

KN10 Ortho-proteogenomics: Multiple proteomes investigation through orthology and a new MS-based protocol

Sébastien Gallien¹, Emmanuel Perrodou², Christine Carapito¹, Caroline Deshayes², Jean-Marc Reyrat³, Alain Van Dorsselaer¹, Olivier Poch², Christine Schaeffer¹ and Odile Lecompte²

1 Laboratory of BioOrganic Mass Spectrometry, UMR 7178, Louis Pasteur University, Strasbourg, France.

2 Institut of Genetic and Molecular and Cellular Biology, Louis Pasteur University, Strasbourg, France

3 Faculty of Medecine, Rene Descarte University, Paris, France

proteomics, mass spectrometry, N-terminal labeling, comparative genomics
Proteomics will play a key role in the genome annotation process

The progress in sequencing technologies irrigates biology with an ever increasing number of genome sequences. The gene repertoire is predicted in silico and conceptually translated into proteins.

Predicted genes exhibit frequent errors, particularly in start codon.

A new “ortho-proteogenomic” approach is presented here for the annotation refinement of multiple genomes at once [1]. It combines comparative genomics with an original proteomic protocol that allows characterizing both N-terminal and internal peptides in a single experiment.

It is based on a TMPP derivatization of proteins before enzymatic digestion. The TMPP derivatized N-terminal peptides display in LC-MS/MS a 50 fold increase of the ESI response factor and a retention time shift toward the less complex part of the chromatogram allowing an easier identification.

This strategy was applied to the Mycobacterium genus with *M. smegmatis* as a reference and identified 946 distinct proteins including 443 characterized N-termini.

It allowed the correction of 19% of the characterized start codons, the identification of 29 proteins missed during the annotation process and the curation, thanks to comparative genomics, of 4328 sequences of 16 other Mycobacterium proteomes.

For classical proteomic studies, the TMPP derivatization is an easy way to increase the sequence coverage for the protein identification.

[1] Gallien et al. *Genome Res.* (2009).