

Improving throughput and reliability of peptide identifications through spectrum quality evaluation

Kristian Flikka¹, Lennart Martens²,
Joel Vandekerckhove²,
Kris Gevaert² and Ingvar Eidhammer³

Keywords: protein identification, mass spectrometry, spectrum quality, mascot

1 Introduction.

Current high-throughput methods for tandem mass-spectrometry based protein identification often generate spectra that may be put into three categories regarding their quality: Clearly high-quality, clearly low-quality and uncertain quality. Clearly high-quality spectra will typically give high scores in database search programs like Mascot [2], or be eligible for high-scoring de novo sequencing. Clearly low-quality spectra will show obvious deficiencies like lack of MS/MS fragmentation, or originate from non-peptide contaminants. The third category contains spectra whose status cannot be easily decided, they may not appear to be identifiable, although this may be caused by unanticipated modifications, or atypical fragmentation of the peptide. Roughly, these three groups often are of equal size.

We have developed a versatile classifier that assesses the quality of a given spectrum, based on a variety of features, using machine-learning techniques. A Bayesian classifier was trained and tested on various proteome data sets from different mass-spectrometry instruments. The classifier can be biased towards different needs, by penalizing erroneous assignments of for example high-quality spectra.

2 Algorithm and Data.

A Bayesian classifier called AODE (Aggregating One-Dependence Estimators) [3] was trained and tested on various real life proteome data sets. The data was obtained by using the COFRADIC [1] technology together with different liquid chromatography (LC) electrospray ionization (ESI) and matrix assisted desorption ionization (MALDI) MS/MS instruments. These included ion-trap, Q-TOF and MALDI-TOF-TOF instruments.

In building training and test data sets, all spectra were submitted to the Mascot search engine, whereby a manually verified identification of a spectrum resulted in being assigned the label “good”. Other spectra were labeled “bad”. Numerous features thought to enable quality assessment were extracted from the spectra, both based on expert experiences and automatic evaluation of different suggested features.

¹Computational Biology Unit, BCCS, and Proteomics Unit (PROBE), University of Bergen, PB7800, N-5020 Bergen, Norway. E-mail: kristian.flikka@uib.no

²Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Department of Biochemistry, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. E-mail (Lennart Martens): lennart.martens@ugent.be

³Department of Informatics, University of Bergen, PB7800, N-5020 Bergen, Norway. E-mail: ingvar.eidhammer@ii.uib.no

3 Results

We were able to distinguish good spectra from bad as shown in Figure 1. The good spectra are the positives. For comparison, a classifier based on a decision tree-like method is included. For the ion-trap data, approximately 70% of the un-identified spectra (true negative rate) can be removed prior to identification, when also removing 5% of the spectra that otherwise would be identified (false positive rate). Classification of thousands of spectra is performed within seconds, as is re-building of the model based on new data sets.

The spectra that are labeled good, but are not identified with Mascot or similar tools, can be further analyzed to look for sequence tags, or unanticipated modifications. Using this methodology, we were able to identify additional proteins, not identified with the primary Mascot identification.

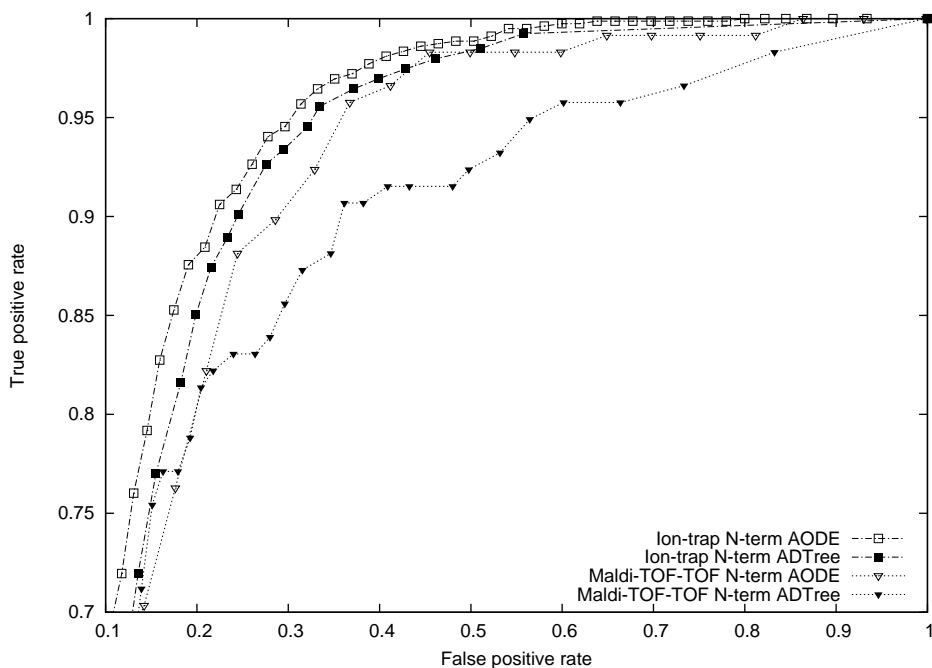


Figure 1: Test results shown as ROC curves for ADTree decision tree and AODE classifier

References

- [1] Gevaert, K. and Vandekerckhove, J. 2004. COFRADIC: the Hubble telescope of proteomics. *Drug Discovery Today: TARGETS*, 3, S16-S22
- [2] Perkins, A.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 18, 3551-3667.
- [3] Webb, G.I., Boughton, J.R. and Wang, Z. 2005. Not so naive Bayes: Aggregating one-dependence estimators. *Machine Learning*, 58, 5-24