Protein Identification as an Information Retrieval Problem

Yiming Yang, Subramaniam Ganapathy, Abhay Harpale
Carnegie Mellon University
Pittsburgh, PA - 15213
{yiming, ssganapa, aharpale}@cs.cmu.edu

ABSTRACT

We present the first interdisciplinary work on transforming a popular problem in proteomics, i.e. protein identification from tandem mass spectra, to an Information Retrieval (IR) problem. We present an empirical comparison of popular IR approaches, such as those available from Indri and Lemur toolkits on benchmark datasets, to representative popular baselines in the proteomics literature. Our experiments demonstrate statistically significant evidence that popular IR approaches outperform representative baseline approaches in proteomics.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous; I.5.2

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Interdisciplinary Investigation, Protein Identification, Tandem Mass Spectra

1 INTRODUCTION

Protein Identification from tandem mass spectra (MS/MS) is a crucial step for many important biomedical applications such as drug discovery and disease biomarker detection, i.e. detection of proteins present in the tissue of an unhealthy individual but usually absent in a healthy individual, for early diagnosis of life threatening conditions such as Cancer.

We formulate the protein identification problem in the form of a query-specific document retrieval problem in IR. Proteins consist of sequence of peptides, each of which is a sequence of amino acids. Drawing an analogy to a text document, we consider the protein as a document, while the constituent peptides are word-tokens, each consisting of amino acids similar to the character alphabet in textual documents. Due to our IR-based formulation, the rich body of research findings in text retrieval would provide meaningful insights into 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.

2 BACKGROUND AND RELATED WORK

The task of protein identification is to find a mapping from the thousands of observed MS/MS spectra, obtained by analyzing the input sample under a Mass Spectrometer, 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 on system-predicted peptides.

Popular peptide identification approaches perform database search by comparing the empirically observed mass spectra of unknown proteins to the theoretical mass spectra of known proteins. The theoretical spectrum of each peptide is derived based on existing knowledge about lower-level chemical properties of amino acid letters. SEQUEST [3] and XTandem! [1] are some of the popular peptide identification approaches that compare theoretical and empirical spectra.

Numerous approaches have been proposed in the proteomics literature for protein identification from the intermediate peptide identification step. The ProteinProphet system by Nesvizhskii et al [2] is a popular method that is commonly used in comparative evaluations on benchmark datasets. ProteinProphet estimates the probability of each protein as a probabilistic-OR function of the constituent peptides as $p_i = 1 - \prod_{j=1,..,J} (1 - q_j)$, where q_j and p_i are the probability of presence of peptide j and protein i in the input sample, and J is the total number of identified peptides. We will refer to this method as prob-OR in the rest of the paper. A recent work by Li et. al. [6] models the input sample as a multi-protein mixture and solves the Maximum-a-Posteriori (MAP) solution for the mixture weights.

3 PROTEIN IDENTIFICATION AS IR PROBLEM

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 [3] . We present the scored peptides using a normalized query vector $\vec{q} = (q_1, q_2, \cdots, q_J)$ such that $\sum_{j=1..J} q_j = 1$. Let $\vec{d}_i = (d_{i1}, d_{i2}, \cdots, d_{iJ})$ be a document vector representing a protein in the database and define within-document term weighting as

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

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07 ...$5.00.

The dot-product similarity in a standard Vector Space Model (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 scoring function is based on the cross entropy between the query and proteins, similar to the KL-divergence-based language model for document retrieval [4]. It is well known in IR research community that such models mimic a probabilistic-AND, i.e., only documents which contain all the query terms will receive positive weight. On the other hand, if we choose $d_{ij}=p_{ij}$ as the term weighting scheme, the dot-product similarity becomes:

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

This is a variant of probabilistic-OR, because a protein receives a positive weight if at least one of the constituent peptides is found in the query.

The connections from probabilistic-OR and probabilistic-AND to conventional VSMs invite 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.

4 EXPERIMENTS

4.1 Datasets

For evaluation and benchmarking of protein identification algorithms, we use standard proteomic mixtures whose MS/MS spectra are publicly available. Table 1 summarizes the datasets. The PPK [5] queries and corresponding protein database is publicly available. However, for the Mark12 and Sigma49 query sets, we created corresponding protein database by contaminating the relevant proteins with 50,000 proteins randomly sampled from the SwissProt protein repository.

Table 1. Dataset characteristics (prot: proteins, pep: peptides)

Data Set	Query Set			Protein Database		
	#spectra	#prot	#pep	#prot	#pep	#relevant proteins
PPK	2995	35	1596	4534	325,812	35
Mark 12	9380	12	1944	50012	5,149,302	12
Sigma 49	12498	49	4560	50049	2,571,642	49

4.2 Experimental Setup

For the representative IR approaches, we choose Indri-based retrieval, and 3 representative approaches from the Lemur toolkit: Cosine (Cosine similarity between query and document), KL (KL-Divergence between query and document), and Okapi retrieval. For evaluating a method on one dataset, we used the remaining two datasets as the validation sets for tuning parameters.

4.3 Results

Table 3 summarizes the performance of the various approaches in terms of Mean Average Precision (MAP) over the three query sets. The IR approaches clearly outperform domain-specific approaches over all the benchmark datasets. We also performed one-sample proportion tests for evaluating the statistical significance of these results and observed statistically significant evidence in favor of our analysis for p-value < 0.01

Table 2. Results summary in average precision

Dataset	Proteomics Approaches		IR Approaches			
	X! Tandem	prob-OR	Cosine	KL	Okapi	Indri
PPK	0.43	0.8	0.84	**0.85**	0.84	0.83
Mark12	0.41	0.66	**0.81**	0.79	0.73	0.76
Sigma49	0.241	0.44	**0.49**	0.48	0.45	0.48
MAP	0.36	0.63	**0.71**	**0.71**	0.67	0.69

5 CONCLUSION

In this paper, we presented 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. The results are highly encouraging: we obtained statistically significant performance improvements by using IR approaches over the representative domain-specific baseline methods. We hope this investigation provides useful information and insights for future research in adapting IR techniques to proteomic applications.

6 REFERENCES

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

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

[3] Moore RE, Young MK, Lee TD. QScore: An algorithm for evaluating SEQUEST database search results, Journal of American Society for Mass Spectrometry, Vol. 13 No. 4

[4] 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, (2001)

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

[6] 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