MS Amanda: A New Scoring System for High Resolution MS/MS Spectra

Viktoria Dorfer°†, Peter Pichler*†, Stephan Winkler°, and Karl Mechtler*

° *University of Applied Sciences Upper Austria; Bioinformatics Research Group, Softwarepark 11, A-4232 Hagenberg, Austria*

* *Research Institute of Molecular Pathology; Protein Chemistry, Dr. Bohrgasse 3, A-1030 Vienna, Austria*

† co-first authors

Abstract

We have developed a new identification algorithm especially designed for high resolution and high accurate tandem mass spectra. In addition to the enormous speed of MS Amanda, which is on average 1.5 ms per spectrum, it is also very accurate, as we observe a high overlap of identified spectra with gold-standard algorithms Mascot and SEQUEST. Furthermore, MS Amanda is able to identify more spectra also when comparing various types of data sets. Soon a Proteome Discoverer Node and a standalone version will be available free of charge.

Problem Statement

Due to recent developments in mass spectrometry instruments including Higher-Energy Collisional Dissociation (HCD) [1], Electron Transfer Dissociation (ETD) [2] and high resolution mass spectrometers such as Orbitraps, the need for efficient and accurate identification algorithms arises. As a consequence, current gold-standard algorithms such as Mascot [3] and SEQUEST [4], which were developed more than a decade ago, might not optimally be suited for these nowadays available types of mass spectra. Changing the tolerated mass error of MS/MS spectra from low accurate (0.5 Da) to highly accurate (0.02 Da) does not have a significant effect on the achieved scores (see Figure 1). This, however, might be expected if a score is considered to be a measure of correctness of the identification (see Figure 2).

(a)

Sequence	Protein Group Accessions	IonScore	First Scan
GAGAVIHTHSK	sp_Q96GX9	32	1318
HFHETTPNK	sp_P08174	40	1349
KRPVIVHR	sp_P26639	34	1366

(b)

Sequence	Protein Group Accessions	IonScore	First Scan
GAGAVIHTHSK	sp_Q96GX9	31	1318
HFHETTPNK	sp_P08174	40	1349
KRPVIVHR	sp_P26639	34	1366

Figure 1: Comparison of a Mascot search with (a) an MS2 tolerance of 0.5 Da and (b) an MS2 tolerance of 0.02 Da. Mascot Ion Scores do not change significantly, although this might be expected.

(a)

Sequence	Protein Group Accessions	AmandaScore	First Scan
GAGAVIHTHSK	sp_Q96GX9	82.48	1318
HFHETTPNK	sp_P08174	137.00	1349
KRPVIVHR	sp_P26639	87.19	1366

(b)

Sequence	Protein Group Accessions	AmandaScore	First Scan
GAGAVIHTHSK	sp_Q96GX9	282.22	1318
HFHETTPNK	sp_P08174	302.26	1349
KRPVIVHR	sp_P26639	237.31	1366

Figure 2: Comparison of an MS Amanda search with (a) an MS2 tolerance of 0.5 Da and (b) an MS2 tolerance of 0.02 Da. MS Amanda scores do change significantly.

The MS Amanda Scoring System

We here propose a new identification algorithm, that is especially designed for tandem mass spectra with high resolution and high mass accuracy: MS Amanda. It is a statistical scoring function where the probability of the null hypothesis – that the peptide spectrum match occurred randomly – is estimated. In addition to the number of matched peaks also other spectra features such as the ion flow are considered. We are able not only to achieve a very high overlap with the results of a Mascot search (see Figure 3) but also to identify additional peptide spectrum matches at 1% FDR (see Figures 4-7) in comparison to Mascot and SEQUEST. Search queries with MS Amanda are very fast, on average 1.5 ms per spectrum; a standalone version and a node for Proteome Discoverer will soon be available free of charge.

References

[1] Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S., Mann, M.: Higher-energy C-trap dissociation for peptide modification analysis. Nature Methods 4, 709-712 (2007)

[2] Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J., Hunt, D.F.: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. PNAS 101, 9528-9533 (2004)

[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 20, 3551-67 (1999)

[4] 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. Journal of the American Society for Mass Spectrometry 5, 976-989 (1994)

[5] Köcher, T., Pichler, P., Swart, R., and Mechtler, K.: Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. Nature Protocols 7, 882-890 (2012)

Identification Results

We have tested MS Amanda on different data sets; we observe a good performance even when compared to the gold-standards Mascot and SEQUEST. All searches have been performed with the same values of mass error tolerances, post translational modifications and on the same forward-decoy database. Results have been trimmed manually to 1% FDR.

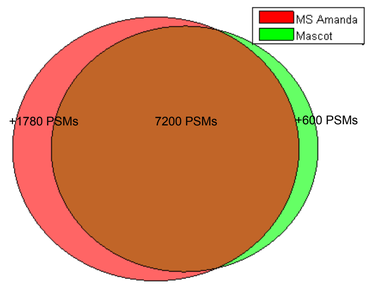

Figure 3: Example of overlap between MS Amanda (red) and Mascot (green) results at 1% FDR.

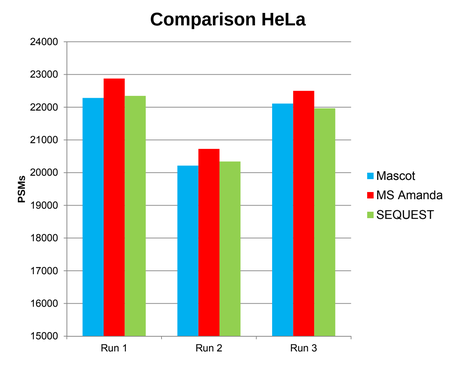

Figure 4: Results for a HeLa dataset at a 5h gradient [5], measured with a Thermo Scientific Orbitrap Velos

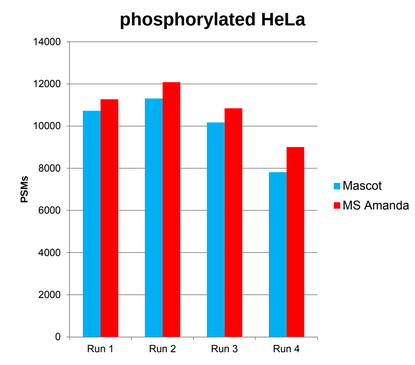

Figure 5: Results for a phosporylated HeLa dataset, measured with a Thermo Scientific Q Exactive

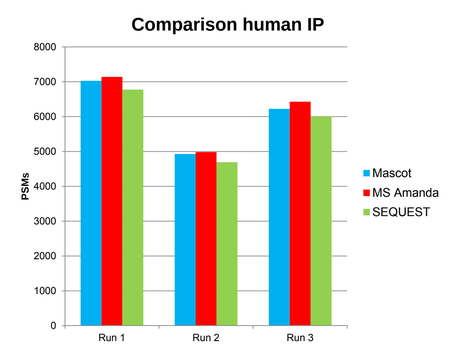

Figure 6: Results for a Human IP dataset, measured with a Thermo Scientific Q Exactive

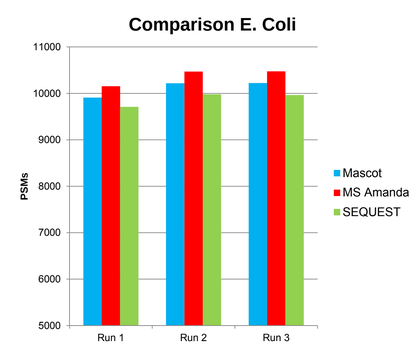

Figure 7: Results for an E.Coli dataset, measured with a Thermo Scientific Q Exactive

Correlation of Scores

Analyzing the correlation of the different scores provided by Mascot, SEQUEST and MSAmanda (see Figure 8) we can clearly see that all three schemes score differently even though there is a big result overlap. This independence suggests that a combination of the results of all three scoring schemes may further improve the result and the number of identified PSMs.

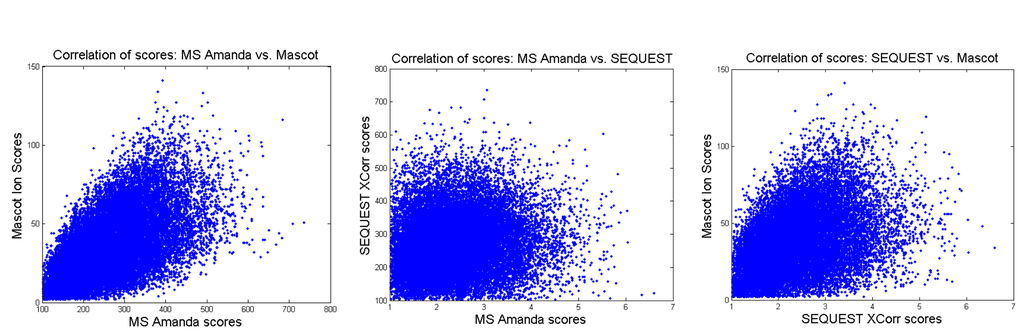

Figure 8: Pearson correlation plot of (a) Mascot Ion Scores vs. MSAmanda scores, (b) SEQUEST XCorr scores vs. MSAmanda scores, and (c) Mascot Ion Scores vs. SEQUEST XCorr scores. These plots indicate a high degree of independency of the scoring schemes, suggesting that a combination of scores may provide additional benefit.

Acknowledgements

This research was only possible through the support of the Protein Chemistry Lab at IMP and the Bioinformatics Research Group at FH OÖ, Hagenberg. Special thanks shall be given to Otto Hudecz and Richard Imre for testing the software, to Ines Steinmacher, Susanne Opravil, and Johannes Fuchs for measuring the data, and to Thomas Taus for his support in node development.