Reliable interpretation from reliable data for complex proteomic characterization is a challenge

Yann Verdier[1], Hidehiro Fukuyama[2], Anne-Maris Hesse[1], Iman Haddad[1], Chieko Makino[2], Jules Hoffmann[2], Jean Rossier[1], Joëlle Vinh[1]
1 ESPCI-ParisTech, Neurobiologie et Diversité Cellulaire, 10 rue Vauquelin, 75231 Paris, France
2 Centre National de La Recherche Scientifiqu, Institut de Biologie Moléculaire et Cellulaire, 15 rue Rene Descartes, 67084Strasbourg

Introduction

LC-MSMS is a method of choice for protein identification in a complex mixture. However, the peak capacity of mono-dimensional LC is too low for very complex mixtures. Beside multidimensional LC, the use of exclusion lists combined with successive analyses has been shown to be very efficient to overcome this limitation. Even with high performance MS such as Orbitrap, the interpretation of the data is a challenge to sort out false positive results.

Material and Methods

Material

3 samples :
Proteins from the IF3 cell line (*Drosophila melanogaster*)
Purified on agarose beads.
0,2 – 1mg protein / sample

LC-MS/MS

Tryptic digestion : directly on the sepharose beads; peptides elution
Liquid chromatography on inverse phase column (Pepmap C18, 75 µm I.D., 15 cm lenght, Dionex) with a flow rate of 220 nL/min.

Mass spectrometer hybrid linear ion trap /Orbitrap (LTQ Orbitrap, Thermofisher, San Jose, CA USA) with a nanospray ion source.
Automatic acquisition between MS and MS/MS
Orbitrap resolution 6000, between 500 and 2000 Da
Followed by 3 MS/MS scan (LTQ) on the 5 most intense peaks. Exclusion 90 sec of the fragmented precursors.

3 analysis / sample, followed by
3 analysis / sample with MS/MS exclusion of ≈200 peptids from the most abundant proteins (±10 ppm on whole LC analysis)

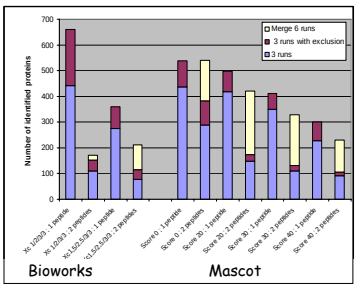
Databank screening

Databank 17.D melanogaster (16535 entries)
Mascot and Bioworks:
peptide mass ± 10 ppm
fragment masse ± 1 Da
variable modifications : carbamidomethyl C, oxidation HW, M, C ; phosphorylation Y
Miss cleavages : 2 ; Enzyme : trypsin
Proteins identified by 1 or 2 peptides
Proteins identified by the same subset of peptides

Mascot conditions: Score peptide >20, >30, >40 OR no filter. Search on Mudpit mode
Bioworks conditions: Xc vs Charge : 1.5, 2.5, 3.0, 3.0 OR 1, 2, 3, 3
For each sample :
6 independant analysis and one Merge analysis (merging the peptides from the 6 runs)

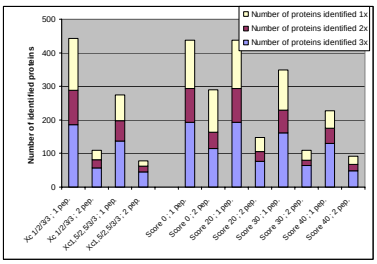
Results

1/ Contribution of the kind of analysis

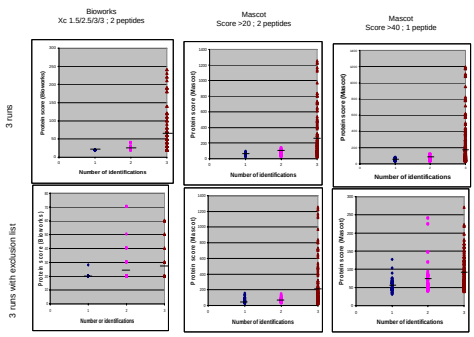


The use of an exclusion list allows the identification of proteins never found without exclusion. For the proteins identified by two peptides, the merge of the 6 runs allows the identification of new proteins, even with stringent criteria.

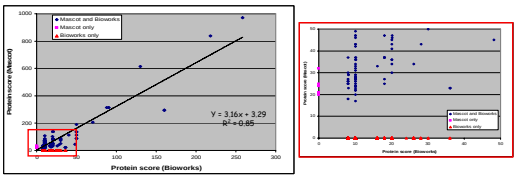
2/ Repetability



Repetability of the protein identification (3 runs without exclusion list). For each parameter used, the ratio of proteins found 3x vary between 1:3 and 1:2. In most of the cases, it is not possible to predict according to the protein score if the protein will be found 1x, 2x or 3x.

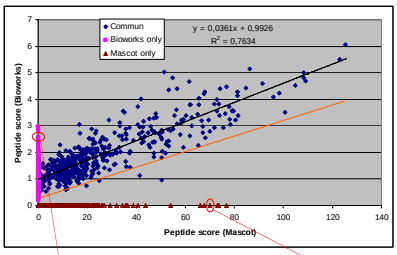


3/ Effect of the search engine on protein identification

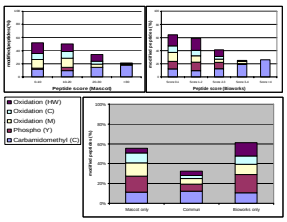


Proteins with a very high score are found by both sofware, however, for the other proteins, some identifications are specifics of Bioworks and some other specific of Mascot.

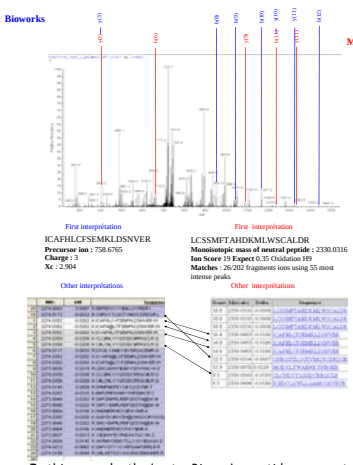
4/ Effect of the search engine on peptides identification



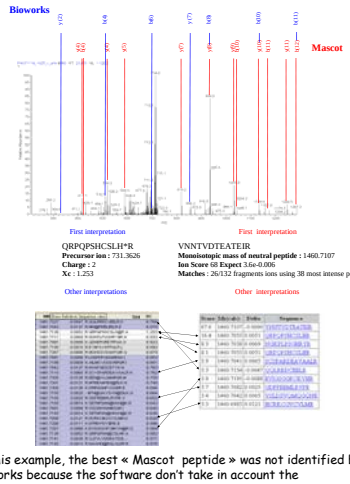
Mascot and Bioworks peptide score from a single run. A lot of low-score peptides are found only with one software. There is no simple rule to explain why some hight-score peptide are found only by one software. In black, linear regression for the peptide found by both sofware; in orange linear regression for all the peptides.



The ratio of modified peptides decreases with the peptide score in both software. This ratio is also correlated with the group of the peptide (found by both sofwares or only one)



In this example, the best « Bioworks peptide » was not found by Mascot.



In this example, the best « Mascot peptide » was not identified by Bioworks because the software don't take in account the processing of the N-terminal M (it should be noted that this peptide is found if we are looking for partial trypsic digestion)

Conclusions

Our study underlines the positive effect of an exclusion list for the LC-MS/MS protein identification in a complex mixture.
The requirement of triplicate analysis and combined search engines for reliable identification has been suggested. The present work gives an estimation of the proteins excluded if we consider only the proteins found in triplicate or by both software.
At the peptide level, validation criteria for databases search parameters should be tuned very accurately so that different search engines will give similar results.

The validation criteria should be considered as a wholein an experimental design, according to the possibility –or not -to use an orthogonal method to validate the protein.