Tutorial for Proteomics Data Submission

Katalin F. Medzihradszky

Robert J. Chalkley

UCSF

Why Have Guidelines?

- Large-scale proteomics studies create huge amounts of data.
- It is impossible/impractical to present all the results.
- The function of the guidelines is to:
    - Provide enough information to be able to explain the experiment.
    - Provide an assessment of the reliability of the results.
    - Provide the data that supports the results, particularly those that have the greatest potential for mis-interpretation, so the readers can manually assess the results that are important to them.

The Problem

- Proteomics experiments are carried out by many different methods, using a variety of instrument types and employing different analysis tools. Hence, many experimental and analysis parameters need to be reported and the parameters required will differ depending on how the experiment was performed and analyzed.

Paris Guidelines

- The present publication guidelines were created at a meeting in Paris in May 2005 by about 30 key people within the proteomics community, including:
  - Academic Researchers
  - Instrument Manufacturers
  - Search Engine Designers
  - Journal Representatives

- The guidelines are published[1] and are on the Molecular and Cellular Proteomics (MCP) website (http://mcponline.org).

- MCP is using these guidelines for reviewing germane submissions.

- Other proteomics journals use different guidelines for a variety of reasons.

- A large percentage of manuscripts submitted to MCP are initially not compliant with the guidelines.

[1] Bradshaw RA, Burlingame AL, Carr S and Aebersold R. *Mol Cell Proteomics* (2006) **5**(5):787-8.

How Does MCP Help Authors Make their Papers Acceptable?

- MCP does try to help/advise authors during the submission and review process on how to make their manuscript compliant.

- MCP has created a checklist based on the Paris guidelines that makes it easier for authors to tell whether they are compliant.

- The journal has created this tutorial to try to explain why certain pieces of information are required and give examples.

- The journal intends to host workshop sessions at conferences to present the guidelines to wide audiences, to answer questions and to try to encourage comments on the guidelines.

- Encourage search engine manufacturers to make it easier to get the required information in the results output.

In the following slides the points on the guidelines’ checklist (in red) will be discussed, point by point, starting with the Experimental section…

Peak Picking Software

In the Experimental Section

- Name of peaklist-generating software and release version (number or date)

- The raw data acquired by the mass spectrometer is converted into a centroided peaklist file for database searching. The program that performs this can do many things, such as:
  - Remove peaks that are below a certain intensity or signal to noise ratio;
  - Require peaks to have a certain resolution in order to make the list;
  - Set a threshold for the maximum number of peaks within a mass range;
  - Assign a charge state to ions;
  - Merge together MS/MS spectra that have the same precursor mass.

In the Experimental Section

Peak Picking (cont’d)

- Parameters used – default vs altered

- Using different peak picking software / parameters will change the subsequent database searching results, so it is necessary to report the software used for peaklist generation and parameters used.

Examples of software

- Extract_msn in Bioworks 3.0 (Thermo)
- Mascot.dll v1.6b19 (Applied Biosystems)
- DataAnalysis 3.2 (Bruker)
- Mascot Distiller v2.1

- As many people use default parameters for whatever software they use, by specifying the version of the software it can be sufficient to state that the default parameters were used.

In the Experimental Section

Search Engine

- Name of the search engine and release version (number or date)

- Search engines change over time:
  - An improved scoring system may be implemented.
  - New ways to filter the results may be included.

Hence, the version number of the search engine is important.

Database Search Parameters

In the Experimental Section

- Enzyme specificity considered
- # of missed cleavages permitted
- Fixed modification(s) (including residue specificity)
- Variable modification(s) (including residue specificity)
- Mass tolerance for precursor ions
- Mass tolerance for fragment ions

Search engines work by:

1. Determining a list of potential peptides from a database that can be formed by the specified parameters and have the correct precursor mass.
2. Determining scores for the matches of the fragment peaks to each of these peptides.

- Changing search parameters will change the number of the potential hits.
- For software using probabilities or Expectation values (E-values) for scoring, it will change the scores.

Search Parameters - Database

In the Experimental Section

- Name of database searched and release version/date

- Entries in a database change over time:
  - New entries are submitted.
  - Old entries may be removed.
  - Multiple entries may be merged into one new entry.

Database examples,

- NCBI nr 2006. 07. 18; 3794285 sequences
- MSDB database updated May 15, 2005, 2011572 sequences
- IPI human database version 3.16, 62322 entries

(Inclusion of entry numbers is discussed on the next slide).

Search Parameters - Database (cont'd)

In the Experimental Section

- Species restriction and justification for searching only a subset of a database
- Number of protein entries in the database *actually* searched

- The number of entries in the database that are searched influences the reliability of the results.
- Probability / E-value scores are calculated on the basis that all the possible identifications for peptides are in the database, and that matching any peptide in the database is equally likely.
    - If a database is searched that contains sequences from proteins that are not possible to be in the sample, then the confidence of matches are going to be lowered.
    - If the database is too small, and so does not contain some possible answers (e.g. human keratin), the search engine will return an over-confident assessment of a match.
    - Probability/ E-value results for searches against a small database can be very inaccurate and unreliable.

In the Experimental Section

Search Parameters (cont'd)

- Threshold score/E-value for accepting *individual* MS/MS spectra
- Justification of the threshold employed
- For large datasets – estimation of false positive rate and how this was calculated.

- All spectra will have a top match, but not all are correct.
- A threshold has to be defined, below which results are discarded.
- This cut-off score has to be statistically justified.
  - e.g. Mascot uses a threshold of 5% probability that a protein identification is incorrect as a default. This threshold is a (probability) score and should be stated in the manuscript.
- Performing searches using a combination of normal and reversed (or random) databases is an effective way of estimating the number of incorrect peptide and protein assignments in the dataset as a whole when employing a given score threshold. However, this requires many spectra to be accurate and, therefore, is only meaningful on large datasets.

Referring to a paper where the authors used the same cut-off values cannot be considered as justification, as most scoring systems are dataset and search parameter dependent.

If PTMs are being reported…

- Software/method used to evaluate site assignment

In the Experimental Section

- Search engines are much more reliable at peptide identification than modification site assignment.
- Search engines are optimized for peptide identification. Optimum parameters for site assignment are different to those for peptide identification.
  - For example, more peaks are often required for site assignment than peptide identification.
- When a search engine does not have evidence to assign the site, it ‘guesses’; i.e. it assigns a site that is consistent with the data even if there is no specific evidence for the site assignment.
- It is difficult to tell which assignments are confirmed by the data and which are one of multiple possibilities based on the data.
- Personal inspection may often be the most reliable tool.

Hence, the journal requires annotated spectra for all peptides where a modification site is being reported.

Peptide Mass Fingerprint Data

In the Experimental Section

- Name of software used for peak-picking and its release version

- The software used for converting the raw data into a peak list must be reported.

  Example software:

  - “flexAnalysis (version 2.0)” for a Bruker MALDI-TOF.

- Different versions of the same software may use different default parameters, so the release version is also important.

PMF Peak Picking

In the Experimental Section

- Parameters and thresholds used for peak-picking; e.g. intensity or S/N threshold, resolution, means of calibrating each spectrum, list of excluded contaminant ions and justification

- In many ways, good peak-picking software and parameters are more important for PMF data than MSMS data, as unmatched peaks are more significant in assessing the reliability of the results.

- It is common to exclude certain peaks from the list to be submitted for database searching; e.g. trypsin autolysis peaks, matrix cluster ions, masses of tryptic peptides from keratin. If this is done, this needs to be stated.

PMF Acceptance Criteria

In the Experimental Section

- The authors must state the threshold used for acceptance of PMF-based identifications.

- Some software uses probability / E-value based scoring. For these, stating a probability / E-value threshold is sufficient.
- For software that does not use a statistical score, in general the threshold will probably include a minimum number of peaks matching and a minimum percentage of peaks or a minimum protein sequence coverage.
- For software that does not use a statistical score, some other evidence in addition to a score must be presented; e.g. the score for the highest non-homologous protein match. A Western blot may also provide this confirmation as well as MS/MS (PSD, CID) analysis of one or more peptides.

Combining Peptides into “Proteins Identified”

In the Experimental Section

- If peptides match to multiple members of a protein family, criteria used for selecting which member to report; i.e. how was the redundancy eliminated/handled (this is an issue for *all* protein databases).
- How were isoforms/individual members of a protein family unambiguously identified.

- Protein databases often contain entries with similar sequences:
  - Proteins from the same species sharing sequence stretches;
  - Equivalent proteins from different species;
  - Multiple entries for one gene product.

- Hence, often the peptides identified match equally well to multiple database entries. In this situation the authors have two options:
  - Report all the proteins the data supports equally well.
  - Choose to only report one accession, in which case they must justify why they chose one particular entry over others; e.g. they have corollary information such as a Western blot that identifies a particular isoform.

In the Experimental Section

Quantitative Studies

- How the quantitation was performed (number of peaks, peak intensity, peak area, extracted ion chromatogram).

- There are many valid mass spectrometry strategies for relative or absolute protein quantitation, both in terms of the method of quantitation, e.g.
  - Isotopic labeling (chemical or metabolic)
  - Number of peptides identified

and the results can be measured in different ways, e.g.

- Peak intensity
- Peak area
- Area of peak from extracted ion chromatogram

The strategy employed must be clearly stated.

In the Experimental Section

Quantitative Studies (cont'd)

- Minimum thresholds required for data to be used for quantitation.
- Outlier datapoints removed. If so, give justification.

- Measurements for weak signals are going to have a lower accuracy.
- There are sometimes other reasons why specific datapoints are excluded (e.g. there was a co-eluting peak with similar m/z).
  - If any data is excluded from the results this must be stated with justification.

Quantitative Studies (cont’d)

In the Experimental Section

- Explanation of statistics used to assess accuracy and significance of measurements.

- Quantitation is usually done at the peptide level but reported at the protein level.
- It is nearly always possible to perform statistical analysis to measure the accuracy / reliability of measurements, e.g.
  - There may be certain protein(s) that are known not to change between samples.
  - Variability in measurements can be determined by assuming all measurements for the same protein should return the same quantitation results at the peptide level.
- Hence, statistics such as means and standard deviations or coefficients of variation can be determined to set a confidence level for whether a measurement represents a statistically significant change.

Quantitative Studies (cont.d)

In the Experimental Section

- Indicate how biological and analytical reproducibility were addressed by experimental design.

- In order to determine whether a change is biologically significant, either the experiment must be repeated with multiple biological preparations, or the change must be independently confirmed by another means; e.g. a quantitative Western blot.

Now that the experiment/s have been adequately described, it is time to present the results…

In the Results Section

Protein Identifications

- For each protein identified the following should be reported in a table:
  - Accession number

- When a frequently studied, highly conserved protein is identified there may be a list of accession numbers that match the peptides identified.
- If only one or a few of this list are reported in the table then this should be indicated and justification for why a particular entry was chosen should be given, e.g.
  - The author had independent evidence that a particular isoform was present.
  - The database entry had the most descriptive protein name
  - The database entry was the best annotated.

Do not report accession numbers of protein constructs or ‘hypothetical proteins’ whenever possible.

Protein Identification (cont'd)

In the Results Section

- Number of *unique* (in terms of amino acid sequence) peptides identified
- % sequence coverage identified from MS/MS data or a list of sequences identified

- Unique means sequences differing in at least 1 amino acid residue:
  - Covalently modified peptides, including N- or C-terminal elongation (i.e. missed cleavages) count as unique; different charge states or multiple fragmentation spectra of the same peptide do NOT count.
- Sequence coverage is calculated by dividing the number of amino acids observed by the protein amino acid length (66/330 = 22%)
  - Observing the same peptide (even differently modified) does not increase sequence coverage.

Do NOT list sequences whose score were below the threshold you stated you were using to guarantee reliable answers in the experimental section.

Example Format for Presenting Protein Identifications Based on Multiple Fragmentation Spectra

In the Results Section

*	*	*	
Accession number	**Unique peptides detected**	**Sequence coverage %**	**Expectation-value**
ADA22_MOUSE	4	17	1.3E-05
Q5SV72_MOUSE	6	10	9.0E-04
Q9Z0P5_MOUSE	2	10	7.5E-06
Q7TNV2_MOUSE	4	9	8.4E-05
ABI1_MOUSE	11	46	1.6E-05
ABI2_MOUSE	6	39	4.8E-05
Q921H8_MOUSE	3	20	9.4E-03

* Required information

•Additional information, such as a protein’s name, function, MW, pI, score, peptide sequences, etc. may be included.

Single Peptide Protein IDs and PTMs

In the Results Section

Create a Table and include:

- Sequence identified
- The precursor m/z and charge
- Score / E-value for this peptide

- If a score / E-value threshold has been employed at the peptide level, the majority of any incorrect assignments will be to peptides that are the only one assigned to a given protein, i.e. ‘one hit wonders’.
- Search engines are significantly more reliable at identifying peptide sequences than identifying sites of modification.

Hence, the journal requires extra information about single-peptide-based protein IDs and PTM assignments.

Example Format for Presenting Single-Peptide-Based Protein Identifications

In the Results Section

*
either/or

* * * *

Acc #	m/z	z	error [Da]	Peptide	Score	Expect
P04264	979.4174	3	1.2	FLEQQNQVLQTKWELLQQVDTSTR	38.4	1.4e-6
P35527	919.4144	2	-0.072	HGVQELEIELQSQLSK	48.7	2.6e-8
P13645	998.8564	2	-0.13	ELTTEIDNNIEQISSYK	37.8	6.4e-5
P02768	722.2113	2	-0.11	YICENQDSISSK	33.3	9.6e-6
P48666	724.2895	2	-0.10	AIGGGLSSVGGGSSTIK	39.4	2.1e-5
Q16195	1044.3822	2	-0.10	GQTGGDVNVEMDAAPGVDLSR	38.2	4.5e-6
P13646	651.2297	2	-0.10	ALEEANADLEVK	36.8	0.0012
P01857	937.3319	2	-0.13	TTPPVLDSDGSFFLYSK	34.6	7.1e-4

* Required information

•Additional information, such as the protein’s name, MW, pI, mass measurement error, number of database entries that contain this peptide, i.e. uniqueness, etc. may be included.

Single-Peptide-Based Protein IDs and PTMs

In the Results Section

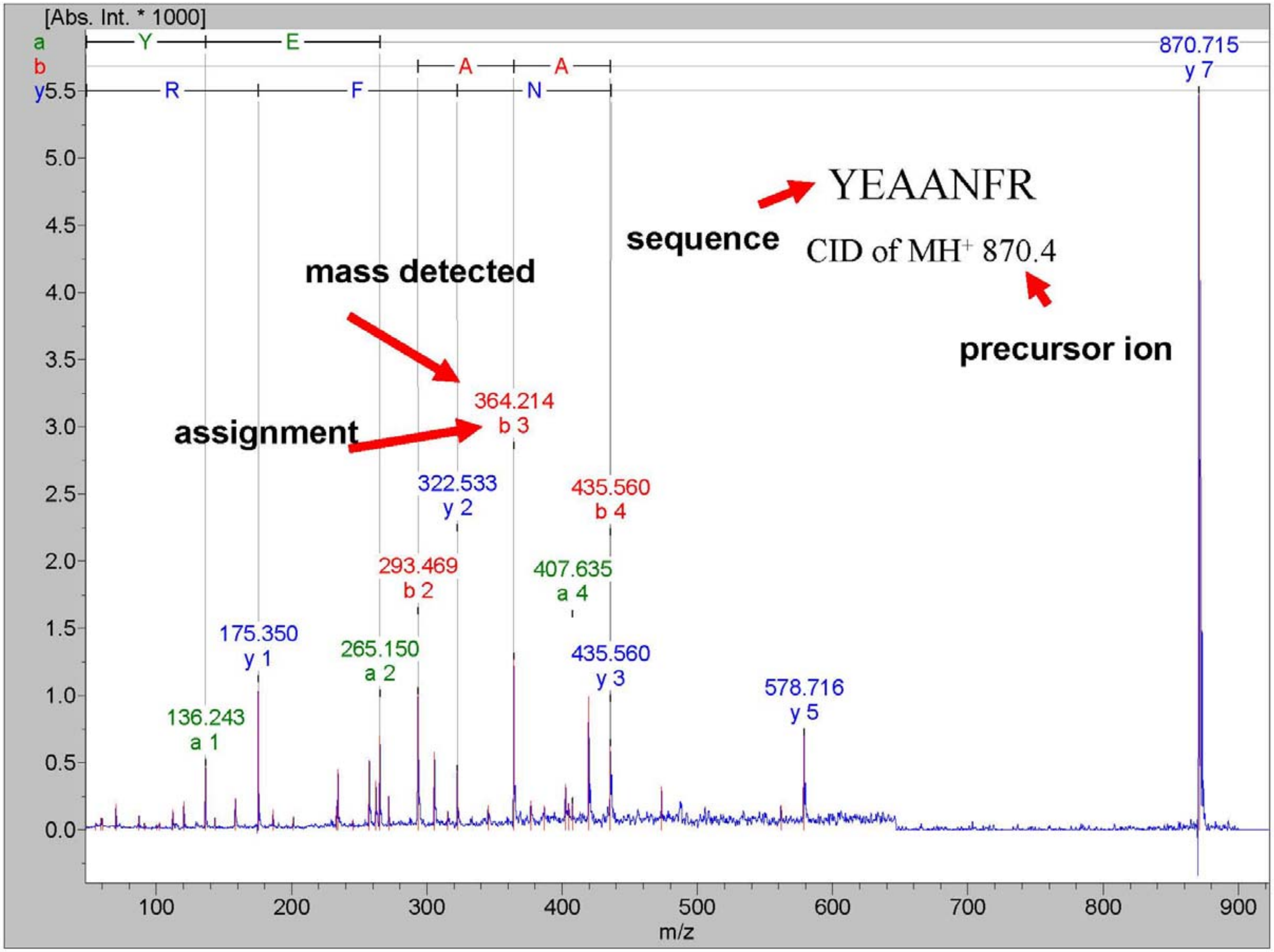
- For each protein or site, a MS/MS spectrum appropriately labeled should be included, with masses detected as well as fragment assignments

- It is recognized that properly labeled spectra are not always readily produced.
- Other acceptable options (in order of preference):
  - Two copies of the same spectrum/page – one labeled with the masses, the other with fragment assignments;
  - Spectrum labeled with masses, accompanied with a Table of the fragment assignments;
  - Fragment assignments provided by the search engine in the spectrum and corresponding masses highlighted in a Table of theoretical fragments (e.g. Mascot results output).

These files can be quite large. If there are large numbers of spectra it is often easier to split them between multiple files. We suggest to not include data on more than 50 proteins in a single file.

Example Presentation of Spectrum of a Single-Peptide-Based Protein ID

In the Results Section



sequence

YEAANFR

CID of MH^+ 870.4

precursor ion

mass detected

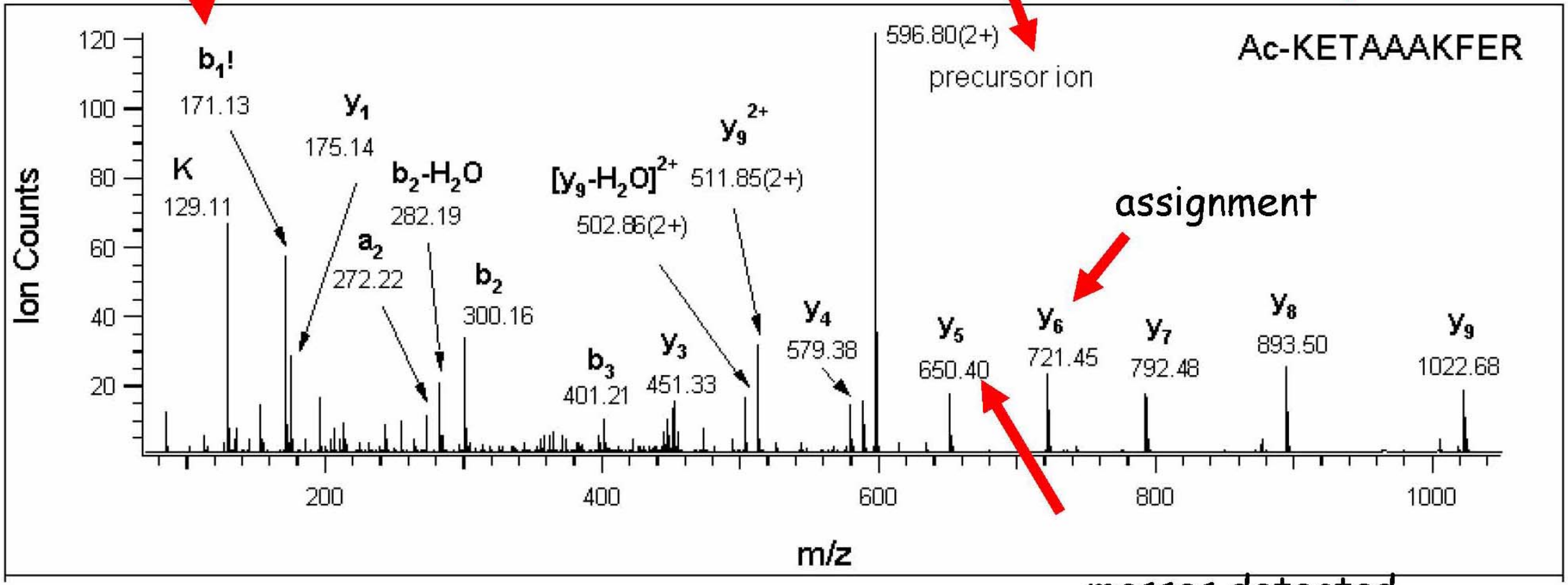
assignment

Example Presentation of Spectrum of a Modified Peptide

In the Results Section

✷ ion(s) that verify the site-assignment

sequence & site of modification



assignment

masses detected

✷Not required, but recommended.

PMF-based Identifications

In the Results Section

For PMF-based IDs

- Number of masses matched
- Number of masses not matched
- % sequence coverage
- Criteria for acceptance

- For PMF data, the number of matched and unmatched peaks are both important for assessing the reliability of a match.
  - Search engines that use a probability based scoring will take both of these factors into account for their scoring.
  - For these search engines, stating the number of peaks identified is sufficient to assess the reliability (obviously along with the probability cut-off as acceptance criterion).

PMF-based Identifications (cont’d)

In the Results Section

- For search engines that do not use a probability based scoring the number of peaks matched and unmatched must be reported.
  - If poor parameters were used for the peaklist generation this will make assignments appear less reliable than they may actually be due to the high number of masses unaccounted for.
- For search engines that do not use probability/ E-value scoring extra information is required to assess the reliability; e.g.
  - Comparison of the score to that of the highest ranking non-homologous protein.
  - Sequence confirmation by MS/MS (the properly labeled spectrum should be submitted) or Edman-sequencing
  - Western blot

Example Presentation of PMF Results

In the Results Section

Could also be a score; probability; score comparison...

*	*	*	* **Acceptance criterion**.
Accession#	Matched /searched	Seq.Cov	Sequence confirmed by PSD or CID^{e}
BAD24662	11/24	28%	^{103}VLQAPGYNGTGKDWALIK^{120}
CAA46635	27/34	41%	^{405}SASGGGFYPPDEVIER^{420}
BAD67179	15/28	37%	^{279}WGGTAGQAFDR^{289}
CAA01746	6/21	46%	^{47}TGTSFPNNDYGIIR^{60}
BAC21011	9/17	13%	^{318}SGIRGDGVGAYSR^{330}
CAH94303	13/14	31%	^{398}DSVLGYADVTLPPGR^{412}
CAB65418	18/47	38%	^{353}ILTDAGYAPIPAEINAK^{369}
AAD30139	19/36	37%	^{121}AAATTQGSGWGVLAYEPVSGK^{141}
AAD30139	13/35	34%	^{52}DKEAWGAINGLQK^{64}

* Required information

- Additional information, such as protein's name, function, MW, pI, may be included.

PMF-based Identifications

In the Results Section

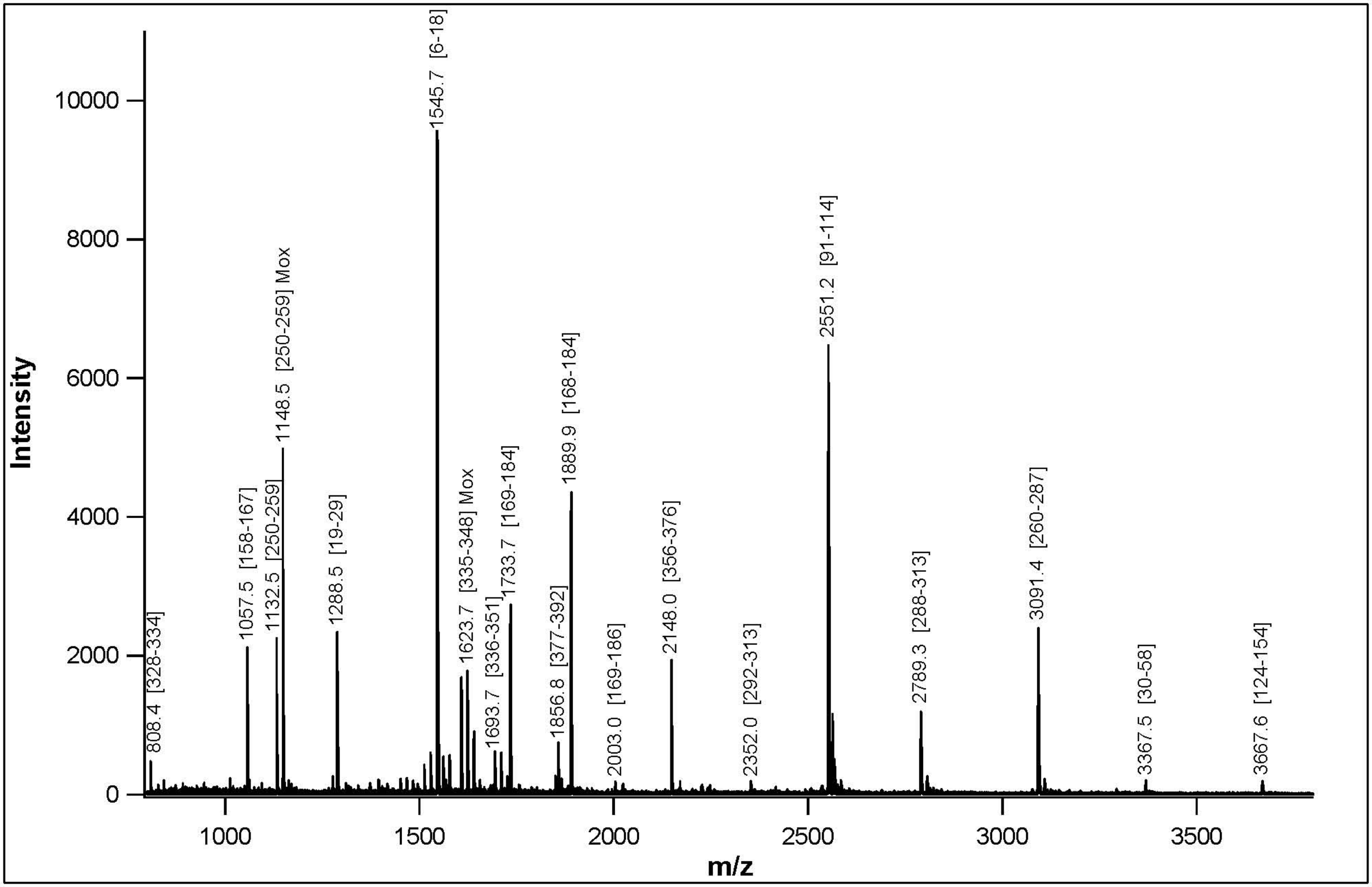
- MS spectra appropriately labeled should be included – masses detected as well as peptide assignments.

- Some software will not produce spectra with mass and assignment labeled peaks.

- Alternative acceptable format: The spectrum with the masses labeled, accompanied with a table with the masses and peptide assignments.

Example Presentation of PMF Spectrum

In the Results Section



Quantitation Results

In the Results Section

- Number of peptides used for protein quantitation measurement

- The reliability of a protein level measurement will be improved by more peptide measurement data points. It is especially important to indicate any protein quantitation measurements that are the result of a single peptide measurement.

In the Results Section

Quantitation Results

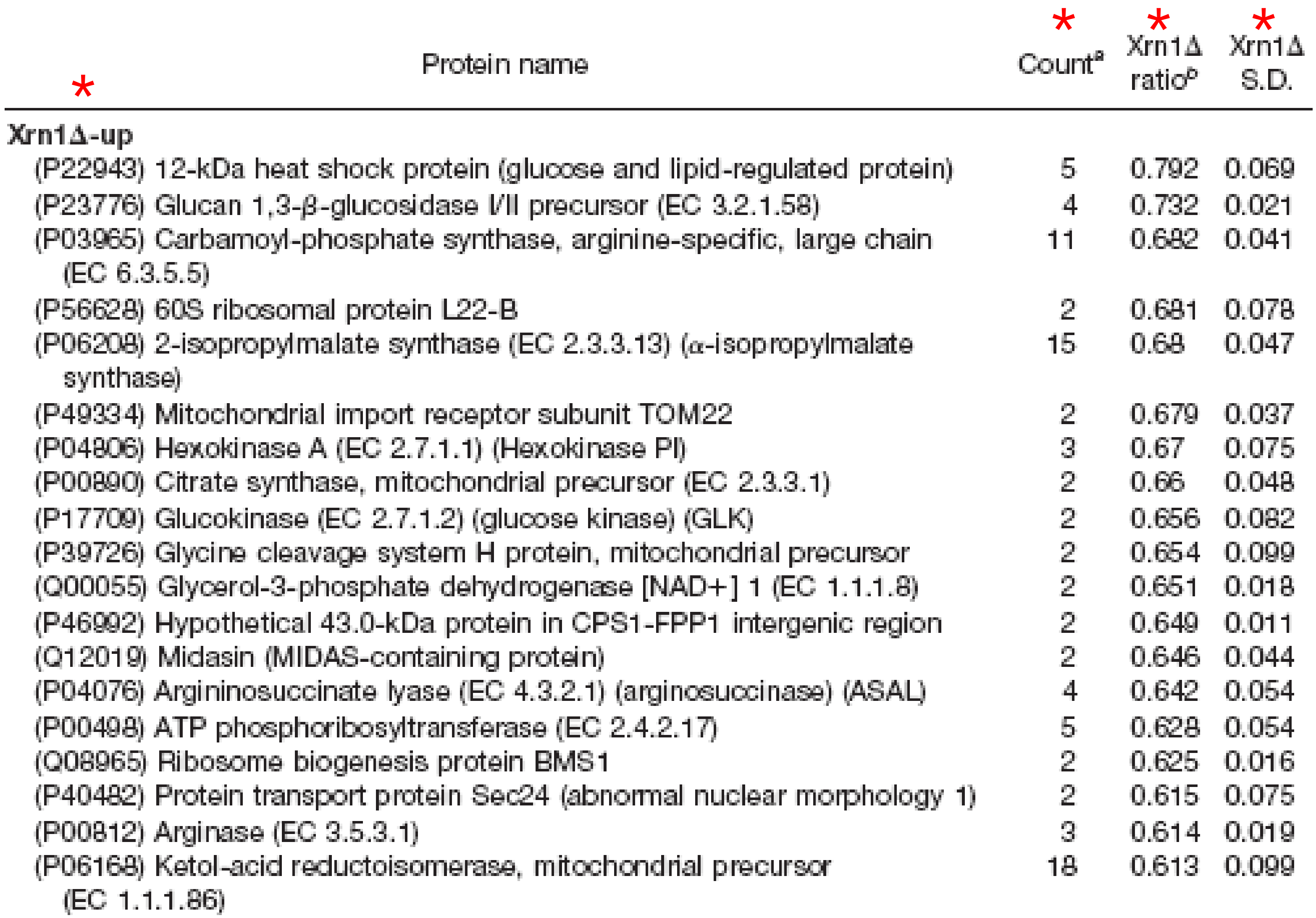
- Protein quantitation measurement and accuracy (e.g. mean and standard deviation)

- There are many factors that can affect the accuracy of measurements:
  - The quantitation method (label-free; SILAC; ICAT etc.)
  - The degree of label incorporation (for isotopic labeling measurements);
  - The amount of sample;
    - Problems with signal to noise for weak signals;
    - Saturation problems for high level samples.

Hence, it is not sufficient to quote measurement accuracies from other studies using a similar method and assume the same accuracy in your study.

The threshold for deciding if a result represents a significant change should be justified in relation to the accuracy of your quantitation measurements.

Example Presentation of Quantitation Data



* Protein name	* Count[a]	* Xrn1Δ ratio[b]	* Xrn1Δ S.D.
Xrn1Δ-up			
(P22943) 12-kDa heat shock protein (glucose and lipid-regulated protein)	5	0.792	0.069
(P23776) Glucan 1,3-β-glucosidase I/II precursor (EC 3.2.1.58)	4	0.732	0.021
(P03965) Carbamoyl-phosphate synthase, arginine-specific, large chain (EC 6.3.5.5)	11	0.682	0.041
(P56628) 60S ribosomal protein L22-B	2	0.681	0.078
(P06208) 2-isopropylmalate synthase (EC 2.3.3.13) (α-isopropylmalate synthase)	15	0.68	0.047
(P49334) Mitochondrial import receptor subunit TOM22	2	0.679	0.037
(P04806) Hexokinase A (EC 2.7.1.1) (Hexokinase PI)	3	0.67	0.075
(P00890) Citrate synthase, mitochondrial precursor (EC 2.3.3.1)	2	0.66	0.048
(P17709) Glucokinase (EC 2.7.1.2) (glucose kinase) (GLK)	2	0.656	0.082
(P39726) Glycine cleavage system H protein, mitochondrial precursor	2	0.654	0.099
(Q00055) Glycerol-3-phosphate dehydrogenase [NAD+] 1 (EC 1.1.1.8)	2	0.651	0.018
(P46992) Hypothetical 43.0-kDa protein in CPS1-FPP1 intergenic region	2	0.649	0.011
(Q12019) Midasin (MIDAS-containing protein)	2	0.646	0.044
(P04076) Argininosuccinate lyase (EC 4.3.2.1) (arginosuccinase) (ASAL)	4	0.642	0.054
(P00498) ATP phosphoribosyltransferase (EC 2.4.2.17)	5	0.628	0.054
(Q08965) Ribosome biogenesis protein BMS1	2	0.625	0.016
(P40482) Protein transport protein Sec24 (abnormal nuclear morphology 1)	2	0.615	0.075
(P00812) Arginase (EC 3.5.3.1)	3	0.614	0.019
(P06168) Ketol-acid reductoisomerase, mitochondrial precursor (EC 1.1.1.86)	18	0.613	0.099

* Required

The Future

- These guidelines will be continuously adjusted to the ever changing needs of the proteomics field.

- As software improves to become easier to output results into compliant formats, it will become easier to properly present proteomics data in any forum.

- As common formats e.g. mzML, AnalysisXML, become universally available, tools will be produced that will be able to automatically extract the relevant information from raw data and search results.

- Once these formatting problems are solved, journal submission guidelines for proteomics data are expected to become similar for all journals.

- Making raw data publicly available is going to become more common.