

# Implementing Spectral Similarity Algorithms for Protein Identification

Hussam Ala Eldeen<sup>1</sup>, David Angulo<sup>2</sup>, Kevin Drew<sup>3</sup>, Dominic Battre<sup>4</sup>, Alex Schilling<sup>5</sup>,  
Eric Puryear<sup>6</sup>, Jennifer Van Puymbrouck<sup>7</sup>, David Jabon<sup>8</sup>, Gregor von Laszewski<sup>9</sup>

<sup>[1]</sup> DePaul University, husamala@yahoo.com

<sup>[2]</sup> DePaul University, dangulo@cti.depaul.edu

<sup>[3]</sup> The University of Chicago, kdrew@uchicago.edu

<sup>[4]</sup> DePaul University, dominic@battre.de

<sup>[5]</sup> University of Illinois at Chicago, aschilli@uic.edu

<sup>[6]</sup> DePaul University, epuryear@gmail.com

<sup>[7]</sup> DePaul University, jvanpuy@gmail.com

<sup>[8]</sup> DePaul University, djabon@depaul.edu

<sup>[9]</sup> Argonne National Laboratory, gregor@mcs.anl.gov

## Abstract

*Database searching for protein identification is an efficient approach for MS/MS spectra identification compared to the existing approaches, such as protein identification with antibodies, and chemical degradation. In order for a database search to give reliable results, the database itself, as well as the searching algorithm, must be reliable. However, MS/MS spectra identification is still imperfect. The Illinois Bio-Grid Mass Spectrometry Database (IBG-MSD), an empirically derived curated and annotated database, along with multiple searching algorithms, addresses this issue and provides a better identification tool. The currently implemented algorithms are K-mutation algorithm, spectral contrast angle, and similarity index algorithm. The search engine allows users to search for post-translationally modified peptide using the K-mutation algorithm implemented. We provide a framework for high speed query processing and for efficient identification task. This new framework allows a community of users integratively to build a larger database and provide more efficient protein identification tools.*

**Keywords:** Proteomics, Mass Spectrometry, Curated Data, Empirical Data, Spectral Similarity.

## 1 Introduction

The ability to identify the amino acid sequence comprising a protein is crucial to bioinformatics. Identifying proteins is a prerequisite for studying other proteomics contexts such as interactions, localization, modification, and protein folding. With the increased popularity of Mass Spectrometry, as a powerful high-throughput tool for protein identification, very large amounts of experimental data are being produced. As a result, the demand for efficient proteins databases and powerful search tools has increased in order to satisfy the needs for analyzing the tremendous amounts of data the mass spectrometer produces.

## 2 Protein Identification Approaches

The natural role of proteins as the actual functional molecules of the cell, and their responsibility for most of the biochemical activity of the cell, motivated multiple approaches for protein identification and for studying their role in the cell. In this paper, we present the most salient protein identification approaches and review their benefits and weaknesses. One powerful approach for protein identification is the utilization of antibodies. However, this approach suffers some drawbacks when it is used on a proteomics scale. This is due to the need for a different antibody for each target protein [7]. Furthermore, this strategy can only be applied to organisms with complete genome sequences and abundant cDNA sequence resources [7].

Chemical degradation is another protein identification approach in which proteins are boiled in concentrated hydrochloric acid and, as a result, broken into their constituent amino acids. This approach gives the amino acid composition of the protein; however, it does not reveal the protein's sequence since all the peptide bonds in the protein are broken.

The third and the primary method for protein identification is mass spectrometry. A mass spectrometer is an instrument that measures the mass-to-charge ( $m/z$ ) ratio of charged ions in vacuum. Mass spectrometric sequencing is considered the leading tool for protein sequencing due to the high speed of spectrometric experiments and to the tremendous amount of data they generate. The high sensitivity of mass spectrometers enables us to characterize even ions in very low abundance. Furthermore, mass spectrometry has the preeminent advantage of its ability to distinguish post-translational modifications, which are modifications that take place when the primary structure of a protein is changed after it has been synthesized by mRNA translation.

### **3 General Motivation**

Given the spectra resulting from the MS/MS, the objective is to find out from which protein it is derived. Two technical ways are available for making use of the MS/MS data and identifying the proteins, *de novo* sequencing and database search.

#### **3.1 Searching Approaches**

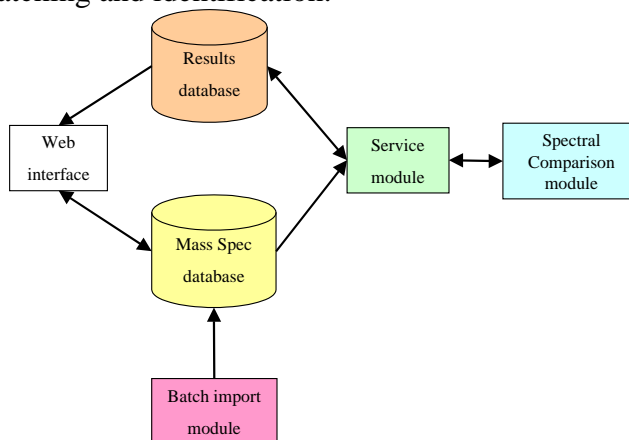
*De novo* sequencing analyzes a spectrum in order to determine its peptide sequence. This involves assigning ions to fragments of the peptide. *De novo* sequencing has the advantage of revealing novel proteins that do not exist yet in the protein databases. However, the nature of the process in which the sample goes through, including tryptic digestion and other factors, adds tremendous amount of noise to the sample. The MS/MS spectral analysis is therefore inappropriate for *de novo* sequencing since *de novo* sequencing requires high quality data for accurate results. The current state of the art for *de novo* sequencing algorithms at best yields inconsistent results for identifying real proteins.

On the other hand, database searches are concerned with comparing the main input, the experimental mass spectra of peptides, to the database spectra where the best fit provides the sequence of the experimental peptide. One of the unique advantages of our peptide identification search engine is that it searches a database of empirically derived spectra which will be described in more detail in a later section. In the context of database searching for protein identification, Mascot [5] and Sequest [1] are considered the leading approaches to MS/MS spectra analysis using database search approach. Mascot uses a probability based scoring method to rank the results of database searches. The score is based on the probability that the observed match of ions is a random event. Sequest uses a cross-correlation approach for the purpose of scoring

uninterpreted product ion spectra to database entries. For database searches, Mascot matches mass spectrometry data against theoretically predicted protein sequence data. Similarly, Sequest generates theoretical ion spectra by producing both mass and intensity data, whereas, our search engine searches empirically derived spectra. On the other hand, de novo [8] and Peaks [4] are the leading searches algorithms using de novo sequencing. De novo follows two stages in identifying proteins: In the first step, it looks for significant ions in the experimental spectra. Next, it creates a list of m/z values and probability of fragmentation at the m/z value. Peaks summarizes its approach in four steps: First, Peaks preprocesses the raw MS/MS data. The second step, computes candidate sequences. Step three re-evaluates each of the candidates of the second step. Finally, Peaks computes a confidence score for each of the top-scoring peptide sequences.

### 3.2 Searching an Empirically Derived Curated Database

IBG-MSD [2] serves as a central repository for empirically derived curated data. The different components of IBG-MSD and their relationships are illustrated in Figure 3.2.1. As mentioned above, searching a theoretical database has a negative impact on the performance accuracy of the protein identification search engine. We provide a tool that removes this negative impact by searching an empirically derived curated database that is capable of analyzing the high throughput mass spectrometry data as well as serving multiple simultaneous incoming queries for experimental spectra matching and identification.



**Figure 1: Components of the IBG-MSD**

Searching a database of valid data is a necessary requirement for obtaining reliable peptide identification. In order to have reliable protein identification, the data stored in the database goes through a curation process. When a user identifies a protein, the spectral data is sent by the IBG-MSD administrator to a curator for data validation. Submitted Mass Spectrum data goes through different stages in which it is curated into a standardized format. The data validation includes, for example, validation of MS instruments specifications such as, mass spectrometer manufacturer, spectrometer model, mass detector, etc. If the result of any of the validation tests is invalid, the submitter is notified and he or she is given the chance to put the data in the standardized format.

The reliability of the method used to generate a theoretical database, for the purpose of MS/MS spectra identification, has been an issue. Even having an ideal search algorithm will not give the best results in the case of defects in the way a theoretical database has been generated.

Searching a theoretical database does not take into account post-translationally modified peptides. About 200 types of covalent modifications of amino acids are known and almost all protein sequences are post-translationally modified [6]. Searching the protein database for exact matches, and disregarding any post-translational modifications has a dramatic effect on the identification process. In this case, the modification information available in the digest will be treated as some type of chemical noise rather than being counted for as part of the sequence. Searching an empirically derived database allows researchers to find accurate matches for their post-translational modified experimental spectra.

One of the characteristics of our identification tool is fast query processing. Comparing a submitted spectra against a dozen million records can require a huge amount of computing time. To increase time efficiency, the submitted query is divided into multiple subqueries, each of which is processed on a different node in a cluster. Furthermore, the database is split into fragments, each of which belongs to a different computational node. The separation allows multiple subqueries to access the database simultaneously, tremendously increasing performance.

### **3.3 Multiple Search Algorithms Available on Public Websites**

Multiple search algorithms have been developed for the purpose of comparing experimental spectra with database spectra. Spectral Contrast Angle treats a given spectrum as a multidimensional vector whose dimensions are the spectrum's peaks [9]. Comparing two spectra is represented here by measuring the  $\theta$  angle between two vectors in multidimensional space. The size of the  $\theta$  angle, which ranges between 0 to 90 degrees, indicates the similarity between the two compared spectra. The smaller the angle, the greater the similarity between the spectra is.

K-mutation algorithm, developed by Pevzner, et al. [6], is another algorithm tool implemented by IBG. Starting from just after the synthesis phase, until the mass spectrometry data is generated, proteins are exposed to multiple modifications that can make changes to the side chain and the main chain of amino acids. The K-mutation algorithm takes an experimental spectrum and searches the database for a spectrum that is the best match among spectra that are at most k mutations or modifications apart from the database spectra. The K-mutation algorithm uses dynamic programming for spectral alignment. By increasing the variable k, spectral alignment allows detecting more and more subtle similarities between compared spectra.

The third algorithm is the Similarity Index algorithm, developed by Wan, et al. This algorithm performs comparison in signal intensity by the smaller intensity for a set of peaks that fall within a given range of masses.

In order to facilitate the process of spectra submission and protein identification, a user-friendly interface is available. The web interface<sup>1</sup> currently allows submitting experimental spectra in DTA format and gives users the choice of running one of the three implemented algorithms for identification. A detailed report including the matching spectra, from the database, their corresponding accession numbers, and other relevant data is produced and returned to the user.

## **4 Future Work**

In the context of Mass Spectrometry, one of the present and future goals of Illinois Bio-Grid is building a community of users and submitters. A larger user group will contribute to the

addition of newly identified protein to the database, which in turn, will provide researchers with a better searching tool and allow a higher chance for identifying their experimental spectra. We are currently working on optimizing the K-Mutation algorithm to work in quadratic time instead of the current cubic running time. In addition, we are working on developing this algorithm to count for peaks' intensity.

## 5 Conclusion

Searching a database of empirically derived curated data increases the chances of identifying experimental spectra, provides a repository of valid data, and unlike searching theoretical databases, allows identifying post-translationally modified proteins. We hope to expand the user community of Mass Spectrometry to assist in providing a larger database and more efficient searching tools.

## Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 0353989, as well as, it was supported in part by the National Institute of Health (NIH) under Grant No. R01 HG3864 and done in conjunction with the Illinois Bio-Grid.

## References

- [1] Eng, Jimmy K.; McCormack, Ashley L.; Yates, John R. III. 1994. *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database*. J Am Soc Mass Spectrum 1994, 5, 976-989
- [2] Illinois Bio-Grid Mass Spectrometry Database. URL <http://illinoisbiogrid.org/MSDB>.
- [3] Kinter, M., and Sherman N. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. Wiley-Interscience: New York, 2000.
- [4] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby and Gilles Lajoie. *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry*, Rapid Commun. Mass Spectrom. 2003; 17: 2337–2342
- [5] Perkins, David N.; Pappin, Darryl J.; Creasy, David M.; Cottrell, John S. (1999) *Probability-based protein identification by searching sequence databases using mass Spectrometry data*. Electrophoresis. 1999, 20, 3551-3567
- [6] Pevzner, P.; DanSik, V.; Tangt, C.; *Mutation-Tolerant Protein Identification by Mass-Spectrometry*. Proceeding of the fourth annual international conference on Computational molecular biology. Pgs 231 – 236, 2000.
- [7] Twyman, R, M. *Principles of Proteomics*. Garland Sience/BIOS Scientific Publisher, 2004.
- [8] Taylor, Alex J.; Johnson, Richard S. (1997) *Sequence Database Searches via de Novo Peptide Sequencing by Tandem Mass Spectrometry*. Rapid Communications in Mass Spectrometry. Vol. 11. 1067-1075. 1997
- [9] Katty X. Wan, Ilan Vidavsky, and Michael L. Gross. *Comparing similar spectra: from similarity index to spectral contrast angle*. Journal of the American Society of Mass Spectrometry, 13:85–88, 2002.

---

<sup>i</sup> The link to the website where experimental spectra could be submitted for identification is <http://www.illinoisbiogrid.org/MSDB/>