

SpectraST: An Open-Source MS/MS Spectra-Matching Library Search Tool for Targeted Proteomics

Henry Lam¹, Eric Deutsch¹, James S. Edes¹, Jimmy Eng^{1,2}, Nichole King¹, Sara Yang³, Jeri Roth³, Lisa Kilpatrick³, Pedatur Neta³, Steve Stein³, Ruedi Aebersold^{1,4}

¹Institute for Systems Biology, Seattle, WA; ²Fred Hutchinson Cancer Research Center, Seattle, WA; ³National Institute of Standards and Technology, Gaithersburg, MD; ⁴Swiss Federal Institute of Technology, Zurich, Switzerland

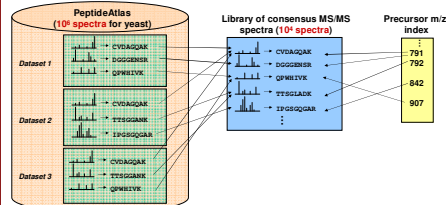
Overview

A notable inefficiency of traditional shotgun proteomics experiments by LC/MS/MS lies in the repeated rediscovery of the same identifiable peptides by sequence search methods. This suggests that a more targeted approach, in which previously observed and identified MS/MS spectra are catalogued and utilized for future identification, should greatly increase the efficiency and enable the study of more complex biological systems. In this approach, a library consisting of the characteristic MS/MS spectra of these observed peptides is first compiled from available shotgun proteomics data. A spectra-matching library search can then be performed to replace or complement time-consuming traditional sequence search. To that end, we have developed an open-source high-throughput MS/MS spectra-matching library search tool, SpectraST.

Introduction to Spectra Library Searching

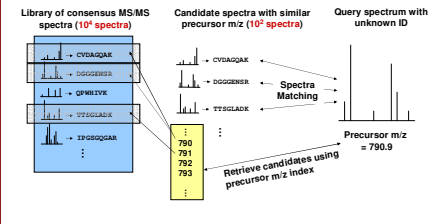
Creation of Consensus Spectra Library

- Make use of PeptideAtlas¹, a compendium of over 1 million MS/MS spectra
- Combine the high-confidence peptide IDs of 4 sequence database search engines: SEQUEST², Mascot³, X!Tandem⁴ and OMSSA⁵
- Group replicate spectra confidently identified to the same peptide
- Create consensus spectra (de-noising, peak averaging, spurious peak removal)
- Annotate and validate consensus spectra
- Build searchable library with indexes for fast retrieval



Spectra-matching library searching

- Given an unknown MS/MS spectrum, the consensus spectra library is screened for candidate spectra with matching precursor m/z values
- All candidate spectra are compared to the unknown for spectral similarity
- Identification is made when highly similar library spectrum is found



SpectraST

SpectraST is developed to build searchable library from consensus spectra and to search query spectra against it. It is:

- Written in C++ and compiled on a LINUX platform
- Open-source, high-throughput and highly resource-efficient
- Command-line executable, with no requirement of additional software
- Based on the spectral dot product as a similarity measure

$$Dot = \sum_{j=1}^n I_{query}(j) I_{library}(j)$$

where $I_{query}(j)$ and $I_{library}(j)$ are the normalized aggregate intensity of peaks within the j^{th} unit- m/z bin of the query spectrum and the corresponding one of the library spectrum respectively, after the spectra are pre-processed to de-emphasize dominant peaks and reduce noise.

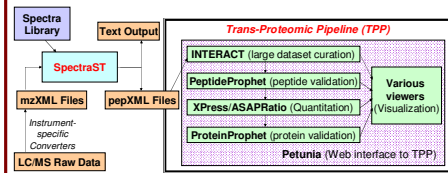
In addition to the dot product, a measure of the separation of the top hit from the runner-up, and a measure of how "biased" the dot product is towards a small number of dominant peaks, are also used in the discriminating function.

All spectra processing and scoring behaviors are customizable with either command-line options or simple extension of the highly-modularized source code.

Integration into the Trans-Proteomic Pipeline (TPP)

SpectraST is fully integrated into the Trans-Proteomic Pipeline (TPP) for standardized and user-friendly data processing, comparison, validation, visualization and storage.

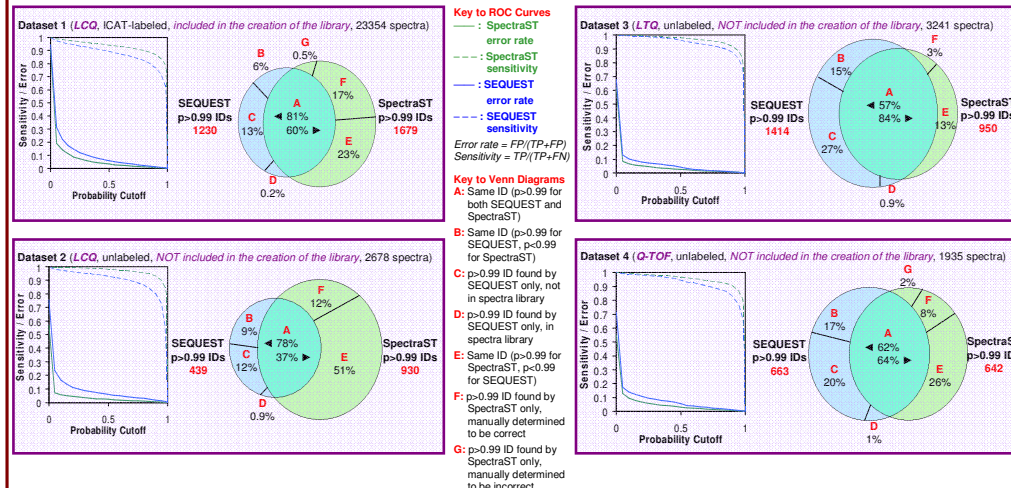
- Probability assignment and false-positive rate estimation (PeptideProphet), quantitation (XPress or ASAPRatio), and protein assignment (ProteinProphet) is done just in the same way as sequence search results.



Comparison to SEQUEST Identifications

The performance of SpectraST is compared against that of SEQUEST for 4 test datasets of yeast extract samples.

- PeptideProphet⁶ is employed to assign probabilities to the highest-scoring hits. A conservative probability cutoff of 0.99, above which a hit is considered positive, is used.



Key observations

- SpectraST is able to obtain better score separation between the good and bad hits, hence better separation between the sensitivity and error curves, than SEQUEST, in all cases.
- SpectraST is able to obtain considerably more highly-confident IDs than SEQUEST for the LCQ-generated datasets (Dataset 1 & 2). The extra hits that are missed by SEQUEST are manually verified to be correct (labeled F). False positives (labeled G) are extremely rare.
- SpectraST seldom missed an ID made by SEQUEST provided the peptide is in the spectra library (labeled D).
- SpectraST performs worse for LTQ and Q-TOF data (Datasets 3 & 4) than for LCQ data (Datasets 1 & 2), mainly because the spectra library is constructed predominantly from older LCQ data and lacks the lower-abundance peptides that can only be seen with better instruments. Improvement is expected with a more comprehensive spectra library.

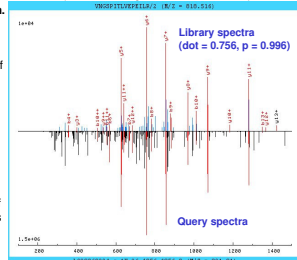
Advantages Over Traditional Sequence Search Engines

- This spectra-matching library searching approach to peptide identification is well suited for **targeted proteomics**, in which one seeks not to discover previously unseen peptides, but rather to find and quantify expected peptides of interest in the sample. However, as data repositories such as PeptideAtlas reach saturation, this approach will become generally useful for high-confidence, high-speed peptide identification.

It offers the following important advantages over traditional sequence (database) search methods:

- Smaller search space**
Only a fraction of the protein sequence database corresponds to observable and identifiable peptides. Most of the sequence search time is wasted on putative peptides that will never be observed by MS.
- Spectra matching is more precise**
Unlike sequence search, peak intensities are accounted for naturally based on direct experimental observation. Global similarity is used rather than selective peak matching. (See Figure to right) As a result, expensive similarity scoring algorithms (e.g. cross correlation) is not necessary to minimize false positives.
- Fast improvement in speed**
One spectrum only takes ~0.1 second to search (compared to ~10 seconds for SEQUEST), due to smaller search space and simpler similarity scoring.
- Library searching is an indirect way of searching multiple sequence search engines**
The spectra library is the union of all high-confidence IDs from 4 search engines, each yields a different set of positive IDs (70% overlap typical). Searching such a library amounts to searching 4 sequence search engines at the same time and combining the IDs.
- Higher confidence in identifications**
More precise spectra matching allows goods hits to be better separated from bad hits. That the peptide IDs have been made multiple times in the past also adds to the confidence.

An example of a confidently-matched query spectrum (as viewed through a viewer built into TPP)



Conclusions

- An open-source, high-throughput, and resource-efficient software tool, SpectraST, is developed to identify peptide MS/MS spectra by a spectra-matching library search approach, well-suited to targeted proteomics applications.
- SpectraST is fully integrated into the widely-used Trans-Proteomic Pipeline (TPP) software suite to facilitate user adaptation. Peptide probability assignment, quantitation, protein assignment, and visualization can be done in the same manner using the existing software tools within TPP.
- SpectraST offers comparable, if not superior, performance to SEQUEST at about 1000x speed.
- The identification rate and accuracy of SpectraST depends on quality and coverage of spectra library. A comprehensive, carefully compiled and validated spectra library is of crucial importance to this approach.

SpectraST will be available for download together with the Trans-Proteomic Pipeline suite of software at <http://tools.proteomecenter.org/>. Spectra libraries can be obtained at <http://www.peptideatlas.org/>. The source code for SpectraST will be available for download at <http://sashimi.sourceforge.net/>.

References

- Peptide Atlas: Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Edes J, Loevenich SN, Aebersold R. **The PeptideAtlas project**. *Nucleic Acids Research*, 34(Database issue): D655-658 (2006).
- SEQUEST: Eng JK, A.L.M. Yates JR. **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *Journal of the American Society for Mass Spectrometry*, 5(11): 976-989 (1994).
- Mascot: Perkins, DN, Pappin, DJ, Creasy, DM and Cottrell, JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis*, 20(18): 3551-3557 (1999).
- X!Tandem: Craig R, Beavis RC. **DATEM: matching proteins with mass spectra**. *Bioinformatics*, 20: 1468-1467 (2004).
- OMSSA: Geer LY, Markey SP, Kowalik JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *Journal of Proteome Research*, 3(5): 958-64 (2004).
- Trans-Proteomic Pipeline: Keller A, Eng J, Zhang N, Li X-J, Aebersold R: **A uniform proteomics MS/MS analysis platform utilizing open XML file formats**. *Molecular Systems Biology*, 1, 17 (2005).
- Targeted proteomics: Kuster B, Schirle M, Mallick P, Aebersold R: **Scoring proteomics with proteotypic peptide probes**. *Nature Review Molecular and Cell Biology*, 6(7): 577-583 (2005).
- PeptideProphet: 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*, 74(20): 5383-5392 (2002).

This work is supported by the National Heart, Lung and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179.