

Improving Computer Interpretation of Linear Ion Trap Proteomics Data Using Scaffold

Brian C. Searle and Mark Turner

Proteome Software Inc., 1336 SW Bertha Blvd, Portland, OR, 97219-2039, United States
E-mail: Brian.Searle@ProteomeSoftware.com, Telephone: (503) 244-6027, Fax: (503) 245-4910

Abstract

The new generation of linear ion traps, for example Thermo's LTQ, have a much shorter duty cycle than older 3-D ion traps, resulting in a five-fold increase in MS/MS scans. One side effect of faster scan rates is, that while a deeper analysis is possible, substantially more uninterpretable spectra are acquired and false positives become a serious concern. To emphasize this point, in control experiments we regularly see linear ion trap data sets with less than 5% of the MS/MS spectra correctly identified, as compared to between 20% and 30% with 3-D ion traps.

The Peptide Prophet algorithm, used by the MS/MS data visualization software Scaffold, is an excellent tool to handle false positives because it assigns probabilities to peptide identifications by comparing their scores to those of clearly false matches. One major hurdle with linear ion trap data is that the increase in uninterpretable spectra places an over emphasis on false matches, decreasing the accuracy of assigned probabilities.

We have extended this method in Scaffold to handle this problem in a novel manner. First, we derive a MS/MS spectrum quality filter specifically tailored to the sample. Using this quality filter we then remove over half of the uninterpretable spectra with no reduction in sensitivity. Finally, we employ the Peptide Prophet algorithm to calculate peptide probabilities using this reduced data set. The accuracy of probabilities is increased because a large portion of uninterpretable spectra has been removed. This reduction in data set size also results in faster computation and the ability to handle significantly more data.

LTQs provide data in overwhelming quantities

The approximately five- to ten-fold increase in MS/MS scans that the LTQ linear trap provides over traditional 3-D traps is simply astonishing. Scientists can now dig deeper into their samples both in terms of sensitivity and faster duty cycle. However, the increased sensitivity comes with a penalty: the LTQ can acquire much more uninterpretable data (Figure 1). Some of these uninterpretable data are rare modified forms of the proteins of interest, but an important majority of the unidentified MS/MS scans are triggered on noise. Two questions remain: first, how do you single out the rare but interesting forms, and second, how do you remove the noise?

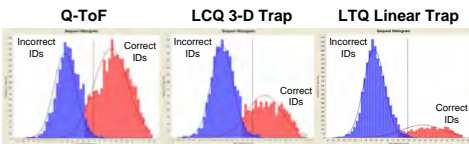


Figure 1: Discriminant score distributions for the same sample control protein sample run on three different MS/MS instruments. Spectra assigned by SEQUEST [1] to the control proteins are red, whereas spectra assigned to other "incorrect" proteins are blue. Despite the fact that LTQ linear traps acquire substantially more data than Q-ToFs and traditional 3-D traps, the linear traps suffer from significantly lower identification rates. This is represented by a lower "correct" distribution compared to the "incorrect" distribution. The dashed lines specify the 95% confidence level.

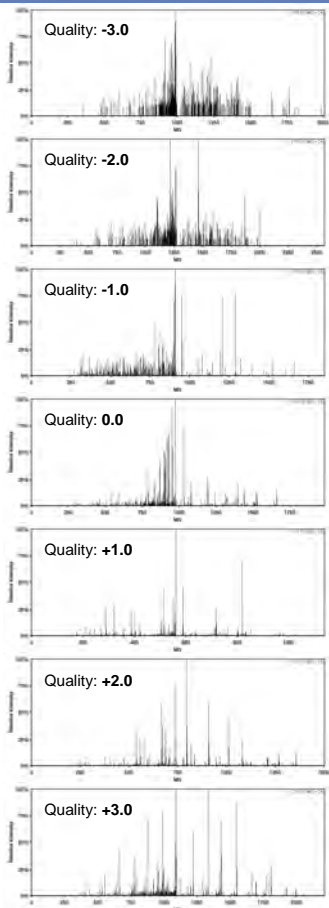
Assessing the Quality of LTQ Spectra

We developed a quality score to help simultaneously draw out interesting yet unidentified spectra, while improving statistical interpretation by removing noise. Our quality score is based on a similar score described by Nesvizhskii et al [2], which trains itself to find the most important spectral features specific to each sample using the database search engine results. Assuming that most identified spectra are of "good" quality, and most unidentified spectra are of "poor" quality, the algorithm uses linear discrimination to identify spectra that the search engine could have identified if it had been given broader criteria (more modifications or sequence substitutions). Most quality scores rely on self directed learning algorithms, such as neural networks or linear discrimination, and suffer from possible overfitting. We've found that using a few hand crafted quality features works better for generalization than a large number of somewhat redundant features. As such, we use the seven features described by Bern et al [3] with an added feature to single out spectra with only one peak. Using the best of both publications, our quality score is quite effective at differentiating between "poor" and "good" quality spectra (Figure 2 and 3).



Figure 2: (Above) The distribution of LTQ spectra across quality scores. Again, spectra assigned by SEQUEST to the control proteins are red, whereas spectra assigned to other "incorrect" proteins are blue. Spectra in green were assigned to "incorrect" proteins, but are assessed by the Quality Score to be of "good" quality. The distribution curves, "good" quality (red line) and "bad" quality (blue line), were assigned without knowing any information about the SEQUEST identifications. Green, unidentified spectra would be good candidates for modification searching and/or *de novo* sequencing. The black dashed line specifies the 20% filter used to remove poor quality spectra.

Figure 3: (Right) Representative +2H MS/MS spectra across a variety of quality score assessments. In an effort to represent the quality score accurately, all but the top spectrum (Quality: -3.0) were chosen because they were assigned to "correct" control proteins.



Using the Quality Score to Improve Probabilistic Interpretations

Peptide Prophet [4], the probabilistic algorithm behind Scaffold [5], was originally designed to work with LCQ 3-D trap results. Working from the data in Figure 2, we set out to use the quality score to remove poor quality spectra, thus improving the Peptide Prophet assessments of LTQ data by making the data distributions resemble LCQ results. Since the assessments are done after the database search engines identify peptides, any peptides in the red region of Figure 2 are automatically kept. We want to keep as much of the unidentified, good quality spectra as possible (the green region), so we chose to remove spectra below the 20% filter (the black dashed line). This filter removes 55% of the incorrectly assigned, poor quality spectra, which improves the score distributions presented to Peptide Prophet (Figure 4). These improved distributions enable Peptide Prophet to more accurately curve fit the distributions and thereby assigning more accurate peptide probabilities. It should be noted that our quality filtering does not affect the characteristics of the score distributions (other than their magnitudes), which means all assumptions about the data made by Peptide Prophet still hold true. Increasing the filter (cutting out more spectra) continues to improve Peptide Prophet at the expense of dropping possible modified peptides.

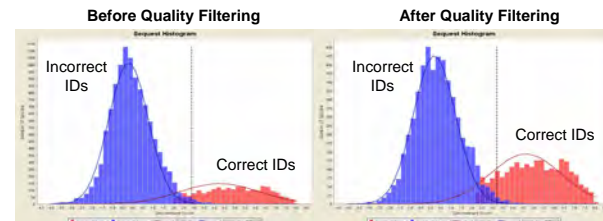


Figure 4: Discriminant score distributions for the SEQUEST interpretation of the LTQ dataset before and after quality filtering. Again, spectra assigned by SEQUEST to the control proteins are red, whereas spectra assigned to other "incorrect" proteins are blue. Because the determinations are made after the fact, 6071 incorrectly assigned spectra (55%) can be safely assigned as "too noisy to interpret" without removing any correctly assigned peptides. The dashed lines specify the 95% confidence level.

Great, but what about the high quality, unidentified spectra?

The removal of "poor" quality spectra allows us to hone in on "high" quality, but incorrectly assigned peptides (the green data) and take advantage of the LTQ's sensitivity. These peptides are perfect candidates for computationally intensive tasks, such as unanticipated modification searches and *de novo* sequencing.

Experimental Data Sets

The Q-ToF and LCQ samples of ten purified proteins of various molecular weights have been described previously [6]. The LTQ sample containing the same proteins was produced using the same methodology generating 13241 MS/MS spectra, which were interpreted with SEQUEST. Charge indeterminate spectra (+2H/+3H) were searched both ways with SEQUEST and the best scoring charge was selected.

Acknowledgements

We especially thank Larry David, Surendra Dasari, Phil Wilmarth and Srinivasa Nagalla for providing us with the data sets used in this study.

References

- Eng, J. K., et al. *J. Am. Soc. Mass Spectrom.* 1994, 5, 976-989.
- Nesvizhskii, A. I., et al. *Mol. Cell Proteomics* 2006, 5, 652-670.
- Bern, M., et al. *Bioinformatics* 2004, 20 Suppl. 1, i49-i54.
- Keller, A., et al. *Anal. Chem.* 2002, 74, 5383-5392.
- Searle, B. C., et al. "Scaffold: A Program to Probabilistically Combine Results...". *ABRF* 2005.
- Searle, B. C., et al. *J. Proteome Res.* 2005, 4, 546-554.