The Need for Guidelines in Publication of Peptide and Protein Identification Data

WORKING GROUP ON PUBLICATION GUIDELINES FOR PEPTIDE AND PROTEIN IDENTIFICATION DATA

Steven Carr‡§, Ruedi Aebersold¶, Michael Baldwin∥, Al Burlingame∥, Karl Clauser**, and Alexey Nesvizhskii¶

Over the past few years, the number and size of proteomic datasets composed of mass spectrometry-derived protein identifications reported in the literature have grown dramatically. This is a direct result of the widespread availability of instruments, methods, and easy-to-use software for collecting large amounts of data and for converting the observed peptide and fragment-ion masses to peptide and then protein identities. In particular, the analysis of samples containing large numbers of proteins by multidimensional liquid chromatography (LC/LC)‡ coupled on-line with tandem mass spectrometry (MS/MS) is now a common component of many biological projects. Clearly it is in the interest of the scientific community to make such data readily available. However, the publication of large proteomic datasets poses new and significant challenges for authors, reviewers, and readers as the need for these guidelines is driven in part by the fact that a significant but undefined number of the proteins being reported as “identified” in proteomics articles are likely to be false positives (2). These incorrect matches probably result most often from the use of low-quality peptide MS/MS data to search the database. However, even high-quality data can produce invalid identifications if, for example, the actual peptide sequence is not in the database being searched. Many different algorithms are being used for peptide and protein assignment (e.g. MSTag, Mascot, SEQUEST, SpectrumMill, Sonar, etc.), and each has unique rules for scoring to move the most probable peptide assignment to the top of the “hit” list. In addition, new filtering criteria are being developed that, when layered onto the results from the above algorithms, help to eliminate a certain additional percentage of false positives (3, 4). It is very important that the users of these tools, our authors, have at least a working understanding of how the algorithm they use works. However, even the judicious use of scoring, threshold parameters, and additional filtering criteria for search engines, while serving the very important purpose of reducing the number of misassigned peptides and proteins, does not eliminate the problem. It is almost always possible to match a MS/MS spectrum to a peptide in the database; the difficult part is validating that the match is correct.

This is not to imply that the situation is bleak. In fact, most assignments of proteins made with high-quality data and using more than a single peptide to identify a protein are likely correct. Furthermore, improved methods are being developed at a rapid pace. Recently, application of statistical methods to validate peptide assignments to MS/MS spectra of peptides has been shown to be a promising approach, and a number of groups are working in this area (2, 5–11). However, these programs are only beginning to be widely used and they are not universally accepted. MCP fully supports continued development and testing of such programs and will publish new search and filtering approaches to make them widely available to the proteomics community. However, in the absence of accepted standards and widely available tools that operate on such standards, there are guidelines that the journal can formulate that would help ensure the publication of high-quality information and to assist readers in being able to make their own assessment of the validity of the assignments in manuscripts. Thus, we introduce the initial set of such guidelines in this issue of the journal. The rationale and purpose of each of these is described below.

The first (and obvious) guideline is to obtain sufficient information from authors to document what search engine was used and how peptide and protein assignments were made using that software.

Guideline 2 defines how peptides should be counted toward the identification of a protein. We do not at this time attempt to deal with the related issue of what constitutes a “unique” peptide with respect to the proteins identified. For

From the ‡Broad Institute, Cambridge, MA 02139; ¶Institute for Systems Biology, Seattle, WA 98103; §§University of California, San Francisco, CA 94143; and **Millennium Pharmaceuticals, Cambridge, MA 02139

Received, April 7, 2004

Published, MCP Papers in Press, April 8, 2004, DOI 10.1074/mcp.T400006-MCP200

1 The abbreviations used are: LC/LC, multidimensional liquid chromatography; MS/MS, tandem mass spectrometry; MCP, Molecular & Cellular Proteomics.

© 2004 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

Molecular & Cellular Proteomics 3.6 531
example, the situation arises with respect to how to use peptides matching to one member of a protein family but not to any other. We are working toward a better definition of both the problem and possible ways for how to deal with this issue (also see explanation of guideline 6, below).

Guidelines 3 and 4 relate to the fact that, regardless of the search engine employed, the risk of a false-positive protein assignment is greater when only a single peptide is used to identify a protein than when multiple peptides (each satisfying the criteria for a good match in the given software) are used to make an identification by database searching (10). Therefore we are increasing the stringency of information required to use single-peptide identifications for protein assignment. This change is not meant to signal that proteins assigned with two or more peptide are automatically correct, only that there is a significantly higher potential for single-peptide assignments to be wrong.

While the accepted standard for peptide identification is now sequence information obtained by MS/MS data, a significant portion of the proteomics community continues to employ peptide mass fingerprinting data for peptide identification. Guideline 5 addresses this type of data and increases the stringency for its acceptance.

Guideline 6 addresses the present difficulties search engines have with counting the number of unique proteins identified based on the peptides found. At present, there is no agreed upon approach. The issue is that essentially the same protein appears in many cases under different names and accession numbers in the databases. In some cases, this is due to redundant entries in the database being searched. However, some of this apparent “redundancy” is biological in that many genomes have multiple copies of similar genes as well as splice variants. In both cases, multiple proteins with similar (if not identical) sequences are identified by the search engine using subsets of the same group of mass spectra. This is a difficult problem for which there is no ideal solution at present. However, it is possible to formulate a practical definition of “similar” sequences and to then group proteins together according to the spectra used to identify them. This guideline addresses this issue.

Achieving the ultimate goal of publishing only high-quality datasets with small and known false-positive rates will require new data analysis tools and methods. As such tools become available, published results will be subject to reanalysis and interpretation. This is a healthy situation for the field, which MCP will assist by moving in the future to require that authors submit their data (in a suitable format) as a condition for acceptance of their manuscript. The goal is to make openly available the MS and LC-MS/MS data in published studies (in a suitable form and with necessary tools) to enable reanalysis, data mining, and development of improved algorithms. The guidelines introduced in this issue are a start. We consider this a work in progress and welcome your comments.

Acknowledgments—We wish to acknowledge the thoughtful contributions of Robert Chalkley, Kirk Hansen, Kati Medzhibrodsky (University of California, San Francisco), Andrew Keller (Institute for Systems Biology), and Ron Beavis (Beavis Informatics, Ltd.).

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: Steven Carr, Director of Proteomics, Broad Institute, 320 Charles Street, Cambridge, MA 02139. Tel.: 617-324-1483; E-mail: scar@broad.mit.edu.

REFERENCES

1. Baldwin, M. (2004) Protein identification by mass spectrometry: Issues to be considered. Mol. Cell. Proteomics 3, 1–9
2. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392
3. Petrits, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., Pasa-Tolic, L., Lipton, M. S., Aubeny, K. J., Strittmatter, E. F., Shen, Y., Zhao, R., and Smith, R. D. (2003) Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. Anal. Chem. 75, 1039–1048
4. Cargile, B. J., Bundy, J. L., Freeman, T. W., and Stephenson, J. L. (2004) Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J. Proteome Res. 3, 112–119
5. MacCoss, M. J., Wu, C. C., and Yates, J. R. (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal. Chem. 75, 5593–5599
6. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome Res. 2, 137–146
7. Fanyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal. Chem. 75, 768–774
8. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. Mol. Cell. Proteomics 2, 96–108
9. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658
10. Nesvizhskii, A. I., and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discov. Today 9, 173–181
11. Sadeghpour, R. G., Liu, H., and Yates, J. R. (2004) Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal. Chem. 76, 1664–1671

GUIDELINES FOR THE PUBLICATION OF PEPTIDE AND PROTEIN IDENTIFICATION DATA IN MOLECULAR & CELLULAR PROTEOMICS

The following guidelines have been developed to specifically address problems associated with articles containing peptide and protein identifications. These guidelines are under development and will likely undergo revisions as experience and the supporting technology develop further. Authors with questions regarding the guidelines are encouraged to contact the Editors prior to submission of their manuscripts for additional discussion and/or clarification.

Manuscripts containing protein identifications based on fragmentation and database searching must provide:

1. The following supporting information:

   • The method and/or program used to create the “peak list” from raw data and the parameters used in the cre-
1. Information regarding the sequence coverage observed for each protein should be provided either in the manuscript or in the supplementary materials. The authors are encouraged to provide a table that lists for each protein the sequences of all identified peptides. At the minimum, the total number of peptides belonging to each protein must be explicitly stated either in the text or in tables presented. To compute this number, different forms of the same peptide are to be counted as only a single peptide. For example, if the same peptide is identified in both $2^+ \text{ and } 3^+$ charge forms, the number of interpreted spectra equals 2, but the count of identified peptides that count toward the protein sequence coverage measure is only 1. Similarly, if multiple forms of a peptide are identified that arise through common sample-handling artifacts (e.g., oxidation of Met and deamidation of Asn), then the count of peptides identified again equals 1. The total number of MS/MS-interpreted spectra assigned to peptides corresponding to each protein can also be provided, but this should not be confused with the protein sequence coverage measure described above.

2. Protein assignments based on single-peptide assignments must include additional information in the table(s). Specifically, we require authors 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 scores for this peptide (in the case of multiple MS/MS spectra assigned the same peptide, this information should be provided only for the best assignment). As noted in guideline 2, assignments of the same peptide to multiple MS/MS spectra of the same charge form are to be counted as single-peptide assignment.

3. In cases where biological conclusions are based on observation of a single peptide matching to a protein or a posttranslationally modified form of that protein, then this identification must be supported by inclusion (in the main body or in supplemental materials) of the corresponding MS/MS spectrum appropriately labeled.

4. Peptide mass fingerprint data will continue to be accepted for peptide identification, but the standard of acceptability will be more stringent than currently allowed. In addition to listing the number of masses matched to the identified protein, authors should also state the number of masses not matched in the spectrum and the sequence coverage observed. 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 encouraged to use and provide the results of scoring schemes that provide a measure of identification certainty, or perform some measure of false-positive rate.

5. Essentially the same protein appears in many cases under different names and accession numbers in the database. When matching peptides to members of such a family, it is the authors’ responsibility to demonstrate that they are aware of the problem and have taken reasonable measures to eliminate redundancy. In cases where a single-protein member of a multi-protein family has been singled out, the authors should explain how the other members of the group were ruled out, if at all. In addition, sometimes proteins are identified from a different species than the studied one. For example, mouse or human protein in a hamster study. If such a protein is included, it also has to be mentioned and justified.

6. MCP strongly encourages (but does not at present require) the submission of all MS/MS spectra mentioned in the paper as supplemental material. We will accept .dta, .pkl, and .mgf files. Because technical aspects of storing large repositories of raw mass spectrometric data has yet to be worked out, MCP cannot at present accept any such data. However, authors are encouraged to provide access to raw MS data using other means, including group websites and public repositories as they emerge.