Merging SpecOMS and X!Tandem identification results using machine learning algorithm in i2MassChroQ

Olivier Langella, Thierry Balliau, Marlène Davanture, Filippo Rusconi, Melisande Blein-Nicolas, Dominique Tessier

Merging SpecOMS and X!Tandem identification results using machine learning algorithm in i2MassChroQ

Olivier Langella*[1], Balliau Thierry[1], Marlène Davanture[1], Filippo Rusconi[1], Mélisande Blein-Nicolas[1], Dominique Tessier[2]

olivier.langella@universite-paris-saclay.fr

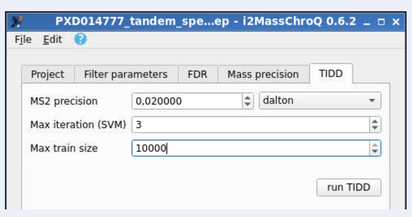
1 - IDEEV - Institut diversité, écologie et évolution du vivant
GQE-Le Moulon, PAPPSO
Université Paris-Saclay, INRAE, CNRS, AgroParisTech,
F-91190 Gif-sur-Yvette, France

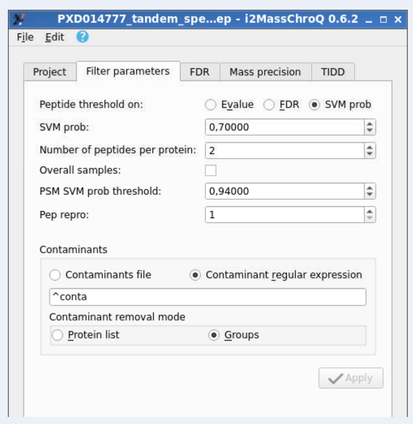
2 - INRAE, UR1268 BIA, F-44316 Nantes, France
INRAE, PROBE Research Infrastructure, BIBS Facility,
F-44316 Nantes, France

ABSTRACT

SpecOMS is an Open Modification Search method published originally in David et al. 2017. SpecOMS compares experimental spectra generated by a discovery proteomics experiment to a whole set of theoretical spectra deduced from a protein database in a few minutes on a standard workstation. The procedure yields identification results that might comprise peptides bearing unknown modifications. X!Tandem is a highly regarded classical identification engine, very efficient but unable to find peptide spectrum matches (PSMs) if the precursor ion's mass does not corresponds to the theoretical peptide's mass. Combining the results obtained by running both SpecOMS and X!Tandem reduces the proportion of undecided PSMs compared to a classical approach. However this combination approach raises a number of questions because the two engines do compute PSM scores very differently. We address this problem using a machine learning algorithm in i2MassChroQ.

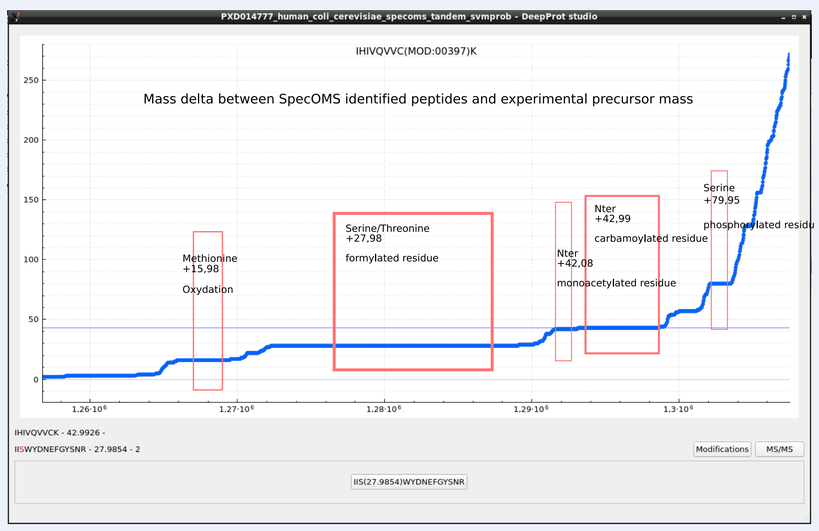
Rescoring PSMs with TIDD (Li et al. 2022)





1) Identification results are loaded by i2MassChroQ in a single project
2) i2MassChroQ rescores each PSMs with a machine learning procedure called TIDD
3) The new score is used to filter PSMs and proteins

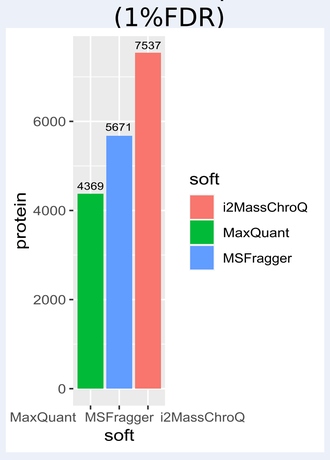
Peptide modifications browser



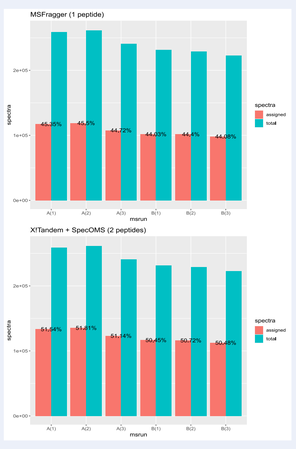
i2MassChroq offers a dedicated user interface to visualize PSMs bearing putative unknown modifications.

Benchmarking software with timsTOF dataset PXD014777

More proteins quantified in both samples (1%FDR)

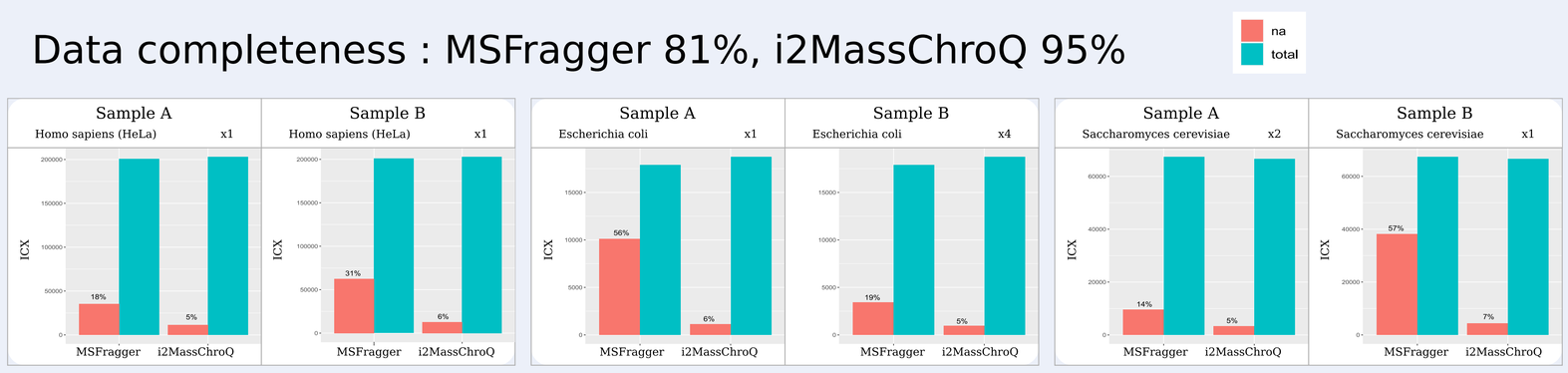


More spectra assigned per MSrun (1%FDR)



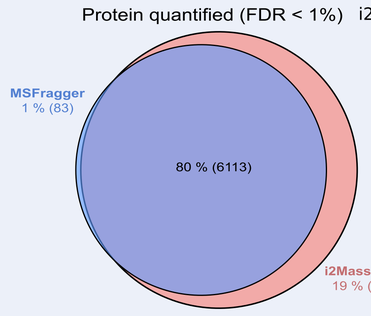
Two samples, A and B containing a mix of protein extracts from three different organisms. 3 replicates each. Samples A and B are 1:1 (*H. sapiens*), 2:1 (*S. cerevisiae*), 1:4 (*E. coli*)

Data completeness : MSFragger 81%, i2MassChroQ 95%

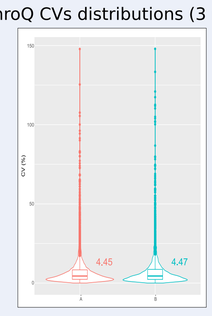


Quantification accuracy

Protein quantified (FDR < 1%)



i2MassChroQ CVs distributions (3 replicates)



i2MassChroQ quantifies more proteins on a wider dynamic range.
SpecOMS and X!Tandem results merged with TIDD yield more peptides per protein (safer identifications, 50% or more spectra assigned per MSrun).
The MassChroQ match-between-run algorithm, specifically designed for timsTOF data, lowers dramatically the proportion of missing values, especially in unfavourable contrasted conditions.

AVAILABILITY

i2MassChroQ source code and binary packages for Windows and Linux are available on the PAPPSO website (http://pappso.inrae.fr/bioinfo/i2masschroq/)
SpecOMS C++ identification engine source code or binary packages for Windows and Linux are available on demand to the PAPPSO facility or on the ForgeMIA project (https://forgemia.inra.fr/pappso/specoms).
The source code is licensed under the GNU General Public Licence v3+.

* presenting and corresponding author

this work is funded by ANR project ANR-18-CE45-0004