Public proteomics data: a (mostly unexploited) gold mine for computational researchers

Dr. Juan Antonio Vizcaíno

Proteomics Team Leader
EMBL-European Bioinformatics Institute
Hinxton, Cambridge, UK

E-mail: juan@ebi.ac.uk
Overview

• Short introduction to proteomics and PRIDE

• Reuse of public proteomics data

• “Big data” approach -> PRIDE Cluster

• Open analysis pipelines
One slide intro to MS based proteomics

Hein et al., Handbook of Systems Biology, 2012
Data resources at EMBL-EBI

Genes, genomes & variation
- European Nucleotide Archive
- European Variation Archive
- European Genome-phenome Archive
- Ensembl
- Ensembl Genomes
- GWAS Catalog
- Metagenomics portal

Gene & protein expression
- ArrayExpress
- Expression Atlas
- PRIDE

Protein sequences, families & motifs
- InterPro
- Pfam
- UniProt

Molecular structures
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Chemical biology
- ChEMBL
- ChEBI

Reactions, interactions & pathways
- IntAct
- Reactome
- MetaboLights

Literature & ontologies
- Europe PubMed Central
- Gene Ontology
- Experimental Factor Ontology

Systems
- BioModels
- Enzyme Portal
- BioSamples
Data resources at EMBL-EBI

Genes, genomes & variation
- European Nucleotide Archive
- European Variation Archive
- European Genome-phenome Archive
- Ensembl
- Ensembl Genomes
- GWAS Catalog
- Metagenomics portal

Gene & protein expression
- ArrayExpress
- Expression Atlas
- PRIDE

Protein sequences, families & motifs
- InterPro
- Pfam
- UniProt

Molecular structures
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Chemical biology
- ChEMBL
- ChEBI

Systems
- BioModels
- Enzyme Portal
- BioSamples

Literature & ontologies
- Europe PubMed Central
- Gene Ontology
- Experimental Factor Ontology

Reactions, interactions & pathways
- IntAct
- Reactome
- MetaboLights
PRIDE (PRoteomics IDEntifications) Archive

- PRIDE stores mass spectrometry (MS)-based proteomics data:
  - Peptide and protein expression data (identification and quantification)
  - Post-translational modifications
  - Mass spectra (raw data and peak lists)
  - Technical and biological metadata
  - Any other related information

- Full support for tandem MS approaches
- Any type of data can be stored
- From July 2017, an **ELIXIR core** resource

http://www.ebi.ac.uk/pride/archive

Martens et al., Proteomics, 2005
Vizcaíno et al., NAR, 2016
Stats (1): Data submissions to PRIDE Archive continue to increase

1,950 datasets submitted to PRIDE Archive in 2016
… and still the number of submitted datasets is growing…
Stats (2): Data growth in EBI resources

![Graph showing data growth in EBI resources](image-url)
ProteomeXchange: A Global, distributed proteomics database

- **Goal:** Development of a framework to allow standard data submission and dissemination pipelines between the main existing proteomics repositories.

http://www.proteomexchange.org

Mandatory raw data deposition since July 2015

Vizcaíno et al., Nat Biotechnol, 2014
Deutsch et al., NAR, 2017
Countries with at least 100 submitted datasets:

- 1019 USA
- 734 Germany
- 492 United Kingdom
- 470 China
- 273 France
- 209 Netherlands
- 173 Canada
- 165 Switzerland
- 157 Australia
- 148 Austria
- **142 Denmark**
- 137 Spain
- 115 Sweden
- 109 Japan
- 100 India

Released: 3462 datasets (66.6%)

Unpublished: 1736 datasets (33.4%)

Top Species represented (at least 100 datasets):

- 2267 *Homo sapiens*
- 765 *Mus musculus*
- 201 *Saccharomyces cerevisiae*
- 169 *Arabidopsis thaliana*
- 154 *Rattus norvegicus*
- 124 *Escherichia coli*

~ 1000 species in total

Stats (3): 5,198 ProteomeXchange datasets in PRIDE

Type:

- 3835 ‘Partial’ submissions (73.8%)
- 1363 ‘Complete’ submissions (26.2%)

Released: 3462 datasets (66.6%)

Unpublished: 1736 datasets (33.4%)

Data volume in PRIDE:

Total: ~400 TB

Number of files: ~670,000

- PXD000320-324: ~4 TB
- PXD002319-26: ~2.4 TB
- PXD001471: ~1.6 TB
Stats (4): PRIDE share in ProteomeXchange (May 2017)

- PRIDE: 5571 (88.2%)
- MassIVE: 516 (8.2%)
- PeptideAtlas: 139 (2.2%)
- jPOST: 86 (1.4%)
PRIDE Inspector Toolsuite: data visualisation/ QC

- PRIDE Inspector - standalone tool to enable visualisation and validation of MS data.
- Build on top of **ms-data-core-api** - open source algorithms and libraries for computational proteomics.
- Supported file formats: mzIdentML, mzML, mzTab (**PSI standards**), and PRIDE XML.
- Broad functionality.

Summary and QC charts

Peptide spectra annotation and visualization

Wang et al., Nat. Biotechnology, 2012
Perez-Riverol et al., Bioinformatics, 2015
Perez-Riverol et al., MCP, 2016

https://github.com/PRIDE-Utilities/ms-data-core-api
https://github.com/PRIDE-Toolsuite/pride-inspector
Overview

• Short introduction to proteomics and PRIDE

• Reuse of public proteomics data

• “Big data” approach -> PRIDE Cluster

• Open analysis pipelines
The “dark” proteome
The “dark” proteome

• Only ~25-30% of spectra in a typical proteomics experiments are identified.

• What does that fraction of unidentified spectra correspond to?
  • For sure, there will be artefacts (e.g. chimeric spectra).
  • Undetected protein variants:
    • What it is not included in the searched database cannot be found.
  • Peptide containing unexpected Post-Translational Modifications (PTMs).

• Big potential to find novel biological relevant “proteoforms”.

Juan A. Vizcaíno
juan@ebi.ac.uk
Danish Bioinformatics Conference
Odense, 25 August 2017
Could any of these “undetected” proteoforms have an important biological function?

Smith et al., Nat Methods, 2013
Reuse of public proteomics data is on the rise!!


Vaudel et al., *Proteomics*, 2016
Data downloads are increasing

Data download volume for PRIDE Archive in 2016: **243 TB**
MS proteomics: Discovery proteomics (DDA)
Public data re-analysis -> Data repurposing

• Individual authors can re-analyze MS proteomics raw data with new hypotheses in mind (not taken into account by the original authors).

• Proteogenomics studies.
• Discovery of new PTMs.
• Meta-analysis studies.
Across-omics -> Proteogenomics approaches

• Proteomics data is combined with genomics and/or transcriptomics information, typically by using sequence databases generated from DNA sequencing efforts, RNA-Seq experiments, Ribo-Seq approaches, and long-non-coding RNAs.
MS proteomics: Proteogenomics

DNA, RNASEq, RiboSeq

Proteogenomics
MS proteomics: ProteoGenomics

Nesvizhskii, Nat Methods, 2014
Examples of repurposing datasets: proteogenomics

Data in public resources can be used for genome annotation purposes -> Discovery of short ORFs, translated IncRNAs, etc
Examples of repurposing datasets: proteogenomics

Also some studies have been performed in model organisms: mouse, rat, *Drosophila*, and other microorganisms (*Mycobacterium tuberculosis*, *Helicobacter pylori*)
Across-omics -> Proteogenomics approaches

- Proteogenomics approaches are increasingly utilized to understand the information flow from genotype to phenotype in complex diseases such as cancer and to support personalized medicine studies.
  - Study of human variation, e.g. in diseases such as cancer.
MS proteomics: ProteoGenomics

- Personal proteomes
- Personal genomes
- in vivo
- in silico
- Personalised medicine
Public datasets from different omics: OmicsDI

- Aims to integrate of ‘omics’ datasets (proteomics, transcriptomics, metabolomics and genomics at present).

PRIDE
MassIVE
jPOST
PASSEL
GPMDB
ArrayExpress
Expression Atlas
MetaboLights
Metabolomics Workbench
GNPS
EGA

http://www.omicsdi.org/

Perez-Riverol et al., Nat Biotechnol, 2017
OmicsDI: Portal for omics datasets
Public data re-analysis -> Data repurposing

- Individual authors can re-analyze MS proteomics raw data with new hypotheses in mind (not taken into account by the original authors).
  - Proteogenomics studies.
  - Discovery of new PTMs.
  - Meta-analysis studies.
Repurposing: new PTMs found

• Some examples (using phosphoproteomics data sets):
  
  • O-GlcNAc-6-phosphate
  • Phosphoglyceryl
  • ADP-ribosylation

\[^{1}\text{Hahne} \& \text{Kuster, Mol Cell Proteomics (2012) 11 10 1063-9}
\]
\[^{2}\text{Moellering} \& \text{Cravatt, Science (2013) 341 549-553}
\]
\[^{3}\text{Matic et al., Nat Methods (2012) 9 771-2}
\]
Public data re-analysis -> Data repurposing

• Individual authors can re-analyze MS proteomics raw data with new hypotheses in mind (not taken into account by the original authors).

• Proteogenomics studies.
• Discovery of new PTMs.
• Meta-analysis studies.
Recent examples of meta-analysis studies

Lund-Johanssen et al., Nat Methods, 2016

Drew et al., Mol Systems Biol, 2017

Juan A. Vizcaíno
juan@ebi.ac.uk
Danish Bioinformatics Conference
Odense, 25 August 2017

Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes
Overview

- Short introduction to proteomics and PRIDE

- Reuse of public proteomics data

- “Big data” approach -> PRIDE Cluster

- Open analysis pipelines
Introduction to Spectrum Clustering

Input Mass Spectra → spectra-cluster algorithm → Consensus spectra (= data reduction)

- Unidentified spectrum
- Spectrum identified as peptide A
- Spectrum identified as peptide B
The spectra-cluster toolsuite

**Clustering**
- Command-line tool, graphical user interface and Hadoop implementation of the spectra-cluster algorithm.
- Stand-alone tools optimised for small datasets

**Development**
- Parser APIs for Java and Python
- spectra-cluster Java API to facilitate the development of new clustering algorithms

**Analysis**
- Growing collection of simple-to-use tools for detailed analysis
- spectra-cluster-py Python framework available for the development of own scripts

[https://spectra-cluster.github.io](https://spectra-cluster.github.io)
PRIDE Cluster - Concept

Originally submitted identified spectra

Consensus spectrum

Threshold: At least 3 spectra in a cluster and ratio >70%.
PRIDE Cluster: Second Implementation

- Clustered all public, identified spectra in PRIDE
- EBI compute farm, LSF
  - 20.7 M identified spectra
  - 610 CPU days, two calendar weeks
- Validation, calibration
- Feedback into PRIDE datasets
- EBI farm, LSF
- Griss et al., Nat. Methods, 2013

- Clustered all public spectra in PRIDE by summer 2015.
- Apache Hadoop.
  - Starting with 256 M spectra.
  - 190 M unidentified spectra (they were filtered to 111 M for spectra that are likely to represent a peptide).
  - 66 M identified spectra
  - Result: 28 M clusters
  - 5 calendar days on 30 node Hadoop cluster, 340 CPU cores
  - Griss et al., Nat. Methods, 2016
One perfect cluster in PRIDE Cluster web

- 880 PSMs give the same peptide ID
- 4 species
- 28 datasets
- Same instruments

http://www.ebi.ac.uk/pride/cluster/
3. Consistently unidentified clusters

Originally submitted spectra

Consensus spectrum

Spectrum clustering

Not identified

Not identified

Not identified

Not identified

Method to target recurrent unidentified spectra
Consistently unidentified clusters (Recurring Unidentified Spectra)

- 19 M clusters contain only unidentified spectra.

- Most of them are likely to be derived from peptides.

- They could correspond to PTMs or variant peptides -> Potential Biomarkers?

- With various methods, we found likely identifications for about 20%.

- Vast amount of data mining remains to be done.
3. Consistently unidentified clusters
PRIDE Cluster as a Public Data Mining Resource

- http://www.ebi.ac.uk/pride/cluster
- Spectral libraries for 16 species.
- Spectral archives (including the Recurring Unidentified Spectra)
- All clustering results, as well as specific subsets of interest available.
- Source code (open source) and Java API
Status of PRIDE Cluster in 2017

- PX Complete
- PX successfully converted
- New Peptide/PTMs
- Number of Identified and non-Identified Spectra
- Number of new clusters
- PRIDE Cluster score distribution
- Number of clusters by modification
- Number of Peptides
- Number of new Peptides
- Number of PTMs
- Number of New PTMs

Refined / Improved pipeline including robust QC checks.
The main focus is not in quantity any longer: Filtering more PSMs \textit{a priori}
Applications of spectrum clustering…

• Applicable to **small groups of “similar” datasets:**
  • Can be used to target spectra that are “consistently” unidentified.
  • Unidentified spectra could represent PTMs or sequence variants.
  • Try “more-expensive” computational analysis methods (e.g. spectral searches, *de novo*).
  • Improve protein quantification.
Overview

• Short introduction to proteomics and PRIDE

• Reuse of public proteomics data

• “Big data” approach -> PRIDE Cluster

• Open analysis pipelines
Open analysis pipelines

- **Goal:** Development of open, reproducible and modular pipelines (based on OpenMS as a starting point) for DDA (Data Dependent Acquisition) approaches.

- Deployment in the EMBL-"Embassy Cloud”, with the goal that in the future, they can be deployed in other cloud infrastructures, and be reused by anyone in the community.

- **Connected to PRIDE, bringing the tools closer to the data.**
  - We can use these pipelines to reanalyse PRIDE data.
Open analysis pipelines
Open analysis pipelines -> In the near future…

• Recent **3-year BBSRC grant** awarded to do the same for **DIA approaches** (to start on December 2017).
  • In collaboration with the Stoller Center (Manchester) (co-PIs Graham, Hubbard & Townsend)

• Recent **4-year Wellcome Trust grant** awarded to do (among other things) pipelines for proteogenomics approaches (to start mid 2018).
  • In collaboration with J. Choudhary (Institute of Cancer Research, London)
Summary
Summary

- Public proteomics datasets are on the rise! Reliable (widely used) infrastructure now exists.

- A lot of possibilities open for reuse of this data.
  - New purposes: proteogenomics, new PTMs,...

- It is possible to mine public data using spectrum clustering looking for new proteoforms (new potential biomarkers?)

- Starting to work in open and reproducible analysis pipelines.
  - Aim: In the future they are made available to everyone in the community.
Acknowledgements: The PRIDE Team

Attila Csordas
Tobias Ternent
Mathias Walzer
Gerhard Mayer (de.NBI)

Johannes Griss
Yasset Perez-Riverol
Manuel Bernal-Llinares
Andrew Jarnuczak

Former team members, especially
Rui Wang, Florian Reisinger, Noemi
del Toro, Jose A. Dianes & Henning
Hermjakob

All data submitters !!!
www.hupo2017.ie