spectra, discovered through an iterative peak selection procedure. At each stage, the peak type which explains the greatest fraction of peaks and intensity is selected. Often a near-correct mass is selected which picks up intensity from the true peak and an isotope (e.g. y2+0.2). This procedure shows the relative importance of peak types in ion trap spectra from tryptic peptides, with the y series much stronger (and somewhat more abundant) than b peaks.

Peaks	Intensity	Ion type	Error
6.27%	23.99%	y	
5.05%	9.80%	b	-0.1
4.46%	9.30%	y+1	
3.39%	3.56%	b+1	0.1
1.42%	4.68%	y2	0.2
2.24%	2.17%	b-h2o	-0.1
0.84%	2.53%	b2	0.3
1.91%	1.78%	y+1+1	
2.01%	1.68%	b-nh3	-0.1
1.15%	1.35%	a	-0.1
0.74%	1.52%	y2+1	0.2
1.28%	0.83%	y-h2o	
1.30%	0.81%	y-nh3	
1.25%	0.80%	b+1+1	-0.1
0.72%	0.88%	y	0.5
0.60%	0.91%	b2-h2o	0.2
0.88%	0.74%	b	0.4
0.60%	0.83%	y2-nh3	0.3
0.53%	0.77%	b2-nh3	0.2

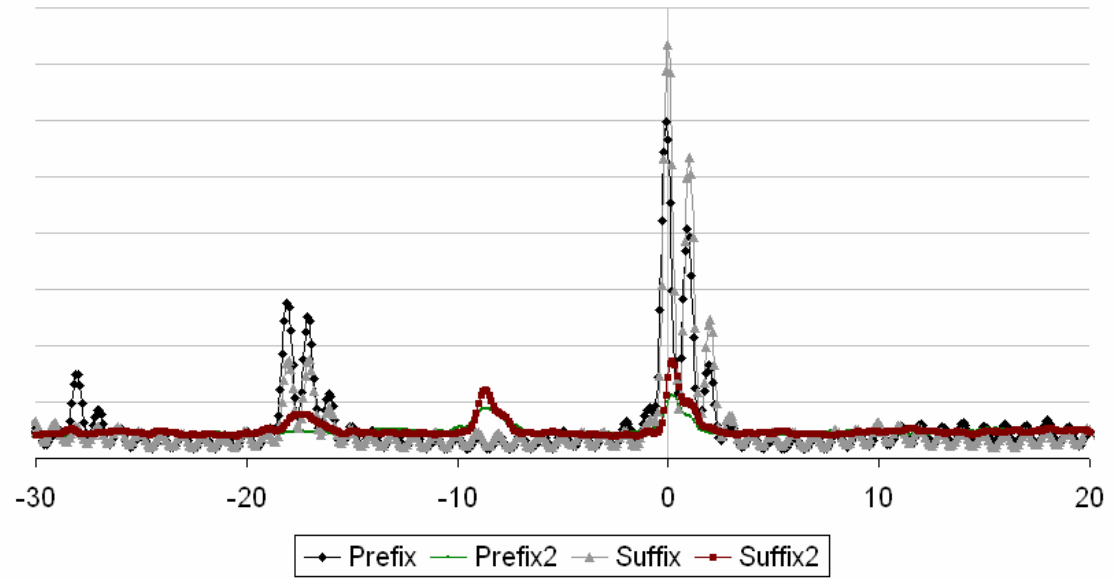


Figure 3.1:Offset frequency function for the four ion series (prefix and suffix series, singly- and doubly-charged). Minor ion types can be detected most easily in subsequent iterations, after noise peaks has been filtered.

I investigated the relationships between various fragment types. For instance, it is very rare to observe a strong y-H_2O peak (corresponding to breakage of two bonds) without a corresponding y peak (corresponding to the more common breakage of a one of the two bonds). I computed the mutual information between each pair of Bayesian nodes. This measurement provides a list of the most informative nodes. Formally, given two nodes X and Y, we compute the normalized mutual information MI(X, Y):

$$MI(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{H(X)}$$

$$H(X,Y) = -\sum_{x,y} p(x,y)\ln(p(x,y))$$

$$H(X) = -\sum_{x} p(x)\ln(p(x))$$

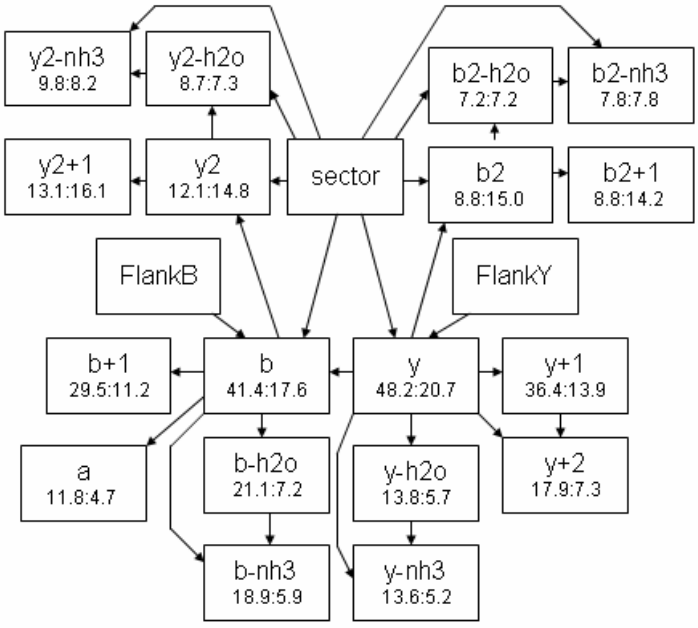


Figure 3.2: Bayesian network for scoring cut points. In nodes corresponding to peaks, the odds that a peak is present (in a charge-2 or a charge-3) spectrum are indicated. The PRM scoring network excludes the "FlankB" and "FlankY" nodes.

I constructed a Bayesian network, using mutual information as a guide (Figure 3.2). The sector (low or high mass range) and flanking amino acids (see Table 3.2) influence fragmentation. Adjustable parameters (such as the number of sectors and the intensity level cutoffs) were tuned by running multiple rounds of tag generation. The output of this scoring model is required for tag generation (Chapter 2), as well as unrestrictive search (Chapter 5). Ongoing research by Samuel Payne applies these same tools to discover the fragment types (and relationships between them) for phosphopeptides.

Table 3.2: Strong flanking amino acid effects, derived from the training corpus. The effects on the C-terminal ion series (left) are not identical to the effects on the N-terminal series (right).

flank (y)	**effect**	**flank (b)**	**effect**
P suffix	enhancer	P suffix	enhancer
K,R suffix	strong suppressor	P,G prefix	strong suppressor
H suffix, P prefix	weak suppressor	H,R suffix	weak suppressor

3.3 Annotation Scoring

Scoring the match between a tandem mass spectrum and a candidate peptide is a key step in database search. This score has two roles. First, it's used to order the candidate peptides, so that the best available match is the first (or only) match reported. Secondly, it's used to order all annotated spectra from high to low confidence. In practice, this second problem is much harder. Typical searches return at most a hundred candidates, which can be compared in a straightforward way. On the other hand, searching a run may generate millions of spectrum annotations, and the possibility for systematic bias exists (e.g. unduly favoring short matches or charge-2 matches).

I tested a wide range of features for spectrum matches. The most informative single feature found (as measured by ROC curve area) is the sum of the PRM scores for all PRMs in a match. This feature is additive, and serves as the scoring function used in the dynamic programming table of the MS-Alignment algorithm. A total of 32 features were considered, including the following:

- Peptide length: Raw or log-scaled.
- PRM score. The total, mean, and median were computed. Scores were computed by excluding 0, 1, or 2 cuts from the peptide termini (where fragmentation is poor).
- Fraction of theoretical b (or y) peaks present
- Fraction of top N peaks which are b or y ions, where N is proportional to parent mass.
- Fraction of total intensity contained in the b series, y series, or both
- Number of tryptic termini (0, 1, or 2)
- Parent mass difference between the spectrum and the candidate peptide
- Self-convolution of b and y peaks, as computed during parent mass correction

Because these features are so similar, using all of them is undesirable. This is true for efficiency reasons (the scoring function must be evaluated millions of times), and because adding too many features can lead to an overtrained model with lower accuracy in practice (the "curse of dimensionality"). Therefore, I performed iterative additive feature selection. At each iteration, the feature which produced the greatest improvement in classification accuracy was added. The procedure ended when the next feature would cause a negligible improvement in accuracy. A total of seven

features were retained for use in the final model. The output of this model is the match quality score (MQScore).

A final feature which is very valuable in distinguishing true from false matches is the delta-score, the difference in score between the top match and its runner-up. The delta-score reflects the degree to which the top spectrum annotation outscores matches achievable by chance. This feature must be computed in the context of a search, after all candidates have been found and ranked by MQScore. A consideration when searching with modifications is that this delta-score should exclude overlapping peptides. For instance, the score difference between "M+16MACDEFGK" and "MM+16ACDEFGK" will be very small (since the theoretical spectra are almost identical). Following [Keller 2002], we take the weighted sum of MQScore and delta-score, and denote this quantity as the f-score for a spectrum match. The empirical distribution of f-scores is fit closely by a mixture model consisting of a gamma distribution (for spurious matches) and a normal distribution (for valid matches).

3.4 Quantifying false discovery rates

In practice, searching a database containing "decoy" records (shuffled protein sequences) is a particularly effective way to quantify the false discovery rate [Higdon 2005, Higgs 2007, Fenyo 2007, Elias 2007]. Normally, InsPecT searches a database containing a 1:1 mixture of valid and shuffled proteins. We compute false discovery rates by examining the collection of hits

above an f-score cutoff. Intuitively: All hits to shuffled proteins are invalid, and for each invalid hit to a shuffled protein, we expect to see one chance hit to a valid protein. Given a collection of X hits to valid proteins and Y hits to invalid proteins, the false discovery rate is simply $\min(1.0, \frac{Y}{X})$. This false discovery rate is reported as the p-value for a match. Separate p-value curves are computed for singly- and doubly-charged spectra, and for triply-charged spectra. Figure 3.3 provides an example of the resulting distribution. Note that the p-value of a match depends on the overall distribution of scores - some mass spectrometry runs generate noisier data (or undergo less filtering at the instrument level), and so matches have a much lower prior probability of being correct.

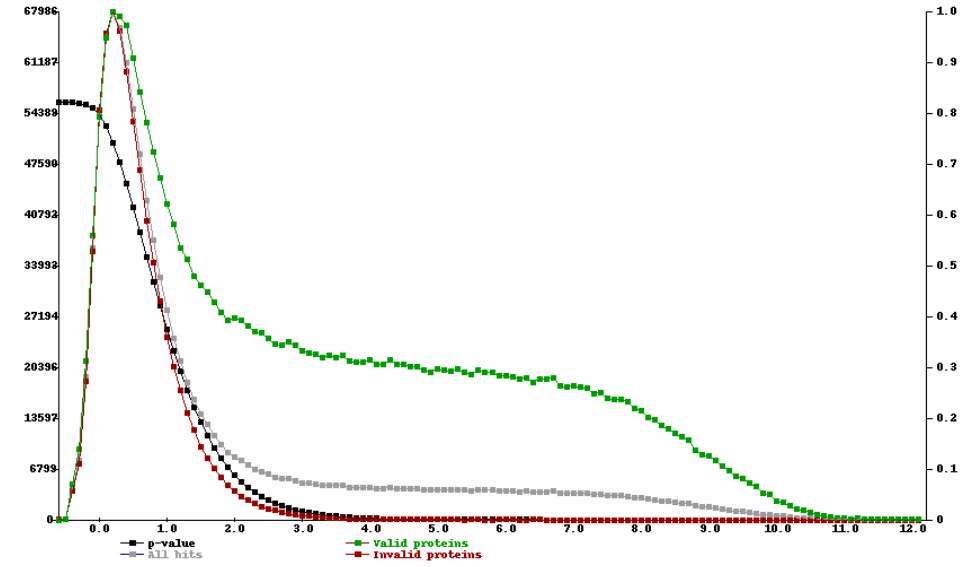


Figure 3.3: Distribution of F-scores for all hits, for hits to valid proteins, and for hits to invalid proteins. This search ran against a 1:1 mixture of valid and decoy proteins, so many chance matches to valid proteins (a strong left "shoulder" of the F-score distribution). This empirical histogram provides us with a false discovery rate for any given F-score cutoff.

As has been noted [Nesvizhskii 2005], a low false discovery rate at the spectrum level does not necessarily correspond to a low false discovery rate at the peptide level, or at the level of proteins. I tested InsPecT's unmodified search on 1 million spectra derived from whole-cell extract of Arabidopsis thaliana. The database was a 1:1 mix of shuffled and valid proteins, including common contaminants (trypsin, and human keratins). Using an F-score cutoff of 0.8, a total of 6,450 peptides and 2,319 proteins were identified at a 5% false discovery rate. We note that peptides with multiple spectra are more likely to be correct (although these additional spectra are not independent events, since spectrum annotation is consistent even when it is incorrect). We compute several features for each distinct peptide species, including the F-score, the MQScore score of a consensus spectrum (see Chapter 6), and the spectrum count. Applying LDA to these features, we obtain 7,057 peptides and at a 5% false discovery rate, an increase of 10%.

Chapter 4: Spectrum pre-processing

This chapter discusses several steps which prepare spectrum for scoring. These include file format conversions, filtering unwanted noise from spectra, and improving measurements of charge and mass of the precursor peptide. These steps occur before tag generation and database search.

4.1 File conversion

Most mass spectrometer vendors use a proprietary binary file format to store peak and intensity information. However, as open-source tools such as InsPecT become more important, and as scientific collaborations between labs increase in scale and complexity, it is best to put spectra in a common file format which can be handled by everyone. This goal has not been achieved yet, though the mzXML file format (and its competitor, mzData) represents a significant step in that direction. The InsPecT toolkit includes scripts to parse mzXML and mzData files (using the expat XML parser), as well as text-based formats such as .dta, .mgf, and .pkl format. Scripts are also provided to convert between file formats, for interfacing with third-party analysis tools.

4.2 Peak filtering

A typical mass spectrum consists of few "trees" (intense peaks) standing above the "grass" (dozens or hundreds of low-intensity peaks). Filtering out the noise peaks speeds up the search, which is O(n) in the number of peaks. In addition, filtration eliminates some of the variability in

peak counts and signal-to-noise ratio, which makes scoring model more relevant across spectra. InsPecT uses a simple window-based filter: A peak is retained if it is one of the top 6 peaks in the window of radius 50Da, centered around itself. A window-based filter is preferable to simply taking the top N peaks from the entire spectrum because the overall intensity varies over the spectrum mass range, with a peak near the center.

4.3 Parent mass correction

Mass spectra from ion trap instruments typically have a mass accuracy on the order of 50 parts per million (ppm). Thus, a peak at 1000Da may be incorrect by up to 0.5Da. This is problematic for peaks in MS1 scans, which may be doubly or triply charged, so that an initial mass error of 0.5 (in Thompsons) translates into a reported parent mass which is incorrect by up to 1.5Da. An Orbitrap or FT instrument typically features greatly improved mass accuracy (5ppm or less), but often reports incorrect parent masses due to selection of +1 or +2 isotopic peaks.

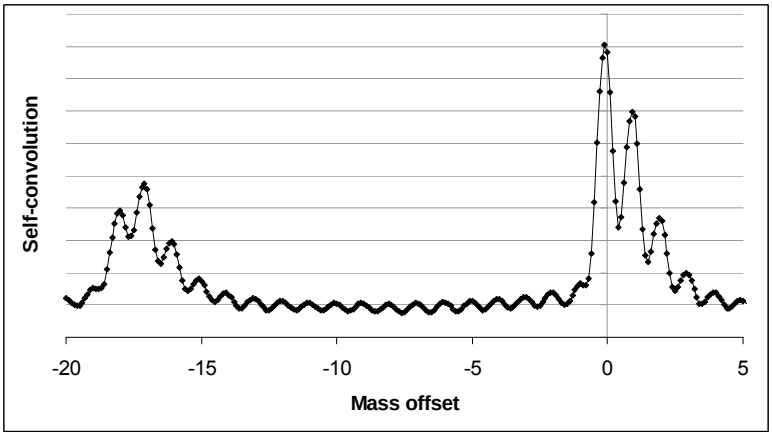
An artifact of dynamic exclusion also gives rise to large parent mass errors. After a peptide species is selected for fragmentation, the instrument excludes the precursor m/z as off-limits for fragmentation for an operator-specified period of time (typically one minute). However, an abundant peptide may have a broad elution peak, whose "shoulders" extend beyond the forbidden m/z range. In this case, the instrument may select the same peptide

(possibly isotopically-enriched) for fragmentation, with an incorrect reported parent mass.

Any improvement in parent mass accuracy is valuable, since it greatly improves the efficiency of peptide identification. For instance, decreasing the radius of the parent mass window from 0.5Da to 0.1Da decreases the number of peptides to be scored by a factor of five. This speeds up the overall search significantly, since over half of the running time is spent in scoring. In practice, masses can generally be corrected by computing the self-convolution of the fragmentation spectrum. For a doubly-charged peptide, we expect to observe pairs of peaks (corresponding to b and y ions) whose masses sum to the (doubly-charged) precursor mass. Similar patterns are expected for neutral loss peaks. Similarly, given a triply-charged precursor, we expect to see pairs of peaks of the form (x, y) where x + 2y equals the parent mass. Formally, we compute a table of binned intensities I, and then compute the b,y self-convolution SC and triply-charged self-convolution SC2 for offset δ. For example, the value SC(δ) measures the overlap between b and y-H_2O ions, and between y and b-H_2O ions.

$$SC(\delta) = \sum_{i=0}^{PM} I(i)I(PM - i + \delta)$$

$$SC2(\delta) = \sum_{i=0}^{PM} I(i)I\frac{(PM - i + \delta)}{2}$$



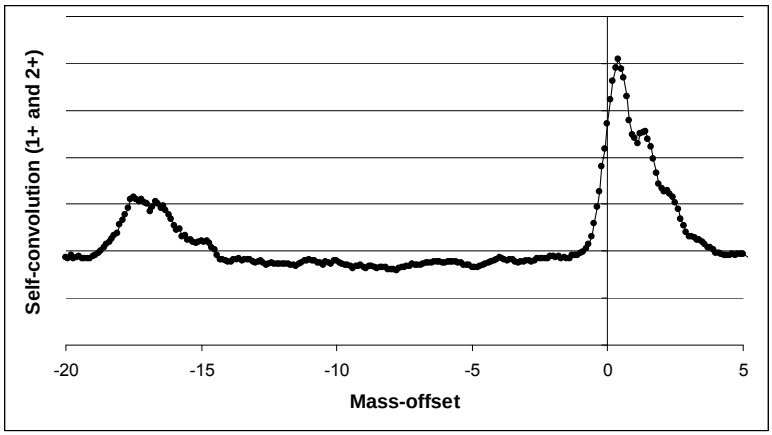


Figure 4.1 (a) Average self-convolution results across 3,000 doubly-charged spectra from the training corpus. The strongest self-convolution (at 0) is produced by b and y ions from the same breakpoint. Other peaks are seen at +1 and +2 (due to isotopic peaks), and at -17 and -18 (due to neutral losses of ammonia and of water). Intuitively, parent mass correction adjusts the parent mass so as to maximize the b,y-convolution. (b) Self-convolution for singly- and doubly-charged peaks from a corpus of 3,000 triply-charged spectra. Peaks are shifted to the right from their theoretical locations due to the presence of +1 isotopic peaks.

To perform parent mass correction, InsPecT considers **candidate** parent masses in a neighborhood near the input mass M, up to a ppm tolerance. The masses M-1 and M+1 are always considered, regardless of the mass tolerance; even on QTOF and FT instruments, mass errors of exactly 1Da are not uncommon. For each mass, InsPecT calculates several

features. The first feature is the difference between the initial mass and the candidate mass. (We take the absolute difference for charge-1 spectra, where there is no noticeable bias in parent mass errors) The b,y-convolution values, SC(δ), are computed for several mass offsets δ: -18Da, -17Da, 0Da, 1Da, 0.5Da, -16.5Da. Each self-convolution is scaled by the spectrum intensity, so that they fall within the range [0, 1]. In addition, the ratio of each self-convolution to the average of the six values is computed; this feature rewards candidate masses where valid mass offsets (particularly 0Da) have much better self-convolution values than invalid mass offsets (0.5Da and -16.5Da). Similarly, values SC2(δ) and ratios are computed for valid (0.4Da, 1.2Da, -17.5) and invalid (-1, 4) offsets.

InsPecT uses a linear discriminant analysis (LDA) model to score each candidate parent mass, and retains the best one or two masses for its search. The training set was derived high-confidence annotations (empirical p-value 0.05) from several searches. Table 4.1 summarizes the results of parent mass correction. A support vector machine (SVM) approach was tried as well [Chang 2001]. However, for this application, LDA was slightly more accurate as well as 25% faster.

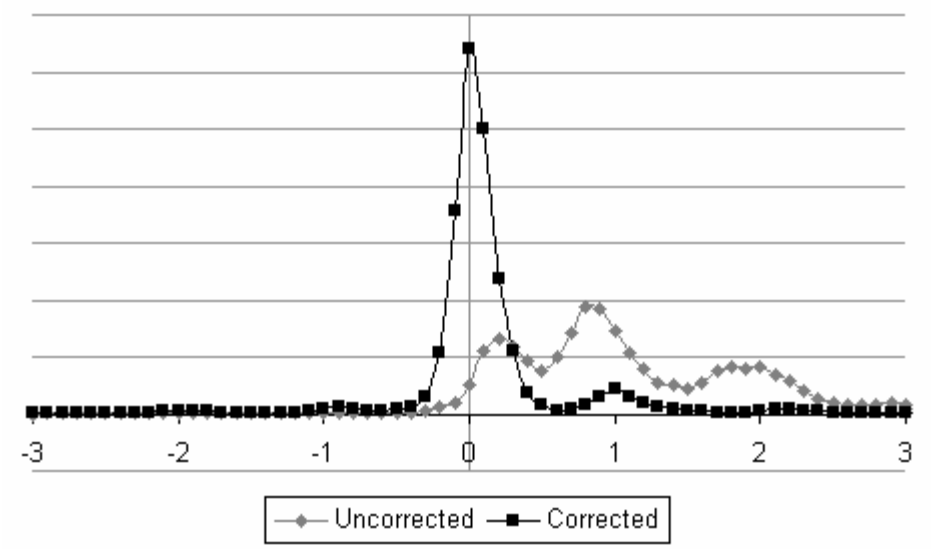


Figure 4.2 Histogram showing the results of parent mass correction on the training corpus. Reported masses were systematically incorrect (with three peaks corresponding to charge 1, 2 and 3); masses are generally corrected within 0.3Da of the true parent mass.

Table 4.1: Results of parent mass correction on the training corpus spectra. Each cell reports the percentage of spectra for which the spectrum's parent mass was within 0.3Da of the peptide's true mass.

	No correction	Correction	Correction and runner-up
Charge 1	60.6	65.2	75.2
Charge 2	9.0	84.9	93.1
Charge 3	2.4	56.0	68.9

4.4 Charge State Determination

Often the charge state of a tandem mass spectrum is unknown. Distinguishing singly-charged spectra is relatively straightforward, but distinguishing between doubly- and triply-charged spectra from ion trap instruments is nontrivial. (We ignore spectra of charge 4 or higher, since such spectra are both rare and very difficult to interpret in LTQ data). A typical approach, employed by tools such as Sequest, is to simply search multiply charged spectra in both charge states, then retain the best match. This has

the disadvantage that it essentially doubles the search time. Thus, it is desirable to determine a charge state before searching a spectrum [Klammer 2005].

InsPecT uses two models for charge state determination. The first distinguishes between singly- and multiply-charged spectra, the second distinguishes between doubly- and triply-charged spectra. When computing these features, we divide the mass window into three sections: Low (from 0 to the precursor m/z, M), medium (from M to 2M) and high (from 2M to 3M). Note that singly-charged spectra should have all intensity in the low region, and doubly-charged spectra should have all intensity in the low and medium regions. The features used to distinguish between charge-2 and charge-3 spectra are as follows:

- Fraction of intensity in each mass region.
- For δ in {-18, -17, 0, 1}, we compute the difference between the self-convolutions SC(δ) for charge 2 and SC(δ) for charge 3.
- For δ in {0.4, 1.2}, we compute the difference between the self-convolutions SC2(δ) for charge 2 and SC2(δ) for charge 3.

Intuitively, these features capture the property that self-convolutions will be much higher for the correct charge state (where b,y-overlap is common) than for the incorrect charge state (where such will occur only by chance). A support vector machine achieves a clear separation (accuracy >99%) between singly- and multiply-charged spectra. The separation between doubly- and

triply-charged spectra is not as complete (accuracy ~95%). Therefore, spectra with borderline scores are searched in both charge states. Since only ~10% of spectra fall in this region, a significant speed savings is still achieved by charge state determination.

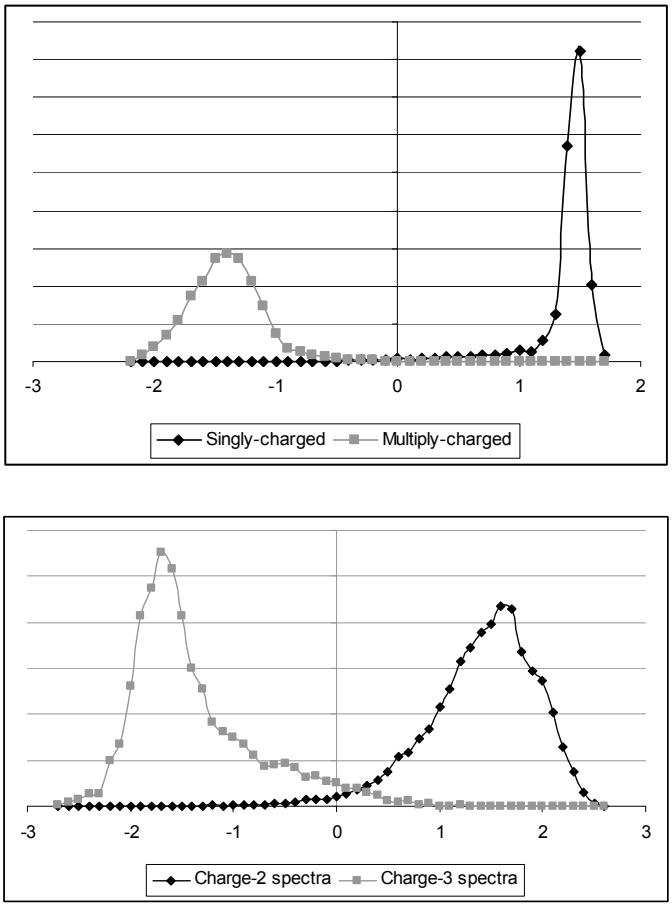


Figure 4.3: Histograms of charge state determination SVM scores. Singly-charged spectra are separated easily. For multiply-charged spectra, values from the middle of the distribution (from -0.5 to 0.5) are searched in both charge states.

Chapter 5: Unrestrictive modification search

5.1 Introduction

Post-translational modifications (PTMs) greatly increase the complexity of the proteome and identifying PTMs is undoubtedly the next big step for proteomics [Shu 2004, Cantin 2004]. However, reliable computational identification of PTMs remains a formidable challenge. The first approach to PTM identification was proposed by Yates et al. [Yates 1995], who advocated the enumeration and scoring of all possible modifications for each peptide from the database. This exhaustive search approach has serious limitations since it can only take into account a few modifications and is prohibitively slow for mutation detection. In other words, a researcher has to "guess" in advance which PTMs are present in the sample. As a result, the current practice is to perform a *restrictive* search for a small set of PTMs and ignore all other PTMs. The question arises whether one can design an *unrestrictive* PTM search algorithm that searches for *all* types of PTMs at once in a *blind* mode, i.e., without knowing which PTMs exist in a sample. Another, more ambitious question is whether one can predict PTMs that are not known yet by mining large MS/MS datasets, something that was never done before. Additionally, the blind PTM identification approach opens a possibility to study the extent and frequency of different types of PTMs, still an open problem in proteomics.

The first blind approach to PTM identification (*spectral alignment*) was proposed by Pevzner et al. [Pevzner 2000, Pevzner 2001]. Recently, Searle et

al. (OpenSea) [Searle 2004] and Han et al. (SPIDER) [Han 2004] proposed yet another approach to blind PTM identification. In contrast to spectral alignment, these approaches rely on de novo interpretation of MS/MS spectra. For example, Han et al. formulate the problem as the identification of a modified peptide that best matches both the de novo interpretation and the database peptide. While this elegant formulation accommodates some de novo sequencing errors, the approach depends critically on a good de novo interpretation. We emphasize the important difference between the spectral alignment approach and these other approaches. Searle et al. use a heuristic branch and bound technique that, in contrast with spectral alignment, (i) does not guarantee the optimal solution and (ii) crucially depends on the quality of de novo reconstruction. Han et al. use a rigorous dynamic programming algorithm (that is similar to spectral alignment in case there is no sequencing errors) but only compare a database peptide against a *single* de novo interpretation of an experimental spectrum. Spectral alignment, on the other hand, compares a database peptide against *every* possible interpretation of an experimental spectrum, thus eliminating dependence on accurate de novo interpretations. In this paper, we describe a new blind approach to PTM identification that extends the spectral alignment approach (code available at http://peptide.ucsd.edu).

Recently, Hansen et al. [Hansen 2005] and Tang et al. [Tang 2005] studied the problem of interpreting peptides with a single modification. In this

case, the mass shift is known in advance and the edit distance is 1 thus allowing one to substitute the dynamic programming with an exhaustive search that analyzes every possible PTM position. We remark that the time complexity of exhaustive search for a single modification is quadratic in the length of the peptide (for each peptide in the database), while spectral alignment can be implemented in linear time (see 5.2).

Identification of all types of PTMs present in a large collection of MS/MS spectra is a difficult task. An even more difficult task is to distinguish between real PTMs and computational artifacts. Using our approach we were able, for the first time, to construct a *PTM frequency matrix* PTM(δ,*a*) that reflects the number of MS/MS spectra in a sample with predicted PTM δ on amino acid *a* for all possible shifts δ and all amino acids *a*. For example, Figure 5.1 represents a fragment of the PTM frequency matrix of IKKb dataset (below). Known chemical modifications dominate the large entries and therefore suggest that PTM frequency matrix is a gateway to the study of modified peptides and mining for still unknown PTM types. PTM frequency matrices represent a new approach to validation that compares *PTM*(δ,*a*) with shift δ at amino acid *a* with *PTM*(δ,*y*) for a “random" shift σ at a “random" amino acid *y*. High values in PTM frequency matrix may point to some still unknown modifications and we provide multiple supporting evidence that they indeed may correspond to previously unknown PTMs rather than artifacts.

a

Offset	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	(N)	(C)	Total
−18		1	19*	10	1*		5	3	1	1		3	4	4	9	9	56	1			44	20	127
1	12		24	24*	17	40	17	34*	4*	77*	2	191*	43	29*	3	19	30*	13*	11	23	4	2	613
14	6			3				4*	93		40		1			7*		12*			90	49	166
16	1*		12*	26	1*	11	2	39*		24*	613*	4*	1*	9		1*		1*	84	1	62	81	830
17			1*	1		3		3*		8*	71*	10*		8		1		1	17	1	1	3	125
22	7		11*	18	8	24	5	16		9	1	11	1	18		16	17	18		8	1	1	188
28	1*	2*						9	154*	21	1					30*		42*			157	33	260
32		2						1			224*	1	1*		5				20		186	1	254
53			36	41		4		17	1	12	1	8	14	3		1	2	5		2			147
71		297								1	1		1								14	1	300
156	2							12		77			6			3	1				85		101

b

AA	Mass	Spectra	Sites	Putative annotation
M,W	16	803	50	Oxidation
C	71	332	9	PAM-cys
M,W	32	248	15	Double oxidation
N	1	225	36	Deamidation
K	28	184	3	Dimethylation
Nonspecific	22	176	64	Sodium
K,M	14	154	4	Methylation
E,D,...	53	130	28	Fe(III)
T,E,D	−18	117	16	Dehydration
V	28	56	2	Dimethylation
K	−57	46	1	Mutation to alanine
S	28	30	1†	Mutation to aspartate
M	38	23	4	Potassium
C	76	22	3	Beta-mercaptoethanol
L	34	17	3	Mutation to phenylalanine
M	−2	21	3	Mutation to glutamate
Nonspecific	1	355	133	Isotopic peaks
L	156	92	1	Truncated K+28
I	16	49	3	Misplaced oxidation
L	17	27	6	Misplaced oxidation
Nonspecific	2	22	17	Isotopic peaks
I	44	20	2	Misplaced K+28 and M+16
L	54	19	7	Shadow of Fe(III)
V	15	18	2*	Mutation to asparagine
M	−32	15	5*	Mutation to valine

c

AA	Mass	Spectra	Sites	Putative annotation
Q,N	1	499	100	Deamidation
N terminus	42	446	12†	Acetylation
N terminus	43	308	98	Carbamylation
Nonspecific	38	229	102	Potassium
C	14	160	14	Methylation
M,W	16	143	37	Oxidation
Q	−17	103	14	Deamidation
S,E,T	−18	94	27	Dehydration
M	58	53	2	Oxidation+acetylation
K	43	47	17	Carbamylation
W	44	23	6	Carboxylation
N	−17	22	5	Deamidation
S	28	21	14	Formylation
S	80	20	5	Phosphorylation
K	58	17	6	Carboxylation
K	72	14	6	Unknown
W	−2	10	3	Cross-linking
R	55	9	5	Unknown

d

AA	Mass	Spectra	Sites	Putative annotation
Nonspecific	22	74	24	Sodium
M	16	43	3	Oxidation
C	−2	39	7	Cross-linking
R	−28	39	1†	Mutation to lysine
Y	16	37	10	DOPA
N	1	35	5	Deamidation
D	−18	17	5	Dehydration
C	12	7	2	Thioproline

Figure 5.1: PTM Selection (a) PTM frequency matrix for IKKb dataset (4,239 PTM annotations total). The first column indicates the mass shift of a modification. The entry for modification δ, and amino acid a refers to the number of times a appeared with modification of mass δ in a top-scoring spectral interpretation. Columns (N) and (C) give the total number of times the modification occurred on the N-terminal and C-terminal amino acids. Only rows with a total of 100 or more are shown. Entries of 50 or more are shown in grey, and entries that can be explained by mutations are starred. (b) Ranked list of modifications for the IKKB data-set, including (below the horizontal line) modifications deemed spurious. (c) Ranked list of modifications on Lens spectra. (d) Ranked list of modifications on ISB spectra. All entries were supported by multiple peptides (except those starred) and by the presence of the unmodified peptide (except those marked with dagger)

The key idea of spectral alignment is to represent the spectrum of a peptide with parent mass *M* as a 0-1 sequence of length *M* that contains a 1 in position *m* for every mass *m* of a spectral peak, and 0s elsewhere (the masses in the spectrum are assumed to be integers). In this representation, a PTM with positive shift δ>0 is simply an insertion of δ zeroes in the sequence, while

a PTM with negative shift $\delta<0$ is simply a deletion of δ zeroes from the sequence. With this model in mind, comparison of two spectra is turned into comparison of two strings in 0-1 alphabet under insertion and deletion operation. Since this is the classical *edit distance* problem in bioinformatics, the spectral alignment approach essentially transfers the power of dynamic programming and sequence alignment algorithms from genomics to mass-spectrometry.

The above paragraph is an over-simplified description of spectral alignment. While it works for comparing spectra in a case-by-case fashion, it is rather slow, does not adequately model some specifics of MS/MS spectra, and therefore has difficulties in analyzing large datasets of MS/MS spectra in an automatic fashion. Pevzner et al. [Pevzner 2001] themselves formulated a number of open problems that need to be resolved to make spectral alignment practical. Here, we describe *MS-Alignment*, a spectral alignment algorithm that resolves those problems, using a number of improvements, as described in Methods.

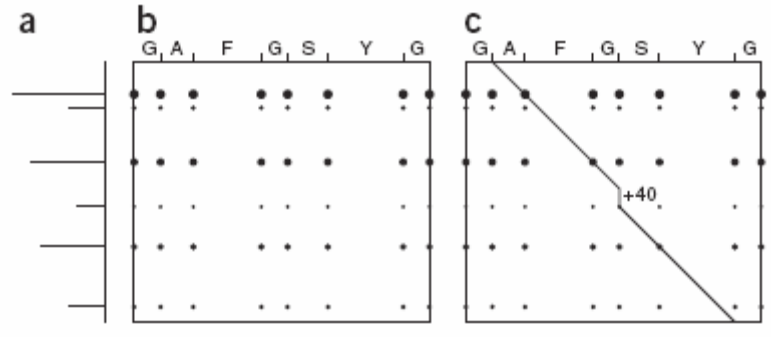


Figure 5.2: Spectral alignment. (a) An example of spectral product for the spectrum S = {71, 100, 218, 315, 402, 532} (shown in a) and the database GAFGSYG. We note that for clarity of the figure, the spectrum S contains only b peaks and noise peaks. (b) The spectral product contains a dot for every mass in the spectrum and every mass of a database prefix. In the figure, the size of the dot is proportional to the intensity of the corresponding peak in the spectrum. (c) An interpretation of the spectrum, such as AFG#SY, corresponds to a path from top to bottom in the spectral product. A path consists of diagonal segments, vertical segments (for modifications with positive mass offsets), and horizontal segments (for modifications with negative mass offsets).

Pevzner et al. reduce the spectral alignment problem to a geometric problem of finding the highest scoring path in the plane. The score of a path is the number of points from the *spectral product* that are on the path, where the spectral product is the set of points corresponding to all pairs of masses from the two spectra. Figure 5.2 illustrates the spectral product for a search against a small database (a database search corresponds to comparisons of two spectra: the experimental spectrum and a theoretical spectrum that consists of all the masses of prefixes of the database). The approach of can be extended by improved scoring of paths. Such scoring requires assignment of (i) intensity-based scores for the points of the spectral product, (ii) penalties for "missing peaks" (prefix masses with no corresponding peak in the spectrum), and (iii) penalties for modifications. MS-Alignment achieves this goal with a rapid running time (see Methods).

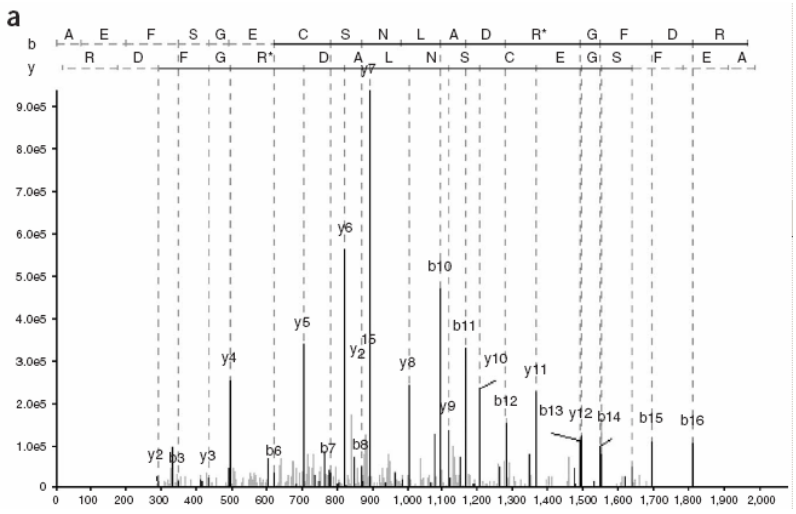
Candidate Validation. Modifications on closely located amino acids often produce similar theoretical spectra, thus making it difficult to identify the exact position of modification. For example, consider a peptide with two consecutive Methionines, one of which is oxidized. Unless the spectrum is of particularly high quality, the two alternative candidates that place the oxidation on either residue will receive very similar scores. However, if these candidates greatly out-score other peptides, we can confidently assign a peptide annotation (with some uncertainty on the oxidation position). Therefore, we categorize search results as being either *incorrect*, δ-*correct*, or *correct*. A δ-*correct* result recovers the correct peptide (possibly with misplaced modifications), while a *correct* result recovers the original peptide sequence and position of modification exactly.

Even with improved scoring, reliable identifications of modifications is challenging. To automate the validation, we construct a *PTM frequency matrix*. For every peptide with predicted modification δ on amino acid *a*, we incremented the count PTM(δ,*a*) in the PTM frequency matrix (Figure 5.1a). We varied δ from -100 to 160 resulting in a 261x20 matrix. If all identified 3571+334x2=4239 PTMs were incorrect, one could assume that PTM frequency matrix represents “noise” with mean value PTM(δ, a) ≈0.8. In reality, while most entries (90 percent of all entries) are zero, a few others represent very high values, a highly unlikely outcome for a random matrix.

One may be tempted to output the list of large entries in PTM frequency matrix as the set of identified PTMs. With this approach, we will certainly miss the rare, yet correct modifications. Surprisingly, even for the common modifications, this approach has pitfalls. For example, *PTM*(16,*I*)=39 may simply be a “shadow" of *PTM*(16,*M*)=612 caused by δ-correct interpretations with misplaced PTM. As another example of the complexity in interpreting this matrix, note the surprisingly many large entries *PTM*(1,*) in the row 1 of PTM frequency matrix. Modifications with mass difference 1 should be viewed with suspicion since some are artifacts produced by errors in parent mass and isotopic peaks, rather than real PTMs. However, one of the entries in the first row *PTM*(1,*N*)=164 is particularly large and we suggest that it represents deamidation of Asparagine rather than a parent mass artifact.

To repair δ-correct annotations, we adopt a “strength in numbers" approach and *rank* all identified PTMs based on the values in PTM frequency matrix after subtracting of “shadows" as described in Methods. This process produces a ranked list of PTMs present in the sample. Figure 5.1b illustrates the utility of this ranking procedure when applied to the PTM frequency matrix in Figure 5.1a.

a



b

Peptide	Count	
28 on K (dimethylation)	Modified	Unmodified
K*LSSPATL	17	0
K*LSSPATLN	6	0
K*LSSPATLNS	55	0
K*LSSPATLNSR	19	0
K*LSSPATLN+1SR	5	0
IK*LSSPATLNSR	1	0
IMLIK*LSSPATLN	1	0
IMLIK*LSSPATLNSR	17	0
IMLIK*LSSPATLN+1SR	1	0
DNDIMLIK*LSSPATLNSR	1	0
TLDNDIMLIK*LSSPATLNSR	1	0
IITHPNFNGNTLDNDIMLIK*LSSPATLNSR	3	0
K*IITHPNFNGN	21	0
K*IITHPNFN+1GN	1	0
K*IITHPNFNGNTL	6	0
K*IITHPNFN+1GNTL	2	0
GNEQFINAAK*IITHPN	1	0
LGEHNIDVLEGNEQFINAAK*	4	20
76 on C (beta-mercaptoethanol)		
SC*ILQEPK	13	37
PQPESVSC*ILQEPK	3	4
32 on M (double oxidation)		
THPNFNGNTLDNDIM*LI	3	3
IITHPNFNGNTLDNDIM*LI	2	0
IITHPNFNGNTLDNDIM*LIK	1	6
M*MALQTDIVDLQ	10	0
M*MALQTDIVDLQR	160	3
M*MALQTD+53IVDLQ	3	0
M*MALQTDIVD+22LQ	3	0

Peptide	Count	
16 on M (oxidation)	Modified	Unmodified
EM*EQAVELCGR	6	5
WREM*EQAVEL	3	0
AWREM*EQAVEL	10	0
LPEM*LK	3	0
DFLSKLPEM*	9	0
DFLSKLPEM*L	38	35
71 on C (PAM-cys)		
C*ISDGKLNEGH	9	1
PDKPATQC*ISDGK	1	0
IPDKPATQC*ISDGK	22	0
LIPDKPATQC*ISDGK	2	0
AGLALIPDKPATQC*ISDGK	104	0
AGLALIPDKPATQC*ISDGKLNE	12	0
AGLALIPDKPATQC*IS+1DGKLNE	1	0
28 on S (mutation to D)		
GPGTS*ILSTWIGGSTR	3	0
FGPGTS*ILSTWIGGSTR	1	0
DIFGPGTS*ILSTWIGGSTR	21	0
DIFGPGTS*ILSTWIGGSTRSISGT	2	0
DIFGPGTS*ILSTWIGGSTRSISGTSMATPHVAGLA	3	0
55 on R (unknown)		
AEFSGECSNLADR*GFDR	5	20
RAEFSGECSNLADR*GFDR	3	3

Figure 5.3: (a) A spectrum from the lens data set supporting placement of an uncharacterized modification of net mass 55 Da at residue R85 of crystallin beta 1. Support from several overlapping peptides provides confirmation of the modification. The peptide sequence is underlined to reflect the presence of runs of multiple peaks in the b and y series. (b) Validation of PTMs described by multiple occurrences of identical or overlapping peptides. The second column is the count of the spectra of modified peptides, whereas the third column is the count of the spectra of unmodified peptides. These examples come from the IKKb data set, with the exception of R+55.

Once the list of PTMs is produced, we can validate particular modification sites. The PTM site supported by multiple spectra is more reliable than a PTM site supported by a single spectrum. In addition, two overlapping peptides validating the same PTM are more robust evidence than two identical spectral annotations for the same PTM. Such “overlapping witness” evidence is most abundant for low-complexity samples, and samples which have been subjected to non-specific digest. We also report, for each modified peptide, the number of times the unmodified peptide was observed. This can provide additional positive evidence for the modified annotation (since the peptide is known to be present), as well as a rough measure of the frequency with which the particular residue is modified.

5.2 Results

We analyzed three experimental datasets (acquired on QTOF and LTQ instruments) and two simulated data-sets as described in Supplementary Methods.

- **IKKb dataset** 45,500 spectra acquired from a digestion of IKKb (inhibitor of nuclear factor kappa B kinase beta) by multiple proteases. (E. Zandi and T. Higashimoto, unpublished data). A database pruning step on Swissprot (54Mb) resulted in a database containing IKKb and proteases (10 proteins, 5kb).
- **Lens dataset** 43,518 spectra acquired from human lens proteins [Searle 2005]. A major component of the lens proteome

comprises of crystallins, which have very little turnover, and acquire modifications with age. When a person ages, the crystallins become insoluble, and the tissue increasingly opaque often leading to cataract. Post-translational modifications are known to play a major role in the process [Searle 2005, MacCoss 2002]. A database pruning step resulted in a database of crystallin proteins that represent the majority of proteins in human lens tissue (20 proteins, 5kb).

- **ISB dataset** 37,000 spectra from a public collection of MS/MS spectra [Keller 2002]. This data-set was chosen as it has been queried extensively, but many spectra remain unannotated. The ISB spectra were searched against a database of 37 proteins (25kb).

Results on IKKb. We analyzed spectra in the IKKb dataset, retaining matches with p-value below 0.05. MS-Alignment identified 8,641 unmodified peptides, 3,571 spectra with a single PTM and 334 spectra with two PTMs. Figure 5.1a shows the PTM frequency matrix while Figure 5.1b presents the ranked list of PTMs. Most large entries in the PTM frequency matrix (highlighted in gray) do indeed correspond to known modifications, validating our approach. Figure 5.1b illustrates that even a single protein (IKKb) has a rich set of PTMs. Figure 5.3b lists some of the modification sites with strong support from multiple overlapping annotations. The IKKb spectra suggest the presence of mutations and modifications missed by traditional database search.

Results on Lens. Our blind search of Lens dataset identified 5,616 unmodified spectra, 3,027 spectra with a single modification, and 244 spectra with two modifications. The majority of annotations (Fig. 5.1c) represent known chemical modifications of crystallins. Evidence was also found for uncharacterized modifications, including a modification of net mass 55Da on Arginine (Fig. 5.3a).

After the blind search discovers the set of PTMs present in a data-set, we perform a restrictive search with a list of identified PTMs using InsPecT [Tanner 2005]. This second pass can use knowledge of fragmentation effects of identified PTMs, such as phosphorylation and oxidized methionine. We re-searched the Lens data-set using the PTMs from Figure 5.1c. We obtained a total of 11,278 spectral annotations with p-value <0.05, resulting in a 40% increase in the number of spectral annotations of modified peptides. The results include 6,672 unmodified spectra, 3,526 spectra with one modification, and 1,080 spectra with two modifications. A total of 519 modification sites were confirmed by two or more spectra. We further filtered this list to modifications which had the modification confirmed by successive “rungs” in the *b* and/or *y* ion ladder. In addition, we rejected any modification of +1 not confirmed by at least one QTOF spectrum. This heavily filtered list contains 378 modification sites supported by 4,442 spectra.

As we are using the same data-set as [Searle 2005], we can compare findings on lens. They report a total of 80 modified sites (44 previously known,

36 novel). On the same data-set, our algorithm found 57 out of the previously reported 80 sites and discovered 322 new PTM sites. In addition, we were able to identify a wider range of modifications (Supplementary Table 1). Of the 23 sites annotated by OpenSea, but not by us, 12 were identified but did not pass our strict validation test (p-value <.05). We therefore argue that while MS-Alignment is more conservative in validating the found PTMs, it still finds many more PTMs than OpenSea. In addition, evidence was found (Fig. 5.3a) for uncharacterized modifications which may be relevant to disease progression.

Results on ISB. Despite the fact that the ISB dataset is arguably the most studied collection of MS/MS spectra to date, very few modifications were reported in this dataset. We found many modifications (Fig. 5.1d) including a surprisingly large number of spectra with *PTM*(-2,*C*). The peptides with modification with shift -2 on Cysteine contain two Cysteine residues, providing strong evidence that these modifications correspond to cross-link formation. Moreover, very few of the original annotations from the ISB dataset (14 out of over 2500) contain cysteines, which is unexpected. We propose the following explanation of the above phenomenon: disulfide linkages in the sample were re-established in an undirected manner to produce a variety of different molecules, each with small concentration and unusual fragmentation properties. The few peptides which contained two Cysteines preferentially formed intra-molecular disulfide bridges and were discovered in blind search.

Indeed, we found that cysteine-containing peptides had strong b- and y-ladders with a gap corresponding to the interval between two Cysteines, an overwhelming evidence in favor of our explanation. A few additional annotations were obtained with modification mass +12 on an N-terminal Cysteine, corresponding to a thioproline conversion which would also protect the cysteine residue from promiscuous cross-linking.

We took the ranked list of identified PTM types and re-searched the ISB spectra using InsPecT against a large database containing the correct proteins as well as all human proteins from the nr database. This search was able to annotate 16% more spectra and 20% more peptides than currently known ISB annotations (these annotations have been added to the ISB webpages at http://www.systemsbiology.org/extra/protein_mixture.html). We therefore argue that a combination of MS-Alignment and InsPecT significantly increases the number of annotated spectra in typical MS/MS samples.

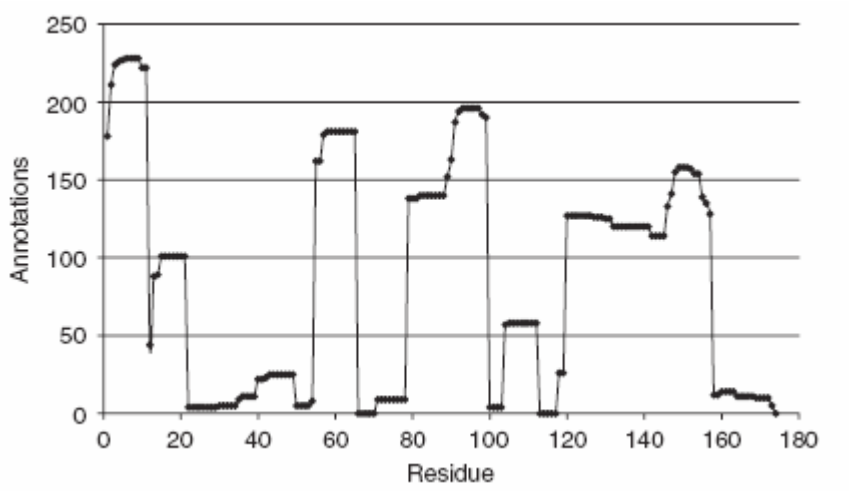


Figure 5.4: Spectrum coverage (c_s) plotted over residues of crystallin αA. Abrupt changes in spectrum coverage are seen at tryptic cuts.

MS-Alignment gives us a rare opportunity to quantify the *modification-rate* at each site. We define coverage c_s of a site *s* as the total number of modified or unmodified peptide spectra that encompass the site. Figure 5.4 plots c_s over crystallin αA protein. The graph approximates a step function with its discontinuities at tryptic cleavage sites. The *modification-rate*, m_s at site *s* is given by

$$m_s = \frac{\#\text{spectra with modification on } s}{c_s}$$

The vast majority of modification events are rare (m_s<10%). However, a high rate of modification of residue K44 in IKKb suggests the presence of a mutation to alanine, since post-translational modification would require the chemically infeasible truncation of the Lysine side chain to a methyl group.

The high rates of modification of residues K97 and K133 of trypsin are likely due to chemical dimethylation performed by the supplier to prevent autodigestion.

We also examined sites with high modification rate in the Lens dataset. We find constitutive acetylation at the N-terminus of most crystallins. We also find that (-17, Q) has high m_s for glutamine residues that are N-terminal in a tryptic fragment.

Recently, a number of methods have been described for identifying PTMs using tandem MS. With few exceptions, these methods generate putative candidates, which must then be manually validated. By incorporating a robust scoring function we take the step towards automated discovery of PTMs from large data-sets of spectra. As the tools mature, and more modified peptides are identified, we can begin to differentiate between different types of modifications by analyzing altered fragmentation patterns and developing PTM-specific score functions. Complete search results are available from http://bioinfo2.ucsd.edu/.

MS-Alignment improves upon earlier spectral alignment algorithms in several ways:

- MS-Alignment is a local or *fitting* version of spectral alignment instead of global alignment as in [Pevzner 2001]. While the global spectral alignment compares a spectrum against *every* possible peptide from a protein, the local spectral alignment compares the spectrum against

entire protein at once by a single pass of the dynamic programming matrix (like in Smith-Waterman algorithm for local sequence alignment). This improvement leads to roughly an order of magnitude speed-up by reducing the time spent evaluating overlapping database peptides.

- MS-Alignment employs a PTM-dependent p-value scoring instead of the rather naive scoring in [Pevzner 2001].

5.3 Methods

Our method has the following steps:

Database Pruning: Note that a typical protein database is very large. We make the reasonable assumption that for every protein present in the spectral data-set, at least one unmodified peptide is present with an identifiable spectrum [Craig 2003]. We use this method to rapidly identify a much smaller subset of expressed proteins.

Candidate generation: We use MS-Alignment to align all spectra against the smaller data-set, and produce a list of candidates (peptides with one or more mass-offsets).

Re-scoring: We use a scoring function (see below) to re-score all the candidates, and discard those with a high p-value.

Candidate Validation: Validating modified peptides is still in early stages due to the lack of the learning samples with annotated PTMs. We use new “PTM frequency matrix" and “overlapping spectral witness" approaches to validate the found PTMs.

We limit our discussion to re-scoring and candidate validation. A complete description of the MS-Alignment algorithm can be found in Supplementary Methods. There has been extensive research of late on improving the scoring of mass spectra of unmodified peptides [Yates 1995, Yates 1995b, Tabb 2003, Perkins 1999, MacCoss 2002, Razumovskaya 2004, Frank 2005, Elias 2004, Havilio 2003]. Unfortunately, these algorithms do not carry over to the identification of modified peptides, and most algorithms for identifying PTMs still consider simpler score functions. The issue of scoring was addressed recently [Tanner 2005]. They revisited restricted PTM search and developed the InsPecT algorithm which (i) incorporates recent advances in scoring into PTM analysis and (ii) is two orders of magnitude faster than other restricted PTM searches. While they report success in revealing many previously unidentified PTMs, their algorithm also has to “guess” the types of PTMs in advance and cannot be run in a blind mode.

The common approach to assigning statistical significance [MacCoss 2003, Nesvizhskii 2003, Anderson 2003, Tanner 2005, Geer 2004] is to combine a number of features, including the percentage of b and y fragments found, the percentage of spectral peaks annotated, the percentage of total ion current in annotated peaks, etc. We use an SVM based approach, similar to [Anderson 2003] to optimally classify the correctly and incorrectly assigned peptides using the features described above. The SVM score for a match is

compared to a histogram of SVM scores over incorrect peptide assignments to produce a p-value for the candidate (See Supplementary Figure 1, 2).

The iterative ranking procedure detects and repairs several type of δ-correct annotations:

- The PTM is shifted to an adjacent residue. A shift of one or two residues affects only one or two *b* and *y* peaks in the theoretical spectrum, so PTM sites may be difficult to pinpoint (particularly near peptide start/end, where peak information is scarce).
- The reported PTM mass is off by 1 Da. Note that a proposed PTM mass is much more likely to be too large than too small, since a candidate with a too-large PTM can still explain a run of secondary isotopic peaks.
- The match may be one residue too short (or too long), and include a PTM which compensates for the missing (or additional) residue. For example, peptide *QA*[111]*EVAHMSQTQEEK* and slightly longer peptide *Q*[-17]*QAEVAHMSQTQEEK* (from the IKKb data-set) produce near-identical spectra, but the latter has a simple explanation (ammonia loss) and the former does not (mass(*Q*)-17 ≈ 111).

On some spectra, our search may fall victim to these pitfalls. Therefore, we adopt an iterative approach that begins with enumerating all spectral annotations whose match score has a p-value of .05 or less. Some spectra will have several such annotations, generally all variants of a single peptide.

Initially, we consider each spectrum to be unlabeled. At each stage, select the (δ,*a*) entry which can annotate the greatest number of unlabeled spectra. Label all the unlabeled spectra which can be annotated using this new PTM, and keep a list *L* of the new peptide annotations. Then, label any unlabeled spectrum carrying an annotation which is a shadow of an annotation in *L*. A spectrum is a shadow of another if it has the same amino acid sequence, but (a) one PTM differs by 1 Da, or (b) a PTM attachment site has been shifted by one residue. This procedure allows true PTMs to “explain away" their shadows. Note that distinct PTMs of similar masses (e.g. +42 and +43) can still be identified, as long as they do not co-occur at the same position. We stop when the next PTM can annotate few new spectra.

Acknowledgments This project was supported by NIH grant NIGMS 1-R01-RR16522. We are grateful to Brian Searle and Larry David for making their Lens dataset available and to Larry David, Katalin Medzihradszky, and Philip Wilmarth for many useful discussions. Production of the lens data-set was supported by National Eye Institute grant EY007755. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. Production of the IKKb data-set was supported by NIH grant R01GM65325 and by the Pew Scholars Program. This chapter initially published as "Identification of Post-translational Modifications via Blind Search of Mass-Spectra". Tsur, D. and Tanner, S. and Zandi, E. and Bafna, V.

and Pevzner, P.A. 2005. Nature Biotechnology **23**: 1562-1567. The dissertation author and Dr. Tsur were primary authors of this paper.

Chapter 6: Peptide and site false discovery rates

6.1 Introduction

Proteins undergo significant post-translational modification in order to modulate their structure, regulate their function, and as part of signaling networks. Modifications such as phosphorylation can be identified from peptide tandem mass spectra [Jensen 2006]. Until recently, search tools required the researcher to specify in advance a list of modifications [Yates 1995, Perkins 1999, Tanner 2005]. More recent work has focused on the identification of modifications in an "unrestrictive" mode, where arbitrary modifications (up to some size limit) are considered [Shevchenko 2001, Liebler 2003, Han 2004, Tang 2005, Tsur 2005, Searle 2004, Savitski 2006, Bern 2007, Havilio 2007]. Earlier work unrestrictive search [Han 2004, Searle 2004] focused on the use of partial *de novo* interpretations followed by local alignment. Our group [Tsur 2005] developed the MS-Alignment algorithm, which identifies modifications without the need for any *de novo* sequencing.

The central challenge of unrestrictive search is that the virtual database of modified peptides is orders of magnitude larger than the original sequence database. Consider a peptide of 10 amino acids, where we permit modifications of integer mass up size up to 251Da. Permitting just one modification per peptide already produces roughly 2,500 modified peptides to choose from. This vast database size makes speed an issue, and enumerating all candidates can become unfeasibly slow. The large database size also

makes it much more difficult to distinguish the correct annotation from a range of quite similar options. Many spectrum annotations are generated which - due to limitations of scoring, and spectrum noise - include the correct peptide, but with a shift in the modification position or mass. We refer to these as δ-correct annotations [Tsur 2005, Tanner 2006]. Estimates of false discovery rates become essential for unrestrictive searches, particularly on large data-sets [Higgs 2007].

Large data-sets, containing millions or tens of millions of spectra, are now routine. Typical searches of such data-sets generate large numbers of spurious modifications, even at a low spectrum-level p-value. On the one hand, even a low false discovery rate (0.01) for a large dataset will result in a significant number of false peptide/protein identifications. [Nesvizhskii 2005]. On the other hand, these large data volumes have the advantage that they contain redundant spectra for the same peptide, and overlapping peptides for many modification sites. This data provides much richer evidence for a modification site than is available from any one spectrum. Therefore, instead of focusing on the accuracy of individual spectra, we focus here on the accuracy of identified peptide and post-translationally modified sites, incorporating evidence from multiple spectra. In prior work, some of us provided a ‘proof of principle’ using a number of heuristics for validation of individual sites, allowing us to discover previously unannotated modifications [Tsur 2005, Tanner 2006]. This approach, though straightforward, requires

significant manual intervention to validate the discovered modifications. With larger training data available, we can use these heuristic ideas as features to develop a tool, PTMFinder, for automated discovery of post-translationally modified sites.

The PTMFinder tool uses a total of seven features to improve the accuracy of PTM identifications. The score of the best available spectrum is an obvious (and useful) feature. Other notable features include: using the consensus spectrum from all spectra available for a putative modified peptide: The difference in score between the putative modified peptide and the best available unmodified peptide (as most peptides are unmodified, annotations should only incorporate modifications if they are required for a satisfactory explanation of spectrum): and the total number of spectra for a peptide species. Additional features are discussed in the methods. The use of features which integrate data across multiple spectra, the careful selection of features, and the focus on modification sites help distinguish our work from earlier work in computing peptide-level accuracy.

We use a decoy database (shuffled protein sequences) to quantify the false discovery rate of our search for PTM sites [Higdon 2005, Higgs 2007, Fenyo 2007, Elias 2007]. Any modification site on a decoy protein is invalid, and their quantity gives an accurate estimate of the number of invalid sites which (by chance) fall within a valid protein. Through this estimate, we are

able to quantify the false discovery rate, a critical consideration in any high-throughput study.

Our tool reflects a shift in emphasis. As mass spectrometric data-sets continue to increase in size and complexity, it will become increasingly rare to focus on accuracy of individual spectral annotations. Instead, the emphasis will shift to presenting a summary of the findings, in terms of peptides/proteins, and their modifications, as well as the combined evidence for these findings. . To illustrate the power of the tool, we present the results of our analysis on several large proteomics datasets, including ~18 million spectra from human samples (primarily HEK293 cell lines), and 1.4 million spectra from the protist *Dictyostelium discoidens*. Thousands of sites of post-translational modification were identified at a 5% false discovery rate (for sites). Comparing our results to databases of known modifications, we confirm over 900 sites previously identified in the human proteome (Table 6.3). Our results also provide evidence for many new phosphorylation sites in the human proteome. The wealth of sites identified from a single tissue source suggests that many post-translational modifications remain to be discovered. The chemical adducts in our results also highlight the particular vulnerability of peptide N-termini to chemical modification in vitro.

The software which performs this analysis, named PTMFinder, is available from our group's webpage (http://peptide.ucsd.edu). Although the analysis here focuses on processing of MS-Alignment search results, in

principle the procedure will work with any tool that performs unrestrictive search and scores putative annotations.

6.2 Methods

Overview

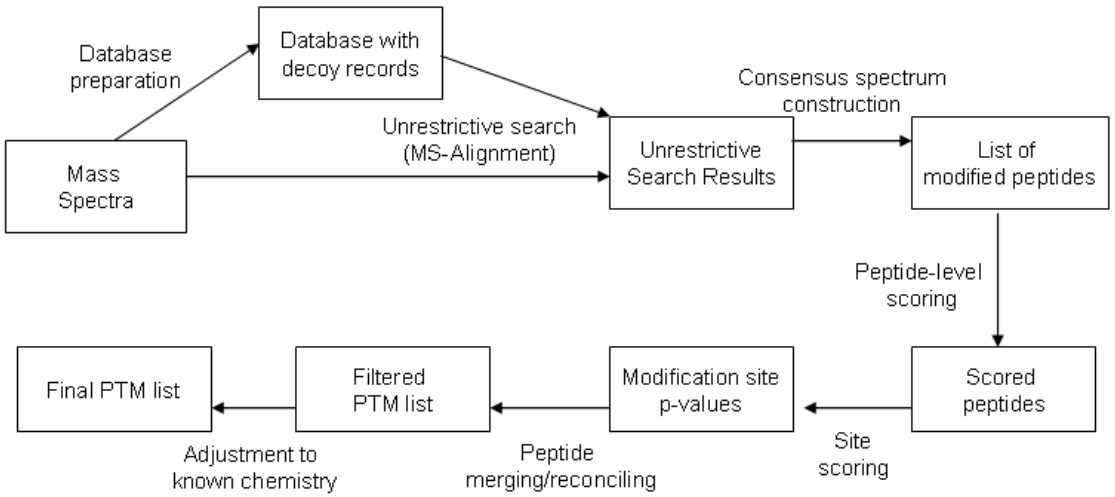


Figure 6.1: Overview of the procedure for high-throughput identification of peptide modifications. The PTMFinder algorithm begins with unrestrictive search results for spectra, and integrates information across multiple spectra (and from multiple peptides) to validate particular modification sites.

In principle, PTMFinder is applicable to the results of any unrestrictive database search tool: a tool that can search a large database for post-translationally modified peptides, and provides a score to quantify the quality of spectrum annotations. We report here the results of PTMFinder on search results generated by the MS-Alignment search algorithm [Tsur 2005], which uses the MQScore from InsPecT.

Figure 6.1 provides a schema for the overall approach. As mentioned, we start with annotations to individual spectra as a key step in the procedure. An unrestrictive search must consider an extremely large "virtual database" of

modifications. Therefore, this is the most computationally-intensive step (a search of 20 million spectra required ~20,000 hours on a cluster of computers). Because search time scales linearly with database size, we filter the sequence database for the species of interest to contain only the proteins present in the sample. Even for highly modified samples such as lens, the majority of tryptic peptides are unmodified. Therefore, a list of proteins present in the sample is reliably generated by a standard search not considering modifications. The database searched also contains a shuffled version of each protein sequence. These shuffled sequences serve as decoy records for quantifying the false discovery rate.

The PTMFinder procedure itself begins by combining all spectra for a particular modified peptide. For each peptide matched by two or more spectra, we generate a consensus spectrum. The consensus spectrum improves mass accuracy by combining multiple measurements. In addition, computing a consensus spectrum filters out most noise peaks, which do not recur consistently [Beer 2004]. Comparisons of peptide annotations to consensus spectra are more effective than those against individual spectra for distinguishing between true and false annotations. Our procedure for constructing consensus spectra uses the same general approach as that used in generation of reference spectrum libraries [Craig 2006, Lam 2007, Frank 2007]. In this way, the analysis is able to consider evidence from multiple spectra. A key idea here is that interpreting a spectrum in the context of a

complete experiment is more effective than interpreting a spectrum in isolation.

We compute a score for each modified peptide using several features. If a peptide is present in multiple charge states, each charge state is scored separately. This analysis used InsPecT's match quality score (MQScore). Most peptide search tools provides a peptide score (e.g. the XCorr value from Sequest), which could be used in this analysis. The features are used to train a support vector machine (SVM) [Chang 2001] using a radial basis function (RBF) kernel. A non-linear basis function is appropriate since the features (Figure 6.2) do not linearly separate the valid and invalid peptides. In the absence of a large, curated collection of post-translationally modified peptides, the training set for the SVM was derived from the Lens search results. We assume spurious matches are randomly distributed throughout the database. We expect very few chance hits to the valid proteins (1% of the total database). Thus, to a first approximation, all modified peptides from a valid protein are true, while all modified peptides from an invalid protein are false. The SVM output for a peptide is the peptide's score. To avoid overfitting, we trained on a randomly-selected collection of 500 valid and 500 invalid peptides.

We then move up another level of abstraction to consider **modification sites**. Each site may be represented by several peptide species due to different charge states, missed tryptic cleavage, and post-digestion breakup of

tryptic samples [Olsen 2004]. These peptide species provide independent confirmation of the modification site. Observation of two spectra for a spurious modification site is not uncommon, since search algorithms are consistent even when they are incorrect. However, observation of multiple peptides for a spurious modification site is very rare, and so overlapping peptides are strongly evidence in support of a modification site [Bandeira 2007]. We use the combined evidence across all peptide species to compute a p-value for each site.

Unrestrictive search

We searched spectra against the database using the MS-Alignment algorithm [Tanner 2006], allowing one arbitrary modification per peptide of size up to ±250Da. Multiply-modified peptides are also detectable by this search if the modifications occur in close proximity. For instance, an N-terminal methionine residue which is both acetylated and oxidized is annotated with a mass shift of +58Da.

Search results were filtered to give a spectrum-level false discovery rate of 5%. This filtering was performed based upon f-score, a weighted sum of Inspect's match quality score (MQScore) and delta-score. A mixture model [Keller 2002] was fitted to the distribution of f-scores, providing a p-value for each match. All spectrum-level filtering was performed using the script PValue.py from the Inspect analysis toolkit.

Consensus spectrum construction

We first consider the top N peaks for a spectrum to be signal peaks, where N is equal to the precursor mass divided by 25. This cutoff is based on an estimated 4 signal peaks per amino acid (with average amino acid mass ~100Da). The intensity of each spectrum is scaled so that (a) the top peak has intensity 20, or (b) the first non-signal peak has intensity 1, whichever scaling factor is lower. This scaling factor compensates for differences in intensity between spectra, while giving a stronger weight to spectra with good signal-to-noise ratios. We then compute the binned intensity (with bin size 0.1Da) of the signal peaks across all spectra.

The consensus spectrum is constructed by converting these binned intensities into a collection of discrete peaks. We iterate over bins from highest to lowest total intensity. When we visit a bin, we create a peak whose mass equals the bin's corresponding mass, and whose intensity is equal to the total intensity from bins within 0.3Da. The intensity of the flanking (assimilated) bins is then set to zero. Finally, the intensity of each peak is scaled by the fraction of spectra where that peak is present (within 0.3Da). This scaling factor rewards peaks which are present (even at medium-to-low intensity) in every spectrum, at the expense of peaks seen in few spectra.

Peptide-level scoring

The final scores for each peptide species is based on a collection of several features. Note that the peptide score may be derived from multiple spectra which re-sample the same peptide, providing a much clearer

annotation than any one spectrum. We note the spectrum with highest MQScore for each peptide. We also compute a score for each consensus spectrum. In addition, we search the consensus spectrum searched (with no modifications allowed) against a large database (Swiss-Prot, downloaded 10/23/2006). We then compute the difference in score between the modified peptide and the best unmodified alternative. The unmodified search runs quickly, and (for a valid PTM) yields a comparatively low-scoring identification. Intuitively, the score difference is large if the modification is necessary for a satisfactory explanation of a spectrum, and small if the modified annotation is similar to matches obtainable by chance.

The final set of features employed was as follows:

- MQScore for the best individual spectrum
- Delta-score for the best individual spectrum
- MQScore of the consensus spectrum
- Delta-score for the consensus spectrum
- Number of spectra annotated with the peptide (log-scaled).
- Peptide length (log-scaled)
- Number of sites from the sample with the same affected amino acid and modification mass (log-scaled)

We now use the peptide scores output from the SVM to compute the probability that each modified peptide is true. The observed distribution of peptide scores is modeled as a mixture of a gamma distribution (for spurious

peptides) and a normal distribution (for valid peptides), in much the same way as the distribution of spectrum scores [Keller 2002]. We note that this mixture model provides us with an estimate of the probability that a (spurious) peptide has score over a threshold.

Site scoring

Consider a modification site supported by distinct peptides P_i. Let $s_i=S(P_i)$ represent the score of peptide P_i. Let X be the event that the modification site is invalid; let Y_i be the event that peptide identification P_i is invalid. If the site is invalid, then each peptide is invalid; these events are independent, we have:

$$P(X) \leq P(Y_1 \wedge Y_2 \wedge ... \wedge Y_n) = \prod_{i=1}^{n} Y_i$$

Incorporating our knowledge of peptide scores:

$$P(X \mid S(P_i) \geq s_i, ..., S(P_n) \geq s_n) \leq \prod_{i=1}^{n} P(Y_i \mid S(Pi) \geq s_i)$$

We now apply Bayes' Theorem, to obtain:

$$P(X \mid S(P_i) \geq s_i, ..., S(P_n) \geq s_n) \leq \prod_{i=1}^{n} \frac{P(S(P_i) \geq s_i \mid Y_i) P(Y_i)}{P(S(P_i) \geq s_i)}$$

The prior probability false, $P(Y_i)$, is estimated when fitting the mixture model during peptide scoring. The values $P(S(P_i) \geq s_i | Y_i)$ and $P(S(P_i) \geq s_i)$ are likewise provided by the peptide scoring model. In this way, we obtain a p-value for each modification site, including those supported by multiple peptide species.

Peptide merging/reconciling

In some cases, we obtain two near-identical variations of the same annotation, with similar theoretical spectra. For instance, on the lens data-set, we often observe peptide -.M+42DVTIQHPWFK.R from the N-terminus of crystallin alpha A. This annotation is correct, the +42 mass offset corresponds to N-terminal protein acetylation. On other spectra, we observe the annotation M.D+173VTIQHPWFK.R. The spectrum is assigned to the correct protein, but the N-terminus is shifted by one residue. This is an example of a δ-correct annotation [Tsur 2005].

We adopt a two-step procedure to repair δ-correct annotations in a principled way. In the first step, we consider merging any two compatible peptides. The peptides to be merged must have the same charge, share a particular protein residue, and have the same prefix and suffix masses relative to the shared residue. The consensus spectrum to be merged must give a comparable match score when annotated with the master peptide. And when both consensus spectra are merged, the resulting consensus spectrum must obtain a high score when annotated with the master peptide.

The second part of the repair procedure is to reconcile those related peptide annotations which cannot be merged. For instance, the peptide R.LMSFRPIC-42SANHKESK.M (from crystallin beta A) cannot be merged with the consensus spectrum R.LMSFRPIC-43SANHK.E, since they differ in length by three residues. In this case, the correct annotation places -43Da on

cysteine residue 117; -43Da is the net mass difference between methylation (14Da) which blocks the expected carboxamidomethylation (57Da). The δ-correct peptide can be reconciled to the correct annotation. Reconciling, like merging, is done only if the revised peptide annotation explains the consensus spectrum (nearly) as well as the original.

After the merging/reconciliation procedure is finished, peptide and site p-values are re-computed, to account for minor shifts in score distributions.

Adjustment to known chemistry

Many annotations from the initial search are δ-correct, and can be trivially repaired. For instance, given an annotation with a +15Da modification on methionine, it is natural to ask whether the common mass shift of +16Da (corresponding to oxidation) would explain the spectrum as well. Therefore, given an annotation, we consider any similar annotations which make use of only common modification types, and - if the site score is improved - retain the correction. The generator for similar annotations may shift the modified residue, shift the peptide endpoints by up to two residues, and alter the modification mass by up to 3Da.

Parent mass errors

Avoiding false positive identifications of small post-translational modifications is a challenge with ion trap data. MS-Alignment requires each peptide to match the parent mass (taken from the spectrum) within a tolerance of 2.5Da. Given an unmodified peptide, if the parent mass is submitted to MS-

Alignment is in error by 3Da or more, then MS-Alignment is forced to apply a modification mass in order to match the true (unmodified) peptide. A typical example: In one run from the HEK293 data-set, the peptide K.SINPDEAVAYGAAVQAAILSGDK.S was repeatedly identified from heat shock cognate 70 kDa protein 8. Several minutes after the peptide was correctly identified (from a scan with precursor m/z 1131.7), a spurious modified peptide K.SINPDE+5AVAYGAAVQAAILSGDK.S was identified from a scan with precursor m/z 1132.6. Comparison of the two spectra reveals that they are almost identical, and that the second spectrum is better explained by the unmodified peptide.

During adjustment to known chemistry, results were post-processed in an effort to filter out these false positives. For each modified peptide species with net modification mass smaller than 10Da, we attempted to annotate the spectrum with the corresponding unmodified peptide. If this improved the annotation score, the modification was filtered from the reported results.

Comparison to known databases

We compared peptide modification sites with those annotated in the Human Protein Reference Database (HPRD) [Peri 2003] and in Uniprot [Boeckmann 2003]. Because we searched IPI (for human samples) or DictyBase (for dictyostelium samples), it was necessary to map from the reference database sequence to the search database. To do this, we extracted the flanking peptide sequence from the reference database, then

found the corresponding sequence in the search database. Next, we considered any sites shared between the set of sites uncovered by our search and those annotated in the reference databases. A shift of one residue or 1Da was accepted in this comparison. The source code for this comparison is available as part of the InsPecT distribution.

Mass spectra

A total of 700,000 tandem mass spectra were acquired from human lens samples, as described previously [Searle 2005, Wilmarth 2006]. These samples include both healthy and cataractous lens tissue. A total of 17 million tandem mass spectra from 747 runs were acquired from H293 cell culture, as described previously [Tanner 2006b]. A total of 1.4 million spectra from 57 runs were acquired from Dictyostelium discoideum whole-cell extract.

Database preparation

Each sample was initially searched against a whole-proteome database with InsPecT, with no post-translational modifications permitted. The list of proteins confidently identified in this search was curated to produce a second-pass database. This reduction in database size greatly reduces the running time [Craig 2003] of the unrestrictive search, important since unrestrictive modification search is computationally intensive. All searches were run with the 10/20/2006 version of InsPecT.

Both human data-sets were searched against the IPI database [Kersey 2004], version 3.22 (10/12/2006). The dictyostelium data-set was searched

against 13,619 records from the DictyBase protein database [Kreppel 2004], downloaded August 28, 2006. Common contaminants (such as human keratins and trypsin) were included in each case. The resulting filtered databases consist of 42 proteins for the lens sample, 2,395 proteins for the dictyostelium sample, and 13,840 proteins for the H293 sample.

Decoy protein records

In order to quantify the false discovery rate of the unrestrictive search, we added spurious proteins to each database. These decoy proteins allow a false discovery rate to be computed without the need for error-prone manual validation of each site. For each protein in the database, a spurious sequence was constructed by shuffling the protein sequence. If the spurious protein contained any peptide of length 8 or more which was shared (ignoring I/L substitutions) with any valid protein, these repeated "words" were re-shuffled as needed. We assume spurious matches are randomly distributed throughout the database. Therefore, given a collection of annotations meeting a score threshold, we can estimate the false discovery rate by simply counting the number of annotations which hit invalid (shuffled) proteins. Since the database of lens proteins was small (only 42 proteins), we added a total of 99 shuffled proteins for each valid lens protein.

Dictyostelium cell preparation

Three cell types of *Dictyostelium discoideum* were studied: vegetative, prestalk and prespore. Cells of the strain NC4 were grown in association with

Klebsiella aerogenes and deposited on 14 nitrocellulose filters for development at 5 x 10^7 per filter (7 x 10^8 cells total). The cells were harvested from the filters at 18hrs (Mexican hat stage), dissociated in 20mM $Na_xK_xPO_4$, 20mM EDTA, 12.5mM 2,3-dimercaptopropanol (*BAL*) and separated on 45% Percoll gradients [Ratner 1983]. The light fraction (prestalk cells) yielded 7.2 x 10^7 cells (22.8%) and the heavy fraction (prespore cells) yielded 2.4 x 10^8 cells (77.2%). The fractions were washed twice in 20mM sodium-potassium phosphate buffer (14.7mM KH2PO4, 5.4mM NaHPO4, pH 6.4) and resuspended in 500µl of the same buffer. 25µl of each was spun down, resuspended in 100µl 0.1% SDS, boiled, and assayed for protein content by Biorad reagent. A volume corresponding to 200µg of each fraction was spun down and resuspended in 200µl 2% Rapigest in TNE buffer, sonicated, diluted with 3 volumes 10mM Tris, heated to 85˚C for 10 minutes, then frozen at –80˚C. Vegetative cells of the strain AX4 were grown axenically (to minimize bacterial protein levels) in HL5 media, and processed as above omitting the dissociation and Percoll gradient steps. Digestion, chromatography, and data acquisition were as described in [Tanner 2006b], producing ~550,000 spectra for each cell type.

Cross-validation

We trained the support vector machine on peptides from one data-set, and tested it against other data-sets, in order to verify that no overfitting

occurred. This ensures that the parameters of our model are applicable across data-sets. Table 6.1 summarizes the results.

Table 6.1: Cross-training between data-sets verifies that the model does not suffer from overfitting. Models trained on one data-set and applied to another (the off-diagonal entries) have comparable accuracy to those trained on a subset of the data they are applied to.

	Train on Lens	Train on Dicty
Test on Lens	91.7%	92.0%
Test on Dicty	81.1%	81.9%

6.3 Results

Data sets

This paper presents the results of the PTMFinder analysis on total of three data-sets. The smallest and simplest data-set comes from human lens. The lens of the eye, consisting primarily of crystallin proteins, is a well-studied system which undergoes both significant post-translational modification and aging-associated chemical damage [MacCoss 2002, Searle 2004]. A previous study of this data-set [Wilmarth 2006] identified many post-translational modifications. However, in prior work, significant manual validation was required, involving additional data (such as elution times) which do not generalize to other modification types and LC/MS/MS protocols. Here, we demonstrate how most of these modifications can be identified in an automated way. In addition to modifications previously reported by our group, we observe evidence for additional modification types. In particular, an apparent gylcosylation (net mass shift 161Da, possibly GlcN) was observed on several glutamine residues.

The second human data-set consists of a very large collection (17 million) of spectra derived from whole-cell extract from the HEK293 cell line. Data-sets of this size are rapidly becoming the norm for proteomic searches. This large data-set has been previously analyzed from the perspective of gene annotation [Tanner 2006b]. Now we are able to turn our attention toward protein modifications. Analysis of this large data-set demonstrates that our analysis procedure scales up to very large samples.

A third data-set consists of 1.4 million spectra from the protist, *Dictyostelium discoideum*. Although *Dictyostelium discoideum* is well established as a model organism for chemotaxis, development, differentiation, and cell communication, relatively little is known about its post translational modifications [Anjard 2005, Chisholm 2004, Chung 2004, Maeda 2004, Van Haastert 2004]. One attractive feature of the organism is the ability to study similarities - and differences - between PTMs in dictyostelium and in other eukaryotes. We investigated this under-studied proteome where few modification sites are known (only 47 sites in Swiss-Prot, most inferred by homology). An interesting related question is about the conservation of post-translational modifications. Much work has been done to examine ancient biochemical pathways, conserved across the tree of life. Most analysis has relied upon sequence analysis. It is reasonable to conjecture that orthologs would also show similar patterns of post-translational modifications. In this

context, our results will show the presence of many modifications conserved across vast evolutionary distances.

Validation Experiments

First, we test the effectiveness of various features in distinguishing valid from invalid modified peptides. We used the lens data-set to measure how well each feature can distinguish peptides from decoy (shuffled) proteins and peptides of valid proteins. The lens consists of just a few dozen proteins, so constructing a very large decoy database (99% shuffled sequence) is feasible. Thus, the lens hits to valid proteins contain very few false positives. Constructing such a large decoy database is inadvisable for most samples, due to the large increase in running time, and possible loss in sensitivity. Hits to invalid proteins do not contain false negatives, since the shuffled database contains no valid peptide sequences.

Table 6.2: Summary of the effectiveness of individual peptide features. Features were scored on their ability to distinguish valid peptides from those on decoy (shuffled) proteins on the lens data-set, where 99% of the database consists of shuffled sequences.

Feature	**ROC curve area (Lens)**
log(SpectrumCount)	0.63
Best MQScore	0.75
Best delta-score	0.80
log(PeptideLength)	0.76
Consensus MQScore	0.75
log(SpectrumCount for this modification type)	0.73
Delta-score of consensus spectrum vs. Swiss-Prot	0.81

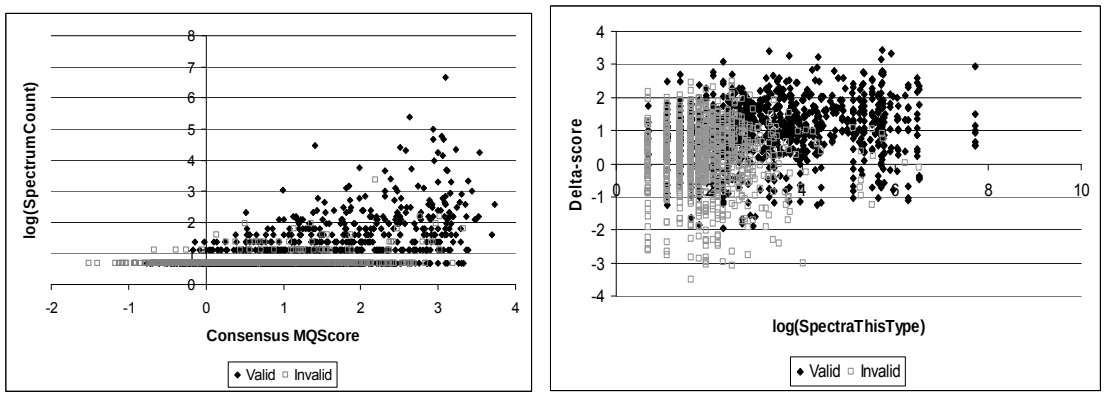


Figure 6.2: Scatterplots show the relationships between pairs of features used to distinguish valid from invalid peptides. Combining several features greatly improves accuracy.

The individual features are described in detail in the Methods. Table 6.2 summarizes the effectiveness of individual using a receiver operating characteristic (ROC) curve. A random feature would have ROC area 0.5, a perfect oracle gives a ROC area of 1.0. Note that search results have already been subjected to a spectrum-level score cutoff; this makes the effect of the MQScore features less dramatic relative to other features. Using peptide length as a feature allows the model to compensate for the correlation of several features with length. Shorter peptide annotations are generally less reliable than long peptide annotations - particularly for modified peptides. Spectra for shorter peptides have fewer peaks demanding explanation, and allowing an arbitrary modification allows *any* peptide to explain the top peak of a spectrum.

The number of spectra with same residue and modification mass reflects the tendency for particular chemical modifications to appear on many

sites throughout the corpus. Note that spectrum count is a valuable feature, but not a perfect one - multiple spectra for the same annotation are not independent, and search tools are consistent even when they are incorrect.

Figure 6.2 illustrates the relationships between some pairs of features. When these features taken together and used in a support vector machine (SVM), the resulting ROC curve has an area of 0.96 (Figure 6.3). In earlier work, we have emphasized the discovery of valid modification types by counting the number of occurrences of the same modification mass on the same amino acid (the "PTM frequency matrix") [Tsur2005]. This remains a valuable feature. However, the full PTMFinder model is much more effective at identifying valid modified peptides than the old approach. In addition, we note that the PTM frequency matrix promotes the discovery of common chemical adducts, such as methionine oxidation, at the expense of less common (and potentially more interesting) modification events or point mutations.

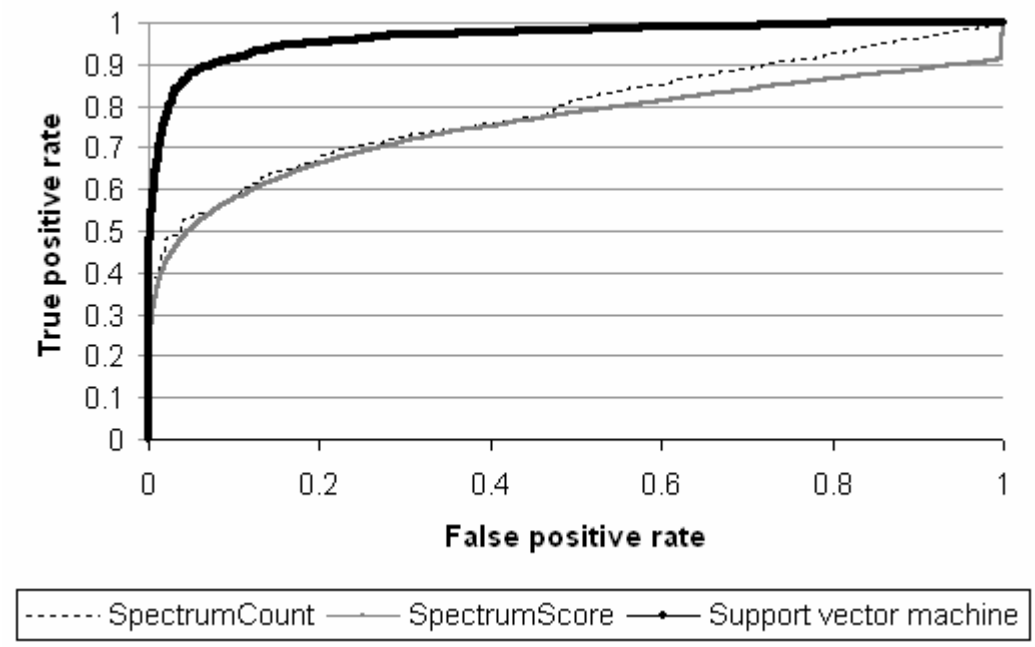


Figure 6.3: ROC curve for categorization of lens peptides using the support vector machine. The accuracy of PTMFinder model is significantly higher than a simple spectrum-level score cutoff. In addition, PTMFinder is more effective than selecting those sites which correspond to the most common modification types (amino acid and mass) by spectrum count.

Adopting this sophisticated model is important for high accuracy. The simplest approach to filtering invalid modified peptides would be to take a very strict spectrum-level false discovery rate, then list all modified peptides which pass this cutoff. However, this naive method is not effective in practice. On the lens data-set, we must reduce the spectrum-level false discovery rate below 1% before the false discovery rate among modified peptides drops below 5%. Only 2,923 modified peptides are found by this approach; the PTMFinder procedure identifies nearly twice as many. The PTMFinder procedure leverages features which are not available when analyzing individual spectra in isolation.

Summary of Results

Table 6.3 presents a high-level summary of the results of the PTMFinder analysis applied on each data-set. Of the peptides (and sites) identified at 5% false discovery rate, we first note those which may be explained by known chemical adducts, such as methionine oxidation. In addition, we highlight those which appear to represent known chemical modification such as lysine methylation. These entries are identified by comparing the modifications observed against a table of "common" modification types (Supplemental table 5), as described in Methods. Of the remaining peptides, some are likely δ-correct annotations which - due to limitations of scoring, and spectrum noise - are better explained by an annotation which shifts the modification position or mass. Others may represent the combination of two post-translational modifications. For instance, a spectrum containing oxidized methionine (+16Da) and an alkylated N-terminus (+42Da) may be annotated with a modification of mass +58Da.

Table 6.3: Summary of modification sites across the three data-sets. For each data-set, we note the sites (and peptides) that appear to correspond to common modification types (either regulated modifications or non-specific chemical damage). The remaining sites may represent novel events, mutations, multiple modifications, or delta-correct annotations. The list of modifications was also compared with large reference databases, to confirm previously-identified modification sites.

	Valid (5% f.d.rate)	Known adducts	Known *in vivo* modifications	Known sites (HPRD / Uniprot)
Lens sites	3,852	1,518	545	
Lens peptides	5,812	2,334	938	
HEK293 sites	24,974	13,949	4,023	933
HEK293 peptides	26,850	15,427	4,229	1183
Dictyostelium sites	12,868	5,095	850	9
Dictyostelium peptides	14,010	5,676	973	18

Modifications shared with the reference databases were identified as described in Methods. We note that the odds of a chance match are quite low. The subset of the IPI database searched contains approximately 5 million residues. The range of modification masses considered yields approximately a 1% chance of matching within 1Da tolerance. Given ~25,000 reference sites and ~25,000 discovered sites, we expect to see approximately one chance match from the HEK293 search results.

Lens data-set

We identified modified peptides from the Lens data-set as described in Methods. A total of 1,088 sites (corresponding to 1,310 distinct peptide species) were obtained at a false discovery rate of 5% (Table 6.4). These sites are reported in Supplemental Table 1.

Table 6.4: Summary of frequent modification types observed on the lens data-set. Modifications with low site-specificity, which may result from chemical damage during processing, are shown in the bottom half of the table.

Mass	Residues	Putative annotation	Sites	Peptides
42	N-terminus	acetylation	99	196
55	R	unknown	92	55
58	K	unknown	82	140
80	S,T,Y	phosphorylation	79	118
-43	C	methylation	69	155
72	K	unknown	43	70
161	Q	hexose adduct	26	42
14	H,K	methylation	24	33
28	K	dimethylation	15	30
43	N-terminus, K	carbamylation	404	615
22	D,E,S	sodium adduct	255	389
-18	D,E,S,T	dehydration	200	284
16	M,W	oxidation	188	311
28	S,T	formaldehyde adduct	131	186
38	D,E	potassium adduct	51	70
-17	N-terminal Q	pyroglutamate	49	94
32	M,W	double oxidation	40	54
-57	C	missing CAM	38	65
-48	M	homoserine lactone	32	54
-17	N	succinimide	31	48
12	W	unknown adduct	30	44
4	W	formaldehyde adduct	24	36

One example of a case that calls for careful scoring is lysine acetylation (mass shift +42Da) and carbamylation (mass shift +43Da). These modifications produce similar theoretical spectra, and are not easily distinguished using ion trap data. However, the consensus spectra for individual sites make it feasible, in most cases, to distinguish between these chemical events. As additional evidence, we observe shifts in retention times (calibrated through measurements on synthetic peptides) which distinguish between these two chemical events. Finally, a neutral loss peak of 43Da from

the carbamylated peptide is observed, providing an additional peak indicative of carbamylation.

HEK293 data-set

A total of 24,974 sites were identified at a p-value of 0.05 or better. After filtering out 13,949 sites which may correspond to known chemical adducts, we retain 4,023 sites which correspond to known *in vivo* modification types (Table 6.3). Table 6.5 summarizes the common modification types observed on this data-set.

Table 6.5: Summary of frequent modification types observed on the HEK293 data-set. Modifications with low site-specificity, which may result from chemical damage during processing, are shown in the bottom half of the table.

Mass	Residues	Putative annotation	Sites	Peptides
80	S,T,Y	phosphorylation	2300	2384
42	N-terminus	acetylation	854	940
14	K,H,R	methylation	386	412
28	K	dimethylation	169	177
57	N-terminus, H	carbamylation	6236	6737
-17	N-terminal Q	pyroglutamate	1880	2242
16	M,W	oxidation	1875	2005
-48	M	homoserine lactone	1072	1348
-18	D,E,S,T	dehydration	1011	1063
-17	N	succinimide	726	827
28	S,T	formaldehyde adduct	519	553
22	D,E	sodium adduct	377	377
43	N-terminus, K	carbamylation	299	313
38	D,E	potassium adduct	228	229
32	M,W	double oxidation	106	113

Figure 6.4 shows the spectrum for one modified peptide, together with the corresponding unmodified peptide spectrum.

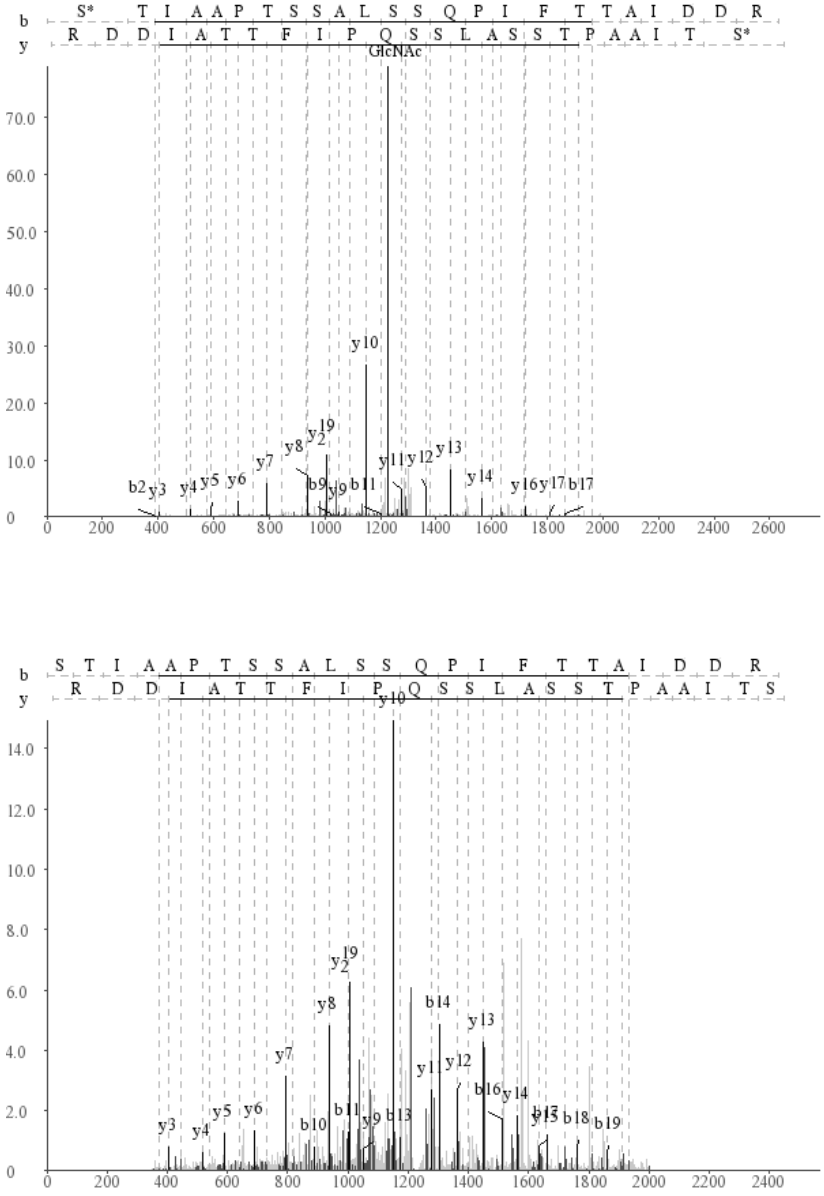


Figure 6.4: Modified (top) and unmodified (bottom) peptides from Eukaryotic translation initiation factor 4 gamma 3 (IPI00646377.1), observed on the HEK293 data-set. A facultative modification of net mass 203Da was observed on residue S277. This modification is presumed to be O-linked GlcNAc; like most glycans, it is extremely labile. The strong peak in the center of the spectrum corresponds to a doubly-charged peptide with the glycan broken off. The similar spectrum from the unmodified peptide is shown for comparison.

A full table of modified sites discovered by our procedure is available in Supplemental Table 3. The human proteome has been extensively annotated.

We extracted from the Uniprot database a total of 5,227 chemical modification sites mapping to proteins in the IPI database (as of 10/2006). The Human Protein Reference Database (HPRD), a human-curated database derived from research literature, includes a total of 18,607 sites mapping to the IPI database [Peri 2003]. (Code to perform this mapping is included in the InsPecT distribution) Only ~38% of modification sites are shared between the two reference databases, which suggests that many more sites remain to be annotated. We compared the collection of sites identified by our procedure to known modification sites, as found in HPRD and Uniprot. A total of 933 of the sites identified from HEK293 were previously annotated in one or both reference databases (Supplemental Table 4). Most of these sites are phosphorylation (817), acetylation (53), or methylation (19). Approximately 10% of the modification sites from the two reference databases were confirmed by this experiment (Figure 6.5).

We note that the reference databases contain primarily phosphorylation, difficult to annotate from CID spectra without performing enrichment methods such as IMAC purifications. The reference databases also contain many glycosylation sites, which are very difficult to map by standard tandem mass spectrometry. Additional experiments involving additional tissue types or organelle enrichment (to increase protein coverage), or using different biological conditions (to probe facultative modifications), could likely confirm

many more modifications. It is clear that much of the dynamic proteome remains to be mapped.

Figure 6.5: Venn diagram of N-terminal acetylation (left) and phosphorylation (right) sites in human proteins. Known sites from two databases (Uniprot and HPRD) are shown, along with sites identified from the HEK293 data-set.

Dictyostelium data-set

We obtained a collection of 13,425 modification sites (for 14,502 distinct peptides) at an empirical false discovery rate of 5% (Supplemental Table 2). Table 6.6 summarizes the modification types observed.

Table 6.6: Summary of frequent modification types observed on the dictyostelium data-set. Modifications with low site-specificity, which may result from chemical damage during processing, are shown in the bottom half of the table.

Mass	Residues	Putative annotation	Sites	Peptides
80	S,T,Y	phosphorylation	296	318
42	N-terminus	acetylation	289	347
28	K	dimethylation	111	131
14	K	methylation	87	92
57	N-terminus, H	carbamidomethylation	2426	2712
-17	N-terminal Q	pyroglutamate	764	865
-17	N	succinimide	558	618
16	M,W	oxidation	535	574
12	N-terminus	unknown N-terminal adduct	475	511
-48	M	homoserine lactone	301	374
22	D,E	sodium adduct	262	268
-18	D,E,S,T	dehydration	248	255
40	N-terminal Q	CAM and pyroglutamate	56	65

Setting aside phosphorylation sites, only nine modification sites were annotated in Swiss-Prot [Boeckmann 2003]. Our search was able to identify three of them. This fraction is commensurate with the percentage of the proteome which is covered by our whole-cell "shotgun proteomics" experimental data. A total of 400,000 residues (6% of all residues in DictyBase) were covered by an accepted peptide annotation. Some of those modifications we observe are well-supported by multiple overlapping peptides. Table 6.7 shows an example of a histidine residue from actin known to be methylated [Vandekerckhove 1980]. A total of ten peptides observed in this data-set all support the attachment of methyl group at the common site. Overlapping peptides are uncommon (except for particularly abundant proteins) given the high specificity of trypsin. Using multiple proteases may improve accuracy further by providing additional overlapping peptides.

Table 6.7: Ten different peptide species witness histidine methylation of Dictyostelium actin. Combining evidence from multiple peptide species gives a site p-value of 6.6x10-12. Fully tryptic peptides are most common, but missed cleavages and post-digest decay produce several other peptide species.

Peptide	Charge	Spectra	Peptide p-value
K.YPIEH+14GIVTNWDDMEK.I	2	13	0.103
K.YPIEH+14GIVTNWDDMEK.I	3	12	0.028
Y.PIEH+14GIVTNWDDMEK.I	2	3	0.052
K.YPIEH+14GIVTNWDD.M	1	2	0.377
K.DSYVGDEAQSKRGILTLKYPIEH+14GIVTNWDDMEK.I	3	1	0.026
K.RGILTLKYPIEH+14GIVTNWDDMEK.I	2	1	0.004
K.YPIEH+14GIVTNWDD.M	2	1	0.673
K.YPIEH+14GIVTN.W	1	1	0.965
P.IEH+14GIVTNWDDMEK.I	2	1	0.027
T.LKYPIEH+14GIVTNWDDMEK.I	2	1	0.065

Genomic analysis has shown that many ancient proteins and domains are conserved, as manifested in sequence similarity. Here we ask whether modifications of proteins are also conserved across large evolutionary distances, by comparing orthologous proteins from Dictyostelium and humans. Modified peptides from Dictyostelium were searched (via BLAST) against the human proteome in order to find homologous modifications. The histidine methylation highlighted in Table 6.7 was also one of the top sites seen in the HEK293 dataset from human samples. Several other conserved modification sites were found between Dictyostelium and humans (Table 6.8). These include two known ubiquitination sites (shown as +114 on lysine). We find a conserved site of phosphorylation on protein RNA polymerase associated protein Leo1 (DDB0186792, Q8WVCO) [Olsen 2006]. We also observe seven conserved sites of N-terminal acetylation. Finally we note the discovery of a conserved modification on histidine of Elongation Factor 2 (EF2_HUMAN, DDB0191363). This site has been previously recognized as a modified histidine targeted by Diphtheria toxin. However, Van Ness reported a modification by 143Da due to a tri-methyl ammonium attached to the amino acid by an ionic bond, which is unlikely to be preserved in a mass spectrometry given the chromatography and collision [Van Ness 1980]. Given the vast evolutionary separation between Homo sapiens and the protist Dictyostelium, these conserved modifications are of particular interest.

Table 6.8: Observed post-translational modifications conserved between *Homo sapiens* and the protist, *Dictyostelium discoidens*. These shared modification sites demonstrate that some protein modifications are conserved across vast evolutionary distances.

Human Protein	Dictyostelium Protein	Residue (human numbering)	Modification Mass	Modification type
IPI00021440	DDB0220457	73	14	Methylation
IPI00784990	DDB0190279	63	114	Ubiquitination
IPI00784990	DDB0190279	48	114	Ubiquitination
IPI00103090	DDB0186792	212	80	Phosphorylation
IPI00375370	DDB0184189	5	42	N-acetylation
IPI00419307	DDB0191258	2	42	N-acetylation
IPI00183508	DDB0216426	36	42	N-acetylation
IPI00220030	DDB0185208	1	42	N-acetylation
IPI00550917	DDB0216426	2	42	N-acetylation
IPI00100160	DDB0234116	2	42	N-acetylation
IPI00005792	DDB0233678	2	42	N-acetylation
IPI00186290	DDB0191363	714	83	Putative dipthamide derivative

6.4 Discussion

We used support from overlapping modified peptides to improve confidence in predicted modification sites. Comparisons to spectra from the equivalent unmodified peptide (where available) may also prove to be valuable [Bern 2007, Bandeira 2007]. Typically the modified and unmodified spectra are similar (Figure 6.3). Some post-translational modifications of great interest have dramatic effects on fragmentation propensities (e.g. phosphorylation). Therefore, application of spectrum comparisons between distinct peptide species must be carried out carefully. We attempted to use, as a scoring feature, the spectral similarity (cross-correlation) between the unmodified and mass-corrected modified spectrum. However, this feature was not effective, and was dropped from this study.

Common modification types

Some types of residue-level modifications are very common *in vivo*. For example, phosphorylation occurs on many different proteins as part of many signaling or regulatory mechanisms. Other modifications, such as methylation, were observed infrequently. We note that glycosylation is generally not detectible from our experiments, due to the poor quality of fragmentation spectra produced glycopeptides. Nevertheless, the whole-cell studies provide a partial overview of modifications across the proteome.

Several modification types are more likely to represent chemical damage than regulated chemical modifications. Modifications such as methionine oxidation can occur after protein extraction, and may not occur at significant rates in vivo. The N-terminus of a peptide fragment is susceptible to several modifications such as carbamylation (observed in the lens sample), carbamidomethylation adducts (observed in the dictyostelium and HEK293 samples), and pyroglutamate formation. These adducts modifications require the peptide's N-terminus to be exposed by proteolytic activity, and so cannot occur in vivo. Given the wide range of chemical insults suffered by a free N-terminus, it seems plausible that one function of N-terminal acetylation (in eukaryotes) or formyl-methionine (in prokaryotes) is to replace the amino group with a less reactive moiety.

Given the vast differences in peptide abundances within a cell, a chemical adduct on a common protein may produce a peptide with much higher concentration than more biologically significant modifications on

proteins expressed at lower levels. Because chemical adducts are so widespread (Table 6.3), filtering them from a blind search is a non-trivial problem. Incorporating modification-specific scores may improve accuracy. The MS-Alignment algorithm permits the assignment of operator-specified scores to particular combinations of mass deltas and amino acids. Further work by our group will examine the effectiveness of this approach.

Parent mass errors

Parent mass errors are a major source of errors and of δ-correct annotations. Three factors contribute to parent mass errors. First, the m/z of a precursor peak from ion trap MS spectrum may be off by 0.5Da, and this m/z error generates a parent mass error two times (for a doubly-charged peptide) or three times larger (for a triply-charged peptide). Secondly, the +1 and +2 isotopic peaks of a precursor may be selected for fragmentation. Particularly for large peptides, these isotopic peaks have strength comparable to that of the base peak. Finally, the mass spectrometer's dynamic exclusion window may force the selection of a minor peak corresponding to the "shoulder" of the actual MS intensity trace. Because the HEK293 data was acquired on LTQ instruments, many more MS1 peaks were selected for fragmentation, so the problem of selecting peaks of skewed masses is more pronounced.

Summary

Understanding post-translational modifications is vital to study of the dynamic proteome of any species. Whole-proteome searches for known and

unanticipated modification sites are now feasible. We emphasize the importance of quantifying the false discovery rate of such studies, using a decoy database. Our tool, PTMFinder, provides a model which accurately distinguishes correct from incorrect modifications. Comparisons of our findings with reference databases, and across multiple species, serve to confirm many putative modification sites. These comparisons also demonstrate the many open questions about when and where proteins undergo modification, what biological purposes these modifications serve, and how chemical damage to experimental samples can be avoided or mitigated.

Acknowledgements

This project was supported by US National Institute of Health grant NIGMS 1-R01-RR16522. S.T. and S.H.P. are supported by NSF IGERT training grant DGE0504645. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. Part of this investigation was supported using the computing facility made possible by the Research Facilities Improvement Program Grant Number C06 RR017588 awarded to the Whitaker Biomedical Engineering Institute, and the Biomedical Technology Resource Centers Program Grant Number P41 RR08605 awarded to the National Biomedical Computation Resource, UCSD, from the National Center for Research Resources, National Institutes of Health. This chapter is in preparation for publication as "Accurate Annotation of Peptide Modifications through Unrestrictive Database Search".

Tanner, Stephen and Payne, Samuel H. and Dasari, Surendra and Shen, Zhouxin and Wilmarth, Philip and David, Larry and Loomis, William F. and Briggs, Steven P. and Bafna, Vineet 2007, in preparation. The dissertation author was the primary author of this paper.

Chapter 7: Improving gene annotation with mass spectrometry

7.1 Introduction

Annotation of protein-coding genes is a key goal of genome sequencing projects. In spite of recent advances in computational gene finding, a comprehensive annotation of protein coding genes remains challenging. In most annotation pipelines, a computationally predicted gene must be confirmed by independent evidence and/or manual validation before it is accepted. The additional evidence is often in the form of conservation across distant organisms or evidence of transcription. This evidence, while compelling, is not sufficient (Ex:Gupta et al., 2004). Conservation across species is not limited to protein coding regions. Roughly 5 - 20% of the human genome is conserved against mouse, of which just 1 - 2% is considered to be coding for proteins (Waterston et al., 2002). Likewise, most cDNA sequences are obtained from single-pass, high-throughput sequencing, and contain sequencing errors, pre-spliced mRNA, as well as untranslated regions. Thus it is hard to determine if every alternative splice form predicted from an EST is also expressed at the protein level. Alternative splicing and overlapping genes present particularly difficult annotation problems. Some estimates suggest that the majority of human genes undergo alternative splicing (Florea et al., 2005, Mironov et al., 1999, Modrek and Lee, 2002).

Therefore, it is customary to provide a conservative genome annotation and then rely upon community efforts to refine annotations and fill in missing

genes. While the genome annotation process is unlikely to be fully automated, high-throughput methods are an important part of any genome annotation strategy. Tandem mass spectrometry is an attractive technique for validating gene predictions. It measures proteins directly, verifying putative gene products at the level of translation. Also, it provides an orthogonal line of evidence, with different error sources than nucleotide-based approaches.

A tandem mass spectrum can be viewed as collection of fragment masses from a single peptide (8-30 amino acids from an enzymatically digested protein). This set of mass values is a 'fingerprint' that identifies the peptide. The spectra are usually not analyzed de novo. Instead, they are compared against peptides from a database of known proteins (Aebersold and Mann, 2003). Much research has been devoted to improving the accuracy of this search by refining scoring (Bafna and Edwards, 2001, Creasy and Cottrell, 2002, Lu and Chen, 2003, Perkins et al., 1999, Sadygov and Yates, 2003, Tabb et al., 2003, Yates et al., 1995b), improving search speed (Craig and Beavis, 2003, Frank et al., 2005), and handling post-translational modifications (Tsur et al., 2005).

In this context, it is natural to ask if we can search translated genomic databases directly. Each match from such a search confirms a genomic locus to be part of a protein-coding gene. This has been proposed in a number of studies (Yates et al., 1995a, Carlton et al., 2002, Kuster et al., 2001, Choudhary et al., 2001, Fermin et al., 2006). However, in eukaryotes,

searching a straightforward six-frame translation is problematic. The typical exon in a multi-exonic gene is short, with an average length of 150bp (50aa). A significant fraction (roughly 25%) of trypsin-digested peptides from eukaryotes span an exon boundary, and so cannot be identified with an ORF database. Predicting the correct introns is a difficult step in gene finding, and such exon-spanning peptides are critical to confirming and annotating splicing. Also, only a small fraction of the genome codes for proteins. A six-frame translation of the human genome has 6Gb residues, while the size of the known human proteome is just 25Mb. Scaling up to a larger database makes searches slower by orders of magnitude. In addition to the issue of speed, searches are known to have a significant error rate, and larger databases incur a higher false positive rate. Polymorphisms are also a potential source of error in such a search.

We overcome these issues with several technical improvements. First, instead of searching translated genomes directly, we search a compact representation of all putative exons, splice variants and polymorphisms. This representation takes the form of a directed acyclic graph which we call the exon graph. Our search is efficient, using a database filtering technique based on tagging (Frank et al., 2005) which extends directly to searching graphs instead of sequences. We also use improved scoring (Tanner et al., 2005, Keller et al., 2002) to keep the false discovery rate at 2.5%. We show that evidence from mass spectrometry can be fed into to computational gene

finding methods to improve gene predictions. An outline of our method is presented in Figure 7.1.

7.2 Methods

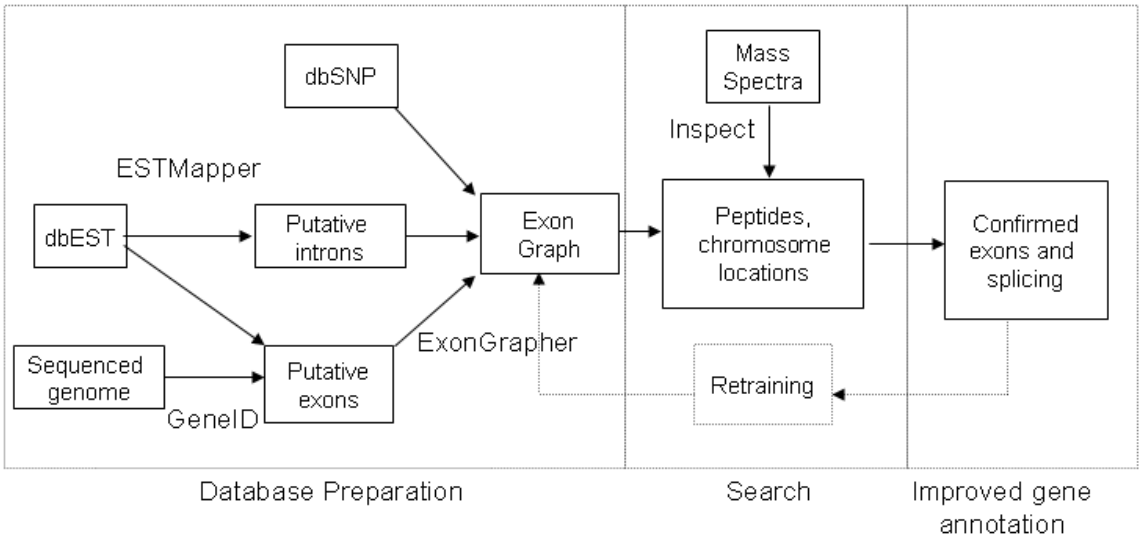


Figure 7.1: Overview of the workflow for genome annotation through mass spectrometry. The exon graph database is constructed without reference to prior annotations of the genome. Putative exons and exon-pairs are generated through EST alignment and de novo predictions; homology maps are another potential source. Peptide matches identify the true exons (and introns) among the gene predictions

Exon and intron predictions

Exon predictions were generated by GeneID (Parra et al., 2000, Blanco et al., 2002) against build 35 (May 2004) of the human genome. All putative exons with a score of -1 or better were retained, producing 4,110,476 exons with considerable overlap. Splice junctions were considered between all pairs of exons with compatible reading frames and intron length between 25 and 20,000 bases. Each interval was linked to the closest intervals with a

compatible reading frame. At most ten introns were considered per genomic position.

We extracted human sequences from dbEST (Boguski, 1993) (6,587,476 sequences). These sequences were aligned against the May 2004 assembly of the human genomic sequence using ESTMapper (Florea et al., 2005). A total of 7,153,771 alignments were generated (including multiple alignments for some sequences). Because genomic contamination and sequencing errors produce noise in EST data, we filtered the set of putative exons and introns. Mappings with sequence identity < 90% or containing cDNA gaps were removed. We also compared each splice junction J against the consensus splice signal using a simple position weight matrix. We discarded any putative intron which (a) occurred in only one EST mapping, and (b) had a poor (5th percentile or less) signal score. Roughly 10% of all introns (336,833) were discarded in this way. The signal score and occurrence count of each intron are stored in the database for later reference. After filtering, 6,923,229 EST mappings were generated, with an average of 2.2 intervals per EST.

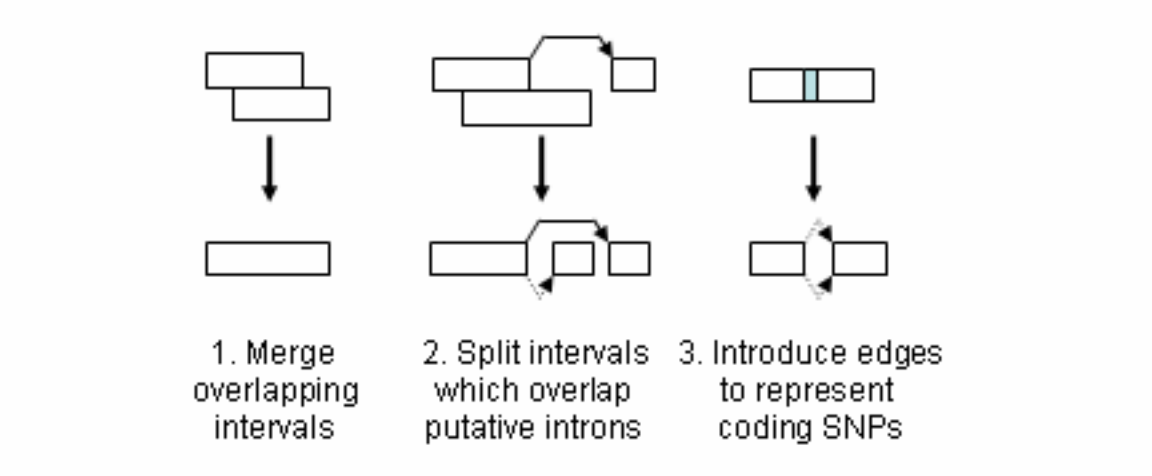


Figure 7.2: Overview of the procedure for turning a collection of putative exons and introns into an exon graph. Adjacent edges are represented by dotted lines, splice events by solid lines.

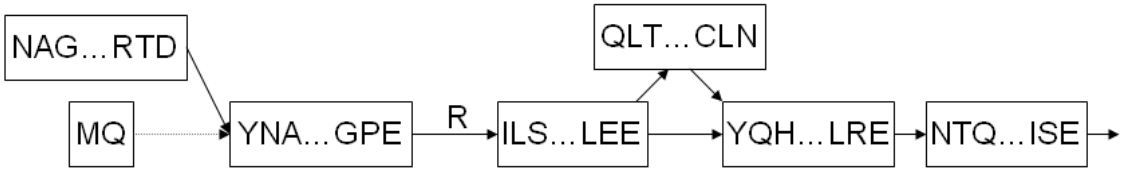


Figure 7.3: A portion of the exon graph for heterogenous nuclear ribonuclear protein K. The labeled edge represents a codon split across a splice junction. The dotted edge is an "adjacent edge" corresponding to a longer form of an exon. Searching the exon graph reveals peptides spanning both outgoing edges from the central node, confirming alternative splicing at the level of translation.

Database construction

Our goal is to build a compact representation of all the exons and introns derived from GeneID and ESTMapper. Let I and J be the collection of intervals and splice junctions for a chromosome strand. The endpoints of interval I_n are denoted as L_n and R_n, with the convention that I_n includes the bases from L_n up to (but not including) R_n. We call a point a *junction point* if it

is an edge of any putative intron. Refer to Figure 7.2 for an overview of the procedure, and Figure 7.3 for an example of the final graph.

Gene prediction algorithms often produce putative exons of various lengths which overlap. Similarly, because ESTs have varying read lengths, it is common for them to map to overlapping genomic intervals. If intervals I_i and I_j overlap, we can merge them into a larger interval without loss of information, so long as:

1. $L_i = L_j$, or $\max(L_i,L_j)$ is not a junction point
2. $R_i = R_j$, or $\min(R_i,R_j)$ is not a junction point.

We perform all such legal merges. This phase greatly reduces the redundancy of the set of intervals. If an interval overlaps the edge of a putative intron, we cut the interval into two sub-intervals at the junction point. At the end of this phase, our set of intervals is disjoint. We now add an edge between any adjacent intervals (I_i and I_j such that $R_i = L_j$). For each putative intron, we add a splice edge between the corresponding intervals. We now incorporate polymorphisms. If an interval contains a coding SNP, we add intervals for each allele. Thus, each SNP produces a “bulge” in the graph.

We derive an exon graph from the genomic interval graph. For each node in the interval graph, add one node to the exon graph for each legal reading frame. Each exon graph node has a protein sequence, and may have an untranslated prefix and suffix. If intervals are joined by an edge, then the

corresponding exons (with compatible reading frame) are similarly joined. Edges are annotated with an amino acid when a codon is split between exons.

In order to remove non-coding “noise” from the database, we remove all nodes and edges which are not part of a coding sequence of length 50 or more. This procedure removes nodes corresponding to translation of EST mappings in the wrong reading frame. The finished exon graph contains a total of 133M amino acids, in 3.5M exons, with 2M splice junctions.

Mass spectra

Proteins were extracted from H293 cell culture. Our standard extraction contains 2% RapiGest (Waters) in TNE buffer. Disulfide bonds were reduced using a final concentration of 2mM TCEP for 30 minutes. A final concentration of 5mM iodoacetamide was used to alkylate sulfhydryl groups. Protein concentration was measured with a Bradford assay. Proteins were then digested with trypsin (1:50) overnight.

An Agilent 1100 HPLC system (Agilent Technologies, Wilmington, DE) was used to deliver a flow rate of 300 nL min-1 to the mass spectrometer through a splitter. Chromatographic separation was accomplished using a 3 phase capillary column. Using an in-house constructed pressure cell, 5 m Zorbax SB-C18 (Agilent) packing material was packed into a fused silica capillary tubing (200m ID, 360 m OD, 20 cm long) to form the first dimension RP column (RP1). A similar column (200m ID, 5 cm long) packed with 5 m PolySulfoethyl (PolyLC) packing material was used as the SCX column. A

zero dead volume 1m filter (Upchurch, M548) was attached to the exit of each column for column packing and connecting. A fused silica capillary (100m ID, 360 m OD, 20 cm long) packed with 5 m Zorbax SB-C18 (Agilent) packing material was used as the analytical column (RP2). One end of the fused silica tubing was pulled to a sharp tip with the ID smaller than 1m using a laser puller (Sutter P-2000) as the electro-spray tip. The peptide mixtures were loaded onto the RP1 column using the same in-house pressure cell. To avoid sample carryover and keep good reproducibility, a new set of three columns with the same length was used for each sample. 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 10mM to 1M ammonium acetate for 20 minutes), followed by high resolution reverse phase separation using an acetonitrile gradient of 0 to 80% for 120 minutes. We have found that a 3D run can provide significantly more resolving power but at the cost of a longer separation time. For 3D, we elute fractions with acetonitrile from RP1 in 10% increments then perform the salt elutions as described above but with a resolving gradient for RP2 of acetonitrile equal to the gradient used to elute from RP1.

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 150C. As a final purification 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.

In addition, we downloaded all human, non-ICAT-labeled spectra publicly available (as of March 2006) in the PeptideAtlas data repository (Desiere et al., 2004). This data consists of spectra from the eryhtroleukemia K526 cell line (Resing et al., 2004), and from the HUPO Plasma Proteome Project (Omenn et al., 2005). The data includes a total of 1.8 million spectra in 621 MS runs, most of them from ion trap mass spectrometers. The HEK293 mass spectra are available from http://bioinfo2.ucsd.edu, together with spectrum annotations.

Database search

The database search proceeds by a modified version of the Inspect search algorithm (Tanner et al., 2005). Given a spectrum, we perform partial de novo reconstruction to generate a peptide sequence tag of three or more amino acids. To accommodate de novo errors, we generate multiple tags, and store them in a trie (Aho and Corasick, 1975). When a tag sequence is found in the database, we perform a depth-first search in the graph to find all

extensions which match the tag's flanking masses. The source code for our software is available from our lab's webpage (http://peptide.ucsd.edu/).

When a tag and its flanking masses are matched, a candidate peptide is produced. Each candidate peptide is scored to compute the probability of that peptide generating the query spectrum (Tanner et al., 2005). Inspect computes match quality scores based upon fragment presence and intensity, and the presence of unexplained "noise" peaks. Once the database scan is complete, the top matches are reported. If the same peptide sequence is observed multiple times, up to ten loci matching the peptide are reported. To improve filtering of incorrect matches, we also consider the difference between the top match score and the score of the next-best peptide (delta-score). To correct for the dependence of delta scores on database size, we take the ratio of a match's delta-score to the average delta-score across all matches. The weighted sum of the match quality score and delta score is called an F-score.

The empirical distribution of F-scores can be fit by a mixture model of a gamma distribution (representing false annotations) and a normal distribution (representing true annotations) (Keller et al., 2002). We select an F-score cutoff which corresponds to a p-value of 0.05 (95% probability of correct annotation).

As an additional measurement of false discovery rate, we constructed a reversed database by reversing the sequences of all nodes and reversing the direction of each edge. We measured an empirical false discovery rate by

searching 700,000 spectra against the reversed databases. Our F-score cutoff yields 1,200 matches on the reversed database, for a false annotation rate of 0.2%. In a search of the forward database, 47,000 spectra passed this same score cutoff. Based on these results, we estimate that 1,200 of the 47,000 spectrum matches against the true database are incorrect, for a false discovery rate of 2.5%. In addition to this filter at the spectrum level, we pay particular attention to exons hit by multiple peptides; no such instances were observed for the search of the reversed database.

Post-processing of the search results was performed to deal with peptides which occur in multiple proteins. We note that in addition to closely-related paralogs, the predicted exons may include some pseudogenes highly similar to their source genes. As an extreme example, the peptide AMGIMNSFVNDIFER (from Histone H2B) is found in over 20 valid and invalid ORFs. Therefore, when measuring coverage, we iteratively select a set of genes. At each stage, the gene which can be used to annotate the greatest number of spectra is selected, and the selected gene “absorbs” all shared peptides. We require at least two peptide hits before judging a protein present. This procedure ensures that redundant or questionable protein records are not selected. When considering alternative splicing, we select multiple isoforms of a protein only if we must do so in order to account for all the peptides matched.

Mapping known proteins to the genome

We wish to ensure the exon graph database captures the exons and introns from known genes. To do this, we produce the full genomic alignment of each protein, including splice junctions. We first identify “seeds”, positions on the genome which appear to match the protein. The chromosome locations are stored for most (54,032) of the IPI database records. In addition, each protein was searched against the repeat-masked human genome using tblastn. Finally, the exon graph was searched for any gene containing length-8 substrings (words) from the full protein; the three records with best coverage were retained as seeds. As a filtering step, we consider only seeds matches that cover at least 30 residues of (an exon from) the source protein.

The heuristic alignment algorithm enumerates 6-mers from the protein found in the six-frame translation of the genomic region of interest. Adjacent hits are merged into putative exons. Using dynamic programming, we find a chain of exons which cover the entire protein. Exons close to each other can be merged, to step over mismatches between the protein sequence and genome. Finally, exon endpoints are refined to capture the best available splice signals.

A total of 56,725 proteins (98%) were mapped against the genome with 95% or better sequence identity. Of these, 37,849 (65%) were mapped with 100% identity. Of the records not successfully aligned, many have no satisfactory “seed” in the tblastn results. Records that represent short signal peptides are often missed in this way (data not shown). Many of the non-

aligned proteins are predicted protein sequences derived from cDNA, which may be chimeric.

Each peptide identified in our database each was compared to the locations of known proteins. If a peptide was found multiple times in the genome, or if two matches had equivalent match scores, we considered each locus. When selecting a locus, the order of preference was as follows: Match to a known gene, match a known gene with SNPs, match a novel single-exon peptide, match a novel intron-spanning peptide. This procedure helps us avoid proposing new exons which correspond to pseudogenes.

Improving gene predictions

Our goal was to demonstrate automated refinement of gene prediction by incorporating MS search results. We selected the GeneID software because it uses a simple two-pass approach to gene prediction. It first predicts a collection of coding exons, then chains these exons into complete genes. Our strategy is to search the exon graph, then boost the scores of exons and introns which correspond to peptides. The assignment of peptide matches to known genes was not used when improving gene predictions.

We first ran GeneID against the human genome, retaining all predicted exons with score -5 or better. We note that exon scores are derived from a log odds ratio; GeneID attempts to avoid incorporating exons with negative scores. We then examined the number of peptide matches which hit each exon, and the p-value of these matches. We note that if the coding sequence

for a peptide spans exons, one of which accounts for just one base pair, there may be several plausible exon pairings which encode the same peptide. Therefore, to reduce false positives, we register an exon hit only if the peptide match is “anchored” by at least 7 base pairs on the exon.

For each exon, we consider three parameters. The parameter c is equal to the number of spectrum annotations which are contained in the exon of interest. The parameter P_a is set to the best p-value of a peptide match covering the splice acceptor of the exon. We set P_a=1 if there are not at least two spectrum annotations covering the acceptor site. Otherwise, we add 0.001 to the p-value to limit the effects of matches with extremely low p-values. Similarly, P_d is the best p-value of a match covering the splice donor. The score S of each exon is modified as follows:

$$S' = S + w_1 \log(1 + c) + w_2(-\log(P_a) - \log(P_d))$$

The weights w_1 and w_2 were tuned to 1.0 and 0.8, respectively, by computing accuracy over a test set of 100 genes from chromosome 1.

For each gene of interest, we extract the genomic interval containing the exons from the gene. We run GeneID in exon-chaining mode to predict a gene on this interval using the original exons, then using the rescored exons.

7.3 Results

Search Algorithm Comparison

We compared the performance of Inspect to that of SpectrumMill (version 3.1, Agilent) on a collection of 800,000 spectra (34 runs) from the

HEK293 data-set. Both tools searched these spectra against the same database consisting of the IPI database, together with the reversed sequence of each protein. We assume that spurious matches are distributed randomly throughout the database. Using this assumption, if 5% of all matches come from reversed proteins, then the false discovery rate among matches from valid proteins is also 5%. Sorting the SpectrumMill matches by score, we obtain 94,633 spectrum annotations (27,845 distinct peptides) at a false discovery rate of 5%.

Sorting the Inspect matches by score, we obtain 135,192 spectrum annotations (43,311 distinct peptides) at this same false discovery rate. These results (40% more spectra, 70% more peptides) indicate that Inspect's filtering and scoring are effective on this data-set.

Exon Graph Construction

One goal in building the exon graph database is to keep the database size as small as possible while still covering all splice variants of all genes. The exon graph contains a total of 134 million amino acid residues, a significant savings over the full length of the EST database (2 billion residues), or the concatenated exon predictions from GeneID (630 million residues). The graph contains a total of ~3M exon nodes and ~8M edges. Modeling possible splicing events as a graph is a familiar formalism (Leipzig et al., 2004, Heber et al., 2002), although our construction of the exon graph differs from previous work (see Methods).

To verify the completeness of the exon graph, we considered the IPI database (version 3.15) as a representative corpus of known human proteins (Kersey et al., 2004). The IPI database contains 25 million residues in 58,099 records. We note that the database is not complete, and contains some hypothetical sequences. We aligned these proteins against the human genome using known genomic locations and BLAST, as described in Methods. We restrict our attention to the 56,725 records mapped to the human genome at 95% or greater sequence identity, the “mapped proteins”. We use this large reference set to estimate the proportion of known genes contained in the exon graph.

The mapped proteins include multiple isoforms of many genes. Counting known proteins that share exons as one gene, we reach a gene count of 32,493, of which 10,583 have multiple isoforms (Supplemental Figure 1). These gene mappings include a total of 442,572 distinct exons. We show later the annotation of peptides corresponding to isoforms which are not contained in the IPI database, but have been deposited in GenBank.

Table 7.1: Coverage of residues, exons and introns from known genes by the exon graph. Our database construction is permissive, and includes many exon variants, in order to capture nearly all proteins. The results of searches against the database confirm specific exons and introns, allowing automated refinement of gene models.

	Residues	Exons	Introns
Total	14715527	258598	220749
EST(%)	90.3	91.9	91.7
GeneID(%)	83.6	80.2	67.7
Combined(%)	95.7	95.6	94.0

For each mapped protein, we determined whether GeneID predictions and/or EST mappings captured the genomic intervals (exons) and putative splice junctions (introns) of the protein. Table 7.1 summarizes the results.

This table reflects the extremely high EST coverage of the human proteome. The exon predictions from GeneID cover most true exons, but the intron coverage is lower. The low intron coverage likely results from the simplistic exon-joining algorithm used in constructing the exon graph. A more sophisticated approach may cover more splice junctions. The exons missed in this construction typically come from the edges of the protein. The coverage rates for first and last exons are 81% for ESTs and 60% for GeneID, significantly lower than the average overall. Further research will target these problematic exons. Given the high coverage of known proteins by the algorithmically derived exon graph, we turn now to the results of mass spectrometric annotation with the exon graph.

Search Results

We obtained ~18.5M spectra from various tissue types and searched them against the exon graph. Searches of this large data-set were run over a grid of 1.6GHz compute nodes (FWGrid Project). The average search time on a node was approximately 2.5 seconds per spectrum. Low-quality matches were filtered out a threshold based on the distribution of match scores (see Methods). Matches shorter than 8 amino acids were discarded due to the difficulty in assigning short peptides to a unique locus.

Each annotation includes the genomic location of the peptide. We compare these loci to the chromosomal locations of known proteins. We then categorize peptide matches based upon their relationship to known genes (see Methods). Recall that the human genome is heavily annotated. Therefore, the degree to which known proteins are covered by annotations from this data-set is a reasonable estimate of our coverage of the full proteome. See Figure 7.4 for an initial breakdown of the results. The majority (89%) of peptides match known genes. Of these, 24% span an exon boundary, confirming splicing events at the protein level. A total of 121 peptides (in 1,517 spectra) span two exon boundaries; these represent cases where a tryptic peptide fully spans a short exon. A total of 11,050 splice events are confirmed by identified peptides. Given that only ~20% of the exon graph corresponds to known proteins, the enrichment for known genes suggests that protein-coding regions of unannotated genomes can be discovered by these methods. Those peptides which do not match known genes may discoveries of novel exons, or novel splicing events; these cases are discussed after the results from known genes.

Protein coverage

The search results include 6,252 proteins confirmed by two or more distinct peptides, and a total of 3,745 proteins matched by five or more distinct peptides. As noted earlier, we select a minimal set of proteins which account

for spectrum annotations. This allows us to avoid listing records corresponding to multiple isoforms of the same protein unless both forms are in fact present.

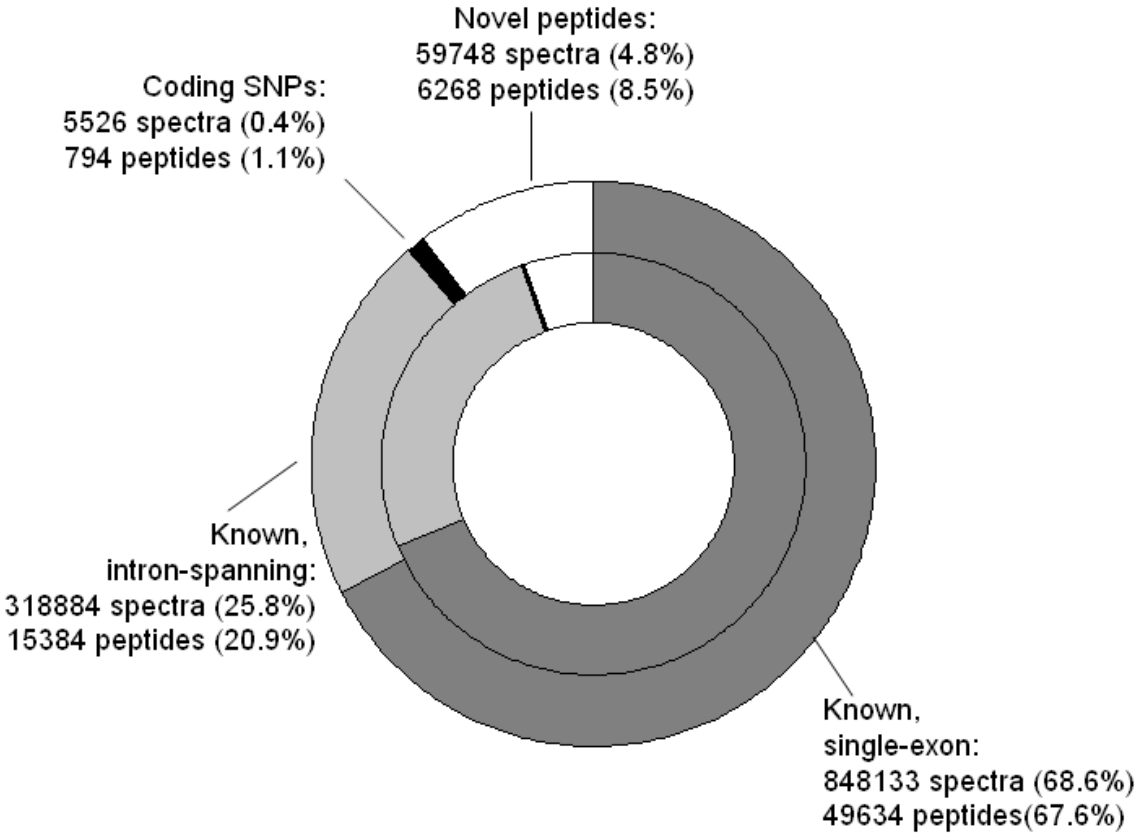


Figure 7.4: Categorization of search results by their relationship to known proteins. The inner ring shows findings at the spectrum level, the outer ring shows findings at the level of distinct peptides. The peptides categorized as “Unknown” include some peptides missing from the IPI database, as well as novel exons.

Because protein abundances within the cell vary greatly, we see extreme variation in the number of spectra matching each protein, with over 25,000 matches from enolase 1, but only one or two matches to other proteins. As with other high-throughput techniques such as cDNA sequencing, the repeated sampling of common elements eventually reaches saturation. We count the number of distinct peptides (from known proteins) discovered for a given number of identifications, and plot the resulting discovery curve. The

discovery rate slows as more peptides are found (Figure 7.5), but is still far from saturation. The discovery curve is fit well by the function $y \sim x^{0.55}$ (correlation coefficient 0.97). Sampling of proteins from more tissue types promises to yield annotations for a wider range of proteins. Based on this discovery curve, we estimate that a tenfold larger data-set should yield high-confidence identifications (five or more distinct peptides) for over 12,000 gene products.

Novel Peptides

Matches to the exon graph which do not correspond to known proteins are potentially of great interest, since they may come from uncharacterized exons or even unannotated genes. We investigated and categorized all peptide matches which are not present in the IPI reference database. We reiterate that searching a larger database increases the likelihood of obtaining a high-scoring match by chance, and we employ several safeguards to filter such matches. First, we use a cutoff based on the false discovery rate (see Methods) to limit the number of such matches. Second, we used the results of a standard database search to filter any novel matches which can be explained away by a known peptide which is missing from the exon graph. An example of a peptide removed by this filtering is LGEHNVEVLEGNEQFINAAK, coded by an intron of protein GI:135523 on the forward strand of chromosome 7. The spectra for this peptide are annotated

by a fragment of porcine trypsin with similar sequence (LGEHNIDVLEGNEQFINAAK).

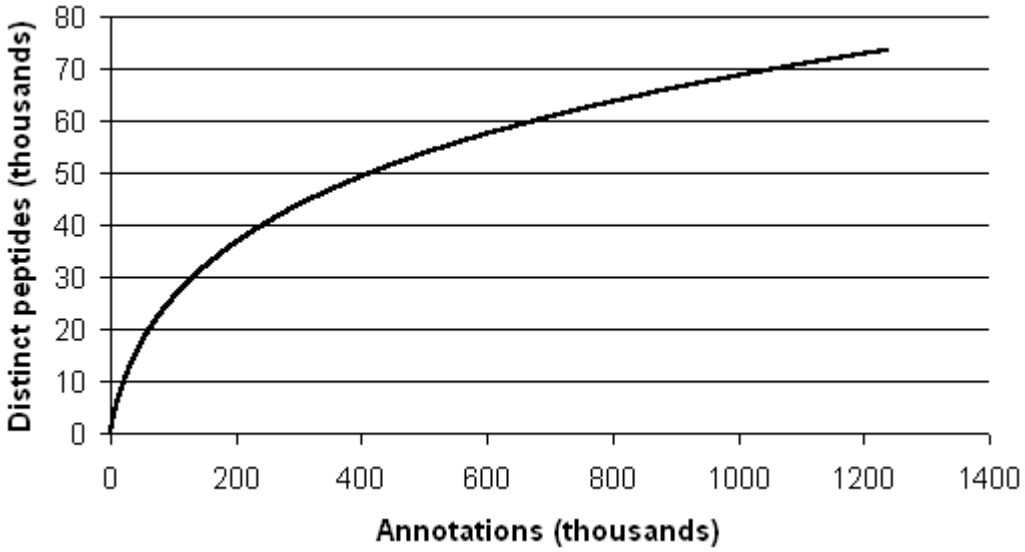


Figure 7.5: Discovery curve, plotting the number of distinct peptides as a function of the number of search hits.

Many of the peptides not present in IPI are present in other isoforms or proteins found in the NCBI non-redundant database. We observe a total of 90 such peptides (1938 spectra). See Supplemental Table 1 for the complete list. These cases illustrate the danger of selecting a limited set of “representative” splice forms for a protein database. After removing such annotations, we retain 58,000 novel spectra (6,100 peptides). We note that incorrect matches are more likely to be novel peptides, since 80% of the exon graph database is novel sequence. Let us conservatively assume the incorrect matches all fall within the novel peptides. Given a 2.5% false discovery rate across all 1.2M annotations, we estimate that 28,000 spectra are correctly annotated by novel

peptides. These correspond to an estimated 3,300 peptides, based on the mean number of spectra per novel peptide. A report of all novel peptides is provided in Supplemental Table 2.

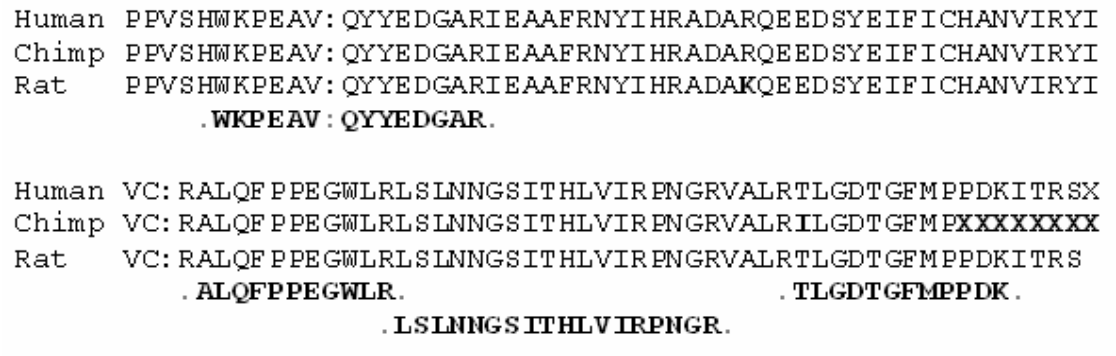


Figure 7.6: Novel exons are supported by peptide identifications and by sequence homology. Above is a multiple alignment for hypothetical protein sequences from chimp (gi:55639283), rat (gi:62531299), and human (genome translation, similar to gi:20070384). Introns are indicated by colons. The peptides identified from mass spectra are indicated below the protein sequence. The novel 3' exon is supported by three peptide identifications, as well as >95% amino acid sequence conservation across species.

In the remainder of our analysis, we restrict our attention to those novel peptides strongly supported by additional lines of evidence. We find evidence for novel exons (or extensions of known exons) in 16 genes. These instances are supported by sequence homology, and by the discovery of one or more peptides in close proximity along the genome. The discovery of translated peptides demonstrates that these sites are indeed exons, and not conserved non-coding sequences. See Figure 7.6 for an example of the evidence for one exon.

Table 7.2: Summary of evidence for additional exons (or exon extensions) in known genes. The genomic coordinates of one peptide representative are shown for each gene. *This exon is present in the updated protein record (GI:113428129)

IPI ID	Gene Symbol	GenBank ID	Spectra	Peptides	Chr.	Location	Annotation
IPI00038698.1	C3orf63	5881256	18	4	3-	56678776-56678842	Two additional 5' exons
IPI00062325.1	SLC3584	18087849	8	2	5+	139926486-139926516	Translation upstream of annotated start
IPI00643156.1	PHF10	74744253	23	1	6-	169936606-169936646	Additional 5' exon
IPI00106642.4	DPYSL2	62087970	75	6	8+	26427785-26427821	Additional 5' exon
IPI00386119.1	SF1	42544130	22	2	11-	64289956-64290070	Different reading frame
IPI00168158.4	C12orf51	74730080	9	4	12-	111183646-111183706	Additional 5' exons
IPI00063242.3	PGAM5	20070384	17	3	12+	131907713-131907749	Additional 3' exon
IPI00004273.5	RBM25	68068009	19	3	14+	72612805-72613862	Extension of 5' exon
IPI00465071.2	TBC1D10B	68534049	35	6	16-	30288483-30288528	Additional 5' exon
IPI00024425.2	KIAA0664	34531906	10	2	17-	2561651-2561693	Additional 5' exon
IPI00164623.4		*34531906*	*10*	*2*	*19-*	*2561651-2561693*	*Additional 5' exon*
IPI00016250.3	FXR2	90177782	13	2	17-	7458719-7458755	Extension of 5' exon
IPI00029863.3	WDR81	74755061	28	6	17+	1575345-1575414	Additional 5' exon
IPI00029863.3		*74759806*	*28*	*6*	*17+*	*1575345-1575414*	*Additional 5' exon*
IPI00295502.3	WIZ	113428129	12	2	19-	15400152-15400188	Exon between exons 3, 4
IPI00295502.3		*89052386*	*12*	*2*	*19-*	*15400152-15400188*	*Exon between exons 3, 4*
IPI00045360.1	CIC	116241300	32	4	19+	47468138-47468195	Two additional 5' exons
IPI00045360.1		*74724286*	*32*	*4*	*19+*	*47468138-47468195*	*Two additional 5' exons*
IPI00258168.6	RBM9	29840825	16	2	22-	34748835-34748901	Additional 5' exon
IPI00158615.5	THOC2	41702296	95	1	X-	122566242-122566278	Additional 5' exon

Table 7.2 summarizes these exon discoveries. While the main purpose of our project is the preliminary annotation of non- or sparsely-annotated genomes, the discovery of new exons on the human genome demonstrates the power of the technique. In most cases, the novel translation is immediately

upstream of known exons. We note that many of the reference protein sequences are derived from cDNA sequences. The 5' portions of such sequences are often inferred or absent due to truncation of cDNA. In addition, predicted translation start sites are often incorrect. With the exon graph, we can use mass spectra not only to confirm translation of these genes, but to correct their sequence annotations. Supplemental table 3 reports the peptide hits to these novel exons, as well as peptides from the known exons of the protein. Supplemental Figure 2 illustrates one such case.

Two peptides were observed which fall within splicing factor 1 (GI:42544130), but not in the annotated reading frame. These peptides are of particular interest since they fall within one of the genomic regions selected by the ENCODE project (Consortium, 2004).

Alternative splicing

Evidence for alternative splicing normally comes from mRNA sequencing projects, which may include prespliced or contaminating sequences. Mass spectrometry data can confirm the presence of specific isoforms in a sample at the protein level. Of our peptide matches, roughly 25% span at least one putative intron. Overlapping exon predictions and EST alignments can produce unreasonably short exons in the database; therefore, we discard peptides undergoing two splice events within 15 base pairs of each other.

We examined our search results for evidence of alternative splicing. We consider all splice donors and splice acceptors which have multiple partners. We ignore matches where the splice boundaries are not part of a known protein, or where the peptide covers six or fewer base pairs on either side of the intron. We highlight a total of 40 instances of alternative splicing in this way. We report these events in Supplemental Table 4.

In 24 of these instances, only one of the two isoforms is present in the IPI database. As s conservative filter, we report such splice junctions only if they are supported by EST evidence and/or supported by sequences in the NCBI non-redundant database.

Polymorphisms

Each known coding SNP produces a “bulge” in the exon graph, where a peptide sequence may not match the genomic sequence. A total of 308 such polymorphisms in known genes were evidenced by at least two spectrum hits (See Supplemental Table 5). For 94 of these cases, both alleles of the SNP were observed. In addition, 221 sites were observed where the observed peptide matches the genomic sequence, rather than the protein from the IPI database. These sites may correspond to SNPs, or simply to sequencing errors. We note that many protein records are derived from error-prone sources such as single-pass cDNA sequencing.

Hypothetical proteins

Many protein records in the IPI database are derived from high-throughput cDNA experiments or computational gene predictions. Identification of peptides from these proteins serve as confirmation that the locus in question is, in fact, a protein-coding gene. We examined all search results which correspond to proteins with annotations of the form “Hypothetical protein” or “Putative protein”. We disregarded any search hits which also match “non-hypothetical” proteins, due either to exons shared with other proteins or multiple occurrences of the peptide within the database. The search results confirm many hypothetical proteins. A total of 224 proteins are matched by a minimum of five spectra from at least two distinct peptides. We omit from this list any sequence present in RefSeq (Pruitt et al., 2005) with an annotation other than “REFSEQ PREDICTED” or “REFSEQ MODEL”. Supplemental table 6 summarizes the results. This may be the first confirmation of these protein sequences at the level of translation. Supplemental Figure 3 shows coverage of one such protein.

Refining gene predictions

Here we address the question: Can de novo gene finding be improved by incorporating evidence from mass spectrometry? Earlier research has demonstrated the effectiveness of incorporating additional lines of evidence, such as comparative genomics, to improve gene prediction (Korf et al., 2001). By searching mass spectra against our database of putative proteins, we accumulate evidence supporting putative exons and introns. When predicting

genes, GeneID first identifies putative exons, then assembles the exons into a collection of genes. We re-score the predicted exons before gene assembly, in an effort to improve the accuracy of gene prediction. We boost exon and intron scores based on the number of spectra matched by corresponding peptides, and on the quality of these matches (see Methods).

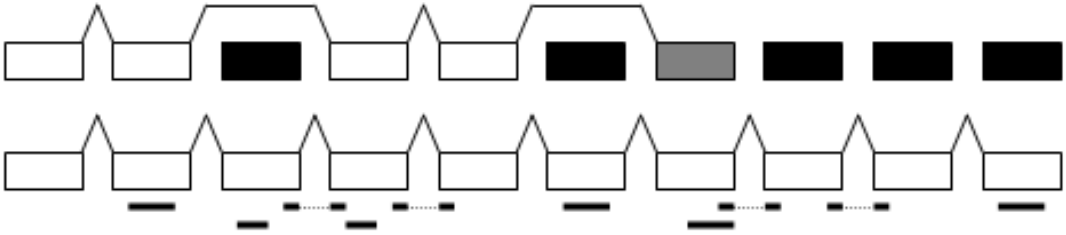


Figure 7.7: Diagram of gene prediction results for IPI00017381.1, before (above) and after incorporation of MS/MS results. Correctly predicted exons are shown in white, missed exons in black. A partially-correct exon is shown in grey. Peptide identifications are indicated below their exons (and spanned introns). After exons are rescored using the identified peptides, the full gene is predicted correctly. (Figure not to scale)

We ran GeneID on the genomic intervals containing 1,386 protein-coding genes. We selected genes for which one or more peptides were mapped to the coding region, and for which a single splice isoform was known (from the IPI database). We then re-scored all predicted exons by incorporating peptide matches from our database search. The sensitivity and selectivity of gene assembly improved (Table 7.3), with a gain of 863 correctly identified exons. The improvements are greatest for proteins that are well-sampled (data not shown). We also note that since we examine a broad selection of gene, including 100 which span over 100,000 base pairs, accuracy on this corpus

may be lower than on other test sets. Figure 7.7 shows an example of a gene prediction improved by this method.

Table 7.3: Integration of mass spectrometry search results improves the gene prediction accuracy. A total of 875 correct exons are added to gene predictions by incorporating MS/MS data.

	Sensitivity	Selectivity
Exons	68.1	75.8
Exons (with rescoring)	74.3	77.2
Nucleotides	84.5	79.5
Nucleotides (with rescoring)	88.5	80.3

In a few cases (20 genes), predictions worsened after rescoring. The peptide annotations used for these genes appear to be correct. In most cases, an incorrect exon (which overlaps the true exon) was boosted and selected for the final gene prediction. One instance of a peptide mapped to an incorrect splice boundary was also observed. Further work will focus on improved incorporation of MS/MS data, and integration of MS/MS search results alongside other data that can corroborate exons (ESTs and comparative genomics). We anticipate that refinement of the algorithm as well as acquisition of additional spectra will improve results.

7.4 Discussion

Delineating the protein-coding genes within a eukaryotic genome remains a complex and labor-intensive process. To cite one example, a human-curated annotation of the human X chromosome required an estimated 15,000 person-hours (Harsha et al., 2005), much of which was spent resolving the set of coding regions. Because automated annotations are the foundation

which biologists later build upon, high-throughput methods to generate and refine annotations are needed. This study demonstrates that with a few mass spectrometry experiments, automated analysis can recapture many of the gene annotations that have been made by painstaking efforts. Even on the extensively-studied human genome, we discover genes and exons which have not yet been deposited in sequence databases. The majority of our data was drawn from two tissue sources (kidney cells, and blood plasma). Consideration of other tissues or enrichment for specific organelles will surely expand our picture of the proteome. On a less thoroughly annotated genome, we expect to see a readout of many more novel genes.

The exon graph is a compact representation of protein splice isoforms and polymorphisms. We observe a near-tenfold reduction in database size between dbEST and the exon graph. We emphasize that this is difficult to accomplish with a typical database, stored in FASTA format. Enumeration of all protein sequences greatly increases search time, and creates confusion when matches to dozens of “records” are explained by one gene. Many databases sidestep the problem by including one or two representative sequences for each protein, but this approach carries omits isoforms and polymorphisms. Algorithmic improvements are one way to reduce redundancy from linear protein databases (Edwards and Lippert, 2004). We believe that, if available in a standard vendor- and tool-independent file format, exon graph databases may be of general interest to proteomics researchers.

We used two data sources which complement each other to construct the exon graph. An advantage of the EST evidence is that it includes evidence for introns. Short exons, or exons with unusual hexamer count, are difficult to identify de novo but may be covered by ESTs. A limitation of EST evidence is that ESTs may not be available for all genes, and may not cover the 5' portion of a gene. Many genes are transcribed only in certain tissues or under certain conditions, and may never have been captured as ESTs. Another drawback of EST data is the presence of unprocessed and truncated transcripts, as well as genomic contaminants. Exon predictions have the advantage that they explicitly indicate reading frame. Database construction proceeds from putative exons and introns, independent of any specific exon prediction method. We are working to integrate other signals including the output from multiple gene finding programs, evolutionarily conserved regions, etc.

Our results include 40 instances of alternative splicing. We emphasize that we have highlighted only those instances where two splicing events are observed at the same locus. These results directly confirm both splice events. Many other peptide identifications are unique to splice isoforms that are not considered standard, giving indirect evidence of alternative splicing. It is notable that many splice isoforms differ by the inclusion of a single amino acid. These are cases where two splice donor (or acceptor) sites are present, separated by 3 base pairs. Some isoforms of biological significance differ by presence or absence of a single amino acid (Tadokoro et al., 2005).

Fully characterizing splice events from tryptic peptides gives rise to a phasing problem which may be avoided by top-down mass spectrometry of complete proteins (Roth et al., 2005). Mass spectrometry can reliably demonstrate the presence of protein isoforms, but confirming their absence is problematic (Godovac-Zimmermann et al., 2005). Sequence-based methods remain important, particularly for splice events that take place in the untranslated region of genes.

Our focus in this paper is on cataloging coding exons and splice events. We note that mass spectrometry can measure other types of information that are invaluable for annotation of genes. These include post-translational modifications (Jensen, 2006), proteolytic cleavages (e.g. of signal peptides) (Gupta et al., 2006), subcellular localization (Dunkley et al., 2006), and relative protein expression levels between tissues (Lill, 2003). These topics are a subject of ongoing research. Our search did not consider post-translational modifications explicitly. Some modified peptides were annotated with a sequence with the same mass as the true (modified) peptide. For example, the putative peptide ASVVAVSDGVIK matches the N-terminally acetylated peptide from cofilin (A+42SGVAVSDGVIK). The putative peptide DELHIVEAEAVYYKGSPIK matches a modified peptide DELHIVEAEAM+16NYKGSPIK from nucleoplasmin. Using other algorithms developed in our lab (Tsur et al., 2005), we are searching these same data-

sets for known and unknown post-translational modifications. Similar studies are underway for bacterial genomes (Gupta et al., 2006).

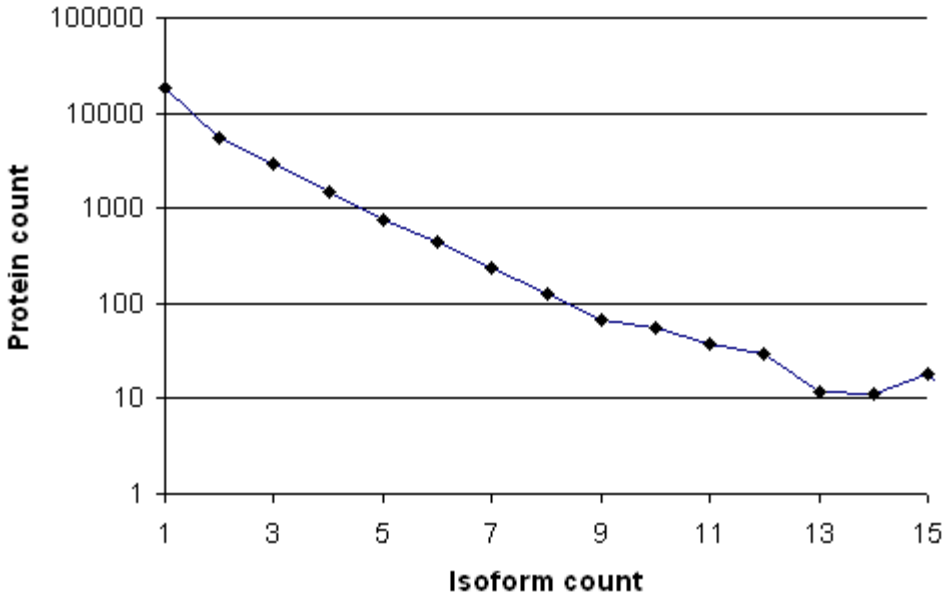
We argue that high-throughput proteomics experiments should accompany each genome sequencing project. Mass spectrometry is a practical technique for annotating protein-coding regions. The search is able to tolerate a substantial overhead of “noise” in exon predictions. In addition, the technique is orthogonal to standard transcript-level methods such as cDNA sequencing. Mass spectrometry complements other experimental methods. With recent advances in instrumentation, the data volume we consider in this paper can be produced in 10 instrument-weeks with two person-weeks of labor. Scaling up mass spectrometry experiments to help annotate a large portion of proteomes is an attractive prospect at feasible cost.

Acknowledgments

S.T. is supported by NSF IGERT training grant DGE0504645. This research was supported in part by NIH (RR016522-04A1), and by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. Part of this investigation was supported using the computing facility made possible by the Research Facilities Improvement Program Grant Number C06 RR017588 awarded to the Whitaker Biomedical Engineering Institute, and the Biomedical Technology Resource Centers Program Grant Number P41 RR08605 awarded to the National Biomedical Computation Resource, UCSD, from the National Center for Research Resources, National Institutes of

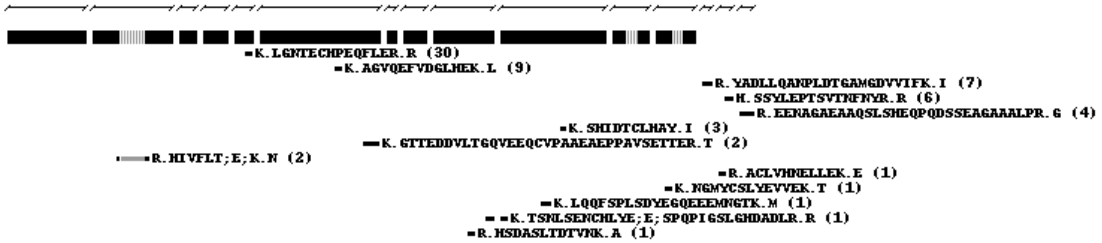
Health. This chapter was published as "Improving gene annotation with mass spectrometry". Tanner, Stephen and Shen, Zhouxin and Ng, Julio and Florea, Liliana and Guigo, Roderic and Briggs, Steven P and Bafna, Vineet 2007. Genome Research 17(2), 231-239. The dissertation author was the primary author of this paper.

Supplemental Figures

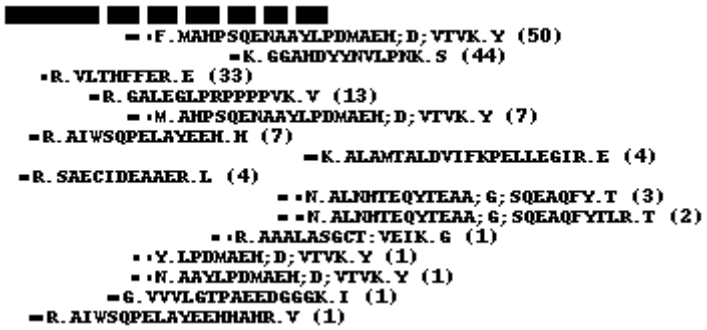


Supplemental Figure 7.1: Histogram of isoform counts in IPI. Two proteins are considered to be isoforms of each other if their genomic intervals overlap. These counts reflect only the isoforms deposited in the IPI database, and under-estimate the number of predicted isoforms.

Supplemental Figure 2:



Supplemental Figure 7.2: Evidence for novel exons upstream of the annotated start site of retinoblastoma-associated protein RAP140 (gi:5881256). Matched peptides are shown below their corresponding genomic location, with spectrum counts indicated in parentheses. Those peptides which match the reference protein sequence are also shown.



Supplemental Figure 7.3: Translation of hypothetical protein IPI00217852.6 is confirmed by annotation of peptides from all seven exons. Three of six splicing boundaries are confirmed at the level of translation. Introns not shown to scale. The number of spectra for each matched peptide is shown in parentheses.

Chapter 8: Generalized gene set queries for microarray analysis

8.1 Introduction

High-throughput experiments such as microarrays generate readings for thousands of genes at once. Individual gene readings measure the degree to which the gene is up- or down-regulated. Microarray experiments contain significant noise, and typically few genes are found whose expression is significantly changed. Recently, several groups have begun to examine microarray experiments from the perspective of biologically related gene sets. We use the term *gene set query* to describe the problem of searching a microarray experiment against database of biologically meaningful sets of genes. Using gene set queries can greatly boost the statistical power of tests for up- and down-regulation. In addition, the results of a gene set query consists of readily interpretable groups, such as the set of all genes annotated with a GO term. Various algorithms such as GSEA [Subramanian 2005] and PAGE [Kim 2005] are available for this type of search. General tools to assist biologists with the analysis are being developed [Kim 2007], and in this context, quantifying the success of particular algorithms becomes important.

In addition to gene set queries against a database of biologically meaningful sets, we may wish to query a database consisting of other experiments. We use the term *gene vector query* to describe a query against a database whose entries are vectors of gene values. Such a query may aim to find related experiments - for example, queries of signatures against the

Connectivity Map corpus were able to identify compounds with similar effects [Lamb 2006].

The naive null model for a gene set query is that the genes in the set are drawn independently from the overall distribution. However, many gene sets of biological interest consist of co-regulated genes. The expression responses of these genes will typically be highly correlated. This tight correlation may cause us to reject the naive null model with high confidence, even in cases where the genes are not differentially regulated. One way to compensate for this interdependence of genes within a set is through permutation of class labels [Tibshirani 2007]. However, a disadvantage of permutation testing is that it requires a large number of replicates. (Note that for experiments involving fewer than 13 microarrays, fewer than 1,000 distinct permutations exist). We find that it is most effective to calibrate p-values for each set (or vector) in the database using a large corpus of experiments. Once this calibration is performed, queries can be performed with higher accuracy than permutation tests, and with less computational cost.

Statistical methods which apply to set queries, such as ANOVA, may not apply to vector queries. If different statistical methods are used, the p-values from gene set queries and gene vector queries may not be comparable. For this reason, we developed a tool that can query either sets or vectors, using a common statistical framework. With this tool, one can query microarray readings against a database of sets (as in GSEA), or query

gene signatures against a database of vectors (as used in the Connectivity Map). In addition, our tool can perform gene vector queries (e.g. to find drugs which offset the transcriptional changes associated with a disease). The source code implementing our query tool, GQuery, is available online at http://bioinfo2.ucsd.edu.

We report the results of validation experiments using publicly available microarray experiments from the GEO corpus. By using pairs of related experiments, we quantify the accuracy of queries, and the effectiveness of p-value calibration.

8.2 Methods

We divide the generalized gene query problem into three steps. The first step is the acquisition of a *reading* for each gene. The second step is the calculation of an *enrichment score* for each gene set (or gene vector) in the database. The third step is the conversion of these enrichment scores into p-values. We focus on the last two steps, since this is where our original contribution lies.

Gene Readings

We obtained several publicly-accessible microarray data-sets from the GEO repository [Barrett 2006, Barrett 2007]. Five pairs, for a total of ten experiments, were used in validation, as follows:

- Muscle: Comparison of muscle tissue from young and aged male (GDS287) and female (GDS472) donors [Welle 2003].

- Malaria: Comparison of whole blood from healthy children, and children with mild or severe malaria (GDS1971)
- AD: Comparison of brain tissue from three stages of progression of Alzheimer's disease (GDS810) [Blalock 2004].
- Glioma: Comparison of gliomas of grade III and grade IV tumors with control (non-tumor) cells (GDS1962) [Sun 2006].
- Obesity: Comparison of skeletal muscle tissue from lean and obese male and female donors (GDS268) [Park 2006].

Before our query can be performed, we need a single reading quantifying the degree of up- or down-regulation of each gene. The gene readings for each gene included in the microarray will be represented as a query vector of length N, whose nth value represents the change in transcription of the nth gene. The levels of transcription of each gene in the public data-sets we used were initially quantified using MAS5 [Welle 2002] or related methods. To quantify the up- or down-regulation of each gene, we employed the Cyber-T algorithm [Long 2001]. The t-statistic itself is retained as the reading for a gene. The t-statistic has the advantage that it reflects both direction (up- versus down-regulation) and confidence. A variety of methods are available to quantify up- and down-regulation [Vardhanabhuti 2006], which can be incorporated similarly. In addition, we tried applying log fold change (which reflects only direction) and the Cyber-T p-value (which reflects only confidence). In addition, p-values from SAM [Tusher 2001] were

computed and tested, and found to give similar results to Cyber-T p-values (data not shown).

Enrichment Scores

A database of gene sets was constructed from several sources: GOA [Ashburner 2000], GenMAPP [Dahlquist 2002], HumanCyc [Romero 2005], the BioCarta pathways database, and the TRANSFAC database. Gene identifiers from the source databases, along with Affymetrix microarrays, are mapped to a collection of common identifiers. Because small gene sets are not statistically significant, our queries ignored any set containing fewer than five genes. A total of 4,256 gene sets of sufficient size were available.

Given a vector of gene readings and a gene set, we considered several statistical models for computing an enrichment score for each set:

- Pearson correlation. We construct a binary membership vector for the set. This membership vector's nth entry is 1 if the nth gene is a member of the set, and 0 otherwise. We then compute the Pearson correlation between the membership vector and the query vector. The enrichment score is the Pearson correlation coefficient, r.
- Spearman correlation. As with Pearson correlation, we first construct a binary membership vector for the set. We then compute a Spearman (rank-based) correlation, ρ, between the membership vector and the query vector. The enrichment score is the variable t, defined as:

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}}$$

- One-way ANOVA.
- Z-score of average reading. By the Central Limit Theorem we know that, for sufficiently large k, the average of k randomly chosen gene readings will follow a normal distribution [Grinstead 1997]. Therefore, given a set of k genes, we can compute the Z-score of the average reading. This approach is modeled closely on the PAGE method [Kim 2005].

The two correlation-based methods have the advantage that they apply equally well to queries against a database of vectors. The accuracy of these four methods was compared on a validation set of experiments.

Calibration of p-values

We must translate the enrichment score of each set into a p-value. Because gene sets differ in their size and in their degree of co-regulation, a raw enrichment score which is significant for one gene set may be insignificant for another. Therefore, we calibrate p-values for a gene set using a corpus of microarray data. Given gene reading vectors X_1, …, X_c, we compute the gene set's enrichment score S_i for each vector X_i as described above, then model the distribution of these scores. The fitted distribution allows us to translate further enrichment scores for this gene-set into p-values.

The Connectivity Map (CMAP) corpus consists of a total of 463 microarray experiments involving the exposure of human cell cultures to various perturbagens [Lamb 2006]. As a second corpus, we obtained all GEO data-sets available for the Affymetrix HG-U133A chip (GPL96) as of February 1st, 2007. The SOFT-format files for each data-set were parsed, and expression differences were measured using Cyber-T for each pair of sample sets which (a) contained three or more entries per set, and (b) were disjoint. To avoid over-representing particular treatments in our corpus, we selected at most three such comparisons per data-set. The resulting corpus contains a total of 285 gene vectors.

Under reasonable assumptions, the theoretical distribution of Pearson correlation scores follows a normal distribution whose variance is inversely proportional to the number of genes [Press 1992]. In practice, the distribution of Pearson correlation scores for biologically relevant gene sets across the corpus is indeed fit well by a normal distribution, but with a standard deviation that varies between gene sets. To enforce a mean of zero, we augment the list of enrichment scores with $-S_i$ for each score S_i. The variance of the enrichment score distribution correlates with size ($r = 0.41$), but is also affected by the degree of co-regulation.

We evaluated the quality of the fit using the Anderson-Darling statistic. The median values of A^2 for the GEO and CMAP corpora were 0.87 and 0.60 respectively, corresponding to a p-value of ~0.3. Similarly, the median

Kolmogorov-Smirnov p-values for the two corpora were 0.57 and 0.41, respectively. The standard deviations of enrichment scores for gene sets across the two corpora are tightly correlated (r = 0.87). This suggests that a sufficiently large and diverse corpus provides a reasonable measurement of the degree to which genes in a set are co-regulated.

The Spearman correlation scores, t, fit a normal distribution (median K-S p-value 0.89). The Z-scores for gene sets also fit a normal distribution (median K-S p-value 0.41). The scores from one-way ANOVA are approximated reasonably well by an exponential distribution.

Evaluation of query algorithm

In order to quantify the performance of our queries, we compared the lists of gene sets obtained for related experiments. Two experiments are considered “related” if they involve similar treatments. If a gene set is found to be enriched in related experiments, we have increased confidence that the gene set is indeed undergoing a biologically relevant change in expression. By contrast, we expect to see few shared gene sets in the query results for unrelated experiments.

We listed the top N gene sets reported as enriched for any of the ten validation experiments. We then enriched any gene sets shared between a pair of experiments. These shared gene sets were considered valid if the experiments were paired (e.g. obesity in male and in female), and invalid otherwise. On average, half of all spurious shared gene sets will fall into

related experiments by chance, and half will come from unrelated experiments. Therefore, the false discovery rate [Benjamini 1995] is equal to 2*F. A successful gene query tool is one which finds obtains many shared gene sets at a given false discovery rate cutoff; we selected a cutoff of 10% for our comparisons. This validation procedure was used to compare the relative effectiveness of enrichment scores and of p-value recalibration methods.

Gene vector queries

The Pearson and Spearman correlation enrichment models apply equally well to queries against a database of vectors. As a test of this procedure, we measured differential expression using Cyber-T for all data sets in the GEO corpus (described above), then performed and all-against-all vector query. We modeled the distribution of correlation values R for a given data-set X with a normal distribution. This enables us to compute the p-value, $P_X(R)$, for a given value of R. When comparing vectors X and Y, the p-value for the association of X and Y is defined as: $P(X,Y) = \sqrt{P_X(R)P_Y(R)}$. This score reflects the significance of a particular correlation R relative to the correlation values observed for X and Y across the entire corpus. In the absence of a training set of query results, we evaluated the query results for several GEO data-sets to determine whether they were biologically reasonable.

8.3 Results

The distribution of enrichment scores for a given gene set across the corpus reflects the co-regulation (formally: the correlation in transcription changes) of the genes across various treatments. Figure 8.1 compares these distributions for two gene sets: A set of 118 genes related to oxidative phosphorylation, and a large set of 1,222 genes related to mRNA processing. As an example of how p-value calibration provides improved query results, consider a data-set (GDS287) comparing muscle tissue from young and aged males. Using a naive query that performs no p-value calibration, we obtained a p-value of 1.2×10^{-34} for the mRNA processing set, much lower than the value for oxidative phosphorylation (6.4×10^{-11}). Similar results were observed in [Kim 2005]. However, after calibration against the CMAP corpus, this ordering is reversed, and the p-value for mRNA processing is no longer significant after correcting for multiple hypothesis testing. Naive queries frequently detect the mRNA processing set as enriched - indeed, it receives an uncorrected p-value below 0.05 in the *majority* of the 463 CMAP experiments. This demonstrates that the naive null hypothesis does not suffice to find enrichment scores significant for a particular gene set.

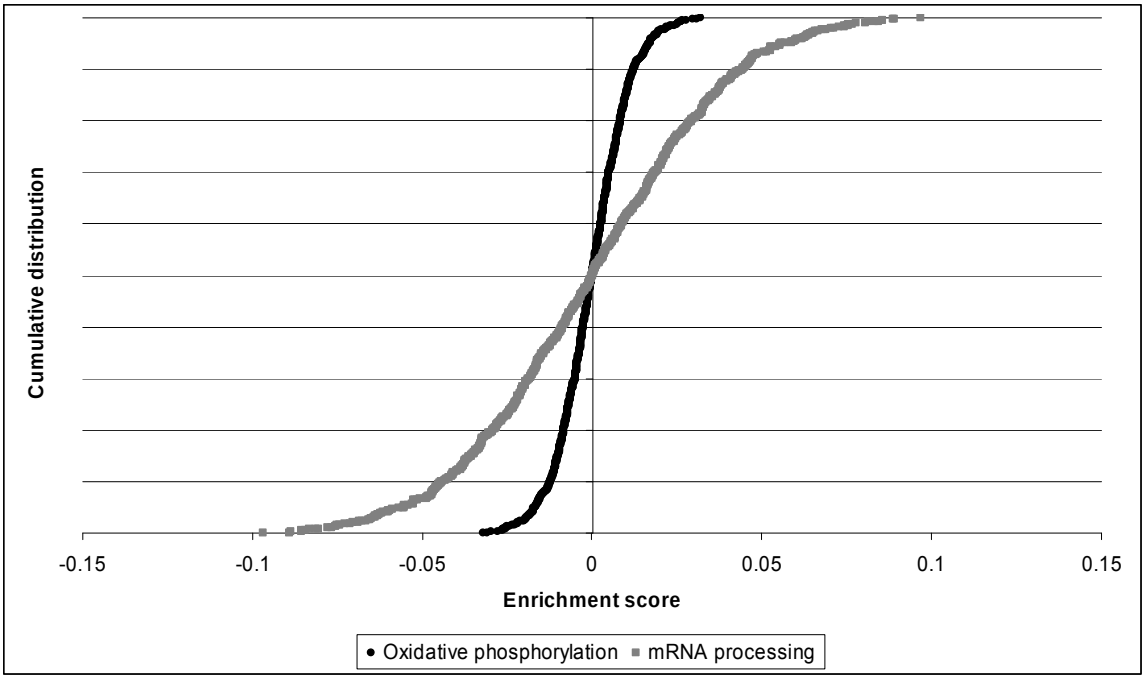


Figure 8.1: Cumulative distribution functions of enrichment scores for two gene sets across the CMAP corpus. By comparing distribution scores against the empirical distribution of scores across the corpus, we obtain an accurate p-value.

As described in the Methods, we computed the false discovery rate for queries across pairs of related and unrelated experiments (Figure 8.2). The results demonstrate that either large corpus provides a reasonable training set for p-value calibration. The GEO corpus has the advantage that it includes a wide array of treatments and tissue types, and that it uses t-scores (available for the GEO corpus) over log fold changes (available for the CMAP corpus). On the other hand, the CMAP corpus is somewhat larger, and has the advantage that it was generated by one lab with high reproducibility. The GEO corpus was arguably more effective, as measured by the slower decrease in accuracy. However, when we list the top 10 gene sets for these

experiments (as measured by product of p-values), the lists reported using calibration against the CMAP corpus were most biologically reasonable.

Calibrating p-values using a corpus of experiments is less expensive computationally than using permutation of class labels, particularly if many queries will be run against the same database of gene sets. The initial corpus calibration is time-consuming (requiring approximately 1 day of running time on a typical desktop PC), but need only be done once for each gene set. Perhaps surprisingly, our results show that calibrating p-values across a corpus of experiments yielded higher accuracy than generating p-values by permuting the class labels. However, we note that permutation of class labels is clearly more effective than no p-value calibration at all.

In a related experiment, we compared the query accuracy obtained when using the t-statistic, Cyber-T p-values, or log fold change as our gene values (Figure 8.3). Queries using the t-statistic or Cyber-T p-value are noticeably more accurate than those driven solely by log fold change. This reflects the large amount of noise in fold-change measurements for genes expressed at a low level. A final validation experiment compared the accuracy obtained using several different enrichment models (Figure 8.4). Pearson correlation is more accurate than Spearman correlation, as might be expected when comparing parametric and non-parametric models.

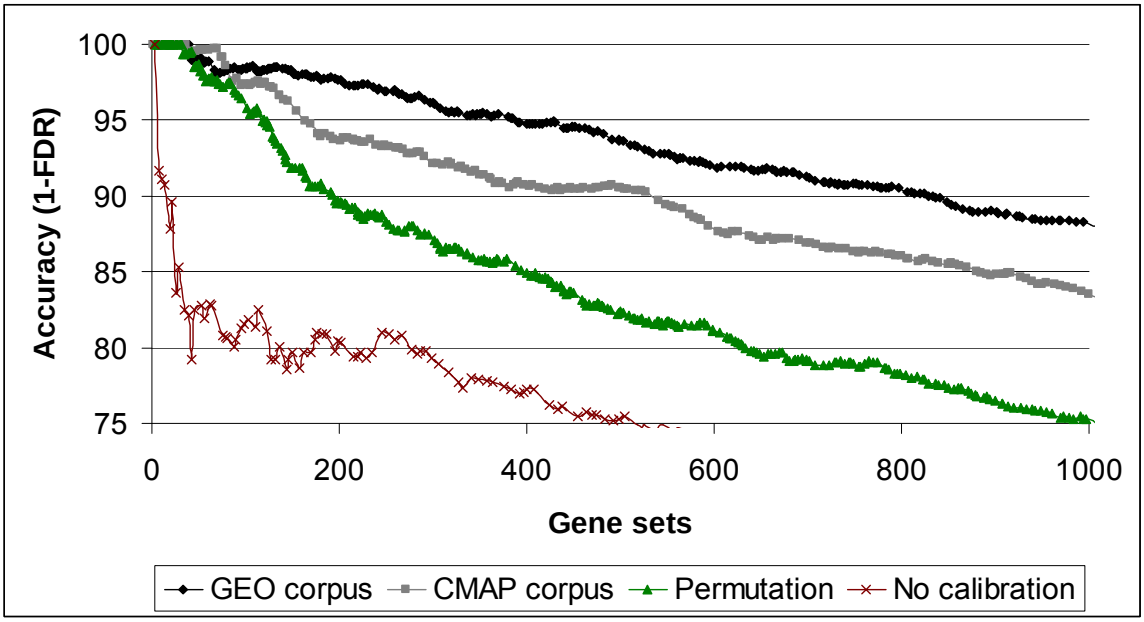


Figure 8.2: Accuracy of gene set queries, as measured by pairs of related experiments. As the p-value cutoff increases, a greater number of gene sets are identified, but at the cost of a decrease in accuracy. If p-values are calibrated using a training corpus, gene set queries produce a significant number of results at high accuracy.

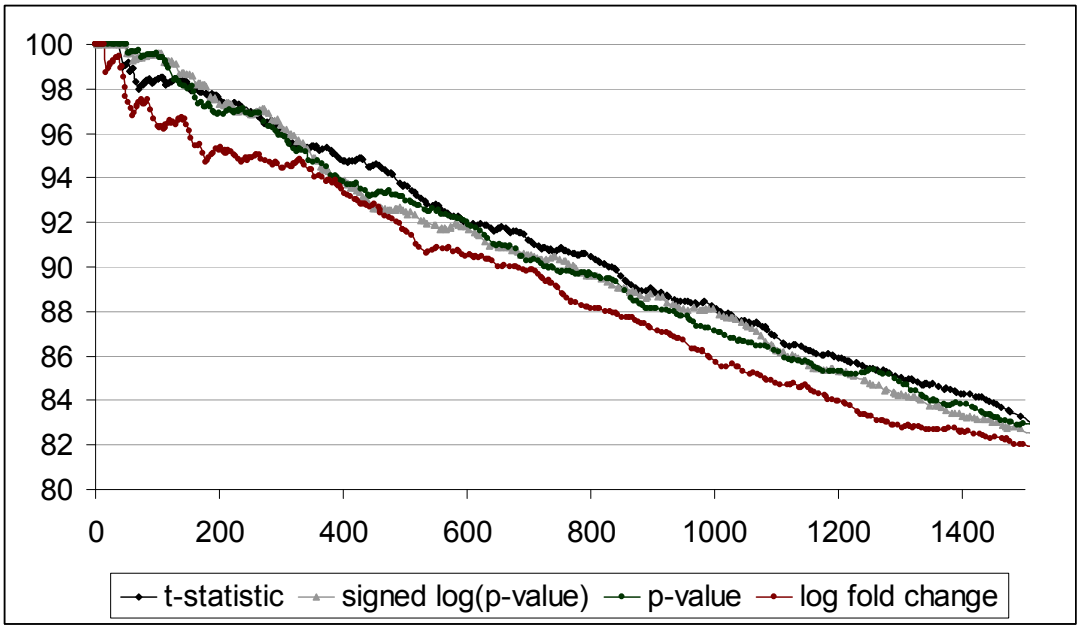


Figure 8.3: Comparison of query accuracy, on the validation set, with p-values calibrated against the GEO corpus. Queries based upon Cyber-T were generally very effective. Using log fold changes as gene values was least effective, due to the significant noise in log fold change for genes with low expression.

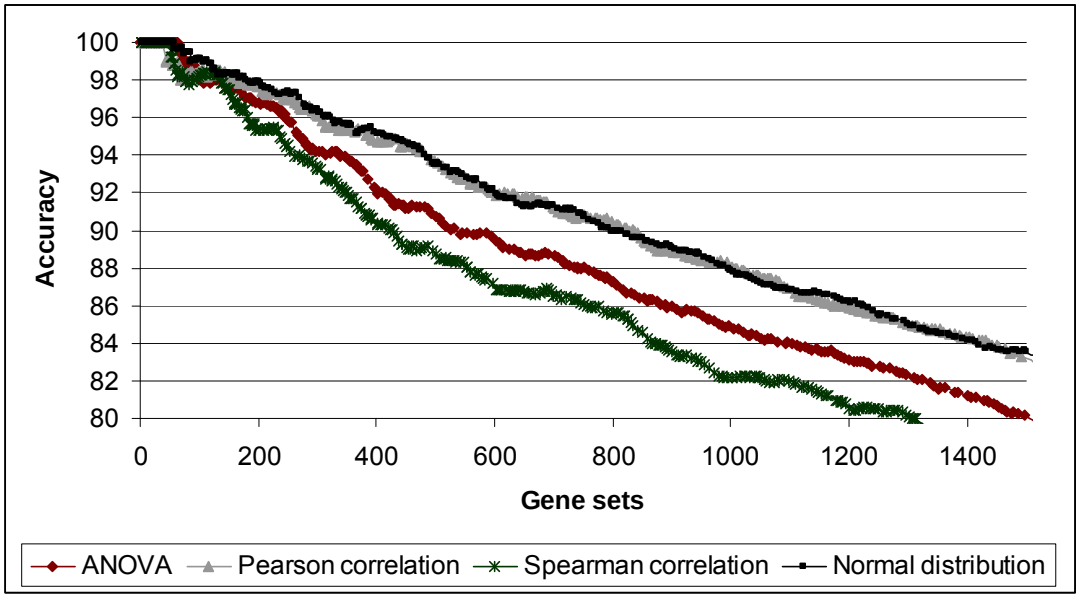


Figure 8.4: Comparison of query accuracy, for the validation set, using various enrichment models. Pearson correlation, and the fit to the normal distribution, obtain very similar results. An additional advantage of Pearson correlation is that it is effective for queries against a database of vectors, as well as gene sets.

Gene set results

Table 8.1 lists several of the top gene sets returned for the five pairs of experiments described in Methods. (A full table of gene sets, together with the corresponding Affymetrix probe IDs, is available in the Supplemental Material). Some gene sets of clear biological interest arise. For example, the set of genes annotated with the biological process "Long-term memory" was down-regulated in the Alzheimer's Disease samples. Gene sets related to immune response were differentially regulated in response to malaria infection. As reported previously [Kim 2005], gene sets related to glycolysis and the TCA cycle are differentially regulated in young and aged muscle.

The original study of the Alzheimer's Disease samples [Blalock 2004] identified several differentially expressed gene sets using a modified Fisher's exact test. Several biological processes were identified again by our study, including downregulation of ATP biosynthesis and GPCR signaling, and upregulation of apoptosis. We believe that the reliability of our results is improved by the use of a parametric statistic, as well as a more reasonable null model.

In practice, some gene sets overlap significantly. For instance, given parent and child GO terms, the gene set corresponding to the child term will be properly contained in the set corresponding to the parent. In examining query results, one should bear in mind that search hits to different gene sets are often not independent events. Our in-house reporting software for gene

set queries reports any significant (>30%) member overlap between sets, omits duplicated gene sets with identical sets of members, and (optionally) filters out gene sets with very high similarity to those already reported.

Table 8.1: Top-scoring differentially expressed gene sets found for pairs of related microarray experiments (see Methods). The p-value reported is the product of the p-values for the two experiments.

Experiment	Rank	p-value	Name	Source
Muscle	1	7.89E-10	Glycolysis_and_Gluconeogenesis	Source24
Muscle	2	6.93E-09	Costamere: CC	Source18
Muscle	3	4.37E-07	superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass	Source44
Muscle	4	4.86E-07	Contractile Fiber Part: CC	Source18
Muscle	5	6.54E-07	Z Disc: CC	Source18
Muscle	6	9.20E-07	Small Leucine-Rich Proteoglycan (SLRP) Molecules	Source5
Muscle	7	1.69E-06	aspartate degradation II	Source44
Muscle	8	4.80E-06	Myofibril: CC	Source18
Muscle	9	4.87E-06	gluconeogenesis	Source44
Muscle	10	5.38E-06	Contractile Fiber: CC	Source18
Malaria	1	1.60E-08	Immune Response-Regulating Signal Transduction: BP	Source18
Malaria	2	1.60E-08	Immune Response-Regulating Cell Surface Receptor Signaling Pathway: BP	Source18
Malaria	3	1.60E-08	Immune Response-Activating Signal Transduction: BP	Source18
Malaria	4	1.60E-08	Immune Response-Activating Cell Surface Receptor Signaling Pathway: BP	Source18
Malaria	5	1.60E-08	Antigen Receptor-Mediated Signaling Pathway: BP	Source18
Malaria	6	1.67E-08	T Cell Receptor Signaling Pathway: BP	Source18
Malaria	7	1.76E-08	Regulation Of T Cell Receptor Signaling Pathway: BP	Source18
Malaria	8	2.69E-08	Regulation Of Antigen Receptor-Mediated Signaling Pathway: BP	Source18
Malaria	9	2.64E-07	Activation Of Csk By cAMP-Dependent Protein Kinase Inhibits Signaling Through The T Cell Receptor	Source5
Malaria	10	5.02E-07	Locomotion: BP	Source18
AD	1	1.13E-11	Proton-Transporting Two-Sector ATPase Complex: CC	Source18
AD	2	1.13E-11	Hydrogen-Translocating V-Type ATPase Complex: CC	Source18
AD	3	9.31E-11	Long-Term Memory: BP	Source18
AD	4	4.91E-10	aspartate degradation II	Source44
AD	5	1.78E-09	Proton-Transporting ATP Synthase Complex: CC	Source18

Table continued

Experiment	Rank	p-value	Name	Source
AD	6	1.78E-09	Proton-Transporting ATP Synthase Complex (sensu Eukaryota): CC	Source18
AD	7	1.78E-09	Hydrogen-Translocating F-Type ATPase Complex: CC	Source18
AD	8	1.95E-09	Hydrogen Ion Transporter Activity: MF	Source18
AD	9	6.44E-09	Monovalent Inorganic Cation Transporter Activity: MF	Source18
AD	10	7.75E-09	Ubiquinol-Cytochrome-C Reductase Activity: MF	Source18

Vector query results

As described in Methods, we computed the correlation of the Cyber-T vectors for all pairs of experiments in the GEO corpus. Given these values, we performed vector queries, to identify all experiments significantly related to a microarray experiment. These experiments may perturb the cell similarly (e.g. exposure to related compounds), or may touch similar pathways with opposite effects (e.g. disease response versus exposure to a treatment). This query involved approximately 800,000 pairwise comparisons, and required 3 CPU days of running time on a computer cluster. We examined the query results for a false discovery rate up to 10%.

We examined the query results for the paired experiments (see Methods). As expected, the two muscle data-sets (GDS287 and GDS472) are related to each other (p-value 3.55×10^{-6}). Two other data-sets were significantly similar - an unrelated separate study of sarcopenia (GDS749) and a study of the effects of exercise on aged muscle (GDS1340). These results

show the effectiveness of exercise in offsetting age-associated muscle loss at the transcriptional level. The AD experiment was similar to one experiment on bipolar disorder (GDS2190), suggesting these disorders may activate a common damage response mechanism. No significant hits were found for the other paired experiments.

Our full GEO-against-GEO results are reported in Supplemental Table 3. Relationships between compounds can be discovered by this kind of undirected data-mining. For instance, a close relationship was observed between experiments exposing a prostate cancer cell line to two different androgens: DHT (GDS2057), and methyltrienolone, or R 1881 (GDS536). Hits were also seen for experiments with the transcriptional changes induced by the estrogen hormone estradiol (GDS1549) and by the estrogen receptor agonist tamoxifen (GDS2367).

Other hits come from experiments with related treatments - comparisons of transcription in blood versus liver (GDS1023) and kidney versus liver (GDS1663) were closely related, presumably due to transcription of liver-specific genes. Some of the confident query hits come from experiments from different labs which applied essentially the same treatments - for instance GDS1549 and GDS2367 both measure the effects of estradiol on breast cancer cell lines. In the future, programmatically examining meta-data (e.g. from MIAME) may allow these unsurprising search hits to be filtered from the reported results.

8.4 Discussion

The high-throughput gene query problem can be formulated in several ways, focusing on either gene set queries or gene vector queries. The use of a parametric statistical framework that handles either gene names or gene values is important, particularly when combining heterogenous data-sources (e.g. microarrays and literature-curated gene lists). Queries can rely upon simple metrics such as Pearson correlation when using t-statistics (rather than log fold changes), and when p-values are calibrated properly.

Calibration of gene set queries against a corpus of experiments provides much more accurate results than using a naive null model. Calibration against a training corpus is certainly not ideal for all situations. In cases where a suitable corpus is not available (e.g. if one is investigating an organism that has not yet been extensively studied), class label permutation is the only practical approach. If a set of genes is not significantly expressed in the training corpus, then the training corpus will not measure the degree of correlation. Therefore, it is desirable to use a training corpus containing as wide a variety of tissues and conditions as possible.

The emergence of large microarray repositories such as GEO provide researchers with a new option: The search for experiments with similar (or opposite) gene changes. Such searches provide an ideal approach to find compounds which offset the gene expression changes associated with

disease states. As with gene set queries, calibration of p-values is crucial when querying by gene values.

The original study of the obesity data [Park 2006] highlighted the differential regulation of GRB14, GPD1, and GDF8. Interestingly, none of these genes are members of the top 50 differentially-regulated gene sets identified by our procedure. The gene SCF, previously reported as significant to glioma development [Sun 2006] was not contained in the gene sets identified on the Glioma data set. We note that our curated gene set database is limited by the available curated knowledge of gene functions. Naturally, our collection of gene sets is incomplete, and includes some noise. As gene set queries become more common, it is important not to lose track of those genes not contained in sets of interest (particularly genes of unknown function). In this context, it is reasonable to highlight those genes which are differentially regulated but *not* contained in any enriched gene set. We carried out this procedure, but the previously highlighted genes were not among the top 500 probe sets (data not shown). These instances show the importance of domain knowledge to focus on results of greatest interest.

Acknowledgments

The authors gratefully acknowledge the assistance and ideas of Liwen Liu and William Reisdorf. This chapter is in preparation for publication as "Generalized gene set queries for microarray analysis", by Stephen Tanner and Pankaj Agarwal.

References

Aebersold, R. and Mann, M., 2003. Mass spectrometry-based proteomics. *Nature*, **422**(6928):198–207.

Aho, A. and Corasick, M., 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, **18**:333–340.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17):3389–3402.

Anderson, D. C., Li,W., Payan, D., and Noble,W., 2003. A New Algorithm for the Evaluation of Shotgun Peptide Sequencing in Proteomics: Support Vector Machine Classification of Peptide MS/MS Spectra and SEQUEST Scores. *Journal of Proteome Research*, **2**(2):137–46.

Anjard C., Loomis W.F., 2005. Peptide signaling during terminal differentiation of Dictyostelium. Proc Natl Acad Sci U S A. **102**(21):7607-11.

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

Bafna, V. and Edwards, N., 2001. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17 Suppl 1**:13–21.

Bandeira, N., Tsur, D., Frank, A., and Pevzner, P., 2007. Protein Identification via Spectral Networks Analysis. *Proc Natl Acad Sci U S A*, **(in press)**.

Barrett, T. and Edgar, R., 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, **411**:352–369.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R., *et al.*, 2007. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, **35**(Database issue):760–765.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**(1):289–300.

Bern, M., Cai, Y., and Goldberg, D., 2007. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal Chem*, **79**(4):1393–1400.

Bern, M., Goldberg, D., 2007. Improved Ranking Functions for Protein and Modification-Site Identifications. Conference on Research in Computational Molecular Biology (RECOMB) 2007.

Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M. et al. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc Natl Acad Sci U S A* 2004 Feb 17;101(7):2173-8

Blanco, E., Parra, G., and Guig´o, R., 2002. Using GeneID to Identify Genes. *Current Protocols in Bioinformatics*, .

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.*, 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**(1):365–370.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M., 1993. dbEST–database for "expressed sequence tags". *Nat Genet*, **4**(4):332–333. Letter.

Cantin, G. and Yates, J., 2004. Strategies for shotgun identification of post-translational modifications by mass spectrometry. *Journal of Chromatography A*, **1053**:7–14.

Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., *et al.*, 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, **419**(6906):512–519.

Chang, C.-C. and Lin, C.-J., 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chisholm, R. L., Gaudet, P., Just, E. M., Pilcher, K. E., Fey, P., Merchant, S. N., and Kibbe, W. A., 2006. dictyBase, the model organism database for Dictyostelium discoideum. *Nucleic Acids Res*, **34**(Database issue):423–427.

Chisholm R. L., Firtel R. A., 2004. Insights into morphogenesis from a simple developmental system. *Nat Rev Mol Cell Biol*. **5**(7):531-41.

Choudhary, J., Blackstock, W., Creasy, D., and Cottrell, J., 2001. Matching peptide mass spectra to EST and genomic DNA databases. *Trends Biotechnol*, **19**(10 Suppl):17–22.

Chung C. Y., Funamoto S., Firtel R. A., 2001. Signaling pathways controlling cell polarity and chemotaxis. *Trends Biochem Sci*. **26**(9):557-66.E

Craig, R. and Beavis, R., 2003. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom.*, **17**(20):2310–6.

Craig, R., Cortens, J., Fenyo, D., and Beavis, R., 2006. Using annotated peptide mass spectrum libraries for protein identification. *J. of Proteome Research*, **5**:1843 –1849.

Creasy, D. and Cottrell, J., 2002. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*, **2**(10):1426–1434.

Creasy, D. M., Cottrell, J.S., 2004. Unimod: Protein modifications for mass spectrometry. Proteomics 4(6):1534-1536.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., and Conklin, B. R., 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, **31**(1):19–20.

Danc´ık, V., Addona, T., Clauser, K., Vath, J., and Pevzner, P., 1999. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*, **6**(3-4):327–342.

Day, R., A.Borziak, and Gorin, A., 2004. Ppm-chain de novo peptide identification program comparable in performance to sequest. In *Proceedings of 2004 IEEE Computational Systems in Bioinformatics (CSB 2004)*, pages 505–508.

Desiere, F., Deutsch, E., Nesvizhskii, A., Mallick, P., King, N., Eng, J., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., *et al.*, 2004. Integration of peptide sequences obtained by high-throughput mass spectrometry with the human genome. *Genome Biology*, **1**(6).

Dunkley, T. P. J., Hester, S., Shadforth, I. P., Runions, J., Weimar, T., Hanton, S. L., Griffin, J. L., Bessant, C., Brandizzi, F., Hawes, C., *et al.*, 2006. Mapping the arabidopsis organelle proteome. *Proc Natl Acad Sci U S A*, **103**(17):6518–6523.

Dunn, O. and Clark, V., 1974. *Applied Statistics: Analysis of Variance and Regression*. Wiley, New York.

Edwards, N. and Lippert, R., 2004. Sequence database compression for peptide identification from tandem mass spectra. In *The 4th Workshop on Algorithms in Bioinformatics (WABI), Bergen, Norway*.

Efron, B. and Tibshirani, R., 2007. On testing the significance of sets of genes. *Annals of Applied Statistics*, **In press**.

Elias, J., Gibbons, F., King, O., Roth, F., and Gygi, S., 2004. Intensity-based protein identification by protein learning from a library of tandem mass spectra. *Nature Biotechnology*, **22**(2):214–219.

Elias, J. E. and Gygi, S. P., 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, **4**(3):207–214.

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**(5696):636–640.

Eng, J., McCormack, A., and Yates, J., 1994. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal Of The American Society For Mass Spectrometry*, **5**(11):976–989.

Farriol-Mathis, N., Garavelli, J.S., Boeckmann, B., Duvaud, S., Gasteiger, E., Gateau, A., Veuthey, A., Bairoch, A., 2004. Annotation of post-translational modifications in the Swiss-Prot knowledge base. Proteomics 4(6): 1537-50.

Fenyo, D., Phinney, B., and Beavis, R., 2007. Determining the Overall Merit of Protein Identification Data Sets: rho-Diagrams and rho-Scores. *J Proteome Res*, e-publication Mar 2007.

Fermin, D., Allen, B., Blackwell, T., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G., and States, D., 2006. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol*, **7**(4).

Florea, L., Francesco, V., Miller, J., Turner, R., Yao, A., Harris, M., Walenz, B., Mobarry, C., Merkulov, G., Charlab, R., *et al.*, 2005. Gene and alternative splicing annotation with air. *Genome Research*, **15**(1):54–66.

Frank, A. and Pevzner, P., 2005. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, **77**:964–973.

Frank, A., Tanner, S., Bafna, V., and Pevzner, P., 2005. Peptide sequence tags for fast database search in mass-spectrometry. *J. of Proteome Research*, **4**(4):1287–1295.

Frank, A. M., Bandiera, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A., 2007. Clustering tandem mass spectra: From spectral libraries to spectral archives. *In preparation*.

Geer, L., Markey, S., Kowalak, J.,Wagner, L., Xu, M.,Maynard, D., Yang, X., Shi,W., and Bryant, S., 2004. Open mass spectrometry search algorithm. *J. Proteome Res*, **3**(5):958–964.

Godovac-Zimmermann, J., Kleiner, O., Brown, L. R., and Drukier, A. K., 2005. Perspectives in spicing up proteomics with splicing. *Proteomics*, **5**(3):699–709.

Grinstead, C. and Snell, J., 1997. *Introduction to Probability*. American Mathmatical Society, OpenURL.

Gupta, S., Zink, D., Korn, B., Vingron, M., and Haas, S., 2004. Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics*, **5**(72).

Han, Y., Ma, B., and Zhang, K., 2004. SPIDER: Software for Protein identification from Sequence Tags with *De Novo* Sequencing Error. In *IEEE Computational Systems Bioinformatics Conference (CSB)*, pages 206–215.

Hansen, B., Davey, S., Ham, A., and D.C., L., 2005. P-mod: an algorithm and software to map modifications to peptide sequences using tandem ms data. *J Proteome Res.*, **4**(2):358–68.

Harsha, H., Suresh, S., Amanchy, R., Deshpande, N., Shanker, K., Yatish, A., Muthusamy, B., Vrushabendra, B., Rashmi, B., Chandrika, K., *et al.*, 2005. A manually curated functional annotation of the human X chromosome. *Nat Genet*, **37**(4):331–2.

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

Havilio, M. and Wool, A., 2007. Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. *Anal Chem*, **79**(4):1362–1368.

Heber, S., Alekseyev, M., Sze, S., Tang, H., and P.A., P., 2002. Splicing graphs and EST assembly problem. *Bioinformatics*, **18**(Suppl 1).

Higdon, R., Hogan, J. M., Van Belle, G., and Kolker, E., 2005. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *OMICS*, **9**(4):364–379.

Higgs, R., Knierman, M., Freeman, A., Gelbert, L., Patil, S., and Hale, J., 2007. Estimating the Statistical Significance of Peptide Identifications from Shotgun Proteomics Experiments. *J Proteome Res*, .

Hogan, J. M., Higdon, R., and Kolker, E., 2006. Experimental standards for high-throughput proteomics. *OMICS*, **10**(2):152–157.

IHGSC Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011):931–945.

Jensen, F. V., 2001. *Bayesian Networks and Decision Graphs*. Springer.

Jensen, O. N., 2006. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, **7**(6):391–403.

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

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

Keller, A., Purvine, S., Nesvizhskii, A., Stolyar, S., Goodlett, D., and Kolker, E., 2002. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, **6**(2):207–212.

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R., 2004. The international protein index: An integrated database for proteomics experiments. *Proteomics*, **4**(7):1985–1988.

Kim, S.B.,Yang, S.,Kim, S.K.,Kim, S.C.,Woo, H.G.,Volsky, D.J.,Kim, S.Y.,Chu, I.S, 2007. GAzer: Gene Set Analyzer. Bionformatics, Apr 27 (Epub)

Kim, S.Y. and Volsky, D. J., 2005. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**:144.

Klammer, A. A., Wu, C. C., Maccoss, M. J., and Noble, W. S., 2005. Peptide charge state determination for low-resolution tandem mass spectra. *Proc IEEE Comput Syst Bioinform Conf*, :175–185.

Korf, I., Flicek, P., Duan, D., and Brent, M., 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**(S1).

Kreppel, L., Fey, P., Gaudet, P., Just, E., Kibbe, W. A., Chisholm, R. L., and Kimmel, A. R., 2004. dictyBase: a new Dictyostelium discoideum genome database. *Nucleic Acids Res*, **32**(Database issue):332–333.

Krueger, I., Meisinger, M., Menarini, M., and Pasco, S., 2006. Rapid systems of systems integration - combining an architecture-centric approach with enterprise service bus infrastructure. In *Proceedings of the 2006 IEEE International Conference on Information Reuse and Integration (IRI'06), Waikoloa, Hawaii, USA.* 51-56.

Kuster, B., Mortensen, P., Andersen, J. S., and Mann, M., 2001. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, **1**(5):641–650.

Lam, H., Deutsch, E., Eddes, J., Eng, J., King, N. Stein, S., and Aebersold, R., 2007. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, **7**:655–667.

Lamb, J., Crawford, E., Peck, D., Modell, J., Blat, I., Wrobel, M., Lerner, J., Brunet, J., Subramanian, A., Ross, K., *et al.*, 2006. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**(5795):1929–1935.

Leipzig, J., Pevzner, P., and Heber, S., 2004. The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Research*, **32**(13).

Liebler, D., Hansen, B., Jones, J., Badghisi, H., and Mason, D., 2003. Mapping protein modifications with liquid chromatography-mass spectrometry and the SALSA algorithm. *Adv Protein Chem*, **65**:195–216.

Lill, J., 2003. Proteomic tools for quantitation by mass spectrometry. *Mass Spectrom Rev*, **22**(3):182–194.

Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W., and Baldi, P., 2001. Improved statistical inference from DNA microarray data using

analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem*, **276**(23):19937–19944.

Lu, B. and Chen, T., 2003. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, **19 Suppl 2**:113–113.

MacCoss, M., McDonald, W., Saraf, A., Sadygov, R., Clark, J., Tasto, J., Gould, K., Wolters, D., Washburn, M., Weiss, A., *et al.*, 2002. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci U S A*, **99**(12):7900–7905.

MacCoss, M., Wu, C., Liu, H., Sadygov, R., and Yates, J., 2003. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem*, **75**(24):6912–6921.

Maeda M., Lu S., Shaulsky G., Miyazaki Y., Kuwayama H., Tanaka Y., Kuspa A., Loomis W. F., 2004. Periodic signaling controlled by an oscillatory circuit that includes protein kinases ERK2 and PKA. *Science*. **304**(5672):875-8.

Mann, M. and Wilm, M., 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**:4390–4399.

Mironov, A., Fickett, J., and Gelfand, M., 1999. Frequent alternative splicing of human genes. *Genome Res.*, **9**(12):1288–1293.

Modrek, B. and Lee, C., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*, **34**(2):177–180.

Nam, D., Kim, S.B., Kim, S.K.,Yang, S.,Kim S.Y.,Chu, I.S., 2006 ADGO: analysis of differentially expressed gene sets using composite GO annotation. Bioinformatics 22(18), 2249-2253.

Nesvizhskii, A., Keller, A., Kolker, E., and Aebersold, R., 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, **75**(17):4646–4658.

Nesvizhskii, A. I. and Aebersold, R., 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, **4**(10):1419–1440.

Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., Baginsky, S., and Aebersold, R., 2006. Dynamic spectrum

quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics*, **5**(4):652–670.

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M., 2006. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**(3):635–648.

Olsen, J. V., Ong, S.-E., and Mann, M., 2004. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics*, **3**(6):608–614.

Omenn, G., States, D., Adamski, M., Blackwell, T., Menon, R., Hermjakob, H., Apweiler, R., Haab, B., Simpson, R., Eddes, J., *et al.*, 2005. Overview of the HUPO plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, **5**(13):3226–3245.

Park, J.J., Berggren, J.R., Hulver, M.W., Houmard, J.A. et al. GRB14, GPD1, and GDF8 as potential network collaborators in weight loss-induced improvements in insulin action in human skeletal muscle. *Physiol Genomics* 2006 Oct 11;27(2):114-21.

Parra, G., Blanco, E., and Guig´o, R., 2000. GeneID in Drosophila. *Genome Research*, **10**(4):511–515.

Patricelli, M.P., Szardenings, A.K., Liyanage, M., Nomanbhoy, T.K., Wu, M., Weissig, H., Aban, A., Chun, D., Tanner, S., Kozarich, J.W., 2007. Functional Interrogation of the Kinome Using Nucleotide Acyl Phosphates. Biochemistry 46: 350-358.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg,M., *et al.*, 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13**(10):2363–2371.

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

Pevzner, P., Danc´ık, V., and Tang, C., 2000. Mutation-tolerant protein identification by mass spectrometry. *J Comput Biol*, **7**(6):777–787.

Pevzner, P., Mulyukov, Z., Dancik, V., and Tang, C., 2001. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res*, **11**(2):290–299.

Pruitt, K. D., Tatusova, T., and Maglott, D. R., 2005. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**(Database issue):501–504.

Purvine, S., Picone, A. F., and Kolker, E., 2004. Standard mixtures for proteome studies. *OMICS*, **8**(1):79–92.

Ratner, D. and Borth, W., 1983. Comparison of differentiating Dictyostelium discoideum cell types separated by an improved method of density gradient centrifugation. *Exp Cell Res*, **143**(1):1–13.

Razumovskaya, J., Olman, V., Xu, D., Uberbacher, E., VerBerkmoes, N., Hettich, R., and Xu, Y., 2004. A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with sequest. *Proteomics*, **4**:961–969.

Resing, K., Meyer-Arendt, K., Mendoza, A., Aveline-Wolf, L., Jonscher, K., Pierce, K., Old,W., Cheung, H., Russell, S., Wattawa, J., *et al.*, 2004. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem*, **76**(13):3556–3568.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D., 2005. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*, **6**(1):R2.

Roth, M. J., Forbes, A. J., Boyne, M. T. n., Kim, Y.-B., Robinson, D. E., and Kelleher, N. L., 2005. Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Mol Cell Proteomics*, **4**(7):1002–1008.

Sadygov, R. and Yates, J., 2003. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem*, **75**(15):3792–3798.

Savitski, M. M., Nielsen, M. L., and Zubarev, R. A., 2006. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics*, **5**(5):935–948.

Searle, B.C., Dasari, S., Turner, M., Reddy, A.P., Choi, D., Wilmarth, P.A., McCormack, A.L., David, L.L., Nagalla, S.R., 2004. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal Chem 76 (8): 2220-2230.

Searle, B., Dasari, S., Turner, M., Reddy, A., Choi, D., Wilmarth, P., McCormack, A., David, L., and Nagalla, S., 2004. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal Chem*, **76**(8):2220–2230.

Searle, B. S., Dasari, S., Wilmarth, P., Turner, M., Reddy, A. P., David, L., and Nagalla, S., 2005. Identification of protein modifications using ms/ms de novo sequencing and the OpenSea alignment algorithm. *Journal of Proteome Research*, **4**:546–554.

Shevchenko, A., Loboda, A., Sunyaev, S., Shevchenko, A., Bork, P., Ens, W., and Standing., K., 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-Quadrupole Time-of Flight Mass Spectrometry and BLAST homology searching. *Analytical Chemistry*, **73**:1917–1926.

Shu, H., Chen, S., Bi, Q., Mumby, M., and Brekken, D., 2004. Identification of phosphoproteins and their phosphorylation sites in the wehi-231 b lymphoma cell line. *Molecular and Cellular Proteomics*, **3**:279–286.

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

Sun, L., Hui, A.M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passanti, A., Menon, J., Walling, J., Bailey, R., Rosenblum, M., Mikkelsen, T., Fine, H.A., 2006. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell, 9(4): 287-300.

Sunyaev, S., Liska, A., Golod, A., Shevchenko, A., and Shevchenko, A., 2003. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem*, **75**(6):1307–1315.

Tabb, D., Saraf, A., and Yates, J., 2003a. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem*, **75**(23):6415–6421.

Tabb, D., Smith, L., Breci, L., Wysocki, V., Lin, D., and Yates, J., 2003b. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem*, **75**(5):1155–1163.

Tadokoro, K., Yamazaki-Inoue, M., Tachibana, M., Fujishiro, M., Nagao, K., Toyoda, M., Ozaki, M., Ono, M., Miki, N., Miyashita, T., *et al.*, 2005. Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of gln in drpla affects subcellular localization of the products. *J Hum Genet*, **50**(8):382–394.

Tang, W., Halpern, B., Shilov, I., Seymour, S., Keating, S., Loboda, A., Patel, A., Schaeffer, D., and Nuwaysir, L., 2005a. Discovering known and unanticipated protein modifications using ms/ms database searching. *Anal Chem*, **77**(13):3931–46.

Tang,W. H., Halpern, B. R., Shilov, I. V., Seymour, S. L., Keating, S. P., Loboda, A., Patel, A. A., Schaeffer, D. A., and Nuwaysir, L. M., 2005b. Discovering known and unanticipated protein modifications using MS/MS database searching. *Anal Chem*, **77**(13):3931–3946.

Tanner, S., Payne, S. H., Dasari, S., Shen, Z., Wilmarth, P., David, L., Loomis, W. F., Briggs, S. P., and Bafna, V., 2007. Accurate annotation of peptide modifications through unrestrictive database search. *In preparation*.

Tanner, S., Pevzner, P., and Bafna, V., 2006. Unrestrictive identification of post-translational modifications through peptide mass spectrometry. *Nat Protocols*, **1**(1):67–72.

Tanner, S., Shu, H., Frank, A., Wang, L., Zandi, E., Mumby, M., Pevzner, P., and Bafna, V., 2005. Inspect: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem.*, **77**(14):4626–4639.

Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P., 2005. Identification of post-translational modifications via blind search of mass-spectra. *Nature Biotechnology*, **23**:1562–1567.

Van Ness, B. G., Howard, J. B., and Bodley, J. W., 1980. ADP-ribosylation of elongation factor 2 by diphtheria toxin. Isolation and properties of the novel

ribosyl-amino acid and its hydrolysis products. *J Biol Chem*, **255**(22):10717–10720.

Van Haastert P. J., Devreotes P. N., 2004. Chemotaxis: signalling the way forward. *Nat Rev Mol Cell Biol.* **5**(8):626-34.Van Ness BG, Howard JB, Bodley JW. ADP-ribosylation of elongation factor 2 by diphtheria toxin. Isolation and properties of the novel ribosyl-amino acid and its hydrolysis products. J Biol Chem. 1980 Nov 25;**255**(22):10717-20.

Vandekerckhove, J. and Weber, K., 1980. Vegetative Dictyostelium cells containing 17 actin genes express a single major actin. *Nature*, **284**(5755):475–477.

Vardhanabhuti, S., Blakemore, S. J., Clark, S. M., Ghosh, S., Stephens, R. J., and Rajagopalan, D., 2006. A comparison of statistical tests for detecting differential expression using affymetrix oligonucleotide microarrays. *OMICS*, **10**(4):555–566.

Venter, J. et al., 2001. The sequence of the human genome. *Science*, **291**(5507):1304–1351.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.*, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915):520–562.

Welle, S., Brooks, A. I., Delehanty, J. M., Needler, N., and Thornton, C. A., 2003. Gene expression profile of aging in human muscle. *Physiol Genomics*, **14**(2):149–159.

Welle, S., Brooks, A. I., and Thornton, C. A., 2002. Computational method for reducing variance with Affymetrix microarrays. *BMC Bioinformatics*, **3**:23.

Wilmarth, P. A. amd Tanner, S., Dasari, S., Nagalla, S. R., Riviere, M. A., Bafna, V., Pevzner, P. A., and David, L. L., 2006. Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: Does deamidation contribute to crystallin insolubility? *Journal of Proteome Research*, 2006.

Yates, J., Eng, J., and McCormack, A., 1995a. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, **67**(18):3202–3210.

Yates, J., Eng, J., McCormack, A., and Schieltz, D., 1995b. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, **67**(8):1426–1436.