An alternative to the SEQUEST cross-correlation scoring algorithm for tandem mass spectral identification through database lookup: the Luck scoring function, and the probability of an unrelated spectra match model

Tema Fridman, [1] Jane Razumovskaya, [2] Nathan Verberkmoes, [3] Greg Hurst [4] and Ying Xu [5]

Keywords: tandem mass spectrometry, scoring function

1 Introduction.

Mass spectrometry represents a leading technology for examining the functional states of proteins in cells [1], [3]. In a typical LC/MS/MS experiment, a protein mixture of interest is digested into peptides which are separated, ionized and introduced into mass spectrometer. Selected peptide ions are isolated and fragmented through collision-induced dissociation (CID). The resultant fragments are then measured for their m/z and intensity values. The fragmentation pattern is then used to identify the parent peptide. The most practical and widely used technique is peptide identification (peptides are then matched to corresponding proteins) through database search (SEQUEST, Mascot software programs). In such methods, peptides from a sequence database are compared to the experimental data. Specifically, theoretical mass spectra are first generated for a set of candidate peptides, and then compared with the experimental spectra using a matching function.

SEQUEST [2] represents one of the most widely used and accurate programs for peptide identification via database searches. Developed several years ago, SEQUEST was not specifically designed for large-scale applications. One of the limiting factors in meeting the needs for genome-scale applications is its (lack of) computing speed.

Here we present a new method with high discriminating power for searching protein sequence databases for peptide identification. The accuracy compares favorably to the SEQUEST scoring function, with better separation between correct and incorrect matches. The algorithm also runs significantly faster than the SEQUEST program, by roughly about two orders of magnitude.

2 The Model.

First we introduce a *Luck* function, that has physical sence of quantitative measure of luck to obtain a certain outcome $wish_k$, given any unimodal probability distribution of possible outcomes:

$$Luck(wish_k) = Sign(wish_k - mode) \log \frac{P(mode)}{P(wish_k)}, \tag{1}$$

[1] ORNL, PO Box 2008 MS 6164, Oak Ridge, TN 37831-6164, USA. E-mail: tfa@ornl.gov

[2] ORNL, PO Box 2008 MS 6164, Oak Ridge, TN 37831-6164, USA. E-mail: rzw@ornl.gov

[3] ORNL, PO Box 2008 MS6131 Oak Ridge, TN 37831-6131, USA. E-mail: verberkmoesn@ornl.gov

[4] ORNL, PO Box 2008 MS6131 Oak Ridge, TN 37831-6131, USA. E-mail:hurstgb@ornl.gov

[5] Biochemistry and Molecular Biology Department, University of Georgia, Athens, GA 30602, USA. E-mail: xyn@bmb.uga.edu

with *mode* being the result that occurs most frequently, and $Sign(x)$ being the sign function (1 for positive x, 0 for $x = 0$, and -1 for negative x).

We developed a model derives the probability distribution of degree of match between a given experimental spectrum (produced by the peptide of interest, the *true* peptide) and a theoretical spectrum from a database peptide, assuming that the theoretical spectrum is produced by a different, unrelated peptide. The resulting probability is a function of the experimental spectrum density, the length of the interval on which the experimental peaks are distributed, the number of theoretical peaks, and their distribution pattern. The model does not take into account the intensity of peaks and the size of the database.

Based on the probability distribution above, we calculate the *Luck* of the match between each experimental-theoretical spectral pair, and use it as a scoring function for the match. As the experimental spectrum has consistently higher degree of match with the theoretical spectrum of the true peptide than with that of an unrelated peptide, we get a substantially higher *Luck* score for the correct match (i.e., you are exceptionally lucky to see that degree of match, if you assume that the process, which generates the match, is random).

3 The experiment and the results.

We have tested our algorithm on a data set of 3771 experimental spectra that resulted from performing an LC-MS/MS experiment on a protein mixture of eight purified proteins.Our peptide database consisted of all possible tryptic peptides generated from the eight target proteins (having 803 peptides), and from the 2873 *S. oneidensis* proteins (having 289,166 peptides). The latter was used as a distractor dataset.

Among 1053 spectra of parent charge one, we identified 526 correctly versus 532 found by Sequest. Among the remaining 2721 spectra of parent charge two and three, we identified 496 versus SEQUEST's 505 for parent charge two, and 221 versus SEQUEST's 227 for parent charge three. The spectra, that SEQUEST identified and *Luck* function did not, have very low Xcorr score. In terms of sensitivity – specificity analysis, *Luck* function performs better than SEQUEST due to greater separation between correct and incorrect identifications.

4 References and bibliography.

References

[1] Aebersold,R., Mann,M. 2003 Mass spectrometry-based proteomics. In: *Nature*, 422(6928):198–207. Review.

[2] Eng,J.K., McCormack,A.L., and Yates III,J.R. 1994 An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. In: *J Am Soc Mass Spectrom*, 5: 976–989.

[3] Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K., Yang,L., Wolting,C., Donaldson,I., Schandorff,S., Shewnarane,J., Vo,M., Taggart,J., Goudreault,M., Muskat,B., Alfarano,C., Dewar,D., Lin,Z., Michalickova,K., Willems,A.R., Sassi,H., Nielsen,P.A., Rasmussen,K.J., Andersen,J.R., Johansen,L.E., Hansen,L.H., Jespersen,H., Podtelejnikov,A., Nielsen,E., Crawford,J., Poulsen,V., Sorensen,B.D., Matthiesen,J., Hendrickson,R.C., Gleeson,F., Pawson,T., Moran,M.F., Durocher,D., Mann,M., Hogue,C.W., Figeys,D., Tyers,M. 2002 Systematic Identification of Protein Complexes in Saccharomyces cerevisiae by Mass Spectrometry. In: *Nature*, 415:180–183.