Universal Metrics for Quality Assessment of Protein Identifications by Mass Spectrometry*

David A. Stead‡, Alun Preece§, and Alistair J. P. Brown‡¶

Increasing numbers of large proteomic datasets are becoming available. As attempts are made to interpret these datasets and integrate them with other forms of genomic data, researchers are becoming more aware of the importance of data quality with respect to protein identification. We present three simple and universal metrics that describe different aspects of the quality of protein identifications by peptide mass fingerprinting. Hit ratio gives an indication of the signal-to-noise ratio in a mass spectrum, mass coverage measures the amount of protein sequence matched, and excess of limit-digested peptides reflects the completeness of the digestion that precedes the peptide mass fingerprinting. Receiver-operating characteristic plots show that the novel metric, excess of limit-digested peptides, can discriminate between correct and random matches more accurately than search score when validating the results from a state-of-the-art protein identification software system (Mascot) especially when combined with the two other metrics, hit ratio and mass coverage. Recommendations are made regarding the use of the metrics when reporting protein identification experiments. Molecular & Cellular Proteomics 5:1205–1211, 2006.

Peptide mass fingerprinting (1) is a widely used protein identification technique that has been automated and applied to protein analysis on a global scale, in particular for the identification of proteins in 2-D1 gel spots which generally contain single proteins or simple mixtures thereof (2). This typically involves in-gel digestion of the proteins with trypsin, extraction of the peptides, and then accurate measurement of their masses by MALDI-TOF MS. To identify the proteins from which the peptides are derived, the list of experimental peptide masses is matched against those predicted from a protein sequence database using knowledge of the specificity of the cleavage reagent and any amino acid modifications that might be expected. It has been suggested that a significant number of protein identifications reported in proteomic studies may be inaccurate because of the low quality MS data used to obtain these identifications (3). This criticism was aimed at large scale LC-MS/MS experiments, but there is no reason to suppose that the same should not apply to peptide mass fingerprinting experiments.

Various software tools are available for obtaining protein identifications from sequence databases using mass spectrometric data, including Mascot (4), MS-Fit (5), ProFound (6), and SEQUEST (7). The scores generated by the various search algorithms for ranking the matches are calculated in different ways, and the results are not directly comparable (Table I). It would be helpful, therefore, to have a generic measure of the quality of protein identifications when assessing the reliability of results reported by different proteomic studies. Such a universal quality indicator could be helpful to users of proteomic data repositories who may wish to gain a rapid estimate of the reliability of the protein identifications contained within large datasets.

Efforts are being made to standardize the way in which results of proteomic experiments are reported. This began with the Proteome Experimental Data Repository (PEDRo) (8) and is currently being driven forward by the Human Proteome Organisation-Proteomics Standards Initiative (HUPO-PSI) (psidev.sourceforge.net/gps/index.html). However, agreement on what constitutes such a standard is still some way off, and in the meantime the journal Molecular and Cellular Proteomics has proposed guidelines to “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 [protein] assignments in manuscripts” (3). For peptide mass fingerprinting, these guidelines suggest reporting the number of peptide masses matched to the identified protein, the number of masses not matched in the spectrum, and the sequence coverage observed. These metrics are readily available and are independent of the search algorithm used.

However, with respect to the establishment of universal reporting standards, there remains a need to demonstrate the utility of any metrics that are proposed for the formal description of the quality of a protein identification result. This was precisely the purpose of this study. We did not attempt to evaluate the various search algorithms (9), but rather we aimed to provide metrics that can be universally applied to the
Peptide Mass Fingerprinting Quality Metrics

TABLE I

<table>
<thead>
<tr>
<th>Mass list no. and description</th>
<th>Masses matched</th>
<th>Masses submitted</th>
<th>Coverage %</th>
<th>Search score (rank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Poor or marginal match</td>
<td>4</td>
<td>15</td>
<td>48</td>
<td>Mascot: 61 (1) MS-Fit: 101 (1) ProFound: 292 (1)</td>
</tr>
<tr>
<td>2) Moderately good match</td>
<td>9</td>
<td>19</td>
<td>23</td>
<td>Mascot: 101 (1) MS-Fit: 2.4e+04(1) ProFound: 7.9e+13(1)</td>
</tr>
<tr>
<td>3) Extremely good match</td>
<td>21</td>
<td>23</td>
<td>41</td>
<td>Mascot: 292 (1) MS-Fit: 7.9e+13(1) ProFound: 1.0 (1)</td>
</tr>
</tbody>
</table>


Identification of proteins using MS data relies on the ability of the search algorithm to discriminate between correct and random matches from a sequence database. The performance of a discriminatory test may be expressed in terms of its sensitivity and specificity, which are measures of the ability of the test to detect what it is testing for (the true positives) and specificity is the ability of the test to reject what it is not testing for (the true negatives) (Table II). The receiver-operating characteristic (ROC) technique (10, 11) plots the fraction of true positive results against the fraction of false positive results for all possible threshold values (Fig. 1). If a number of tests (and therefore ROC curves) are to be compared, the area under the curve (AUC) is a useful discriminator. An AUC value of 1.0 indicates a perfect test in which there is no overlap in the distributions of the group scores. An AUC value of 0.5 (Fig. 1, diagonal) indicates that the scores for the two groups do not differ and that the test is unable to discriminate between the two conditions. We used the ROC technique to test the ability of various metrics resulting from a peptide mass fingerprint search to discriminate between correct and random protein matches and to optimize their combination in novel ways using a training dataset. The validity of the metrics was then confirmed using three independent datasets.

Experimental Procedures

**Training Dataset**—The data used to optimize the quality metrics came from 329 peptide mass fingerprinting experiments performed by Aberdeen Proteomics Consortium for the Functional Genomics of Microbial Eukaryotes to identify fungal proteins in 2-D gel spots (www.abdn.ac.uk/proteomics/). Total soluble proteins were extracted from various fungal cells in culture, the proteins were separated by 2-D PAGE, and selected spots were identified by peptide mass fingerprinting essentially as described previously (12). In each experiment a monoisotopic mass list, with known systematic peaks removed, was submitted to a Mascot peptide mass fingerprint search over a fungal sequence database, typically NCBInr. Default parameters used in the searches were: taxonomy = fungi; enzyme = trypsin; fixed modifications = carbamidomethyl (Cys); variable modifications = oxidation (Met); peptide tolerance = 50 ppm; mass values = MH⁺, monoisotopic.

**Test Dataset 1**—Initial validation of the quality metrics used data from 96 peptide mass fingerprinting experiments to identify Streptomyces coelicolor proteins in 2-D gel spots performed by an independent laboratory. These data, which were kindly provided by Andrew Hesketh (John Innes Centre), are deposited in the public repository, PEDRoDB (13), and are available for download from pedrodb.man.ac.uk:8080/pedrodb/pages/Browse.jsp. Mass lists (from which trypsin autolysis peaks had already been removed) were extracted under the curve (AUC) is a useful discriminator. An AUC value of 0.5 (Fig. 1) indicates that the scores for the two groups do not differ and that the test is unable to discriminate between the two conditions. We used the ROC technique to test the ability of various metrics resulting from a peptide mass fingerprint search to discriminate between correct and random protein matches and to optimize their combination in novel ways using a training dataset. The validity of the metrics was then confirmed using three independent datasets.

**Training Dataset**—The data used to optimize the quality metrics came from 329 peptide mass fingerprinting experiments performed by Aberdeen Proteomics Consortium for the Functional Genomics of Microbial Eukaryotes to identify fungal proteins in 2-D gel spots (www.abdn.ac.uk/proteomics/). Total soluble proteins were extracted from various fungal cells in culture, the proteins were separated by 2-D PAGE, and selected spots were identified by peptide mass fingerprinting essentially as described previously (12). In each experiment a monoisotopic mass list, with known systematic peaks removed, was submitted to a Mascot peptide mass fingerprint search over a fungal sequence database, typically NCBInr. Default parameters used in the searches were: taxonomy = fungi; enzyme = trypsin; fixed modifications = carbamidomethyl (Cys); variable modifications = oxidation (Met); peptide tolerance = 50 ppm; mass values = MH⁺, monoisotopic.

**Test Dataset 1**—Initial validation of the quality metrics used data from 96 peptide mass fingerprinting experiments to identify Streptomyces coelicolor proteins in 2-D gel spots performed by an independent laboratory. These data, which were kindly provided by Andrew Hesketh (John Innes Centre), are deposited in the public repository, PEDRoDB (13), and are available for download from pedrodb.man.ac.uk:8080/pedrodb/pages/Browse.jsp. Mass lists (from which trypsin autolysis peaks had already been removed) were extracted.
from the XML (extensible markup language) file and submitted to Mascot peptide mass fingerprint searches using the following default parameters: database = NCBInr November 5, 2005 (3,021,705 sequences); enzyme = trypsin; fixed modifications = carbamidomethyl (Cys); variable modifications = oxidation (Met); peptide mass tolerance = 75 ppm; number of missed cleavages = 1. The two groups for the ROC analysis were generated as described above for the training dataset.

Test Dataset 2—Further validation used data from 44 peptide mass fingerprinting experiments that identified unique proteins from *Clostridium difficile* in 2-D gel spots. These data were kindly provided by Anne Wright and Neil Fairweather (Imperial College London) and Jerry Thomas (University of York) and had been acquired using a MALDI-TOF-TOF instrument (Applied Biosystems 4700 Proteomics Analyzer). The correct protein identifications were defined by Mascot MS/MS ion searches using tandem MS data collected concurrently. Mascot PMF searches were performed over a database generated from the current preliminary gene prediction (European Molecular Biology Laboratory format) for *C. difficile* (3676 sequences). These sequence data were produced by the *C. difficile* Sequencing Group at the Sanger Institute and can be obtained from ftp.sanger.ac.uk/pub/pathogens/CD. The Mascot search parameters used were: taxonomy = none; enzyme = trypsin; maximum number of missed cleavages = 1; fixed modifications = carbamidomethyl (Cys); variable modifications = oxidation (Met); mass filter = none; peptide mass tolerance = 50 ppm.

**RESULTS AND DISCUSSION**

A peptide mass fingerprint search returns a list of candidate matches, each of which displays a number of characteristics that may be used to rank the results and to inform a decision as to whether or not the protein has been correctly identified. The search scores take some of these characteristics into account, but different algorithms calculate the scores in different ways (that are less than transparent to the user), making it impossible to compare them directly without rerunning the searches. With this in mind, we hypothesized that characteristics present in all peptide mass fingerprinting search reports could be used to compare the quality of the resulting protein identifications. Such generic metrics include the number of masses matched, the number of masses not matched in the spectrum, and the sequence coverage, all of which have been suggested as minimum reporting requirements for protein identifications by peptide mass fingerprinting (3). Another validation method is to compare the peptide mass errors with a calibration error curve for the mass spectrometer (14). Given that MALDI-TOF instruments can measure masses with high precision, but their calibration may vary from spot to spot, correctly assigned peak masses should tightly fit the calibration error curve. Some search engines, e.g. *Aldente* (www.expasy.org/tools/aldente/) (15) and AutoMS-Fit (specifically the IntelliCal™ feature provided with Protein Solution 1) (Applied Biosystems), incorporate algorithms that correct for such systematic errors.

We calculated the HR using the number of masses matched and the number of masses not matched in the spectrum. This created a metric that estimates the protein signal-to-noise ratio.
ratio in the mass spectrum. A poor quality spectrum with a low signal-to-noise ratio would be expected to contain many unmatchable peaks, whereas it should be possible to eliminate much of the noise from a good quality spectrum by appropriate use of the peak detection threshold. To test this we generated ROC plots for the hit ratio using the test datasets (Figs. 2a, 3a, and 4a). The hit ratio was found to be a moderately good discriminator of correct and random database matches.

Using ROC analysis we compared the discriminatory ability of sequence coverage with that of the hit ratio. Sequence coverage did not perform as well as the hit ratio (Figs. 2a, 3a, and 4a). Sequence coverage by itself has limited value for assessing the quality of protein identifications because the percentage of sequence coverage that can be achieved by peptide mass fingerprinting is inversely proportional to the size of the protein. However, this relationship can be normalized by multiplying the sequence coverage by the predicted protein molecular weight to give what we term the MC. It should also be noted that sequence coverage measurements depend on the accuracy of the protein sequence in the database. Mature proteins may be shorter than the gene sequence predicts. However, such a discrepancy would be apparent from the coverage map, for example as shown in the Protein.
View of a Mascot report. Mass coverage describes a different aspect of the quality of a protein identification, namely the amount of protein (in kDa) that is covered by the matched peptides. This metric is related to hit ratio because the value of both metrics increases as the number of matched peptides increases. ROC analysis revealed that MC performed better than HR for discriminating between correct and random matches (Figs. 2a, 3a, and 4a).

In processing samples (typically 2-D gel spots) for peptide mass fingerprinting, the experimenter aims to perform a limit digest of the proteins with a cleavage reagent of known specificity (e.g. trypsin). Should this be achieved, none of the peptides would be expected to contain a missed cleavage site. Using this knowledge, we invented a novel metric termed the ELDP that is calculated by subtracting the number of matched peptides that contain a missed cleavage site from the number of matched peptides that do not. Therefore, the greater the value for ELDP the more complete the digestion (other factors being equal). We predicted that genuine matches would have a positive ELDP value (limit-digested peptides in excess), whereas incorrect matches would not show any bias in this direction and in practice will tend to give a negative ELDP value (incompletely or randomly digested peptides in excess). Our analyses of the training and test datasets showed that this prediction was correct (Fig. 5a). It is important to note that the input parameters for the search should not specify the maximum number of missed cleavages as zero, otherwise the diagnostic power of the ELDP metric is lost. It is recommended that the maximum number of allowable missed cleavages (per peptide) be set to one for all peptide mass fingerprinting searches unless there is good reason to do otherwise.

The novel metric ELDP discriminated accurately between correct and random matches in the ROC analyses even when used alone (Figs. 2b, 3b, and 4b). Its diagnostic performance was better than the Mascot score (Fig. 5b), which is widely regarded as the “industry standard” software system for protein identification. When we combined ELDP with HR and MC to create the PMF score, the discriminatory power of ELDP was increased still further, resulting in a ROC curve for PMF score approaching that of an ideal test (AUC = 1.0) (Figs. 2b, 3b, and 4b and Table III). For example, using all four datasets (n = 581), a cutoff for PMF score of 79 gave a specificity of 99% (five false positives) and a sensitivity of 78% (128 false negatives).

Given the exceptional utility of ELDP for describing the quality of protein identifications by peptide mass fingerprinting, we recommend that any reporting standards proposed for proteomic experiments, such as the Minimum Information about a Proteomics Experiment (MIAPE) (psidev.sourceforge.net/gps/index.html), include all the information required to calculate this metric. This requires (a) the number of peptides matched and (b) the number of these peptides that contain a missed cleavage site or alternatively (c) the number of missed cleavage sites in each matched peptide. Ideally all of the quality metrics described above (HR, MC, and ELDP) would be reported by future versions of the search algorithms, and a universal scoring system, perhaps based on the PMF score, would be adopted throughout the field. Furthermore these metrics have utility outside peptide mass fingerprinting experiments. The metrics described above can be calculated equally well for protein identifications based on tandem mass spectrometric data (peptide fragment fingerprinting), the only difference being that each peptide has been matched on the basis of its fragmentation pattern rather than by its intact
Fig. 5. Distributions of the ELDP values and Mascot scores for all protein identifications. a, frequency of ELDP value returned by correct (gray bars) and incorrect (black bars) matches. The minimum, maximum, and mean ELDP values in the combined datasets (n = 581) were -4, 20, and 6.4 for correct matches and -8, 4, and -1.3 for incorrect matches, respectively. b, frequency of Mascot score returned by correct (gray bars) and incorrect (black bars) matches. The minimum, maximum, and mean Mascot scores in the combined datasets (n = 581) were 25, 354, and 113.7 for correct matches and 21, 75, and 44.3 for incorrect matches, respectively.
mass. We envisage a bioinformatics tool that provides a graphical output of the metrics for browsers of proteomic data repositories that can facilitate comparisons of the quality of different datasets and allow the user to set acceptance criteria (e.g. PMF score >79) for individual protein identifications.

Acknowledgments—We give special thanks to Laura Selway, Jan Walker, and Zhikang Yin (Aberdeen Proteomics, Consortium for the Functional Genomics of Microbial Eukaryotes) for major contributions toward the generation of the fungal datasets. We thank Andrew Hesketh (John Innes Centre, Norwich, UK) for making the *S. coeli-
color* dataset available in PEDRo and acknowledge the work of Norman Paton and colleagues at the University of Manchester in developing and maintaining the PEDRo database. We are indebted to Anne Hesketh (John Innes Centre, Norwich, UK) for making the *S. coeli-
color* dataset available in PEDRo and acknowledge the work of Norman Paton and colleagues at the University of Manchester in developing and maintaining the PEDRo database. We are indebted to Anne Wright and Neil Fairweather (Centre for Molecular Microbiology and Infection, Department of Biological Sciences, Imperial College Lon-
don), to Jerry Thomas (Proteomics Laboratory, Technology Facility, Department of Biology, University of York), and to Hanjo Lim (Sanofi- Aventis US, Bridgewater, NJ) and John Yates III (The Scripps Research Institute, La Jolla, CA) for providing the *C. difficile* and *M. jannaschii* datasets respectively, without which this work would not have been possible. Finally we also thank Binling Jin (Computing Science, University of Aberdeen) for stimulating discussions of data quality concepts in proteomics within the Qurator project (www.
quator.org).

* This work was supported by UK Biotechnology and Biological Sciences Research Council Grants BBS/B/13764, BBS/B/06679, BB/ C510391/1, and BB/C501176/1 and Engineering and Physical Sciences Research Council Grants GR/S67609/01. 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. Tel.: 44-1224-555883; Fax: 44-1224-555844; E-mail: al.brown@abdn.ac.uk.

REFERENCES


