Published online 3 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D619–D622***
doi:10.1093/nar/gkn863

Reactome knowledgebase of human biological pathways and processes

Lisa Matthews[1], Gopal Gopinath[1], Marc Gillespie[1,2], Michael Caudy[1], David Croft[3], Bernard de Bono[3], Phani Garapati[3], Jill Hemish[1], Henning Hermjakob[3], Bijay Jassal[3], Alex Kanapin[1], Suzanna Lewis[4], Shahana Mahajan[5,6], Bruce May[1], Esther Schmidt[3], Imre Vastrik[3], Guanming Wu[1], Ewan Birney[3], Lincoln Stein[1,7] and Peter D'Eustachio[1,6,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, [2]College of Pharmacy and Allied Health Professions, St. John's University, Queens, NY 11439, USA, [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [4]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [5]Hunter College, New York, NY 10010, [6]NYU School of Medicine, New York, NY 10016, USA and [7]Ontario Institute for Cancer Research, Toronto, ON, Canada M5G0A3

Received September 17, 2008; Revised October 14, 2008; Accepted October 16, 2008

ABSTRACT

Reactome (http://www.reactome.org) is an expert-authored, peer-reviewed knowledgebase of human reactions and pathways that functions as a data mining resource and electronic textbook. Its current release includes 2975 human proteins, 2907 reactions and 4455 literature citations. A new entity-level pathway viewer and improved search and data mining tools facilitate searching and visualizing pathway data and the analysis of user-supplied high-throughput data sets. Reactome has increased its utility to the model organism communities with improved orthology prediction methods allowing pathway inference for 22 species and through collaborations to create manually curated Reactome pathway datasets for species including *Arabidopsis*, *Oryza sativa* (rice), *Drosophila* and *Gallus gallus* (chicken). Reactome's data content and software can all be freely used and redistributed under open source terms.

EXPANDED COVERAGE OF HUMAN PATHWAYS

The current release of Reactome (version 26, September 2008) covers approximately 12.5% of 20 000 curated UniProt human proteins, a 2.7-fold increase over the last three years. Forty-six major domains of human biology, such as apoptosis, the HIV and influenza life cycles, DNA replication, transcription, hemostasis and carbohydrate metabolism are annotated, as are normal functions of 1005 proteins associated with OMIM disease phenotypes (http://www.ncbi.nlm.nih.gov/omim/).

IMPROVED TOOLS, SOFTWARE AND DATA MODEL

Revised orthology prediction methods

The OrthoMCL clustering procedure (1,2) (http://reactome.org/electronic_inference.html) applied to data from OrthoMCL DB (http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi), Version 2, is used to identify orthologs of curated human proteins in each of 22 evolutionarily divergent species for which high-quality whole-genome sequence data are available (3). In line with changes in the OrthoMCL clustering procedure, only the longest transcript of each gene is considered, and a gene-based rather than a protein-based method is used to map the Ensembl identifiers used by OrthoMCL to the UniProt accessions used in Reactome. These changes have improved our success rate for electronic inference without measurably affecting accuracy.

Improved tools for analysis of large-scale data sets, data-mining and modeling

SkyPainter. The current version of the SkyPainter tool allows users to more effectively visualize the functional relationships among genes identified in large-scale

*To whom correspondence should be addressed. Tel: +1 212 263 5779; Fax: +1 212 263 8166; Email: deustp01@med.nyu.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

experiments (Figure 1A and B). For a user-submitted list of genes, each reaction arrow on the reaction map is colored according to the number of user-specified genes whose products participate in the reaction. In addition, hypergeometric testing is now used to display statistically over-represented events in the event hierarchy. Statistically over-represented events may also be viewed as an ordered list and a mapping from submitted identifiers to reactions is provided. The colored reaction maps can be downloaded in publication quality PNG, SVG or PDF format.

Biomart. The newly implemented Reactome BioMart tool (www.biomart.org) facilitates data mining, cross-database analysis and large-scale analysis of gene function. A user can formulate queries across selected data sets (pathways, reactions and complexes) specifying data attributes and filters to narrow searches. For example, querying the Reactome dataset, a user can identify all complexes that contain a given protein (Figure 1C). Alternatively, a user can link a query of the Reactome 'complexes' dataset to a UniProt proteome query and retrieve sequences of all the proteins in these complexes.

Popular searches, e.g. all reactions or proteins or genes in a given pathway, all pathways inferred for a given species or all reactions or pathways involving a set of specified genes can be launched with predefined 'canned' queries. Additional context-sensitive help documentation is under construction.

Changes in the Reactome data model. These have been minimized, to facilitate data curation and use of the knowledgebase. Two additions are a 'black box' reaction class to allow the annotation of events for which not all defining attributes can be provided and an 'entityOnOtherCell' attribute to allow description of individual events and complexes that span two cells.

COLLABORATIONS

We are actively collaborating in the creation of model organism Reactome projects for *Arabidopsis*, rice, *Drosophila* and chicken. The Arabidopsis Reactome (http://arabidopsisreactome.org/), developed with the human Reactome data model, software and curation tools, contains seven manually curated pathways and 311 pathways inferred from data in KEGG (http://www.genome.ad.jp/kegg/pathway.html) (4) and MetaCyc (http://metacyc.org/) (5). Arabidopsis Reactome events have been electronically projected onto five other predicted plant proteomes using the OrthoMCL prediction method described above (6). Curation of *Drosophila* and chicken reactions is underway in collaboration with FlyBase and the Gallus Knowledge Extraction and Annotation project using the Reactome infrastructure and curation protocols; the first releases of these databases are planned for late 2008.

In collaboration with the BioHealthbase group at University of Texas Southwestern Medical College, the life cycle of Influenza virus has been curated and reviewed and is publicly accessible as a segment of Reactome (7).

In collaboration with the Protein Ontology group—PRO (8), Reactome annotations of complexes and physiological states of proteins are being applied to build a hierarchy of proteins and their functional forms. For example, protein objects used to illustrate TGF-β signaling in PRO use corresponding protein objects in the Reactome annotated pathway. The Reactome and PRO collaboration, as a source for curated annotations of proteins, extends the application of Reactome to a larger ontology community.

INCREASED DATA INTEGRATION AND USER SUPPORT

We are increasing the accessibility of Reactome data through data integration and improved online documentation. Reactome pathway annotations are now included in the Pathway Interaction Database (http://pid.nci.nih.gov/) (9). Reactome's online documentation is now available as WIKI pages that provide easy access to user, author and curator guides as well as glossaries and standard operating procedure (SOP) guidelines. In addition, an Editorial Calendar lists modules being prepared for the knowledgebase, their planned release dates and contact information for module curators.

To support data mining, analysis and modeling of Reactome content by other groups, individual reactions and pathways can be exported in SBML (http://sbml.org/Main_Page) (10), Protégé (http://protege.stanford.edu), Cytoscape (http://www.cytoscape.org/) (11) and BioPax (http://www.biopax.org/) (levels 2 and 3) formats. The entire data content of Reactome can be downloaded as a MySQL database or in SBML or BioPax 2 and 3 formats. A SOAP based Web Services API is now available to access the Reactome data. Details about this API are provided in many forms including a 'Flash' tutorial and a PDF user's guide at http://www.reactome.org/download/index.html.

FUTURE DIRECTIONS

A long-term objective of the Reactome project is to provide users with intuitive graphical representations of pathways and reactions. Toward this goal, Reactome has developed a beta version of an entity-level pathway visualization tool: through enhanced navigation features including improved zooming, scrolling and event highlighting, this tool provides a detailed graphical representation of the relationships among the molecules participating in reactions (Figure 1D).

Hover-over molecule/reaction descriptions allow quick and easy scanning of the molecular details of each reaction. An improved search option with an auto-complete feature at the top of the page allows directed queries over reaction and pathway names, entity names, identifiers and GO molecular function terms.

New author tool with pathway diagram editing capability

We are in the process of developing a user interface that couples the powerful navigation and search features of the entity level pathway visualization tool with software that

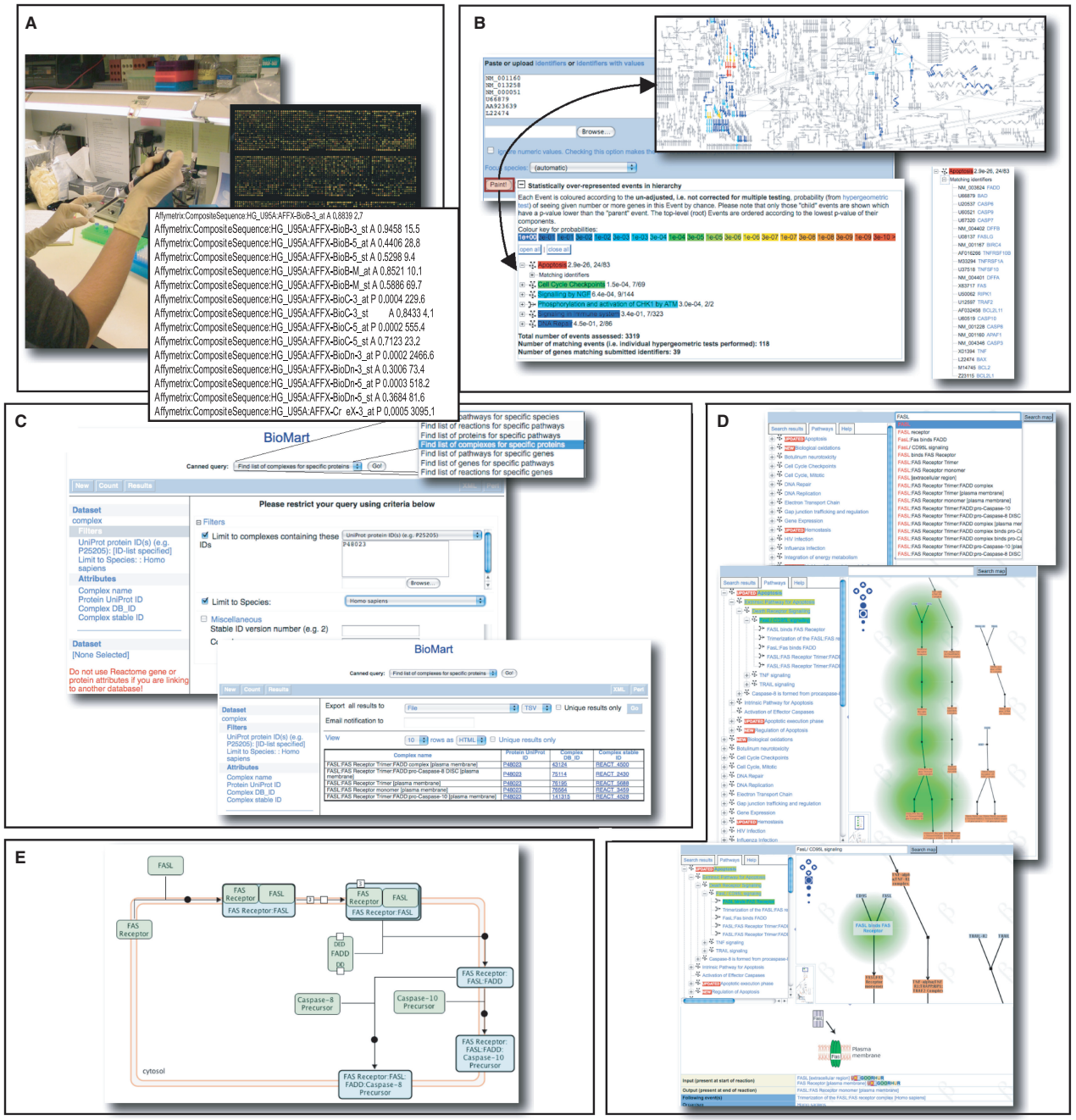

Figure 1. Practical applications of Reactome tools in data analysis. Gene sets derived from large-scale experiments such as microarray analyses (**A**) can be analyzed using the SkyPainter (**B**). The reaction arrows in the reaction map are colored according to the number of genes in the user-supplied data set whose products participate in the reaction. Statistically over-represented events (ones with significantly more user-specified participants than would be expected if the user-specified genes were randomly distributed over events) are highlighted in color and can be viewed as an ordered list. A mapping of submitted identifiers to reactions is provided. These colored reaction maps can be downloaded in publication quality PNG, SVG or PDF formats. Here, Apoptosis is the overrepresented pathway. (**C**) BioMart can be used to learn more about genes in a data set. To identify the complexes that contain a protein/proteins of interest, the 'complexes' data set is selected, the protein Uniprot identifier(s) are entered as a filter and complex names/identifiers are selected as attributes to be displayed. By selecting the 'Dataset' button (bottom left), users can combine searches across additional databases such as UniProt to retrieve additional data, e.g. the amino-acid sequences of the individual proteins making up the complexes annotated in Reactome. In this example, numerous FASL:FAS receptor complexes have been identified. (**D**) The entity-level pathway visualization tool can be used to identify events involving FASL:FAS receptor complexes. Searching on FASL produces a hit list presented in the hierarchy panel. The selected event/entity, Fasl/CD95L signaling is displayed on the map highlighted in green. Reactions are represented as arrows. Reaction names are displayed by clicking on the 'nodes'. Molecules are represented as boxes and their names are displayed upon scroll over. Selecting an event in the event hierarchy centers the map on that event and highlights it. Selecting a reaction on the map highlights that reaction in the event hierarchy. A zoom/scroll box is available in the upper left. A 'birds-eye view' of a pathway is provided in a box at the lower left. Dragging the box in this view repositions the focal point of the zoomed view. The entire pathway can be moved within the window by clicking/dragging it. A collapsible details section at the bottom of the page provides the option to view the event page description of the selected reaction or pathway. (**E**) A part of the FASL/CD95L signaling pathway drawn using the new author tool. Diagrams include a graphical representation of the subcellular localization of the reactions and their constituent molecules as well as the stoichiometry, states and post-translational modifications and binding features of these molecules.

draws intuitive 'textbook style' illustrations of pathways. To this end, we have recently developed an SBGN (http://www.sbgn.org) based version of the Reactome author tool that can create such diagrams (Figure 1E). Currently, this has been released as a beta test version and is available as a Java Web Start application at the Reactome development site: (http://brie8.cshl.edu:8080/ReactomeTools/AuthorTool/authorToolLaunch.html).

ACKNOWLEDGEMENTS

We are grateful to many scientists who collaborated with us as authors and reviewers to build the content of the knowledgebase, and for the helpful comments of the reviewers of this manscript.

FUNDING

The development of Reactome is supported by a grant from the US National Institutes of Health (P41 HG003751) and a grant from the European Union Sixth Framework Programme (LSHG-CT-2003-503269). Funding for open access charge: US National Institutes of Health P41 HG003751.

Conflict of interest statement. None declared.

REFERENCES

1. Li,L., Stoeckert,C.J. Jr. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **9**, 2178–2189.
2. Chen,F., Mackey,A., Stoeckert,C.J. Jr. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
3. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
4. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
5. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.
6. Tsesmetzis,N., Couchman,M., Higgins,J., Smith,A., Doonan,J.H., Seifert,G.J., Schmidt,E.E., Vastrik,I., Birney,E., Wu,G. *et al.* (2008) Arabidopsis Reactome: a foundation knowledgebase for plant systems biology. *Plant Cell*, **6**, 1426–1436.
7. Squires,B., Macken,C., Garcia-Sastre,A., Godbole,S., Noronha,J., Hunt,V., Chang,R., Larsen,C.N., Klem,E., Biersack,K. *et al.* (2008) BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Res.*, **36**, D497–D503.
8. Natale,D.A., Arighi,C.N., Barker,W.C., Blake,J., Chang,T.C., Hu,Z., Liu,H., Smith,B. and Wu,C.H. (2007) Framework for a protein ontology. *BMC Bioinformatics*, **8 (Suppl 9)**, S1.
9. Schaefer, C.F. (2006) An introduction to the NCI Pathway Interaction Database. *NCI-Nature Pathway Interaction Database*, doi:10.1038/pid.2006.001.
10. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al.* (2003) The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
11. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocols*, **2**, 2366–2382.