Large-scale multi-omic data integration and analysis: challenges and opportunities

Biomedical Informatics and Computational Biology Research Symposium
January 17, 2014

Tim Griffin
tgriffin@umn.edu
Outline

- Historical perspective on multi-omics: yesterday and today
- Informatic challenges in multi-omics
- A solution: The Galaxy framework
- Galaxy in use
  - Proteogenomics
  - Metaproteomics
- Concluding thoughts
Acknowledgements

University of Minnesota
Ebbing de Jong
Joel Kooren
Sri Bandhakavi

University of Minnesota Supercomputing Institute
John Chilton (Penn State)
Ben Lynch
James Johnson
Getiria Onsongo
Bart Gottschalk

Dr. Joel Rudney

University of Wisconsin-Madison
Gloria Shenykman
Dr. Lloyd Smith

International collaborators
Dr. Ira Cooke (La Trobe University)
Dr. Lennart Martens (Ghent University)
Dr. Conrad Bessant (Queen Mary University of London)

Funding
NSF, NIH
Starting point: connecting the “-omes” of biology

- Integrating ‘omic data (i.e. multi-omic data) reveals new molecular connections and cause/effect relationships
Historical perspective: multi-omics circa 2002

- ICAT labeling for quantitative proteomics
- LCQ mass spectrometer
- DNA microarray containing ~6200 yeast ORFs
Flash-forward: New and improved ‘omics technologies

DNA Genome

RNA Transcriptome

Protein Proteome

Metabolite Metabolome

High-throughput sequencing

High resolution mass spectrometry
Technology example: MS-based proteomics

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)
Protein identification from MS data

Raw MS/MS spectrum

Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification
Realizing comprehensive/reproducible proteome?

Orbitrap Fusion mass spectrometer

Single LC-MS data acquisition in triplicate!

*Anal. Chem.* 2013, 85, 11710–11714
Converging technologies lead to new multi-omic possibilities

**Proteogenomics**

High throughput sequencing data, (genomic, RNA-seq)  
*annotation optional*

Comprehensive protein sampling by MS

- Genome annotation
- Gene expression regulation
- Protein variants in disease
- Functional outcomes of genome mutation

*Proteogenomics*

![Diagram of proteogenomics workflow]

- **Genomic sequence**
  - Confirmation of gene models
- **Identified peptides**
  - Correction of gene models
  - Identification of new ORFs

*L. BioSyst., 2011, 7, 284–291*
Converging technologies lead to new multi-omic possibilities

**Metaproteomics (aka Community Proteomics)**

- **Pre-processing**
  - Gene prediction and high-throughput sequencing
  - Filtering MS/MS data

- **Metaproteome analysis**
  - Protein database searching
  - Spectral libraries
  - De novo searching and homology search
  - Quantification methods

- **Post-processing**
  - Protein inference and taxonomic assignment
  - Functional and metabolic pathway analysis
  - Data storage and online data repositories

- Genomic sequences
- 16s RNA sequences

- Characterizes collection of proteins expressed by the community offering insight into conferred biochemical functions
• Tools needed to solve a multidimensional, integrated puzzle
Challenge: use and integration of disparate software

- Mastery of many different software
- Diverse hardware needs
- Compatibility of input/output data
- Handling large data files

Software for genomic/transcriptomic assembly

Software for dB assembly and peptide sequence matching

Software for integration of data and interpretation

A solution: The Galaxy Framework

• A web-based, community developed bioinformatics framework/platform/workbench

• Originally designed to address issues in genomic informatics including:
  • Software accessibility and usability
  • Analytical transparency
  • Reproducibility
  • Scalability
  • Share-ability: complete sharing of even complex workflows

• In a nutshell: Galaxy provides an open framework into which disparate software programs can be deployed, integrated into customized workflows for typical to advanced applications, which can be shared in their entirety with other users

A (free) supermarket for ‘omics software?
Extending Galaxy for multi-omics: GalaxyP

PROTEOMICS: GALAXY-P
- Peak List Processing
- ProteinPilot
- Statistical Validation
- Quantitation
- MaxQuant
- Bumbershoot Tools
- Metaproteomics
- Proteogenomics
- Get Data

PROTEOMICS: COMMUNITY
- Peptide Shaker
- mzMatch
- OpenMS
- ProTK
- Utilities
- Visualization
- FASTA Manipulation
- Adapt

BIOINFORMATICS

Data Conversion
- msconvert
- MGF formatter
- OpenMS tools

Raw MS data → Converted data (mzML)

Database search
- XITandem
- OMSSA
- MyriMatch
- MaxQuant
- ProteinPilot
- Target-decoy dB generator

Peptide sequence matches

Data filtering/organization
- Peptide/Protein
  - Prophet (TPP tools)
- FDR calc (e.g. Mayu)
- Scaffold

Inferred proteins

Quantification
- iQuant (isobaric labeling)
- SILACAnalyzer

Galaxy-P 101: Building up and using a proteomics workflow

Galaxy-P 201: Building up and using a proteomics workflow

University of Minnesota
Driven to Discover™
Example application: proteogenomics

- Database Generation*
- Peaklist generation
- Database search
- First-step
- Two-step
- Identifying peptides from translated nucleotide db
- Automated BLAST-P search*
- Peptide-Spectral-Match Evaluation
- Genomic context analysis*
Proteogenomics: protein database generation

Galaxy genomic software

Gloria Sheynkman
UW-Madison

Mol Cell Proteomics (2013) 12, 2341-53
Assessing novelty: automated BLAST-P processing

- Automatic searching of thousands of peptides against BLAST-P using criteria for small peptides (8-30 aa) and large (> 30 aa); flexible to different stringencies for “novel” sequences
Visualizing novel peptide hits

- IGV compatible: Peptide-to-genome viewer
Putting it all together

- 150 step workflow using diverse software, integrated and automated in Galaxy
Metaproteomic workflow

1. **Translation**
   - Translation of host protein database
   - Translation of microbial protein database

2. **Database Merge**
   - Merge host protein database with microbial protein database

3. **Host Protein Database**
   - Host protein database

4. **Microbial Protein Database**
   - Microbial protein database

5. **Target Sequence DB**
   - Target sequence database

6. **Fractions of Mass Spectra**
   - Fractions of mass spectra

7. **Spectrum Matching**
   - Spectrum matching

8. **Long Peptides (>30 aas)**
   - Long peptides (>30 amino acids)

9. **Short Peptides (<30 aas)**
   - Short peptides (<30 amino acids)

10. **Filter Peptides**
    - Filtered peptides

11. **Micropeptides**
    - Micropeptides

12. **Data Processing**
    - Data processing

13. **BLAST Output**
    - BLAST output

14. **NO BLAST = remote BLAST**
    - NO BLAST = remote BLAST

15. **BLAST-P Analysis**
    - BLAST-P analysis

16. **MEGAN**
    - Submit for MEGAN analysis

17. **Submit to UniPept**
    - Submit to UniPept


---

**University of Minnesota**

**Driven to Discover**
Increasing microbial peptide identifications

- Addressing the large database challenge: 2-step database searching

**STEP 1**

- Search high mass accuracy peaklists and associated MS/MS against “target” version of human + HOMD or human + translated RNA sequences (~0.8 \( \times 10^6 \) to 2.8 \( \times 10^6 \) sequences)

- Accession numbers associated with all microbial peptide or EST-translated sequence peptide identifications from first search, were merged with human database to create a target-decoy database so that FDR can be calculated in the second step.

**STEP 2**

- Refined target-decoy database of human + bacterial proteins or translated RNA sequences (~1.5 \( \times 10^6 \) sequences)

- Search high mass accuracy peaklists against refined database

- Distinct peptide sequences from spectra identified at 5% local FDR.

- Validate high confidence bacterial peptide matches or potential alternative site isoforms via BLAST search

- Analyze with downstream metaproteomic or proteogenomic analysis tools.

Taxonomic analysis

• Output compatible with bioinformatic tools (MEGAN)

Bacterial phyla

KEGG pathways

(Joel Rudney)

*Proteomics* 2012, 12, 992–1001
Concluding thoughts: A new paradigm in publishing?

Old paradigm

2 Materials and methods

2.1 Salivary supernatant dataset

Salivary supernatant was collected and pooled from six healthy subjects who refrained from eating or drinking for 90 min. Proteins were analyzed using ProteoMiner™ (Bio-Rad Laboratories, Hercules, CA, USA) for DRC, multidimensional peptide fractionation, and an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) as described in Bandukwala et al. 2009 [8]. Additional 45 RAW files generated from ProteoMiner™ Library-2 treated saliva were also analyzed.

2.2 Two-step method for peptide sequence matching and protein identification

RAW files generated (200 total) from the LTQ-Orbitrap salivary supernatant dataset were processed using the MaxQuant (v1.0.13.13) "Quant" module to generate .MSS files [8, 22-24]. Individual files and iso .MSS files corresponding to each RAW file were converted to Mascot generic format (MGF) and searched using ProteinPilot + 4.0 (ProteinPilot Software 4.0; Revision: 140085; Paragon Algorithm: 4.0.0.0, 148003; AB SCIEX, Framingham, MA). Protein searches were conducted using LTQ-Orbitrap subprogram instrument settings. Other parameters used for the search were as follows: Sample Type: Identification; Cyclization: None; Digestion: Trypsin; LFQ: Biological Modification; Search effort: Thorough.

In the first step, all 200 RAW files were searched against a database consisting of all the trimmed human oral microbial genomic sequences from the Human Oral Microbiome Database (HOMD) [25], along with the human IPI v3.5.2 database and contaminant proteins (1,687,416 total protein sequences) [26]. The ProteinPilot searches generated a group file that was used to generate a Protein Report from the peptide sequence matches. All non-human protein sequences that were identified at the threshold of at least 66% Confidence (0.47 Proteins) in the first step were merged with the Human IPI v3.5.2 database along with contaminant proteins to generate a “refined” FASTA database for the second step.

In the second step, all 200 RAW files were searched against a “Target Decoy” version of the FASTA database mentioned above, by appending the reversed protein sequences to the forward sequences, resulting in a database containing 1,932,724 total protein sequences. Parameters for ProteinPilot

New paradigm: Transparent, complete and usable by others
Concluding thoughts

• “Big Data” repositories: Workflow framework (e.g. Galaxy) offers a way to store and use analytical tools/workflows with raw multi-omic data

• Better ways needed to integrate ‘omic data repositories to realize benefits of multi-omics

• Academic-industry partnerships: a way forward in solving data analysis challenges in multi-omics and Big Data?