La qualité des données et des résultats en analyse protéomique

Pierre-Alain Binz

Swiss Institute of Bioinformatics,
Geneva, Switzerland

EMBNet course, 5 Mars 2004

Here are my results:

Can I believe in them?
Are they meaningful?

That’s not the question:
But:
Can others believe in them?
**Why to talk about quality in Proteomics?**

Proteomics was mainly technology development, now it goes to biological interpretation

Publications are difficult to reproduce

Reduce propagation of errors

Allow integration of information

---

**Tasks/needs for Bioinformatics in Proteomics**

**Process handling:**
- Sample and information tracking, workflow integration tools (LIMS)
- Signal detection (MS peaks, spots, …)

**Interpretation of experimental data:**
- Image analysis tools (qualitative and quantitative sample comparison)
- Protein identification, characterization tools (matching, data mining, scoring, prediction, analysis, validation)
- Predict and associate protein forms as members of pathways

**Information source:**
- Databases (sequences, families, structure, function, pathways, 2-DE maps, MS data, DNA arrays, LIMS DB…)}
Complexity in proteomics

Heterogeneous physicochemical properties:
• Multiple protein forms: splicing variants, processing events, PTMs
• Wide range of pI, Mw, solubility, concentration

Complex interactions:
• Protein/protein, protein/DNA, protein/chemicals

Variable, dynamic systems:
• Proteomes differ from individual to individual
• Proteomes vary as function of environment (time, drugs, stress, …)

The protein ABC? OK, which one?

I have identified the protein ABC

splicing variants
truncations, fragments
discrete and heterogeneous PTMs
What is identification, what is characterization?

**Identification:**
matching experimental results with a proteomics database

**Characterization**

- **FUNCTION:** Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for collagenase and thrombin.
- **SUBCELLULAR LOCATION:** Extracellular.
- **TISSUE SPECIFICITY:** Plasma.
- **POLYMORPHISM:** The most common is that of the M2 allele, which is the most frequent allele. Other frequent alleles are: M1A 26 to 23%, M2 10 to 11%, M1B 14 to 15%.
- **DISEASE:** The major physiological function of AAT is the protection of the lower respiratory tract against proteolytic destruction by human leucocyte elastase (HLE). A hereditary deficiency of AAT is associated with a 20-30 fold increased risk of developing chronic obstructive pulmonary disease.
- **DEFICIENCY:** The normal inhibitor in individuals homozygous for the Z or M1 allele can result in the development of chronic emphysema or infantile liver cirrhosis.
- **DISEASE:** Variant Pittsburgh is the cause of bleeding diathesis.

**22 spots in plasma 2-DE**
Proteomics today: a couple of types of biological questions but also:

- many proteomes
- many different proteins
- many different protein forms per protein
- many workflows
- many different instrumentations
- many bioinformatics tools

Proteomics Workflows using Mass Spectrometry: complementarity

1) Classical 1-DE/2-DE -- spot excision -- protein identification
   +: >1000 protein forms detected, PTMs, quantitation
   -: limits for incompatible protein forms

2) Molecular scanner from 1-DE/2-DE
   +: idem 2-DE, contextual info

3) MudPIT and similar
   +: no gels (virtually no incompatible protein forms)
   -: identify peptides, not protein forms
   -: reproducibility due to complexity

4) ICAT and similar
   +: idem MudPIT
   -: only Cys-containing proteins
   -: no differentiation of protein forms

5) SELDI
   +: good in fast, simple, rapid, selectivity
   -: Mw range limited, complexity limited

6) Protein interactions, protein arrays...

Identification and characterisation: What approaches, what tools?
Proteomics Workflows using Mass Spectrometry: complementarity

<table>
<thead>
<tr>
<th>Method</th>
<th>Identification</th>
<th>Characterisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Classical 1-DE/2-DE</td>
<td>PMF, MS/MS</td>
<td>PTM, sequence alterations</td>
</tr>
<tr>
<td></td>
<td>-- spot excision</td>
<td>quantitation on separation step</td>
</tr>
<tr>
<td></td>
<td>-- protein identification/ characterization</td>
<td></td>
</tr>
<tr>
<td>2) molecular scanner from 1-DE/2-DE</td>
<td>PMF, MS/MS</td>
<td>PTM, sequence alterations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>quantitation with isotope labels,</td>
</tr>
<tr>
<td>3) MudPIT and similar</td>
<td>MS/MS</td>
<td>no distinction of protein forms,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no quantitation (15N)</td>
</tr>
<tr>
<td>4) ICAT and similar</td>
<td>MS/MS</td>
<td>no distinction of protein forms,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>quantitation with isotope labels</td>
</tr>
<tr>
<td>5) SELDI</td>
<td>~ no</td>
<td>selection is part of the process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>relative quantitation of signals</td>
</tr>
<tr>
<td>6) Protein interactions, protein arrays...</td>
<td>~</td>
<td>detection of binding partners</td>
</tr>
</tbody>
</table>

Protein identification /characterization variables

- various sample preparation
- various MS technologies (MALDI-MS, ESI-MS/MS, ...)
- various tools
- various parameters
- various databases

different results with variable confidence
**H. ducreyi** proteins identified by 2D LC (requiring at least 1 significant peptide)

- **498** MALDI – 4700 Proteomics Analyzer
- **372** ESI - QSTAR™ Pulsar System

578 total unique proteins identified

**Mascot sequence recovery from LC-MS/MS on ESI-QTOF**

Sequence Coverage: 24%

**PeptIdent sequence recovery from PMF on MALDI-TOF**

33% of sequence covered: Q9Y2X3
What is correct?

Only those validated by identification with two methods?

Every identified protein entries / peptides?

What validation criteria?
How to represent your confidence?

Quality in Proteomics: quid?

• Appropriate choice of sample and technologies
• QC procedures (+/- controls, replicates)
• Reduce human errors
• Manage data
• Detect and consider levels of accuracy in databases
• Detect bioinformatics tools weaknesses
• Interpret correctly / believe in results
• Compare with others (compatibility issues)
Quality in Proteomics: searches on the web

In general, difficult to find:
- homogeneous protocols,
- validity limits of technologies,
- quality criteria for interpretation.

Medline abstracts:
Only hints; papers SHOULD describe in Material and Methods section

In the ABRF web forum:
Query quality and proteomics: 176 hits; only a few about ways to validate and qualify a result or a method

Google search:
Many hits, few real descriptions

Quality in Proteomics: searches on the web

Google search (2/2):

Some Proteomics core labs says that they deliver protein identification results after applying quality criteria …

Foundation of the German Society for Proteomics Research:
Aims to establish technology standards (quality criteria)

ESF workshop on data integration

Some grant proposal guidelines

Proteomics Standards Initiative
How to improve confidence and quality?

Use appropriate samples / controls
Adjust threshold values
Perform more than once
Use different approaches
Check consistency
Get more information
Improve the tools
Have a critical eye
How to improve confidence and quality?

- Use appropriate samples /controls
- Adjust threshold values
- Perform more than once
- Use different approaches
- Check consistency
- Get more information
- Improve the tools
- Have a critical eye
Some quality criteria

The following criteria were set for considering an identification as positive in MS-Fit database searching: (a) at least four matching peptide masses; (b) at least 50% of the measured masses must match the theoretical masses; (c) 40 p.p.m. or better mass accuracy


FROM THE ABRF DISCUSSION FORUM:
Briefly, we search all data on PeptideSearch and ProFound using the following search parameters:

1. Taxonomy: all kingdoms
2. Modifications: none
3. Missed cleavage sites: 1
4. Mass tolerance: 0.3 Da or 0.015%, monoisotopic
5. MW range - from ≤BD to 2x the SDS PAGE estimated MW.

The primary criteria we use for an identification are a ProFound score of 1.0 for the top ranked protein and a minimum sequence coverage of 20% - with both criteria having to be met. The median sequence coverage for the 90 proteins identified was 34%.

(Kenneth Williams (Kenneth.Williams@yale.edu), 1998)
From the ABRF discussion List" <ABRF@list.abrf.org>
Subject: Re: Manual Validation of MS/MS spectra

In general, I agree with Steve's criteria for manual validation of MS/MS. However, I use a different set of criteria for the initial thresholds, when using data searched with Sequest alone. To decide on the cut-off threshold, we determined the range of XCorrs that we got with a random sequence (at the suggestion of Jimmy Eng, we simply inverted the protein sequences in our database, so that they read from C to N-terminus--this nicely randomizes the database, without changing the composition or protein sizes). This will tell you what threshold you need to eliminate random chance hits (for us, this threshold for +1 ions is 2.1, for +2 ions its 2.5, for +3 its 3.1, which allow about 1% of bad data through, when I want more stringent, I use 2.3/2.7/3.3). However, there is good data below these thresholds--I've seen good data down as far down as XCorr 1 for singly or doubly charged or 1.3 for triply charged, when working with weak or noisy spectra. I've seen bad data above these thresholds (particularly for what we call decoys--where I'm using LCQ "tree" data, and the decoy is the "fake" charge form set up by the computer--of course, at the beginning, you don't know which is the decoy and which is the correct, "main" ms/ms data file.)

One thing I do is search against mascot as well.

Katheryn Resing

Date: Mon, 13 Jan 2003 14:46:26 -0500
From: "Christian, Rob" <Rob_Christian@mspeople.com>
To: 'ABRF Discussion List' <ABRF@list.abrf.org>
Subject: RE: Manual Validation of MS/MS spectra

Benjamin - some other information you can use is the mass assignment of the product ions and the presence of appropriate immonium ions. The latter will only be available if you are using a triple quadrupole or a QTof type instrument. If you have an accurate mass instrument, such as a QTof, mass assignment of the product and precursor ions can be extremely powerful information. You should look for continuity in the mass errors as you progress through a series of b or y ions. For example, a spectrum might contain 10 ions that match the masses of y ions for a peptide. If 7 of these ions had mass errors of 5ppm and 3 had 50 ppm errors then, unless the instrument is incorrectly calibrated, this data should be further scrutinized.

Hope this helps

Best Regards,

Rob Christian, Ph.D.
How to improve confidence and quality?

Use appropriate samples / controls
Adjust threshold values
Perform more than once
Use different approaches
Check consistency
Get more information
Improve the tools
Have a critical eye

Analyzing populations of gels with Melanie

Population based comparison

Disease

Disease 2 or Control
**Results Summary**

- **DM** = ob/ob diagnostic markers
- **TM** = Rosi target markers
- **SM** = Rosi side effect markers

And now what? Believe? Understand? Validate?

⇒ Identify and quantify these targets

107 Differentially Expressed Proteins
How to improve confidence and quality?

Use appropriate samples /controls
Adjust threshold values
Perform more than once
Use different approaches
Check consistency
Get more information
Improve the tools
Have a critical eye

Separation of nucleolar proteins

• one-dimensional
• two-dimensional

97 KDa
14 KDa

pI = 4
pI = 7

97 KDa
14 KDa

Faculté de Médecine
Université de Genève
Hôpitaux Universitaires de Genève

Courtesy of Alexander Scherl
Annotated 2-DE Gel

- 46 annotated spots
- 35 different proteins

Annotated SDS-PAGE gel

SwissProt entries
- Hypothetical protein (Y052_Human)
- Proliferating-cell nuclear antigen P120 (No1_Human)
- Antigen NGP1 (NGP1_Human)
- Hypothetical protein (Y062_Human)
- Probable ATP dependent RNA helicase DDX10 (DD10_Human)
- DNA directed RNA polymerase I 135 kDa polypeptide (RPA2_Human)
- Periodic tryptophan protein 2 homolog (PWP2_Human)
- U5 small nuclear ribonucleoprotein component (U5S1_Human)

TrEMBL entries
- DHM1-like protein (Q6ud53)
- Hypothetical 115.7 kDa protein (Q6h0a0)

NCBI entries
- DEADH (Asp-Glu-Ala-Asp/His) box polypeptide 24 (GI:9966805)
- Similar to KIAA0266 gene product (GI:12654625)
- Hypothetical protein FJ00419 (GI:8923388)
Functional classification

Hypothetical function
44

known function
108

unknown function
61

Sequence analysis by BLAST

97% of known proteins have nucleolar localization

Chromatin structure, 6
Ribosomal proteins, 33
Ribosomal Biogenesis, 44
Others, 15
Unpredictable, 61
Proliferation, 6
DNA-PK complex, 5
Fibrous proteins, 11
Chaperones, 7
Translation factors, 4
mRNA metabolism, 21

Faculté de Médecine
Université de Genève
Hôpitaux Universitaires de Genève
Courtesy of Alexander Scherl
GlycoMod: prediction of glycosilations

FindMod: prediction of PTMs

FindPept: prediction of non-specific cleavages, contaminants, etc

PeptideMass: calculation of theoretical peptide masses

How to improve confidence and quality?

Use appropriate samples / controls
Adjust threshold values
Perform more than once
Use different approaches
Check consistency
Get more information

Improve the tools
Have a critical eye
Use of the predictors

PMF: Aldente:
Hough transform to correct for bad calibration,
score from machine learning,
PTMs from SWISS-PROT

MS/MS: POPITAM
swarm intelligence to look for sequence tags
Popitam

Protein Or Peptide Identification using TAndem Mass spectrometry
**MS-MS: Popitam**

1) **INTERPRETING**
- Ionic hypothesis
  - \( b^-\text{NH}_3 \)
  - \( b^+ \text{H}_2\text{O} \)
  - \( y'' \)
- Ionic m/z
- Singly charged b-ions

2) **STRUCTURING**
- Directed Acyclic Graph (Spectrum Graph)

3) **COMPARING**
- Source MS/MS peak list
- Interpreted peak list
- Peptide sequence database
- Identification

---

**Graph**

- All amino acid tags and complete sequences

\[
\text{LVNELTEFAK} \\
\text{LVNE} \quad \text{FAK}
\]

\[
\text{QNTHSP}
\]

- Finding sections in the graph which best explain theoretical peptides
Significant advantages:

- structuring the source data
- use information coming from the peak succession
- Understand the instrument specific ion series
- No calibration necessary
- Tag approach: No precursor m/z necessary to start, Merging tags and looking for PTMs
How to improve confidence and quality?

Use appropriate samples / controls
Adjust threshold values
Perform more than once
Use different approaches
Check consistency
Get more information
Improve the tools
Have a critical eye
How to deal with dataflows?

What about international efforts?

How to improve confidence and quality?
And now, you can publish . . .

What introduction?
   All that allow to understand the relevance.
What information / material and methods?
   All what is useful to understand the experiments.
What data / results?
   All what is useful to be reviewed.
What discussion ?
   All that argue for an appropriate interpretation

What about data that do not fit in a paper form?
Proteomics Standards Initiative

Mission Statement

The Proteomics Standards Initiative (PSI) aims to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification.

The Proteomics Standards Initiative was founded at the HUPO meeting in Washington, April 28–29, 2000 (see Science 286, 629). As a first step, the PSI will develop standards for two key areas of proteomics: mass spectrometry and protein-protein interaction data. We will also participate in the development of a standardized protein sequence format.

Contents:

1. News and future events
2. Work groups
3. 1. Protein-Protein interaction
   2. Mass spectrometry
   3. General proteomics format
4. Publications
5. Previous events
6. Member lists

The Big Picture (PAB)

 PSI MS focuses
Proteomics data integration

- Proteomics data today: “Publish and vanish”
- Need to develop infrastructure to exchange, analyse and archive proteomics data across different, fast-evolving technologies
- Long term: Develop modular standard for functional genomics in collaboration with MGED
- Initial focus on limited, feasible domains:
  - Mass spectrometry
  - Protein-protein interactions

Mass Spectrometry Topics Strategy

- Minimal requirements for publication (MIAME like)
- XML for proteomics (MAGE, HUP-ML, others)
  - Controlled vocabulary
- Database schema (PeDRo and others)
- Tools (export, import, converters, queries)
  - Open source repository
- Database(s)
  - Repository(ies)
- Access to data (test sites)
- Quality and user requirements
- Coordinations
Status

- Input formats from
  - Bruker-Daltonics, Inc.
  - Ciphergen Biosystems, Inc.
  - Micromass-Waters
  - Protagen, Inc.
  - Thermo LabSystems
  - Institute for Systems Biology
  - U Manchester (Pedro*)

- Draft format developed during EBI workshop July 21-25
- Presented at HUPO congress Montreal, October 2003
- Format improved and available on http://psidev.sourceforge.net

Conclusion:

Biological question
Choice of sample(s)
Workflow(s) capabilities and choice
Validation
Data management

Thank you for your attention!