A NOVEL COMPUTATIONAL APPROACH APPLICABLE TO HUMAN MICROBIOME STUDIES – URINARY TRACT MICROBIOME EXAMPLE*

Authors: Antonio Starcevic*, Janko Diminic, Mario Cindric, Kristina Perica, Marina Cepnja,
 Ena Melvan, Daslav Hranueli, Jurica Zucko, *astar@pbf.hr

Parallel implementation of natural language processing based algorithm for proteomics mass spectrometry experiments. New approach to database searching in order to identify peptides whose sequences are deposited in a data-bases. Utilizes latent semantic indexing – a strictly mathematical approach independent of MS platform. Effectively handles large, sparse, ambiguous and contradictory data. Able to deliver correct results even with spectra produced by older generation tandem mass spectrometers. Solves problem of „gaps” caused by missing b/y ion peaks.

RESULTS & CONCLUSION

Metad sample containing only S. aureus and E. coli	Clinical urine sample (E. coli) > 10e6 CFU	Clinical urine sample (E. coli) > 10e6 CFU	Clinical urine sample 3 (P. aeruginosa) > 10e6 CFU
0 79163.0 Escherichia coli	0 219.0 Acinetobacter baumannii	0 58894.0 Escherichia coli	0 44.0 Pseudomonas
1 3901.0 Escherichia coli O154:H4	1 807.0 Escherichia coli	1 5080.0 Salmonella enterica subsp. enteric 15.0 Chlamydia trachomatis	
2 5524.0 Staphylococcus aureus	2 213.0 Lactobacillus kegelii	2 2854.0 Escherichia coli O154:H4	2 217.0 Pseudomonas aeruginosa
3 3556.0 Staphylococcus aureus subsp. aureus	3 200.0 Lactobacillus	3 3084.0 Salmonella enterica subsp. enteric 3 100.0 Escherichia coli	
	4 249.0 Enterobacteriaceae	4 2249.0 Salmonella enterica subsp. enteric 4 11.0 Bacillus	
	5 190.0 Staphylococcus aureus	5 1484.0 Salmonella enterica subsp. enteric 5 150.0 Escherichia coli	
	6 131.0 Bacillus cereus	6 1658.0 Klebsiella pneumoniae	6 5.0 Staphylococcus
	7 215.0 Acinetobacter	7 1375.0 Staphylococcus	
	8 149.0 Neisseria meningitidis	8 1339.0 Escherichia coli O157:H7	
	9 29.0 Mycobacterium abscessus	9 1348.0 Klebsiella pneumoniae subsp. pneumoniae	
	10 25.0 Lactobacillus	10 900.0 Salmonella enterica subsp. enterica serovar Typhimurium	
	11 6.0 Trichomonas vaginalis	11 514.0 Escherichia coli	
		12 555.0 Escherichia coli O157	

Method can be applied not only for database matching of peptides, but for matching entire species/strain proteomes obtained by MS/MS. This works for isolated strains as well as mixtures so called „microbiomes” or „metaproteomes”. This example shows results of analysis from one controlled experiment (two strains grown separately and mixed in same ratio) and of three urine samples taken from different hospitalized patients (some with catheter, some without). What differs from current approaches is that urine samples were analyzed directly without culturing. One sample was even frozen on minus 70°C.

Novel parallel LSI approach	Current state of art
<ul style="list-style-type: none"> Generic. Can be applied to any tandem mass spec. Easy to use, low CPU demands. Fast – vectorized representation of spectra in lower dimensionality. Massively parallel. No BLAST. LSI bag-of-words approach is an effective solution for fragmented time-series type of data with missing peaks. 	<ul style="list-style-type: none"> Platform dependant Computationally demanding Complicated workflow Linear relationship between search time and database size No satisfactory algorithm for short fragmented peptide sequences – rely on BLAST

One of the biggest obstacles of modern proteogenomics is the process of comparing large-scale MS/MS data with ever growing databases. Method here presented is offering a possible solution for this problem. Utilisation of natural language processing algorithms and alienation from compromise solutions taken from genomics (e.g. BLAST) allow us to be faster, simpler, more flexible and up-to date with constantly growing databases compared to any other solution as far as we know it. Further more, this method is completely generic, able to upgrade any existing tandem mass spec (MALDI or ESI) without significant costs related to hardware upgrade and personnel training.

REFERENCES:

- [1] M. HELMY, M. TOMITA, Y. ISHIHAMA, Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics, Genes, Genomes and Genomics 6 (2012): 76–85.
- [2] P. W. FOLTZ, Using Latent Semantic Indexing for information filtering, in Proceedings of the ACM Conference on Office Information Systems (COIS), 1990, pp. 40–47.
- [3] Federal Bureau of Investigation (J. Edgar Hoover), The Science of Fingerprints: Classification and Uses, Project Gutenberg EBook, 2006 (accessed 4/3/2012).

*This research is funded by HRZZ (Croatian Science Foundation) project “Clinical proteomics of microorganisms”