Thesis proposal

Aaron A. Klammer

Department of Genome Sciences

University of Washington

June 2, 2005

1 Research Plan

Tandem mass spectrometry (MS/MS) is a particularly useful technology for the identification of proteins in complex mixtures (McCormack et al., 1997; Yates, III, 1998b,a). An essential step in this technology is the fragmentation of a protonated peptide and detection of the resulting fragment ions in the form of a mass spectrum (Arnott et al., 1993). Due to the complex chemistry of peptide fragmentation, the pattern of peaks in such a spectrum can be predicted only qualitatively: an exact prediction of spectrum peak heights, or even which peaks will be present or absent, has proven elusive (Brancia et al., 2001; Dancik et al., 1999; Elias et al., 2004; Wan and Chen, 2005; Zhang, 2004; Schutz et al., 2003).

I plan to test two closely related hypotheses in an attempt to address this problem. The first is that an improved predictive model of mass spectrum peak intensity, trained on actual mass spectrometry data, will provide insight into the complex chemistry of protonated peptide fragmentation. In addition, it is likely such a probabilistic intensity model will provide information about which peptide has generated a particular spectrum. Thus, my second hypothesis is that a realistic probabilistic intensity model will be useful for improving identification of unknown peptide fragmentation spectra, especially in conjunction with a sequence database search step. To test these hypotheses, I propose the following three specific aims:

Aim 1: Implement and train a probabilistic graphical model of peptide fragmentation. Several methods for predicting the mass spectrum for a given peptide have been developed, but the most commonly used have relied heavily on expert knowledge to design simple heuristics (Eng et al., 1994; Field et al., 2002), or to set probabilities within a larger model (Bafna and Edwards, 2001). Only recently have models trained on actual mass spectrometry data been pursued (Dancik et al., 1999; Elias et al., 2004; Wan and Chen, 2005; Zhang, 2004; Havilio et al., 2003). Most have incorporated some knowledge of fragmentation pathways; none to my knowledge have integrated specific fragmentation probabilities into a comprehensive model of fragmentation.

The peptide fragmentation process can be represented as a Bayesian network—that is, a directed, acyclic graph, in which the nodes represent events (in this case, the formation of fragment or ion species) and the edges represent conditional probability dependencies. By training such a

model on mass spectrometry data, I will gain insight into the fragmentation pathways that are more or less likely to occur, and thus into overall peptide fragmentation chemistry. Different parameterizations of the model will provide insight into which factors are important for predicting peptide fragmentation.

Aim 2: Generate a set of high-confidence spectrum-peptide associations from peptides with controlled composition under controlled conditions. Any computational model benefits from experimental validation. While I will use data from outside sources, it is likely that the model will make predictions for which there is no available data, especially for certain classes of experimental conditions, or for certain peptide compositions. Hence, to validate my intensity model, I will acquire spectra from a variety of peptides. Some spectra will be from artificially synthesized peptides of defined composition. Other spectra will be obtained from chemically modified digests of whole proteomes. These various methods will obtain a broader sampling of the space of all valid peptide-spectrum matches than is available in existing databases.

Aim 3: Incorporate the intensity model into a sequence database search algorithm. Peptide fragmentation chemistry is interesting in itself, but it also finds a direct application in the identification of unknown peptides in complex mixtures. A probabilistic model of peptide fragmentation will be particularly useful in this context. While the model described in Aim 1 gives the conditional probability of a spectrum given a peptide, it can easily be inverted to yield the probability of a peptide given a spectrum. Peptide identification using this probability is then a matter of finding the peptide that has maximal probability given the spectrum, or the peptide that has maximal probability given several spectra.

2 Background

A major goal in modern biology is the identification and characterization of the cell's entire protein complement, or proteome. Towards this end, mass spectrometry (MS)-based technologies offer the ability to rapidly identify proteins in complex mixtures (McCormack et al., 1997; Yates, III, 1998b,a; Pandey and Mann, 2000; Mann et al., 2001; Aebersold and Goodlett, 2001; Aebersold and Mann, 2003; Tyers and Mann, 2003). An essential step in this technology is the fragmentation

of a protonated peptide and detection of the resulting fragment ions in the form of a mass spectrum (Arnott et al., 1993). Here, I first review pertinent papers from the large body of peptide fragmentation literature. Then, I discuss some common algorithms that identify peptides from mass spectrometry data and their limitations, with special emphasis on sequence database search algorithms.

2.1 Peptide fragmentation

The study of fragmentation patterns of peptides under low-energy conditions has long been the focus of intense research. A general consensus has emerged surrounding the mechanism for fragmentation known as the "mobile-proton" model (Wysocki et al., 2000; Dongre et al., 1996). In the model, peptide fragmentation is caused by migration of a proton to a location on the peptide backbone.[1] Fragmentation occurs by a nucleophilic attack of an adjacent amide carbonyl group to form the familiar N-terminal b-ion and C-terminal y-ion (Roepstorff and Fohlman, 1984; Polce et al., 2000). Under low-energy conditions, the b-ion and y-ion fragments co-exist in a loose complex; the two members of this dimer compete for the proton, with assignment of charge being determined by the proton affinities of the two ions (Paizs and Suhai, 2004).

The fragmentation event can be influenced by numerous factors, including but not limited to charge state (Dongre et al., 1996), the residues present (Qin and Chait, 1999; Summerfield et al., 1997), the size of a peptide, collision energy (van Dongen et al., 1999), the peptide's fugacity[2] (Downard and Biemann, 1995), degrees of freedom (Dongre et al., 1996), and even the peptide conformation (Wu et al., 1995). The influence on cleavage that has been most closely studied is the effect of amino acid residues on the probability of cleavage occurring at a particular backbone amide bond. These trends are often divided into two general categories: those present when the peptide has a proton that is "mobile", meaning that there is an excess of protons relative to basic residues; and those trends that occur when all protons are "non-mobile", that is, when all protons are sequestered at the amino terminus or at a basic residue (especially arginine).

In general, when a proton is non-mobile, amide bonds are cleaved more selectively than when

[1] There appears to be some debate over whether protonation of the amide oxygen or amide nitrogen precedes amide bond cleavage. Paizs and Suhai (2004) holds that the amide oxygen does not play a substantial role. Regardless, fragmentation occurs by a nucleophilic attack of an adjacent amide carbonyl group to form an N-terminal b-ion and C-terminal y-ion

[2] Fugacity is the ability of a molecule to leave an electrospray drop and escape to the gas phase

the proton is mobile. For example, when peptides with non-mobile protons are acetylated at arginine (abolishing this residue's proton-sequestering ability, and making the peptide's protons more mobile), then non-selective cleavage is increased (Dongre et al., 1996). Selective cleavage can occur C-terminal to aspartate in non-mobile proton peptides; if no aspartate is present, non-mobile proton peptides generally cleave non-selectively (Gu et al., 2000; Qin and Chait, 1999). Other trends for non-mobile proton peptides include selective cleavage and a novel b-ion formation C-terminal to histidine (Wysocki et al., 2000).

Recently, efforts have been made to systematically sample peptides using small databases of tryptic peptides in order to detect broad trends affecting peptide cleavage. These studies have confirmed many trends found with the previous directed mechanistic studies. For example, histidine presence increases cleavage C-terminal to aspartate (Huang et al., 2002). Statistical comparison of tryptic spectra confirms trends showing that cleavage is more likely N-terminal to proline and C-terminal to histidine, as well as showing enhanced cleavage C-terminal to glycine and serine (Tabb et al., 2003; Breci et al., 2003).

Recent enhancements to the mobile proton model include the introduction of a third category, that of "partially-mobile" protons, where the number of protons is more than the number of arginine present but less than the total number of arginine, histidine and lysine. Such partially-mobile peptides show increased cleavage N-terminal to proline and C-terminal to aspartate (Kapp et al., 2003).

Many of these trends have been incorporated into theoretical spectrum generation. Probably the most extensive computational implementation of peptide fragmentation is Zhang's implementation of the mobile proton model (Zhang, 2004), which provides inspiration for the graphical model presented in the Methods section. My first goal will be to implement Zhang's model in a probabilistic framework.

An issue that is related to but independent of modelling peptide fragmentation probabilities is the attempt to characterize the noise associated with fragmentation, and the variability of individual peptide spectra. Efforts to improve peptide identification with SEQUEST have shown that averaging spectra is inferior to searching each spectrum individually (Venable and Yates, III, 2004), implying that signal in each individual spectrum is washed out upon taking the mean intensity of peaks. Analysis of noise in spectral peaks has shown that most of the variability can be accounted

for by shot-noise in a Poisson distribution (MacCoss et al., 2001). Shot noise is noise due to random fluctuations in the number of ions detected by the mass spectrometer, proportional to the square root of the peak intensity.

2.2 Peptide identification algorithms

Algorithms for protein identification using mass spectrometry can be broken down into several sub-categories. In the first category belong mass fingerprinting algorithms, which examine MS scans and identify peptides solely from their m/z values without fragmentation (Zhang and Chait, 2000; Papping et al., 1993; Clauser et al., 1999). In the second category are algorithms that take advantage of peptide fragmentation for identification, in the form of MS-MS scans. These MS-MS based peptide identification algorithms can be further divided into two sub-categories: *de novo* sequencing algorithms that determine the sequence of a peptide directly from a fragmentation spectrum, without use of a database (Dancik et al., 1999; Pevzner et al., 2000; Taylor and Johnson, 1997; Johnson and Biemann, 1984; Bartels, 1990; Fernandez de Cossio et al., 1995); and database search algorithms, which use the information in a spectrum, combined with sequence information from a protein database, to determine a peptide's identity. Of course these divisions are not absolute: some algorithms combine information from multiple peptides found in MS scans with MS-MS fragmentation information (Tabb et al., 2002; Edwards and Lippert, 2002); and the distinction between *de novo* and database search algorithms can be blurry (Mann and Wilm, 1994). Regardless, I will focus on the sequence database search algorithms, which are the most pertinent to my research proposal.

Sequence database search algorithms use a protein sequence database to constrain the peptides that might possibly have generated a spectrum. One of the most commonly used of these algorithms is SEQUEST (Eng et al., 1994; Yates, III et al., 1995), which creates a theoretical spectrum for each peptide. To increase speed, SEQUEST creates a spectrum only for peptides within a certain mass tolerance of the measured precursor peptide mass; in addition, it compares spectra using a preliminary score before using a more computationally expensive (and more accurate) normalized dot product score. SEQUEST uses a uniform intensity model for b-ions and y-ions, even though it is known that spectra differ greatly from uniform intensity (Brancia et al., 2001). Furthermore, creating a theoretical spectrum limits SEQUEST to only one spectrum per peptide. In reality, each

peptide is capable of producing a distribution of spectra. The advantage of using a distribution as opposed to a single spectrum is analogous to the advantages of estimating an unknown using a posterior distribution from a Bayesian approach rather than using a maximum likelihood estimate. Bayesian approaches allow incorporation of prior knowledge into the posterior probability estimates.

Other algorithms such as ProteinProspector and MS-Tag (Clauser et al., 1999) exploit high-precision peptide mass measurements from the MS scan combined with immonium ion information about amino acid composition from the MS-MS scan to determine peptide identity. For example, Craig and Beavis (2003) present a multi-step process to search a database looking for incomplete enzymatic hydrolysis, non-specific hydrolysis and chemical modifications of residues. The algorithm further constrains candidate peptides by assuming that at least one tryptic peptide is generated from each protein. Some algorithms combine a partial sequencing peptide sequence through *de novo* sequencing and then use this to search database. Along these lines, PeptideSearch (Mann and Wilm, 1994) determines short peptide subsequences from the MS scan. Lutefisk performs partial *de novo* sequencing followed by a homology based sequence database search with CIDentify (Taylor and Johnson, 1997).

Several algorithms generate a probability for the peptide-spectrum match. Mascot (Perkins et al., 1999) calculates the probability that the observed fragment ions for a given peptide mass are chance events, yielding a p-value; unfortunately, the details of this algorithm are not made clear in the paper. The search algorithm that most closely approximate the probabilistic approach outline in this paper is SCOPE (Bafna and Edwards, 2001). Bafna and Edwards (2001) present a probability model that splits spectrum generation into two steps, a fragmentation step and a measurement step. They formulate the calculation of a probability density function Ψ for $Pr(S|p)$ (the probability of a spectrum given a peptide) as

$$\Psi(S|p) = \sum_{F \subseteq F(p)} \Psi(S|F,p) Pr(F|p) \tag{1}$$

for a spectrum S, peptide p, and fragmentation event F from fragmentation space $F(p)$ (all possible fragment ions from a particular peptide p). Bafna *et al.* present an algorithm for calculating the summation for $\Psi(S|p)$ above in $O(|F(p)|k + k^2)$ time. For the measurement error of the fragment ions, $\Psi(S|F,p)$, they use a normal distribution. Bafna and Edwards (2001) do not train on actual

data, because of its scarcity; instead they set the parameters for $Pr(F|p)$ using expert knowledge. This is a major weakness of their approach. In addition, the model does not take into account fragment ion intensities.

To my knowledge, the only algorithms that use an intensity model for probabalistic assessment of peptide-spectrum matches are those by Havilio et al. (2003) and Elias et al. (2004). The Havilio et al. (2003) paper is plagued by a questionable methodology: the authors mix testing and training, and seem to use a rather artificial metric for measuring their algorithm performance. The decision tree approach of Elias et al. (2004) met with some success in improving over other algorithms such as SEQUEST; however, they do not harness knowledge of peptide fragmentation pathways explicitly, and they only model b-ion and y-ions.

More recently, Wan and Chen (2005) develop a hidden Markov model (HMM) modelling spectrum generation using only b-ions and y-ions, where the (somewhat artificial) time dimension is sequential steps along the peptide backbone. They modeled peak m/z measurement noise using a normal distribution; however, they did not take into account different peak intensities.

In addition to the battery of sequence database search algorithms described above, there are a number of post-processing algorithms that seek to improve the predictions of the initial peptide-spectrum matches generated by these algorithms (in most cases SEQUEST). Anderson et al. (2003) use a support vector machine (SVM) to classify peptide-spectrum assignments as either true or false using SEQUEST metrics and other statistics as discriminatory features. DTASelect (Tabb et al., 2002) processes SEQUEST results to find a parsimonious set of proteins that can account for the identified peptides. PeptideProphet (Keller et al., 2002) uses Fisher's linear discriminant and the expectation-maximization algorithm (EM) to process SEQUEST search results to compute the probability that a peptide-spectrum assignment is correct. A related software package, ProteinProphet (Nesvizhskii et al., 2003) uses EM to find minimal list of proteins that can account for identified peptides.

A major problem with sequence database search algorithms is their inability to handle post-translational modifications in time that is not exponential in the number of modifications. This factor aggravates another problem with database search algorithms: a large fraction of spectra in proteomics experiments are not identified (Simpson et al., 2000). It is my hope that a more sophisticated and subtle approach to intensity modelling of spectra will decrease this fraction of

unidentified spectra. I am encouraged in thinking this is possible by the successes of Elias et al. (2004) and Zhang (2004).

3 Preliminary Findings

I present some preliminary data which indicates that the prediction of mass spectrometry peak intensities from peptide fragmentation models is far from a solved problem. Other than this preliminary data, the bulk of my research to date has not focused directly on modelling peptide fragmentation, but has been concerned with two closely related issues regarding improvements to peptide identification using mass spectrometry. The first is computational and focuses on improving mass spectrum charge-state assignment algorithms; this research has been submitted for publication. The second is experimental and is concerned with improving protein digestion conditions to maximize protein identification.

3.1 Peptide fragmentation is an unsolved problem

The state of the art in modelling peptide fragmentation for mass spectrometry is the mobile-proton model (Wysocki et al., 2000; Dongre et al., 1996) implemented by Zhongqi Zhang (Zhang, 2004). In Fig. 1 I show a pair of spectra generated from this model, along with observed spectra from an actual mass spectrometer. The observed spectra were assigned to the peptides from an *E. coli* soluble protein fraction digest using the algorithms SEQUEST and DTASelect with parameters set for a low (less than 0.1%) false-discovery rate. Together, these spectra demonstrate both the power and limitations of Zhang's implementation.

In Table 1, I compare the spectra to one another using the following similarity metric (Zhang, 2004):

$$s(S,\hat{S}) \quad = \quad \frac{\sum_{i=1}^{k}\sqrt{S_i\hat{S}_i}}{\sqrt{\sum_{i=1}^{k}S_i\sum_{i=1}^{k}\hat{S}_i}} \tag{2}$$

where s is the similarity between an observed spectrum $S = (S_1 \cdots S_k)$ (binned into k one m/z bins) and an analogous predicted spectrum $\hat{S}$. A score of 1 indicates a perfect match; a score of 0 indicates no similarity. Two unrelated spectra (E2 and V2) have a similarity of 0.392, for example.

Table 1: Comparison of two groups of four spectra using Zhang's similarity metric (Eqn. 2). Two spectra (E1 and V1) are predicted by Zhang's MassAnalyzer v.1.02 Zhang (2004), the other six (E2-4 and V2-4) are observed spectra collected on a linear ion trap ThermoFinnigan LTQ mass spectrometer (see Fig. 1 for spectra).

	E1	E2	E3		V1	V2	V3
E2	0.550	—	—	V2	0.775	—	—
E3	0.703	0.746	—	V3	0.980	0.799	—
E4	0.674	0.763	0.877	V4	0.907	0.775	0.914

While Zhang's metric is not the only possible similarity metric, in order to give his method the greatest chance of success, I will use it here.

The model does well for predicting the spectra of certain peptides, such as VVVTGLGML-SPVGNTVESTWK. Notice the striking similarity between the three experimentally generated spectra and Zhang's spectrum. For peptide VVVTGLGMLSPVGNTVESTWK, the similarity between the predicted spectrum and the observed spectra is remarkably high, and in one case close to a perfect match (0.980). For other peptides, however, the model's predicted spectrum is unsatisfactory, as for peptide EIELEDKFENMGAQMVK. The predicted spectrum for this peptide does not closely match the observed spectra, yielding similarity scores closer to random matches. A readily noticed feature of the observed spectra is the familiar decrease in intensity at high and low m/z values relative to the precursor ion m/z in the observed. This feature appears to be wholely lacking from the predicted spectrum; thus, an improved model might incorporate m/z relative to the precursor ion into the probability of detecting a fragment ion. It is because of Zhang's successes and failures like these that I see modelling mass spectra peak intensities as a tractable—yet not fully solved—problem.

3.2 Peptide charge state determination

Most recently, the main focus of my research has been refining methods for determining charge state for tandem mass spectra. A typical 24-hour microcapillary shotgun proteomics experiment on a ThermoFinnigan LTQ mass spectrometer can produce more than 250,000 spectra. Each of these spectra must be searched against a sequence database—a step which is often rate-limiting. Low-resolution spectra from multiply charged peptides further aggravate this problem, because

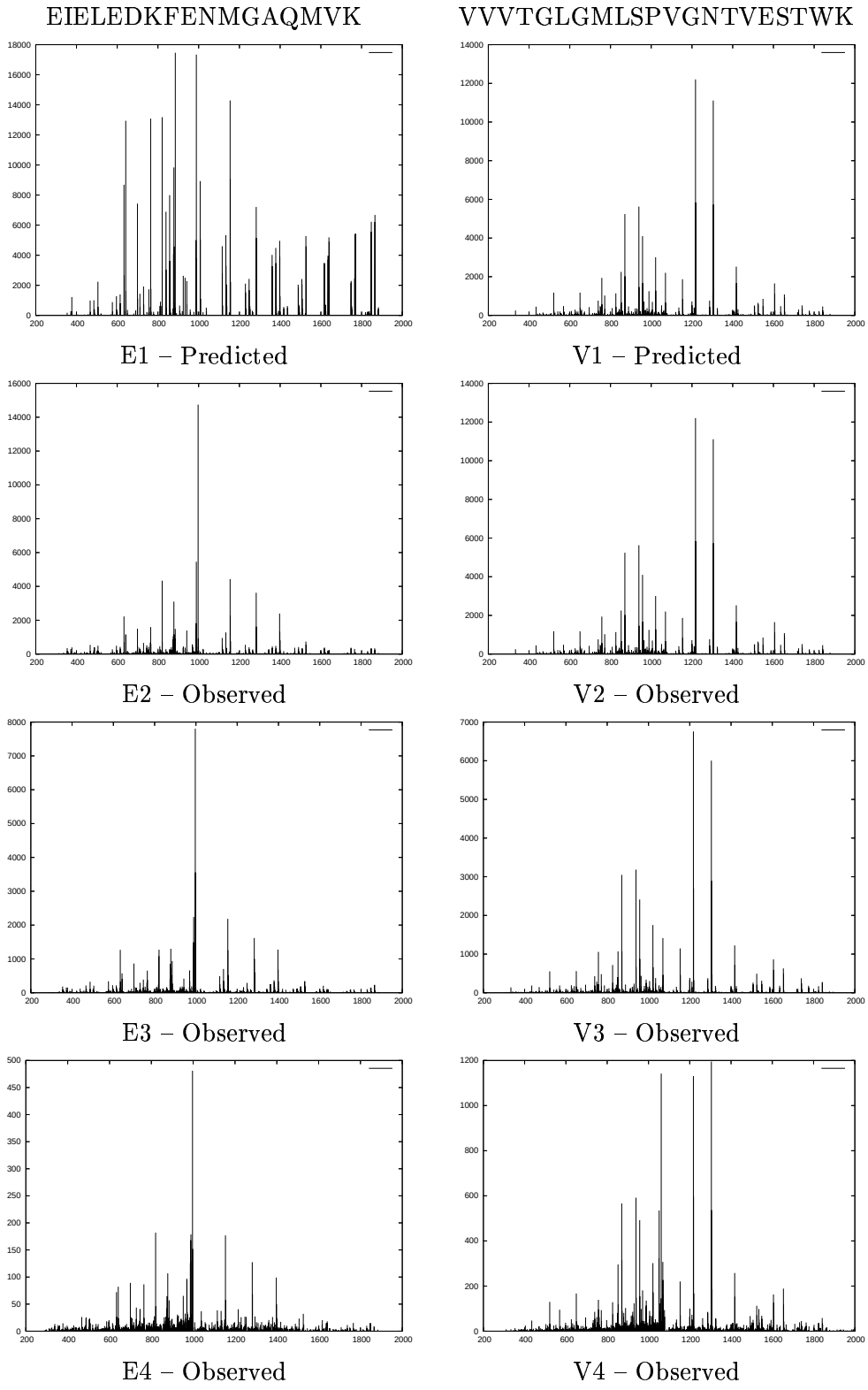

Figure 1: Two predicted and six observed fragmentation spectra for two peptides. Two groups of four spectra are shown, four from peptide EIELEDKFENMGAQMVK (E1-4) and four from peptide VVVTGLGMLSPVGNTVESTWK (V1-4). The first two spectra (E1 and V1) are predicted spectra from MassAnalyzer v.1.02, an algorithm developed by Zhang (2004). The other six are spectra collected on a linear ion trap ThermoFinnigan LTQ mass spectrometer from the same two peptides.

these spectra must be searched multiple times, once for each possible peptide charge state.

To eliminate this inefficiency, I used a support vector machine (SVM) classifier to quickly and reliably classify multiply-charged MS/MS spectra as resulting from either +2 or +3 peptides. By classifying 80% of multiply charged spectra, the SVM obtains a 40% reduction in processing time while maintaining an average of 99% of both peptide and protein identifications across two platforms. These results have been submitted for publication (see attached Klammer et al. (2005)).

This project has influenced my thinking regarding the intensity modelling project in several ways. It has given me an understanding of common fragmentation pathways, even though I did not model them probabilistically. It has also helped me to appreciate the difficulty in generalizing conclusions across different mass spectrometry platforms. An accurate model of fragmentation must be flexible enough to take this cross-platform variability into account.

3.3 Improved digestion for protein identification

My second area of research has focused on improving protein identification by modifying a common experimental protein digestion protocol. This research was motivated by the observation that most μLC-MS-MS experiments detect only a fraction of a sample's proteins. One cause of this poor detection rate is incomplete protein digestion, which in turn can be affected by several factors, including poor protein solvation or denaturation, inadequate enzyme concentration or insufficient reaction time. To test how these factors might interfere with digestion, and hence protein identification, I systematically modified and tested a standard digestion protocol on the *E. coli* proteome. A key modification of the protocol involved the replacement of solution-phase trypsin with a column of immobilized trypsin beads. These beads were intended to increase digestion efficiency by increasing local enzyme-substrate ratio.

Improved digestion was indicated in the trypsin column digestion by a three-fold increase in protein identification, increased peptide sequence coverage of proteins and an almost five-fold increase in identification of low-level proteins. This work demonstrated a substantial improvement over existing methods and suggests alteration of current protein digestion strategies. These results will soon be submitted for publication (see attached Klammer and MacCoss (2005)).

This project did not have direct relevance to intensity modelling. However, pursuing it has given me useful experimental background that will undoubtedly help me to execute Aim 2, the

generation of high-confidence peptide-spectrum matches from controlled peptides.

4 Methods of Procedure

In this section, I first present the probabilistic graphical model of peptide fragmentation and spectrum peak intensity. Next, I outline some specific experimental procedures I will perform to validate the model and test its predictions. Finally, I show how such a model will be efficiently incorporated into a sequence database search algorithm.

4.1 Aim 1: Implement and train a probabilistic graphical model of peptide fragmentation.

Here I describe a Bayesian network to probabilistically model fragmentation of a peptide within the mass spectrometer. I begin with a brief description of Bayesian networks. I then describe the specific network proposed for the fragmentation of a single peptide. Finally, I describe how these probabilities can be used to calculate the probability of a spectrum given a peptide, $Pr(S|p)$.

Bayesian networks are a form of graphical model that encode the conditional independencies of a system of random variables. A Bayesian network is a directed, acyclic graph, in which the nodes are random variables representing events and the edges represent conditional dependencies between these variables. A conditional probability table (CPT) is stored at each node containing the probability of the node taking on each of its possible values given values in its parent nodes (i.e. nodes with arrows pointing to it). If designed properly, such networks offer an efficient means of calculating the joint probability distribution function over the random variables. Moreover, tools exist that allow these models to be trained from data, enabling them to learn previously unknown trends.

Peptide fragmentation can be modeled relatively easily with a Bayesian network because, unlike many networks of chemical reactions, there are no back-reactions: ions only fragment, they do not re-form. By training such a model on mass spectrometry data, I will gain insight into the fragmentation pathways that are more or less likely to occur, and thus into overall peptide fragmentation chemistry. Different parameterizations of the model will provide insight into which factors are important for predicting peptide fragmentation.

4.1.1 Detailed model description

Conceptually, the model will divide peptide fragmentation into multiple steps, each conditionally dependent on the previous (See Fig. 2). There are four main parts to the fragmentation model. The first part models the distribution of peptides that could have produced the spectrum. In the initial implementation, the peptide is fixed, so this part of the model is not used, but it is included here for completeness. The second and main part of the model involves peptide fragmentation into b-ions and y-ions, followed by secondary fragmentation of these ions to form internal ions. The third part models neutral losses and isotopic distribution. Finally, the fourth part models measurement error and noise. At the leaf nodes, the model specifies the expected distribution of observed peaks generated from a single peptide.

The first phase models the distribution of peptides that could give rise to a spectrum. While the initial implementation of the model will fix the identity of the original peptide, future applications will involve a distribution over many peptides. The identity of the peptide depends on an observed m/z value and will be a string generated from a particular alphabet having a certain number of residues and charge.

In the second phase I model fragmentation. Each peptide has a certain probability of fragmenting into a particular b-ion/y-ion pair depending on the identity of that ion and a particular fragmentation parameterization (discussed below). The charge in the original peptide is allocated between the b-ion and y-ion depending on these ion's chemical composition and a charge-allocation parameterization. For example, a +2 ion can produce a +1/+1 pair, a +0/+2 pair or a +2/+0 pair. After fragmentation, there is a node that takes into account the probability of charge loss,i.e. loss of $H+$. This fragmentation process is then repeated. Each of the b-ions and y-ions undergo an additional round of fragmentation, charge-allocation and charge-loss. The products of the fragmentation include b-ions and y-ions as well as internal fragment ions (ions that do not contain the peptide's original N- or C-terminal ends).

The third phase models losses of small neutral molecules and isotopic distribution. Three different forms of neutral losses are modeled: CO, H_2O and NH_3, up to the maximum consistent with the composition of the fragment ion. The losses are treated independently, and only depend on the chemical group participating in the loss. Multiple fragmentations will be the joint probability of

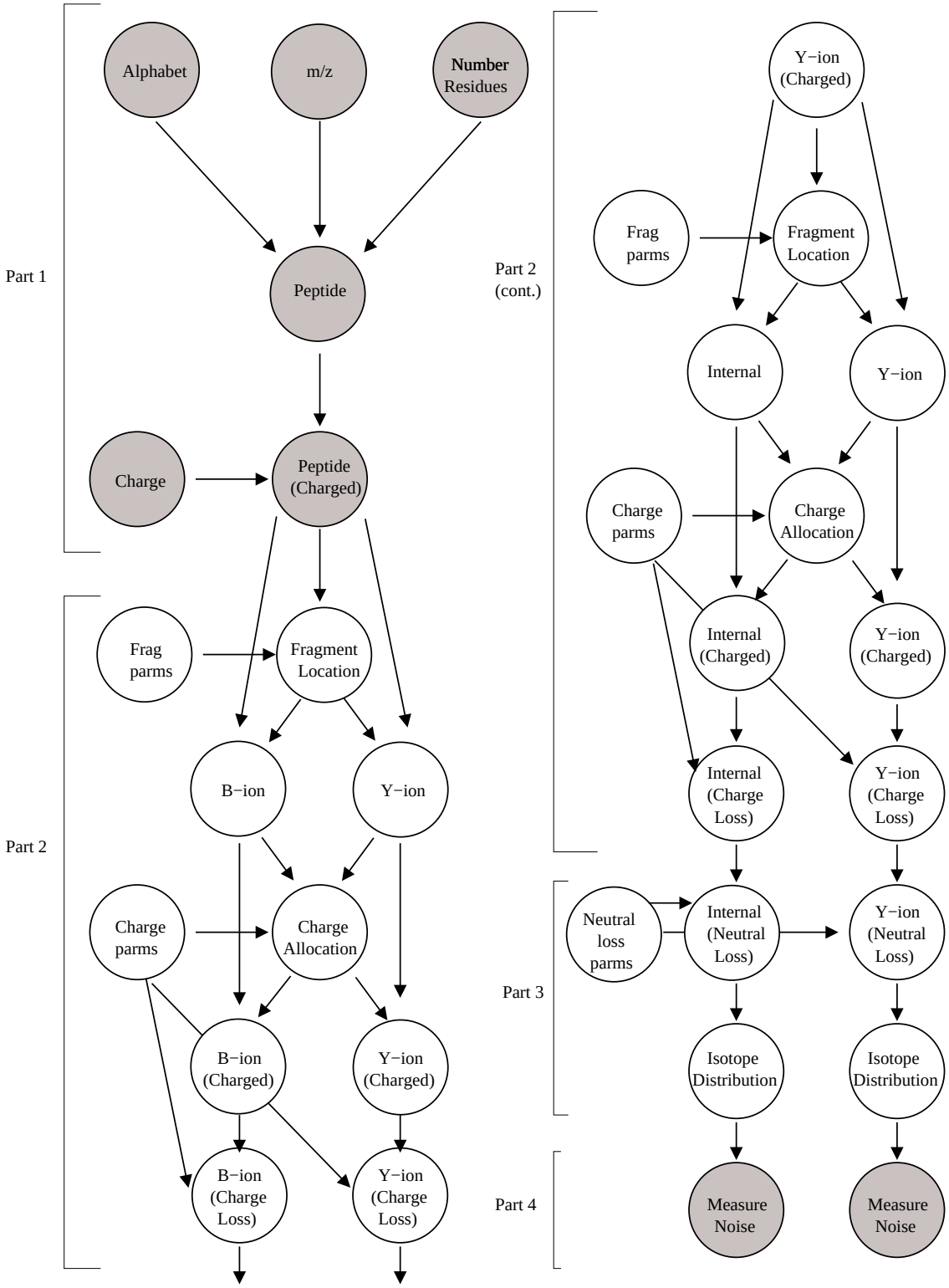

Figure 2: A graphical model for fragmentation of a single peptide. On the right is shown a continuation of the pathway for a y-ion only, for simplicity; the b-ion fragmentation pathway is analogous. Shaded nodes indicate observed variables: for calculating $Pr(F|p)$, the identity of the charged peptide and measured m/z values are known.

the occurrence of the multiple neutral losses. The following set of nodes model isotopic distributions. Each peptide has a probability of containing an element of a non-standard mass; this will be modeled using a product of multinomial distributions, one for each of the elements in the peptide (S, O, N, C or H), where the possible values in each distribution are the possible isotopes of each of element (Yergey, 1983).

The final nodes account for instrument noise. The resolution of current ion-trap mass spectrometers such as the ThermoFinnigan LTQ is 0.3 m/z units. Each peak will have some probability of being shifted by between -0.3 and 0.3 m/z units following a Gaussian distribution, which is reasonable approximation to the true distribution (Wan and Chen, 2005). In addition, the height of the peaks will be discretized to some number of partitions, and will be modeled with a Poisson distribution that should account for majority of noise (MacCoss et al., 2001). For other mass spectrometers, noise will be modelled taking into account their respective m/z measurement precisions.

4.1.2 Model parameterization

A key element of the model is the way in which it parameterizes both fragmentation and charge allocation. For fragmentation, the initial parameterization will assign equal probability of fragmenting at each location. A subsequent parameterization will have a parameter for the contribution to activation energy E_a of each residue adjacent to a cleavage site, for example, E_a^{Gly-C} for the contribution of glycine C-term to the site, for a total of 40 parameters. The probability for fragmentation $P(f_i)$ at site C-term to residue i will then be proportional to the rate of cleavage k at the site where

$$k = Ae^{-E_a/RT} \tag{3}$$

for ion concentration A, gas constant R, temperature T and activation energy $E_a = E_a^{Xaa-N} + E_a^{Xaa-C}$, where the E_a terms are for the N-term and C-term amino acids, respectively. Extending this idea, the next set of fragmentation parameters will be based on each residue *pair* adjacent to a cleavage site, for 400 parameters in the form $E_a^{Xaa-N,Xaa-C}$, but otherwise with probabilities calculated as above. Further extensions involve including residues two residues away from the cleavage site in the activation energy calculations, for a total of 440 parameters, 400 of the form

$E_a^{Xaa-N,Xaa-C}$ and 40 for $E_a^{Xaa-N-1}$ or $E_a^{C+1=Xaa}$. Finally, I will incorporate the parameters of Zhang's mobile proton model implementation, including, at first, only charge directed fragmentation and then both charge directed and charge remote fragmentation. All cleavage probabilities will be normalized over the entire fragment ion. The parameterizations described here are derived from ideas from Tabb et al. (2003); Kapp et al. (2003) and Zhang (2004). Solving for the E_a parameters in the fragmentation step will give each probability a direct scientific interpretation.

Charge allocation will have a similar staged parameterization, based on the identity of residues near the fragmentation at first, and then all residues in either fragment. These values will involve a parameter for gas-phase basicity of each residue G_B^{Gly}, for that of glycine, for example. The probability of a specific fragment ion of length n retaining a charge will then be proportional to $\sum_i^n e^{G_B^{r_i}/RT}$, where r_i is the ith residue of the fragment, normalized for the two fragment ions. Solving for the G_B parameters in the charge allocation steps will give each probability a direct scientific interpretation.

4.1.3 Model training

The model outlined here will be trained on mass spectrometry peak intensity data, using a standard graphical-model training algorithm such as the Junction-Tree Algorithm (Jordan, 2005). This peak intensity data will come from data sets of high-confidence peptide-spectrum matches, generated locally and externally. Locally-generated peptide-spectrum matches will rely on algorithmic assignments, applying some stringent threshold to SEQUEST-DTASelect assignments, and validating these assignments with a secondary algorithm, such as Mascot. Externally-generated manually-curated data sets exist (e.g. that of Keller et al. (2002)) and in previous studies have produced somewhat better results than algorithmically generated associations (Klammer et al., 2005), so these will be used where available. Zhang (2004) uses more than 5,000 manually curated ions (4000 peptides) to train his model. Thus, training the model described here, which includes noise for each peak, and would rely at least partially on less-reliable algorithmically generated associations, will require an order of magnitude more data. Locally available high-confidence redundant SEQUEST peptide-spectrum associations currently exceed (FIXME) [3], which should be sufficient. Completion of Aim 2 will generate additional data on the order of 5,000 partially redundant peptide-spectrum

[3] Personal communication from Gennifer Merrihew, MacCoss lab.

assignments per day of mass spectrometer time.

The model allows us to calculate $Pr(f|p)$, the probability of a single fragmentation pattern from a peptide. We wish to calculate $Pr(s|p)$, which is really a sum over all possible fragmentation events within the spectrum. This sum could be calculated using an algorithm analagous to the one proposed by Bafna and Edwards (2001), except adapted for graphical models (see Eqn. 1).

4.1.4 Model testing

The model will be tested in at least two ways. The first will use Zhang's similarity metric to ask the question: Are the spectra predicted by the model with high probability for a given peptide highly similar to that peptide's observed spectrum? Are these high-probability spectra more similar to the observed spectrum than spectra predicted using other methods? Examples of other predicted spectra will include a uniform intensity model for b- and y-ions, a SEQUEST-like predicted spectrum, and Zhang's predicted spectrum. The second means of testing the model will involve training on a given peptide's high-confidence peptide-spectrum matches and determining if other spectra from the same peptide have high probabilities, and if spectra from different peptides have low probabilities. A particularly satisfying result would be if the test peptide spectrum differs from the model's maximum likelihood spectrum for that peptide, but nonetheless has high probability assigned to it by the model, thus validating the distribution as opposed to maximum likelihood approach.

4.1.5 Assumptions of the model

In designing the model, I have made several simplifying assumptions:

1. A neutral loss from a given residue depends only on that residue.
2. Each peptide fragmentation can be modeled independently.
3. There are no reverse fragmentation reactions (i.e. no peptide synthesis).
4. There are no simultaneous fragmentation events.
5. The effective temperature of ions can be ignored.
6. Neutral losses can be modeled separately from (and after) peptide fragmentation.

7. There is a maximum of two fragmentation events per peptide.

Assumptions 1-4 are made by other spectrum modelling algorithms, such as Zhang's mobile proton implementation Zhang (2004). Assumptions 5-7 are not absolutely necessary, but simplify the graphical model considerably.

4.2 Aim 2: Generate a set of high-confidence spectrum-peptide associations from peptides with controlled composition under controlled conditions.

While much of the probabilistic model described above can be validated with existing data, the model will undoubtedly make new predictions. Testing these predictions will likely require new data. Consequently, I will conduct four series of experiments to generate new data to test the model; all should yield useful but qualitatively different data.

In the first experiment series, I will obtain a set of high-confidence spectrum-peptide associations from a small number of synthesized peptides (approximately two dozen). The specific peptides used will depend greatly on the predictions made by the probabilistic model. If it detects trends that are not present in the literature (e.g. enhanced cleavage C-terminal to alanine), then peptides will be synthesized to test these specific predictions in detail. If the model only confirms known trends (e.g. the known enhanced cleavage C-terminal to histidine), then the peptides will be designed to broadly confirm these trends. In either case, the peptides will be designed in pairs so that each member of a pair differs from the other member by only one feature. One possible peptide pair would include a peptide that contains an influential residue (such as arginine) and one that does not, or a peptide of length n and another of length $n+1$ with an inserted neutral residue.

While analyzing peptides of controlled composition offers a way of obtaining relatively confident peptide-spectrum associations, a broad sampling of peptides in this manner would quickly become prohibitive. In order to efficiently and systematically test the effects of specific residues on peak intensity, I will employ a second, related approach: whole proteome chemical modification. By treating protein samples with simple techniques, residues can be modified in ways known to affect spectrum peak intensity. For example, the negative charge on aspartate or glutamate can be abolished by methyl esterfication of their carboxylic termini using methanolic HCl. Alternatively, the positive charge on lysine or arginine can be eliminated by acetylation of their amino termini

using acetic anhydride. The spectra generated from these modifications can be identified simply by searching against a sequence database with a currently existing algorithm such as SEQUEST while taking the modified residue masses into account. Spectra from the modified and unmodified peptides will be directly compared, offering a control for testing the effects of specific changes in residue composition across an entire proteome.

The third experiment series will determine noise levels for the mass spectrum peaks for a few (three or four) tryptic peptides over thousands of spectra under similiar conditions. Peptides will be eluted continuously into the mass spectrometer, providing distributions for the chemical and instrument noise for specific peaks within the spectra. This data can be used for learning the variability of peak m/z and intensity, as well as providing a way of measuring false positives and negatives.

The fourth experiment series will search a small space of varying experimental conditions. Pertinent variables include peptide concentration; solvent conditions such as salt or organic concentration; the presence or absence of other peptides, including complex proteome digests; ion collision energy; and electrospray voltage. Manipulating these qualities will allow measurement of the effects these physical properties have on spectrum peak intensity.

4.3 Aim 3: Incorporate the intensity model into a sequence database search algorithm.

The probabilistic model described in Aim 1 and tested using data generated in Aim 2 will likely find its most useful application in a peptide identification algorithm. The graphical model generates a probability distribution over all spectra S given a peptide p, or $Pr(S = s|P = p)$. For peptide identification from a spectrum, I need to be able to calculate the inverse, or $Pr(P = p|S = s)$ for a given spectrum s. These two values are related by Bayes's theorem:

$$Pr(p|s) = \frac{Pr(s|p)Pr(p)}{Pr(s)} \tag{4}$$

This formula requires knowledge of two priors, $Pr(p)$ and $Pr(s)$. In the initial implementation, I will use uniform priors across all peptides and spectra. The peptide most probably associated with a given spectrum would then be $argmax(Pr(p|s))$. Unfortunately, calculating $Pr(p|s)$ for

all peptides within a sequence database is very computationally expensive. To solve this common problem, many search algorithms restrict the search to peptides within a small m/z or mass window around the m/z or mass of the precursor peptide, combined with a quick preliminary scoring step (Eng et al., 1994; Perkins et al., 1999). This approach is equivalent to using a prior $Pr(p) = 0$ for all peptides outside the mass window and a uniform prior for all peptides within the window. Some other possible priors include those that reflect the relative abundance of peptides within a sequence database (Perkins et al., 1999), or that take into account peptide trypticity, probable peptide levels, probable peptide elution time, peptide precursor mass measurement error rate, or other factors that might skew the probability of observing p. If calculating $Pr(s|p)$ is fast (which will be unknown until implementation of the model), then the mass window approach will likely work fine, and regardless will be useful for comparing the measurement against other search algorithms.

If, on the other hand, calculating $Pr(s|p)$ is found to be too slow, then our probability distribution approach offers a means of further reducing the number of candidate peptides to be searched. Each peptide will have some most probable path through its model, which represents the maximum likelihood peak or handful of peaks (as opposed to an entire spectrum). Using dynamic programming, these highest peaks can be calculated more rapidly than the full probability distribution for all peaks. Thus, I can restrict peptides by calculating the full conditional probability $Pr(p|s)$ only if there is sufficient overlap between the highest peaks in the spectrum and the peptide's predicted highest peaks, according to some heuristic. This approach increases speed, but it has the disadvantage of unnecessarily eliminating some peptides, thus increasing false negatives; where the costs of this decrease in sensitivity outweigh the benefits of increased speed will be determined by the specific peptide identification problem.

By taking the maximum likelihood spectrum $argmax Pr(s|p)$, rather than using the full spectrum distribution, I can incorporate this probabilistic model into a framework such as that used by SEQUEST (see Section 2.1). Demonstrating the validity of the maximum likelihood spectrum estimate will be an important part of validating the model.

A final (and particularly useful) aspect of having a value $Pr(p|s)$ is that it can be directly incorporated into further probabilistic frameworks. An obvious application is to use $Pr(p|s)$ to calculate the probability of a *collection* of spectra being produced by a particular *protein*. For example, the probability $Pr(p|s)$ could be incorporated as a Fisher kernel (Jaakkola and Haussler,

Table 2: Proposed schedule of research.

Project	Q3 05	Q4 05	Q1 06	Q2 06	Q3 06	Q4 06	Q1 07
Model Design Refinement	X						
Model Implementation	X	X					
Experimental Validation		X	X	X			
Model Design Refinement				X			
Search Implementation					X		
Search Evaluation					X	X	
Thesis writing						X	X

1998) into an SVM classifier for identifying valid peptide-spectrum matches. Describing algorithms for these applications is beyond the scope of this thesis proposal. It is nonetheless clear that there are promising applications for the results of this research.

4.4 Plan of research

Table 2 shows a proposed schedule of research. I will begin by implementing and training the spectrum graphical model. This will be followed by generation of in-house data for further experimental validation of the model. I will incorporate the model into a sequence database search algorithm and evaluate this algorithm. Finally, I will write my thesis, defending on April 1, 2007.

5 Alternative Approaches

In this section, I present two alternative approaches, one for Aim 1 and one for Aim 3. In each case, rather than first modelling the fragmentation process and then using that model to make the desired predictions (peak intensity values in Aim 1 and valid peptide-spectrum matches in Aim 3), I use a learning algorithm known as the support vector machine (SVM) (Boser et al., 1992; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000) to attack each problem directly. For Aim 1, the alternative to the graphical model is to use an SVM regression to predict intensity values from a peptide. For Aim 3, the alternative to using the graphical model for sequence database search is to use an SVM to classify peptide-spectrum matches as either "good" or "bad." I do not provide an alternative approach for Aim 2.

5.1 Alternative Approach to Aim 1: SVM regression on intensity values

SVM regression offers an alternative to using the graphical model to solve the problem of predicting mass spectrum peak intensities from a peptide. SVM regression is a machine learning algorithm that can be trained to estimate a function that maps pairs of inputs to real-valued outputs. In this case, I would train the SVM regression to predict peak intensity given a b-ion peptide sequence B and a y-ion sequence Y. While using only b-ions and y-ions is an oversimplification, modifications are possible that would allow more sophisticated comparisons. A separate regression could be performed for the b-ion intensities and the y-ion intensities. Predicting a spectrum would simply involve iterating over all possible b-ion/y-ion pairs from a peptide and using the SVM to predict the intensity value at each ion's m/z location.

To perform this regression I define a kernel that will measure similarity between a pair of sequence fragmentation events $K((B,Y),(B',Y'))$. A simple example of such a kernel would compare the residues immediately flanking each cleavage, returning a similariy of 1 when they match and 0 when they do not. More sophisticated sequence similarity kernels are possible. Examples include the spectrum kernel (Leslie et al., 2002) and the mismatch kernel (Leslie et al., 2003); an optimal kernel might combine these with domain knowledge, such as the gas-phase basicity of the b-ions and y-ions, their length, and presence or absence of important residues (such lysine, arginine or histidine).

The disadvantage of using the SVM regression approach is that it does not provide any easily accessible insight into the probabilities of various peptide fragmentation pathways. The SVM will also likely be slow computationally, but this might be true for the probability model as well.

5.2 Alternative Approach to Aim 3: SVM classification of peptide-spectrum matches

An alternative to using the graphical model to solve the problem of identifying peptides from mass spectra is to use an SVM to classify "good" and "bad" peptide-spectrum matches. An SVM is a machine learning algorithm that can be trained to classify data into two categories (as opposed to the regression described in above). In this case, I will train the SVM to discriminate between correct and incorrect peptide-spectrum matches, following the work of Anderson et al. (2003).

For the SVM to learn to classify each peptide-spectrum match correctly, I would first extract a vector of features $\hat{f}$ from each known peptide-spectrum training example (P, S). These features should discriminate between positive and negative examples, either alone or in conjunction with others. I will then compare each vector using a kernel $K(\hat{f}, \hat{f}')$, effectively projecting each vector into a higher-dimensional feature space, where the SVM algorithm finds a maximum-margin hyperplane separating the two classes. The specific kernels tested will be a polynomial kernel of degree d, $K(X, Y) = (X \cdot Y + 1)^d$; and a radial basis kernel of width σ, $K(X, Y) = exp((X \cdot Y)/2\sigma^2)$. The resulting hyperplane is then defined by a weighted subset of the training example feature vectors, and can be used to classify unknown peptide-spectrum matches as either positive or negative, allowing peptide identification.

In this particular problem, the features used to discriminate between peptide-spectrum matches would take into account the presence or absence of the ions predicted for each peptide. One feature might be the percentage of all predicted b-ions observed; another the percentage of predicted y-ions; another the fraction of neutral losses of H_20, NH_3 or CO observed. These features could also include the percentage of peaks predicted from cleavage events C-terminal or N-terminal to particular amino acids, or between particular residue pairs. In fact, an SVM trained on such features would make an excellent benchmark comparison for the probability assigned to peptide-spectrum matches by the graphical model.

The simple features outlined here for use by the SVM are not likely to fully detect the subtleties of peptide-spectrum matches; these subtleties might, however, be detectable by a probabilistic graphical model that takes into account peptide fragmentation.

6 Significance

Each of the three components of my planned research has clear potential benefits for advancing both scientific knowledge and human welfare. The illumination of peptide fragmentation chemistry from Aim 1 is the first and most obvious benefit. Even if the model developed from my research is not the final word on this subject, the information gleaned from a probabilistic approach trained on real data could be directly incorporated into other models of peptide fragmentation.

Another problem addressed by this research is the widely recognized paucity of high-quality

publicly available mass spectrometry data (Prince et al., 2004). The peptide-spectra matches generated from peptides of controlled composition in Aim 2 will be made available to other researchers, mitigating the data scarcity problem. This data will be useful to researchers in developing their own models of peptide fragmentation or spectrum identification algorithms. I am not aware of any data regarding systematic chemical modifications of a whole proteome is publicly available; this data will be a particularly valuable resource for other researchers.

Finally, the research presented in Aim 3 address some of the problems that currently plague protein identification using mass spectrometry. One example is the large number of spectra that are unidentified in a typical mass spectrometry analysis. More accurate models of peptide fragmentation will increase sensitivity, reducing the number of unidentified spectra. Increased sensitivity will also help with one of mass spectrometry's main weaknesses: a lack of dynamic range. Another problem addressed by this research is the inadequacy of current methods of assigning confidence levels to peptide-spectrum matches. A probabilistic model eliminates this problem by providing an easily interpreted probability for a peptide-spectrum match.

From a broader perspective, anything that assists protein identification will be useful for addressing myriad biological problems. Improved protein identification could be useful for biomarker detection and distinguishing disease phenotypes. Ultimately, the research presented here contributes to complete knowledge of an organism's proteome, knowledge of which would have inestimable value both for modern biologists and human welfare.

References

Aebersold, R. and D. R. Goodlett (2001). Mass spectrometry in proteomics. *Chemistry Reviews 101*, 269–295.

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

Anderson, D. C., W. Li, D. G. Payan, and W. S. Noble (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), 137–146.

Arnott, D., J. Shabanowitz, and D. F. Hunt (1993). Mass spectrometry of protein and peptides: sensitive and accurate mass measurement and sequence analysis. *Clinical Chemistry 39*(9), 2005–2010.

Bafna, V. and N. Edwards (2001). SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics 17*, S13–S21.

Bartels, C. (1990). Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environmental Mass Spectrometry 19*, 363–368.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th Annual ACM Workshop on COLT*, Pittsburgh, PA, pp. 144–152. ACM Press.

Brancia, F. L., A. Butt, R. J. Beynon, S. J. Hubbard, S. J. Gaskell, and S. G. Oliver (2001). A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis 22*(3), 552–559.

Breci, L. A., D. L. Tabb, J. R. Yates, III, and V. H. Wysocki (2003). Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Analytical Chemistry 75*(9), 1963–1971.

Clauser, K. R., P. R. Baker, and A. L. Burlingame (1999). Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry 71*, 2871.

Craig, R. and R. C. Beavis (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid communications in mass spectrometry 17*, 2310–2316.

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge UP.

Dancik, V., T. Addona, K. Clauser, J. Vath, and P. Pevzner (1999). *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology 6*(3-4), 327–342.

Dongre, A. R., J. L. Jones, A. Somogyi, and V. H. Wysocki (1996). Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. *Journal of the American Chemical Society 118*, 8365–8374.

Downard, K. M. and K. Biemann (1995). Charging behavior of highly basic peptides during electrospray ionization. A predilection for protons. *International Journal of Mass Spectrometry 148*, 191–202.

Edwards, N. and R. Lippert (2002). Generating peptide candidates from amino-acid sequence databases for protein identification via mass spectrometry. *WABI*, 68–81.

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

Eng, J. K., A. L. McCormack, and J. R. Yates, III (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry 5*, 976–989.

Fernandez de Cossio, J., J. Gonzales, and V. Besada (1995). A computer program to aid the sequencing of peptides in collision-acitvate decomposition experiments. *Comp. Appl. Biosci. 11*, 427–434.

Field, H. I., D. Fenyo, and R. C. Beavis (2002). Radars, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics 2*, 36–47.

Gu, C., G. Tsaprailis, L. Breci, and V. H. Wysocki (2000). Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Analytical Chemistry 72*, 5804–5813.

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

Huang, Y., V. H. Wysocki, D. L. Tabb, and J. R. Yates, III (2002). The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptide. *International Journal of Mass Spectrometry 219*, 233–244.

Jaakkola, T. and D. Haussler (1998). Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, San Mateo, CA. Morgan Kauffmann.

Johnson, R. J. and K. Biemann (1984). Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed. Environmental Mass Spectrometry 18*, 945–957.

Jordan, M. I. (2005). Graphical models.

Kapp, E. A., F. Schutz, G. E. Reid, J. S. Eddes, R. L. Moritz, R. A. J. O'Hair, T. P. Speed, and R. J. Simpson (2003). Mining a tandem mass spectrometry database to determine trends and global factors influencing peptide fragmentation. *Analytical Chemistry 75*, 6251–6264.

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

Klammer, A. A. and M. J. MacCoss (2005). Improved protein identification with mass spectrometry using an immobilized trypsin column.

Klammer, A. A., C. C. Wu, M. J. MacCoss, and W. S. Noble (2005). Peptide charge state determination for low-resolution tandem mass spectra.

Leslie, C., E. Eskin, and W. S. Noble (2002). The spectrum kernel: A string kernel for SVM protein classification. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein (Eds.), *Proceedings of the Pacific Symposium on Biocomputing*, New Jersey, pp. 564–575. World Scientific.

Leslie, C., E. Eskin, J. Weston, and W. S. Noble (2003). Mismatch string kernels for SVM protein classification. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems*, Cambridge, MA, pp. 1441–1448. MIT Press.

MacCoss, M. J., M. J. Toth, and D. E. Matthews (2001). Evaluation and optimization of ion-current ratio measurements by selected-ion-monitoring mass spectrometry. *Analytical Chemistry 73*(13), 2976–2984.

Mann, M., R. C. Hendrickson, and A. Pandey (2001). Analysis of proteins and proteomes by mass spectrometry. *Annual Review of Biochemistry 70*, 437–473.

Mann, M. and M. Wilm (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry 66*, 4385–4564.

McCormack, A. L., D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. R. Yates, III (1997). Direct analysis and identification of proteins in mixtures by LC-MS/MS and database searching at the low-femtomole level. *Analytical Chemistry 69*(4), 767–776.

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

Paizs, B. and S. Suhai (2004). Fragmentation pathways of protonated peptides. *DOI:10.1002/mas.20024*.

Pandey, A. and M. Mann (2000). Proteomics to study genes and genomes. *Nature 405*, 837–846.

Papping, D. J. C., P. Hojrup, and A. J. Bleasby (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol. 3*, 327–332.

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

Pevzner, P. A., V. Dancik, and C. Tang (2000). Mutation-tolerant protein identification by mass spectrometry. *Journal of Computational Biology 7*, 777–787.

Polce, M. J., D. Ren, and C. Wesdemiotis (2000). Dissocation of the peptide bond in protonated peptides. *Journal of the American Society for Mass Spectrometry 35*, 1391–1398.

Prince, J. T., M. W. Carlson, R. Wang, P. Lu, and E. M. Marcotte (2004). The need for a public proteomics repository. *Nature Biotechnology 22*, 471–472.

Qin, J. and B. Chait (1999). Collision-induced dissociation of singly charged peptide ions in a matrix-assisted laser desorption ionization trap mass spectrometer. *International Journal of Mass Spectrometry 191*, 313–320.

Roepstorff, P. and J. Fohlman (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry 11*, 601.

Schutz, F., E. A. Kapp, R. J. Simpson, and T. P. Speed (2003). Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochemical Society Transactions 31*, 1479–1483.

Simpson, R. J., L. M. Connolly, J. S. Eddes, J. J. Pereira, R. L. Moritz, and G. E. Reid (2000). Proteomic analysis of the human colon carcinoma cell line (lim 1215): development of a membrane protein database. *Electrophoresis 21*, 1707–1732.

Summerfield, S. G., A. Whiting, and S. J. Gaskell (1997). Intra-ionic interactions in electrosprayed peptide ions. *International Journal of Mass Spectrometry 162*, 149–161.

Tabb, D. L., M. J. MacCoss, C. C. Wu, S. D. Anderson, and J. R. Yates, III (2003). Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical Chemistry 75*, 2470–2477.

Tabb, D. L., W. H. McDonald, and J. R. Yates, III (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *Journal of Proteome Research 1*(1), 21–26.

Tabb, D. L., L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin, and J. R. Yates, III (2003). Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Analytical Chemistry 75*, 1155–1163.

Taylor, J. A. and R. S. Johnson (1997). Sequence database searches via *de novo* peptie sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry 11*, 1067–1075.

Tyers, M. and M. Mann (2003). From genomics to proteomics. *Nature 422*, 193–197.

van Dongen, W. D., J. I. T. van Wijk, B. N. Green, W. Heerma, and J. Haverkamp (1999). Comparison between collision induced dissociation of electrosprayed protonated peptides in the up-front source region and in a low-energy collision cell. *Rapid communications in mass spectrometry 13*, 1712–1716.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Adaptive and learning systems for signal processing, communications, and control. New York: Wiley.

Venable, J. D. and J. R. Yates, III (2004). Impact of ion trap tandem mass spectra variability on the identification of peptides. *Analytical Chemistry 76*(10), 2928–2937.

Wan, Y. and T. Chen (2005). A hidden markov model based scoring function for mass spectrometry database search.

Wu, Q., S. V. Orden, X. Cheng, R. Bakhtiar, and R. D. Smith (1995). Characterization of cytochrome c variants with high-resolution FTICR mass spectrometry: correlation of fragmentation and structure. *Analytical Chemistry 67*(14), 2498–2509.

Wysocki, V. H., G. Tsaprailis, L. L. Smith, and L. A. Breci (2000). Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of the American Society for Mass Spectrometry 35*, 1399–1406.

Yates, III, J. R. (1998a). Database searching using mass spectrometry data. *Electrophoresis 19*, 893–900.

Yates, III, J. R. (1998b). Mass spectrometry and the age of the proteome. *Analytical Chemistry 33*, 1–19.

Yates, III, J. R., J. K. Eng, A. L. McCormack, and D. Schieltz (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry 67*, 1426–1436.

Yergey, J. A. (1983). A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry 52*, 337–349.

Zhang, W. and B. T. Chait (2000). Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Analytical Chemistry 72*, 2482–2489.

Zhang, Z. (2004). Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry 76*, 3908–3922.