

The Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx)

Sandra Orchard^{1*}, Lukasz Salwinski², Samuel Kerrien¹, Luisa Montecchi-Palazzi¹, Matthias Oesterheld³, Volker Stümpflen³, Arnaud Ceol⁴, Andrew Chatr-aryamontri⁴, John Armstrong⁵, Peter Woollard⁵, John Salama⁶, Susan Moore^{6,20}, Jérôme Wojcik¹⁹, Gary D. Bader⁷, Marc Vidal⁸, Michael Cusick⁸, Mark Gerstein⁹, Anne-Claude Gavin¹⁰, Giulio Superti-Furga¹¹, Jack Greenblatt¹², Joel Bader¹³, Peter Uetz¹⁴, Mike Tyers¹⁵, Pierre Legrain¹⁶, Stan Fields¹⁷, The GO Consortium¹⁸, Michael Gilson²¹, Chris Hogue⁵, Hans-Werner Mewes³, Rolf Apweiler¹, Ioannis Xenarios¹⁹, David Eisenberg², Gianni Cesareni⁴, Henning Hermjakob¹

1. EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK
2. UCLA-DOE Institute for Genomics & Proteomics, UCLA, Los Angeles, USA
3. Institute for Bioinformatics, GSF - National Research Center for Environment and Health, Neuherberg, Germany
4. Dept. of Molecular Biology, University of Rome Tor Vergata, Rome, Italy
5. GlaxoSmithkline R&D, Stevenage, UK
6. Blueprint Initiative, Samuel Lunenfeld Research Institute, Ontario, Canada.
7. Memorial Sloan-Kettering Cancer Center, New York, USA
8. 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.
9. MB&B Department, Yale University, New Haven, CT, USA
10. EMBL Heidelberg, Germany
11. CeMM Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
12. Banting and Best Department of Medical Research, University of Toronto, Ontario, Canada
13. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
14. Institute of Toxicology and Genetics, FZK, Germany
15. Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada
16. Commissariat à l'Energie Atomique CEDEX, France.
17. Howard Hughes Medical Inst. Dept of Genome Sciences & Medicine, Univ. of Washington, WA, USA
18. [go@geneontology.org](http://go.geneontology.org)
19. Serono International S.A., Geneva, Switzerland
20. National University of Singapore, Clinical Research Centre, Singapore
21. University of Maryland Biotechnology Institute, Rockville, MD, USA

*** To whom correspondence should be addressed.**

Tel: +44 (0)1223 494 675

Fax: +44 (0)1223 494 468

Email: orchard@ebi.ac.uk

Abstract

A wealth of molecular interaction data is available in the literature, ranging from large-scale datasets to a single interaction confirmed by several different techniques. The data is all too often reported either by free text or a single table and is often missing key pieces of information, essential for a full understanding of the experiment. Here we propose MIMIx, the Minimum Information required for reporting a Molecular Interaction Experiment. Adherence to this reporting guideline will result in publications of increased clarity and usefulness to the scientific community and support the rapid, systematic capture of molecular interaction data in public databases, thus improving access to valuable interaction data.

Introduction Deciphering the molecular mechanisms of cell function relies, to a large extent, on tracing the multitude of interactions between the numerous components of living cells. A number of public databases strive to capture the ever-increasing amount of molecular interaction data described in the literature. During the process of manual curation, the raw data is extracted from a published paper or from a submitted manuscript and systematically transferred into the database.

Initially interaction databases worked in isolation and according to their own internal standards and data formats. Since no one database can achieve complete coverage of all known molecular interactions, the user may need to download and combine multiple datasets from two or more different databases to answer a specific question. Until recently, such users were forced to first parse the data into a common format and then try to remove redundant entries themselves. However, in 2004 a number of major databases jointly published a community standard data model for the representation and exchange of protein interaction data¹. This data model, jointly developed by members of the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO)², has already been adopted by major public domain interaction databases; datasets can be downloaded from many of these databases in PSI-MI XML format and further analysed using a number of PSI-MI compatible tools such as Cytoscape³, ProViz⁴ and PIMWalker⁵.

Building on the PSI MI standard, major public interaction databases have formed the International Molecular Interaction Exchange consortium (IMEx , <http://imex.sf.net>). These databases, currently BIND⁶, DIP⁷, IntAct⁸, MINT⁹, MPact (MIPS)¹⁰, 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. These databases strive to jointly achieve as near complete coverage as possible of the interaction data published in the literature, a task that is greatly hindered by inconsistencies and missing information in the published interaction reports. It would often be a minor task for the data producers to provide an additional key piece of information, the lack of which can lead to both misinterpretation of the information by scientists reading the

published article and a time-consuming and error-prone attempt to derive this by the curation team of a molecular interaction database. The lack of key information is not due to ill intention or even oversight, but simply due to the lack of a community consensus on the information required to appropriately describe a molecular interaction. Thus we present MIMIx as a discussion base, and as a compromise between the necessary depth of information to describe all relevant aspects of the interaction experiment, and the reporting burden placed on the scientists generating the data. The primary intention of a MIMIx-compliant dataset is not to reproduce an interaction experiment from a database record, but to allow database users to quickly assess and focus on data relevant to them, and then link to the source publications for the full experimental context. On the other end of the complexity scale, the PSI MI format, which is adopted by all IMEx partners, allows a much richer representation of an interaction than required in MIMIx, and IMEx partners welcome data submissions using the full complexity of the PSI MI format to provide a detailed account of a molecular interaction experiment.

Molecules The single major cause of time and data loss when extracting information from a paper, be it from peer reviewed literature or author-maintained websites and databases, is the use of ambiguous molecule identifiers, such as gene names, by authors in describing the molecules with which they are working. Based on anecdotal evidence from database curators, up to 70% of overall curation time can be spent on mapping molecule identifiers unambiguously to well-characterised database entries. For example, an author may fail to clearly indicate both the gene name and species from which the molecule in question originated. All this information would be implicit if the author were to use molecule identifiers generated by the major databases. An author may be working on “human p56^{lck} protein” which may be clearly identified both by the preceding verbal description or by quoting either the UniProtKB¹¹ accession number P06239 or RefSeq¹² NP_005347 indicating both gene and species. However, it is not uncommon for an author to merely refer to “lck cloned in a mammalian expression vector” or thank a collaborator for supplying an unspecified clone, with no indication as to whether the protein source is human, murine, bovine or rat.

The use of database accession numbers, in addition to gene or protein names, becomes even more important when common names are themselves indefinite – human PI3-kinase p85 subunit may appear to be a unique reference, but does it refer to the alpha subunit (P27986) or the beta subunit (O00459) which are two very distinct gene products? Such errors will almost certainly result in the paper in question not being added to a curated dataset and may also mislead the reader as to the actual construction of the experiment. Similarly, it is important to state if an interaction described in one organism has actually been modelled from an interaction detected between similar participants in a related organism, for example an interaction between a rat and a human protein been used to infer a human-human protein interaction. This means clearly describing the constructs used, including the organism of origin of the sequence. Important information on splice variants may also be lost if the author fails to give full details of the protein sequence in use.

Therefore we request that all molecules should be identified by the use of a *database accession number* from a public database.

For proteins, UniProt or RefSeq are strongly recommended, for genes Ensembl¹³ or Entrez Gene¹⁴, for chemical entities PubChem¹⁴ or ChEBI¹⁶. Nucleotide sequence database accession numbers (DDBJ/EMBL/GenBank, www.insdc.org) identify specific nucleic acid transcripts and give additional information as to the source and the class of nucleic acid under investigation. Where a molecule description is not available from these databases, identifiers from other public databases may be used, in particular model organism databases. For a full list of databases permitted, please refer to the relevant section of the PSI MI controlled vocabulary (see below), which also provides unified names for these resources.

An annotated protein or nucleic acid sequence may vary with time as the original submitters update their coding sequence prediction programs, frameshifts are identified and correction or resequencing is undertaken. This may invalidate the mapping of specific sequence positions, for example where deletion mutants or binding domains are described. Thus we request the addition of *version numbers*, either of the molecule, for example P06239.5, or of the database, to the MIMIx record.

While the identification of molecules by accession number is precise, it may be unwieldy to refer to “UniProt:P06239.5” instead of “lck” in the text of a manuscript. To satisfy both precision and readability, we recommend that accession number and *molecule name* used in the text be associated either in the submitted database record, or at least at their first occurrence, for example “...lck (UniProt:P06239.5) ...”.

A key element in the description of an interaction experiment is the role a molecule has in the interaction. MIMIx requests the classification of the molecule role in two aspects, the *biological role*, for example enzyme or enzyme target, and the *experimental role*, for example bait or prey. For both of these, the PSI MI standard (<http://psidev.sf.net/mi/rel25>) defines a comprehensive controlled vocabulary.

Experiment The MIMIx experiment description implements the core requirement of the “HUPO Minimum Information About a Proteomics Experiment (MIAPE)” guidelines (in this issue) and aims to capture the aspects of an interaction experiment which are necessary to classify and critically assess the results and their interpretation. This domain is likely to be further refined in the future, as other, technology-specific MIAPE modules evolve. The attributes we currently consider essential are as follows: The *host organism* describes the system in which the interactions have been detected. The host organism should be described by an NCBI tax id, and contain further specification like cell line or tissue descriptors. The *interaction detection method* accurately describes the method by which the interaction has been determined, for example tandem affinity purification (MI:0676).

The *participant detection method* names the experimental procedure for the detection of the molecules participating in the interaction, for example peptide mass fingerprinting (MI:0082).

Beyond these essential requirements we strongly recommend that authors provide additional detail on molecule sources, sample preparation, and further relevant experimental parameters using the detailed controlled vocabularies provided by the PSI MI standard.

Interaction The PSI MI standard provides a formal frame for a detailed description of an interaction, including both qualitative parameters and quantitative parameters, for example dissociation constants. However, this data is often not available, and thus MIMIx only requires two elements for the description of an interaction, the list of *molecules participating* in the interaction, characterised as above, and a *quality assessment*. In particular in large scale experiments, resulting interactions are usually assigned a quality score, which might be derived from data gained in the experiment itself, or from additional data outside the experiment. For the inclusion into public databases, it is essential that this reliability score is easily accessible. Ideally, not only the resulting score, but also the raw data that has been used to derive the score should be reported, so that users can perform alternative quality assessments.

Data Deposition Curators of the major molecular interaction databases work to collect and archive data from journal publications. While a systematic reporting of published interaction data according to the above guidelines would already enormously increase the efficiency of the curation task, literature curation after publication is only a second best option. As part of the MIMIx guidelines, we request that all reported interaction data be deposited in a publicly available molecular interaction database prior to publication. This has benefits for all parties involved. The databases will be able to work more efficiently, and will have a more direct access to the data producer to resolve unclear issues. The community will benefit from more, and more precise information in the databases, as database records can be checked directly by the domain expert, namely the data producer. Journals and data producers will benefit from consistently formatted database records, which can be included in the supplementary material of a publication. Accession numbers issued by a database and included in the journal will allow direct access to the data in the database, and allow a quick connection to related data in the database, for example other records on the same molecules. And finally, the data producers and journals will gain exposure for the publication through cross-references from the database records.

IMEx databases offer several options for data deposition

(<http://imex.sf.net/deposition.html>). The submission of fully formatted PSI MI XML files is recommended for large-scale data producers, who usually have the data available in in-house databases anyway. For smaller scale experiments, a pre-formatted Microsoft Excel spreadsheet file is available, with instructions on how to complete it. In addition to technical systems like the OLS ontology browser¹⁵ and a system for the automatic validation of PSI MI XML files (<http://www.ebi.ac.uk/intact/validator>), database curation teams provide assistance in all stages of the data deposition process, for example in

the correct use of the detailed controlled vocabularies used to characterise an interaction. We particularly encourage early contacts to database curation teams, to embed appropriate data collection already in the experiment planning stage.

In addition to the biological data, each data deposition must be accompanied by the minimal administrative data, namely *contact email*, *publication title*, *first author*, and the *publication identifier*, usually a Pubmed or Digital Object (<http://www.doi.org>) identifier. In the pre-publication stage, a journal-specific identifier can be used to provide a unique identification of the manuscript accompanying the data deposition.

To optimise the use of public resources, IMEx partners have developed common curation guidelines, and have agreed to synchronise their curation work and to exchange all user-submitted data, building up a network of stable, well co-ordinated molecular interaction databases freely accessible to the community. While accession numbers for deposited interactions will be issued within five working days of providing all necessary data, all deposited data will only be released on publication of the associated manuscript, or at the request of the data provider.

The MIMIx guidelines presented here, as well as the PSI MI format and the corresponding controlled vocabularies are not static, they will evolve based on community requirements in the context of a rapidly developing science. To get involved in the further development of MIMIx or the PSI MI standard, please refer to the mailing lists at <http://psidev.sf.net>.

Inset/Box:

The Minimum Information about a Molecular Interaction Experiment (MIMIx) checklist. Example data is taken from Croze et al¹⁷. and this example may be seen in full as a MIMIx-compatible submission at <http://imex.sf.net> in Excel, XML and HTML format. The full paper, annotated to the richer IMEx standard, can be viewed in the IntAct database (www.ebi.ac.uk/intact) using accession numbers EBI-958406, EBI-958452, EBI-958498 and EBI-959602.

- **Manuscript:**
A submission must contain the essential administrative information:
 - contact email:
 - publication title:
 - first author:.
 - publication identifier:
- **Experiment:**
Each experimental setup should be described separately, with the following parameters:
 - **Host system:**
The host organism in which the interaction took place.
Identificaton by NCBI tax id.
Example: Yeast (TaxID:4932)
Further specification of cell line or tissue is recommended.
 - **Interaction detection method:**
The method by which the interaction has been detected.
Root term MI:0001.
Example: two hybrid (MI:0018).
 - **Participant Identification method:**
The method by which the interaction participants have been determined.
Root term MI:0002.
Example: nucleotide sequence (MI:0078).
- **Interaction:**
 - **Participant list:**
The list of all molecules participating in the interaction. The list can contain one to many elements. Each molecule should be characterised by:
 - **Database.**
Root term: MI:0444.
Example: UniProt (MI:0486)
 - **Accession number from that database.**
Example: P48551
 - **Version number. Optional.**
 - **Name. The common name of the molecule, as used in the manuscript.**
Example: IFN- α R β L
 - **The species of origin for the molecule. Identified by NCBI tax id.**
Example: 9606.
 - **Biological role.**
The biological role of the molecule in the interaction.

- Root term: MI:0500.
 - Example: neutral component (MI:497)
 - Experimental role.
The experimental role of the molecule in the interaction.
Root term: MI:0495.
Example: bait (MI:0496)
- Confidence:
A confidence value attributed to the interaction.
The confidence attribution system must be described in the manuscript. Ideally, the raw data for the confidence assignment should be available.
Optional.

Controlled vocabularies are an essential part of the characterisation of a molecular interaction in PSI MI format. Elements of these controlled vocabularies are referred to as MI:xxxx above. The complete controlled vocabularies can be accessed at <http://psidev.sf.net/mi/rel25>, or interactively at <http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI>¹⁶.

References

1. Hermjakob, H. et al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol* 22, 177-83 (2004).
2. Orchard, S. et al. Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4-6, 2005. *Proteomics* 6, 738-41 (2006).
3. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-504 (2003).
4. Auber, D., Iragne, F., Mathieu, B., Nikolsky, M. & Sherman, D. in *ECCB 2003* (2003).
5. Meil, A., Durand, P. & Wojcik, J. PIMWalker: visualising protein interaction networks using the HUPO PSI molecular interaction format. *Bioinformatics* 4, 137-139 (2005).
6. Bader, G. D., Betel, D. & Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 31, 248-50 (2003).
7. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32 Database issue, D449-51 (2004).
8. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res* 32 Database issue, D452-5 (2004).
9. Zanzoni, A. et al. MINT: a Molecular INTeraction database. *FEBS Lett* 513, 135-40 (2002).
10. Pagel, P. & al., e. The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 832-834 (2005).
11. Bairoch, A. et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33 Database Issue, D154-9 (2005).
12. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* 31, 34-7 (2003).
13. Birney, E. et al. An overview of Ensembl. *Genome Res* 14, 925-8 (2004).
14. Wheeler, L. & al., e. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 34, 173-180 (2006).
15. Cote, R. G., Jones, P., Apweiler, R. & Hermjakob, H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 7, 97 (2006).
16. de Matos, P., Ennis, M., Darsow, M., Guedj, M., Degtyarenko, K., Apweiler, R. (2006) ChEBI - Chemical Entities of Biological Interest, *Nucleic Acids Res.*, Database Summary Paper 646.
<http://www3.oup.co.uk/nar/database/summary/646>
17. Croze E, Usacheva A, Asarnow D, Minshall RD, Perez HD, Colamonici O (2000) Receptor for activated C-kinase (RACK-1), a WD motif-containing protein, specifically associates with the human type I IFN receptor. *J Immunol.*165, 5127-5132