Spectral Library Generating Function for Assessing Spectrum-Spectrum Match Significance

Mingxun Wang[1,2] and Nuno Bandeira[1,2,3]

[1]University of California, San Diego, Dept. of Computer Science and Engineering, 9500 Gilman Dr., La Jolla, CA, 92093, USA
{miw023, bandeira}@ucsd.edu
[2]Center for Computational Mass Spectrometry, CSE, UCSD
[3]Skaggs School of Pharmacy and Pharm. Sci., UCSD

Tandem mass spectrometry (MS/MS) continues to be the technology of choice for high-throughput analysis of complex proteomics samples. While MS/MS spectra are commonly identified by matching against a database of known protein sequences, the complementary approach of spectral library searching [1, 2, 3] against collections of reference spectra consistently outperforms sequence-based searches by resulting in significantly more identified spectra. But despite this demonstrated superior sensitivity, the development of methods to determine the statistical significance of Spectrum-Spectrum Matches (SSMs) in peptide spectral library searches is still in its early stages.

The most common approach to controlling the False Discovery Rate (FDR) in both database search [4] and spectral library search [5] is the Target-Decoy approach where one extends the database/library of true peptides with a complement of sequences/spectra from 'random' peptides and uses matches to the latter to estimate the number of false matches to true sequences/spectra. But while these FDR approaches continue to be very valuable in correcting for multiple hypothesis testing in large-scale experiments, they provide little to no insight on the statistical significance of individual SSMs or Peptide Spectrum Matches (PSMs).

The estimation of the significance of SSMs is currently hindered by difficulties in finding an appropriate definition of 'random' SSMs to use as a null model when estimating the significance of true SSMs. We propose to avoid this problem by changing the null hypothesis – instead of determining the probability of a random match with a score $\geq T$, our approach determines the probability that a *true* match has a score $\leq T$. To this end, we explicitly model the variation in instrument measurements of MS/MS peak intensities (using a reference spectral library and a set of matching experimental MS/MS spectra) and show

how these models can be used to determine a theoretical distribution of SSM scores between reference and query spectra of the same molecule. While the proposed Spectral Library Generating Function (SLGF) approach can be used to calculate theoretical distributions for any additive SSM score (e.g., any dot product), we further show how it can be used to calculate the distribution of expected cosines between reference library and replicate query spectra. To assess the statistical significance of a score of a SSM between a reference library and unknown query spectrum, we used these SLGF calculated theoretical cosine score distributions to derive a p-value. In evaluating SLGF we explored both the accuracy of the theoretical distributions as well as SLGF's usefulness in the context of spectral library search. First we show that these expected cosine distributions did indeed approximate empirical score distributions and note that further work is necessary to enable more accurate theoretical distribution calculations. Second, using these p-values as scores, we demonstrate that these SLGF-based SSM p-value scores significantly outperform current state-of-the-art spectral library search tools such as SpectraST [1] in our test dataset. We additionally provide a detailed discussion of the multiple reasons behind the observed differences in the sets of identified MS/MS spectra.

Acknowledgements This work was supported by the National Institutes of Health grant 3-P41-GM103484 from the National Institute of General Medical Sciences.

References

[1] Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., Aebersold, R.: Development and validation of a spectral library searching method for peptide identification from ms/ms. Proteomics **7** (2007) 655–667

[2] Wang, J., Pérez-Santiago, J., Katz, J.E., Mallick, P., Bandeira, N.: Peptide identification from mixture tandem mass spectra. Mol Cell Proteomics **9** (2010) 1476–1485

[3] Dasari, S., Chambers, M.C., Martinez, M.A., Carpenter, K.L., Ham, A.J., Vega-Montoto, L.J., Tabb, D.L.: Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. J. Proteome Res. **11** (2012) 1686–1695

[4] Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods **4** (2007) 207–214

[5] Lam, H., Deutsch, E.W., Aebersold, R.: Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. J Proteome Res **9** (2010) 605–610