

## Proteomics Data Types

### Introduction

Proteomics is defined as “the analysis of the expression, localizations, functions, and interactions of the proteins expressed by the genetic material of an organism”. Proteomics deploys a diversity of technologies to collect, separate and identify proteins. High throughput technologies contribute a significant increase in the volume and complexity of proteomics data. Thus, a common standard for representation of proteomics data and an agreed minimum required level of detail are required to facilitate the analysis, dissemination and exchange of proteomics data. The Human Proteomics Organization (HUPO) Proteomics Standards Initiative (PSI) has been working to develop a number of proteomics data standards. These standards will serve as a guide for accommodating the NIAID ARC proteomics data submissions and dissemination. This document describes the information requested for data submitted by PRCs.

### General Information

General information about any experimental data submitted. This information is required for data mining and interoperability between the different data types and outside databases.

Additional details maybe required for each data type and included in SOPs.

- a. Project
  - Project Name (Required)
  - Project Description (Required)
  - Contact Information (Required)
  - Publication (PubMedID, or manuscript)
- b. Experiment
  - Title of the Experiment (required)
  - Experiment Design Description (required)
  - Experiment Type (a controlled vocabulary)
  - Experiment Design Type (a controlled vocabulary)
  - Experiment Factors ( a high level list of the important parameters and conditions used in the experiment., more details are required for each data type)
- c. Biological Sample Information
  - Organism
    - Scientific Name (required)
    - NCBI Taxonomy ID (required)
    - Common Name
    - Strain Subtype Line
  - Sample
    - Description (required)
    - Sample Type (controlled vocabulary i.e. cell\_lysate, genomic\_DNA..)
    - BioSource Type (controlled vocabulary i.e. fresh\_sample, frozen\_sample..)
    - Tissue\_type
    - Cell\_type
    - Age

Cell\_cycle\_phase  
Developmental stage  
Cell\_component  
Cell\_line  
Genetic modifications (brief description for transgenic organisms)  
Provider

- Manipulation Of Biological Samples (Protocols)
  - Growth Protocol
    - a. Culture\_condition
    - b. Condition\_temp
    - c. Environment
  - Biological Sample Treatment (Protocols)
    - a. Metabolic\_label
    - b. Treatment
- Extract Information (Protocols)
  - c. Extraction Protocol
  - d. Purification Protocol
  - e. Labeling Protocol
- Experimental Factor for each sample. For example this could be the time factor.

d. Analytical Technique (see below for each data type)

## Mass Spectrometry Data

According to HUPO PSI web site, “The MIAPE requirement represents a subset of the total information available from a proteomics experiment, containing just enough information to assess the provenance and relevance of a set of methods, results and conclusions”. MIAPE has clearly described what information is required. The MS data submitted by PRCs should comply with MIAPE.

### 1 Files Required

The following files, or there equivalents, should be submitted.

- Metadata file, formatted in key value pairs or XML. This file describes the essential project, experiment, sample and protocol (SOP) information.
- Primary Peak list and intensity data file xxx.dta file.
- Primary SEQUEST or MASCOT search output file (xxx.out)
- File containing a list of peptide and protein identifications the PRC considers valid by their filtering criteria and thresholds if different from the primary search output.
- Summary of the analysis and a description of table headers and fields in each file.
- Lists of the peptides and proteins included in the final analyzed results and conclusions.
- Spectrometry image data (xxx.raw) files are not required, but if not provided must be stored by the PRC and made available if requested.

## 2 MIAPE Requirements

### 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
  - Machine manufacturer, model and date
  - Significant customizations (summary)
  - Resolution for all MS levels for which data are presented; state '10% valley' or 'FWHM'
  - Calibration method and compound(s)
  - Estimated mass accuracy (ppm) for all MS levels for which data are presented
- b) Control and analysis software
  - Software name and version
  - Switching criteria (tandem only)
  - Isolation width (global, or by MS level)
  - Location of 'parameters' file

### 2. Ion sources

- a) Electrospray Ionisation (ESI)
  - Supply type (static, or fed)
  - Scan cycle times (fed only)
  - Solvent flow rate and composition
  - Interface manufacturer, model and catalog number (where available)
  - Sprayer type, coating, manufacturer, model and catalog number (where available)
  - Tip and cone voltages (V)
  - Acceleration voltage for each MS level (V)
  - Whether in-source dissociation performed
  - Nebulising gas and pressure (bar), if any
- b) MALDI
  - Plate composition (or type)
  - Matrix composition (if applicable)
  - Deposition technique
  - Grid voltage (V)
  - Acceleration voltage for each MS level (V)
  - PSD (or LID/ISD) summary, if performed
  - Whether extraction was delayed
  - Laser type (e.g. nitrogen), wavelength (nm), pulse energy ( $\mu\text{J}$ ), attenuation, focus diameter ( $\mu\text{m}$ ), pulse duration (ns at FWHM), frequency (Hz) and average shots fired per spectrum

### 3. Post-source componentry

- a) Ion optics, 'simple' quadrupoles, hexapoles
  - No parameters to be captured

- b) Time-of-flight drift tube (TOF)
  - Reflectron status (on, off, none)
- c) Ion trap
  - Final MS exponent achieved
- d) Collision cell
  - Gas type and pressure (bar)
  - Collision energy
- e) FT-ICR
  - As for 'Ion trap' (3c) and 'Collision cell' (3d) combined, no further parameters required
- f) Detectors
  - Detector type
  - Detector sensitivity
  - Rate of data acquisition

#### 4. **Peak list generation and annotation**

For this section, if software other than that listed in 1b (Control and analysis software) is used to perform a task, the producer, name and version of that software must be supplied in each case

- a) Generation of 'stick' spectrum
  - Location of source ('raw') file including file name and type
  - Parameters triggering the generation of peak lists from raw data, where appropriate
  - Acquisition number (from the 'raw' file) of all acquisitions combined in the peak list, the total number combined and whether summed or averaged
  - Relative times for all acquisitions combined in the peak list (electrospray only)
  - Identifying information for the target area (MALDI-like methods only)
  - Signal-to-noise estimation and method
  - Smoothing; whether applied, parameters
  - Percentage peak height for centroiding; or algorithm used, if appropriate
  - Background threshold, or algorithm used
  - Base peak m/z, where appropriate
  - Whether charge states calculated, spectrum deconvoluted and peaks deisotoped (with methods described as appropriate)
  - Metastable peaks removed, if applicable
  - Ion mode for this spectrum
  - MS level for this spectrum
  - Precursor m/z and charge, with the full mass spectrum containing that peak
  - m/z and intensity values
- b) Quantitation for selected ions
  - As for 'Generation of stick spectrum' (4a), plus the following information
  - Experimental protocol, canonical reference where available with deviations
  - Number of combined samples analyzed
  - Quantitation approach (e.g. integration)
  - Normalization technique

- Location of quantitation data, with file name and type (where appropriate)

For MIAPE information see: <http://psidev.sourceforge.net/gps/index.html>

For MIAPE MS detail see: [http://psidev.sourceforge.net/gps/miape/MIAPE\\_MS\\_2.0.pdf](http://psidev.sourceforge.net/gps/miape/MIAPE_MS_2.0.pdf)

## Microarray Data

### 1. Introduction

This document outlines the Minimum Information About a Microarray Experiment (MIAME) a result of a MGED effort to codify the description of a microarray experiment. It tries to specify the information needed to allow someone to completely reproduce an experiment that was performed elsewhere. MIAME is a semi-formal textual description of the information that should be provided for each type of data. Usage of controlled vocabulary is recommended whenever possible.

For more information on MIAME see: <http://www.mged.org/Workgroups/MIAME/miame.html>

### 2. Data Formats

MIAME strongly suggests that the data be submitted in a standard format. Submitted data should include an Array Description File (ADF), the raw data i.e. scanner or imager and feature extraction output (providing the images is optional) and normalized data should be submitted. An ADF (Array Description file) is usually a tab-delimited-file devised to provide a consistent framework for representing a microarray layout and all relevant information attached to it. Examples of how the ADF file should be constructed are shown in Table 1 (for Oligonucleotide) and Table 2 (for cDNA).

An alternative for exchanging array designs is to use MAGE-ML.

For a MIAME checklist see: [www.mged.org/Workgroups/MIAME/MIAMEchecklist\\_cgh.pdf](http://www.mged.org/Workgroups/MIAME/MIAMEchecklist_cgh.pdf)

For details on MAGE-ML see: [www.mged.org/Workgroups/MAGE/mage.html](http://www.mged.org/Workgroups/MAGE/mage.html)

### 3. MIAME Requirements

#### 3.1 Hybridization Design

Contains any information pertaining to the physical parameters of hybridization.

1. Hybridization Procedures and Protocols
  - 1.1. Hybridization Protocol
  - 1.2. Washing Protocol
  - 1.3. Staining Protocol

#### 3.2 Measurement Data Design

Contains experimental results (raw data). Includes gene expression matrix.

1. Data

- 1.1. The raw data, i.e. scanner or imager and feature extraction output (providing the images is optional).
- 1.2. The normalized and summarized data, i.e., set of quantifications from several arrays upon which the authors base their conclusions
2. Data Extraction and processing Protocols
  - 2.1. Image scanning hardware, software, processing procedures and parameters.
  - 2.2. Normalization, transformation and data selection procedures and parameters.

### 3.3. Array Design

This sections contains all information concerning physical attributes of the arrays themselves

1. General Information
  - 1.1. Array Design Name
  - 1.2. Technology Type
  - 1.3. Substrate Type
2. Array Description Format
  - 1.1. Array Protocol

### Protein Protein Interaction Data.

The HUPO PSI Molecular Interaction (PSI-MI) describes a widely accepted molecular interaction data exchange format and is highly recommended. PSI is following a leveled approach to building this specification. Level 1 describes protein interactions at a basic level that cover a large amount of currently available data (e.g., BIND, DIP etc.). Subsequent levels will add capability to represent new molecular interaction information, such as among small molecules, DNAs and RNAs. PSI-ML is being incorporated into the BioPax format for system biology.

For more information on PSI-MI see: <http://psidev.sourceforge.net/mi/rel25/#submission>

For PSI-MI XML documentation see: <http://psidev.sourceforge.net/mi/xml/doc/MIF.html> .

The AC strongly recommends exchanging interaction data in the PSI-MI format as it is compatible with most all current interaction databases. However, we can accept older tab delimited interaction formats as long as the AC and MIAPE required metadata are also submitted.

### Proteomics data type of PRCs

Currently the PRCs have submitted some experimental and sample data. The submitted data are enumerated in Table 3, and they cover liquid chromatography coupled with mass spectrometry (LS-MS), yeast two-hybrid system (Y2H), microarray and genomic clone data types. For capturing and

disseminating those submitted proteomics data, the Admin Center has developed a proteomics data type definition document. The proposed proteomics DTD is simplified in comparison to the PEDRo modeling diagram.

**Table 1. Oligonucleotide array description file example:**

Feature				Reporter						Biological annotation			
Coordinates on Array				Reporter ID (user defined) Oligo ID	Bio-sequenc e Type	Sequence	DDBJ/ EMBL/ Genbank	Reporter Usage	Control Type	ID	Designation	Related Gene Symbol, if appropriate	Database Entry
Meta Col	Meta Row	Col	Row										
1	1	1	1	Cy3Cy5	Oligo	AAAAAAA AAAAAAA AAAA		Control	Positive	C001_01	Labeled oligo		
1	1	2	1	M00868_01	Oligo	ACCAGCA GATACCT CCTTG	D83002	Experiment al		C002_01	Gene	ALK	LocusID 11682
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	6	10	8	M00264_01	Oligo	ATGTCCG TTGAATT GG	D83002	Experiment al		C002_01	Gene	ALK	LocusID 11682
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	6	11	8	M02404_01	Oligo	AGTGGCG AGGAGGA GGAC	L11065	Experiment al		C449_01	Gene	OPRK1	LocusID 18387
4	6	12	8	M03172_01	Oligo	CCACCAC CAAGACC TACTCC	U34891	Experiment al		C450_01	Gene	KLRA9	LocusID 16640

**Table 2 cDNA array description file example:**

Feature				Reporter						Biological annotation			
Coordinates on Array				Reporter ID (user defined) HGMP Ref	Biosequence Type	Clone ID	DDBJ/ EMBL/ Genbank	Reporter Usage	Control Type	ID	Designation	Related Gene Symbol	Database Entry
Meta Col	Meta Row	Col	Row										
1	1	1	1	370503	cDNA clone	IMAGE 32017	R17905	Experimental	–	C1	Gene	FNTA	LocusID2339
1	1	2	1	370504	cDNA clone	IMAGE 296283 1	BC005866	Experimental	–	C2	Gene	MLH1	LocusID 4292
1	1	3	1	370505	Genomic clone	Cosmid 9H11	L40416	Control	Positive	–	–	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	8	24	12	380696	cDNA clone	IMAGE 521448 3	BC028215	Experimental	–	C285	Gene	PTEN	LocusID 5728

Table 3: The data submitted by PRCs

Organization	Dataset Description	Metadata Availability	Data Type	Size	Submit Date	Database Status	Display Status	Organism
<b>UOM</b>	46 microarray chip data	No	Microarray	128MB	Sep-05	test instance	test instance	Bacillus anthracis
<b>PNNL</b>	MS data (8Gb)	Yes	MS	8G	Sep-05	test instance	test instance	Salmonella typhimurium 14028
<b>PNNL</b>	MS data (128Gb)	Yes	MS	128G	Nov-05	test instance	test instance	Salmonella typhimurium LT12
<b>UOM (Scripps)</b>	MS data from 4 experiments	Yes	MS	0.7G	Sep-05	test instance	test instance	Bacillus anthracis
<b>Scripps</b>	Structure data on protein Nsp7	Yes	Structure data	<1MB		No	No	Bacillus anthracis
<b>Myriad</b>	Y2H searches	No	Y2H		Sep-05	No	No	Bacillus anthracis and Y. pestis