New strategy for bacterial species identification in Urinary Tract Infection using Artificial Intelligence on Ultrafast LC-MSMS-DIA runs
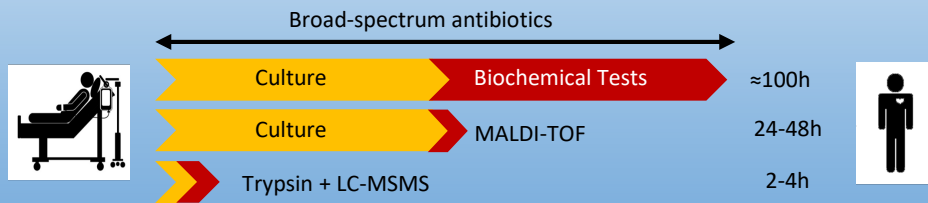
Florence Roux-Dalvai[1]; Mickaël Leclercq[1]; Elsa Claude[1]; Marion Narbeburu[1]; Tabiwang N. Arrey[2]; Nicolai Bache[3]; Clarisse Gotti[1]; Claire Dauly[2]; Dorte B. Bekker-Jensen[3]; David Bouyssié[4]; Maurice Boissinot[5]; Michel G. Bergeron[5]; Arnaud Droit[1]
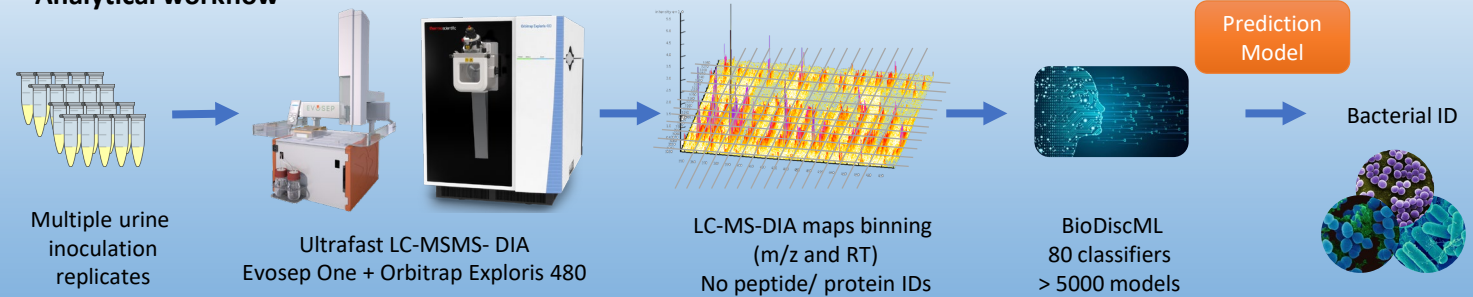
[1]Proteomics platform and Computational Biology Laboratory - CHU Québec Université Laval Research Center, Québec, QC; [2]Thermo Fisher Scientific, Bremen, Germany; [3]Evosep Biosystems, Odense, Denmark; [4]Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse, CNRS, Toulouse, France; [5]Infectiology Research Center - CHU Québec Université Laval Research Center, Québec, QC

Introduction

The standard pipelines to identify bacterial species in Urinary Tract Infections (UTI) require a long step of bacterial culture, during this time patients receive broad- spectrum antiobiotics known to increase bacterial resistance in the whole population. Our projet aims to use LC-MSMS analysis to avoid the bacterial culture step.

Broad-spectrum antibiotics

Culture	Biochemical Tests	≈100h
Culture	MALDI-TOF	24-48h
Trypsin + LC-MSMS		2-4h

Analytical workflow

Prediction Model

Multiple urine inoculation replicates → Ultrafast LC-MSMS- DIA Evosep One + Orbitrap Exploris 480 → LC-MS-DIA maps binning (m/z and RT) No peptide/ protein IDs → BioDiscML 80 classifiers > 5000 models → Bacterial ID

Short run length assessment

Test of 3 run lengths on 4 species found in 70% of UTI (*E.coli, E.faecalis, K.pneumoniae, S.agalactiae*), 285 samples.

Run length	21 min	11.5 min	5.6 min
Throughput (samples/day)	60	100	200
Blank discrimination	Ibk (5) MCC= 0.935	IB1 (2) MCC = 0.997	Pegasos (5) MCC = 0.889
Species discrimination	2 models (17) MCC = 0.975	NaiveBayes (10) MCC = 0.955	Kstar (5) MCC = 0.967
Final accuracy MCC*	0.955 +/- 0.026	0.976 +/- 0.016	0.928 +/- 0.082

MCC = Matthew Correlation Coefficient

Binning size

Various binning sizes in m/z and RT have been tested. Addition of RT dimension improves to final accuracy while keeping a reasonable computational effort.

Binning	m/z 0.1 Da	m/z 0.01 Da	m/z 0.1 Da RT 20 sec	m/z 0.1 Da RT 10 sec
Nb total features (computational effort)	950 000	9 500 000	1 800 000	2 700 000
Final accuracy (MCC)	0.798 +/- 0.086	0.928 +/- 0.082	0.977 +/- 0.027	0.971 +/- 0.003

Confusion matrices

Very good performance of the 2 parts of the prediction model (blank discrimination followed by species discrimination). The good prediction are found on the diagonal line.

Final MCC = 0.971 +/- 0.003

Binning: m/z 0.1 Da; RT 10 sec
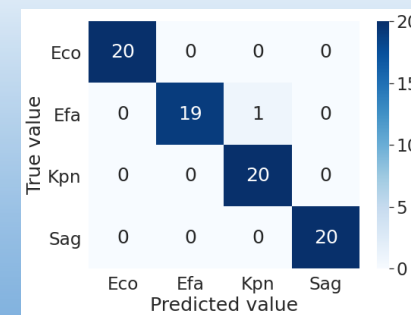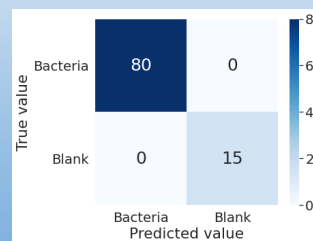
Conclusion

MALDI-TOF allows a high throughput of analyses but with a long turnaround time to get the bacterial ID. In our previously published work (2), we used machine learning on peptide IDs to reduces this time to 4h but for a maximum of 50 samples per day. Our new strategy performs extremely well on both sides to offer a fast bacterial identification compatible with the **high throughput of clinical laboratories**.

	MALDI-TOF	Peptide ID + ML	Raw data + ML
Total Time	24-48h	< 4h	<3h
Throughput	100s /day	< 50/day	**200/day**

References

(1) Leclercq et al., Front. Genet, 2019; (2)Roux-Dalvai et al., MCP, 2019