## Poster G-20

**Enhanced protein identification and gene annotation from MS/MS data**


**Authors:**
Jens Allmer (*Institute for Plant Biochemistry and Biotechnology, Westfälische Wilhelms University, Münster, Germa*)
Bianca Naumann (*Institute for Plant Biochemistry and Biotechnology, Westfälische Wilhelms University, Münster, Germa*)
Monica Zhang (*Institute for Plant Biochemistry and Biotechnology, Westfälische Wilhelms University, Münster, Germa*)
Michael Hippler (*Institute for Plant Biochemistry and Biotechnology, Westfälische Wilhelms University, Münster, Germa*)

**Short Abstract:** A novel error tolerant computer algorithm is presented which enables high throughput proteomics data mining in genomic databases, using mass spectrometric data. This method successfully identifies peptides that are split by an intron when deduced from genomic DNA or originate from alternative splicing while adding confidence to the individual identifications.

**Long Abstract:**

Today, genomic sequences for many organisms such as human, mouse, and yeast are readily available. These sequences code for the proteome which may however be diversified by post translational modifications and alternative splicing of the encoded proteins. The aim of proteomics is to elucidate the proteome including functional and expressional components. Mass spectrometry is clearly the method of choice when identifying and sequencing peptides and proteins from complex mixtures (Mann and Jensen 2003). Cell extracts are usually purified by one or multiple methods before the separated proteins are proteolytically cleaved into peptides. These peptides are then submitted to mass spectrometry (MS) and tandem MS analysis. With the rapid evolution of mass spectrometers, which yields ever more and more precise data, it becomes evident that today the computational analysis presents the bottleneck in current proteomics research.

A new high throughput computational platform was established which combines de novo amino acid sequence predictions with subsequent error-tolerant database search and database search algorithms that correlate mass spectra with theoretical spectra derived from sequence databases. The GenomicPeptideFinder (GPF) searches de novo predictions in the six-frame translation of a genomic database in an error tolerant fashion (Allmer et al. 2004). The results are in-turn validated by a database search algorithm. Thus GPF interconnects the two approaches and enables detection of peptides which are split by an intron when deduced from genomic DNA or which originate from alternative splicing.

A fully automated computational pipeline with database backend was developed to join all necessary software tools. PEAKS (Ma et al. 2003) was used to perform de novo sequencing. Sequest (Eng et al. 1994) was used to correlate the mass spectra to databases. The GPF predictions, which present one result from the pipeline in form of a FASTA database, were afterwards validated using Sequest.

This platform, established as AutoMS, http://www.automs.de.ms, was used to analyze mass

spectra from a one dimensional (1D) sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE) of the thylakoid fraction of the unicellular green alga Chlamydomonas reinhardtii, which is an emerging model organism for proteomics research. As revealed by its genomic sequence, Chlamydomonas possesses a complex exon-intron structure thus making it especially suitable for genomic data mining in respect to gene annotation.

Enriched thylakoid membranes were isolated from crude cell extracts and separated via sucrose density centrifugation. Then 1D-SDS-PAGE was performed with the thylakoid fractions of the sucrose gradient. The bands were excised, digested in-gel with trypsin, and the resulting peptides were then submitted to mass spectrometry via nano-flow liquid chromatography.

From 200 1D-SDS-PAGE bands, 11735 distinct peptide sequences, which were generated by the proteomics platform and displayed a significant Sequest cross correlation coefficient, could be imported into the database. 698 of these peptides are putatively split by an intron on the “genomic level”, as detected by GPF. Furthermore, concerted action of PEAKS and GPF lead to the supporting identification of 334 peptides also identified with Sequest alone, thus adding confidence to the individual peptide identification.

For protein identification from the detected peptides two different strategies were employed. All proteins supported by two or more distinct peptides were considered confidently identified. Proteins that were identified by the combined effort of PEAKS, GPF, and Sequest, were also considered confidently identified if they were supported by only one supporting peptide. This was done because PEAKS and Sequest employ two distinct and complementing algorithms. Additional confidence arises through PEAKS predictions and GPF database search since identification via this strategy requires at least 5 correctly identified amino acids and a close match in the precursor mass. Taken together these parameters significantly raise the confidence in the identifications with respect to the size of the database.

This approach led to the identification of 206 distinct proteins. 77 of these proteins were identified with a single supporting peptide. The incorporation of the peptides that were suggested to be split by an intron when deduced from genomic DNA led to the identification of 27 gene models which are either incorrect or for which alternative splice variants may exist. 3569 (~30%) of the detected peptides could not be matched to existing gene models, therefore suggesting that the available gene models are far from being sufficient. This is inline with recent studies which assessed prediction accuracy of gene structure prediction programs (Reboul et al. 2003). It is said that about 20% of the gene models are predicted correctly (Flicek et al. 2003; Parra et al. 2003).

We propose, that the proteomics approach developed in the present study will facilitate gene annotation and will complement gene model prediction algorithms as well as EST data mapping. Since the predictions generated by this method are based on experimental data they may also aid in validation of existing gene models.

Allmer J, Markert C, Stauber EJ, Hippler M (2004) A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases. FEBS Lett 562(1-3): 202-206.

Eng J, McCormack AL, Yates JR, 3rd (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J Am Soc Mass Spectrom 5(11): 976-989.

Flicek P, Keibler E, Hu P, Korf I, Brent MR (2003) Leveraging the mouse genome for gene

prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res 13(1): 46-54.
Ma B, Zhang K, Hendrie C, Liang C, Li M et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17(20): 2337-2342.
Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21(3): 255-261.
Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW et al. (2003) Comparative gene prediction in human and mouse. Genome Res 13(1): 108-117.
Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M et al. (2003) C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet 34(1): 35-41.