D452–D455 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh052

IntAct: an open source molecular interaction database

Henning Hermjakob*, Luisa Montecchi-Palazzi[1], Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron[2], Bernd Roechert[3], Peter Roepstorff[4], Alfonso Valencia[5], Hanah Margalit[6], John Armstrong[7], Amos Bairoch[3], Gianni Cesareni[1], David Sherman[8] and Rolf Apweiler

European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK, [1]Universita Tor Vergata, Rome, Italy, [2]Max Planck Institute for Molecular Genetics, Berlin, Germany, [3]Swiss Institute of Bioinformatics, Geneva, Switzerland, [4]University of Southern Denmark, Odense, Denmark, [5]National Center of Biotechnology, Madrid, Spain, [6]The Hebrew University of Jerusalem, Israel, [7]Glaxo Research and Development Limited, Stevenage, UK and [8]University of Bordeaux, France

Received August 15, 2003; Revised and Accepted September 23, 2003

ABSTRACT

IntAct provides an open source database and toolkit for the storage, presentation and analysis of protein interactions. The web interface provides both textual and graphical representations of protein interactions, and allows exploring interaction networks in the context of the GO annotations of the interacting proteins. A web service allows direct computational access to retrieve interaction networks in XML format. IntAct currently contains ~2200 binary and complex interactions imported from the literature and curated in collaboration with the Swiss-Prot team, making intensive use of controlled vocabularies to ensure data consistency. All IntAct software, data and controlled vocabularies are available at http://www.ebi.ac.uk/intact.

INTRODUCTION

Protein interactions provide a valuable resource for the elucidation of cellular function, and protein interaction studies have been a focus of recent biomolecular research. Experimental technologies like two-hybrid systems (1) or tandem affinity purification (2) allow the generation of large quantities of interaction data. However, practically all medium- to large-scale projects develop their own systems for the storage, representation and analysis of protein interaction data. In addition to the duplication of work, this results in a high degree of incompatibility between different protein interaction data sets. IntAct provides a comprehensive, open source database and analysis system for protein interactions which can be locally installed and adapted to the needs of the local organization, thereby reducing development time, and promoting consistency of interaction data sets through the use of the same infrastructure and annotation system.

DATABASE STRUCTURE AND ACCESS

Data model

The IntAct data model has three main components: Experiment, Interaction and Interactor. An Experiment groups a number of Interactions, usually from one publication, and classifies the experimental conditions in which these Interactions have been generated. An Experiment may have only a single interaction, or hundreds of interactions in the case of large-scale experiments. An Interactor is a biological entity participating in an Interaction, usually a protein, but potentially also a DNA sequence, or a small molecule. An Interaction contains one or more Interactors participating in the Interaction. The representation of interactions is not limited to binary interactions; data on multi-protein interactions, e.g. the results of tandem affinity purification experiments (3), can be represented as one interaction, without artificially splitting them up into several binary interactions.

Controlled vocabularies

A curated protein interaction database integrates data from many sources, and a major database maintenance task is to ensure data consistency. Data attributes, e.g. experimental methods, must be annotated in a consistent manner to ensure the data remain correct and searchable. Naming variations like 'yeast two hybrid' and 'Y2H' are easy to understand for a scientist, but they are very difficult to search for in a consistent manner. In addition, it is often desirable to group data according to specific annotations into broader categories, for example to search for all interactions that have been derived by different types of protein complementation assay, thus covering both classical two-hybrid assays and related technologies. To support these requirements, IntAct intensively uses controlled vocabularies instead of free text attributes. Where possible, existing reference systems like the NCBI taxonomy database (4) or Gene Ontology (GO) (5) are used. For a number of attributes specific to protein interactions, new controlled vocabularies have been developed in the IntAct

*To whom correspondence should be addressed. Tel: +44 1223 494671; Fax: +44 1223 494468; Email: hhe@ebi.ac.uk

Table 1. IntAct controlled vocabularies

Vocabulary domain	Description	Number of terms
Interaction detection method	The experimental method used to determine an interaction. Example: yeast two hybrid	104
Participant detection method	The experimental method used to determine the interactors (proteins) participating in an interaction. Example: peptide mass fingerprinting	14
Interaction type	Type of interaction. Example: aggregation	30
Sequence feature type	Type of a sequence feature. Example: binding site	73
Sequence feature detection method	Experimental method to determine a sequence feature. Example: alanine scanning	20

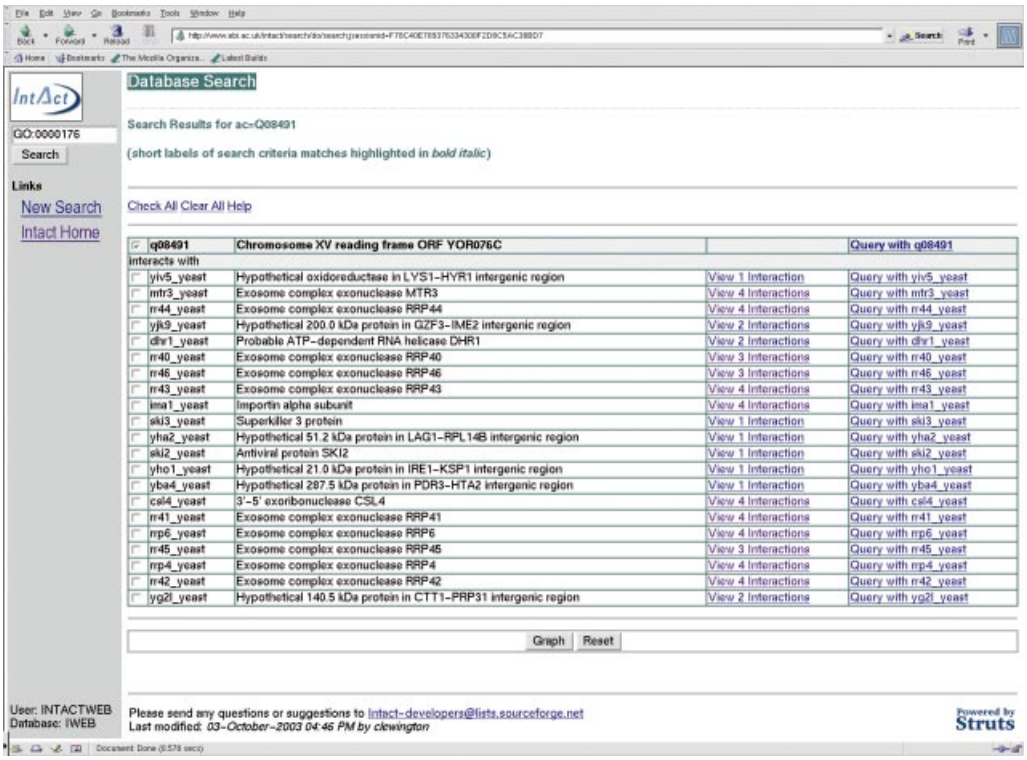

Figure 1. IntAct search interface, binary view. For each query protein, this view shows all its potential interaction partners, together with a short description. For each interaction partner, a link allows to access the experiment view, displaying all the experiments that provide evidence that these two proteins interact.

project, together with extensive definitions and cross-references (Table 1). To support generalized queries, these vocabularies have a hierarchical structure, following the GO format, where terms that are higher in the hierarchical structure represent more general terms. All controlled vocabularies are publicly available in GO format.

Database access

IntAct currently provides a simple search interface that searches the database by name, by IntAct accession number, or by identifiers of external databases, for example GO, Swiss-Prot (6) and SGD (7).

The data retrieved can be displayed in two views, a binary view (Fig. 1) and an experiment view (Fig. 2). For a given protein, the binary view displays all its known interaction partners and their minimal textual description. This may quickly give indications of potential functional roles for uncharacterized proteins. For each pair of interactions, the number of supporting experiments is indicated, and a link allows switching to the experiment view, which shows a detailed view of all the experiments supporting the given binary interaction. In both views, all terms from controlled vocabularies are hyperlinked, providing direct access to their definitions. Both views allow the selection of specific proteins and their display in the graphical view (Fig. 3). This view shows proteins in the context of their local interaction networks. For clarity, only the local interaction neighbourhood up to a given depth is shown, but the network can be expanded, either globally, or only around specific proteins of interest. Any of the displayed proteins can be selected as the new centre of the interaction network. A separate panel of the graphical view displays all GO terms which are annotated to proteins in the displayed interaction network. Any of these GO terms can be selected, and all proteins which have this GO term or any of its child terms annotated, will be highlighted. This functionality provides a quick method to interactively explore the functional context of proteins.

The emerging Molecular Interaction (MI) standard developed by the Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO) is a joint

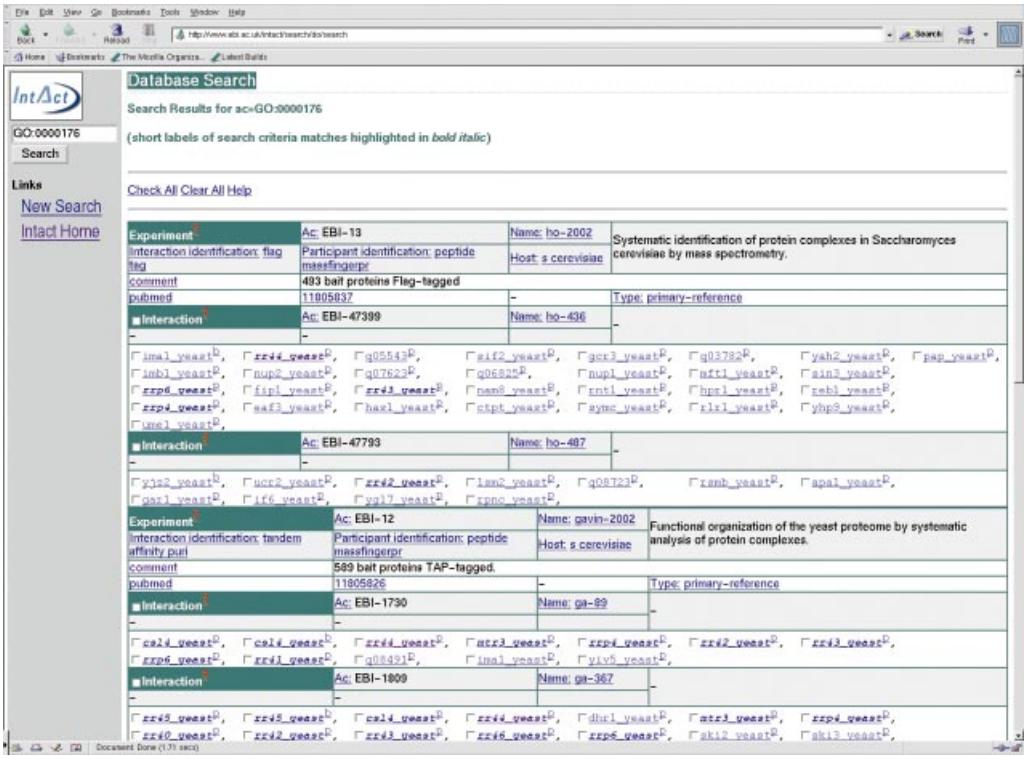

Figure 2. IntAct search interface, experiment view. This view provides details on the context in which interactions have been derived, in particular the experimental methods by which the interaction and the interacting proteins have been determined, and a list of relevant interactions that have been determined in this experiment. All controlled vocabulary terms are hyperlinked to their definition, and field descriptors are hyperlinked to their definition in the online user manual. The query used in this example is the GO accession number GO:0000176. Proteins that are annotated with this GO term are shown in the context of their interactions, and are shown in bold italics.

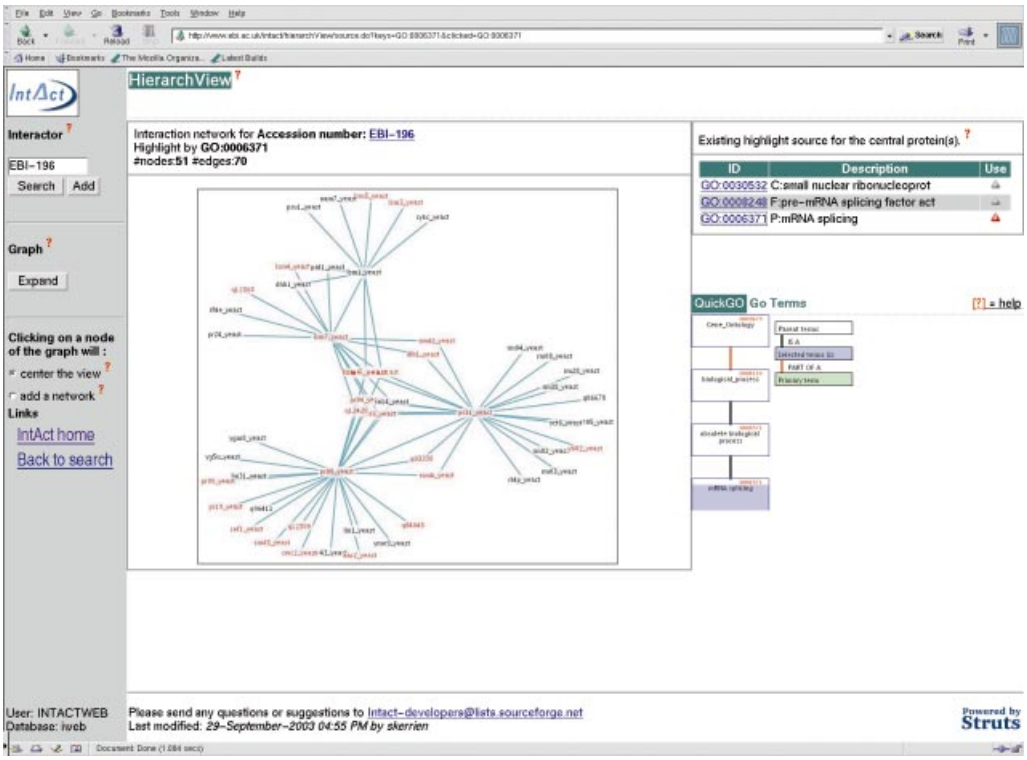

Figure 3. Graphical interaction network viewer. The protein lsm6_yeast (Swiss-Prot Q06406) is shown in the context of its local interaction network. Using the GO browser in the right-hand panel, the GO term 'mRNA splicing' has been selected, and all proteins that have this GO term or child terms annotated, are highlighted.

development of major interaction data providers for the XML representation of molecular interaction data (8). The IntAct project co-develops and supports the PSI MI standard and provides both a web service and a simple URL-based interface, which allow direct computational access to retrieve interaction networks in PSI XML format.

Data content and submission

IntAct currently contains ~2200 interactions, mainly imported from large-scale experiments (3,9), and a growing number of interactions extracted from the literature by the IntAct and Swiss-Prot curation teams. We accept and encourage submission of protein interaction data to IntAct in PSI MI format at datasubs@ebi.ac.uk.

IMPLEMENTATION

IntAct is Java based, and all required internal and external components are publicly available. Using the OJB (http://db.apache.org/ojb/) object-relational mapping tool, IntAct can access either Oracle (http://www.oracle.com/) or Postgres (http://www.postgresql.org/) relational database systems. Based on the database access layer, the IntAct components are independent modules using the Struts (http://jakarta.apache.org/struts/) framework to provide a uniform web interface. The graphical view uses the open source Tulip system (http://www.tulip-software.org/) for efficient graph layout.

DISCUSSION

Storage, retrieval and visualization of protein interaction data is provided by several publicly available databases, in particular BIND (10), DIP (11), MINT (12), CYGD (13) and STRING (14). While the data models and some of the features of these databases have influenced IntAct, additional features such as the representation of binary and *n*-ary interactions in the same object, the tight integration of controlled vocabularies into the data model, and the GO term highlighting in the network browser are to our knowledge unique additions to the set of analysis tools available to the scientific community. In addition, IntAct has been explicitly developed to support local installation. Through its open source modular architecture, IntAct provides a framework which allows easy installation and adaptation to local needs, while the use of well-defined controlled vocabularies supports data consistency between installations.

We are currently developing a system to automatically synchronize the data of IntAct nodes in different installations, to facilitate data exchange between cooperating IntAct installations. This will be similar to the data exchange between the EMBL/GenBank/DDBJ nucleotide databases (15–17). In the framework of the PSI (8), we are also exploring possibilities for regular data exchange between major protein interaction databases, as a means to overcome the current fragmentation of interaction data.

ACKNOWLEDGEMENTS

IntAct is funded by EU grant number QLRI-CT-2001-00015 under the RTD programme 'Quality of Life and Management of Living Resources'.

REFERENCES

1. Legrain,P. and Selig,L. (2000) Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.*, **25**, 32–36.
2. Rigaut,G., Shevchenko,A., Rutz,B., Wilm,M., Mann,M. and Seraphin,B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17**, 1030–1032.
3. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
4. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.
5. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
6. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
7. Weng,S., Dong,Q., Balakrishnan,R., Christie,K., Costanzo,M., Dolinski,K., Dwight,S.S., Engel,S., Fisk,D.G., Hong,E. *et al.* (2003) *Saccharomyces* Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.*, **31**, 216–218.
8. Orchard,S., Hermjakob,H. and Apweiler,R. (2003) The Proteomics Standards Initiative. *Proteomics*, **3**, 1374–1376.
9. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
10. Bader,G.D., Betel,D. and Hogue,C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
11. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S. and Eisenberg,D. (2002) DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.
12. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
13. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
14. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
15. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
17. Miyazaki,S., Sugawara,H., Gojobori,T. and Tateno,Y. (2003) DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res.*, **31**, 13–16.