Protein Identification from Tandem Mass Spectra with Probabilistic Language Modeling

Yiming Yang, Abhay Harpale, and Subramaniam Ganapathy

Carnegie Mellon University,
Pittsburgh, PA – 15217, United States

Abstract. This paper presents an interdisciplinary investigation of statistical information retrieval (IR) techniques for protein identification from tandem mass spectra, a challenging problem in proteomic data analysis. We formulate the task as an IR problem, by constructing a "query vector" whose elements are system-predicted peptides with confidence scores based on spectrum analysis of the input sample, and by defining the vector space of "documents" with protein profiles, each of which is constructed based on the theoretical spectrum of a protein. This formulation establishes a new connection from the protein identification problem to a probabilistic language modeling approach as well as the vector space models in IR, and enables us to compare fundamental differences in the IR models and common approaches in protein identification. Our experiments on benchmark spectrometry query sets and large protein databases demonstrate that the IR models significantly outperform well-established methods in protein identification, by enhancing precision in high-recall regions in particular, where the conventional approaches are weak.

Keywords: Proteomics, Information Retrieval

1 Introduction

Statistical pattern matching technologies have been successfully applied to many real-world problems. Among those, text-based information retrieval (IR) is perhaps one of the most intensively studied and highly successful areas. Computational biology is another important area where pattern matching plays a key role in various forms of data analysis. This paper presents an interdisciplinary investigation, focusing on how to generalize good ideas and successful technologies in one domain (IR) into new insights and novel solutions in another (computational proteomics). Specifically, we focus on the problem of protein identification from detected peptides in tandem mass spectra.

Protein identification is important for discovering biomarkers linked to diseases, therapeutic outcomes and individualized drug toxicity. Tandem mass (MS/MS) spectra, generated by a chemical process over complex biological samples such as tissues or blood, contain rich information about proteins and peptides which are constituents of proteins. Protein identification from MS/MS data is typically carried

out in two steps. Step 1 is to predict peptides based on observed empirical spectra, and step 2 is to predict proteins based on the predicted peptides. Many technical solutions have been developed for the peptide identification step in the past two decades, including commercially available software [1], [2], [4], [5], [6], [7], [8]. However, for the second step, the current literature is relatively sparse. Particularly, few interdisciplinary investigations were conducted for exploring the potential of advanced IR technologies in solving the mapping from predicted peptides to proteins. Addressing this research gap is the primary motivation of this paper, and we focus on the second step in particular, i.e., the mapping from system-predicted peptides to the true proteins in large protein databases.

Why would it be beneficial to bridge the technical and methodological gaps between the fields of protein identification and text retrieval? At first glance, the two tasks look totally different. However, at a higher level of abstraction, the two tasks and related technical solutions have important properties in common. If we consider peptides as words, proteins as documents, and peptide identification from spectra as a query generation process, then the mapping from predicted peptides to proteins in a protein database is just like ad-hoc retrieval in a vector space, albeit a particularly high-dimensional one. A database with tens of thousands of proteins would contain tens of millions of unique peptides. Some common peptides could be considered in analogous to stop-words in text, while the majority of peptides are much rarer, form a highly skewed distribution over proteins. This means that the rich body of research findings in text retrieval would provide meaningful insights into how to weight peptides in proteins, how to combine peptide-level evidence into predictions of proteins, and how to leverage state-of-the-art IR methods directly or with adaptation, including efficient inverted indexing, effective term weighting schemes, smoothing and dimensionality reduction techniques, choices of similarity measure in retrieval models, well-understood evaluation metrics, and standardized software toolkits like Lemur and Indri [9][10]. In order to leverage those potentials we need a good understanding of the background knowledge and related literature, including how biological samples, MS/MS spectra, peptides and protein sequences are related to each other, what kinds of technical solutions have been developed for peptide identification from spectra and for protein identification from predicted peptides, how the current solutions have been evaluated and compared, what the strengths and weaknesses of those methods, and how can we apply or adapt retrieval techniques to create better solutions. Achieving such an understanding is the primary contribution we target in this paper. Specifically, our main contributions can be listed as:

1) A new formulation of the protein-prediction task that enables probabilistic modeling with joint use of well-established peptide identification techniques and domain knowledge about protein sequences, as well as the rich developments in IR on language modeling and vector space models.
2) A comparative theoretical analysis of two probabilistic models for combining peptide-level evidence in the prediction of in-sample proteins: one is the well-establish method by Nesvizhskii et al., which scores candidate proteins based on estimated probability of a Boolean OR function, and the other is a language-modeling method that we propose, which uses the estimated probability of a Boolean AND instead. We show that the former has a weakness in

discriminating true positives from false positives in the high-recall region of the system's predictions, and that the latter addresses such a weakness in a principled way.

3) A thorough evaluation on standard MS/MS datasets and large protein databases for comparison of methods, including probabilistic OR, probabilistic AND, cosine-similarity with a robust TF-IDF term weighting scheme, and the commonly used X!Tandem commercial software in proteomic applications. We observed significant performance enhancement by the IR models tested over the conventional methods in the current literature of protein identification from tandem mass spectra .

The rest of the paper is organized as follows. Section 2 outlines related background and representative approaches in protein identification from tandem mass spectra. Section 3 defines our new framework and discusses its connection to well-established techniques in text retrieval. Section 4 introduces three benchmark query sets (spectrometry samples) and the corresponding large protein databases for empirical evaluation. Section 5 reports the experiments and the results. Section 6 summarizes the main findings and conclusions.

2 Background and Related Work

Statistical approaches for data analysis with tandem mass spectra is an important and fast growing area in recent computational biology research. Analogous and complementary to micro-array data which are highly informative for analyzing gene-level activities under various conditions, MS/MS spectra contain rich information about proteins which are potentially responsible for diseases, therapeutic responses and drug toxicity [3]. MS/MS spectra are generated using liquid chromatography where a sampled tissue or a blood drop is digested into peptides which are segments of protein sequences. The peptides are further separated into ionized fragments and analyzed to produce MS/MS spectra. Each spectrum is a list of spikes: the location of each spike is the mass/charge (m/z) ratio of an ionized fragment, and the magnitude of the spike is the abundance or intensity of the fragment. An input sample is typically a mixture of multiple proteins but the exact number of proteins is unknown in advance. Standard MS/MS datasets for benchmark evaluations were typically constructed for sample mixtures that contain a dozen or a few dozens of proteins [19]. The numbers of MS/MS spectra obtained from those samples are in the range of a few thousands. The task of protein identification is to find a mapping from the few thousands of observed MS/MS spectra to the true proteins in the input sample. It is typically accomplished in two steps: first, identify the peptides based on observed spectra; second, predict proteins based by system-predicted peptides.

In peptide identification research, database search techniques have been commonly used to select a candidate set of peptides based on the degree of matching between the "theoretical" (expected) mass spectra of candidate peptides in a protein database and the empirical spectra in the input sample [1],[2],[4],[5],[6], [7], [8]. The theoretical spectrum of each peptide can be automatically derived by rules from the amino acid sequences of proteins. Each known protein has a unique amino acid

sequence, which can be segmented by rules into peptide-level subsequences. The theoretical spectrum of each peptide can also be automatically generated based on existing knowledge about lower-level chemical properties of amino acid letters. The number of unique peptides in a protein database can be very large. For example, we found roughly 5 million unique peptides in a sample of 50,000 proteins as typical. By comparing each theoretical spectrum against the empirical spectra in an input sample, a system obtains a confidence score of each candidate peptide. Further, applying some threshold to the confidence scores yields peptide assignments by the system. The similarity measures differ from system to system. SEQUEST, for example, one of the most commonly used commercial programs in practical applications as well as in comparative evaluations of peptide identification methods on benchmark datasets, employs a Fourier Transform cross-correlation strategy [4]. X!Tandem is another popular open-source software for peptide/protein identification and has been commonly used as a baseline in comparative evaluations. It uses aggressive thresholds to reduce false-alarms and to enhance computational efficiency, and produces a ranked list of proteins for each input sample [21].

In protein identification based on system-predicted peptides from MS/MS spectra, the ProteinProphet system by Nesvizhskii et al [12] is among the most commonly used in comparative evaluations of methods on benchmark datasets. This system uses SEQUEST-predicted peptides as the input, and converts the non-probabilistic confidence scores by SEQUEST to the probabilistic scores for peptide assignments. Specifically, they used the Expectation-Maximization (EM) algorithm to obtain a mixture model for true positives and false positives in system-predicted peptides. Some empirical evaluations [6] showed performance improvement by the score refinement method over that of the original SEQUEST. ProteinProphet estimates the probability of each protein being present in the input sample using the probability that at least one of the constituent peptides in the protein is a corrected assignment to the sample. To be explicit, suppose $q_j \in [0,1]$ is the estimated probability of the presence of peptide j in the input sample. The probability that a protein i is present in the input sample is calculated in ProteinProphet as

$$p_i = 1 - \prod_{j=1}^{J} (1 - q_j) \ .$$

This formula calculates the estimated probability of the Boolean-OR function over the peptide-level evidence, assuming that the occurrence (being present or not) of each peptide is an *identically independently distributed* (i.i.d.) random event with $q_j \in [0,1], \forall j$. If any constituent peptide of a protein is predicted as present by the system, we have $q_j = 1, \exists j$ and $p_i = 1$ as the consequence. Clearly, the protein scoring function in ProteinProphet is the estimated probability for Boolean OR logic. We will refer to this method as prob-OR in the rest of the paper. A refined version of this method is also supported by the system, i.e., an EM algorithm is used to find hidden groups of proteins, and the peptide probabilities are estimated conditioned on the hidden groups of proteins instead of individual proteins.

Other work of a similar nature in protein identification includes that by MacCoss et al. [11] who used a modified version of SEQUEST to generate peptide assignments

with normalized scores, and performed protein-level predictions with a prob-OR equivalent operation. Moore et al. [13] pursued a different but heuristic approach: after aggressive removal of low-scoring candidate peptides, the product of the scores of the remaining peptides that constitute a protein sequence is used to estimate the quality of the match for the protein. Theoretical comparison of their method with the probabilistic models (including Nesvizhskii et al., and others) is difficult because their scoring functions are heuristically or procedurally defined, not explicitly probabilistic; empirical comparison was not reported on the other hand. The recent work by Li et. al. [22] presents another interesting alternative which predicts proteins by modeling the input sample as a multi-protein mixture and finding the Maximum-a-Posteriori (MAP) solution for the mixture weights. They used Gibbs sampling as an approximation method because solving MAP exactly is computationally intractable. Although no theoretical upper/lower bound is guaranteed by the approximation, an empirical evaluation on a new (their own) dataset shows improved results over that of Nesvizhskii's method (prob-OR). However, repeating this comparative evaluation has been difficult as the dataset is not publicly available, and no sufficient details were published about how to reconstruct the dataset from publicly available protein databases. Other indirectly related work includes CHOMPER [14], INTERACT [15] and DTASelect [16], which focus on visualization and filtering tools for manual interaction in protein identification, and Mascot [3] and Sonar [17] which focus on commercial tool development.

3 Methods

The desiderata for a new approach are: 1) a theoretically justified function (or family of functions) for combining peptide-level evidence, and 2) higher performance in standard metrics such as average precision, compared to the best results reported in the MS/MS literature. To address these objectives we turn to modern IR approaches for mapping predicted peptides to proteins.

Notice that the commonly used prob-OR type of functions in protein scoring has a potential weakness. That is, it has the tendency to produce many false alarms due to an overly simplistic assumption because if any constituent peptide of a protein is detected, then the protein is assumed as a correct assignment. As an alternative, we propose a mapping with stronger constraints, i.e. using a probabilistic AND (prob-AND) function to combine evidence in predicted peptides. More precisely, we propose to

1) translate the predicted peptides into an empirical in-sample distribution of peptides as observed in the MS/MS spectra;
2) use the relative frequencies of peptides in the amino acid sequence of each protein as the protein profile; and
3) measure the Kullback-Leibler (KL) divergence of each protein-specific distribution of peptides from the sample distribution of peptides.

These steps together accomplish a prob-AND mapping from the predicted peptides to candidate proteins with probabilistic scores.

3.1 Data representations

The input to our protein identification system is a set of peptides with confidence scores which are produced by a well-established method for peptide identification from a sample of MS/MS spectra [6]. We present the scored peptides using a vector $\vec{q} = (q_1, q_2, \cdots, q_J)$ whose elements $q_j \in [0,1]$ are normalized so that they sum to one, and J is the number of total unique peptides being identified. For convenience, we call vector $\vec{q}$ the "query" for protein search. Notice that a peptide identification method may not generate normalized scores. In that case, we translate scores as following:

$$a = \min\{q_1, q_2, \cdots, q_J\}, \quad b = \max\{q_1, q_2, \cdots, q_J\},$$

$$q_j^{'} = \frac{q_j - a}{b - a}, \qquad q_j^{''} = \frac{q_j^{'}}{\sum_{t=1}^{J} q_t^{'}}.$$

We also define a normalized vector (profile) for each protein in the target database (DB) as:

$$\vec{p}_i = (p_{i1}, p_{i2}, \cdots, p_{iJ}), \quad p_{ij} = \frac{n_{ij}}{\sum_{j=1}^{J} n_{ij}},$$

where n_{ij} is the count of peptide j in protein i. Notice that query normalization is generally not required in text retrieval methods because it does not effect the ranking of documents given a query. Similarly, in our mapping from a "bag" of system-predicted peptides to protein profiles, query normalization does not affect the ranking of proteins given a query. However, with explicit normalization of both the query vector and protein profiles we can intuitively interpret the mapping criterion based the KL-divergence between the two types of vectors (Section 3.2).

We smooth the peptide probabilities using a Dirichlet prior [18], modifying the elements as

$$p_{ij} = \frac{n_{ij} + \mu \pi_j}{\sum_{j=1}^{J} n_{ij} + \mu}$$

Parameter μ controls the degree of smoothing, and $\pi_j \in [0,1]$ is calculated as:

$$\pi_j = \frac{\sum_{i \in DB} n_{ij}}{\sum_{t=1}^{J} \sum_{i \in DB} n_{it}}.$$

Smoothing is a crucial step in the formulation of protein profiles. As discussed earlier, the peptide identification step identifies peptides in the sample which are the result of the protein cleavage, i.e. breaking of protein into its constituent peptides upon the reaction with a chemical cleaving agent (e.g. Trypsin). It is not guaranteed that each

protein breaks at every peptide boundary (phenomenon known as miscleavage), and consequently, not every constituent peptide is necessarily observed and not every observed component is necessarily a valid peptide. Smoothing therefore is necessary for assigning non-zero weights to unobserved peptides, just as the out-of-vocabulary words need to be handled in language modeling for document retrieval. To ensure that all the observed components (both valid peptides and peptide concatenations) are taken into account, we simulated miscleavages in creation of protein profiles.

Why do we construct protein profiles in the above way? Because we want to leverage the domain knowledge about amino acid sequences and establish the mapping from peptides to proteins accordingly. Ideally, we would like to have a large training set of MS/MS spectra with explicitly labeled correspondences to positively and negatively related proteins in a target database, which would enable supervised learning of the conditional distribution of peptides given a protein in MS/MS samples. However, such a large training set would be very expensive to produce and is not currently available in open-source benchmark datasets for protein identification evaluations. The only knowledge we have for relating predicted peptides for an input sample to the proteins in a target database are 1) the peptide occurrences in amino acid sequences of proteins, and 2) the expected (theoretical) spectrum of each valid peptide. Thus, we stay with the unsupervised setting for the mapping problem, i.e., by constructing a peptide-based profile for each protein, and by conducting proximate-search over protein profiles given a synthetic query. The normalization of both the query vector and the profile vectors of proteins enables probabilistic interpretation for the mapping criterion, and avoids an unjustified bias of favoring longer proteins (i.e. with a larger number of constituent peptides) over shorter ones, as present in the prob-OR approaches. As for the need of smoothing, it is well understood in statistical learning theory and practice that model smoothing is particularly important when the feature (input variable) space is very large and the observed data is highly sparse. In our problem, the feature space consists of a large number of peptides, with a skewed distribution over a modest number of protein sequences. For example, the PPK benchmark dataset (Section 4) contains 4,534 proteins and 325,812 unique peptides. This means that most protein profiles are both high-dimensional and extremely sparse, and that appropriate smoothing is necessary for successful mapping from a query to candidate proteins.

3.2 Protein scoring based on prob-AND

The choice of scoring criterion is obviously crucial for successful ranking of proteins given a query. We may consider the presence or absence of a peptide in the predicted list as a random variable, where the randomness comes from both the sampled protein mixture, and the noisy process of generating MS/MS spectra from the protein mixture. Consequently, we may view vector $\vec{q}$ as the empirically observed in-sample distribution of peptides in an unknown protein mixture. Similarly, we may view vector $\vec{p}_i$ as the "theoretical" peptide distribution in a specific protein, derived based on the amino acid sequences of proteins in a target database, and existing knowledge (rules) about how protein sequences decompose to peptides. We use the cross entropy

to score each candidate protein with respect to the query. The cross entropy of the two distributions is defined as:

$$H(\vec{q} \,||\, \vec{p}_i) \equiv -\sum_{j=1}^{J} q_j \log p_{ij} = -\sum_{j=1}^{J} q_j \log\left(q_j \frac{p_{ij}}{q_j} \right)$$

$$= -\sum_{j=1}^{J} q_j \log q_j + \sum_{j=1}^{J} q_j \log\left(\frac{p_{ij}}{q_j} \right)$$

$$= H(\vec{q}) + D(\vec{q} \,||\, \vec{p}_i)$$

The cross entropy decouples into two terms as shown in the last line above: the first term $H(\vec{q})$ is the entropy of the query, which is the same for every protein; the second term $D(\vec{q} \,||\, \vec{p}_i)$ is the Kullback-Leibler (KL) divergence that determines the relative ranking of proteins with respect to the query. A smaller KL divergence means a better matched protein for the query.

We use prob-AND as the abbreviation of the proposed method. KL divergence has been commonly used in language modeling (LM) approaches for ad-hoc text retrieval with probabilistic ranking of documents given a query. It is a function proportional to the log-likelihood of the query conditioned on the document model under the term independence assumption in a multinomial process. Let $\vec{x} = (x_1, x_2, \cdots, x_J)$ be the vector whose elements are the within-query term frequencies. The log-likelihood is estimated as:

$$\log \Pr(\vec{x} \,||\, \vec{p}_i) = \log \prod_{j=1}^{J} p_{ij}^{\,n_{ij}} = \sum_{j=1}^{J} x_j \log p_{ij}$$

$$= N_x \sum_{j=1}^{J} \frac{x_j}{N_x} \log p_{ij}$$

where the scaling factor $N_x = x_1 + x_2 + \cdots + x_J$ is the total count of term occurrences in the query. Except for the scaling factor (which is constant given a query), the log-likelihood and KL divergence are identical. Therefore, using the multinomial probabilistic model to rank documents for a query makes no difference compared to using the negation of the KL divergence as the metric. With respect to ranking proteins given a set of predicted peptides, the only difference is that the within-query term frequencies are not directly observed but are predicted instead. Nevertheless, the connection between the log-likelihood function and KL divergence shows clearly that the logic being used for assembling partial evidence (from individual terms or peptides) is probabilistic AND, not probabilistic OR. In other words, KL divergence imposes stronger constraints in the mapping from predicted peptides to proteins. Probabilistic AND and KL divergence have not been studied for protein identification in the current literature, to our knowledge.

3.3 Connections to other vector space models for text retrieval

How are prob-AND and prob-OR related to conventional retrieval Vector Space Models (VSMs) in IR? In fact, they are closely related. Let $\vec{d}_i = (d_{i1}, d_{i2}, \cdots, d_{iJ})$ be a document vector and define within-document term weighting as

$$d_{ij} = \log p_{ij} \equiv \log \Pr(\text{term } j \mid \text{doc} i),$$

The dot-product similarity in a standard VSM is calculated as:

$$sim(\vec{q} \cdot \vec{d}_i) = \vec{q} \cdot \vec{d}_i = \sum_{j=1}^{J} q_j \log p_{ij}.$$

This is exactly the same formula in the prob-AND model, i.e., scoring function based on the cross entropy. On the other hand, if we choose $d_{ij} = p_{ij}$ as the term weighting scheme, the dot-product similarity becomes:

$$sim(\vec{q} \cdot \vec{d}_i) = \vec{q} \cdot \vec{d}_i = \sum_{j=1}^{J} q_j p_{ij} = \sum_{j \in \text{query}} q_j p_{ij}$$

This is a variant of soft OR. That is, a document (or protein) receives a positive weight where as long as any of its terms (or peptides) is found in the query. In a further extreme setting of $d_{ij} = I_{ij}$ which is the indicator function with $I(i, j) = 1$ if peptide j is a constituent of protein i and $I(i, j) = 0$ otherwise, we have:

$$score(\vec{d}_i \mid \vec{q}) = \sum_{j \in \text{query}} q_j I(i, j)$$

It also mimics the Boolean OR logic in a soft manner, obviously. There soft-OR scoring functions are closely related to the prob-OR metric in ProteinProphet which we analyzed in Section 2.

The connections from prob-OR and prob-AND to conventional VSMs invites a question: are they better choices than other variants of VSM, e.g., the commonly used cosine similarity with TF-IDF term weighting scheme? Since the latter is not a probabilistic scoring function, direct theoretical comparison on the basis of probabilistic modeling is impossible. However, an empirical comparison between these VSM variants would be highly informative and practically important for a thorough investigation on the applicability and effectiveness of advanced IR techniques in solving the protein identification problem. Hence, we report such a comparative evaluation in Section 5.

4 Datasets

For evaluation and benchmarking of protein identification algorithms, we use standard proteomic mixtures whose MS/MS spectra are publicly available. Purvine et al in 2003 introduced a standardized proteomics dataset to support comparative evaluation which consists of a query set of MS/MS spectra from a mixture of 12

proteins and 23 peptides[1] and a search database consisting of 4534 proteins [19]. The dataset was designed to mimic the complexity of large scale proteomics experiments and to serve as a standard in proteomics research. We refer to this dataset as PPK, after the authors Purvine S, Picone AF and Kolker E [19].

We also created two more datasets, called Mark12+50000 and Sigma49+50000, respectively. The Mark12+50000 dataset consists of a query set of MS/MS spectra from a 12-protein mixture (from Invitrogen, Carlsbad CA) called the 'Mark12 Electrophoresis Standard', and a target protein database which we name as M50000. The Sigma49+50000 dataset consists of the query set of MS/MS spectra from a 49 protein mixture (from Sigma-Andrich, St. Louis MO) and a target protein database which we name as S50000. Both query sets were provided by the Mass Spectrometry Research Center at Vanderbilt University and have been used as standard benchmarks in proteomics research. The target databases were generated by us by drawing two random samples from the SwissProt[2] protein database, which contains over 280,000 protein sequences, and then adding Mark12 query-set proteins to one sample and Sigma49 query-set proteins to the other sample. We chose the size (50,000) of the target protein databases to be comparable to those used in actual proteomic analyses. Tables 1 and 2 summarize the datasets[3].

Table 1. Query set statistics

Query Set	#spectra	#proteins	#peptides
PPK (queries)	2995	35	1596
Mark12	9380	12	1944
Sigma49	12498	49	4560

5 Experiments

We conducted a comparative evaluation with controlled experiments for three models: prob-OR, prob-AND, and a standard VSM model (supported by the Lemur) which uses TF-IDF ("ltc") for within-document term weighting and cosine similarity for the scoring function. We name the last method "TFIDF-cosine". We also used the popular X!Tandem software (available online) to generate an alternative baseline.

[1] The query set was generated from 12 proteins and 23 peptides. Each of the peptides was treated as a single-peptide protein in evaluation yielding a total of 35 proteins.

[2] http://expasy.org/sprot/

[3] Datasets will be made publicly available to support comparative evaluation and benchmarking at the following URL: http://nyc.lti.cs.cmu.edu/clair/datasets.htm

Table 2. Protein database statistics

Protein DB	#proteins	#peptides	#relevant proteins
PPK (protein DB)	4534	325,812	35
M50000	50012	5,149,302	12
S50000	50049	2,571,642	49

5.1 Experimental Settings

To ensure a controlled setting, all the four methods share the same query generation process. We used the publicly available software of SEQUEST [4] and the PeptideProphet[4] pipeline to predict peptides from MS/MS data, producing the queries shared by all the methods except X!Tandem for retrieving proteins. For the experiment with X!Tandem, we use the inbuilt peptide and protein identification tools in the open-source software package. When evaluating a method on one dataset, we used the remaining two datasets as the validation sets for tuning parameters. For example, when PPK is used as the test set, we tuned the smoothing method and μ (smoothing parameter) in prob-AND on Mark12+50000 and Sigma49+50000 as the validation datasets[5]. Based on the results, we chose the Dirichlet prior over Laplace as the smoothing method and μ=5000 as the smoothing parameter.

5.2 Metrics

The output of each method is a ranked list of predicted proteins for a pre-specified MS/MS dataset and a protein database. Applying a threshold to the ranked list of each method yielded binary decisions and shifting the threshold enables us to calculate precision values at different levels of recall. Using TP (true positives), FP (false positives), FN (false negatives) and TN (true negatives) to denote the counts of predictions in the four corresponding categories, the performance at a fixed threshold is measured as:

$$\text{Recall} = TP / (TP + FN),$$

$$\text{Precision} = TP / (TP + FP)$$

[4] PeptideProphet is a part of the TransProteomicPipeline, a publicly available software toolkit for protein identification at http://tools.proteomecenter.org/software.php

[5] Note that the 50000 proteins in Sigma49+50000 and Mark12+50000 are different independent samples drawn from Swissprot.

To evaluate the ranking ability of each method, we computed its average precision (over all recall levels) per query, and then the mean over all queries. This produces the standard MAP score for each method.

5.3 Main Results

The performance of the four methods in average precision is summarized in Table 3.

The main observations are the following:

- Prob-OR had a relatively weak performance, with the MAP score significantly below the levels of all the other methods except X!Tandem. This observation supports our theoretical analysis (Sections 7) on the weakness of the protein scoring functions based on Boolean-OR in assembling peptide-level evidence – they are not sufficiently powerful for discriminating true positives from false positives.

Table 3. Results summary in average precision (bold case indicates best performance)

Dataset	prob-AND	prob-OR	TFIDF cosine	X!Tandem
PPK	**0.87**	0.8	0.84	0.43
Mark12	0.77	0.66	**0.81**	0.41
Sigma49	0.48	0.44	**0.49**	0.241
MAP	**0.71**	0.63	**0.71**	0.36

- Prob-AND is among the two best methods (the other is TFIDF cosine) on average, with a MAP score of 0.71. It outperformed the prob-OR method significantly on all the datasets, successfully addressing the main weakness of the latter.
- The TFIDF-cosine method performed equally well as Prob-AND. This is not surprising from the view point of text retrieval model analysis. It has been well-understood that the conventional vector space model (VSM) using cosine and TFIDF term weighting is a good approximation of language modeling with a multinomial assumption and the Dirichlet prior of corpus-level term distribution [18]. And the latter is the foundation of our prob-AND approach. On the other hand, it is the first time that the conventional VSM is examined in protein identification and compared with prob-AND. We are pleased to see both methods worked equally well on average, and both superior to prob-OR as a strong baseline in the computational proteomics literature.
- X!Tandem, one of the most popular publicly available protein identification program that is commonly used as a comparative baseline algorithm, performed inferior to the other methods on all three datasets in our

experiments. It has been reported in the peptide/protein identification literature that X!Tandem differs from SEQUEST significantly in the identified peptides (and proteins). X!Tandem usually suffers with poor recall (sensitivity) in the peptide identification step as compared to SEQUEST based approaches [20], as a result of an aggressive thresholding strategy for computational efficiency and for reducing false alarms in the peptide identification step. Our results of X!Tandem agree with the previously reported findings in this sense.

5.4 Performance in high-recall regions

While average precision or MAP is well-accepted in evaluations of IR models, they may not be sufficiently informative for judging how much the protein identification systems would help biologists in reality. Notice that for biologists to verify the validity of the system-predicted proteins, wet-lab experiments would be needed and the cost would be much higher than what is required for a user to check through a ranked list of documents. In other words, dealing with a large number of false alarms would be too costly and hence impractical in proteomic data analysis. With this concern, we further analyze the performance of the methods in the high-recall (80%, 90% and 100%).

Table 4 shows the average numbers of false positives (FP) for each method at fixed levels of recall; the average is computed over the three datasets.

Table 4. Results summary in false positive counts (averaged over the 3 datasets) at fixed levels of recall

	Average Number of False Positives		
Recall	prob-AND	prob-OR	TFIDF-cosine
80%	28	52	28
90%	74	1002	96
100%	17746	16631	16586

It can be observed that all the methods achieved 80% recall with a relative small number of FP, which is quite encouraging. However, to achieve 90% recall, the FP number of prob-OR increased from 92 (at 80% recall) to 1002 which is unacceptably high, while prob-AND and TFIDF-cosine retain their low-FP behavior. At the 100% recall level, all the methods produced a large number of FP, which is not too surprising. X!Tandem did not reach any of the recall levels higher than 60% on all the 3 datasets, thus it is not included in the table.

5.5 Statistical significance tests

We conducted one-sample proportion tests for comparing the error rates at 90% recall levels of the protein identification methods. Table 5 summarizes the results.

Table 5. Significance test summary: each element in the matrix indicates the number of datasets (out of 3) on which System A significantly outperforms System B with a p-value < 0.01

A \ B	prob-AND	prob-OR	TFIDF-cosine	X!Tandem
prob-AND		3	1	3
prob-OR	0		0	3
TFIDF-cosine	0[6]	3		3
X!Tandem	0	0	0	

Comparing the two strongest methods, i.e., prob-AND and TFIDF-cosine, each of them significantly outperformed the other on one of the three datasets, and performed equally well on the remaining dataset. Comparing prob-OR with all the others, it significantly underperformed prob-AND and TFIDF-cosine on all three datasets. X!Tandem performance was inferior to all other approaches on all the datasets.

6 Conclusion and Future Work

In this paper, we present the first interdisciplinary investigation on how to leverage the rich research insights and successful techniques in IR to better solve the challenging problem of protein identification from tandem mass spectra. We formulated the problem (the mapping from system-predicted peptides to proteins) as an ad-hoc retrieval task, proposed a prob-AND model for combining peptide-level evidence in protein retrieval, and conducted a thorough evaluation of these models in comparison with a well-established method (prob-OR by Keller et al.) and a common baseline method (X!Tandem) in the field of protein-identification and a successful vector space model (TFIDF-cosine) in IR. The results are highly encouraging: we obtained significant performance improvements by the prob-AND models and the VSM model over the representative baseline methods. We hope this investigation provides useful information and insights for future research in adapting IR techniques to proteomic applications, and invites new ideas for further improvements from both the IR community and the computational proteomics community.

Several extensions of the presented work are possible, including modeling the queries as a mixture of proteins. Such approaches are likely to rely on sampling and greedy approximation strategies as explicitly modeling mixtures of thousands of proteins is computationally intractable. One such approach by Li et. al. [22] uses the Gibbs Sampling strategy to overcome the computational limitations. It might also be possible to reduce the search space of mixtures is by grouping proteins based on co-occurrences and modeling queries as mixture of such protein groups. We would like to explore such approaches in the future..Other important extensions of the presented work include addressing the issues caused by incorrect cleaving of protein sequences

[6] In the published version, this cell contains a 1. That is a typo.

into peptides, leveraging n-gram peptides in extended protein profiles, and applying supervised or semi-supervised classification and functional analysis to predicted proteins in different types of MS/MS data samples, e.g., cancerous vs. normal. Also, Nesvizhskii et al. have found that using Expectation Maximization (EM) as an additional step for finding hidden groups of proteins and for dealing with degenerate peptides can improve the performance of the prob-OR method. That suggests a potential way to further improve prob-AND and the other methods similarly by deploying the additional EM step, which is an interesting topic for future research.

ACKNOWLEDGMENTS

We thank the Mass Spectrometry Research Center at the Vanderbilt University for providing the Mark12 and Sigma49 query sets, and many useful suggestions on processing the data as well as insightful analysis on related literature. This work is supported in parts by the National Science Foundation (NSF) under grants EIA-0225656 and IIS-0704689. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17(Suppl 1):S13-21 (2001)

[2] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 20(9):1466-7 (2004)

[3] Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis, Vol. 20 Pages 3551-3567 (1999)

[4] Eng JK, McCormack AL, Yates III JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrum 5:976-989 (1994)

[5] Friedman, T., Razumovskaya, J., Verberkmoes, N., Hurst, G, Protopopescu, V. and Xu, Y. The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. J Bioinformatics and Computational Biology, Vol. 3. No.2, 455-476 (2005)

[6] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Analytical Chemistry, Vol. 74 Pages 5383-5392 (2002)

[7] Sadygov R, Yates III J. A Hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases, Anal Chem 75:3792-3798 (2003)

[8] Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. (2005) ProbIDtree: an automated software program capable of identifying multiple peptides from a

single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics. 2005 Nov 5(16):4096-106 (2005)

[9] http://www.lemurproject.org/indri/

[10] http://www.lemurproject.org/

[11] MacCoss MJ, Wu CC, Yates III JR. Probability-based validation of protein identifications using a modified SEQUEST algorithm, Analytical Chemistry, Vol. 74 Pages 5593-5599 (2002)

[12] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by Tandem mass spectrometry, Analytical Chemistry, Vol. 75 Pages 4646-4658 (2003)

[13] Moore RE, Young MK, Lee TD. QScore: An algorithm for evaluating SEQUEST database search results, Journal of the American Society for Mass Spectrometry, Vol. 13 No. 4 Pages 378-386 (2002).

[14] Eddes JS, Kapp EA, Frecklington DF, Connolly LM, Layton MJ, Moritz RL, Simpson RJ. CHOMPER: a bio-informatics tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies, Proteomics, Vol.2 No. 9 Pages 1097-1103 (2002)

[15] Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation induced microsomal proteins using isotope-coded affinity tags and mass spectrometry, Nature Biotechnology, Vol. 19 No. 10 Pages 946-951 (2001)

[16] Tabb DL, Hayes MacDonald W, Yates III JR, DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics, Journal of Proteome Research, Vol.1 No.1 Pages 21-26 (2002)

[17] Field HI, Fenyo D, Beavis RC. RADARS, a bio-informatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database, Proteomics, pages 36-47 (2002).

[18] Zhai C and Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of ACM SIGIR'2001, pages 334-342 (2001)

[19] Purvine S, Picone A F, Kolker E. Standard Mixtures for Proteome Studies. OMICS, Vol. 1 No. 1:79-92 (2004)

[20] Eugene A. Kapp, Frédéric Schütz, Lisa M. Connolly, John A. Chakel, Jose E. Meza, Christine A. Miller, David Fenyo, Jimmy K. Eng, Joshua N. Adkins, Gilbert S. Omenn, Richard J. Simpson. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. Proteomics, Vol 5, Issue 13, Pages 3475-3490

[21] Robertson Craig and Ronald C. Beavis. Tandem: Matching Proteins with mass spectra. Bioinformatics, 2004, 20, 1466-7.

[22]Yong Fuga Li, Randy J Arnold, Yixue Li, Predrag Radivojac, Quanhu Sheng, and Haixu Tang. A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. RECOMB 2008, LNBI 4955, pp. 167-180, 2008