Minimum Information About a Proteomics Experiment (MIAPE): Minimum requirements for reporting proteomics experiments

Version 1.0, 6th January, 2005.

Both the generation and the analysis of proteome data are now widespread, and high-throughput approaches are commonplace. Protocols continue to increase in complexity, as both methods and technologies evolve. A standardised minimum reporting requirement for proteomics experiments, analogous to the MIAME guidelines for transcriptomics, is required to facilitate the analysis, exchange and dissemination of proteomics data.

This document states the principles underlying the specification of the data and metadata that should be captured from proteomics experiments. It also describes the series of accompanying 'MIAPE modules', each of which contains the minimum reporting requirement for a specific technique such as liquid chromatography, mass spectrometry or proteoinformatic analysis. It is anticipated that these modules will evolve over time, and increase in number, to take account of changes to experimental technologies and practice.

Introduction

The burgeoning of the various gene and genome sequence databases is well documented[1,2]. In recent years we have also witnessed increasing interest in functional genomics. Now proteomics, the protein-oriented complement of genomics, is maturing with the development of a wide range of proteome analysis methodologies, many operating in a high-throughput mode. Proteomics workflows most frequently consist of one or more separations performed on samples by chromatography or gel electrophoresis, examination of separands by mass spectrometry, then peptide and protein assignment through bioinformatic analysis of the generated mass spectra[3,4] (examples of the kinds of data produced are shown in Figure 1). However, there are many technologies available, each enabling the analysis of different portions of a given proteome (differentiated by mass, charge or localisation); these generate heterogeneous representations of the data and most are clearly still evolving. Additionally, the future holds the promise of many new technologies; improved prefractionation and depletion techniques, affinity binding 'chips', and novel mass spectrometer components, for example.

Standardised representations of gene, genome and transcriptome data are now well established, and databases and tools for their analysis are widely used[5,6]. However, for proteome data the situation is less advanced. This is predominantly because, as stated, the field is only now maturing; this continuing dynamism has previously made it difficult to fully define the key data from an experiment. Both transcriptomics and proteomics generate data that are only meaningful set in the correct biological and methodological context: There are many different subsets of the 'total' (*i.e.* all tissues, all states) proteome of an organism, just as there are many related (if not ostensibly correlated) patterns of mRNA transcription and turnover, each distinguished both by cell type and condition. Therefore to understand an analysis, perform comparisons between data sets or derive statistics from their aggregation, it is crucial to understand the biological context. Equally, one cannot reliably compare images of electrophoresis gels without knowing how those gels, and the samples run on them were prepared; neither can one compare mass spectra without knowledge of the sample analysed, the make and model of the instrument and its configuration; nor can one compare the results of sequence homology searching where the reference database and its version are not known. Both the biological and methodological contexts are clearly crucial to the formation of expectations about the content of a data set; if for example either the state of the organism, or one the processing methods negate

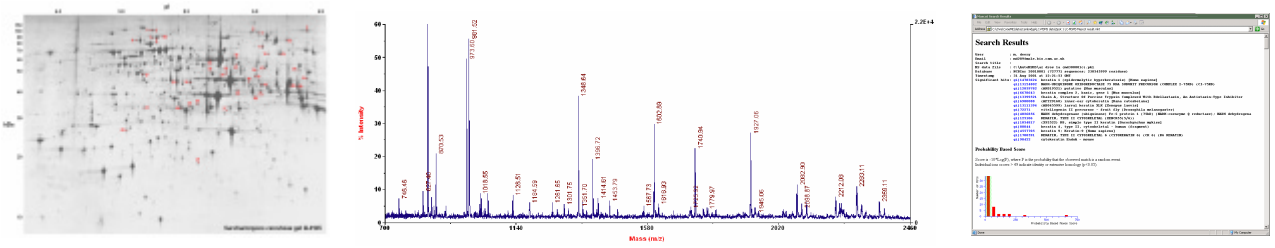

Figure 1. Examples of the kinds of data generated by proteomics experiments: (L to R) an annotated 2D-gel image; a mass spectrum; a protein identification search report.

the presence of a protein of interest in the sample studied, then one would want to be aware of that.

These context-sensitivities necessitate that the data generated by an experiment be accompanied by a rich set of metadata (data about the data). The use of paper citations as proxies for such metadata is a hindrance to users of data sets; impeding the reassessment of results, and obstructing automated search processes. More importantly, in the absence of agreed standards the requirements of various journals will certainly differ; as a result, expected information may be obscured, or absent.

It is clear then, that in addition to providing data, reporting a proteomics experiment should entail the provision of a representative set of the available metadata, making explicit both where samples came from and how analyses were performed. Therefore it is appropriate to define the minimum information about a proteomics experiment (hereafter, MIAPE) that should be provided to accompany data and conclusions. The definition will allow the users of a data set containing this minimum information to quickly establish both its provenance and relevance; it will also facilitate the development of effective search, analysis and quality control tools, because both developers and users will know what to expect from 'MIAPE-compliant' data sets and repositories (*i.e.* in terms of the information that should be available).

Many journals now require that papers reporting transcriptome experiments are accompanied by the MIAME-defined[7] set of metadata as a prerequisite for publication[8,9,10]. An equivalent requirement for some proteomics data has been drafted and implemented by at least one journal[11]. Of course the precise level of detail required will vary within the legion of techniques to be found in proteomics workflows: Most protocols generate a long list of parameters, many of which are unlikely to be of subsequent interest; all of these require classification as 'important' or not. Note that the MIAPE guidelines do not deal directly with the contentious issue of data quality, but do require the provision of sufficient information to allow that aspect to be independently assessed.

To guide the specific decisions on the data and metadata that should be required by each MIAPE module, two very general criteria were employed:

1. ***Sufficiency***. MIAPE should require sufficient information to unambiguously describe the experimental context of a dataset; thereby allowing a user to understand and critically evaluate the results and their interpretation, and in principle to reproduce the work.
2. ***Practicability***. Achieving MIAPE compliance should not be so burdensome as to prohibit the widespread use of the guidelines.

There is then a clear trade-off between the depth to which an experiment could be described and the time the average experimentalist can be expected to devote to generating an appropriate description. The resulting reporting requirement is perhaps reminiscent of the schema underlying a minimal Laboratory Information Management System (or LIMS), in that in addition to the data, metadata is required that would normally remain confined to the laboratory in which the data were produced, such as who performed the work, what hypothesis drove it and so on. The prevalence of computer technology in the lab is a boon here; for example it is expected that instrument and LIMS vendors will embrace these guidelines, simplifying greatly the process of reporting some elements of a workflow

through the provision of MIAPE-compliant export facilities. The provision of standard file formats and an accompanying ontology by the HUPO Proteomics Standards Initiative, as discussed in the final section of this paper, will facilitate this.

Operating Procedures

The aims of this document are firstly, to make explicit the scope, purpose and manner of use of the modular MIAPE guidelines; secondly, to lay out the principles underlying their production; and thirdly, to describe the first tranche of technology-specific modules to be produced. This document will be stable, as the principles described herein are technique-independent and should therefore remain valid in perpetuity. The associated documents (the modules) will be more labile; they will both evolve over time and increase in number, to track changes in experimental techniques.

Initial versions of all modules are generated by the Proteomics Standards Initiative (drawing heavily on community expertise) and then validated by expert committees. Once published to the PSI website at http://psidev.sf.net/gps/miape/, they can be discussed on an open email list (details of which can be found on that website). It is expected that this open discussion will drive both the evolution of extant modules and the genesis of new ones, as appropriate.

The MIAPE Modules

Each of the MIAPE modules relates broadly to the use of a particular technology such as gel electrophoresis, liquid chromatography or mass spectrometry. To build the MIAPE requirement for any proteomics workflow, the experimenter should retrieve the most recent versions of the appropriate modules for the technologies used and concatenate these into a single, bespoke reporting requirement for that entire workflow. The first modules to be deployed will be:

- ***Study design and sample generation***
  Experimental motivation; factors of interest; numbers of replicates; origin and preprocessing of biological material; relation to other studies; administrative details pertaining to the gestalt, such as the list of modules used.
- ***Prefractionation and sample treatment***
  The use of various techniques to fractionate or deplete a sample; the preparation of a sample for further separation or analysis (for example, by mass spectrometry).
- ***Liquid chromatography***
  The use of columns, of all scales and flow rates.
- ***Gel electrophoresis***
  The use of gel-based electrophoretic separation techniques, single- or multi-dimensional, native or denaturing; the use of 'electroblotting'; gel image acquisition, processing and analysis (to identify spots, measure relative intensities, or warp images to align them).
- ***Mass spectrometry***
  The use of a mass spectrometer; the generation of peak lists from raw data; quantitation based on the use of an isotopic or chemical label (the application of that label is a 'sample treatment' though, and is therefore captured elsewhere).
- ***Proteoinformatics***
  The use of processing engines to analyse MS data. This includes for instance search engines that assign peptides, proteins or biological class membership to spectra; the matching of those peptides or proteins, or *de novo* sequence, against protein sequences from a named database or other appropriate database; the use of quality control measures (*e.g.* evaluation of the significance of the processed results).

There are significant absences from the above list, which will be addressed. One notable omission is the use of protein arrays ('chips') to assay the presence, interactions or abundance of a protein using either antibody-based affinity binding, or some permutation of a protein-ligand interaction. A second, somewhat larger omission is the suite of methods used to investigate naturally-occurring protein-protein and other protein-ligand (*e.g.* small molecule) interactions; TAP tagging, yeast two-hybrid experiments; protein chips and so on (note though that far-Western blots can be described in the 'Gel electrophoresis' module). Guidelines for reporting protein interaction investigations will be generated after further consultation.

Legacy data

The intent of these guidelines is to ensure that the reporting of proteomics experiments requires a certain level of detailed annotation to be supplied. The problem of legacy, or 'pre-guidelines' data sets will be significant in scale and difficult to address. It is clear that a lack of annotation does not mean that a data set is without worth (although of course the quality of the work will be more difficult to establish), so the following principles should be applied when reannotating such legacy data:

1) The data set should be reannotated, with reference to the appropriate MIAPE modules, to the fullest extent possible; the data set should then be flagged as legacy, and some indication given of where the reporting requirements have not been met (*e.g.* a summary of missing items).
2) Data and metadata should *never* be created to supplement the real data in a file. Even the simple interrelating of parts of a data set can be hazardous. The *only* allowable additions are those that serve to indicate the absence of real data (*e.g.* the use of the word 'null' or the phrase 'not available').

Discussion

We believe that the MIAPE guidelines represent a sensible compromise between our stated criteria of sufficiency and practicability, in that they require a moderately rich description without being overly burdensome. Much of the required data should be readily available in the laboratory, especially as instrument, analysis software and LIMS vendors implement standards-compliant export facilities. Additionally, a substantial portion of the captured metadata will be common to many experiments, permitting some economies of scale. The guidelines therefore provide a sound base for repository and tool developers to work from, and a rational framework for journals and funders to consider enforcing.

Several benefits will arise from the widespread acceptance of MIAPE: All compliant data sets will contain sufficient information to quickly establish the provenance and relevance (to the researcher) of a data set, and to allow non-standard searches; additionally, tools will quickly appear that afford easy access to, and analysis of large numbers of such data sets. This tool development will be facilitated by the XML-based data transport formats (mzData, mzIdent, Gel-ML and the more general PSI-ML) and the supporting ontology generated by the Proteomics Standards Initiative; for further details of these standard formats browse to the PSI website at http://psidev.sf.net/.

Once the MIAPE guidelines are stable, application development will proceed apace, as will the integration of proteomics databases with each other, and with other resources (such as the major sequence databases), providing sophisticated search and analysis tools to practitioners of proteomics and to the wider research community. We are aware that guidelines from other domains (*e.g.* transcriptomics[7]) do partially overlap with MIAPE; for example on study design, sample collection or statistical analysis. Where this is the case we will coordinate with the appropriate organisations to agree common standards, thereby expediting the process of assembling compatible data sets from such disparate domains.

In this age of genome- and proteome-scale experiments, the need to standardise the process of reporting biological experiments is evident, if we are to extract full value from our activities[12,13]. It is our hope that this document and the modules accompanying it will begin to fulfill this need for the proteomics community, increasing the value of both individual pieces of work and of the general, diverse corpus to which we all contribute.

References

[1] Kulikova, T. *et al*. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. **32,** Database issue: D27–30 (2004).

[2] Galperin, M.Y. The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res. **32,** Database issue: D3–22 (2004).

[3] Wilkins, M.R., Williams, K.L., Appel, R.D. & Hochstrasser, D.F. (Eds.) *Proteome research: new frontiers in functional genomics* Springer, Berlin (1997).

[4] Pennington, S.R. & Dunn, M.J. (Eds) *Proteomics. From Protein Sequence to Function.* BIOS, Oxford (2001).

[5] Attwood, T.K. The quest to deduce protein function from sequence: the role of pattern databases. Int. J. of Biochem. & Cell. Biol. **32**, 139–155 (1999).

[6] Birney, E., Clamp, M., Hubbard, T. Databases a nd tools for browsing genomes. Annu. Rev. Genomics Hum. Genet. **3**, 293–310 (2002)

[7] Brazma, A. *et al*. Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. Nature Genetics **29**, 365–371 (2001).

[8] Nature **419**, 323 (2002)

[9] Lancet **360**(**9338**), 1019 (2002)

[10] http://www.cell.com/misc/page?page=authors

[11] Carr, S. *et al.* The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data . Mol. Cell. Proteomics **3**(**6**), 531-3 (2004).

[12] Prince, J.T. *et al*. The need for a public proteomics repository. Nature Biotechnology **22**(**4**), 471-472 (2004).

[13] Noble , W.S. Data hoarding is harming proteomics. Nature Biotechnology **22**(**10**), 1209 (2004).