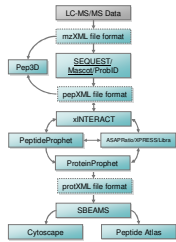


New Functionality for the Trans-Proteomic Pipeline: Tools for the Analysis of Proteomics Data

Luis Mendoza¹, Iliana Avila-Campillo¹, Eric Deutsch¹, James Eddes¹, Jimmy Eng², Tim Galitski¹, Robert Hubley¹, Andrew Keller³, Nichole King¹, Xiao-jun Li⁴, Parag Mallick⁵, Alexey Nesvizhskii⁶, Patrick Pedrioli⁷, Paul Shannon¹, David Shteynberg¹, Joshua Tasman¹, Julian Watts¹, Bernd Wollschlaedl⁷, Ning Zhang¹, and Ruedi Aebersold⁷

¹Institute for Systems Biology, Seattle, WA; ²Fred Hutchinson Cancer Research Center, Seattle, WA; ³Rosetta Biosoftware, Seattle, WA; ⁴Homestead Clinical, Seattle, WA; ⁵Department of Biochemistry, UCLA, Los Angeles, CA; ⁶Department of Pathology, University of Michigan, Ann Arbor, MI; ⁷Institute for Molecular Systems Biology (ETH), Zurich, Switzerland

OVERVIEW

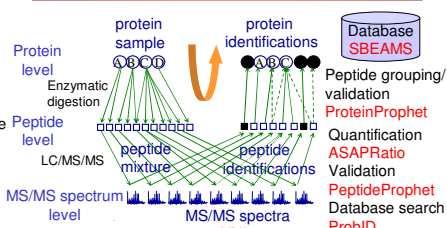


- We developed a data analysis pipeline that provides an automated, reliable, consistent, and objective analysis of high-throughput quantitative LC-MS/MS data
- The Trans-Proteomics Pipeline (TPP) includes software tools for MS data representation, MS data visualization, peptide identification and validation, protein identification, quantification, and annotation, data storage and mining, and biological inference
- We present an overview of the TPP and describe newly available functionality.
- All software tools are available under an open source software license at tools.proteomecenter.org

INTRODUCTION

Protein Identification by High Throughput MS/MS

High throughput LC-MS/MS is capable of simultaneously identifying and quantifying thousands of proteins in a complex sample. The consistent and objective analysis of the obtained large amounts of data is challenging and time-consuming. Over the past few years, we developed a data analysis pipeline that facilitates and standardizes such analysis.



DISCUSSION

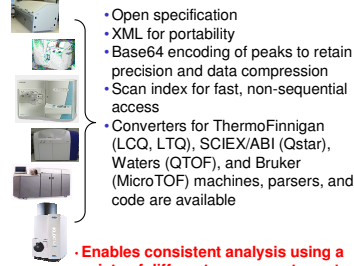
- Our software pipeline accepts as input raw LC-MS/MS data and outputs biological insight.
- It significantly lowers the hurdle of carrying out MS-based proteomics studies.
- All software tools are available under an open source software license at tools.proteomecenter.org.
- The TPP has been downloaded over 2,500 times via various distribution channels
- Free email support for the installation and operation of these tools is also available, as is a searchable knowledge base
- Over 230 members have joined the support and discussion email list
- 1,500+ messages have been posted on more than 400 threads
- Training on how to use these tools is provided three or four times a year at the Seattle Proteome Center. See www.proteomecenter.org/course.php.
- A free CD containing the source codes and the publications of these tools is being distributed at this meeting

REFERENCES

- mzXML:** Pedrioli *et al.*, *Nat. Biotechnol.* 22, 1459-1466, 2004.
Pep3D: Li *et al.*, *Anal. Chem.* 76, 3856-3860, 2004.
Probid: Zhang *et al.*, *Proteomics* 2, 1406-1412, 2002.
INTERACT & XPRESS: Han *et al.*, *Nat. Biotechnol.* 19, 946-951, 2001.
PeptideProphet: Keller *et al.*, *Anal. Chem.* 74, 5383-5392, 2002
ASAPRatio: Li *et al.*, *Anal. Chem.* 75, 6648-6657, 2003.
ProteinProphet: Nesvizhskii *et al.*, *Anal. Chem.* 75, 4646-4658, 2003.
SBEAMS: www.sbeams.org
Cytoscape: Shannon *et al.*, *Genome Res.* 13, 2498-2504, 2003.

COMMON FILE FORMAT

Different mass spectrometers output raw data in a variety of proprietary formats. **mzXML** is a common file format for mass spectrometry data:

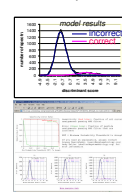


- Open specification
- XML for portability
- Base64 encoding of peaks to retain precision and data compression
- Scan index for fast, non-sequential access
- Converters for ThermoFinnigan (LCQ, LTQ), SCIEX/ABI (Qstar), Waters (QTOF), and Bruker (MicroTOF) machines, parsers, and code are available

- Enables consistent analysis using a variety of different mass spectrometers

PEPTIDE VALIDATION

PeptideProphet can be used as a 2nd step to compute probabilities that peptides assigned to MS/MS spectra are correct

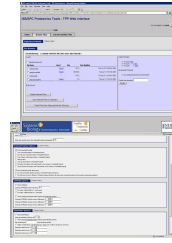


- Majority of peptide assignments by search engines are incorrect
- Manual validation is time-consuming, subjective and impossible to compare
- Applies statistical principles to automate peptide validation
- Validates peptide assignments by SEQUEST, COMET and Mascot

- Robust: learns distributions of search scores and peptide properties among correct and incorrect results
- Accurate: probabilities are true measures of confidence

GRAPHICAL USER INTERFACE

Petunia provides a consistent and easy-to-use, password protected, web accessible interface to the tools mentioned above, as well as others.

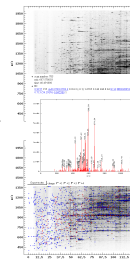


- Client/server architecture enables users to remotely run analyses and browse and manage files
- State-of-the-art command monitoring provides constant status updates
- Eliminates the need to learn the sometimes cryptic command-line arguments
- Parameter validation helps avoid many common user mistakes

- Facilitates the use of the Tools, and enables access to a network dedicated analysis server

VISUALIZATION

Pep3D displays LC-MS data in two-dimensional density images:



- Peptide abundance is represented by the darkness of its spot
- CID attempts and/or identified peptides can be highlighted
- Embedded links lead to CID spectra and peptide information
- Applications include sample evaluation, optimization of LC-MS/MS systems, and feature discovery

- Visualizes LC-MS data in easy-to-interpret images

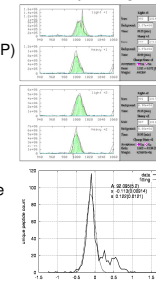
QUANTIFICATION

ASAPRatio and **XPRESS** calculate relative abundance of proteins labeled with heavy and light (2 channel) isotope tags:

- Evaluate peptide ratio from multiple charge states (ASAP)
- Apply statistical methods to evaluate protein ratios and standard deviations
- Quantify ICAT, SILAC, and many other samples

Libra performs quantification on MS/MS spectra that have multi-reagent labeled (4 channel) peptides, such as iTRAQ labeled samples.

- Compute protein ratios automatically and accurately



DATA STORAGE & MINING

SBEAMS-Proteomics stores large amounts of data generated from high throughput proteomics studies in an efficient manner:



- Database is part of the Systems Biology Experiment Analysis Management System Project
- Data products of the analysis pipeline are ingested into the database
- Data exploration, annotation, and correlation with other experiments can all be managed
- Interface allows flexible analysis of the data: analysis across multiple experiments

- Enables complex queries across multiple experiments and data types

SEQUENCE DATABASE SEARCH

Peptides are assigned to MS/MS spectra by evaluating eligible sequences in a database. New search algorithms can take advantage of a variety of mass spectrometer types:

- **SEQUEST & Mascot** – industry standard search engines
- **COMET** – database search engine running on compute cluster
- **Probid** – Bayesian approach to peptide scoring function
- **ProbidTree** – multiple peptide identifications from a single, co-fragmented spectrum

- Optimal peptide search results with a variety of mass spectrometers

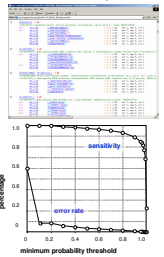


User interface
Web server

PROTEIN IDENTIFICATION

ProteinProphet takes as input a list of peptides and probabilities and infers the proteins in the sample:

- Groups peptides according to their corresponding protein
- Adjusts individual peptide probabilities to account for new protein grouping information
- Finds simplest list of proteins sufficient to explain all observed peptides (Occam's razor approach)
- Computes accurate protein probabilities
- Displays ASAPRatio results



- Allows meaningful comparison between results of different experiments

BIOLOGICAL INFERENCE

Cytoscape is a cross-platform software package for the integration, visualization, and algorithmic analysis of disparate large-scale biological data sets:



- Integrates LC-MS/MS data with other types of data such as mRNA expression and/or protein-protein interaction
- Analyzes all information in the context of biological networks
- Graphical approach to the extraction of biological insight
- Works in conjunction with the SBEAMS database
- Is available as an open source software at www.cytoscape.org

- Integrates proteomics data with other biological information to gain deep biological insight