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

Authors: Antone Starcic*, Janko Domic, Mario Cindric, Kristina Perica, Marina Ceprijan, Ena Mihalic, Daslav Hranjec, Jurica Zucko. *astar@gib.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 database. 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**

Automatically generated and updated database for peptide matching. Simulates enzyme digestion followed by MS/MS ionization producing only b and y ions. Fragments are assigned to blocks which span instruments resolution shift errors (e.g. ±0.3–0.5 Da limit interval). Reducing matrix having peptides for columns and ion blocks for rows is being decomposed and a low rank approximation is found. This allows for fast and simple queries where cosine similarity serves as a measure of similarity between analyzed spectra and database match. This approach is extremely fast and parallel allowing thousands of simultaneous queries being performed on entire database in split seconds.

One of the biggest obstacles of modern proteomics 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 elimination of 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 (MALD or ESI) without significant costs related to hardware upgrade and personnel training.

**REFERENCES:**


*This research is funded by HRZZ (Croatian Science Foundation) project: "Clinical proteomics of microorganisms."