Building a low-budget public resource for large-scale proteomic analysis

Anoop Mayampurath
Computation Institute
April 17, 2014
The world of proteomics

Proteins carry out most cellular activity, including control (regulation) of transcription, translation, and replication of DNA.
Alzheimer’s Disease

- Most common late-onset Alzheimer’s gene: APOE (e2, e3, and e4)
- Plaque accumulation but do not develop Alzheimer’s
- Repressor element-1 silencing transcription factor

ARTICLE

doi:10.1038/nature13163

REST and stress resistance in ageing and Alzheimer’s disease

Tao Lu¹, Liviu Aron¹, Joseph Zullo³, Ying Pan¹, Haeyoung Kim¹, Yiwen Chen², Tun-Hsiang Yang¹, Hyun-Min Kim¹, Derek Drake¹, X. Shirley Liu², David A. Bennett³, Monica P. Colaiácovo¹ & Bruce A. Yankner¹
A little background on mass spectrometry-proteomics
Proteins: $P_1, P_2, P_3$

Protease digestion

Peptides: $p_1, p_2, p_3, p_4, p_5, p_6$

Protein identification

Peptide identification

MS/MS

Separation

Fragmentation

Shen C et al. Bioinformatics 2008;24:202-208
LC-MS/MS clinical proteomic challenges

Biology

– Complexity of the proteome
– Dynamic range of plasma proteins
Baker et al. Genome Medicine 2012 4:63
LC-MS/MS clinical proteomic challenges

Computational
- lack of a framework capable of performing large-scale proteomic analysis
The workflow
Software List

platforms, pipelines and libraries

CPAS
LIMS and analysis tools for proteomics data (includes msinspect)

CPPP
Central Proteomics Facilities Pipeline [1] (demo here)

GenePattern
platform for integrative genomics and proteomics (includes PEPPeR [2] and other tools for proteomics)

InSilicoSpectra
open source proteomics library (of Perl functions) [3]

liibz
a fast implementation of box intersection for correspondence estimation in peak picking, alignment, etc.

Mass-up
utility with full GUI for proteomics data analysis, particularly MALDI-TOF

MASSyPup
a lightweight Linux Live distribution prepackaged with XTandem, mMass, mzmine, PepNovoUnIvo, PeptideShaker, mascovid, XCMS etc.

mspira
MS data processing in Ruby, including mzML reader/writer, in-silico digestion, isotopic pattern calculation etc. [4]

OpenMS
library for the analysis, reduction and visualization of LC-MS/MS data

PAPPSO
Plateforme d'Analyse Protéomique de Paris Sud-Ouest

Proteins
pipeline LIMS for proteomics experiments and analysis

Proteomatic
platform for creating MS/MS data analysis workflows using scripts [5]

ProtaWizard
open source library for proteomics tools development (supports mzML) [6]

pymzML
Python module to parse mzML data based on cElementTree [7]

Pyteomics
framework for proteomics data analysis, supporting mzML, MGF, pepXML and more [8]

QuPEx
integrated environment for storage, analysis and integration of proteomics data (requires login) [9]

QuProteomics
set of routines for analyzing proteomics data, an XML database to store the results and a user interface

TOPP
the OpenMS protein identification quantitation pipeline

TPP
Institute for Systems Biology "Trans-Proteomic Pipeline"

XCMS
software package (in R) for metabolite profiling from LC/MS data
About

OpenMS is an open-source software C++ library for LC/MS data management and analyses. It offers an infrastructure for the rapid development of mass spectrometry related software. OpenMS is free software available under the three clause BSD license and runs under Windows, MacOSX and Linux.

It comes with a vast variety of pre-built and ready-to-use tools for proteomics and metabolomics data analysis (TOPPTools) and powerful 2D and 3D visualization (TOPPView).

OpenMS offers analyses for various quantitation protocols, including label-free quantitation, SILAC, iTRAQ, SRM, SWATH, ....

It provides built-in algorithms for de-novo identification and database search, as well as adapters to other state-of-the art tools like X!Tandem, Mascot, OMSSA, etc.

It supports easy integration of OpenMS built tools into workflow engines like Knime, Galaxy, WS-Porado, and TOPPAS via the TOPPtools concept and a unified parameter handling (CTD).
Get Galaxy-P

The Minnesota Supercomputing Institute provides a public Galaxy-P server capable of limited analyses, testing, and demonstration. For heavy use you will likely need to install your own instance of Galaxy-P or spin up a Galaxy-P cluster on the cloud.

Public Server

**Advantages** Immediately accessible. Easiest option for publicly sharing data and pages.

**Limitations** Limited computational and disk resources. Potential problems associated with uploading protected or sensitive data to any public resource.

[usegalaxyp.org](http://usegalaxyp.org)

Install Your Own

**Advantages** Full control of computational resources. Easy to modify existing tools or add your own. Use our open source Galaxy-P tools targeting commercial applications that are not available on the public server. Right now these include ProteinPilot and Scaffold.

**Limitations** Because of its flexibility, Galaxy can be time-consuming to install and maintain. Galaxy-P adds more tools and can be configured to utilize remote Windows resources adding additional complexity.

[Install Instructions](#)

Take to the Cloud

**Advantages** BioCloudCentral provides an interface for creating a Galaxy-P cluster and CloudMan a likewise easy-to-use interface for creating a Galaxy-P cluster on the fly.

**Limitations** Currently no access to Windows such as MaxQuant or vendor-specific tools.

[Launch Now](#)

What is Galaxy-P?

*Galaxy-P* is a multiple ‘omics’ data analysis platform with particular emphasis on mass spectrometry data analysis.

News

For the latest Galaxy-P news, please follow us on Twitter.
proteomics.globusgenomics.org
An example: multiple myeloma

<table>
<thead>
<tr>
<th>Metric</th>
<th>One fraction (~800 MB)</th>
<th>All 22 fractions (~16GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall Time</td>
<td>19 minutes</td>
<td>24 minutes</td>
</tr>
<tr>
<td>Total CPU Time</td>
<td>19 minutes</td>
<td>418 minutes</td>
</tr>
<tr>
<td>AWS Nodes Used (m2.4xlarge)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>On-demand Cost ($1.64)</td>
<td>$1.64</td>
<td>$4.92</td>
</tr>
<tr>
<td>Spot instance Cost ($0.14 per hour)</td>
<td>$0.14</td>
<td>$0.42*</td>
</tr>
</tbody>
</table>

Current pricing for off-campus LC-MS/MS is as follows:
- University of Illinois at Chicago (UIC): Orbitrap-Velos ($100/sample run), ($75/hourly)
- Mayo Clinic Proteomics Resource Center (MPRC): Orbitrap-LTQ ($100/sample run)
- Northwestern University (NU) Proteomics Core: Orbitrap-Velos ($140/sample run)
- Northwestern University (NU) Proteomics Center of Excellence (PCE): Velos Orbitrap Elite with ETD, Velos-FT12T Ultra.
# Future work

<table>
<thead>
<tr>
<th>Functional category</th>
<th>Tools currently available</th>
<th>Tools to be incorporated in future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Tools</td>
<td>uf-mzML, MGF-DTA file converters</td>
<td>msconvert, Decon2LS (console version), DeconMSn, MultiAlign(command line), SuperHirn</td>
</tr>
<tr>
<td>Search Tools</td>
<td>X! Tandem, OMSSA, Mascot, MSGF+</td>
<td>SpectraST, Morpheus (command line), MyriMatch</td>
</tr>
<tr>
<td>Identification Tools</td>
<td>PeptideProphet, ProteinProphet, iProphet, Validator-MAX</td>
<td>MaxQuant, IDPicker3</td>
</tr>
<tr>
<td>Quantification Tools</td>
<td>Quantifier, XPRESS, Libra, ASAPRatio</td>
<td>MaxQuant, IDPQuantify</td>
</tr>
<tr>
<td>Other</td>
<td>PepXML/ProtXML to Table</td>
<td>Tools for: processing proteomics catalogs, combining outputs from different workflows, intelligent inclusion predictors.</td>
</tr>
</tbody>
</table>
Workflow integration

Label-free

Labeled

X!Tandem

$\uparrow$

SpectraST

$\uparrow$

Morpheus
Globus Proteomics catalog
Acknowledgements

• Sam Volchenboum
• Kolbrun Kristjansdottir
• Don Wolfgeher
• Alex Rodriguez
• Kyle Chard
• Ravi Madduri
• Paul Dave