nature biotechnology

PERSPECTIVE

The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data

Henning Hermjakob[1,22], Luisa Montecchi-Palazzi[1,2,22], Gary Bader[3], Jérôme Wojcik[4], Lukasz Salwinski[5], Arnaud Ceol[2], Susan Moore[6], Sandra Orchard[1], Ugis Sarkans[1], Christian von Mering[7], Bernd Roechert[8], Sylvain Poux[8], Eva Jung[9], Henning Mersch[1,10], Paul Kersey[1], Michael Lappe[1], Yixue Li[11], Rong Zeng[12], Debashis Rana[13], Macha Nikolski[19], Holger Husi[15], Christine Brun[14], K Shanker[17], Seth G N Grant[15], Chris Sander[3], Peer Bork[7], Weimin Zhu[1], Akhilesh Pandey[17], Alvis Brazma[1], Bernard Jacq[14], Marc Vidal[18], David Sherman[19], Pierre Legrain[4], Gianni Cesareni[2], Ioannis Xenarios[20], David Eisenberg[5], Boris Steipe[21], Chris Hogue[6,21] & Rolf Apweiler[1]

A major goal of proteomics is the complete description of the protein interaction network underlying cell physiology. A large number of small scale and, more recently, large-scale experiments have contributed to expanding our understanding of the nature of the interaction network. However, the necessary data integration across experiments is currently hampered by the fragmentation of publicly available protein interaction data, which exists in different formats in databases, on authors' websites or sometimes only in print publications. Here, we propose a community standard data model for the representation and exchange of protein interaction data. This data model has been jointly developed by members of the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO), and is supported by major protein interaction data providers, in particular the Biomolecular Interaction Network Database (BIND), Cellzome (Heidelberg, Germany), the Database of Interacting Proteins (DIP), Dana Farber Cancer Institute (Boston, MA, USA), the Human Protein Reference Database (HPRD), Hybrigenics (Paris, France), the European Bioinformatics Institute's (EMBL-EBI, Hinxton, UK) IntAct, the Molecular Interactions (MINT, Rome, Italy) database, the Protein-Protein Interaction Database (PPID, Edinburgh, UK) and the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING, EMBL, Heidelberg, Germany).

In almost every domain of molecular biology, efficient access to databases presenting the current status of factual knowledge is essential to efficient research planning, avoidance of duplicated work and successful interpretation of experimental results. However, the degree of database maturity varies significantly from domain to domain. For nucleotide sequence data and macromolecular structures, well-established databases[1–4] exist, and the data are synchronized among the major data providers. In addition, the public databases essentially reflect the complete publicly available knowledge, because data producers are strongly encouraged to submit their data to a public database before publication. In the much younger field of microarray-based gene expression data, the Microarray Gene Expression Data

[1]European Bioinformatics Institute, EBI-Hinxton, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [2]Department of Biology, University of Rome 'Tor Vergata', Via della Ricerca Scientifica, 00133 Rome, Italy. [3]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 460, New York, New York 10021, USA. [4]Hybrigenics SA, 180 avenue Daumesnil, 75012 Paris, France. [5]Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, 611 Young Drive East, Los Angeles, California 90095, USA. [6]Samuel Lunenfeld Research Institute, 1060-600 University Avenue, Toronto, Ontario M5G 1X5, Canada. [7]European Molecular Biology Laboratory, Structural and Computational Biology Program, Meyerhofstrasse 1, D–69117, Heidelberg, Germany, and Max Delbrück Center for Molecular Medicine, Department of Bioinformatics, PO Box 740238, 13092 Berlin-Buch, Germany. [8]Swiss Institute for Bioinformatics, CMU-Rue Michel-Servet 1, 1211 Geneva, Switzerland. [9]LG Informatics, Aventis Pharma Deutschland, Königsteiner Strasse 10, 65812 Bad Soden, Germany. [10]Faculty of Technology, University of Bielefeld, Universitätsstrasse 25, D-33615 Bielefeld, Germany. [11]Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 319 Yueyang Road, Shanghai, Peoples Republic of China 200031. [12]Proteomics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, Peoples Republic of China 200031. [13]Department of Genetics, Cambridge University, Downing Street, Cambridge, CB2 3EH, UK. [14]CNRS Laboratoire de Génétique et Physiologie du Développement–IBDM, 31, chemin Joseph Aiguier, 13402 Marseille, Cedex 9, France. [15]Division of Neuroscience, University of Edinburgh, 1 George Square, Edinburgh EH8 9JZ, UK. [16]Institute of Bioinformatics, International Tech Park, Whitefield Road, 560 066 Bangalore, India. [17]McKusick-Nathans Institute of Genetic Medicine and Department of Biological Chemistry, Johns Hopkins University, 600 N. Wolfe Street, Baltimore, Maryland 21287, USA. [18]Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Smith 858, 44 Binney Street, Boston, Massachusetts 02115, USA. [19]Laboratoire Bordelais de Recherche en Informatique CNRS UMR 5800, Domaine Universitaire, 351, cours de la Libération, 33405 Talence Cedex, France. [20]Serono International SA, Chemin des Aulx 14, CH-1228 Plan-les Ouates, Geneva, Switzerland. [21]Departments of Biochemistry and Molecular and Medical Genetics, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto M5S 1A8, Canada. [22]These authors have contributed equally to this work. Correspondence should be addressed to H.H. (hhe@ebi.ac.uk).

Published online 30 January 2004; doi:10.1038/nbt926

Box 1 Graphical representation of XML document structure

XML is a format to structure, describe and interchange data. For a detailed description of XML and links to tutorials, see http://www.w3c.org/XML/.

Example of a minimal XML document:

```
<protein>
      <name> Innexin Inx6 </name>
      <sequence length="481">
MYAAVKPLSNYLRLKTVRIY</sequence>
</protein>
```

An XML document consists of elements. Elements begin with the element name in angle brackets '<element>' and end with '/' and the element name in angle brackets '</element>'. These two tags enclose the content of the element, data and/or further elements. Elements may also have attributes, which are shown within the opening tag, like the length attribute in the example above.

The structure of an XML document can be described by an XML schema (http://www.w3c.org/XML/Schema), which has a similar function to the grammar in a natural language; it describes which elements are allowed in which parts of an XML document. To graphically document the PSI MI XML schema, we have used XMLSpy 5 (http://www.xmlspy.com/).

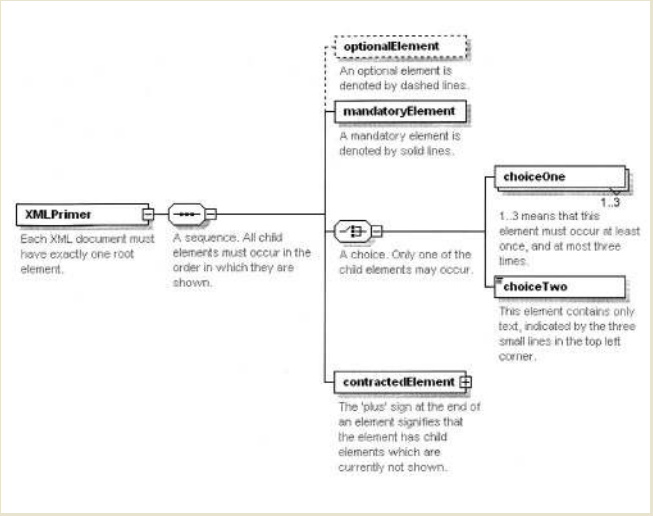

Figure 5 Graphical representation of XML document structure. The figure displays all the graphical elements that are used in **Figures 1** and **2** to describe the structure of the PSI MI format.

(MGED) Society has successfully developed the 'Microarray Gene Expression-Markup Language' (MAGE-ML)[5] and the 'Minimum Information About a Microarray Experiment' (MIAME)[6] standards, and many journals now encourage authors to support transcriptome publications with MIAME-compliant data[7]. In the case of proteomics, no generally accepted databases exist yet, but an increasing awareness of the importance of data standardization is reflected in community efforts and publications, in particular the mass spectrometry work group of the PSI[8], and the PEDRo model, which aims to describe proteomics experiments in sufficient detail to allow users to recreate any of the experiments whose results are stored in the format[9].

For protein interactions, several well-established databases exist, in particular the BIND[10], the DIP[11], the Comprehensive Yeast Genome Database (CYGD)[12], the MINT database[13] and STRING[14]. Although they are well documented, and, for example, the detailed BIND specification[15] has been published, the databases are not synchronized with each other, and their data formats are incompatible. Additionally, company and academic websites provide supplementary data from publications in yet more formats. To collect all publicly available protein interaction data, a time-consuming process of reformatting and mapping must be undertaken. As an initial step to improve this situation, we propose the Molecular Interaction (MI) extensible markup language (XML) format developed by the HUPO PSI as a community standard for the representation and exchange of protein interaction data.

The PSI was founded at the HUPO meeting in Washington in April 2002 (ref. 16), with the aim of developing standards for the representation of proteomics data, initially focusing on mass spectrometry and protein-protein interactions. The PSI MI XML format presented here has been jointly developed by the members of the protein interaction work group during two meetings hosted by the European Bioinformatics Institute (Hinxton, UK)[17,18], and through subsequent open collaborative development.

The PSI MI format is a database-independent exchange format. The supporting data providers intend to offer data downloads and analysis tool results in PSI MI format, so that end users can more easily combine data from different sources for performing or storing results from their own data analysis. As a long-term aim, we will explore the possibility of regular data exchange between data providers to create a network of synchronized protein interaction databases.

The PSI MI format is being developed using a multi-level approach similar to that used by the Systems Biology Markup Language[19]. Level 1, presented here, provides a basic format suitable for representing the majority of all currently available protein-protein interaction data. It allows the representation of both binary and more complex interactions, and the classification of experimental techniques and conditions. To allow efficient development and implementation of the PSI MI standard, Level 1 does not contain detailed data on interaction mechanisms or full experimental descriptions. Though these data are essential to the full understanding of cellular function, the current availability of this kind of data is often very limited. Subsequent PSI MI levels will encompass more detail, and will add support for additional interacting entities, in particular small molecules and nucleic acids.

Although a common data exchange format is a key requirement for an efficient exchange of protein interaction data, it does not by itself guarantee data compatibility. It is essential to ensure standardized use of the data attributes through documentation and controlled vocabularies. The PSI MI format contains detailed documentation within the XML schema itself, which is automatically extracted as an easily accessible web page and accompanied by a detailed documentation on the PSI MI context (http://psidev.sourceforge.net/mi/xml/doc/user/). To standardize the contents of data attributes, the PSI MI format makes extensive use of controlled vocabularies or ontologies. External systems, such as the Gene Ontology[20] and the US National Center for Biotechnology Information (Bethesda, MD, USA) taxonomy, are referenced where possible. Detailed controlled vocabularies have been developed for the PSI MI format for several key protein interaction data attributes, such as experimental method.

Structure of a PSI MI record

Box 1 provides a minimal introduction to XML and **Figure 5** introduces the graphical elements we use to visualize the structure of the PSI MI format. **Figure 1** presents the structure of the PSI MI format itself. **Figure 2** shows an example of an XML file.

The root element of a PSI MI XML file is the 'entrySet,' which contains one or more entries (see **Fig. 1**). Each 'entry' is a self-contained unit. This allows easy concatenation of the contents of multiple files into a single file by simply adding all the 'entry' units into an 'entrySet.'

Each 'entry' describes one or more protein interactions. The 'source' element describes the source of the entry, normally the data provider. It optionally contains a release number and a release date assigned by the data provider.

The 'availabilityList' provides statements on the availability of the data, usually copyright statements. In the current version, the availability statements are free text. The PSI MI format might later be extended to provide predefined availability statements.

The 'experimentList' contains experiment descriptions ('experimentDescription'). Each 'experimentDescription' describes one set of experimental parameters, usually associated with a single publication. In large-scale experiments, normally only one parameter, often the bait (protein of interest), is varied across a series of experiments. The PSI MI format describes the constant parameters (*e.g.*, experimental technique) in an 'experiment Description,' whereas the variable parameters (*e.g.*, the bait) are described in the 'interaction' element, which is part of the 'interactionList' (see below).

The 'interactorList' describes a set of interactors participating in an interaction. In the current version of the PSI MI standard, interactors are proteins. We plan to extend this to other types, for example small molecules, in future versions. The 'proteinInteractor' element describes the 'normal' form of a protein commonly found in databases like Swiss-Prot and TrEMBL[21], consisting of the administrative data, such as name, cross-references, organism and amino acid sequence.

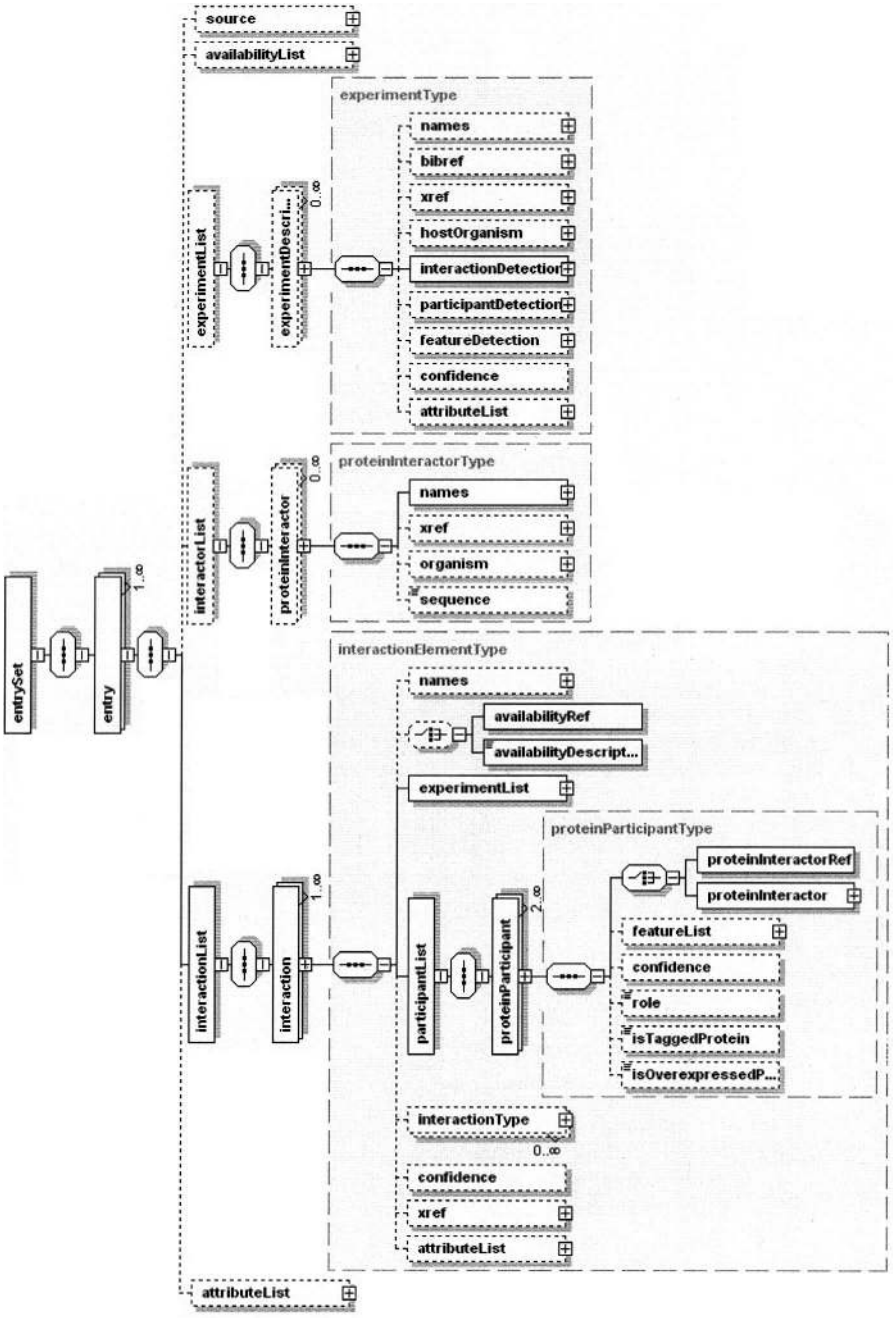

Figure 1 Graphical representation of the PSI MI format. Some elements have been omitted for clarity (indicated by a '+' in a rectangular box), and some elements have been rotated to provide a compact figure. The symbols are described in **Figure 5**. The major PSI MI elements are described in the text.

The 'interactionList' contains one or more 'interaction' elements. This is the core of the PSI MI format and the only mandatory element of an entry. Each 'interaction' contains an 'availability Description' (a description of the data availability, which may be a copyright statement), and a description of the experimental conditions under which it has been determined ('experimentList'). An interaction also contains a 'confidence' attribute. Different measures of confidence in an interaction have been developed, for example, the paralogous verification method[22], and the Protein Interaction Map (PIM) biological score[23]. To accommodate different methods, the confidence element contains a unit attribute and the confidence value itself.

Each interaction has a 'participantList,' containing two or more 'proteinParticipant' elements (that is, the proteins participating in the interaction). Each 'proteinParticipant' element contains a description of the molecule in its native form, either by reference to an element of the 'interactorList,' or directly in an 'proteinInteractor' element.

Additional elements of the participant element describe the specific form of the molecule in which it participated in the interaction. The 'featureList' describes sequence features of the protein, normally

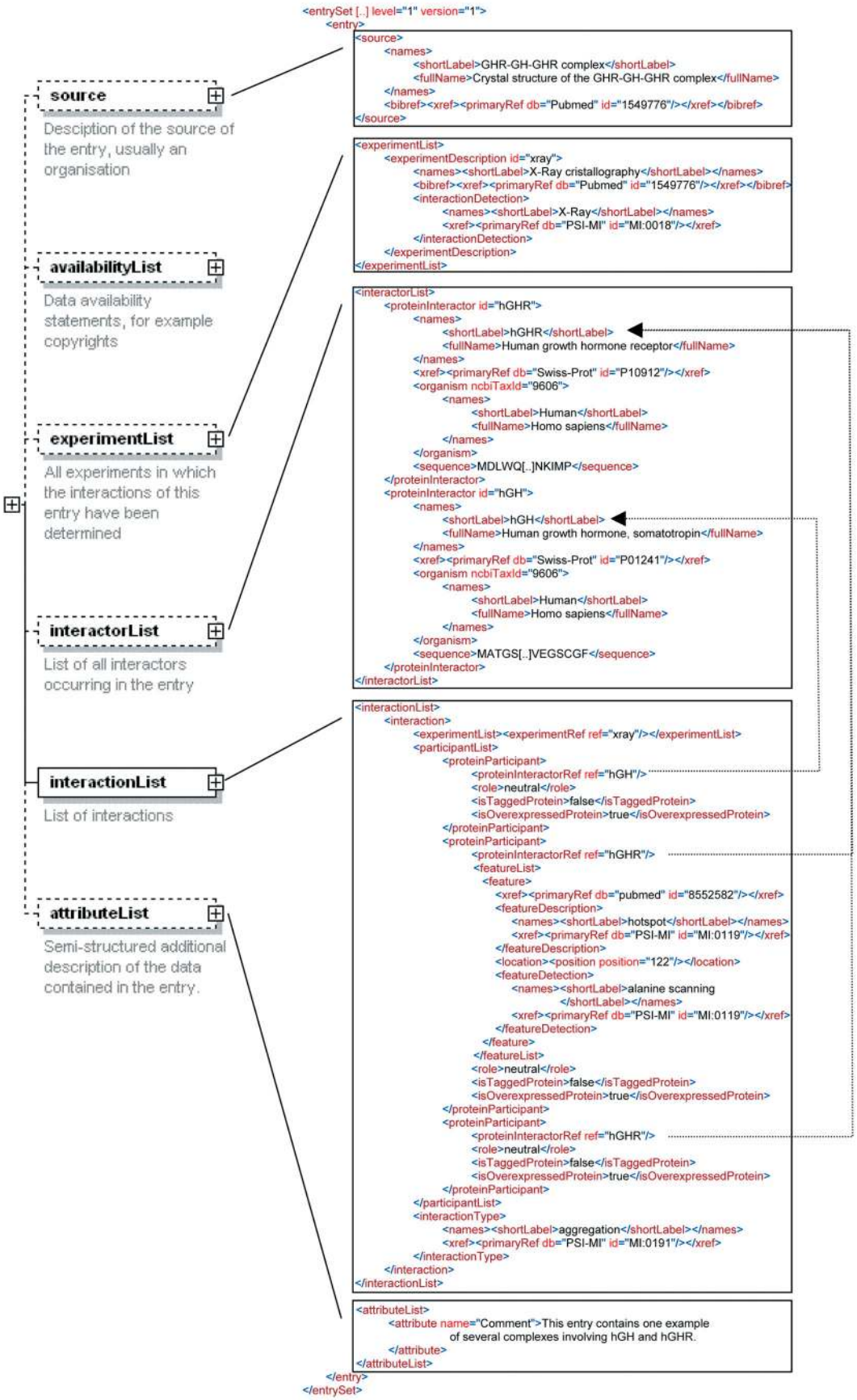

Figure 2 PSI MI example file. The figure shows a complete PSI MI file in compact form and how sections in the file correspond to the structural elements shown in **Figure 1**. The described heterocomplex consists of one human growth hormone (hGH) and two human growth hormone receptor (hGHR) molecules. The arrows to the right indicate the use of references in the compact form. The hGHR is only described once in the 'interactorList,' but is referred to twice in the 'interaction' element. In the expanded form of the PSI MI format, each reference to a protein is replaced by the full protein description. As an example of a feature list, the first instance of hGHR has a 'hotspot' annotated (a residue essential for binding).

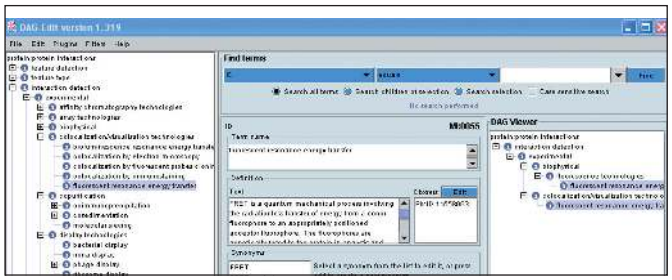

Figure 3 'Interaction detection' controlled vocabulary. The controlled vocabulary for describing methods for the detection of interactions is shown partially expanded in the DAG-Edit tool (http://www.geneontology.org/).

binding domains or post-translational modifications relevant for the interaction. The 'role' describes the particular role of the protein in the experiment—usually whether the protein was a bait or prey.

The 'attributeList' elements are placeholders for semi-structured additional information the data provider might want to transmit. They contain simple tag-value pairs and provide an easy mechanism to extend the PSI MI format. For each new level of the format, we will survey frequently used additional tags in attribute lists for additional element inclusion.

The PSI MI format can be used in two forms: compact and expanded. In the compact form, all interactors (proteins), experiments and availability statements are described once in the respective list elements, and then referred to only by reference from the individual interactions in the 'interactionList.' The compact form allows a concise, nonrepetitive representation of the data, in particular for large data sets.

In the expanded form, all proteins, experiments and availability statements are described directly in the interaction element. As a result, each interaction is a self-contained element providing all necessary information. The expanded form results in larger files, but is more suitable for conversion to displayed data (*e.g.*, hypertext markup language (HTML) pages). The sample file in **Figure 2** is shown in compact form.

Controlled vocabularies

The PSI MI format is designed for data exchange by many data providers. It is therefore important to ensure that both the data format (syntax) and the meaning of the data items (semantics) are consistent and well defined. Without the standardization of data items as part of a community standard, data sets that are generated by the combination of data from different sources will quickly become difficult to search and to use. For example, the terms 'yeast two-hybrid,' 'Y2H' and '2H' are all used in the literature to refer to the yeast two-hybrid technology, making it difficult to retrieve this experimental subset. To address this problem, we use controlled vocabularies instead of free-text attributes where possible. Five controlled vocabularies have been developed for Level 1 of the MI format: 'interaction type,' 'sequence feature type,' 'feature detection,' 'participant detection,' and 'interaction detection.'

The controlled vocabularies are provided in Gene Ontology format and are linked from the Open Biological Ontologies (OBO) site (http://obo.sourceforge.net/). The name space used as a prefix for the OBO identifiers is 'MI.' All terms have definitions and, whenever appropriate, are supported by literature references. The controlled vocabularies have a hierarchical structure, higher level terms being more general than lower level terms. This has advantages for both annotation and querying of the data. Annotation can be done on the desired level of detail, a phosphorylation can be annotated as 'phosphorylation' (MI:0170), or if known, as the detailed type of phosphorylation, for example using the subterm '3′-phospho-L-histidine' (MI:0175). A querying tool can take advantage of the hierarchical structure of the controlled vocabularies, and return all protein objects that have the term 'phosphorylation' or any of its subterms annotated.

The attributes described by terms from the controlled vocabularies are not independent of each other; term selection for one attribute may limit choice for another attribute. We do not model these semantic dependencies explicitly for PSI MI Level 1, for simplicity and ease of use, and let the data providers control the data integrity. However, we invite suggestions from the community for explicit modeling of these interdependencies in a full ontology in the next level of the PSI MI format.

The controlled vocabulary 'interaction type' describes the type of connection between molecules. As PSI MI Level 1 goals are restricted to protein-protein physical interactions, this controlled vocabulary is currently populated only by a single term, 'aggregation.' For future releases of PSI MI, we plan to add additional terms, such as 'enzymatic modification.'

The 'feature type' contains four main protein sequence properties that are relevant for the binding of the interacting proteins: binding site, post-translational modification (PTM), mutation and hot spot. We list 66 residue specific post-translational modifications, which are cross-referenced with the RESID database (http://www.ncifcrf.gov/RESID/)[24].

To describe the method by which a feature has been detected, a 'feature detection' term can be used. The 'feature detection' method is the first of the three controlled vocabularies concerning experimental technologies. If an experimental method is used in more than one context, it is defined only once. For instance, the term 'X-ray' is part of both the 'interaction detection' and 'feature detection' controlled vocabularies.

To model the experimental procedure that supports the interaction itself, we use two attributes, the 'interaction detection' and the 'participant detection.' The 'interaction detection' controlled vocabu-

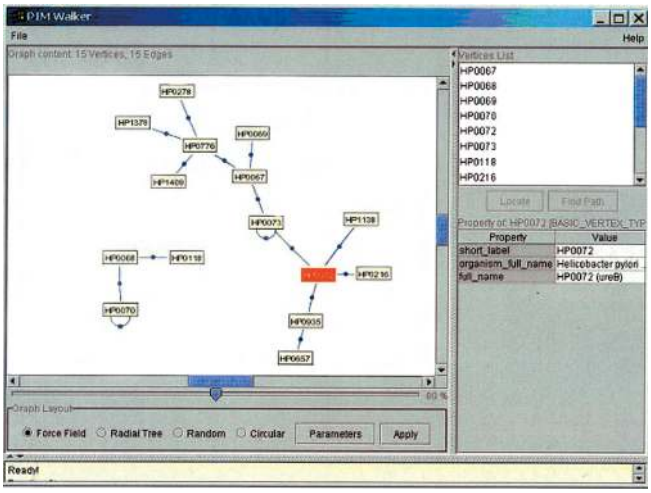

Figure 4 The PIMWalker network visualization tool. PIMWalker graphically displays the interaction network described in a PSI MI file. The network is represented as a graph (vertices are interactors and edges are interactions) and can be displayed using different layouts (force field, radial tree, circular). The user can search for paths between two interactors in the network and display all the attributes of interactors and interactions. PIMWalker is available from http://pim.hybrigenics.com/pimwalker/.

lary lists technologies that can be used to infer that two or more proteins form a molecular aggregate. This vocabulary of more than 80 terms has a hierarchical structure based on a limited number of high-level terms that group similar methods and reflect commonly used classifications and technical distinctions. As one method may be a specialization of more than one technology, a term may have more than one parent. For example, the fluorescence resonance energy transfer[25] method (MI:0055) is both a fluorescence technology (MI:0051) and a colocalization/visualization technology (MI:0023).

Figure 3 shows the 'interaction detection' controlled vocabulary partially expanded in DAG-Edit (http://www.geneontology.org/doc/GO.tools.html), a tool to manipulate hierarchical, controlled vocabularies. The 'participant detection' controlled vocabulary lists >20 methods commonly used to establish the identity of the interacting partners, for example peptide mass fingerprinting (MI:0082).

The combination of controlled vocabulary terms, together with the binary elements 'isOverexpressedProtein' and 'isTaggedProtein' achieves a very compact, but highly expressive experiment description, enabling meaningful searches and data analysis. The controlled vocabularies described here are not static; they will be maintained and updated by the HUPO PSI workgroup to reflect new experimental methodologies, or requirements from the community, in a manner similar to the maintenance of the Gene Ontology. This will ensure consensus on the inclusion of new terms by the user community, a high degree of flexibility to define terms when they are needed, and the avoidance of vague or ambiguous categories such as 'other methods.' Any requests and contributions should be directed to the mailing list: psidev-vocab@lists.sf.net.

Available data, support tools and submissions

A set of PSI MI formatted sample files is available from http://psidev.sf.net/mi/xml/sampleData/. This page shows both XML data files, and human-readable HTML versions of these files. Static or dynamically generated data sets are available from several resources, in particular DIP, HPRD, Hybrigenics, IntAct and MINT.

Handling of PSI MI formatted interaction data is supported by several tools. Extensible stylesheet language transformation (XSLT) scripts allow conversion between the compact and expanded data representation, and the conversion to HTML, whereas the Java-based Psimaker tool converts between tab-delimited and PSI formatted data. The Cytoscape, ProViz (Protein Interaction Visualization) and Hybrigenic's PIMWalker (see **Fig. 4**) interaction network analysis tools already support PSI MI formatted input data.

We strongly encourage submission of experimental protein interaction data to one of the PSI partner sites. BIND, DIP, IntAct and MINT all currently accept PSI MI formatted submissions by e-mail, whereas HPRD offers a web-based submission tool.

We expect these resources to be the nucleus of a rich set of PSI MI resources developed by the scientific community. For a full list of currently available data sources, tools and submission sites please see http://psidev.sourceforge.net/mi/xml/doc/user/.

DISCUSSION

In this article, we have presented the HUPO PSI MI data exchange format. This format has been jointly developed by members of the HUPO PSI, which includes among others the BIND, Cellzome, Dana Faber Cancer Institute[26], DIP, HPRD[27], Hybrigenics, IntAct[28], MINT and PPID[29] protein interaction data providers. We expect the definition of a common data format to both significantly facilitate the comparative analysis of protein interaction data and promote data exchange between data producers, databases, journals and end users.

In addition to a standard data format, a key requirement for the successful comparative analysis of data from different sources is the standardization of terminology. Together with the XML format, we present a set of controlled vocabularies that provide well-defined terms for key attributes of the description of protein interactions, thus providing a framework for the standardization of not only the format, but also the contents of PSI MI formatted data.

The PSI MI format is developed in a multi-level approach. Level 1, presented here, provides a compact format expressive enough to represent the vast majority of currently publicly available protein interaction data. The relative simplicity of the format allows quick implementation by data providers, and rapid development of support tools, as exemplified by the software listed above. It also allows the usage and support of the MI standard by smaller organizations with moderate bioinformatics support. The PSI MI format will evolve towards more detailed descriptions and more complexity, but we plan to maintain Level 1 as a practical, easily accessible exchange format.

Future levels of the format will include a more detailed description of the interaction type and interaction process, as well as detailed experiment description. For the latter, we plan to share the sample and experiment description components of the MAGE-ML standard. We will also extend the format to include more types of interacting molecules, in particular RNA, DNA and small molecules. Although the format is explicitly limited to interaction data, it is being developed in consultation with the BioPAX consortium (http://www.biopax.org/) to allow compatibility with future pathway standards.

PSI developments are pursued in an open community process, all project resources are hosted at the open-source site (http://sourceforge.net/). We invite active participation in the further development of the PSI MI format and tool set. For news, mailing lists, and the latest versions, please see the PSI website (http://psidev.sf.net/).

ACKNOWLEDGMENTS

This work was supported partially by EU grant number QLRI-CT-2001-00015 under the Research and Technological Development program 'Quality of Life and Management of Living Resources'. The PSI meetings were supported by the Human Proteome Organization. The work in the University of Rome 'Tor Vergata' was supported by grants from Associazione Italiana per la Ricerca sul Cancro and grant GTF02011 from Telethon. M.L. is supported by the European Molecular Biology Laboratory International PhD program and Biotechnology and Biological Sciences Research Council grant 8/C19399. Y.L. and R.Z. are supported by grants 2001AA233031, 2002CB512801, 110CB510209. M.V.'s laboratory is supported by grants from the US National Cancer Institute and National Human Genome Research Institute. L.M.-P. would like to thank Jens Pedersen, Claudia Bagni, Benedetta Mattei, Elena Santonico, Federico Demasi and Michael Ashburner for contributions to the controlled vocabularies. Emmanuel Cézanne, Sébastien Cros, Claire Even, Nicolas Jolibert, Sandrine Marquès, Christophe Roumegous, Patrick Sablayrolles and René Thomas-Nelson contributed to the development of the PSI XSLT utilities. The collaborative development process has been facilitated by the infrastructure provided by Source Forge.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Miyazaki, S., Sugawara, H., Gojobori, T. & Tateno, Y. DNA Data Bank of Japan (DDBJ). *Nucleic Acids Res.* **31**, 13–16 (2003).
2. Stoesser, G. *et al.* The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids. Res.* **31**, 17–22 (2003).
3. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank. *Nucleic Acids Res.* **31**, 23–27 (2003).
4. Westbrook, J., Feng, Z., Chen, L., Yang, H. & Berman, H.M. The Protein Data Bank and structural genomics. *Nucleic Acids Res.* **31**, 489–491 (2003).
5. Spellman, P.T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, research0046.1–0046.9 (2003).
6. Brazma, A., *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).

7. Ball, C.A. Microarray Gene Expression Data (MGED) Society: standards for microarray data. *Science* **298**, 539 (2002).
8. Orchard, O., Hermjakob, H. & Apweiler, R. The Proteomics Standards Initiative. *Proteomics* **7**, 1374–1376 (2003).
9. Taylor, C.F. *et al.* A systematic approach to modeling, capturing and disseminating proteomics experimental data. *Nat. Biotechnol.* **21**, 247–254 (2003).
10. Bader, G.D., Betel, D. & Hogue, C.W.V. BIND, the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
11. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
12. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
13. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
14. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
15. Bader, G.D. & Hogue, C.W. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477 (2000).
16. Kaiser, J. Proteomics. Public-private group maps out initiatives. *Science* **296**, 827 (2002).
17. Orchard, S., Kersey, P., Hermjakob, H. & Apweiler, R. The HUPO Proteomics Standards Initiative meeting: towards common standards for exchanging proteomics data. *Comp. Funct. Genomics* **4**, 16–19 (2003).
18. Orchard, S. *et al.* Progress in establishing common standards for exchanging proteomics data: the second meeting of the HUPO Proteomics Standards Initiative. *Comp. Funct. Genomics* **4**, 203–206 (2003).
19. Hucka, M. *et al.* The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
20. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
21. Boeckmann, B. *et al.* The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
22. Deane, C.M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* **1**, 349–356 (2002).
23. Rain, J.-R. *et al.* The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
24. Garavelli, J.S. The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res.* **31**, 499–501 (2003).
25. Day, R.N., Periasamy, A. & Schaufele, F. Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus. *Methods* **25**, 4–18 (2001).
26. Reboul, J. *et al. C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**, 35–41 (2003).
27. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
28. Hermjakob, H. *et al.* IntAct—an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455 (2004).
29. Husi, H. & Grant, S.G. Construction of a Protein-Protein Interaction Database (PPID) for Synaptic Biology. in *Neuroscience Databases: A Practical Guide.* (R. Kotter, ed.) 1–62 (Boston/Dordrecht/London, Kluwer Academic Publishers, 2002).