Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

DEVELOPMENT AND IMPLEMENTATION OF THE PSI MI STANDARD FOR MOLECULAR INTERACTION

Samuel Kerrien

EMBL Outstation - Hinxton
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge, CB10 1SD
UNITED KINGDOM

Henning Hermjakob

EMBL Outstation - Hinxton
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge, CB10 1SD
UNITED KINGDOM

ABSTRACT

The pool of molecular interaction data is growing fast but nevertheless remains fragmented. Combining together data coming from heterogeneous sources is a crucial step towards a deeper understanding of the cell machinery. The Proteomics Standard Initiative offers mature standards (PSI-MI) to facilitate the exchange and analysis of Molecular Interaction data. After introducing the details of the latest version of the PSI-MI data model, we will present the implementation of PSI-MI in the IntAct project, which offers a platform for management and analysis of interaction data. Finally we will give some insight into realistically using molecular interaction data as a foundation for other research.

1 INTRODUCTION

In the past years, we have witnessed the amount of scientific publications reporting molecular interaction data increase considerably. The amount of data reported can vary from a single interaction to several thousands (e.g., Giot et al. 2003 reported more than 22.000 binary interactions). However, data is not reported in a consistent manner, often given as supplementary tables which do not allow systematic interpretation.

From these observations has arisen the Proteomics Standards Initiative (Hermjakob et al. 2004a), a work group of the Human Proteome Organization (HUPO) which jointly developed, amongst others, a Molecular Interaction standard (PSI-MI) aiming at facilitating data comparison and exchange.

The consortium brought together a representative sample of the community including experimentalist, data and software providers, machine vendors and publishers.

2 THE PSI STANDARD FOR MOLECULAR INTERACTIONS

The first version of the molecular interaction standard was published in 2004 and was designed to accommodate protein-protein interaction data (Hermjakob et al. 2004a). Keeping in mind that the standard would ultimately be used as a container for exchanging interaction data, the model was implemented as an XML Schema.

Despite its youth, the field of Interactomics is already vast and complex. The repertoire of experimental methods available to the scientific community is constantly growing as new protocols are being developed to overcome the shortcomings of their predecessors. PSI-MI does not aim at describing the experimental procedure associated with the resulting interaction data, though it aims at reporting which technologies were used to produce the result.

Should PSI-MI be the model of authority for molecular interaction, it must be flexible enough to cope with the continuous evolution of the data and particularly the experimental methods used to characterize the details of interactions. This was addressed by developing a hierarchical ontology and seamlessly integrating it in the data model. This approach has also the advantage to disambiguate the description of a concept, for instance, when referring to the interaction detection method ‘yeast two hybrid’, we encourage the user of the data model to specify as well an ontology identifier in addition to its name, so here: MI:0018.

The data model versioned 1.0 (aka. PSI-MI 1.0) has now been in use for more than two years and has been broadly accepted by many of the data providers (e.g., interaction databases) as well as software providers. In the following part, we are going to give a short descriptive introduction to the first PSI-MI 1.0 and then introduce the numerous changes that gave birth to PSI-MI 2.5.

2.1 XML Schema in a Nutshell

Extensible Markup Language (XML) is a standard for representing and exchanging data. A detailed specification is available from <http://www.w3.org/XML>.

Below is an XML document illustrating the syntax:

```
<root>
   <element>
      <subElement attribute="value" />
   </element>
   <otherElement>example</otherElement>
</root>
```

XML Schemas (see Figure 1) is a technology allowing to define the structure and the data types of an XML document. Users can define rules that the XML document has to implement. e.g., position of an element in the hierarchy, cardinality, and type of value held by an attribute. A detailed specification is available from <http://www.w3.org/XML/Schema>.

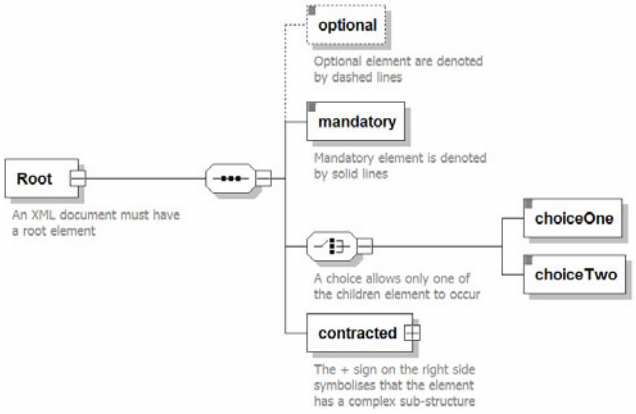

Figure 1: Structure of a Simple XML Schema (Generated Using XML Spy 2006 Home Edition)

XML is well supported by freely accessible software and libraries allowing to read and process an XML document, to validate a document against an XML Schema or even to transform a document into some other format.

All these advantages make XML Schemas a very good tool for building community standards for data exchange.

2.2 PSI-MI 1.0

The schema is composed of a list of entries, each of which is articulated around three main components: Experiment (`experimentDescription`), Protein (`proteinInteractor`) and Interaction (`interaction`).

Experiment - represents the experimental procedure that was used to discover molecular interactions. It includes the description of the `hostOrganism` in which the experiment was carried out and the protocols details such as interaction, interactor and feature detection method. An experiment can also contain references to scientific publications from which the interaction data originated.

Protein - describes the biological entity that was observed to be involved in an `interaction`. PSI-MI does not aim at giving an exhaustive description of the protein but rather to refer to a reference database such as the UniProt Knowledge (Apweiler et al. 2004) Base or NCBI RefSeq (Pruitt et al. 2005). A reference (called `xref` in the data model) is composed of a database (ontology term) and an identifier. Optionally an `organism` and a `sequence` can be added to further describe the protein.

Interaction - Provides details on the interaction itself, as well as the specific forms of the interacting proteins. To do so, it uses the definition of a protein encapsulated into a `participant`. While a `proteinInteractor` describes the abstract, "database" representation of a protein, a `Participant` describes the specific form of the protein in the `interaction`. For instance, the `Role` of one protein can be *bait* where the other(s) will be *prey(s)*. Four roles are allowed: *bait*, *prey*, *neutral* and *unspecified*. The participant can also describe accurately which part of the molecule's chain was physically interacting by using concepts called `feature` and `range` (e.g., amino acid 10 to 15 of protein sequence). It is also possible to specify if a protein was tagged or over-expressed.

Further optional elements of an interaction are annotation topics with free text comments, for example a description of the biological function of the complex.

The core components (protein, interaction, experiment, feature, etc.) share a set of common properties:

Identifier – This is solely used for internal referencing of a component and allows its reuse within the scope of an entry e.g. the definition of an experiment can be reused across many interactions.

Names – Allows the definition of a short memorable name, a long name as well as a list of synonyms.

Attributes – Should the schema not allow you to model some specific data, a system of attributes composed of key-value pairs allow the addition of structured data to an object.

In a nutshell, the PSI-MI 1.0 schema has managed to capture a large quantity of datasets by providing a generic and flexible way to describe protein-protein interaction. However as time passes, experimental protocols are becoming more complex and numerous, pipelines allowing discovering of other types of interacting molecule such as DNA, RNA, or small molecules have matured considerably. The reported details of interactions have also gained in complexity and depth: it is now getting more and more common to find numerical parameter of an interaction describing, for instance, its kinetics.

Cross references – Allow to define reference to external resource, e.g., an experiment may have a reference to a scientific publication, or a protein to a protein database.

We are now going to describe how the PSI-MI standard has adapted to the development of the data produced by experimentalists.

2.3 PSI-MI 2.5

Through a series of examples, we will illustrate the new extensions of PSI-MI 2.5 (see the data model in Figure 2).

Enforced use of identifier over free text – Throughout the schema, wherever we were trying to describe a term of the ontology we made possible the addition of the identifier of the term in addition to its name. A name can be misspelled, capitalized in different ways, etc. All these potential errors can be solved by providing a non ambiguous way to describe a term. For instance, when describing a protein reference:

```
<xref>
  <primaryRef db="uniprot"
              dbAc="MI:0486"
              id="P60953"/>
</xref>
```

Extended naming – Many of the major object types had the facility to be described using names and aliases. We have introduced a way to give a type to an alias using a controlled vocabulary. For instance when describing a protein, one can now give additional information such as gene name or synonym.

Generic interactor – In order to allow the representation of other types of `interactor` than proteins, `proteinInteractor` became `interactor` and was parameterized with a controlled vocabulary: `interactorType`. Thus when a new type of molecule is needed, we can simply extend the ontology without altering the schema.

List of participant detection methods – In some cases, experimentalists have used more than one method to identify the participants of an `interaction`. The schema now allows the definition of more than one by declaring them at the `participant` level, consequently overriding the one declared at the experiment level.

Experimental preparation – some experimental protocols require molecules to be engineered. PSI-MI 2.5 has introduced a way to describe the experimental treatment and status of interaction participants (e.g., expression level, method of delivery of the molecule into the cell, etc.).

Modeled interaction – It is not unusual to find reported in a scientific publication molecular interactions inferred on the basis of protein orthology (Persico et al. 2005). Should we need to model that information, an interaction's `participant` can describe a list of experimental interactors which refers to the ones that were used experimentally. A flag (`modeled`, which can take the values `true` or `false`) at the level of the `interaction` states to the user the nature of the data.

Confidence value – the concept of confidence value of interactions has been extended to represent more faithfully the data coming from scientific literature. Indeed, it happens that an experimentalist splits a set of interactions into multiple subclasses of various reliability. The complexity of such classification scales from a simple low/high confidence classification (Giot et al. 2003) to the more complex Hybrigenics Predicted Biological Score PBS® (Rain et al. 2001), which features five distinct classes. It can easily be modeled in PSI-MI 2.5 by specifying a list of `confidence` class at the experiment level and then annotate interactions using the predefined classes.

Numerical parameters – the advance of proteomics technologies make it easier for experimentalists to capture numerical details of molecular interactions such as the equilibrium dissociation constant of an interactor (aka. Kd) and many other kinetics values. A list of parameters can be stored at the level of an `interaction`, the `parameter` type (Kd, Ka, etc.) as well as the corresponding unit of measurement (concentration, time, etc.) are controlled by specific ontologies which were developed for that purpose.

Participant role – the role of an interaction participant can differ according to the technology used. Also, PSI-MI 1.0 was quite restrictive as only four roles were allowed and modifying this constraint would have altered the schema. PSI-MI 2.5 has split the role into experimental (e.g., bait, prey, etc.) and biological (e.g., enzyme, target, etc.) role but also made them generic by introducing controlled vocabularies.

Features – A `feature` is the description of the relevant subsequence of an interaction's participant. PSI-MI 2.5 has extended the concept by allowing some degree of fuzziness in the definition of the range's boundaries (called `featureRange`). A few examples of range are:

- Amino acid 3 to 15.
- From between 4 and 10 to 23.
- From 6 to less than 142.

Representation of complexes – The modeling of a molecular interaction can be convoluted, sometimes requiring the description of complex sub-units that are later assembled to form larger interactions. In order to cope with this requirement, the hierarchical build-up of molecular interactions was introduced. Interactions can be used as an interactor, and thus they can be reused in the context of other interactions.

In Figure 3, the left part shows a graphical representation of a complex assembly. The right part shows how the assembly was modeled using sub-interactions. The shapes labeled P1, P2, P3 are proteins and I1, I2 are interactions.

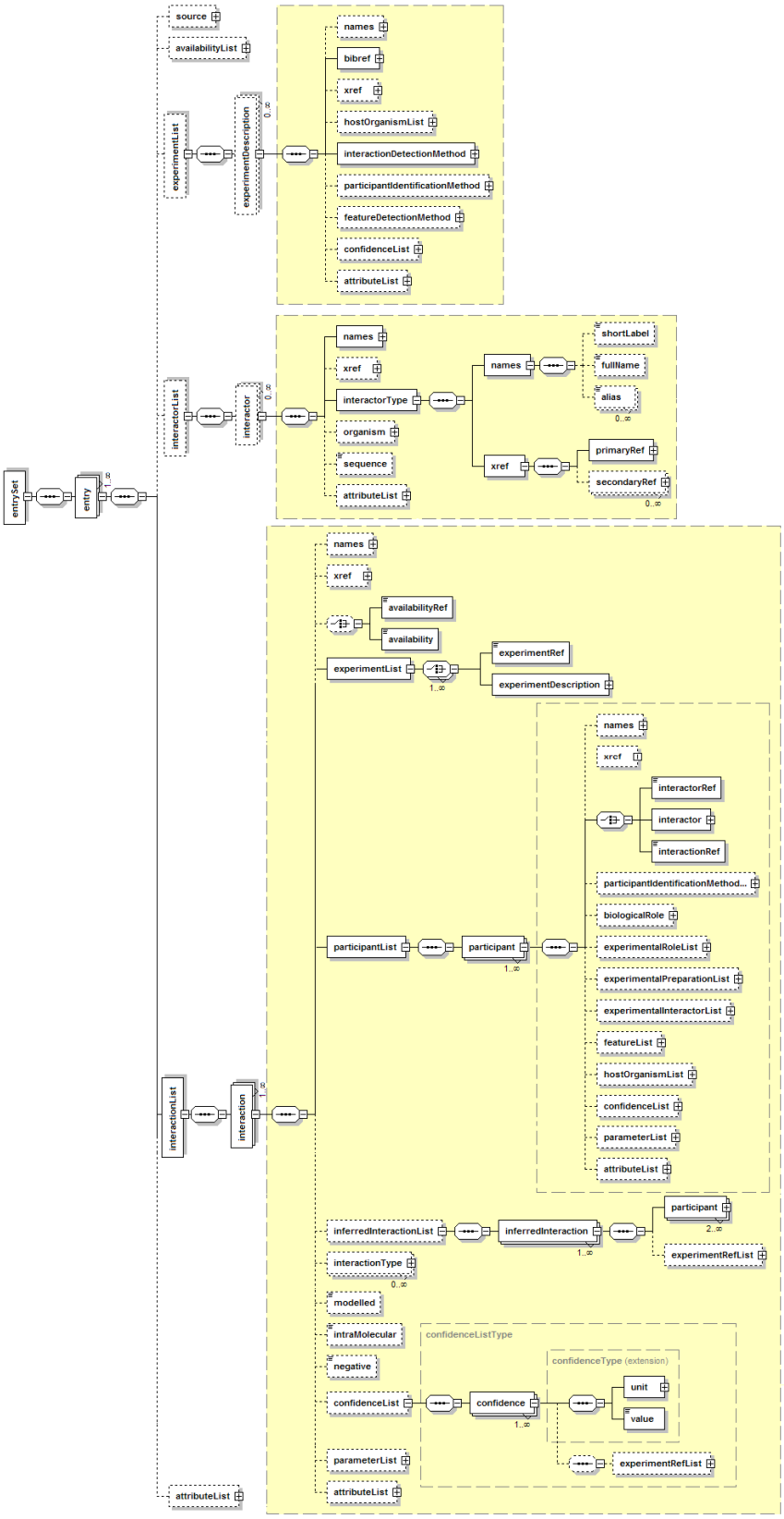

Figure 2: The PSI-MI 2.5 Data Model

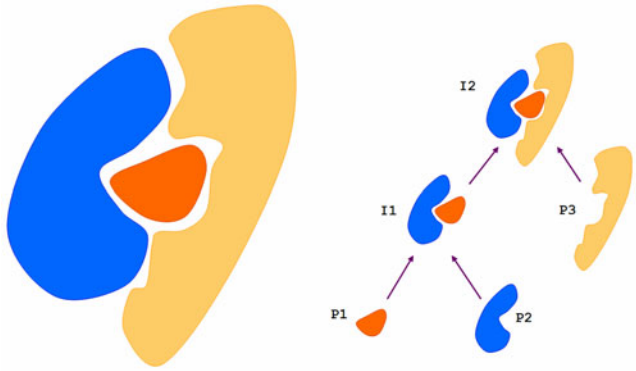

Figure 3: Representation of an Interaction Built Hierarchically

Expansion of controlled vocabularies – As one can deduce by observing the difference between PSI-MI 2.5 and its predecessor, controlled vocabularies lie at the heart of the data model and their integration has undeniably given extra flexibility and ease of maintenance to the data model. This explains the growth of the ontology from 5 to 16 classes and from 400 to over 700 terms.

PSI-MI has consequently succeeded in adapting to the evolution of the field of molecular interaction by providing ways to model data commonly generated by experimentalist. Furthermore, the use of controlled vocabularies has enabled the expansion of the data model without sacrificing its stability.

2.4 The Proteomics Standard Initiative Development Process

The PSI makes heavy use the of the Sourceforge (see <http://sourceforge.net>) infrastructure for hosting project web pages, mailing lists, bug and request trackers. They can all be accessed via the Sourceforge web site: <http://sf.net/projects/psidev>.

The development of the molecular interaction standard has two tracks:

Data model – The PSI-MI schema follows a biannual development cycle (more details in Figure 4), the goal being to avoid users having to continuously adapt their software to the evolution of the schema. Thus, over the past four years, only two major versions have been released (v1.0 and v2.5).

Figure 4: Roadmap Leading to the Release of a PSI-MI Schema

Ontology – The PSI controlled vocabularies are maintained continuously so they can adapt quickly to new technologies and user requests. Any user can submit new terms via the appropriate sourceforge request tracker and mailing list. Specialists in the field process user requests and strive to complete them within five working days. If the request is successful, the new term(s) is/are added to the master copy of the ontology and made available through the PSI web site.

PSI is currently making heavy use of the Open Biomedical Ontology (OBO) format. It is a well established language that benefits of numbers of supporting tools such as Dag-Edit or Obo-Edit (Gene Ontology Consortium 2006).

The Proteomics Standard Initiative is an open community, should you be interested in contributing in the development of its standards, you can find more information on the PSI web site <http://psidev.sourceforge.net>.

2.5 Tools

There is a number of tools supporting PSI-MI that have been developed over the past year. We are going to describe a few here, but a more exhaustive list can be found on the PSI web site <http://psidev.sf.net/mi/rel25>.

XML maker/flattener – Allows the intuitive conversion of a tabular format such as an Excel file into PSI-MI formatted data. It supports both PSI-MI 1.0 and 2.5.

PSI Schema Validator – Provides a generic framework for checking on the syntax and semantic use of data modelled after an XML schema. It currently implements an extension for PSI-MI 2.5, including a set of rules checking on the correct usage of controlled vocabularies.

Interaction network visualization – There are many visualization tools that are supporting PSI-MI 1.0 as an input format such as Proviz (Iragne et al. 2005), PIM Walker (Formstecher et al. 2005) and Cytoscape (Shannon et al. 2003). The latter being the only one supporting PSI-MI 2.5 so far.

2.6 INTACT – AN IMPLEMENTATION OF THE PSI MI 2.5 STANDARD

2.6.1 Aims

IntAct (Hermjakob et al. 2004b) is an extensible open source framework for molecular interactions. The project provides a public repository populated with experimental results from both direct submission and curated literature. Additionally, a set of applications is available to manage the IntAct repository and analyze the data. IntAct was developed from the beginning to support local installation, so one could easily install a local molecular interaction pipe-

line, import private datasets as well as a selection of the data available publicly from the European Bioinformatics Institute's repository (currently over 1300 publications curated to a high level of detail). Eventually it is possible to use the provided software to visualize and analyze the data.

2.6.2 Analysis Tool Example: Pay-As-You-Go

Exploring the interactome is one of the primary challenges in the post-genomic era. The topology information gained from the coverage of interaction space can shed light on the likely function and structure of proteins within the network (Vitkup et al. 2001). A key challenge in the planning of large scale experiments is the bait selection. IntAct provides an analysis tool allowing the generation of a list of the 'best baits' which are expected to yield the highest return on experimental effort - i.e. those proteins which form 'hubs' in the interaction network. These lists are generated using the Pay-As-You-Go strategy (Lappe and Holm 2004) which detects and prioritizes those proteins which have the highest likelihood of being hubs based on the current data within IntAct for various species. To illustrate the experimental effort which could be saved by using the strategy implemented, take the Human interactome as an example. It is estimated that 50,000 experiments would be required to cover the entire human interactome following a purely random strategy for the selection of targets. The information gained here is close to linear for each experiment. This could be drastically reduced to less than 10,000 for 90% coverage of the entire human interactome by using a near-optimal bait selection strategy. This would save years of experimental effort. This of course relies on the timely deposition of experimental data into the IntAct database in order that the Pay-As-You-Go algorithm remains up-to-date and effective. Should you be willing to submit data, please contact <datasubs@ebi.ac.uk> for curator assistance.

2.6.3 PSI-MI Integration

IntAct has integrated the PSI-MI data model as its core unit for data exchange, we are now going to discuss a few cases where the PSI-MI format is used in IntAct.

The software suite integrates a 2D visualization engine allowing the display of a molecular interaction network built from interactors selected by the user. One can also explore the network in the context of GO and InterPro (Mulder et al. 2005) annotation of its components. It is also possible to expand or contract a network to widen the context of the interaction studied. Should the user wish to analyze that same network in other software, it is possible to download the data in PSI-MI format and load it into a more interactive stand alone application such as Cytoscape (Shannon et al. 2003) or Proviz (Iragne et al. 2005).

All of IntAct's manually curated data are made available in both PSI-MI 1.0 and 2.5. In order to satisfy the user community, we have chosen to group the data in different ways: by species and by publication. IntAct's web interface allows direct download of data from a specific publication in both formats.

IntAct can be locally installed and interaction data can be added to a local repository by uploading PSI-MI data. The data file release on the IntAct FTP can be used for that purpose.

By choosing the PSI-MI standard as its main vector for handling molecular interaction data, IntAct has considerably increased its exposure to the community:

- One can use IntAct to aggregate data coming from multiple PSI-MI compliant data source.
- Users are free to use IntAct's high quality data in any PSI-MI compliant software.

Finally, IntAct is part of the International Molecular interaction Exchange consortium (IMEx) where the major data providers, including BIND (Alfarano et al. 2005), DIP (Salwinski, et al. 2004), MINT (Zanzoni et al. 2002) and MIPS (Guldener et al. 2006), have agreed to exchange their curated literature data using the PSI-MI 2.5 format. A strict guideline defining the use of the PSI-MI data model was agreed upon by all IMEx partners.

More information is available online at the IMEx web site: <http://imex.sourceforge.net>.

3 DISCUSSION

The Proteomics Standard Initiative gathers experimentalists, interaction databases, software providers, and publishers in the joint development of standards for data representation in proteomics. The PSI-MI standard has now been widely accepted for the representation of molecular interactions.

This is a clear step forward new discoveries, yet many challenges are still to be overcome. For instance, if we knew consistently the details of interactions such as cellular context (disease, etc.), kinetics, protein state (phosphorylation, etc.), post translational modifications and concentrations of its participants we could more realistically start thinking about modeling virtual cells and its response to a given stimulus. However, this data is still not very often made available in scientific literature and this may impede the development of accurate simulation models.

It is important to note that PSI-MI was specifically developed for data exchange, yet considering its ability to encompass very detailed information about molecular interactions, it becomes a potent candidate for simulation model and software. Given the appropriate set of controlled vocabularies and the flexibility of the PSI-MI schema, one could use the standard to model molecular interactions

coming out of a simulation model. Consequently, PSI-MI should not be seen as a competitor of SBML (Finney and Hucka 2003) and Cell ML (Lloyd et al. 2004) but complementary to these effort.

However, if one is to reuse interaction data for the purpose of other research, interaction reliability is of utmost importance as experimental results can be of varying quality. For instance, high throughput protein-protein interaction screens, such as those utilizing yeast two hybrid methods, increase the chance of identifying artifactual partners by testing exhaustively arbitrary protein-protein interactions. Those include the partners that can physically interact but that are never in close proximity to one another in the cell because of distinct subcellular localization or expression at different times during the life cycle. All these factors can lead to the observation of either false negatives (interactions that cannot be detected under the conditions used) or false positives (physical interactions without biological meaning). Fortunately, computational methods have been developed to help assessing quality of interaction data, amongst which the analysis of GO classification of interacting partners show that they tend to display similar features in term of function or cell localization (Schwikowski et al. 2000), and RNA expression profile and paralogous verification (Deane et al. 2002).

The quality of molecular interaction data is very likely going to increase as experimentalists strive to reach full coverage of the interaction space of a given organism. Meanwhile it is important to combine the existing data and increase the coverage of curated literature. The International Molecular Exchange consortium has emerged for that very purpose. It promises to provide a network of stable, synchronized and freely accessible databases, and will serve as a way to jointly capture all published molecular interaction data in a manner similar to the successful global collaborations for protein and DNA sequences and for macromolecular structures. This cumulative effort should result in an overarching repository that is broader in scope and deeper in information than any individual efforts and one that scientists can use to better understand issues of health and disease or in the development of new drugs and therapeutics.

ACKNOWLEDGMENTS

Funded by EU grant number QLRI-CT-2001-00015 under the RTD programme “Quality of Life and Management of Living Resources” and EU contract no 21902 “Felics - Free European Life-Science Information and Computational Services”.

REFERENCES

Alfarano, C., C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. Ouellette, and C. W. Hogue. 2005. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research* 33: D418-24.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. 2004. Uniprot: The universal protein knowledgebase. *Nucleic Acids Research* 32: D115-9.

Deane, C. M., L. Salwinski, I. Xenarios, and D. Eisenberg. 2002. Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics* 1: 349-56.

Finney, A., and M. Hucka. 2003. Systems biology markup language: Level 2 and beyond. *Biochemical Society Transactions* 31: 1472-3.

Formstecher, E., S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J. A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M. P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Rosse, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, and L. Daviet. 2005. Protein interaction mapping: A drosophila case study. *Genome Research* 15: 376-84.

Gene Ontology Consortium. 2006. The gene ontology (go) project in 2006. *Nucleic Acids Research* 34: D322-26.

Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M.

Rothberg. 2003. A protein interaction map of drosophila melanogaster. *Science* 302: 1727-36.

Guldener, U., M. Munsterkotter, M. Oesterheld, P. Pagel, A. Ruepp, H. W. Mewes, and V. Stumpflen. 2006. Mpact: The mips protein interaction resource on yeast. *Nucleic Acids Research* 34: D436-41.

Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler. 2004a. The hupo psi's molecular interaction format--a community standard for the representation of protein interaction data. *Nature Biotechnology* 22: 177-183.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. 2004b. Intact: An open source molecular interaction database. *Nucleic Acids Research* 32: D452-5.

Iragne, F., M. Nikolski, B. Mathieu, D. Auber, and D. Sherman. 2005. Proviz: Protein interaction visualization and exploration. *Bioinformatics* 21: 272-4.

Lappe, M., and L. Holm. 2004. Unraveling protein interaction networks with near-optimal efficiency. *Nature Biotechnology* 22: 98-103.

Lloyd, C. M., M. D. Halstead, and P. F. Nielsen. 2004. Cellml: Its future, present and past. *Progress in Biophysics & Molecular Biology* 85: 433-50.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C. H. Wu. 2005. Interpro, progress and status in 2005. *Nucleic Acids Research* 33: D201-5.

Persico, Maria, Arnaud Ceol, Caius Gavrila, Robert Hoffmann, Arnaldo Florio, and Gianni Cesareni. 2005. Homomint: An inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* 6: S21.

Pruitt, K. D., T. Tatusova, and D. R. Maglott. 2005. Ncbi reference sequence (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 33: D501-4.

Rain, J. C., L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. 2001. The protein-protein interaction map of helicobacter pylori. *Nature* 409: 211-5.

Salwinski, L., C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. 2004. The database of interacting proteins: 2004 update. *Nucleic Acids Research* 32: D449-51.

Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of protein-protein interactions in yeast. *Nature Biotechnology* 18: 1257-61.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* 13: 2498-504.

Vitkup, D., E. Melamud, J. Moult, and C. Sander. 2001. Completeness in structural genomics. *Nature Structural & Molecular Biology* 8: 559-66.

Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. 2002. Mint: A molecular interaction database. *FEBS Letters* 513: 135-40.

AUTHOR BIOGRAPHIES

SAMUEL KERRIEN is a software engineer working in the Proteomics Services of the European Bioinformatics Institute. He is primarily working on IntAct, a framework for molecular interactions. Since 2002, he has also been involved with the Proteomics Standard Initiative in the development of the Molecular Interaction standard (PSI-MI). More recently, he took part in the creation and development of the International Molecular Exchange Consortium (IMEx). His e-mail address is: <skerrien@ebi.ac.uk>.

HENNING HERMJAKOB, M.Sc., studied Bioinformatics at the University of Bielefeld, Germany. He worked at the German National Research Center for Biotechnology in Braunschweig and joined the Swiss-Prot team at the European Bioinformatics Institute in 1997. He is now managing the Proteomics Services Team at EBI, which provides databases and tools for the deposition, distribution and analysis of proteomics and proteomics-related data, such as the IntAct protein interaction database and the PRIDE protein identification database. As co-chair of the HUPO Proteomics Standards Initiative (PSI) and executive member of the British Society for Proteome Research (BSPR), he actively contributes to the co-ordination of data standards in Proteomics. His e-mail address is <hhe@ebi.ac.uk>.