APPLICATION NOTE

SILVER Helps Assign Peptides to Tandem Mass Spectra Using Intensity-Based Scoring

Francis D. Gibbons
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA

Joshua E. Elias and Steven P. Gygi
Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, USA

Frederick P. Roth
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA

Tandem mass spectrometry is commonly used to identify peptides (and thereby proteins) that are present in complex mixtures. Peptide identification from tandem mass spectra is partially automated, but still requires human curation to resolve "borderline" peptide-spectrum matches (PSMs). SILVER is web-based software that assists manual curation of tandem mass spectra, using a recently developed intensity-based machine-learning approach to scoring PSMs, Elias et al. [4]. In this method, a large training set of peptide, fragment, and peak-intensity properties for both matched and mismatched PSMs was used to develop a score measuring consistency between each predicted fragment ion of a candidate peptide and its corresponding observed spectral peak intensity. The SILVER interface provides a visual representation of match quality between each candidate fragment ion and the observed spectrum, thereby expediting manual curation of tandem mass spectra. SILVER is available online at http://llama.med.harvard.edu/Software.html. (J Am Soc Mass Spectrom 2004, 15, 910–912) © 2004 American Society for Mass Spectrometry

Tandem mass spectrometry methods, such as liquid chromatography combined with tandem mass spectrometry (LC-MS/MS), are commonly used to identify the peptides (and therefore proteins) present in complex mixtures. Peptide identification is accomplished either by de novo prediction of peptide(s) that are consistent with the observed spectrum [1], or by comparing the observed spectrum with the spectra predicted for peptides in a genome-derived database. Software using the latter approach has been more commonly adopted, with SEQUEST [2] and Mascot [3] being popular examples. However, exhaustive peptide identification still requires human intervention to resolve "borderline" peptide-spectrum matches (PSMs). "Borderline" PSMs are those for which the match *might* be a good one, but established scoring criteria are unable to make a confident positive call. Manual curation to resolve these cases represents a bottleneck in high-throughput proteomics.

Published online May 10, 2004

Address reprint requests to Dr. F. P. Roth, Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, SGMB-322, 250 Longwood Ave., Boston, MA 02115, USA. E-mail: fritz_roth@hms.harvard.edu

Elias et al. [4] have developed a probability-based score which, when combined with scoring criteria used by either SEQUEST or Mascot, is superior to either method alone. At the core of this algorithm are probabilistic decision trees that estimate the probability distribution of peak intensity for a given fragment ion, conditioned on properties of the peptide and fragment ion. Two trees, trained respectively on correctly and incorrectly matched PSMs, are used to derive a log-odds score (LOD) measuring agreement between each fragment ion of a candidate peptide and the observed tandem mass spectrum. The LOD scores for fragment ions of a given candidate peptide can be summed to give an overall match score for that peptide. SILVER (for Spectrum Intensity Likelihood ViewER) provides this information in a visual format that assists manual spectrum curation.

Overview and Examples

Using SILVER is simple. It requires two input files to be uploaded by the standard CGI file-upload mechanism: (1) A short list of candidate peptides produced by the initial peptide identification software, and (2) the ob-

1044-0305/04/$30.00
doi:10.1016j.jasms.2004.02.011

Received December 17, 2003
Accepted February 6, 2004

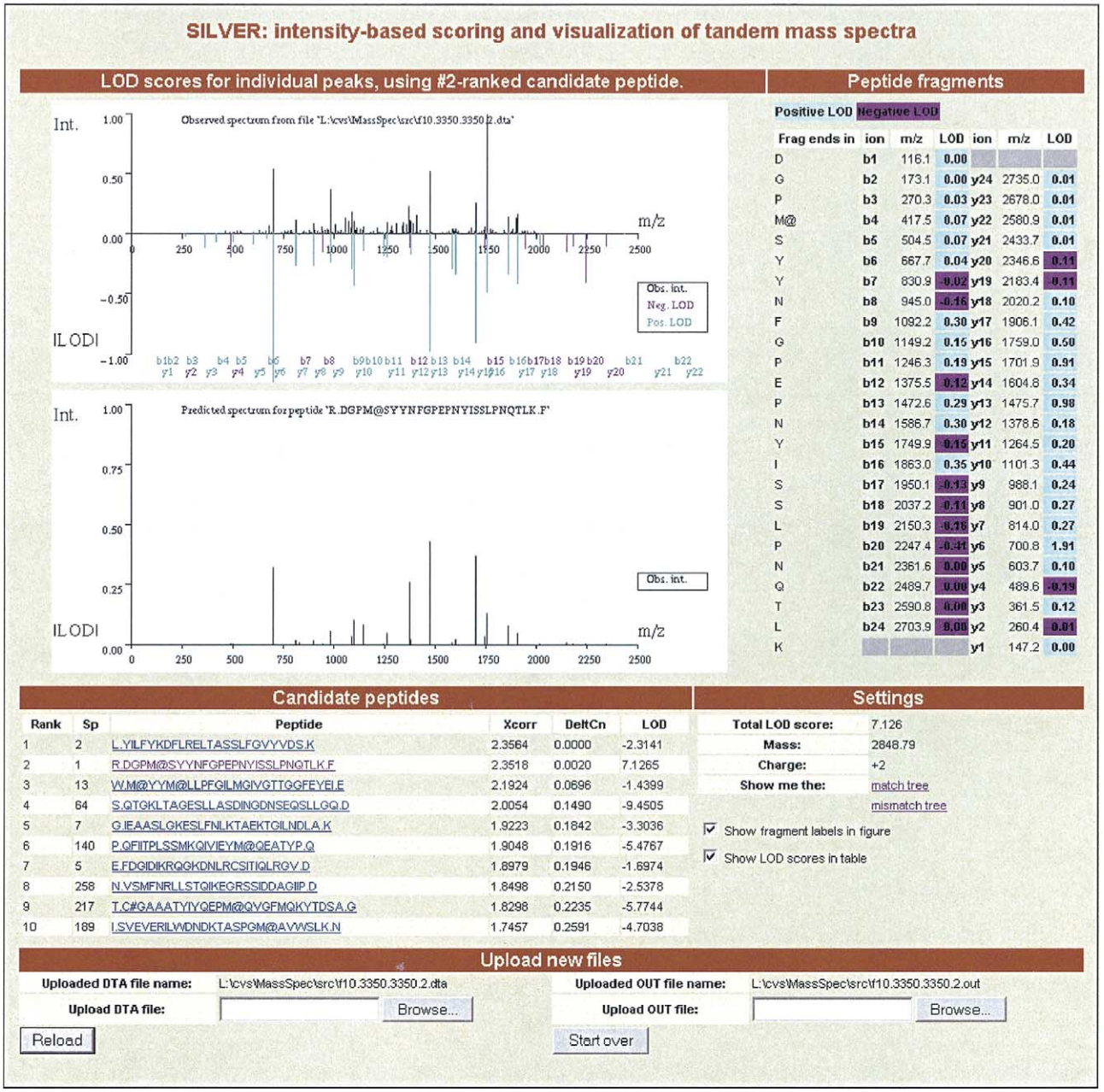

Figure 1. SILVER runs in any web browser, displaying a single page. This figure shows a candidate-peptide list produced by SEQUEST, however the format for the list is completely generic. These are borderline peptide-spectrum matches (PSMs), which cannot be distinguished by applying standard criteria (e.g., Xcorr > 2.0 and ΔCn > 0.08), and the intensity-based LOD score has been shown to be particularly effective at increasing confidence in such cases [4].

served tandem mass spectrum (e.g., the ".DTA" file, in the case of SEQUEST). The only requirement for the candidate peptide list is that each peptide be on a separate line, and that it be the first item on that line. SEQUEST produces .OUT files which satisfy this format. It is not difficult to obtain such information from MASCOT output, for which we supply scripts that the user can download and run.

SILVER's output consists of two figures and three tables, as shown in Figure 1. The uppermost figure shows the observed spectrum. One peptide at a time may be selected from the candidate list (the top peptide is chosen by default). For the selected peptide, a list of potential fragment ions (currently restricted to *b*- and *y*-type ions) is generated, together with predicted *m/z* values. These are shown under "Peptide fragments". For each fragment ion, a LOD score is calculated as described in [4] to measure compatibility of the predicted fragment ion based on the observed peak intensity at the appropriate *m/z* position. Positive scores (shaded cyan/blue) indicate that the observed intensity is more likely to arise if the candidate peptide were

correctly rather than incorrectly matched. The reverse is true for negative scores (shaded magenta). For visualization purposes, the negative absolute value of the LOD score is plotted using the same horizontal axis as the observed spectrum. Both positive and negative LOD scores are shown descending from the horizontal axis, with positive scores in cyan (blue), negatives in magenta. Longer blue lines indicate a better match; longer magenta lines indicate a poorer match. Optional labels, color-coded by LOD score, indicate with their left-hand edge the location of each potential fragment. The unified color scheme allows fast visual comprehension of the scoring structure. For comparison, the spectrum predicted using the expectation value of the intensity probability distribution at the appropriate leaf node of the "match" tree is shown below the observed spectrum, with the *m/z* axes aligned. The probabilistic decision trees currently used by our software to calculate LOD scores were trained using over 27,000 high-confidence spectra (over 800,000 fragment ions), filtered for peptide redundancy. These spectra were collected on ThermoElectron LCQ DECA and DECA XP ion-trap instruments, as described previously [4].

The lower part of the page contains two tables. Candidate peptides (one per line) are shown on the left, along with additional information derived from the peptide identification software initially employed. Peptides are ranked in the order in which they appear in the input file. Here, since the uploaded list was produced by SEQUEST, they are listed in decreasing order of that program's Xcorr score. The total LOD score for each candidate peptide is also shown. By clicking on any candidate peptide, the page is reloaded with that peptide as the default, with an updated fragment table and images. This feature makes it easy to move between candidates and compare scores. The "Settings" table allows some customization (e.g., fragment labels in the upper figure are off by default) and provides access to images of the match and mismatch trees used to compute the LOD score.

SILVER is implemented as a set of Python [5] classes, and runs on our web server as a CGI application. It loads a candidate peptide list and spectrum file, calculates fragment and peptide LOD scores, and generates a visual display in under two seconds. Reloading the page by clicking on a hyperlinked peptide is accomplished in less than one second. SILVER is available online at http://llama.med.harvard.edu/Software.html, and includes several examples illustrating its use.

Discussion

We are currently developing decision trees that make use of predicted fragment ions other than *b*- and *y*-type, and which account for neutral losses and protein modifications such as phosphorylation. These will be incorporated into SILVER. Also, SILVER currently assumes candidate peptides to have a +2 charge state, as these represent 70% of all peptides derived from a tryptic digest that result in tandem mass spectra on commonly used ion-trap instruments [6]. We plan to develop additional probabilistic decision trees to allow for candidate peptides in other charge states. The code is freely available to academic users upon request. It would be easy to add the capability to analyze other scoring methods, such as Havilio et al. [7], or Dancik et al. [8].

Acknowledgments

This work was supported in part by NIH HG000 41 (SPG), NIH NRSA 5T32CA86878 from the NCI (JEE), and by an institutional grant from the HHMI (FPR, FDG).

References

1. Pevzner, P.A. *Computational Molecular Biology, 1st ed.*; MIT Press: Boston, Massachusetts, 2000, pp 229–249.
2. Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
3. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999,** *20,* 3551–3567.
4. Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotech.* **2004,** *22,* 214–219.
5. Van Rossum, G.; Drake, F. L. *An Introduction to Python*. 2003: Network Theory Ltd: Bristol, UK.
6. Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003,** *2,* 43–50.
7. Havilio, M.; Haddad, Y.; Similansky, Z. Intensity-Based Statistical Scorer for Tandem Mass Spectrometry. *Anal. Chem.* **2003,** *75(3),* 435–444.
8. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De Novo Peptide Sequencing via Tandem Mass Spectrometry. *J. Comp. Bio.* **1999,** *6,* 327–342.