Instructions to Authors Regarding Required Manuscript Content and Publication Guidelines for Molecular and Cellular Proteomics

The following guidelines describe the information required by the journal for all articles dealing with mass spectrometric analyses designed for protein, peptide or posttranslational modification (PTM) identification and their quantification regardless of whether the study is biological or clinical in focus.

Authors must include information in the manuscript that will allow readers and reviewers to assess the validity and reproducibility of their data, as well as the significance of their observations. This information includes a description of - and rationales for - the overall experimental design (e.g. the number of statistically independent, technical or biological replicates, the sample size, etc.). In addition, detailed information as to how raw mass spectrometry data was converted into a format for database searching, the search engine employed in data processing, the database(s), the scoring function(s), the false discovery rate (FDR) and how it was calculated and what statistical methods were used to infer significance must be clearly stated and justified. These topics are dealt with in detail, below. It is very important that authors also review the document entitled Checklist for publication of peptide & protein identification data in MCP. Failure to provide the required information will result in significant delays and rejection of the manuscript until compliance is achieved.

If your manuscript describes a clinical study or a glycomics study, please see in addition the documents entitled Publication Guidelines for Preparing Manuscripts Describing Research in Clinical Proteomics and Guidelines for the publication of glycomic studies, respectively.

Manuscripts that use previously generated datasets for software or algorithm development should cite the original literature reference(s) related to the datasets and provide the name and web address of the repository used and the specific data set identifiers. Datasets not associated with previously published studies are subject to all of the review guidelines described for herein for research papers.
Instructions to Authors Regarding Required Manuscript Content and Publication Guidelines for *Molecular and Cellular Proteomics*

I. INFORMATION THAT MUST BE PROVIDED IN THE EXPERIMENTAL METHODS SECTION

**Experimental Design and Statistical Rationale**: Authors must include a subsection in the Experimental Methods section with the header “Experimental Design and Statistical Rationale”. In this section, clearly state:

- the total number of samples analyzed and described in Results
- the number of technical, process and/or biological replicates performed. If no replicate analyses were performed, clearly state why this should be considered acceptable for your study.
- the numbers and types of controls employed
- the methods used for randomization (if appropriate for your study)
- rationale for the choice of sample numbers, replicates, etc. Fully describe and/or reference the statistical tests used for these analyses and provide reasons for choice of statistical tests used.

**Database Search Parameters And Acceptance Criteria For Identifications**: For all manuscripts describing either MS/MS or Peptide Mass Fingerprint (PMF) analyses, authors must detail in the Experimental Methods section of the manuscript how raw mass spectrometry data were converted into a format for database searching, the search engine employed in data processing, the database(s) used, the scoring function(s) employed, the false discovery rate (FDR) and how it was calculated and the statistical methods used to infer significance. Specifically, please describe the following:

- **Peak Lists**: The method and/or program (including version number and/or date) used to create the "peak lists" from the original data and the parameters used in the creation of this peak list, particularly any processing which might affect the quality of the subsequent database search. Examples include smoothing, any signal-to-noise thresholding, charge states assignment or de-isotoping, etc. In cases where additional customized processing of the collections of peak lists has been performed, e.g. clustering or filtering, the method and/or program (including version number) should be referenced.

- **Search Engine**: The name AND version (or release date) of all programs used for database searching must be provided.

- **Sequence Database or Spectral Library**: The name AND version (or release date) of all sequence database(s) or spectral libraries used must be listed. If a database or library was compiled in-house, a complete description of the source of the sequences or spectra is required and the software used for library generation. The number of entries actually searched from each database or library must be included. If
the database or library used is very small (<1000 entries) or excludes common contaminants, justification must be specifically provided since this may generate misleading assignments and an inaccurate false discovery rate estimate.

- **Enzyme specificity:** A description of all enzymes used to generate peptides, including the number of missed and non-specific cleavages (e.g. semi-tryptic) permitted, must be listed.

- **Fixed modification(s):** A list of all modifications considered (including residue specificity) must be given.

- **Variable modifications:** A list of all modifications considered (including residue specificity) must be given. If there fixed or variable modifications were not specified, this should be so stated.

- **Mass tolerance for precursor ions.**

- **Mass tolerance for fragment ions** (not required for PMF data).

- **Known contaminants excluded** (particularly for PMF data): All omitted peaks from pre-designated contaminants (or if any of these fragments are used for calibration) must be identified.

- **Threshold score/expectation value:** Criteria used for accepting individual spectra should be stated along with a justification.

- **False Discovery Rates at Peptide and Protein levels:** For large scale experiments, the results of any additional statistical analyses that estimate a measure of identification certainty for the dataset, or allow a determination of the false discovery rate, e.g., the results of decoy searches or other computational approaches.

### II. INFORMATION THAT MUST BE PROVIDED IN THE RESULTS SECTION OR IN SUPPLEMENTAL MATERIAL

**Protein And Peptide Identification:**

- **All peptide sequences assigned:** A list (in one or more Tables), noting any deviation from the expected enzyme cleavage specificity, must be provided.

- **Precursor charge and mass/charge:** These parameters should be listed for each peptide assignment in the same table.

- **All modifications observed.**
Instructions to Authors Regarding Required Manuscript Content and Publication Guidelines for *Molecular and Cellular Proteomics*

- **Number of matched and unmatched masses**: For PMF data, the total number of peaks, both matched and unmatched, should be listed in the identification table.

- **Score(s)**: The relevant score (depending on the software used) and any associated statistical information obtained for searches conducted must be provided for each peptide.

- **Protein accession number and sequence database or spectral library source**: (Note: If the identifications are presented only at the peptide level, then protein level information may be omitted.)

- **Count of the number of distinct peptide sequences assigned to each protein**: When computing this number, multiple matches to peptides with the same primary sequence should be counted as a single distinct peptide, including multiple matches that represent different precursor charge states or modification states. Any alternative assumptions must be justified.

- **Protein sequence % coverage**: This value should be expressed as the number of amino acids spanned by the assigned peptides divided by the sequence length X 100. Alternatively, a derived protein identification probability can be given.

- (Note: all articles reporting protein identifications minimally must supply the accession numbers, number of peptides for each identification and the percent coverage in either the manuscript proper or the supplemental material submitted to the journal. Deposition of this information only in a public repository does not meet the journal requirement.)

- **Ability to view annotated mass spectra**: For all proteins identified on the basis of one unique peptide spectrum, (a peptide mass fingerprint spectrum counts as one spectrum), the ability to view annotated spectra for these identifications must be made available. This can be achieved in one of three ways:
  
  - Submission of all spectra and search results to a public results repository that is equipped with a spectral viewer prior to submission of the manuscript to the journal. This information will appear as a hyperlink in the published article. Please see the document entitled “Information on how to provide access to annotated spectra” for details on how to comply.
  
  or

  - Submission (with the manuscript) of spectra and search results in a file format that allows visualization of the spectra using a freely-available viewer.
  
  or

  - Submission (with the manuscript) of annotated spectra in an ‘office’ or PDF
Instructions to Authors Regarding Required Manuscript Content and Publication Guidelines for Molecular and Cellular Proteomics

Note: Files submitted through the online manuscript submission process should be less than 100 MB in size. If files are greater than 100 MB in size, the journal recommends depositing the file in a suitable repository, such as member of the proteomeXchange consortium [http://www.proteomexchange.org], then supplying the dataset identifier in the manuscript and in the cover letter accompanying the manuscript submission. Authors must provide in the cover letter at the time of submission the USER NAME and PASSWORD for all deposited data, regardless of the repository selected, if the data is not publicly available during the review process. Failure to do so will cause the manuscript to fail the compliance check and delay the final decision on the article. Posting of results on the author’s website as the sole source of this data does not satisfy this requirement, as the ability to anonymously access the data is necessary for the review process.

Posttranslational Modifications: Studies focusing on posttranslational modifications (PTMs) require specialized methodology and documentation to assign the type(s) and site(s) of the modification(s). The guidelines in this section apply to PTMs that occur under physiological conditions and to which biological significance may be assigned, such as phosphorylation, glycosylation, etc. as well as purposefully induced chemical modifications of central importance to the results of the study, such as chemical cross-linking. These guidelines do not apply to common modifications arising from sample handling or preparation such as oxidation of Met or alkylation of Cys. In addition to the tabular presentation(s) of the data described in guideline II, the following information is required:

- **The site(s) of modification:** Within each peptide sequence, all modifications must be clearly located (unless ambiguous; see below) and the manner in which this was accomplished (through computation or manual inspection) must be described.

- **A justification for any localization score threshold employed.**

- **Ambiguous assignments:** Peptides containing ambiguous PTM site localizations must be listed in a separate table from those with unambiguous site localizations. In cases where there are multiple modification sites and at least one is ambiguous, then these peptides should be listed with the ambiguous assignments. Ambiguous assignments must be clearly labeled as such.

Examples of ambiguities include:

- Modified peptides in which one or more modification sites are ambiguous.
Instructions to Authors Regarding Required Manuscript Content and Publication Guidelines for Molecular and Cellular Proteomics

- Instances where the peptide sequence is repeated in the same protein so the specific modification site cannot be assigned.
- Instances in which the same peptide is repeated in multiple proteins, e.g. splice variants and paralogs (See also Section IV).
- Isobaric modifications (e.g., acetylation vs. trimethylation, phosphorylation vs. sulfonation etc), where the possibilities may not be distinguished. Examples of methods able to distinguish between these include mass spectrometric approaches such as accurate mass determination, observation of signature fragment ions (e.g. m/z 79 vs. m/z 80 in negative ion mode for assignment of phosphorylation over sulfonation), or biological or chemical strategies.
- Annotated, mass labeled spectra: Spectra for all modified peptides must be either submitted to a public repository or accompany the manuscript as described in Section II, above and in more detail in the document “Information on how to provide access to annotated spectra”.

Protein Inference From Peptide Assignments: Since protein identification experiments that are based on proteolytic digestion and subsequent characterization of the resulting peptides result in the loss of connectivity between these peptides and their protein precursors, identifications based on the assignment of peptide sequences can result in a combination of two possible outcomes: distinct peptides that map to only one protein sequence or peptides that are common to more than one protein sequence (protein group) arising, for example, from alternative splicing. When identifications are of the latter type, authors are required (in addition to the tabular presentation(s) of the data described in guideline II) to:

- Provide accession numbers (or other identifiers) for all proteins that were combined into the group. Authors should justify any cases where a single protein from a protein group has been singled out or when asserting that more than one indistinguishable member of a protein group is actually present.
- Provide a summary list of common peptides belonging to each protein group and those distinct to a specific protein.
- State (and justify) if proteins are identified from a different species than the one being studied. For example, identification of a mouse or human protein in a hamster study.
Peptide and Protein Quantification: Manuscripts presenting quantitative proteomic results from mass spectrometric analyses must provide the following information:

- All relevant quantification data (as part of the peptide and protein identification tables), along with a description of how the raw data were processed to produce these measurements.

- A description of how the analytical reliability of measurements was validated using technical replicates and statistical methods. Citation of standard methods or specialized software may be used. However, it is essential to demonstrate that the data contained in the manuscript actually conform to the same models.

- A description of how the biological reliability of measurements was validated using biological replicates, statistical methods, independent experiments, etc. Studies based on a single biological experiment are generally not acceptable (except as a dataset to test bioinformatic systems). If a biological replicate from the same source cannot be performed (e.g. patient sample), a large enough number of similar biological samples, appropriately justified, must be performed in order to enable sound conclusions.

  Note: For any article reporting quantification of mass spectrometric data, the above three requirements are mandatory and must be included in either the manuscript or supplemental material.

- A description of the treatment of relevant systematic error effects such as interference from overlapping precursor ions, incomplete isotope labeling, bias correction for pipetting error, etc.

- A description of the treatment of random error issues such as outlier rejection and the categorical exclusion of data by thresholds; for example, based on signal to noise or minimum ion counts.

- Proper estimates of uncertainty and the methods used for the error analysis.

Quantification of many proteins or peptides generally results in the need to use some form of multiple hypothesis testing correction. Whenever possible, confidence in protein quantification should be provided for each individual protein rather than the global dataset. Any conclusions drawn or hypothesis generated from the quantitative data in the manuscript must be in concert with the determined estimate of uncertainty.

- If a component is not being identified by database searching in a particular experiment, assurance of the identity of the analyte being measured and the specificity (e.g. presence/absence of interference) with which it is measured must be provided. This
particularly applies to intensity-based methods such as SELDI, selected reaction monitoring / multiple reaction monitoring (SRM/MRM) and accurate mass and retention-time (AMT) based methods.

• A description of the way multiple isoforms in a protein group were quantified.

• For spectral counting measurements, in addition to the above guidelines, additional details should be provided such as whether numbers of peptides or spectra were counted, whether modified peptides, semi-tryptic peptides or shared peptides were counted, and whether or not dynamic exclusion was used, etc.

**Quantification of peptides, modified peptides and proteins using targeted MS approaches.**

III. SUBMISSION OF RAW DATA TO A PUBLIC REPOSITORY

All mass spectrometric output files in the original instrument vendor file format must be deposited, at the time of first submission of the paper, in a publicly accessible site that is independent of the authors' control. Data should be deposited in a manner that requires username and password for access. Access information is to be provided to the Editors at time of submission, and it will be made available to Reviewers as part of the manuscript review process. Data conversion to an open format such as mzML is encouraged if software capable of reading the instrument vendor file format is not widely available. In all cases, the spectra are expected to be provided in a form prior to any processing that might affect the quality of subsequent interpretation as described in the peak list guideline (See section I). The editors of MCP recognize that uploading large datasets can sometimes engender unforeseen difficulties and authors encountering problems should contact the Bethesda office (mcp@asbmb.org) for advice and/or assistance. Authors will not be penalized for delays resulting from such difficulties. Requests for exemptions (or delays not related to technical problems) from this requirement must be made in writing to the Data Management Editor <chalkley@cgl.ucsf.edu> at the time of submission. Embargoed deposits must be made publically available at the time of publication.

e.g., The MS proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifier PXDxxxxxx (submitter to supply the correct number).

Further information regarding this requirement can be obtained by contacting mcp@asbmb.org.