09/02/06

FINAL REPORT

PROPOSED PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS

1. 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.

2. 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, multiple matches to peptides with the same primary sequence count as one, even if they represent different charge states or modification states;
- 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;

3. 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.
4. Manuscripts presenting any conclusions citing quantitative proteomic results should contain the following information:

   The experimental design and data analysis methods should be described in sufficient detail to enable critical assessment of the reliability of the results and conclusions. All relevant quantitative data must be made available. The amount of data and method description is considered sufficient when it is possible to reproduce the reported results or critically reanalyze the data. Citation of standard methods covered in publications or specialized software may be used. However, where any deviations exist and for all methods not covered by citation, authors must thoroughly describe:

   - The methods for how the biological reliability of measurements was validated using biological replicates, statistical methods, independent experiments, etc.
   - The methods for how the analytical reliability of measurements was validated using technical replicates and statistical methods.
   - The treatment of relevant systematic error effects such as peptides shared by multiple proteins, interference from overlapping precursor ions, incomplete isotope labeling, bias correction for pipetting error, etc.

- 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.

- All quantitative results upon which conclusions are based must bear proper estimates of uncertainty and the methods for the error analysis should be clearly described.

The absence of thorough validation of both analytical and biological results, including error analysis should result in rejection. Results with associated protein identifications or post-translational modifications must also follow the respective guidelines.

5. 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 must be presented. 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 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 engine score for this peptide. Frequently more than one possible site of modification exists within a given peptide sequence. Assignment of specific site(s) of modification requires observation of fragment ions that distinguish among 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 Ser residues in parentheses is phosphorylated, but the spectra do not permit assignment of which one). The number of detected modifications in each peptide (e.g, 1, 2 or 3 phosphates) must also be included in the table.

   In all cases involving the assignment of a posttranslational modification(s) , we require that copies of the annotated, mass labeled spectra for those modified peptides be submitted electronically together with the manuscript for review purposes. Authors are required to present representative spectra of posttranslationally modified peptides in the body of the text and the remaining annotated spectra as supplemental material. In addition, authors are encouraged to provide the corresponding peak (m/z and intensity) lists for review.

6. 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.

7. 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 an uncertainty 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 specific 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.

8. 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.