MIAPE: Mass Spectrometry Informatics

Pierre-Alain Binz[1,2]*, Robert Barkovich[3], Ronald C. Beavis[4], David Creasy[5], David M. Horn[6], Randall K. Julian Jr.[7], Sean L. Seymour[8], Chris F. Taylor[9], Yves Vandenbrouck[10]

[1] Swiss Institute of Bioinformatics, Rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland
[2] GeneBio SA, Av. de Champel 25, Geneva, Switzerland
[3] Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051, USA
[4] Biomedical Research Centre, University of British Columbia, Vancouver, BC, Canada. V6T 1Z3
[5] Matrix Science Ltd, 64 Baker Street, London W1U 7GB, UK
[6] Agilent Technologies, 5301 Stevens Creek Blvd., Santa Clara, CA 95051, USA
[7] Indigo BioSystems, Inc., Indianapolis, IN, USA
[8] Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA
[9] European Bioinformatics Institute, Hinxton, UK
[10] CEA, DSV, DRDC, laboratoire de Biologie, Informatique et Mathématiques, Grenoble, F-38054 France

* Corresponding author

Abstract

"MIAPE - Mass Spectrometry Informatics" is one module of the Minimal Information About a Proteomics Experiment (MIAPE) documentation system. MIAPE is developed by the Proteomics Standards Initiative of the Human Proteome Organisation (HUPO-PSI). It aims at delivering a set of guidelines representing the minimal information required to report and sufficiently support assessment and interpretation of a proteomics experiment. This MIAPE – Mass Spectrometry Informatics module is the result of a joint effort between the Mass Spectrometry group of HUPO-PSI, the Proteomics Informatics group of HUPO-PSI and the proteomics community. It has been designed to specify a minimal set of information to document a mass spectrometry-based peptide and protein identification and characterization experiment. As for all MIAPE documents, these guidelines will evolve and will be made available on the PSI website at the url http://psidev.info.

MIAPE: Mass Spectrometry Informatics

Version 1.1, 5th February, 2008.

This module identifies the minimum information required to report the use of protein and peptide identification and characterisation software to analyse the data produced by mass spectrometry experiments, sufficient to support both the effective interpretation and assessment of the data and the potential recreation of the work that generated it.

Introduction

This document is one of a collection of technology-specific modules that together constitute the Minimum Information about a Proteomics Experiment (MIAPE) reporting guidelines produced by the Proteomics Standards Initiative. MIAPE is structured around a parent document that lays out the principles to which the individual reporting guidelines adhere. In brief, a MIAPE module represents the minimum information that should be reported about a data set or an experimental process, to allow a reader to interpret and critically evaluate the conclusions reached, and to support their experimental corroboration. In practice a MIAPE module comprises a checklist of information that should be provided (for example about the protocols employed) when a data set is submitted to a public repository or when an experimental step is reported in a scientific publication (for instance in the materials and methods section). The MIAPE modules specify neither the format that information should be transferred in, nor the structure of the repository/text. However, PSI is not developing the MIAPE modules in isolation; several compatible data exchange standards are now well established and supported by public databases and data processing software in proteomics (for details see the PSI website www.psidev.info).

The correct analysis of the data produced by mass spectrometry is key to the generation of reliable biological knowledge. Peptide Mass Fingerprints and Peptide Fragment Fingerprints are two types of data that can be used in peptide and protein identification, quantitation, structural characterisation and the investigation of protein modifications. The heterogeneity of sample content and complexity on one hand, and the heterogeneity of mass spectrometers (type, sensitivity, accuracy, efficiency of sample introduction, ionisation, processing) on the other hand have a strong impact on the type, amount and quality of experimental information to be analysed by protein and peptide identification and characterisation software. This is highlighted by the complexity of input parameters required by such software; the results generated are similarly complex in terms of both data structure and pertinence. These guidelines for the reporting of the use of such software do not prescribe that all of that information be captured; and given the diversity of tools currently available, the utility of such detail is clearly open to question.
However, it is possible to specify (generic) parameters that are representative of the way in which the software was used, and to contextualise the data generated; enabling a better-informed process of assessment and interpretation. For a full discussion of the principles underlying this specification, please refer to the MIAPE parent document, which can be found on the MIAPE website http://psidev.info.

These guidelines cover the use of protein and peptide identification and characterisation software and the data generated. They do not cover the mass spectrometry that generated the data, or the reduction of 'raw' profile data to peak lists; these details are captured in separate MIAPE modules.
Note also that these guidelines do not cover all the available features of a protein and peptide identification and characterisation tool (for example, some of the less frequently used parameters, types of spectra or other experimental data). Items falling outside the scope of this module may be captured in complementary modules, which can be obtained from the website. Note that subsequent versions of this document

may have altered scope, as will almost certainly be the case for all the MIAPE modules.

The following section, detailing the reporting guidelines for the use of protein and peptide identification and characterisation software, is subdivided as follows:

1. General features; the software employed.
2. Input data and parameters.
3. The output from the procedure; the list of peptides and proteins identified, characterised or quantified.
4. Interpretation and validation.

Reporting guidelines for protein and peptide identification and characterisation software

1. General features

a) Global descriptors
- Date stamp (as YYYY-MM-DD)
- Responsible person (or institutional role if more appropriate); provide name, affiliation and stable contact information
- Software name, version and manufacturer
- Customisations made to that software
- Availability of that software
- Location of the files generated; parameter files, spectral data (input/output)

2. Input data and parameters

a) Input data
- Description and type of MS data
- Availability of MS data (source of data, file format)

b) Input parameters
- Databases queried; description and versions (including number of entries searched)
- Taxonomical restrictions applied
- Description of tool and scoring scheme
- Specified cleavage agent(s)
- Allowed number of missed cleavages
- Additional parameters related to cleavage
- Permissible amino acids modifications
- Precursor-ion and fragment ion mass tolerance for tandem MS (when applicable)
- Mass tolerance for PMF (when applicable)
- Thresholds; minimum scores for peptides, proteins (probabilities, number of hits, other metrics)
- Any other relevant parameters

3. The output from the procedure

The procedure might generate all or part of the elements described below (identified proteins, identified peptides, quantization information). Select the elements that apply.

a) For identified proteins
- Accession code in the queried database
- Protein description
- Protein scores
- Validation status
- Number of different peptide sequences (without considering modifications) assigned to the protein
- Percent peptide coverage of protein
- Identity of supporting peptides
- In the case of PMF, number of matched/unmatched peaks

b) For identified peptides
- Sequence (indicate any deviation from the expected protein cleavage specificity)
- Peptide scores
- Chemical modifications (artefactual) and post-translational modifications (naturally-occurring); sequence polymorphisms with experimental evidence (particularly for isobaric modifications)
- Corresponding spectrum locus
- Charge assumed for identification and a measurement of peptide mass error
- Other additional information, when used for evaluation of confidence

c) Quantitation for selected ions
- Quantitation approach (*e.g.* 4plex-iTRAQ, ICAT, cICAT, COFRADIC)
- Quantity measurement (*e.g.* integration of signals, use of signal intensity)
- Data transformation and normalisation technique (description of method and software)
- Number of replicates (biological and technical)
- Acceptance criteria (including measure of errors)
- Estimates of uncertainty and the methods for the error analysis, including the

treatment of relevant systematic error effects and the treatment of random error issues Results from controls (when described)

4. ***Interpretation and validation***
   - Assessment and confidence given to the identification and quantitation (description of methods, thresholds, values, etc,)
   - Results of statistical analysis or determination of false positive rate in case of large scale experiments
   - Inclusion/exclusion of the output of the software are provided (description of what part of the output has been kept, what part has been rejected)

Summary

The MIAPE: Mass Spectrometry Informatics minimum reporting guidelines for the use of protein and peptide identification and characterisation software specify that a significant degree of detail be captured, for peptide mass fingerprinting experiments as well as peptide fragment fingerprinting processes. However, it is clear that providing the information required by this document will enable the effective interpretation and assessment of the usage of such software, the qualitative and quantitative assignment of proteins and peptides and potentially, support experimental corroboration. Much of the information required herein may already be stored in an electronic format, or exportable from the instrument; we anticipate further automation of this process.

These guidelines will evolve. To contribute, or to track the process to remain 'MIAPE-compliant', browse to the website at http://psidev.info

Appendix One. The MIAPE-MSI glossary of required items

Classification	*Definition*
1. General features — (a) Global descriptors	
Date stamp	The date on which the work described was initiated; given in the standard 'YYYY-MM-DD' format (with hyphens).
Responsible person or role	The (stable) primary contact person for this data set; this could be the experimenter, lab head, line manager *etc*.. Where responsibility rests with an institutional role (*e.g.* one of a number of duty officers) rather than a person, give the official name of the role rather than any one person. In all cases give affiliation and stable contact information.
Software name, version and manufacturer	For each software used: The trade name of the software used for the identification and/or characterization work, together with the version name or number according to the vendor's nomenclature and vendor name.
Customisations	Any (*i.e.* affecting behaviour) modification made to the original code and functionality of the software.
Availability of the software	The references of the vendor or public url if a publicly available version has been used.
Location of the files generated	The location of the data generated. If made available in a public repository, describe the URI (for instance an url, or the url of the repository and the information on how to retrieve the data). If not made available for public access, describe the contact person reference or source and the internal coordinates of the data.
2. Input data and parameters – (a) input data	
Description and type of MS data	Provide a short description that can refer to the data in the experiment (e.g. LC-MS run1). Specify if the submitted data are raw full traces (either as proprietary binary format or exported readable format) or if they are peaklists files. In all cases, specify the original format (i.e. Bruker .yep, Applied Biosystems .wiff, mzData, mzML, dta (Sequest format), mgf (Mascot generic format) and mzXML (Institute for Systems Biology xml format), other…)
Availability of MS data	Specify the source data. Provide either a URI or the location of the files, or their availability.
2.. Input data and parameters – (b) input parameters	
Database queried	The description and version of the sequence databank(s) queried, If the databank(s) is/are not available on the web, describe the content of this/these databank(s), including the number of sequences.
Taxonomical restrictions	Specify the amplitude of the subset of the databank(s) (for instance, "mammals", a NCBI TaxId, a list of accession numbers). Specify the number of entries searched.
Description of tool and scoring scheme	Descriptor of the scoring algorithm in the search engine (such as ESI-TRAP in Mascot, ESI – ion trap HCTultra in Phenyx, ESI-ION-TRAP in MS-Tag, Ion Trap (4 Da) in X !Tandem, etc)
Specified cleavage agent(s)	Describe the cleavage agent as available on the search engine. If the cleavage agent rules have been defined by the user, describe the cleavage rules)
Allowed number of missed cleavages	Allowed maximum number of cleavage sited missed by the specified agent during the in-silico cleavage process.

Additional parameters related to cleavage	This includes, for instance the consideration of semi-specific cleavages (occurring on only one terminus), or other parameter that is relevant to the generation of peptides.
Permissible amino acids modifications	Specify the amino acid modifications that have to be considered in the search, according to the search engine list and mode of consideration (fixed or variable for instance). If the user has added custom modification to the predefined lists, specify the modification and the associated rules.
Precursor-ion and fragment-ion mass tolerance for tandem MS (when applicable)	For tandem MS queries, specify the mass tolerance applied to precursor ions and to fragment ions submitted to the search engine (when applicable).
Mass tolerance for PMF (when applicable)	For PMF and other MS queries, specify the mass tolerance applied to the search engine (when applicable)
Thresholds; minimum scores for peptides, proteins	Describe the parameters associated to the selection of peptide and protein hits retained in the output. These might include, for instance, a minimal score, a maximum p-value, a maximum number of proteins in the output, a minimum number of peptides to match a protein, etc.
Any other relevant parameters	Any application-specific parameters that are to be set for a search and that have an impact on the interpretation of the result for that experiment.
3. The output from the procedure – (a) for identified proteins	
Accession code in the queried database	Protein Accession Code such as a Uniprot-SwissProt AC number (P34521) (not an ID such as YM45_CAEEL), or a ncbi gi number (12803681). In case of the concept of protein group is applied (i.e. a list of protein that share a number of identical peptides among those identified), the description of the protein group might replace the list of individual Accession Codes
Protein description	Protein description field from the database.
Protein scores	Values as reported by the search engine.
Validation status	For all protein hits in the search, specify if accepted without post-processing of search engine/de-novo interpretation (accept raw output of identification software) or if manually accepted as valid or as rejected (false positive).
Number of different peptide sequences (without considering modifications) assigned to the protein	Do consider the number of different peptide sequences. A peptide identified as unmodified and for instance containing an oxidized Methionine is counted as one. A peptide with one missed cleavage and one included peptide with no missed cleavages are considered as two.
Percent peptide coverage of protein	Expressed as the number of amino acids spanned by the assigned peptides divided by the sequence length.
In the case of PMF, number of matched/unmatched peaks	Describe the number of matched m/z values associated to the identified protein, and the number of unmatched signals (or the total number of m/z values in the original spectrum/spectra)

Other additional information, when used for evaluation of confidence	This might include retention time, multiplicity of peptide sequence occurrences, flanking residues, etc.
3. The output from the procedure – (b) for identified peptides	
Sequence (notify any deviation from the expected protein cleavage specificity)	Primary sequence of the matched peptides (include number of missed cleavages or if the peptide is issued from semi-specific cleavage or fully unspecific cleavage).
Peptide scores	Scores values and any associated statistical information as available in the output.
Chemical modifications (artefactual) and post-translational modifications (naturally-occurring); sequence polymorphisms with experimental evidence (particularly for isobaric modifications)	Occurrence and position of amino acid modifications, being artefactual (such as oxidized Met or Carbamidomethylated Cys), post-translational (such as myristoylated amino acid), issued from an amino acid mutation or a frame shift with respect to the sequence in the database. When a choice among possible isobaric modifications are reported, a justification should be applied (I/L, acetylation/trimethylation
Corresponding Spectrum locus	Reference to the experimental spectrum.
Charge assumed for identification and a measurement of peptide mass error	Description of the precursor ion charge state and mass deviation (either expressed as m/z difference or as recalculated mass difference).
3. The output from the procedure – (c) quantitation for selected ions	
Quantitation approach (*e.g.* 4plex-iTRAQ, ICAT, cICAT, COFRADIC)	Describe the experimental protocol used for the quantitation experiment (refer to the sample prep section)
Quantity measurement	Specify the way the quantitation is measured (*e.g.* integration of signals, use of signal intensity)
Data transformation and normalisation technique	Describe the method and software: What are the input intensity values, how have they been filtered, transformed, normalised and processed?
Number of replicates (biological and technical)	As part of the protocol.
Acceptance criteria (including measure of errors)	Describe the evaluation method applied to the quantitation software (or manual calculation) result. Describe the acceptance criteria and measures of variability.
Estimates of uncertainty and the methods for the error analysis, including the treatment of relevant systematic error effects and the treatment of random error issues	Include estimates of uncertainty and error tolerances. Include description of treatment of relevant systematic error effects and the treatment of random error issues (such as labelling efficiency, instrument detector saturation, presence of non-unique peptides in MS/MS spectra, interference of distinct peptides with overlapping isotopes or other known or suspected sources of quantification outliers).

Results from controls (when described)	In case control experiments have been made, or in case internal controls have been incorporated in the protocol, describe the result of the analysis of these controls.
4. Interpretation and validation	
Assessment and confidence given to the identification and quantitation (description of methods, thresholds, values, etc.)	Describe the approach the software (when applicable) used for assessing confidence to the identification and quantitation (when applicable).
Results of statistical analysis or determination of false positive rate in case of large scale experiments	Describe the result of any applied statistics to estimate false positive rate in case of "large scale experiments".
Inclusion/exclusion of the output of the software are provided (description of what part of the output has been kept, what part has been rejected)	Provide access to the output data.