Reactome: An integrated expert model of human molecular processes and access toolkit

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

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

Summary

The behaviour of pervasive molecular processes in human biology can be studied through the large-scale modeling of the molecular events that define them. Constructing detailed models of such extent and scope is a considerable undertaking well beyond the reach and capability of individual efforts, due to the range of expertise required. Reactome (http://www.reactome.org) is an open-access project that collaborates with field experts to integrate their pathway knowledge into a single quality-checked human model. This resource dataset is systematically cross-referenced to major molecular and literature databases, and is accessible to the community in a number of well-established formats. Various tools have been developed to facilitate querying and interaction with this content. The salient features of the annotation strategy are discussed here, and examples of pathway and genomic data integration using flexible interfacing methods from the associated toolkit are also presented.

1 Introduction

The complexity of the molecular network at the basis of human physiology stands as an important reminder that 'everything is connected to everything else'(First Law of Ecology [1]). Physiology studies the change in key parameters such as pressure, tension, temperature and electrolyte concentration brought about by adaptations of this molecular network that alter the electrical and material properties of tissue fabric.

Building a detailed functional molecular model to gain further understanding of the behaviour of this vast physiological network is a formidable task. Representation, even of the most basic and fundamental of processes such as the homeostasis of H^+ levels in body fluids (described in Fig. 1.), requires extensive mechanistic knowledge of both cause and effect, spanning a number of molecular pathways and organ systems.

* corresponding author – email: bdb@ebi.ac.uk

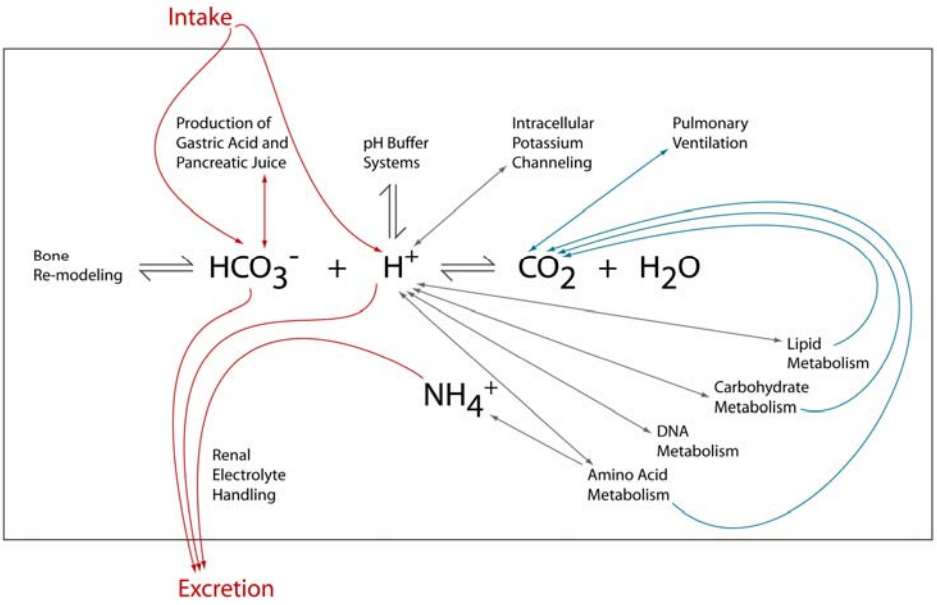

Figure 1: Synopsis of different factors contributing to acid-base balance. A considerable number of cellular functions are sensitive to alterations in pH, such that H^+ concentration is one of the most tightly controlled parameters in the body. Functions that are sensitive to pH include glycolysis, the pentose phosphate cycle, DNA synthesis, K^+ and Ca^{2+} channel activity, gap junction conductance and haemoglobin oxygen affinity. Apart from the tight interplay of lungs and kidneys in the elimination of CO_2 and HCO_3^- respectively, a substantial proportion of renal acid elimination occurs in the form of NH_4^+, created through the breakdown of the amino acid glutamine by the enzyme glutaminase. As the availability of glutamine in the blood also depends largely on glutamine metabolism in the liver, the differential regulation of glutaminase gene expression in the liver and kidneys to acidosis is a key mechanism in the body's pH homeostasis. Adapted from [2].

A pathway model provides the starting point for a number of investigations. With increasing volumes of high-throughput protein interaction and gene expression results, it is crucial to interpret such data in the functional context of a standard pathway reference framework. The ability to map experimental results onto a curated model, is therefore a key step to gaining insight through the correlation with pathway-specific knowledge. The structure of the model is also indicative of the expected behaviour of its components. Modeling of protein and small molecule connectivity thus provides a way to analyse crosstalk and feedback loops that determine the functional interdependencies between network elements [3]. Therefore, integrating detailed knowledge of physiological mechanisms enables the logical analysis of their pathways, as well as the identification of optimal intervention target points for further scientific enquiry and biotechnological development.

However, the creation of a molecular model of sufficient quality and breadth to address for instance, pH homeostasis, is hampered by a number of production issues. The first is securing the biological expertise necessary to describe molecular mechanisms ranging from amino acid metabolism and DNA synthesis to bone remodeling and renal electrolyte handling [2]. Secondly, given the resources invested in such an undertaking, the model then requires (a) regular maintenance and updating, as well as (b) packaging in a manner that is accessible to and adaptable by the scientific community.

The main objective of the Reactome Knowledgebase [4] is to provide a scalable solution to these production issues by integrating verifiable functional pathway data into a unified human

model under constant expert and editorial supervision. At a software level, components of the Reactome toolkit are used by a number of query applications within BioMart (http://www.biomart.org) [5], ENFIN (http://www.enfin.org) [6] and caBIG (https://cabig.nci.nih.gov) [7, 8] frameworks, ensuring further integration with these large systems projects. In this work, we discuss (a) the curation methodology, and (b) those portable tools that have been created to simplify access to this resource.

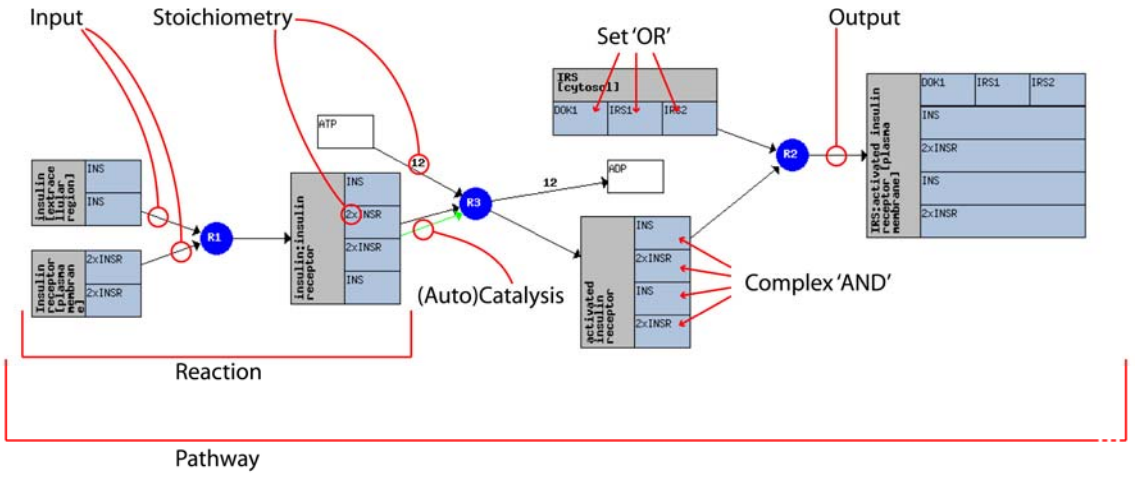

Figure 2: The interactions of five Complexes, one Set and two small molecules described over three consecutive Reactions (filled blue circles) that form part of the larger Insulin Receptor Cascade Pathway. Protein rectangles are shaded blue, small molecules are white. The reaction 'R3' depicts the autophosphorylation of the hormone-bound receptor. On the left hand side, note how the Insulin (INS) and Insulin Receptor (INSR) gene symbols each label two protein fragments that dimerize and tetramerize respectively to create the functional form of the proteins. The Set represents three human IRS proteins that play a role in binding the activated receptor at this stage in the cascade.

2 Methods

2.1 Model building

The Reactome model of human molecular biology consists of a broad descriptive graph in which biological function is defined in terms of a change in molecular structure, often brought about by some participating catalyst that brokers this step (Fig. 2). Such a descriptive unit is called a Reaction. This simple graph structure connects molecules present before a function is carried out (the input) to those entities produced from its outcome (the output). This provides a meaningful context for the interaction of proteins with each other and with other molecules. A set of reactions in Reactome, usually consecutive and interlinked, can be grouped to form a Pathway. A list of top-level pathways is featured on the main panel of the Reactome website shown in Fig. 3. Reactions and Pathways are collectively known as Events.

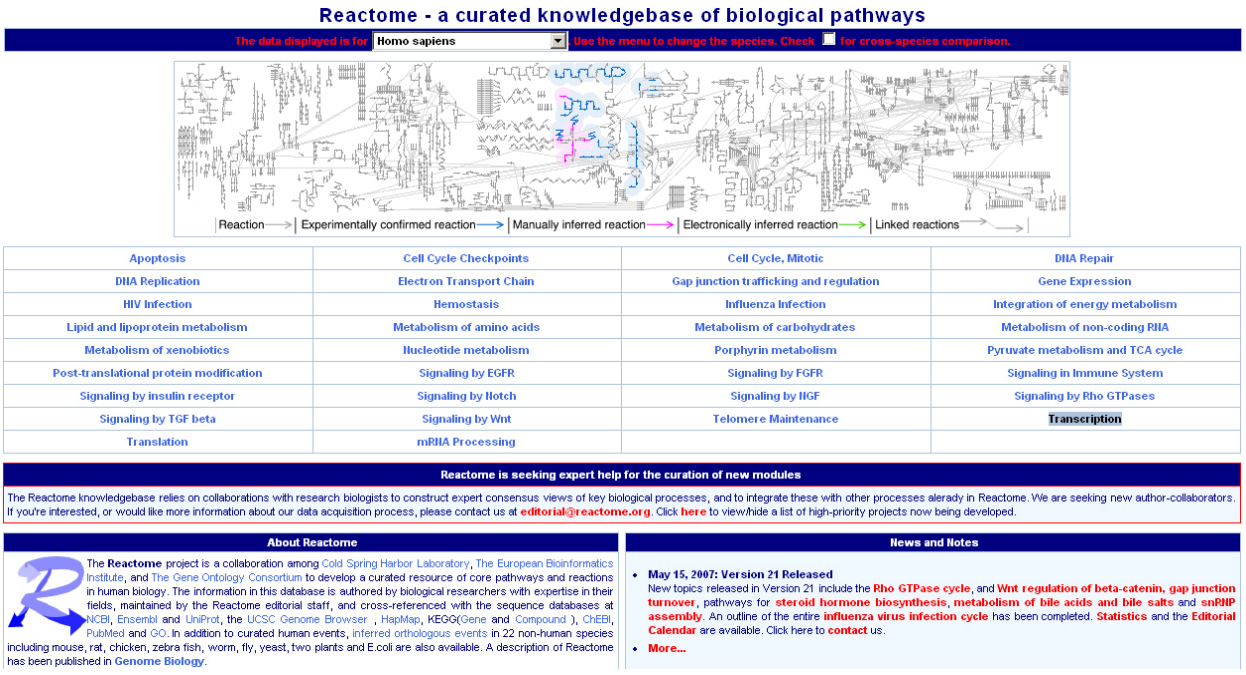

Figure 3: The Reactome home page. This view highlights the Transcription Pathway on the 'Sky' panel and in the pathway table beneath it.

In collaboration with Reactome, the expert biologist plays a central role in extending this graph model by creating new Pathways on a particular topic module in the style of a formal literature review process. The structure and content of such Pathways are constructed under the direct supervision of the expert to reflect current consensus in the field. Apart from the functional details ingrained in the graph connectivity of the model *per se*, Events are packaged with additional information and links to external resources. Protein and small molecule entities are cross referenced with accession identifiers to a number of well-established databases (e.g. UniProt [9], KEGG [10], ChEBI [11], PubChem [12]), while both Entities and Events are further linked to standard 'Molecular Function', 'Biological Process' and 'Cellular Component' ontology terms found in the GO vocabularies [13]. Any component of the new module may be further qualified by the expert biologist through the association of key literature references, as well as the annotation with original diagrams and summaries to highlight items of interest. This publication process is completed through review by a second independent expert who checks for quality and clarity and suggests refinements, prior to release.

2.2 Modeling strategy

The data available through the Reactome website focuses on human pathways. Therefore, when creating this annotation, the expert author's priority is to utilize interaction evidence based on experiments carried out using human cells. However, for practical and ethical reasons, a substantial amount of research has been carried out on non-human 'model' organisms. This poses a problem of applicability for the expert to solve – which insights derived from one species can be legitimately projected onto another? For instance, are pathways in a mouse hepatic cell line identical to those in the human liver? What lessons learnt from the study of eye development in Drosophila are applicable to human embryology?

In those 'non-human' cases in which applicability to human is ascertained, our protocol is to construct Events pertaining to the model organism interactions first. These are annotated using the original literature reference as evidence. The corresponding molecules from human

are then selected to create a new set of inferred Events that point to the equivalent lower organism annotation as evidence.

2.3 The model toolkit

The detailed description of biological function in terms of molecular change requires a detailed representation of biological structure. Therefore, the eloquence of the process model in Reactome owes much to the following five key properties of Physical Entity representation (see Fig. 2):

1) **Strict referencing** uses external accession identifiers to enable interaction tracking for any molecule across the entire model. This also establishes an independent measure of growth and coverage for the Reactome network (Fig. 4). In the case of molecules that are not sequence-based, the ChEBI database (Chemical Entities of Biological Interest) plays a key role by providing expertise and curatorial support to the addition of new small molecules in Reactome.

2) **Granularity:** While the re-arrangement of structure in a small molecule can be described in terms of a change from one ChEBI accession ID to another, it is more complicated to represent protein modification in discrete form. The states of phosphorylation or palmitoylation of a protein, to mention just two instances, can not be distinguished on the basis of a change in its UniProt accession ID. The same holds true for protein cleavage into fragments, as well as sequence polymorphism. In Reactome, any shift from the original form of the protein results in the creation of a new Physical Entity. The start and end amino acid position is recorded in the case of fragmentation. If a particular residue is modified, the nature of the new chemical group is referred to in terms of the corresponding ChEBI accession ID.

3) **Localization**: A number of biological processes are strictly partitioned such that the transition of a molecule from one compartment to another may have profound effects, and is therefore held under strict control (as in the case of signaling triggered by the influx of calcium ions into the cytosol). As compartment type is one of the basic defining features of a Physical Entity, a transport Reaction is able to simply map, as input and output, two distinct Entities that refer to the same molecular accession ID but have different localization properties. The range of compartments utilized by Reactome is a subset of standard GO 'Cellular Component' terms of subcellular locations.

4) **Equivalence**: A specific role in a Reaction may be assumed equally well by a number of equivalent molecules. For instance, different isoforms of regulatory and catalytic components of an enzyme dimer may exist (e.g. PI3K). Another example may involve a large family of hormones binding differentially to a corresponding set of related receptors (e.g. FGF receptors). On similar lines, it may be required to represent the number of different molecules transported by the same membrane channel (e.g. bile salts co-transport with sodium). The use of Sets in Reactome does away with the necessity of depicting every possible combinatorial Reaction instance, without losing any of the detail such an Event is required to convey.

5) **Assembly**: The formation of molecular complexes is a mainstay in representing a number of biological scenarios. All types of Physical Entity, including small molecules, proteins, Sets and any other complex can be used as a component for assembly.

The potential descriptive space of the Physical Entity is therefore considerable, being roughly the product of (1) the set of small molecules and chemical groups, (2) all possible protein

fragments in all species, and (3) all cellular compartments. Any number of Physical Entities may feature in one Reaction, in an input, output or catalytic role. The skill essential to the Reactome curatorial process is to match the requirements of the expert biologist using the appropriate descriptive instruments from this data model palette.

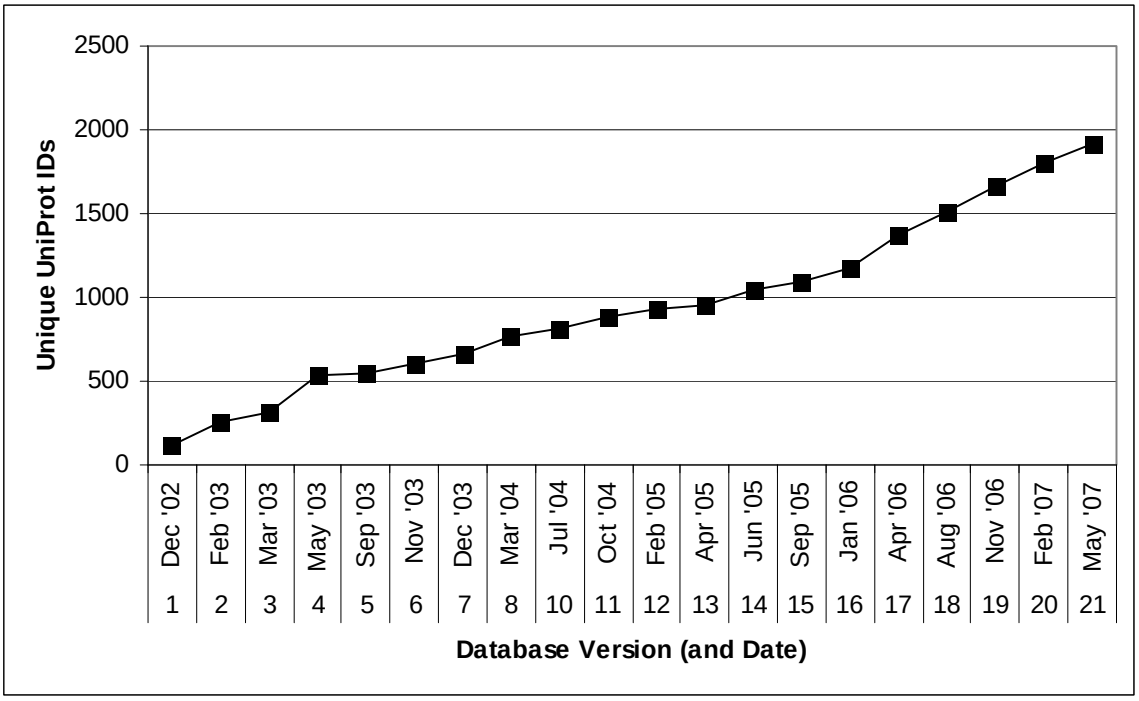

Figure 4: The expanding coverage of the published Reactome human model in terms of unique UniProt Accession IDs over the past 21 releases.

2.4 The access toolbox

A number of complementary tools have been developed to access, maintain, update and broadcast this material. In view of the open access nature of this toolkit, it is feasible to install all Reactome software locally (http://www.reactome.org/download/index.html) to carry out customized operations for analysis and curation. Consortia working on *Drosophila melanogaster* (http://www.flybase.org) [14] and *Arabidopsis thaliana* (http://www.agron-omics.eu) [15] biology have recently adopted these tools in support of their pathway annotation efforts.

All Event network and Physical Entity data is served through MySQL, and accessed using Application Programming Interface (API) tools, available as Perl, Java and SOAP-based webservices kits. For instance, integrative tools and services developed within the ENFIN modeling and caBIG cancer consortia also build upon this connectivity framework. The Perl and Java API classes drive the website and curation clients respectively but such APIs can also be integrated into any implementation to access and manipulate Reactome data tables directly. The Perl toolkit in particular integrates closely with APIs that connect to the Ensembl [16] resource, facilitating cross-talk between genomic and pathway network data. For instance, the combined use of Reactome and EnsEMBL APIs permits the seamless enumeration of gene pairs from the same chromosome that participate in the same reaction in Reactome (results are discussed below and illustrated in Fig. 5).

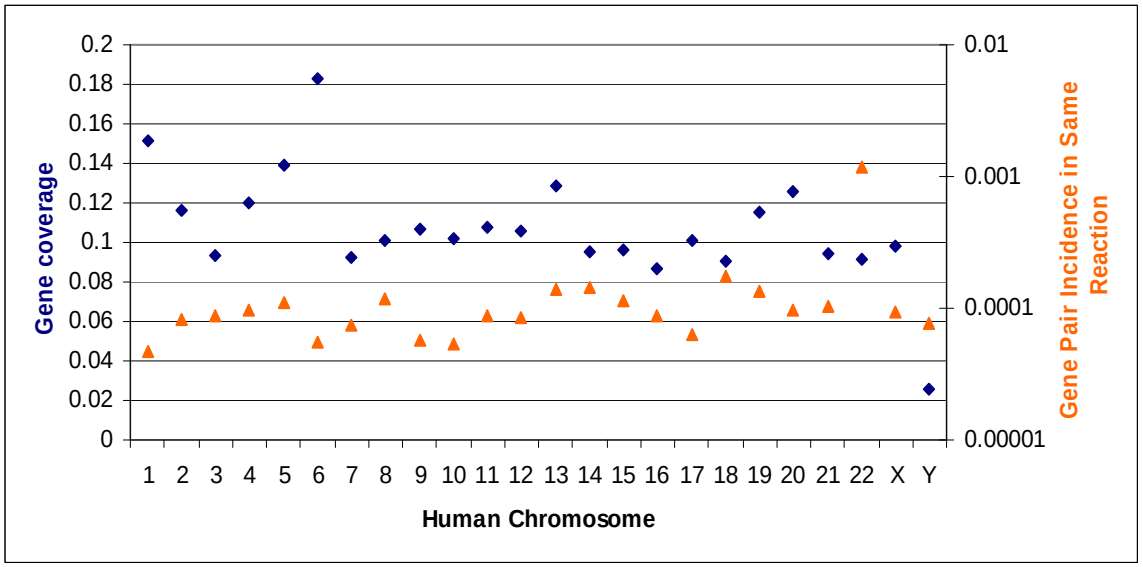

Figure 5: Examples of integrated genome-pathway analyses. For every human chromosomal gene set: (a) the number of genes mapped to Reactome Physical Entities in relation to the total number of genes per chromosome in Ensembl (left Y axis, dark blue), (b) incidence index of a pair of gene products from the same chromosome involved in the same Reaction (right Y logarithmic axis, orange).

The main Reactome website provides an orderly and hierarchical presentation of pathway annotation, consisting of a series of panels with author reviewed diagrams and summaries, as well as hyperlinked Reaction depictions. There is also extensive documentation on how to use Reactome resources. The site also supports a number of query functions and other services, such as:

1) **Visualization**: The top graphical panel of the website (Fig. 3), known as the 'Sky', lays out Pathway constellations of all Reactions for user 'point-and-click' interaction, providing a global context for every human molecular event.

2) **Simple Searches**: The website's 'SkyPainter' provides, amongst other things, a simple interface to highlight Reactions on the Sky given a submitted list of recognized accession IDs for sequence, small molecule, and other established data types (e.g. GO, InterPro, Affymetrix, MIM etc). As identifiers may also be followed by numeric qualifiers, this tool is well suited to summarizing the influence and effects of differential gene expression, for example. Submissions with multiple numeric columns, such as a time course series, are rendered as an animated movie. Filtered lists of Event and Physical Entity entries can be generated using the text query tool in the 'Extended Search' section.

3) **Complex Queries**: Reactome has recently launched a BioMart service (Fig. 6). This simple, federated query system is designed for use with large datasets such as Ensembl, Hapmap [17] and UniProt. It is based on the 'star' implementation of database schemas where a single main table is linked to different dimension tables [18]. The overall simplicity of these schemas enables fast data retrieval of complex queries such as 'what genes are involved in a particular pathway?'

4) **Orthology Mapping**: The OrthoMCL method [19] uses protein sequence matching and clustering to produce a set of orthologs and recent paralogs between two species. These orthology maps play a central role in automatically projecting human Reactome Events onto a number of lower organism gene sets, providing suggested pathways across which their products may interact. These electronically inferred resources for

lower organisms can be explored and queried through the identical web interface used to navigate human data ('Sky' included).

5) **Export Services and External Links**: Through the website, every Event can be individually exported in a number of well-established formats such as SBML (http://sbml.org), BIOPAX (http://biopax.org) and Cytoscape (http://cytoscape.org) or repackaged in PDF/RTF for printing and perusal. Gene products map to external database records such as UniProt, KEGG, Entrez Gene, MIM, RefSeq, HapMap and the EnsEMBL and UCSC genome browsers.

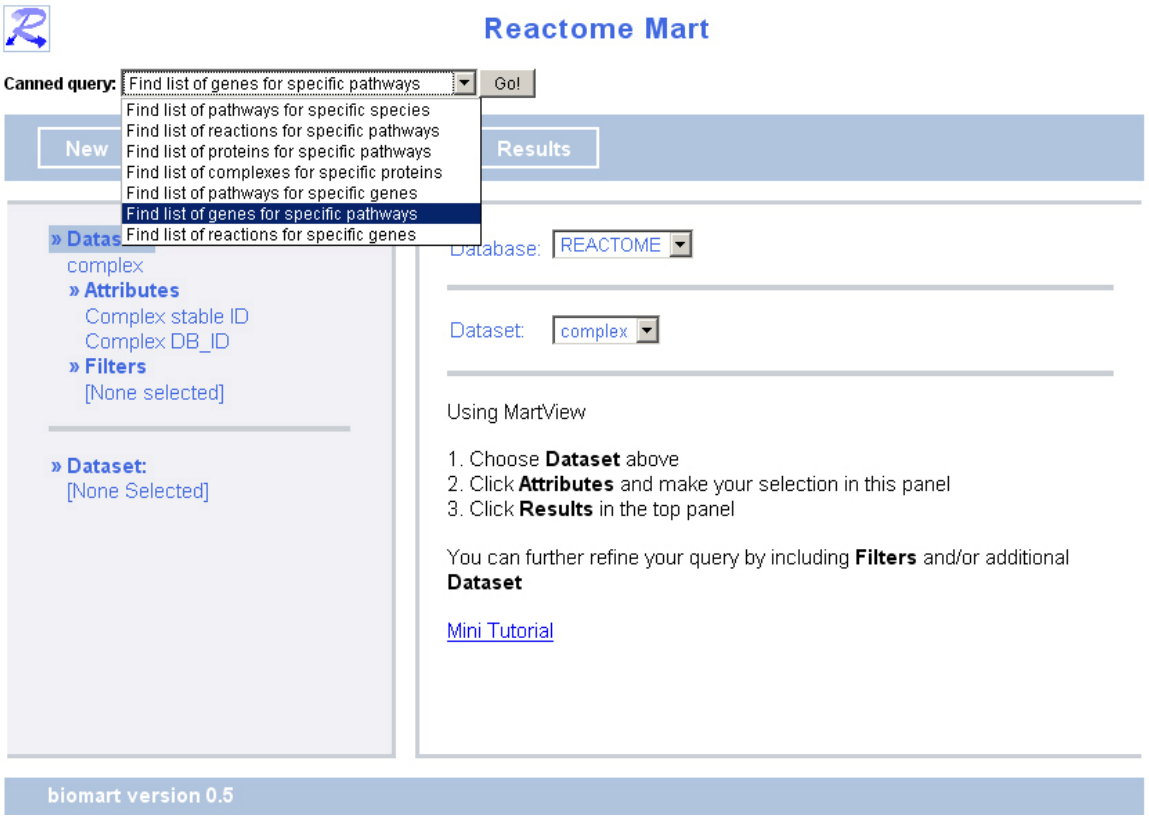

Figure 6: The Reactome Mart interface showing details of 'canned' queries available through a pull-down menu.

3 Results & Discussion

A major theme emerging from biomedical research in recent years is the multifactorial origin of many diseases (e.g. [20, 21]). This feature is thought to reflect the concerted evolution of a number of genes responsible for our survival on the one hand, and rapidly changing environmental pressures on the other. As the effort to establish the genetic basis of disease intensifies, single genes and their products are under close scrutiny to determine both biological role and their individual contribution to pathology and morbidity. The challenge today is to integrate this accumulated knowledge to provide the 'bigger picture' – a global functional context in which every human gene has a well-defined place.

The successful elucidation of major disease processes depends on making full use of knowledge acquired so far in the discovery of novel associations between the sequence and functional properties of key human molecules. However, the selective recovery and interpretation of information from the literature is largely inaccessible to computational

mining methods. Although much of biological knowledge is carefully written up and recorded, it is also dispersed over a number of literature sources in disparate formats, emphasis, styles and indeed levels of quality. Expertise is therefore required to reclaim knowledge that is credible, well established and reliable. The collaboration between field biologists and the Reactome editorial style of curation guarantees more objectivity to this process and ensures consistent standards throughout the model. In Reactome, the unit of knowledge sought for inclusion is that molecular interaction or modification that has a definite and manifest biological purpose.

The key integrative thrust of the Reactome effort is focused on establishing a deep and robust connectivity between established external 'dictionaries'. The integration of complex relationships between biological dictionary elements (in UniProt and Chebi, for instance), ontologies (e.g. GO), as well as the assimilation of third-party maps (e.g. Ensembl cross-references, OrthoMCL matches), puts available resources to optimal use while providing the requisite accessibility to the scientific community. The Mart tool enables users to generate flexible custom integrations, and the SBML and BioPax export tools facilitate the integration of Reactome resources into federated community databases as these develop.

Reactome is undertaking a systematic collaborative review of human molecular processes to create a central reference dataset that is both very readable and amenable to large-scale analysis. Given the wide connectivity to other data sets, a number of approaches using pathway knowledge to understand the effect of variation in sequence and expression are now feasible. For example, recent work has cross-referenced Reactome data to investigate the pathway distribution of somatic mutations involving protein kinase genes from diverse human cancers [22].

Collaborative software development also provides ample scope and means to interpret this molecular network in a genomic context. Fig. 5 illustrates the combined use of EnsEMBL and Reactome Perl APIs to relate Reaction annotation to the chromosomal origin of the gene products involved. Human chromosome 6 appears to have received the most coverage. This is partly explained by the presence of a considerable number of well-characterized immune-related genes about which biological expertise is more accessible. Based on the Reactome annotation so far, proteins pairs arising from human chromosome 22 appear more likely to occur in the same Complex or Reaction, although this does not necessarily imply physical interaction. The presence of DNA and RNA processing genes on this chromosome (RNA polymerases in particular), annotated over a number of Pathways including those pertaining to the Cell Cycle, Viral Infection, DNA Replication, Transcription and Translation, is largely responsible for this tendency.

All data generated by Reactome, as well as the tools constructed to achieve its goals, are openly accessible. The salient aspects of this project's experience and optimized procedures are also available as tutorials and other forms of documentation via the 'User Guide' and download area, as well as on the http://wiki.reactome.org site. An active mailing list is supported on help@reactome.org for user enquiries.

Such wealth of accessible information ingrained in the human Reactome events is particularly opportune as graph-based methods are increasingly relying on verified interaction networks to predict protein function [23] and consequent disease phenotypes [24]. To date, Reactome has released detailed interconnected depictions of molecular pathways involving about 10% of UniProt's human protein complement. The challenge ahead is to extend good quality coverage while providing effective integration with tissue expression and kinetic molecular models [25], thus opening more avenues to investigate tissue-specific pathologies and targeted drug design.

4 Acknowledgements

We gratefully acknowledge all authoring and reviewing biologists who provide their expertise and resources for the annotation and quality assessment of the Reactome model, and Geeta Joshi-Tope, Beth Nickerson, and Marcela Tello-Ruiz for their work on the initial stages of this project. Reactome is supported by a grant from the US National Institutes of Health (P41 HG003751), a grant from the European Union FP6 (LSHG-CT-2003-503269), and a subcontract from the EBI Industry Programme.

5 References

1. Commoner, B., *The Closing Circle*. 1971: Knopf, New York.
2. Lang, F., *Acid-Base Metabolism*, in *Comprehensive Human Physiology*, R. Greger, Windhorst, U., Editor. 1996, Springer: Berlin. p. 1571-1584.
3. Klamt, S., et al., *A methodology for the structural and functional analysis of signaling and regulatory networks.* BMC Bioinformatics, 2006. **7**: p. 56.
4. Vastrik, I., et al., *Reactome: a knowledgebase of biological pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.
5. Durinck, S., et al., *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.* Bioinformatics, 2005. **21**(16): p. 3439-40.
6. Kahlem, P. and E. Birney, *Dry work in a wet world: computation in systems biology.* Mol Syst Biol, 2006. **2**: p. 40.
7. Fenstermacher, D., et al., *The Cancer Biomedical Informatics Grid (caBIG^{TM}).* Conf Proc IEEE Eng Med Biol Soc, 2005. **1**: p. 743-6.
8. Kakazu, K.K., L.W. Cheung, and W. Lynne, *The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research.* Hawaii Med J, 2004. **63**(9): p. 273-5.
9. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D187-91.
10. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
11. de Matos, P., et al., *ChEBI - Chemical Entities of Biological Interest.* NAR Molecular Biology Database Collection, 2007(646).
12. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2007. **35**(Database issue): p. D5-12.
13. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
14. Crosby, M.A., et al., *FlyBase: genomes by the dozen.* Nucleic Acids Res, 2007. **35**(Database issue): p. D486-91.
15. Bevan, M. and S. Walsh, *The Arabidopsis genome: a foundation for plant research.* Genome Res, 2005. **15**(12): p. 1632-42.
16. Birney, E., et al., *Ensembl 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D556-61.
17. IHC, *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
18. Kasprzyk, A., et al., *EnsMart: a generic system for fast and flexible access to biological data.* Genome Res, 2004. **14**(1): p. 160-9.

19. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

20. Barnetche, T., P.A. Gourraud, and A. Cambon-Thomsen, *Strategies in analysis of the genetic component of multifactorial diseases; biostatistical aspects.* Transpl Immunol, 2005. **14**(3-4): p. 255-66.

21. Talmud, P.J., *How to identify gene-environment interactions in a multifactorial disease: CHD as an example.* Proc Nutr Soc, 2004. **63**(1): p. 5-10.

22. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.

23. Sharan, R., I. Ulitsky, and R. Shamir, *Network-based prediction of protein function.* Mol Syst Biol, 2007. **3**: p. 88.

24. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.

25. Le Novere, N., et al., *BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.* Nucleic Acids Res, 2006. **34**(Database issue): p. D689-91.