Broadening the Horizon – Level 2.5 of the HUPO-PSI Format for Molecular Interactions

BMC Biology

Software

Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions

Samuel Kerrien*[1], Sandra Orchard[1], Luisa Montecchi-Palazzi[1], Bruno Aranda[1], Antony F Quinn[1], Nisha Vinod[1], Gary D Bader[2,3], Ioannis Xenarios[4], Jérôme Wojcik[4], David Sherman[5], Mike Tyers[3], John J Salama[6], Susan Moore[6,7], Arnaud Ceol[8], Andrew Chatr-aryamontri[8], Matthias Oesterheld[9], Volker Stümpflen[9], Lukasz Salwinski[10], Jason Nerothin[10], Ethan Cerami[11], Michael E Cusick[12], Marc Vidal[12], Michael Gilson[13], John Armstrong[14], Peter Woollard[14], Christopher Hogue[15], David Eisenberg[10], Gianni Cesareni[8], Rolf Apweiler[1] and Henning Hermjakob[1]

Address: [1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]Banting & Best Department of Medical Research and Terrence Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario, Canada, [3]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario, Canada, [4]Merck Serono, 9 chemin des Mines, 1211 Geneva, Switzerland, [5]Laboratoire Bordelais de Recherche en Informatique, ENSI Électronique, Informatique et Radiocomm. de Bordeaux, France, [6]The Blueprint Initiative of Mount Sinai Hospital, 600 University Avenue, Toronto, ON, M5G 1X5, Canada , [7]National University of Singapore, Office of Life Sciences (OLS), Centre for Life Sciences, Singapore, [8]Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, Rome, Italy, [9]Institute for Bioinformatics, GSF – National Research Center for Environment and Health, Neuherberg, Germany, [10]UCLA-DOE Institute for Genomics & Proteomics, UCLA, LA, USA, [11]Computational Biology Center, Memorial Sloan-Kettering Cancer Center 1275 York Avenue, Box 460, New York, NY, USA, [12]Center for Cancer Systems Biology (CCSB) and Department of Cancer, Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA, USA, [13]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, MD, USA, [14]Glaxo Smithkline Medicines Research Centre, Gunnels Wood Road, Stevenage, Herts, UK and [15]Dept. of Biochemistry, University of Toronto, Toronto, Ontario, Canada

Email: Samuel Kerrien* - skerrien@ebi.ac.uk; Sandra Orchard - orchard@ebi.ac.uk; Luisa Montecchi-Palazzi - luisa@ebi.ac.uk; Bruno Aranda - baranda@ebi.ac.uk; Antony F Quinn - aquinn@ebi.ac.uk; Nisha Vinod - nvinod@ebi.ac.uk; Gary D Bader - bader@cbio.mskcc.org; Ioannis Xenarios - ioannis.xenarios@isb-sib.ch; Jérôme Wojcik - Jerome.Wojcik@serono.com; David Sherman - david.sherman@enseirb.fr; Mike Tyers - tyers@mshri.on.ca; John J Salama - jjsalama@gmail.com; Susan Moore - susanmoore197@hotmail.com; Arnaud Ceol - arnaud.ceol@uniroma2.it; Andrew Chatr-aryamontri - aryamontri@hotmail.com; Matthias Oesterheld - m.oesterheld@gsf.de; Volker Stümpflen - v.stuempflen@gsf.de; Lukasz Salwinski - lukasz@mbi.ucla.edu; Jason Nerothin - jason@mbi.ucla.edu; Ethan Cerami - cerami@cbio.mskcc.org; Michael E Cusick - Michael_cusick@dfci.harvard.edu; Marc Vidal - marc_vidal@dfci.harvard.edu; Michael Gilson - gilson@umbi.umd.edu; John Armstrong - john.t.armstrong@gsk.com; Peter Woollard - peter.m.woollard@gsk.com; Christopher Hogue - cwvhogue@gmail.com; David Eisenberg - david@mbi.ucla.edu; Gianni Cesareni - cesareni@uniroma2.it; Rolf Apweiler - apweiler@ebi.ac.uk; Henning Hermjakob - hhe@ebi.ac.uk

* Corresponding author

Published: 9 October 2007

BMC Biology 2007, **5**:44 doi:10.1186/1741-7007-5-44

Received: 19 February 2007
Accepted: 9 October 2007

This article is available from: http://www.biomedcentral.com/1741-7007/5/44

Abstract

Background: Molecular interaction Information is a key resource in modern biomedical research. Publicly available data have previously been provided in a broad array of diverse formats, making access to this very difficult. The publication and wide implementation of the Human Proteome Organisation Proteomics Standards Initiative Molecular Interactions (HUPO PSI-MI) format in 2004 was a major step towards the establishment of a single, unified format by which molecular interactions should be presented, but focused purely on protein-protein interactions.

Results: The HUPO-PSI has further developed the PSI-MI XML schema to enable the description of interactions between a wider range of molecular types, for example nucleic acids, chemical entities, and molecular complexes. Extensive details about each supported molecular interaction can now be captured, including the biological role of each molecule within that interaction, detailed description of interacting domains, and the kinetic parameters of the interaction. The format is supported by data management and analysis tools and has been adopted by major interaction data providers. Additionally, a simpler, tab-delimited format MITAB2.5 has been developed for the benefit of users who require only minimal information in an easy to access configuration.

Conclusion: The PSI-MI XML2.5 and MITAB2.5 formats have been jointly developed by interaction data producers and providers from both the academic and commercial sector, and are already widely implemented and well supported by an active development community. PSI-MI XML2.5 enables the description of highly detailed molecular interaction data and facilitates data exchange between databases and users without loss of information. MITAB2.5 is a simpler format appropriate for fast Perl parsing or loading into Microsoft Excel.

Background

Molecular interaction data is a key resource in modern biomedical research, and molecular interaction datasets are currently generated on a large scale, demonstrating from one to tens of thousands of interactions per experiment. These interaction data sets are represented in many different forms, from simple pairs of protein names to detailed textual descriptions, and are collected in various databases, each with their own database schema. In 2004, the HUPO Proteomics Standards Initiative developed and published the PSI-MI XML1.0 format for molecular interactions [1] as a community format for the exchange of protein interaction data. This format had been jointly developed by major producers of protein interaction data and by data providers, among them BIND [2], DIP [3], IntAct [4], MINT [5], MIPS [6] and Hybrigenics [7]. The PSI-MI XML1.0 format was widely implemented and supported by both software tool development and data providers.

The PSI-MI format was explicitly intended to develop in an incremental fashion. As a first step, version 1.0 focused exclusively on protein interactions, and provided only very limited support for quantitative parameters, such as kinetic measurements. As a direct result of requests from users, database groups and data providers, the HUPO-PSI work group for molecular interactions has significantly extended the capabilities of the original PSI-MI format resulting in version 2.5, presented here. The main features that have been added to the format broaden the range of interactor types, extend the descriptions that can be made of both experimental conditions and features on participating molecules, and add kinetic and modelled interaction parameters.

Results

We have developed version 2.5 of the PSI-MI XML schema for molecular interactions, extended the associated controlled vocabularies and updated the tools supporting the format. In the following pages, we will provide a general description of the PSI-MI XML2.5 format, pointing out the major changes and additions with respect to level 1.0. A complete, detailed documentation of the PSI-MI XML2.5 schema is available at [8].

The XML schema

The root element of PSI-MI XML2.5 is the *entrySet*. It contains one or more *entry* elements. An *entry* is the core element of PSI-MI XML2.5, describing one or more interactions with all associated data as a self-contained unit (Figure 1). Thus, several PSI-MI XML2.5 files can easily be merged by inserting all *entry* elements into a single *entrySet*. The *entry* contains a source element describing the origin of the data in the *entry*, usually an organisation such as a database provider. The following three elements, *availabilityList*, *experimentList*, and *interactorList*, are container elements for repetitive elements of interactions. The next element, *interactionList*, contains all interactions described in the *entry*.

The PSI-MI XML2.5 schema allows two different representations of interactions – compact and expanded. In the compact form, the repetitive elements of a larger set of interactions, namely copyright statements, experiment descriptions, and interactors (proteins, small molecules, etc), are only described once, in the respective list elements. The interaction elements in the *interactionList* then only refer to the previously defined interactors and experiments through elements of type *idRef*, similar to the normalised representation in a relational database. The compact form is more suitable for larger datasets, where, for example, one protein would be referred to by multiple interaction elements.

Alternatively, the expanded form of the PSI-MI XML2.5 schema can be used. In this form, the *availabilityList*, *experimentList*, and *interactorList* are not used, and all *experiment*, *interaction*, and *availability* elements are provided within the relevant *interaction* elements, potentially repeating the same description multiple times throughout

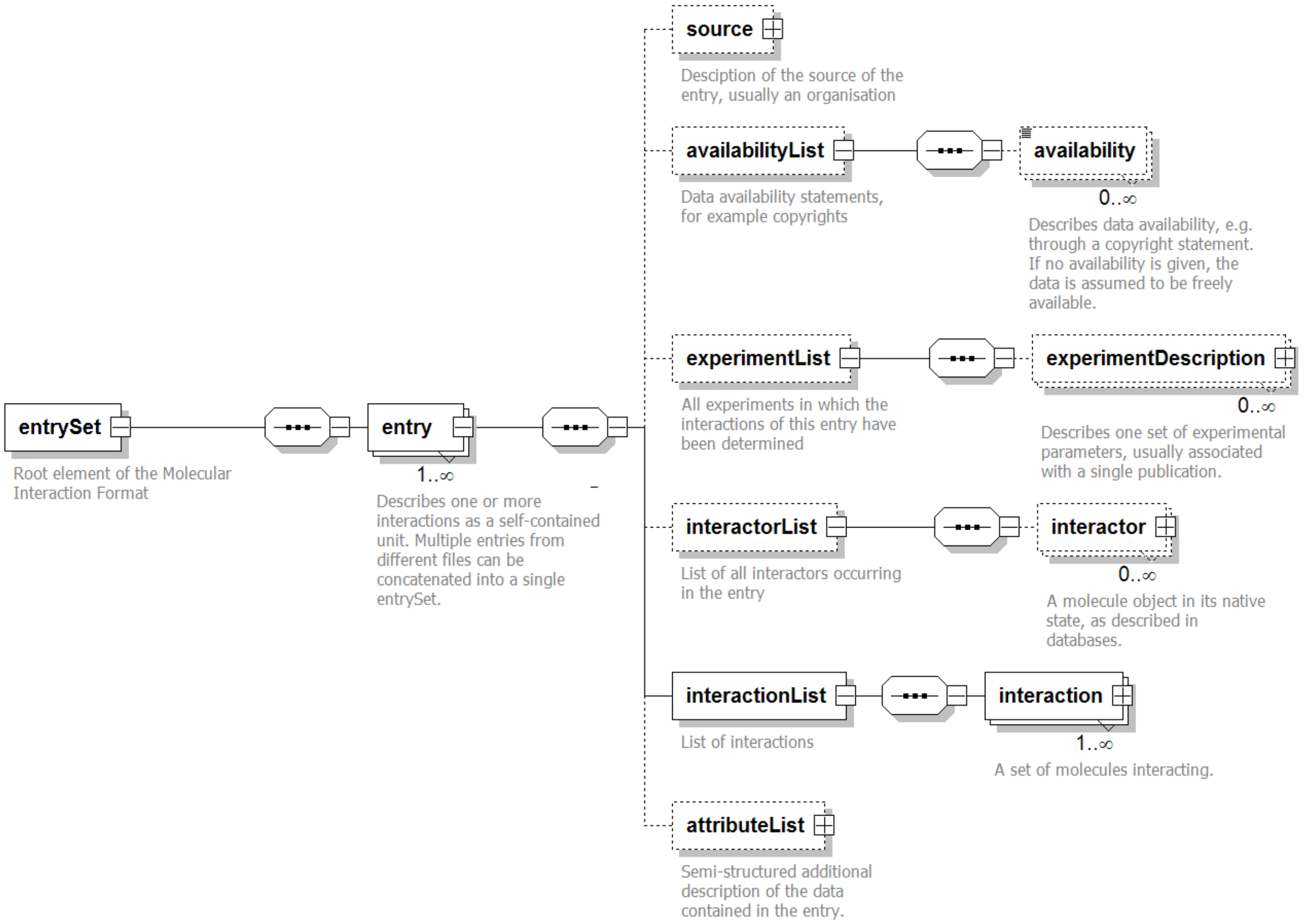

source
Desciption of the source of the entry, usually an organisation
availabilityList
Data availability statements, for example copyrights
availability
0..∞
Describes data availability, e.g. through a copyright statement. If no availability is given, the data is assumed to be freely available.
experimentList
All experiments in which the interactions of this entry have been determined
experimentDescription
0..∞
Describes one set of experimental parameters, usually associated with a single publication.
entrySet
Root element of the Molecular Interaction Format
entry
1..∞
Describes one or more interactions as a self-contained unit. Multiple entries from different files can be concatenated into a single entrySet.
interactorList
List of all interactors occurring in the entry
interactor
0..∞
A molecule object in its native state, as described in databases.
interactionList
List of interactions
interaction
1..∞
A set of molecules interacting.
attributeList
Semi-structured additional description of the data contained in the entry.

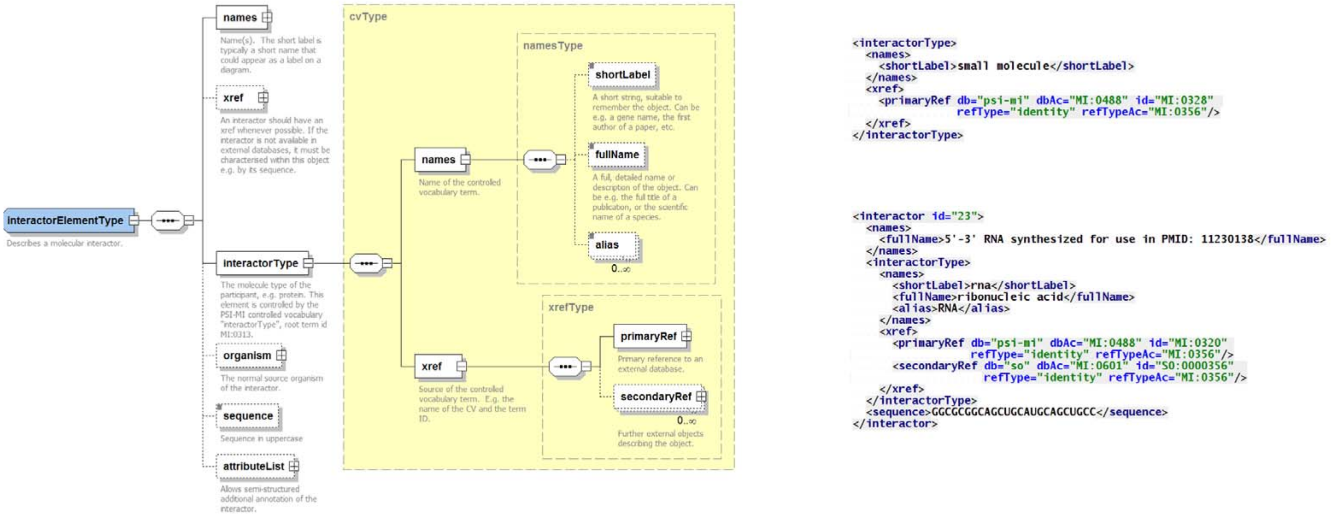
Figure 1
Graphical representation of the PSI-MI XML2.5 format. Some elements have been collapsed for clarity (indicated by a '+' in a rectangular box).

an *entry* in a manner comparable to a renormalized relational database. The expanded form is more suitable for smaller datasets, as it groups all related data closely together. The format presents further advantages in that interactions are now self-contained units thus rendering streaming of data much simpler and parsing more efficient. The compact and expanded form can be automatically inter-converted by XSLT scripts as described in the 'tools' section.

An *experimentDescription* provides a set of experimental parameters, which is usually associated with a single publication. However, a single publication will result in more than one *experimentDescription* element in PSI-MI XML2.5 representation, if the authors have used different technologies to confirm interactions. In the PSI-MI group, there is broad agreement that the experimental technologies to demonstrate a particular interaction are a key element in interpreting molecular interaction data. However, a full description of experimental detail, as specified in recent journal guidelines [9,10], is beyond the scope of the current PSI-MI XML2.5 schema. The *experimentDescription* focuses on a classification of experiments according to the methods used. Descriptions of experimental details, such as sample preparation or instrumentation used, will be modelled in detail in a future version, probably in the context of the maturing Functional Genomics Experiment (FuGE) [11] data model. As with PSI-MI XML1.0, the three method classifiers on experiment level are the *interactionDetectionMethod*, the *participantIdentificationMethod* and the *featureDetectionMethod*. Further key elements of the *experimentDescription* are the *bibref* bibliographic reference to the data source, and the *hostOrganismList* describing the environment in which the experiment has been performed. The latter is described using the NCBI taxonomy identifiers, with the addition of -1 to indicate '*in* *vitro*', -2 to indicate 'chemical synthesis', -3 to indicate 'unknown', -4 to indicate '*in vivo*' and -5 to indicate '*in silico*'.

An *interactor* element describes a molecule participating in an *interaction*. In version 1.0, the corresponding element was named *proteinInteractor*, as version 1.0 was only intended to represent protein interactions. The original plan was to add additional interactor types, e.g. *rnaInteractor*, as separate elements in future versions of the schema, however, it quickly became apparent that a full modelling of all relevant interactor types, for example mRNA, tRNA, rRNA and so on, would make the schema unnecessarily repetitive and complex. Thus, version 2.5 contains a lightweight representation of an interacting molecule, with the key element *interactorType*, which qualifies an interactor with a term from the PSI-MI controlled vocabulary, for example 'protein' (MI:0326; see Availability and requirements for instructions on how to access this and other entries) or 'small nucleolar rna' (MI:0607) (Figure 2). This allows a compact, yet very expressive representation of an interactor. The schema does not attempt to provide a detailed representation of interacting molecules – *interactor* elements will usually contain only basic information and refer to external databases for detailed descriptions. The controlled vocabulary for *interactorType* currently comprises >25 terms and can be easily expanded should a user wish to describe an additional *interactor* type, for example phospholipids, without any change to the XML schema. Additional cross-references to external resources allow for a more detailed description of the interactor, for example to UniProtKB (MI:0018) [17] to further describe a protein molecule or ChEBI (MI:0474) [18] for a chemical entity (Figure 3). The schema also now supports the hierarchical build-up of complexes from component sub-complexes, in that a sub-complex can be recycled as an *interactor* using the appropriate controlled vocabulary term and internal referencing (Figure 4). This allows the visualisation of a series of consecutive events, for example in the case where a receptor must form from two or more subunits before its agonist can bind.

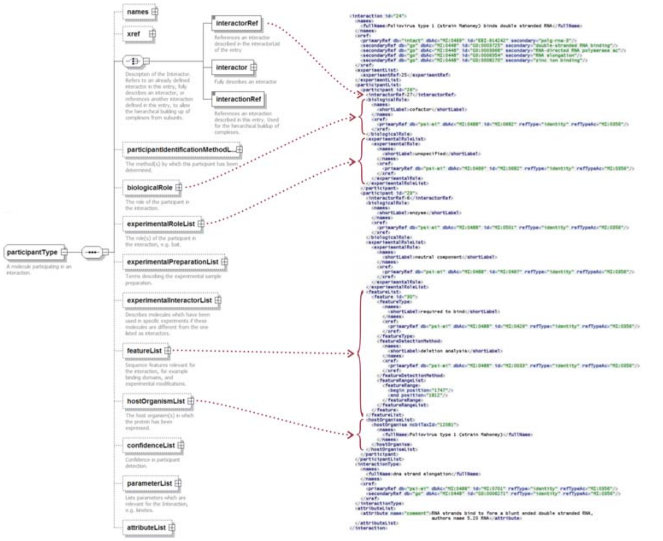
The interaction element describes one interaction. Within the interaction element, the experimentList either refers to or describes all experiments in which this interaction has been determined. The participantList lists all the molecules participating in an interaction. In the PSI-MI schema, an interaction can have one (e.g. in an autophosphorylation), two (e.g. in a yeast two-hybrid experiment) or multiple (e.g. in a tandem affinity purification (MI:0676) experiment) participants. While an interactor is the molecule in its 'canonical form' as represented in an external database, a participant element describes the specific instance of the interactor in the interaction. The participant must refer to an underlying interactor. In version 2.5, the role element of participant has been split into a biologicalRole, such as 'enzyme' (MI:0501), 'enzyme target' (MI:502) or 'cofactor' (MI:0682), and an experimentalRole, for example 'bait' (MI:0496) or 'prey' (MI:0498) (Figures 3 and 4). In addition, it is now possible to include multiple experimentalRoles for a given participant, each referring to one or more experiments in which the participant had this role. This can be used when an interaction in a yeast two-hybrid (MI:0018) experiment


InteractorElementType
Describes a molecular interactor.
names
Name(s). The short label is typically a short name that could appear as a label on a diagram.
xref
An interactor should have an xref whenever possible. If the interactor is not available in external databases, it must be characterised within this object e.g. by its sequence.
interactorType
The molecule type of the participant, e.g. protein. This element is controlled by the PSI-MI controlled vocabulary "interactorType", root term id MI:0313.
organism
The normal source organism of the interactor.
sequence
Sequence in uppercase
attributeList
Allows semi-structured additional annotation of the interactor.
cvType
names
Name of the controlled vocabulary term.
xref
Source of the controlled vocabulary term. E.g. the name of the CV and the term ID.
namesType
shortLabel
A short string, suitable to remember the object. Can be e.g. a gene name, the first author of a paper, etc.
fullName
A full, detailed name or description of the object. Can be e.g. the full title of a publication, or the scientific name of a species.
alias
0..∞
xrefType
primaryRef
Primary reference to an external database.
secondaryRef
0..∞
Further external objects describing the object.

```
<interactorType>
  <names>
    <shortLabel>small molecule</shortLabel>
  </names>
  <xref>
    <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0328"
                refType="identity" refTypeAc="MI:0356"/>
  </xref>
</interactorType>
```

```
<interactor id="23">
  <names>
    <fullName>5'-3' RNA synthesized for use in PMID: 11230138</fullName>
  </names>
  <interactorType>
    <names>
      <shortLabel>rna</shortLabel>
      <fullName>ribonucleic acid</fullName>
      <alias>RNA</alias>
    </names>
    <xref>
      <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0320"
                  refType="identity" refTypeAc="MI:0356"/>
      <secondaryRef db="so" dbAc="MI:0601" id="SO:0000356"
                    refType="identity" refTypeAc="MI:0356"/>
    </xref>
  </interactorType>
  <sequence>GGCGCGGCAGCUGCAUGCAGCUGCC</sequence>
</interactor>
```

Figure 2
Use of *interactorType* within the PSI-MI XML2.5 schema to describe the molecular class of two participating molecules – a small molecule interactor and a RNA – using the appropriate controlled vocabulary term


names
xref
interactorRef
References an interactor described in the interactorList of the entry.
interactor
Fully describes an interactor
interactionRef
References an interaction described in this entry. Used for the hierarchical buildup of complexes.
Description of the Interactor. Refers to an already defined interactor in this entry, fully describes an interactor, or references another interaction defined in this entry, to allow the hierarchical building up of complexes from subunits.
participantIdentificationMethodL...
The method(s) by which this participant has been determined.
biologicalRole
The role of the participant in the interaction.
experimentalRoleList
The role(s) of the participant in the interaction, e.g. bait.
participantType
A molecule participating in an interaction.
experimentalPreparationList
Terms describing the experimental sample preparation.
experimentalInteractorList
Describes molecules which have been used in specific experiments if these molecules are different from the one listed as interactors.
featureList
Sequence features relevant for the interaction, for example binding domains, and experimental modifications.
hostOrganismList
The host organism(s) in which the protein has been expressed.
confidenceList
Confidence in participant detection.
parameterList
Lists parameters which are relevant for the Interaction, e.g. kinetics.
attributeList

```
<interaction id="24">
  <names>
    <fullName>Poliovirus type 1 (strain Mahoney) binds double stranded RNA</fullName>
  </names>
  <xref>
    <primaryRef db="intact" dbAc="MI:0469" id="EBI-814242" secondary="polg-rna-3"/>
    <secondaryRef db="go" dbAc="MI:0448" id="GO:0003725" secondary="double-stranded RNA binding"/>
    <secondaryRef db="go" dbAc="MI:0448" id="GO:0003968" secondary="RNA-directed RNA polymerase ac"/>
    <secondaryRef db="go" dbAc="MI:0448" id="GO:0006354" secondary="RNA elongation"/>
    <secondaryRef db="go" dbAc="MI:0448" id="GO:0008270" secondary="zinc ion binding"/>
  </xref>
  <experimentList>
    <experimentRef>25</experimentRef>
  </experimentList>
  <participantList>
    <participant id="28">
      <interactorRef>27</interactorRef>
      <biologicalRole>
        <names>
          <shortLabel>cofactor</shortLabel>
        </names>
        <xref>
          <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0682" refType="identity" refTypeAc="MI:0356"/>
        </xref>
      </biologicalRole>
      <experimentalRoleList>
        <experimentalRole>
          <names>
            <shortLabel>unspecified</shortLabel>
          </names>
          <xref>
            <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0682" refType="identity" refTypeAc="MI:0356"/>
          </xref>
        </experimentalRole>
      </experimentalRoleList>
    </participant>
    <participant id="29">
      <interactorRef>6</interactorRef>
      <biologicalRole>
        <names>
          <shortLabel>enzyme</shortLabel>
        </names>
        <xref>
          <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0501" refType="identity" refTypeAc="MI:0356"/>
        </xref>
      </biologicalRole>
      <experimentalRoleList>
        <experimentalRole>
          <names>
            <shortLabel>neutral component</shortLabel>
          </names>
          <xref>
            <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0497" refType="identity" refTypeAc="MI:0356"/>
          </xref>
        </experimentalRole>
      </experimentalRoleList>
      <featureList>
        <feature id="30">
          <featureType>
            <names>
              <shortLabel>required to bind</shortLabel>
            </names>
            <xref>
              <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0429" refType="identity" refTypeAc="MI:0356"/>
            </xref>
          </featureType>
          <featureDetectionMethod>
            <names>
              <shortLabel>deletion analysis</shortLabel>
            </names>
            <xref>
              <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0033" refType="identity" refTypeAc="MI:0356"/>
            </xref>
          </featureDetectionMethod>
          <featureRangeList>
            <featureRange>
              <begin position="1747"/>
              <end position="1812"/>
            </featureRange>
          </featureRangeList>
        </feature>
      </featureList>
      <hostOrganismList>
        <hostOrganism ncbiTaxId="12081">
          <names>
            <fullName>Poliovirus type 1 (strain Mahoney)</fullName>
          </names>
        </hostOrganism>
      </hostOrganismList>
    </participant>
  </participantList>
  <interactionType>
    <names>
      <fullName>dna strand elongation</fullName>
    </names>
    <xref>
      <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0701" refType="identity" refTypeAc="MI:0356"/>
      <secondaryRef db="go" dbAc="MI:0448" id="GO:0006271" refType="identity" refTypeAc="MI:0356"/>
    </xref>
  </interactionType>
  <attributeList>
    <attribute name="comment">RNA strands bind to form a blunt ended double stranded RNA,
                             authors name 5.20 RNA</attribute>
  </attributeList>
</interaction>
```

Figure 3
Graphical representation of section of the PSI-MI XML2.5 schema describing *participant* with the corresponding lines of XML containing sample data from [26]

has been observed twice, with the participant bait/prey roles being reversed in the repeat experiment. The experimentalPreparationList allows the provision of terms from a controlled vocabulary to describe key aspects of the experimental preparation of each participant prior to the detection of the interaction, for example a protein might be an 'in vitro translated protein' (MI:0589).

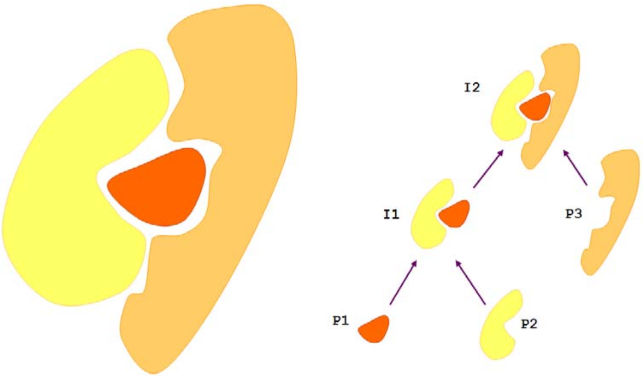


Figure 4
Assembly of a mature complex (I2) from two protein components (P1 and P2) that bind to form an intermediate assembly (I1) with which a third protein (P3) interacts. I1 is both the result of an interaction and a subsequent interactor in the PSI-MI2.5 format.

The method used for participant identification was described only at the experiment level in version 1.0. Whilst this is sufficient in most cases, it has been shown to be a limitation for experiments where molecules have been identified by different methods, for example in the case of an'antibait coimmunoprecipitation' (MI:0006) where some of the prey proteins could be identified by 'western blot' (MI:0113), and additional interacting molecules subsequently identified by 'mass spectrometry' (MI:0427). In PSI-MI XML2.5, the *participant* element contains a *participantIdentificationMethodList* where this information can be mapped.

Experimental constraints often require researchers to determine interactions in model systems. As an example, an experiment might be performed using mouse proteins, but the aim of the experiment is to make a statement about the homologous interaction (interolog) in human. The *experimentalInteractor* element allows the reporting of this fact. The *experimentalInteractor* will point to the model interactor used, while the human protein is referred to by the mandatory '*interactorRef*' element of the *participant*. Interactions of this type are referred to as 'modelled' interactions, and additionally flagged by the *modelled* element on the interaction level.

The *featureList* element describes the sequence features of the *participant* that are relevant for the interaction, using the appropriate term from the corresponding controlled vocabulary, for example 'DNA binding domains' (MI:0688) or experimental modifications such as 'biotin tag' (MI:0239). Features are classified by a controlled vocabulary linked from the *featureType* element. The *featureRangeList* describes the location of a feature on the sequence. This element has undergone extensive remodelling in version 2.5 to allow the representation of discontinuous sequence features and features of undetermined or general location such as 'n-terminal' (MI:0340).

The *negative* flag of an *interaction* is set to *true* if a particular interaction has been explicitly shown not to occur under the described conditions. Although it is optional (default value is *false*), it is obviously essential to correctly interpret this element when using PSI-MI formatted datasets.

The *confidenceList* provides an estimate of the confidence that this interaction actually occurs. The development of method-specific and method-independent measures for interaction confidence measures is an ongoing research topic, and thus the *confidenceList* allows the provision of multiple confidence measures for an interaction, usually based on different methods. A typical case would be one confidence measure provided by the original publication, and a separate confidence measure implemented by a public database.

In version 1.0 of the PSI-MI schema, the modelling of quantitative parameters was knowingly neglected, as very little quantitative data on protein interactions was available from public databases. In the current version, the *parameterList* on both *interaction* and *participant* level allows the flexible addition of kinetic parameters, which are becoming more available now through technologies like 'surface plasmon resonance' (MI:0107), and are being systematically collected by many resources, for example the BindingDB database [19].

The major XML schema elements contain a set of recurring standard elements, listed below.

(1) The *id* attribute provides an identifier to the object that must be unique to that object within the PSI-MI file. An object can be repeated, however, if the extended form of the schema is used. The *id* object is not necessarily stable, stable identifiers are provided in the *xref* element, see below.

(2) The *names* element to provide short labels (LCK_HUMAN), fully detailed names (Proto-oncogene tyrosine-protein kinase LCK), and synonyms (p56-LCK).

(3) The *xref* element to reference external database entities. Each *xref* element contains a single primary and optionally multiple secondary references that enable the referencing of both a primary data source, and a related secondary source such as a method publication. The specific reference type can be provided as an attribute *refType*, which is itself defined by a controlled vocabulary. Learning from experience gained in implementing version 1.0, database cross-references now use the extended PSI-MI CV2.5 controlled vocabulary to ensure consistent naming of external database resources, avoiding problems with the unification of, for example, 'Swiss-Prot', 'swissprot', 'sp', 'uniprot' and 'uniprotkb'. This resource is now identified by the attribute *db* and the optional attribute *dbAc*, which contains the accession number for the database in the PSI-MI controlled vocabulary.

(4) The *attributeList* element provides the option of flexibly adding content to PSI-MI elements. Each attribute has a name derived from a controlled vocabulary, and a free text value. As the attribute *name* must be selected from the PSI-MI controlled vocabulary, the *attributeList* contents remain semi-structured, and allows easy searching for specific topics.

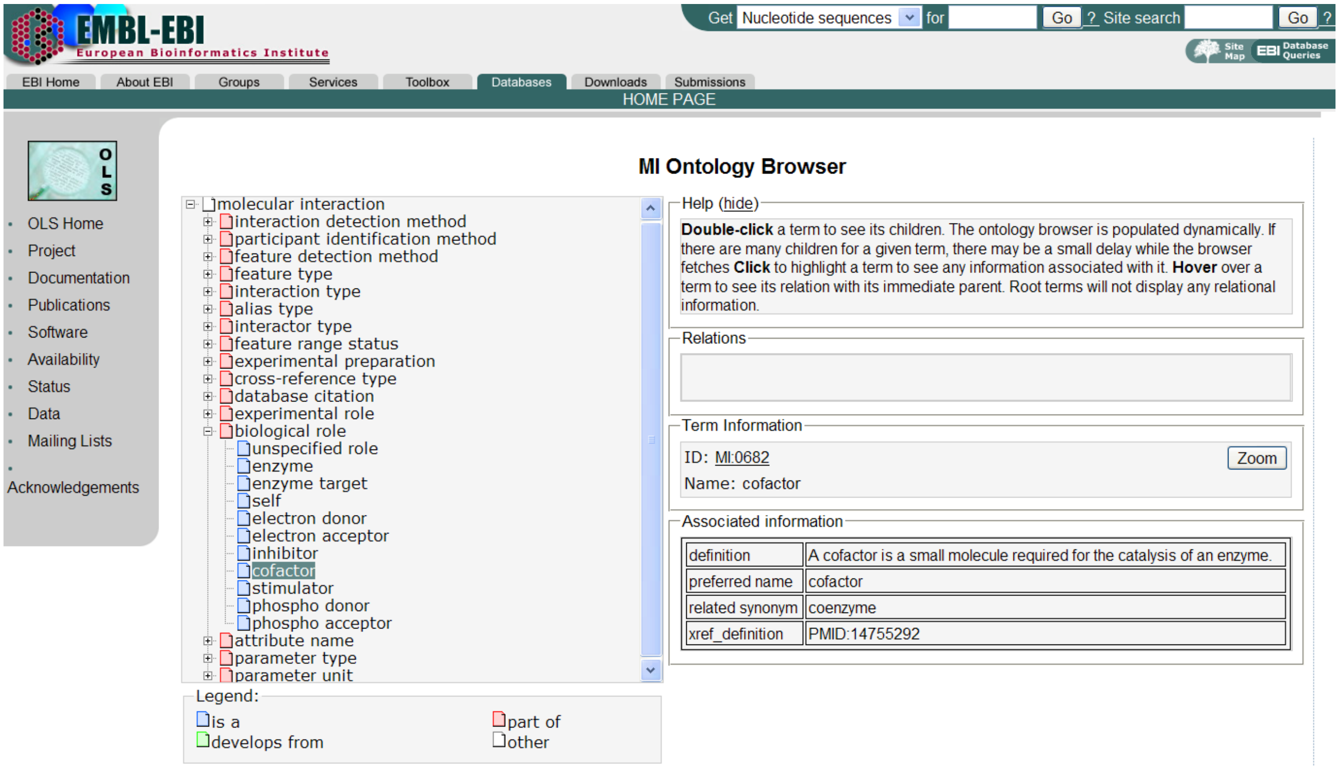
Controlled vocabularies

Controlled vocabularies (CVs) are used throughout the PSI-MI schema to standardize the meaning of data objects. Their use ensures that the same term used throughout a description by a data producer, instead of a synonym or alternative spelling, and also that the interpretation of the meaning of that term remains consistent between multiple data producers and data users. In order to achieve this, all terms have definitions and, where appropriate, are supported by a literature reference. The controlled vocabularies have a hierarchical structure, higher level terms being more general than lower level descriptors, allowing annotation to be performed to an appropriate level of granularity whilst also enabling search tools to return all mapped objects to both parent and child terms, if required (Figure 5).

The descriptive ability of the PSI-MI controlled vocabularies has increased dramatically since their original release concurrent with PSI-MI version 1.0. The number of root terms has increased from 5 to 16 and the depth of each CV has been augmented by the addition of both new terms and the improved definitions of existing term. For example, on the release of PSI-MI 1.0, only a single term 'aggregation' was listed in the controlled vocabulary with the root 'interaction type'. This term proved to be too ambiguous to adequately describe the interaction annotated in major interaction databases, and unpopular with users, so it was made obsolete. The term 'interaction type' (MI:0190) now has three top-level terms 'genetic interaction' (MI:0208), 'physical interaction' (MI:0218) and 'colocalization' (MI:0403) all of which have numerous children. This allows an interaction to be described as, for example, an 'enzymatic reaction' (MI:0414) or even for the annotator to be as specific as 'dna strand elongation' (MI:0701). Expansion of the CVs has resulted in these replacing elements in the PSI-MI XML1.0 that were previously Boolean such as *isOverexpressedProtein* and *isTaggedProtein*. This now allows a much richer description of each term, such as the ability to state precisely which tag has been used and whether it was placed at the C-terminus or N-terminus of the molecule.

The CVs are maintained in the Open Biomedical Ontologies (OBO) [12] flat file format (psi-mi25.obo) with the namespace prefix 'MI' and are available both on the PSI-MI and the OBO websites or can be browsed using the Ontology Lookup Service (see below). The CVs are designed and actively developed to rapidly respond to changes in technology and methodology. This process is overseen by an elected editorial board and users can request new terms via a tracker on the PSI-MI website. We aim at responding to user requests within two weeks.

As backward compatibility between the PSI-MI XML1.0 and 2.5 was not fully achievable, given the support for new 2.5 features, a final version of the 1.0 CVs were frozen and are available from the PSI-MI website. A mapping file from CV2.5 to CV1.0 has been supplied on the PSI-MI web pages to enable providers to release data in 1.0 for-



Double-click a term to see its children. The ontology browser is populated dynamically. If there are many children for a given term, there may be a small delay while the browser fetches **Click** to highlight a term to see any information associated with it. **Hover** over a term to see its relation with its immediate parent. Root terms will not display any relational information.

ID: MI:0682
Name: cofactor

definition	A cofactor is a small molecule required for the catalysis of an enzyme.
preferred name	cofactor
related synonym	coenzyme
xref_definition	PMID:14755292

Figure 5
A representative section of the PSI-MI controlled vocabulary displayed in the Ontology Lookup Service.

mat but there will inevitably be data loss with reverse mapping and it is envisaged that this will increase as the CV2.5 further develop, so migration to 2.5 is recommended at the earliest possible opportunity.

Tools

The development and maintenance of tools to enable the use of this format by a wide number of users is a community effort to which many people have contributed. All tools are freely available and can be accessed and downloaded from the PSI-MI web pages. These include applications to view and validate the use of the schema, to enable graphical representation of interaction network, to convert between data formats and to facilitate the use of controlled vocabularies. All tools are specific to the PSI-MI XML format apart from the validator.

PSI-MI XML view

XML is a powerful means by which to model complex data while preserving human readability. However, due to the complexity of the PSI-MI schema we felt it important to provide an easier way to visualise the data. XSLT scripts have been made available in order to convert XML data files to HTML, thus providing user-friendly representation of the data.

Conversion between expanded and compact form of the schema

A set of XSLT scripts has been made available to the community, thus allowing users to convert PSI-MI XML2.5 between compact and expanded forms.

Validator

The PSI-MI data model representation is based on an XML Schema that represents how the data should be formatted. Schemas provide the means to define cardinality of XML elements and give type to their attributes but do not allow the enforcement of complex semantic rules. In order to overcome this limitation, a generic validation framework was created and a specific extension developed for the PSI-MI XML2.5 format. It allows the definition of very detailed rules such as checking that the biological role of participants involved in a yeast two-hybrid (MI:0018) experiment is either bait (MI:0496) or prey (MI:0498). As illustrated in the example above, the framework provides ways to interface with OBO format ontologies. Several clients currently allow the validation of PSI-MI XML2.5 data. These include a Java API that enables the embedding of the validator into any third party application (this is the most versatile application), a command line interface in addition to a graphical interface bundled with the Java

API, and a web application that allows the uploading of a PSI-MI data file and reporting of both syntactic and semantic discrepancies [13] (after validation, it automatically creates the HTML view described above).

Ontology Lookup Service (OLS)

Many ontologies are made available using the OBO format. OLS is an AJAX-based ontology viewer that also acts as a portal to access all the controlled vocabularies currently maintained on the Open Biomedical Ontologies website, enabling the searching of the PSI-MI controlled vocabulary by term name, id or synonyms [20]. Autocompletion provides a user-friendly search mechanism. A programmatic interface allows the embedding of OLS into any SOAP enabled application.

Java XML parser

Many molecular interaction databases have chosen the PSI-MI format for providing data to their users. In order to ease the development of tools exploiting this data, a Java library providing a user friendly data model and the core functionalities for reading and writing PSI-MI XML2.5 data. Furthermore the parser supports XML streaming, thus providing a versatile and cursor based approach for retrieving interactions, interactors or experiments. The parser has been made available on the PSI-MI website.

XMLMaker/Flattener

The XMLMaker/Flattener is a Java application that converts any XML schema into tab-delimited ASCII format (flat-files) and vice versa, given a user-defined mapping that can be saved and reused on subsequent files. A PSI-MI mapping can be readily created to inter-convert PSI-MI 1.0 or 2.5 XML files to simple flat files.

We recognize that some of the tools listed above have limitations with respect to memory requirements when dealing with large data files and expanded/compact forms of the schema. The latter problem can be addressed by choosing the appropriate XSLT script.

Relationship with other community resources

The PSI-MI standard is only one of several inter-related efforts of the proteomics community to develop standards for data and XML-interchange formats [14]. These standards have been developed as part of the broader standards community and care is being taken to avoid overlaps, such that if a process is described in one standard it need only be referred to from others employing the same technology. For example, the controlled vocabularies used throughout all of the PSI-XML schemas are mapped to the schema, rather than being embedded within them to enable use of other useful CVs as they arise. For example, should a particular mass spectrometry technique be used to identify a participant, the appropriate term can be taken from the PSI-Mass Spectrometry CV rather than separately listed within the PSI-MI CVs.

BioPAX

The Biological Pathway Exchange (BioPAX) format is a collaboratively developed data exchange format for biological pathway data [15], that currently uses the PSI-MI ontology internally for modelling associated molecular interactions [16]. All PSI-MI entries annotated to 'physical interaction' map to the BioPAX *physicalInteraction* class.

Cytoscape

Cytoscape is a powerful open-source resource for analysing and visualizing biological networks [21]. The Cytoscape user community has developed numerous plugins allowing the extension of its functionalities in the area of data format compatibility and network analysis. Cytoscape now allows users to load molecular interaction data in PSI-MI XML1.0 and 2.5 formats without installing additional extensions, enabling data retrieval from one or more databases and its subsequent integration with data from other sources such as high-throughput expression experiments.

The tabular form: MITAB2.5

The PSI-MI XML2.5 format allows a detailed representation of fully annotated interaction records both for inter-database and database-end user data transmission. However, to support many use cases, such as fast Perl parsing or loading into Microsoft Excel, that only require a simple, tabular format of interaction records, the MITAB2.5 format was defined as part of PSI-MI 2.5. The MITAB2.5 format (Additional files 1, 2, 3, 4, 5) has been derived from the tabular format provided by BioGRID [22]. MITAB2.5 only describes binary interactions, one pair of interactors per row. Columns are separated by tabulators and the contents should be as follows:

1. Unique identifier for interactor A, represented as databaseName:ac, where databaseName is the name of the corresponding database as defined in the PSI-MI controlled vocabulary, and ac is the unique primary identifier of the molecule in the database. Identifiers from multiple databases can be separated by '|'. It is recommended that proteins be identified using stable identifiers such as their UniProtKB or RefSeq [24] accession number.

2. Unique identifier for interactor B.

3. Alternative identifier for interactor A, for example the official gene symbol as defined by a recognised nomenclature committee. Representation as databaseName:identifier. Multiple identifiers separated by '|'.

4. Alternative identifier for interactor B.

5. Aliases for A, separated by '|'. Representation as databaseName:identifier. Multiple identifiers separated by '|'.

6. Aliases for B.

7. Interaction detection methods, taken from the corresponding PSI-MI controlled vocabulary, and represented as databaseName:identifier (methodName), separated by '|'.

8. First author surname(s) of the publication(s) in which this interaction has been shown, optionally followed by additional indicators, e.g. 'Smith-2005-a'. Separated by '|'.

9. Identifier of the publication in which this interaction has been shown. Database name, usually PubMed (MI:0446), taken from the PSI-MI controlled vocabulary, represented as databaseName:identifier. Multiple identifiers separated by '|'.

10. NCBI Taxonomy identifier for interactor A. Database name for NCBI taxid taken from the PSI-MI controlled vocabulary, represented as databaseName:identifier. Multiple identifiers separated by '|'. Note that in this column, the databaseName:identifier(speciesName) notation is only there for consistency. Currently no taxonomy identifiers other than NCBI taxid are anticipated, apart from the use of -1 to indicate '*in vitro*', -2 to indicate 'chemical synthesis', -3 to indicate 'unknown', -4 to indicate '*in vivo*' and -5 to indicate '*in silico*'.

11. NCBI Taxonomy identifier for interactor B.

12. Interaction types, taken from the corresponding PSI-MI controlled vocabulary, and represented as dataBaseName:identifier(interactionType), separated by '|'.

13. Source databases and identifiers, taken from the corresponding PSI-MI controlled vocabulary, and represented as databaseName:identifier(sourceName). Multiple source databases can be separated by '|'.

14. Interaction identifier(s) in the corresponding source database, represented by databaseName:identifier.

15. Confidence score. Denoted as scoreType:value. There are many different types of confidence score, but so far no controlled vocabulary. Thus the only current recommendation is to use score types consistently within one source. Multiple scores separated by '|'.

16. Allfurther columns are currently undefined.

All columns are mandatory. Missing values (empty cells) are marked by '-'.

MITAB2.5 API

In order to ease the conversion from PSI-MI XML2.5 to MITAB2.5, a Java API was developed so data collected from various data providers can be easily integrated. A user-friendly graphical interface allows the user to drag and drop a set of PSI-MI XML2.5 files and convert them into one (or many) MITAB2.5 file(s). Additionally, whenever interactions involving more than two participants, for example many coimmunoprecipitation experiments, are encountered, either Spoke or Matrix expansion can be applied to the data in order to result in only binary interactions.

Data availability

Interaction data is available in PSI-MI XML2.5 format from DIP, IntAct, MINT, MIPS, BioGRID and the HPRD databases. MINT, IntAct and DIP also supply data in MITAB2.5 format. Building on the PSI-MI standard, major public interaction databases have formed the International Molecular Interaction Exchange consortium (IMEx) [23]. These databases, currently DIP, IntAct, MINT and MPact (MIPS), have started to share the curation load and aim to regularly interchange data curated to the same common standards, in a manner similar to the well-established pattern followed by the nucleotide sequence databases. IMEx downloads are currently available from the DIP, IntAct and MINT FTP sites.

Submissions

A number of databases actively encourage direct submissions by authors. Members of the IMEx Consortium host a number of tools and resources to assist in this process [25]. Any data submitted via this website will be exchanged between participating databases and made available, on publication of the accompanying manuscript, from all database websites in both PSI-MI XML2.5 and MITAB2.5 formats.

Sample entry

Sample data throughout this paper have been taken from [26] (see also Availability and requirements). The full data set can be viewed as supplementary materials in both PSI-MI 2.5 XML or MITAB2.5 formats or can be viewed as a fully curated dataset in the IntAct database (EBI-914159, EBI-914232, EBI-914360).

Conclusion

The PSI-MI XML1.0 format provided the molecular interaction community with a common mechanism for data exchange that enabled data to be shared between multiple sources, integrated, analysed and visualised by a range of software tools. Its rapid and widespread acceptance showed that this was a much needed resource, but it also soon became apparent that version 1.0 was limited in scope and that the community required a broader, more flexible, format. We have presented PSI-MI XML2.5, with the increased expressive power, which allows more detailed and complete representation of interaction data, including the ability to express interactions involving any interactor type. The accompanying controlled vocabularies have been expanded and improved to further enable detailed annotation of interaction data. Tool development has continued in parallel and a suite of programming resources now exist to assist the user in producing and manipulating valid PSI-MI XML2.5 and MITAB2.5 files.

The original stated intention was to develop PSI-MI in a multi-levelled approach, with each level being backward compatible with earlier versions. Due to the complexity of interaction data already available in the public domain, this aim proved impossible to achieve, with a rewrite of version 1.0 proving necessary to enable all this information to be expressed within the interchange format. The restructuring of the XML format required a concomitant major reorganization of the controlled vocabulary. Version 1.0, and the accompanying CVs, has therefore been frozen and is maintained in support of users who have already started to use this format, but it is recommended that all users switch to the more versatile version 2.5 as soon as possible. Whilst the controlled vocabulary will continue to quickly adopt to the needs of the community, the PSI-MI XML2.5 format is expected to stay stable at least for two to three years, as with version 1.0, which was published in early 2004.

The benefits of PSI-MI adoption are clear, including better cooperation between public domain databases, culminating in the formation of the IMEx consortium to share the load of collecting vast amounts of interactions from the literature. Scientists can now download and combine the contents of multiple databases to populate a local resource using a single data format, enabling easy comparison with experimental results generated in their own laboratory.

There is still much work to be undertaken in further developing both the controlled vocabularies and accompanying tools. A PSI common query interface API for all participating databases, PSICQUIC is under development. Work is also in hand to investigate use of the current format to enable the use of the PSI-MI XML2.5 format for transferring information on antibody properties such as specificity, cross-reactivity, affinity and avidity. This will require extensions to many of the current controlled vocabularies such as those describing interactors and also extending the current kinetic parameters.

The PSI effort requires support from the user community to continue to expand, thus is actively seeking input and advice from all quarters. Anyone wishing to become involved is invited to visit the PSI website [14], to participate in the discussion groups listed, and to contribute to the further development of community standards for proteomics data.

Availability and requirements

The PSI XML2.5 schema, documentation, and tools are freely available from the PSI MI website [8]. The entries throughout the text can be accessed by appending them to the root URL http://www.ebi.ac.uk/ontology-lookup/?termId=xx:XXXX where xx is the two letter code before the colon and XXXX is the four digit code after the colon.

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

SK, SO contributed to the schema development, controlled vocabulary development, the tools development and the writing of the manuscript. LM-P, AC, MO, VS, LS, JN contributed to the schema development, controlled vocabulary development and the tools. BA, AFQ, NV, GDB, EC contributed to the tool development. IX, JW, DS, JJS, SM, MEC, JA, PW, AC-A contributed to the schema development and the controlled vocabulary development. MT, MG contributed to the schema development. HH, MV, CH, DE, GC, RA conceived of the concept. All authors read and approved the final manuscript.

Additional material

Additional file 1

PSI-MI XML2.5 example*. Data from [26] represented in PSI-MI XML2.5 format in compact form.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-5-44-S1.xml]

Additional file 2

PSI-MI XML2.5 example in HTML*. Data from [26] represented in PSI-MI XML2.5 format in compact form, converted to HTML.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-5-44-S2.html]

Additional file 3

PSI-MI XML2.5 example*. Data from [26] represented in PSI-MI XML2.5 format in expanded form.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-5-44-S3.xml]

Additional file 4

PSI-MI XML2.5 example in HTML*. Data from [26] represented in PSI-MI XML2.5 format in expanded form, converted to HTML.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-5-44-S4.html]

Additional file 5

MITAB2.5 example*. Data from [26] represented in MITAB2.5 format.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-5-44-S5.txt]

Acknowledgements

This is a community effort with voluntary contributions from many organisations not directly funded for this work. In addition, this work has been supported by the European Commission under FELICS, contract number 021902 (RII3), within the Research Infrastructure Action of the FP6 'Structuring the European Research Area' Programme; by the European Commission under the INTERACTION PROTEOME grant, by grant GM070064 from the NIH to MKG; and by the DOE (grant DE-FC03-02ER63421) and NIH (grant 1 R01 GM071909) to LS, JN, and DE. CH, JJS, and SM were funded by the BIND project grants from the Ontario Research & Development Challenge Fund and by Genome Canada through the Ontario Genomics Institute, as well as by the Economic Development Board of Singapore. We acknowledge Michael Müller for his help on the development of the Java XML Parser.

References

1. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, *et al.*: **The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data.** *Nat Biotechnol* 2004, **22:**177-183.
2. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, *et al.*: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33:**D418-424.
3. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32:**D449-451.
4. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, *et al.*: **IntAct – open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35:**D561-565.
5. Chatr-Aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res* 2007, **35:**D572-574.
6. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast.** *Nucleic Acids Res* 2006, **34:**D436-441.
7. **Hybrigenics** [http://www.hybrigenics.fr]
8. **The HUPO Proteomics Standards Initiative, Molecular Interaction work group** [http://www.psidev.info/index.php?q=node/60]
9. Bradshaw RA, Burlingame AL, Carr S, Aebersold R: **Reporting protein identification data: the next generation of guidelines.** *Mol Cell Proteomics* 2006, **5:**787-788.
10. Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Gorg A, Hecker M, Huber LA, Langen H, Link AJ, Paik YK, *et al.*: **Guidelines for the next 10 years of proteomics.** *Proteomics* 2006, **6:**4-8.
11. Jones AR, Pizarro A, Spellman P, Miller M, FuGE Working Group: **FuGE: Functional Genomics Experiment Object Model.** *OMICS* 2006, **10:**179-184.
12. **Open Biomedical Ontologies** [http://obo.sf.net]
13. **The IntAct validator** [http://www.ebi.ac.uk/intact/validator]
14. **The HUPO Proteomics Standards Initiative** [http://www.psidev.info]
15. **BioPAX** [http://www.biopax.org]
16. **BioPAX wiki** [http://www.biopaxwiki.org/cgi-bin/moin.cgi/PSI-MI_Conversion]
17. The UniProt Consortium: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35:**D193-197.
18. **ChEBI – Chemical Entities of Biological Interest** [http://www.oxfordjournals.org/nar/database/summary/646]
19. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK: **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.** *Nucleic Acids Res* 2007, **35:**D198-201.
20. Cote RG, Jones P, Apweiler R, Hermjakob H: **The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries.** *BMC Bioinformatics* 2006, **7:**97.
21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
22. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34:**D535-539.
23. **The International Molecular Exchange Consortium(IMEx)** [http://imex.sf.net]
24. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35:**D21-22.
25. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, *et al.*: **The Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx).** *Nat Biotechnol* 2007, **25:**894-898.
26. Hobson SD, Rosenblum ES, Richards OC, Richmond K, Kirkegaard K, Schultz SC: **Oligomeric structures of poliovirus polymerase are important for function.** *EMBO J* 2001, **20:**1153-1163.