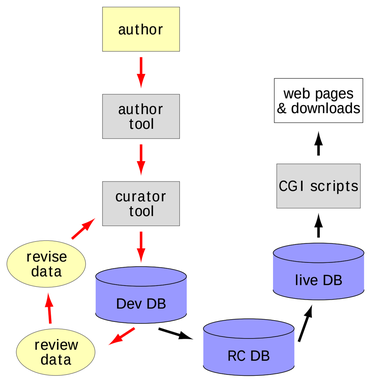
Reactome - a knowledgebase of human biological pathways

Peter D'Eustachio[1,2], David Croft[3], Bernard de Bono[3], Gopal Gopinath[1], Marc Gillespie[1,4]
Bijay Jassal[3], Lisa Matthews[1], Esther Schmidt[3], Imre Vastrik[3], Guanming Wu[1], Suzanna Lewis[5], Ewan Birney[3], Lincoln Stein[1]

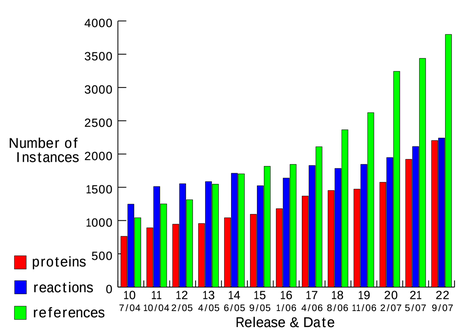
[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor NY 11724; [2]NYU School of Medicine, 550 First Avenue,New York NY 10016; [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD United Kingdom; [4]College of Pharmacy and Allied Health Professions, St. John's University, 8000 Utopia Parkway, Queens NY 11439; and [5]Lawrence Berkeley National Laboratory, 1 Cyclotron Road 64R0121, Berkeley CA 94720

The Reactome knowledgebase of human biological processes is
Reductionist. All of biology can be represented as events that convert input physicalentities into output physical entities located in compartments.
A generic parts list. Tissue and state specificity of events are not captured.
Qualitative. Kinetic parameters and data are not captured.
Human-centric. Experiments may use reagents from diverse sources, but our focus is on human biological processes.
Manually curated. Events are annotated by expert curators, and linked to published data.
Open source. All data and software are freely available at www.reactome.org.

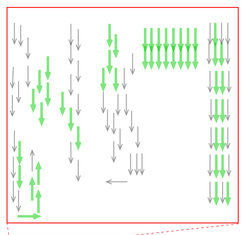
Curation and release workflow



Manual curation statistics, 2004-present

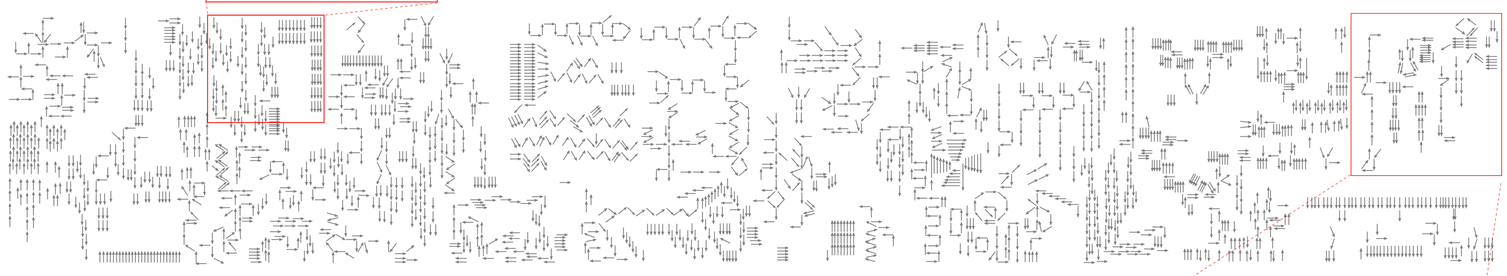


Curated human reactions as a template for manual curation of other species: a projected chicken reactome

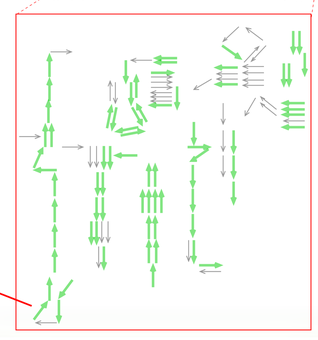


signaling pathways

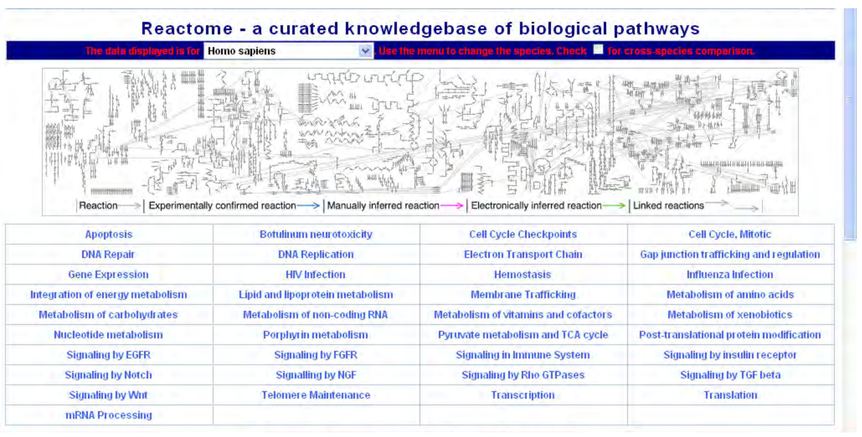
human reaction for which G. gallus counterpart could be inferred
(no computed human-to-G. gallus inference)

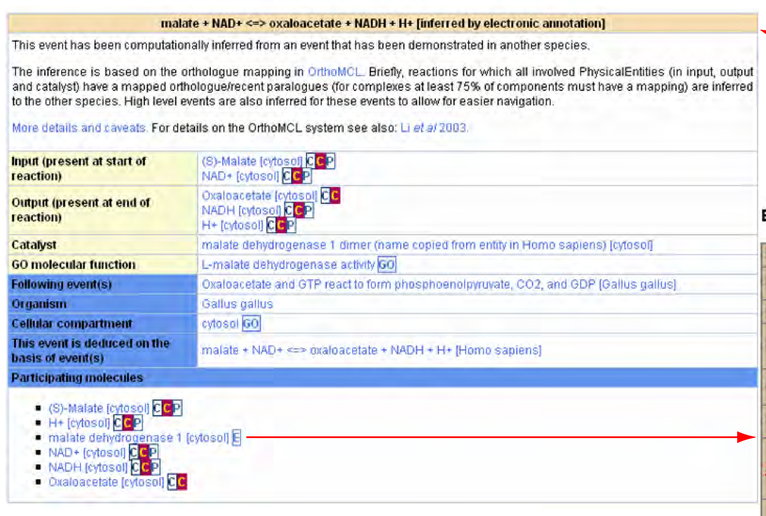


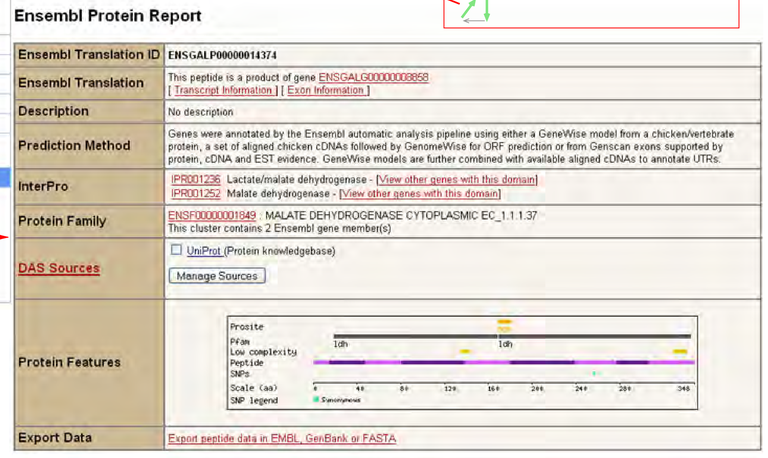
glucose metabolism



Reactome as an online textbook of human cell and molecular biology







Curation of human biological processes

All the information in Reactome comes from expert curation. Reactome curators collaborate with independent research scientists who are recognized experts in these areas to annotate human biological processes. While the Reactome data model can accommodate alternative, controversial versions of a single biological process, as a matter of editorial policy, to maximize the value of Reactome as a data mining resource for users, experts are asked to construct views of processes that reflect current expert consensus. The expert and a curator work together to create an electronic outline to define the exact scope of the biological process to be annotated and to identify and order the reactions that comprise the process. The expert uses a graphical application called the Reactome Author Tool to add molecular detail to the outline, e.g., the identities and subcellular locations of the molecules that participate in each reaction, the role of each molecule (input, output, regulator, catalyst), the compositions of multimolecular complexes, the order of reactions within a pathway, citations of key primary research publications, and brief free text descriptions of each reaction and pathway. The curator then uses another graphical application called the Reactome Curator Tool to revise this material and integrate it into the Reactome data scheme. Molecules are linked to their corresponding reference entities and, where appropriate, organized into sets, catalyst activities are linked to GO molecular function terms, and links are created between the new reactions and ones already in Reactome. This information is then uploaded directly from the curator tool into the Reactome development database, so that it can be reviewed by the expert author and other Reactome curators, viewing it on the development version of the Reactome web site. The curator then revises the material as appropriate using the curator tool. Finally, the module is peer-reviewed on the development web site, by one or more bench biologists selected by the curator in consultation with the author. The peer review is open and the reviewers are acknowledged in the database by name.

Reactome follows a quarterly release schedule. The process of creating a release database starts with extracting the finished modules a nd associated information into a separate "slice" database. Quality assurance scripts are run to check the completeness and consistency of the data. If necessary, material in the development database is revised, and a new "slice" database is generated. We then add cross-links to other relevant external resources and the new database is made available via the public web site.

The database now contains annotations of 2203 human proteins, involved in 2241 reactions that in turn are grouped into 37 modules. This material can be browsed as an on-line textbook at www.reactome.org.

Inference of pathways in other species

Each Reactome release includes computationally inferred pathways and reactions in 22 non-human species including Mus musculus, Tetraodon nigroviridis, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Arabidopsis thaliana, Oryza sativa, Plasmodium falciparum, and Escherichia coli. Together they represent more than 4000 million years of evolution and span the major branches of life.

The inference process begins with the set of peer-reviewed curated human reactions in the pre-release database. We project these curated reactions onto the genomes of the selected species using protein similarity clusters derived by the OrthoMCL method. OrthoMCL finds reciprocal best similarity pairs of proteins for each protein and pair of species, as well as "reciprocal better" similarity pairs within species. The latter are proteins that are more similar to each other within the same species than to any protein in the other species. These pairs are entered into a similarity matrix, normalized by species, and clustered using a Markov chain length algorithm to generate sets of related proteins that include both orthologues and recent paralogues that post-date the divergence of the two species. Each curated human reaction is checked to ask whether the proteins involved in it have at least one orthologue or recent paralogue in the other species. In the case of protein complexes, we relax this requirement so that a complex is considered to be present in the other species if at least 75% of its protein components are present in the other species. For each reaction that meets these criteria, we create an equivalent reaction for the species under consideration by replacing all human protein components with their OrthoMCL counterparts from the second species. If a human protein corresponds to more than one protein in the second species, a protein set named 'Homologues of ...' is used as the corresponding component of the reaction in the second species.

The rate of reaction inference ranges from 87% for human-to-mouse to 10% for human-to-Micrococcus. Comparison of reactions inferred in this way for Saccharomyces to expert-curated Saccharomyces reactions indicates a false positive rate (we incorrectly predict a nonexistent yeast event) of 22% and a false negative rate of 28% (we fail to predict an authentic yeast event).

These inferred reactions and pathways are integrated with their curated human counterparts for display on the Reactome web site, linked to Ensembl entries for the non-human proteins. This projection of the Reactome human data set can thus serve as a template for rapid manual curation of events in a non-human species of interest. A current project to annotate Gallus gallus (chicken) is outlined in the top panel on the right

Projection of high-throughput data onto Reactome

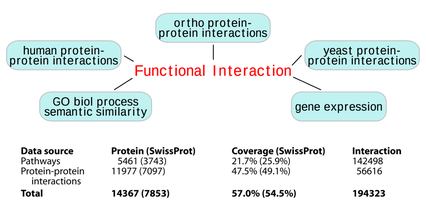
The SkyPainter tool allows a user to overlay expression data onto the Reactome event set to explore patterns in the expression data. The lower panel on the right shows the projection of protein expression data for human pancreatitis and pancreatic cancer generated by Chen, Brentnall, and colleagues, and the identification of a protein up-regulated in cancer but not pancreatitis.

Incorporation of high-throughput data into Reactome

While manual curation remains the gold standard for generating high-quality annotations linking human proteins and their functions, we have also sought approaches to allow the incorporation of protein-protein interaction data into Reactome. Work now underway is aimed at developing a scoring system using a Naïve Bayes Classifier. Five predicators are used for the classifier: human protein-protein interactions, interactions predicated from fly and worm, interactions predicted from yeast, gene coexpression from DNA microarray data sets, and sharing of GO biological process annotations. To train the Classifier, we have generated a positive dataset from reactions in six pathway databases, and a negative dataset using protein pairs that are from the same pathway but that do not have functional interactions. Using these external data sources, we have increased our protein coverage from 6% to 57%. Validation studies are underway to assess the quality of these predictions.

A web display now under development will allow proteins thus identified as "functional interactors" of annotated ones to be displayed as clusters of entities linked to the annotated ones.

Naive Bayes Classifier

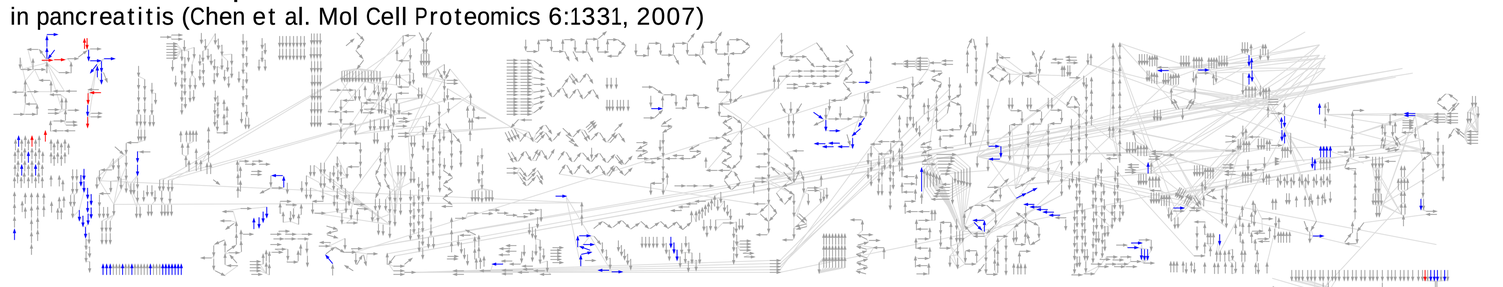


Data source	Protein (SwissProt)	Coverage (SwissProt)	Interaction
Pathways	5461 (3743)	21.7% (25.9%)	142498
Protein-protein interactions	11977 (7097)	47.5% (49.1%)	56616
Total	**14367 (7853)**	**57.0% (54.5%)**	**194323**

Reactome as a tool for analysis of gene expression data sets: protein expression in pancreatitis and pancreatic cancer

Protein overexpression

in pancreatitis (Chen et al. Mol Cell Proteomics 6:1331, 2007)



in pancreatic cancer (Chen et al. Gastroenterology 129:1187, 2005)