Title: Protein interaction data curation: the International Molecular Exchange (IMEx) consortium.

Authors: Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H

Journal: Nature methods

Year: 2012 Apr

Volume: 9

Issue: 4

Pages: 345-50

DOI: 10.1038/nmeth.1931

Published in final edited form as:
Nat Methods. 2012 April ; 9(4): 345–350. doi:10.1038/nmeth.1931.

Protein Interaction Data Curation - The International Molecular Exchange Consortium (IMEx)

Sandra Orchard[1], **Samuel Kerrien**[1], **Sara Abbani**[2], **Bruno Aranda**[1], **Jignesh Bhate**[3], **Shelby Bidwell**[4], **Alan Bridge**[5], **Leonardo Briganti**[6], **Fiona S. L. Brinkman**[7], **Gianni Cesareni**[6,8], **Andrew Chatr-aryamontri**[6,9], **Emilie Chautard**[10,11], **Carol Chen**[12], **Marine Dumousseau**[1], **Johannes Goll**[4], **Robert E. W. Hancock**[12], **Linda I. Hannick**[4], **Igor Jurisica**[13], **Jyoti Khadake**[1], **David J. Lynn**[14], **Usha Mahadevan**[3], **Livia Perfetto**[6], **Arathi Raghunath**[3], **Sylvie Ricard-Blum**[11], **Bernd Roechert**[5], **Lukasz Salwinski**[2], **Volker Stümpflen**[15], **Mike Tyers**[9,16], **Peter Uetz**[17,18], **Ioannis Xenarios**[5,19,20], and **Henning Hermjakob**[1]

[1]European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK [2]University of California, Los Angeles Department of Energy Institute for Genomics and Proteomics, Los Angeles, California, USA [3]Molecular Connections Pvt. Ltd., Kandala Mansions, 2/2 Kariappa Road, Basavangudi, Bangalore - 560004, India [4]The J. Craig Venter Institute, Rockville, MD 20850, USA [5]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, CMU 1, Rue Michel Servet, 1211 Geneva 4 [6]Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, Rome Italy [7]Department of Molecular Biology and Biochemistry, 8888 University Drive, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6 [8]Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS)-Fondazione S. Lucia, Rome, Italy [9]Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh, Scotland [10]Institut de Biologie et Chimie des Protéines, UMR 5086 CNRS – Université Lyon 1, IFR 128 Biosciences Gerland Lyon Sud, 7 passage du Vercors, 69367 Lyon Cedex 07, France [11]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 800, Toronto, Ontario, Canada, M5G 0A3 [12]Centre for Microbial Diseases and Immunity Research, 232 - 2259 Lower Mall, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4 [13]Ontario Cancer Institute, the Campbell Family Institute for Cancer Research, and Techna Institute, University Health Network, 101 College Street, Toronto, Ontario M5G 1L7, Canada [14]Animal & Bioscience Research Department, AGRIC, Teagasc, Grange, Dunsany, Co. Meath, Ireland [15]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Bioinformatics and Systems Biology (MIPS), Neuherberg, Germany [16]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada [17]Proteros Biostructures, Martinsried, Germany [18]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA, USA [19]Vital-IT group, SIB Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland [20]University of Lausanne, Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland

Abstract

The IMEx consortium is an international collaboration between major public interaction data providers to share curation effort and make a non-redundant set of protein interactions available in a single search interface on a common website (www.imexconsortium.org). Common curation rules have been developed and a central registry is used to manage the selection of articles to enter

Correspondence to Sandra Orchard, (orchard@ebi.ac.uk) EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK Fax: +44-(0)1223-494-468.

into the dataset. The advantages of such a service to the user, quality control measures adopted and data distribution practices are discussed.

Introduction

Protein-protein interactions are a key element in our understanding of molecular biology. However, in contrast to areas of activity such as DNA sequencing or protein structural analysis, the systematic capture of published molecular interaction data into public domain repositories is still in its infancy. This is not due to lack of resources in this domain. As of December 2011, the PathGuide resource [1] listed more than 100 protein-protein interaction (PPI) –related databases. Although many of these focus on predictions of potential interactions or interologue-mapping, rather than experimentally determined interactions, the level of activity suggests ample resources. However, most of these resources are independently funded and pursue their goals in isolation. As a result, accessing all publicly available molecular interaction data, even on a specific biological or biomedical topic, is a challenging, time-consuming task requiring the user to query multiple resources, each with a different interface, many using different identifiers and often containing redundant data from overlapping sets of publications.

Efforts to address this problem began ten years ago with the development of a common file format for representing protein interaction data. The Minimum Information about a Molecular Interaction eXperiment (MIMIX) guidelines were then published [2] defining a checklist of the information to be supplied when describing experimental molecular interaction data in a journal article. In parallel to this, the curation strategies of a select group of molecular interaction databases, the IMEx Consortium, were coordinated to create a single non-redundant set of homogeneously curated protein interaction data, as discussed in this article.

A Common Data Format and the IMEx Consortium

The issue of the individual data resource formats maintained by the separate resources has largely been addressed by the efforts of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) [3]. In 2002, a number of protein interaction data providers, among them BIND [4], DIP [5], Hybrigenics [6], IntAct [7], MINT [8], and MIPS [9], set out to develop a common file format for the representation of protein interaction data. This resulted in the creation of the HUPO PSI-MI XML format [10], which is now widely implemented, and has since been expanded to enable the interchange of all forms of molecular interaction data [11]. This enables the user to download, combine, visualize and analyze data in a single format from multiple resources. It has since been supplemented by a simplified tabular format, MITAB [11].

While a common data format is a key step in providing consistent, user-friendly access to publicly available molecular interaction data, it is only a first stage. Until recently, all interaction databases independently curated interaction data publications, on occasion resulting in several alternative datasets derived from a single publication, due to the implementation of different curation strategies. In addition to the use of scarce public funding for the duplication of expensive manual database curation, the differences in the datasets can leave the user bewildered about which to regard as the correct interpretation of the data within the paper. To address this issue, five molecular interaction databases agreed in September 2005 on a long-term co-ordination of their curation strategies. The framework for this collaboration was the International Molecular Exchange Consortium (IMEx) which currently comprises DIP [5], IntAct [7], MatrixDB [12], MINT [8], MPIDB [13], I2D [14], InnateDB [15] and Molecular Connections (www.molecularconnections.com) as full

members, with BioGRID [16] as an observer member. A full IMEx member commits to producing a relevant number of records curated to a common IMEx standard whereas an observer member is a prospective IMEx consortium member working with the full members to produce the curation rules and improve curation quality. The aims of IMEx are to coordinate curation to avoid redundant work on the same data, increase curation coverage and synchronize curation strategies to ensure consistency of data across all IMEx member databases. Since 2005 an increasing number of these databases have been working together to generate a single set of curation rules to ensure both the quality and consistency of annotation across the IMEx databases. As a result of many detailed IMEx consortium discussions, a single joint IMEx curation manual (www.imexconsortium.org/curation) has been agreed upon and made publicly available. This forms the basis for the curation by all IMEx partner databases and at all levels uses the controlled vocabularies developed by the HUPO-PSI [10,11].

Curation strategy and coverage

Currently interaction databases contain a considerable amount of redundant data, i.e. the same paper curated by multiple resources, often to differing depths of curation or following different annotation strategies. As stated above, one of the major aims of the IMEx Consortium is to present the user with a non-redundant set of data to search, namely each paper should be present only once in the IMEx set, with the protein-protein interaction information it contains having been fully captured following a consistent set of rules.

Initially, the IMEx databases agreed to share the curation workload based on journal selection. Each IMEx database selected one or more journals to curate, with the aim of representing all relevant protein-protein interaction data published in that journal in the database within a reasonably short time frame, normally less than three months from publication. The journal(s) were selected by the databases and largely reflect their particular areas of interest or editorial connections. Table 1 shows the current IMEx journal coverage. There is no pre-selection of data from particular organisms, although, in practice, the well-studied model organisms such as human, mouse, *Arabidopsis, Saccharomyces cerevisiae* and *E. coli* also tend to be the best represented in the scientific literature available for curation.

Whilst articles from targeted journals form the baseline of IMEx curation, most databases curate additional publications, the choice usually based on scientific collaborations, curator expertise, or to reflect the specialization of thematic databases such as MatrixDB and MPIDB. As an example, IntAct recently curated a targeted dataset on interactions of proteins that play a role in Alzheimer's Disease [17]. Until recently, these targeted curation efforts were not coordinated between the IMEx members. However, in 2010 we released IMExCentral, a web service which enables IMEx partners to reserve any publication for curation, either manually through a web interface, or through a web service directly from our curation tools. Based on this tool, we are now also coordinating curation of all individual publications outside of the journal curation commitment.

IMEx members are currently working on releasing a non-redundant set of all papers curated to IMEx standards by the participating databases since 2006. Key large-scale papers, such as the Giot protein interaction map of Drosophila melanogaster [18], and the Rual [19] and Stelzl [20] human protein-protein interaction networks have been recurated to the existing IMEx standard and released to the dataset. More recent large-scale papers are routinely added to the dataset and users are encouraged to propose additional publications for curation.

Several of the participating databases contain data curated to different depths (see below) or which were curated during the period whilst the IMEx rules were under development. As can be seen in Table 2, the majority of the participating databases still have a wealth of curated papers which have yet to be released to the IMEx dataset. A major aim of 2012 will be to identify archival papers appropriate for release through the IMEx website and, if necessary, re-curate these to current IMEx standards. Papers curated by MINT and IntAct as training and test datasets for the BioCreative competitions [21,22] have already been released as part of this process. Where a paper has previously been redundantly curated, i.e. it has been annotated by more than one IMEx database, IMExCentral will only allow one copy of the paper to acquire an IMEx accession number and will alert the databases if a second resource attempts to register the same publication.

IMExCentral already allows participating databases to encourage and manage the annotation of directly submitted data as an integral part of the publication process. Authors may submit data to any IMEx database. A common identifier allocated by IMExCentral (IM-xxxx), will allow data users to access the dataset, subsequent to publication, both in the original resource and via the IMEx website. Should identical data be offered to more than one member database, this will immediately be highlighted by the IMExCentral service when a database attempts to register a second copy of the same dataset.

In addition to deposition of new experimental data, IMEx users can also request curation of specific publications via the IMEx web site (http://www.imexconsortium.org), for example if they notice a well-known interaction missing from the IMEx databases or to establish the currently know interactions for a particular research target.

Curation depth

The IMEx partners have committed to a "deep" curation model, which aims to capture the full experimental detail provided in the interaction report, as this is often essential to assess interaction context and confidence. In fact, it has become increasingly clear that minor changes in experimental detail may have dramatic effects on the outcome of an interaction experiment [23]. Fig. 1a shows a breakdown of the major interaction detection methods used to identify protein interactions represented within the IMEx dataset, as defined in the PSI-MI controlled vocabulary (www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI). IMEx members refer to all interactions between two molecules as binary interactions and these are classified by the type of binding described (Fig.1b). ‘Association’ indicates that the interaction is from an experimental method which identifies a loose co-complex, within which all the members may not have been identified, typically by co-immunoprecipitation or pulldown from an *in vivo* sample. ‘Physical association’ indicates the interaction has been identified by a method indicating a tighter complex, but again in which all the members may not have been identified, for example protein complementation assays such as yeast 2-hybrid. ‘Direct interaction’ indicates that the two molecules are known to be in actual physical contact with each other. Such data is only taken from *in vitro* methodologies and does not include yeast 2-hybrid assays, but we acknowledge that when performed properly yeast 2-hybrid assays are strong evidence of a direct interaction. Experimental molecular features such as affinity tags, labels, and functional protein modifications including post-processing of the transcript or phosphorylation sites are mapped to the given sequence, as are binding domains and interacting residues. Author-provided confidence scores are also documented, where this data is available. Where essential data as required by the MIMIx guidelines [2] are not available, we aim to obtain the data from the corresponding author, or mark the manuscript as containing data that cannot be annotated. The IMEx Consortium collates experimental evidences from any species for which interaction data is available (Fig. 1c).

Quality control

Curation rules are only useful if they are consistently applied, and the IMEx consortium is gradually implementing measures for mutual quality control. The PSI validator [24], a tool which executes a set of rules based on the PSI-MI ontology to check XML files, provides not only syntactic checking of released files, but also semantic checking, validating the use of the correct controlled vocabularies as well as more complex, context-dependent rules. A set of validator rules ensuring compliance with the IMEx curation manual have been developed and are now publicly available for use by consortium members, data submitters and, indeed, any user of the PSI-MI XML format.

Cross-curation exercises have been undertaken, and will remain an ongoing regular exercise, with a number of ‘challenging’ papers being selected for annotation by all participating databases. The resulting download files are compared and issues discussed to ensure curation rules and controlled vocabularies are used consistently across databases. Alternatively, rules/vocabularies may be modified to address challenging issues.

In addition, each month, one database selects a paper for discussion by the collaboration, initially via a Wiki page, but if problems cannot be resolved, then they can be discussed in person by phone conference or face-to-face meetings. In this way, rules can be generated to address new technologies or variations on accepted methodologies. Finally, approximately 20 papers highlighted by the iRefIndex database [25] as being curated by more than one IMEx database have been compared. The redundant curation predated the formation of IMEx and the exercise confirmed that the current IMEx curation rules and internal quality control measures would have addressed the vast majority of problems identified.

Data dissemination

Many collaborative curation projects, for example UniProt, Gene Ontology annotation, or wwPDB, exchange data on a regular basis, with the data from each partner being copied to all other partners. However, the regular full copying of complex records from multiple partners, and in particular the management of the updates and deletions of both interaction records and the underlying sequences to which they are mapped, is highly resource-consuming in terms of both computational load and staff. While the IMEx partners have been increasingly collaborating since 2005, we only recently entered IMEx ‘production mode’ with the regular release of IMEx records and requiring sharing of curated interaction data between partners.

Recently, a standard interface for direct computational access to standards-compliant molecular interaction data resources, the PSI Common Query Interface (PSICQUIC) was developed [26]. PSICQUIC supports simultaneous querying of multiple participating molecular interaction databases.

The IMEx partners decided to use the distributed PSICQUIC system as the basis for IMEx data dissemination to minimize the data exchange overhead [26]. Delivery of the IMEx set of interaction records to the IMEx partners and individual member database websites is done through a tagging process. Only IMEx partners may use the IMEx tag and only records presented in a registered PSICQUIC service tagged as an IMEx record and with an IMEx accession number will be part of this IMEx dataset. Each IMEx partner operates a PSICQUIC server, and a PSICQUIC client can query all partners for IMEx data matching a given query, providing an up-to-date view of all relevant data from all IMEx partners.

Because the PSICQUIC service can query data from non-IMEx as well as IMEx members, uses can access either the tagged IMEx subset or all of the available interaction data.

However, when searching across the entirety of data available through PSICQUIC, it is currently difficult to separate out experimentally proven binary pairs from predicted interactions, functional associations or the results from text-mining. This data is also highly redundant, in that the manually curated data in primary databases is re-exported by several integrative databases such as iRefIndex [25], APID [27] and STRING [28]. Unfortunately, much of the experimental detail may be lost during the integration process although a link back to the primary database record is usually provided. For example, at the time of going to press, PMID:17923092 appeared in 5 resources when searched for in PSICQUIC, and in many of these resources, it is not clear that the majority of data in this paper derives from genetic interference assays (MI:0254) as the data in integrative databases can lack the detailed information required to make this clear. In the IMEx set, each interaction publication appears once only, with experimental detail and the protein constructs clearly defined. Users are encouraged to access and search the IMEx dataset via PSICQUIC, either directly on the website (www.imexconsortium.org) or via member database websites.

In addition to the interactive PSICQUIC access, all IMEx data is also available for full download in PSI-MI XML or MITAB tabular formats. All IMEx data from all partners is freely available without any restrictions.

In the future, we expect a significant increase in the coverage of IMEx records, in particular through ongoing curation, and the acquisition of new IMEx partners, but also through the "upgrade" of existing archival records to IMEx records as discussed above. In particular, we will validate and where necessary re-curate widely used large scale interaction data sets as IMEx records. Most importantly, however, we aim to shift the focus of IMEx curation from the curation of manuscripts after publication, to pre-publication, in collaboration with all relevant stakeholders. The curation of data prior to publication, in a direct dialog with the authors, ensures a data representation that is both factually correct, and optimally aligned with the authors' view of the data. Through inclusion of IMEx accession numbers in the publication, and data release synchronized with the publication of the manuscript, both parties benefit from increased visibility, and users benefit from timely access to this comprehensive, annotated and accurate protein interaction data.

Why is the IMEx Consortium necessary?

As previously stated, a significant number of interaction databases exist, which attempt to capture PPI data from the literature using different curation strategies. In addition to this, there are now a number of ‘composite’ databases, which contribute no novel manual curation but instead merge the work of other resources. Other databases take a median strategy and import selected data from curated resources whilst adding to this by their own annotation efforts. There are also datasets of predicted protein interactions, using a variety of information sources. Attempts to merge data are often frustrated by the differing strategies adopted by the databases, in particular when mapping ambiguous protein descriptions given in the text to identifiers in sequence databases. Even when both gene name and species are stated, which is often not the case [2], authors rarely clearly define which isoform of the protein they are dealing with, even when this information is known. Databases deal with this ambiguity in several ways, either by mapping to a gene identifier and sacrificing all ability to map to a specific isoform (BioGrid) or by selecting one transcript, usually the longest (BIND), which makes it impossible to indicate when this is an ambiguous or a specific mapping, or by utilizing the canonical sequence displayed by UniProtKB (IntAct, MINT, DIP, MatrixDB, I2D, MPIDB).

A further cause of apparent differences between databases is caused by their varying policies to describe interactions demonstrated between protein constructs from different species e.g.

human-mouse. Most databases report the data as performed by in the experiment, others choose to model this onto a single organism, for example human (HPRD) [29]. Additionally, databases may only partially curate a publication, extracting only content which relates to their specific area of interest (InnateDB, HPRD, MPIDB). Whilst none of these policies are in any way wrong, they do create difficulties when attempting to reconcile redundancies between databases. A recent report suggested agreement between databases may be only 54% for curated interactions and 71% in protein identifications and attributed much of this to the difficulties described above [30]. The effect of curation errors cannot be ignored but a recent re-curation exercise showed this to range from 2–9% for a number of different databases [31].

The authors firmly believe that the policy followed by the IMEx Consortium of taking a coordinated, collaborative rather than competing approach to the intergation of protein-protein interaction data provides the best possible service to the user community. We not only achieve a much broader coverage of the interaction literature published in each calendar year than a single database working in isolation can achieve, but we also provide the research community with a single point of access to the data, removing the need to combine records from different databases. The quality control measures, both internal and cross-database, being developed by the consortium minimize curation error, and by supplying data consistently mapped to external reference resources, eliminate errors potentially introduced when identifiers are remapped by third-party resources.

The IMEx records are mapped to the UniProtKB canonical sequence [32] when the isoform is ambiguous, and to the specific isoform identifier when known, with the corresponding entity in RefSeq [33], mapped at the sequence level, also referenced. To facilitate coordination among resources, we use a scientific publication as the basic unit of IMEx curation. If a publication is curated within IMEx, it is curated in full, harvesting all reported protein-protein interactions into the database, rather than, for example, only those relevant for a specific disease. This enables full data traceability and where possible, we provide even more fine-grained data source information, by annotating figure or supplement numbers from which the data have been extracted. The quality control measures currently being implemented will also act to bring down the curation error rates cited above and improve data quality. Turinsky *et al.* concluded that “Many of the discrepancies we identified should in the future be eliminated if the IMEx guidelines are widely followed.” [30]

Outlook

We believe that by establishing a network of closely collaborating interaction data resources with a common data representation, query interface, and shared curation rules, we are creating a novel, reliable and highly visible infrastructure for protein interaction data collection that will motivate data producers, funding agencies and journals to increasingly make interaction data deposition an integral part of the publication process. Enforcing quality control checks across the partner databases will improve data quality, and clear statements of our curation policies will make these transparent to users, and ensure consistency across the entire IMEx dataset. Regular meetings enable the review of these curation rules and will allow us to rapidly respond to new data types such as quantitative data and dynamic interactions. The IMEx consortium is open to the participation of new partners, and all data producers are encouraged to submit their data to one of the IMEx partners prior to publication. For detailed information on IMEx membership and data deposition, please see http://www.imexconsortium.org.

Acknowledgments

Funding

This work has been supported by the European Commission under PSIMEx, contract number FP7-HEALTH-2007-223411, and in part by AntiPathoGN, contract number HEALTH-F3-2009-223101, SLING, contract number 226073, APO-SYS, contract number FP7-HEALTH-2007-200767), the Ontario Research Fund (GL2-01-030), Canada Foundation for Innovation (CFI #12301 and CFI #203383) and National Institutes of Health grant GM071909.

Competing Financial Interests

Authors Jignesh Bhate, Usha Mahadevan and Arathi Raghunath are employees of Molecular Connections, a contract annotation company, however all work reported here is freely available without restriction.

References

1. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic acids research. 2006; 34:d504–d506. [PubMed: 16381921]
2. Orchard, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nature Biotechnology. 2007; 25:894–898. This reference outlines the information an author needs to include in a manuscript to allow successful curation by a database
3. Orchard S, Hermjakob H. The HUPO proteomics standards initiative - easing communication and minimizing data loss in a changing world. Briefings in bioinformatics. 2008; 9:166–173. [PubMed: 18065433]
4. Alfarano C, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic acids research. 2006; 33:d418–d424. [PubMed: 15608229]
5. Xenarios I, et al. Nucleic acids research. 2002; 30:303–305. [PubMed: 11752321]
6. Rain JC, et al. The protein-protein interaction map of Helicobacter pylori. Nature. 2001; 409:211–215. [PubMed: 11196647]
7. Kerrien S, et al. The IntAct molecular interaction database in 2012. Nucleic acids research. 2012; 38:d525–d531. [PubMed: 19850723]
8. Ceol A, et al. MINT, the molecular interaction database: 2009 update. Nucleic acids research. 2009; 40
9. Guldener U, et al. MPact: the MIPS protein interaction resource on yeast. Nucleic acids research. 2006; 34:d436–d441. [PubMed: 16381906]
10. Hermjakob H, et al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nature biotechnology. 2004; 22:177–183.
11. Kerrien S, et al. Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. BMC biology. 2007; 5:44. [PubMed: 17925023] Describes the current version of the formats and controlled vocabularies used by the IMEx consortium.
12. Chautard E, Fatoux-Ardore M, Ballut L, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database. Nucleic acids research. 2011; 39:d235–d240. [PubMed: 20852260]
13. Goll J. MPIDB: the microbial protein interaction database. Bioinformatics (Oxford, England). 2008; 24:1743–1744.
14. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 2007; 8:R95. 2007. [PubMed: 17535438]
15. Lynn DJ. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular systems biology. 2008; 4:218. [PubMed: 18766178]
16. Breitkreutz BJ, et al. The BioGRID Interaction Database: 2008 update. Nucleic acids research. 2008; 36:d637–d640. [PubMed: 18000002]
17. Perreau VM, et al. A domain level interaction network of amyloid precursor protein and Abeta of Alzheimer's disease. Proteomics. 2010; 10:2377–2395. [PubMed: 20391539]

18. Giot L, et al. A protein interaction map of Drosophila melanogaster. Science. 2003; 302:1727–1736. [PubMed: 14605208]

19. Rual JF, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437:1173–1178. [PubMed: 16189514]

20. Stelz, lU, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005; 122:957–968. [PubMed: 16169070]

21. Chatr-aryamontri A, et al. MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. Genome Biol. 2008; Suppl 2(9):s5. [PubMed: 18834496]

22. Leitner F, et al. The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. Nat Biotechnol. 2010; 28:897–899. [PubMed: 20829821]

23. Chen YC, Rajagopala SV, Stellberger T, Uetz P. Exhaustive benchmarking of the yeast two-hybrid system. Nat Methods. 2010; 7:667–668. [PubMed: 20805792]

24. Montecchi-Palazzi L, et al. The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. Proteomics. 2009; 9:5112–5119. [PubMed: 19834897]

25. Turner B, et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford). 2010 baq023.

26. Aranda B, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods. 2011; 8:528–529. [PubMed: 21716279] Describes the method by which a non-redundant dataset can be made available to users by the IMEx consortium.

27. Prieto C, De Las Rivas J. APID: Agile Protein Interaction DataAnalyzer. Nucl. Acids Res. 2006; 34:W298–W302. [PubMed: 16845013]

28. Szklarczyk D, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011; 39:d561–d568. [PubMed: 21045058]

29. Keshava Prasad TS, et al. Human Protein Reference Database-2009 update. Nucleic Acids Res. 2009; 37:d767–d772. [PubMed: 18988627]

30. Turinsky AL, et al. Literature curation of protein interactions: measuring agreement across major public databases. Database (Oxford). 2010 baq026.

31. Salwinski L, et al. Recurated protein interaction datasets. Nat Methods. 2009; 6:860–861. [PubMed: 19935838]

32. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Research. 2011; 39:d214–d219. [PubMed: 21051339]

33. Sayers EW, et al. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2010; 38:d5–d16. [PubMed: 19910364] *Key papers

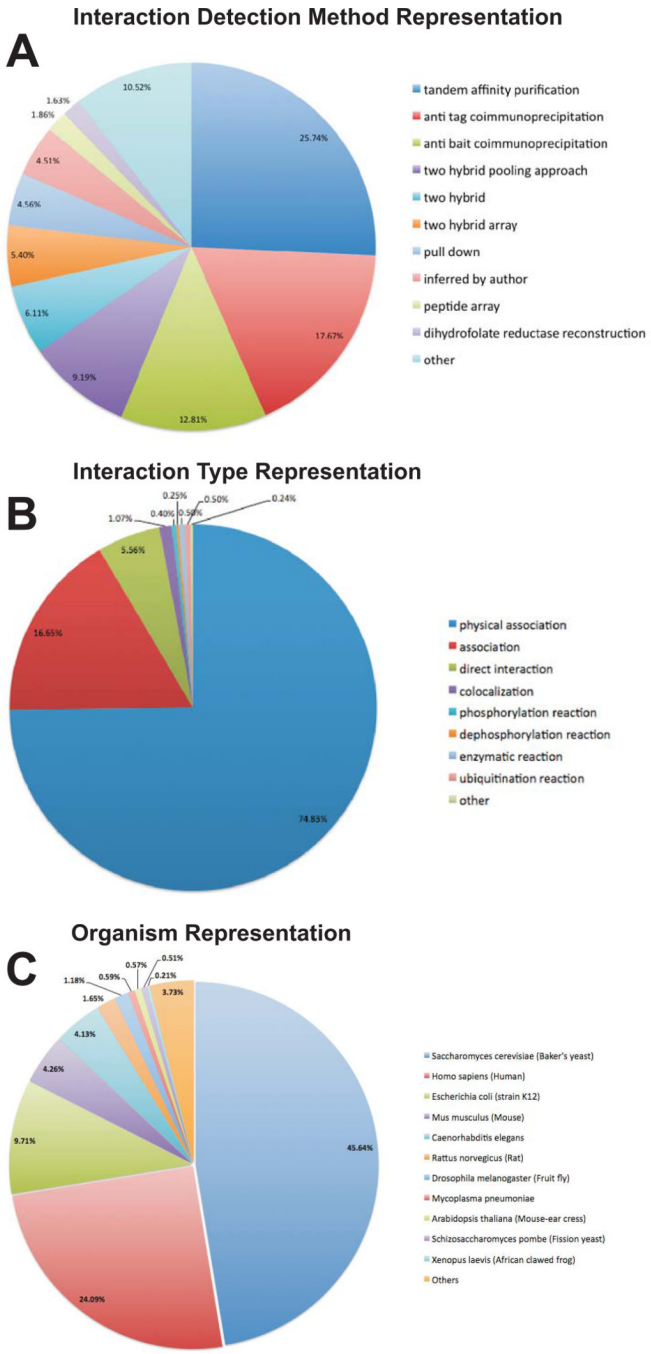

Figure 1. Overview of the IMEx dataset

(**a**) Interaction detection methods currently represented in the IMEx dataset.
(**b**) Types of interaction data represented in the IMEx dataset.
(**c**) The range of species for which data is available in the IMEx dataset.
Data taken from December 2011

Table 1

Current journal coverage by IMEx members

Journal Name	Period of Coverage	Database
Cancer Cell	01/2006– present	IntAct
Cell	01/2006– present	IntAct
FEBS Letters	01/2005– present	MINT
EMBO Journal	01/2006– present	MINT
EMBO Reports	01/2006– present	MINT
J. Bacteriology	08/2007–present	MPIDB
J. Molecular Signalling	11/2006–present	Molecular Connections
Matrix Biology	01/2009–present	MatrixDB
Molecular Cancer	09/2010–present	Molecular Connections
Molecular Microbiology	08/2007–08/2009	MPIDB
Nature Immunology	10/2010 – present	InnateDB
Nature Structural and Molecular Biology	01/2006–present	DIP
Oncogene	09/2010	I2D
PLoS Biology	01/2003–present	DIP
Proteomics	01/2005–present	IntAct
Structure	01/2006–present	DIP

Table 2

Total number of publications curated each year by each database in each calendar year and the number of those released to date to the IMEx dataset. Data available until end of 2011.

	MINT		IntAct		DIP		MPIDB		MatrixDB		Molecular Connections		I2D		InnateDB	
	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx	Curated papers	Exported to IMEx
2001	278	6	0	0	0*	0										
2002	185	0	0	0	0*	0										
2003	131	1	110	0	0	0										
2004	538	29	348	0	3,005*	0										
2005	439	120	519	1	0*	0										
2006	557	401	1,294	236	0*	0										
2007	268	259	715	87	899	0	813	0								
2008	466	251	756	138	771	771	0	0							1,038**	
2009	574	211	478	123	621	3	183	183							808**	
2010	957	152	348	130	542	542	0	0	36	36	18	6	42	27	2,596**	
2011	614	284	447	160	615	615	5	5	41	25	17	17	58	58	676**	27

*records curated prior to the formation of IMEx Consortium, exact release date cannot be tracked

**publications not exported to IMEx are curated to MIMIx standards