

Homology-driven Proteomics in Organisms with Unsequenced Genome by AUTOMATED LC-MS/MS de novo sequencing and MS Blast search

Patrice Waridel¹, V. Surendranath¹, H. Thomas¹, A. Frank², P. Pevzner² and A. Shevchenko¹

¹ Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany, ² Department of Computer Science and Engineering, UC San Diego, CA 92093-0114, USA
waridel@mpi-cbg.de

The characterization of proteomes of many important model organisms, especially within plant and insect kingdoms, is hampered by the paucity of genome sequences and remarkable phylogenetic diversity of proteins in wild-bred species. Here we report a strategy for automated identification of proteins from organisms with unknown genome by a combination of nanoLC-MS/MS, automated de novo sequencing and Mass Spectrometry driven BLAST (MS BLAST) sequence similarity search.

Tryptic digests of proteins separated by 1D or 2D electrophoresis are analyzed by a linear trap instrument LTQ (ThermoElectron) coupled to a nanoLC system (Dionex). The entire pool of 5,000 to 10'000 MS/MS spectra is filtered by a pattern recognition algorithm to remove spectra of trypsin and keratins peptides, as well as non-peptide background. The remaining spectra are interpreted de novo by PepNovo software, which takes about 0.5 sec per spectrum. The resulting redundant, degenerate and partially inaccurate sequence candidates are submitted to the web-accessible MS BLAST tool for protein identification. As protein identification relies on the similarity of peptide sequences (rather than on their identity), sequence polymorphism that commonly occurs in wild-bred species is tolerated.

The method was validated by automated analysis of proteins from the alga *Dunaliella salina*, the bug *Triatoma infestans* and the moth *Cerodirphia speciosa*, from which a number of proteins were identified by cross-species matching to known protein homologues from other model organisms.