Revised Draft Guidelines for Proteomic Data Publication

July 14, 2005

Dear Colleague:

The attached draft is a proposed set of guidelines, drafted by a group of scientists, engineers and bioinformaticians at a recent workshop held in mid May at the Maison de la Chimie in Paris, France, for preparing, reviewing and publishing data, primarily generated from MS/MS experiments, that deals with the identification of proteins and peptides. The purpose of this endeavor was to formulate standards that would give practioneers and editors alike an appreciation of what a diverse group of stakeholders in this activity considered to be the necessary level of information to reasonably insure the integrity of assignments and thus preserve the accuracy of the scientific record. While the number of participants in the workshop was necessarily small, the number of interested individuals potentially affected is not, and it was the unanimous opinion of all present that this draft should be widely circulated for additional comments, criticism and suggestions. Accordingly, it is being posted, circulated and otherwise distributed as broadly as possible and any recipient should feel to pass it on to any and all parties with an interest in this subject. Comments should be sent to me at where they will be collected and returned to the original working group for consideration, discussion and assimilation. The period of comment will be approximately three months and will end October 15th. The final version, when released by the Committee, will then be circulated to Editors and Publishers of all journals with an interest in proteomics and protein identification for their consideration and, it may be hoped, ratification. It is anticipated that this process will be complete by the end of this calendar year.

By way of added information, *Molecular and Cellular Proteomics* will post these draft guidelines on its web site (www.mcponline.org) along with a list of the Paris participants and the power point presentations made at that meeting that were used as background material for the preparation of the draft. Other individuals and organizations may feel free to do so as well.

We hope that you will take some of your valuable time to look at this draft and to give us the benefit of your wisdom and experience in the form of comments and suggestions. I believe I speak for all those involved when I say that the value of these guidelines will be in improving and strengthening protein identification data and the substantial value that that information will have in future applications. Thus achieving a true community standard is a highly desirable goal.

Sincerely,

Ralph A. Bradshaw

For the workshop participants

DRAFT

Proposed Publication Guidelines for the Analysis and Documentation of Peptide and Protein Identifications The following supporting information should be included with the manuscript:

- The method and/or program (including version number) used to create the "peak list" from the raw data and the parameters used in the creation of this peak list, particularly any that might affect the quality of the subsequent database search. Examples include whether smoothing was applied, any signal-to-noise criteria, whether charge states were calculated or peaks de-isotoped, 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.
- The name and version of the program(s) used for database searching and the values of search parameters. Examples include precursor-ion mass tolerance, fragment-ion mass tolerance, modifications allowed for, any missed cleavages, protein cleavage chemistry (if any), etc.
- The name and version of the sequence database(s) used. If a database was compiled in-house, a complete description of the source of the sequences is required. The number of entries actually searched from each database should be included. Authors should justify the use of a very small database or database that excludes common contaminants, since this may generate misleading assignments.
- Methods used to interpret MS/MS data, thresholds and values specific to judging certainty of identification, whether any statistical analysis was applied to validate the results, and a description of how applied.
- For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, *e.g.* the results of randomized database searches or other computational approaches.

Information for each protein sequence identified should specify the following:

- accession number and database source;
- score(s) and any associated statistical information obtained for searches conducted;

- sequence coverage, expressed as the number of amino acids spanned by the assigned peptides divided by the sequence length;
- the total number of peptides assigned to the protein. To compute this number, different forms of the same peptide are to be counted as one peptide;
- for results from the searching of MS/MS data, the following should be specified for each peptide match:
  - peptide sequence, noting any deviation from the expected protein cleavage specificity;
  - modifications
  - precursor mass, charge and mass error observed;
  - score(s) and any associated statistical information;

Additional potentially valuable information could include the retention time of each peptide, the observation of multiple charge states, multiple observations of the same peptide, flanking residues, start and end positions of peptides in proteins, and any platform-specific information.

Manuscripts presenting studies based on quantitative proteomic data obtained through the use of stable isotope labeling methods or other means such as direct measurement of mass spectrometric signal intensity must contain the following information:

- A thorough description of the experimental design, including the biological sample size and number of technical replicates of such samples or preparations derived thereof so that (bio)statistical methods may be used to assess independently the significance of the results presented. Studies in which the number of biological and/or technical replicates equals one, can generally not be accepted particularly if only few or a single peptide is used for quantification. In exceptional circumstances, other lines of evidence such as time or dose dependent experiments may be acceptable instead of technical replicates.
- A detailed description of the methodology employed for quantification including figures of merit specifically on accuracy, precision, sensitivity, dynamic range and linearity as they apply to a given experimental setting (*e.g.* serum *vs.* tissue analysis). This information may also be provided by referencing an appropriate methods paper. Authors must also specify the acceptance criteria for data from which quantitative information is derived (*e.g.* minimal signal to noise, ion currents, ion counts, chromatographic peak area, etc.). In addition, there must be an appropriate treatment of relevant systematic bias (such as isotope labeling efficiency) and non-systematic bias (such as instrument detector saturation, non-unique peptides, interference of distinct peptides with overlapping isotopes or other known or suspected sources of quantification outliers, etc.). Furthermore, any data transformation or processing must be documented along with the name and version of the employed software tools as well as the parameters used in such data transformations.
- Quantitative information must generally be associated with the identity of a protein unless there is strong supporting evidence that this information is not required for the interpretation of the results obtained. The appropriated guidelines for protein identification should be observed (see points 1–3).
- Reported quantification values must be associated with appropriate measures of confidence (including the number of measurements), uncertainty (error) and reference point (*e.g.* normalization of protein loading across multiple samples, absolute quantification standards, etc.). Particularly, the method of error estimation must be clearly described (*e.g. p*-values, standard deviations, correlation coefficients etc). Where there is no statistical basis to provide meaningful error estimates, representative spectra or ion chromatograms must be provided in the supplementary information.

In order to be of general or specific interest to the scientific community, biological interpretations or hypothesis generated on the basis of quantitative protein changes, must be validated to an appropriate level in the same manuscript. This may take the form of checking for consistency with the scientific literature or by the presentation of further substantiating data at least on a subset of the presented findings.

Authors are also encouraged to refer to two recent reviews by D.F. Ransohoff relevant to the interpretation of proteomic data (*Nat. Rev. Cancer* **4,** 309–314 (2004) and *Nat. Rev. Cancer* **5,** 142–149 (2005).

Authors are also encouraged to make custom software available to the community.

Studies focusing on posttranslational modifications require specialized methodology and documentation to assign the presence and the site(s) of modification. Certain modifications are also nominally isobaric (*e.g.* acetylation *vs.* trimethylation, phosphorylation *vs.* sulfation). If one of these modifications is being reported, then evidence for assigning a specific modification over another needs to be presented. Examples for methods for distinguishing 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 sulfation) or biological or chemical strategies.

In the tabular presentation of the data, authors are required to show 1) the sequence of the peptide used to make each such assignment, together with the amino acids N- and C-terminal to that peptide's sequence, 2) the precursor mass and charge (not just *m/z*) observed, and 3) the search scores for each peptide. Frequently more than one possible site of modification exists within a peptide. Assignment of specific site(s) of modification requires observation of fragment ions

that distinguish between the possible sites. When ambiguity with regard to the modification site cannot be resolved, then the ambiguity must be explicitly shown in the tables, *e.g.* ALEG(sss)YLLK where one of the three serine residues in parentheses is phosphorylated, but information in the spectrum does not permit assignment of a specific residue.

Copies of the annotated, mass labeled spectra for each modified peptide must be submitted electronically together with the manuscript for review purposes. Authors are encouraged to present all or representative spectra of posttranslationally modified peptides either in the body of the text or as supplemental material. In addition, authors are encouraged to provide the corresponding peak (*m/z* and intensity) lists for review.

While more reliable results for peptide identification are generally produced by MS/MS data, in selected circumstances, such as analysis of 2D gel spots, peptide mass fingerprinting can be an effective choice of technique for protein identification. For each identification, an annotated mass spectrum must be supplied. We also encourage the submission of the peak lists for review. In the tabular presentation of the results the authors must supply: 1) the number of matched peaks, 2) the number of unmatched peaks, and 3) the sequence coverage. In addition to the score for the top match they must also show the score for the highest ranked hit to a non-homologous protein. They must describe the parameters and thresholds used to analyze the data (see guideline 1, above), including mass accuracy, resolution, means of calibrating each spectrum, and exclusion of known contaminant ions (keratin, etc.). Authors are required to use and provide the results of scoring schemes that provide a measure of identification certainty, or perform some measure of the false-positive rate.

Identical peptide sequences can be included in multiple unique protein sequences due to biological variation such as single amino acid variants, alternative splice forms, homologs, orthologs and paralogs. Other reasons for apparent redundancy in protein sequence database entries are the inclusion of sequence fragments and sequences with errors. Apparent redundancy can also occur due to clerical errors arising from the merger of multiple sequence databases or identical protein sequences appearing under different names or accession numbers.

Experimental strategies based on proteolytic digestion of protein mixtures introduce the complication of loss of connectivity between peptides and their protein precursors. Assignment of peptide sequences results in two outcomes; *distinct peptides* that map to only one protein sequence or *shared peptides* that map to more than one protein sequence. Detection of shared peptides introduces a paradox between the possibility that a shared peptide can be mapped to more than one protein sequence (bioinformatics redundancy) versus the possibility that more than one precursor is in the original protein mixture (physical redundancy). The apparent ambiguity in peptide assignment requires reporting of a protein group. When assembling peptides into proteins and protein groups, authors should adhere to principles of parsimony, *i.e.* describe the minimum set of protein sequences that adequately accounts for all observed peptides. While the identification of shared peptides implies that multiple related protein sequences are present, the initial assumption should be that only a single form is being detected. Authors should explain and be able to justify cases where a single protein from a protein group has been singled out or that more than one member of a protein group is present. When reporting a summary list of peptides belonging to each protein group, peptides shared among multiple proteins and those unique to a protein should be clearly indicated. In addition, sometimes 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. If such an orthologous protein is included, the circumstances should be mentioned and justified.

It is strongly encouraged (but not yet required) that all MS/MS spectra mentioned in the paper be submitted as supplemental material. Journals will vary in their ability to handle this information and authors are encouraged to provide access to raw MS data using other means, including group websites and public repositories, as they emerge, in addition to the journal itself.