Pathway Tools Omics Viewer II

PathwayExplorer

While generation of high-throughput expression data is becoming routine, the fast, easy, and systematic presentation and analysis of these data in a biological context is still an obstacle. To address this need, we have developed PathwayExplorer, which maps expression profiles of genes or proteins simultaneously onto major, currently available regulatory, metabolic and cellular pathways from KEGG, BioCarta and GenMAPP. PathwayExplorer is a platform-independent web server application with an optional standalone Java application using a SOAP (simple object access protocol) interface. Mapped pathways are ranked for the easy selection of the pathway of interest, displaying all available genes of this pathway with their expression profiles in a selectable and intuitive color code. Pathway maps produced can be downloaded as PNG, JPG or as high-resolution vector graphics SVG. The web service is freely available at https://pathwayexplorer.genome.tugraz.at; the standalone client can be downloaded at http://genome.tugraz.at.

We used state-of-the-art Java technologies to develop PathwayExplorer. It is an entirely Java-based application client using a three-tiered-architecture that ensures a clean separation between the presentation front-end, business and database back-end layer. The business layer, a Java application client conforming to the J2EE specification, performs the calculations and search functions and can be accessed by the presentation layer in two ways: (i) an application client using SOAP (simple object access protocol); and (ii) a web browser-based application client running on a web server using JSP technology. The database layer called PathwayDB is based on an Oracle DBMS (data base management system), which is also portable to freely available MySQL or PostgreSQL DBMS. Frequently changing information is kept in flat files, which are obtained and constantly updated from NCBI (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/) and KEGG (ftp://ftp.genome.ad.jp/pub/kegg/ligand/).

PathwayDB minimizes the ambiguity among its gene identifiers. The fact that almost all identifiers are relationally and hierarchically linked allows it to specify the gene element nodes with only one kind of identifier, which constitutes the top of the hierarchical identifier tree. All the gene identifiers that lie below the root identifier can be linked to it later by using external

data sources. We have successfully integrated KEGG, BioCarta and GenMAPP pathways into PathwayDB by using only the minimum information necessary. This comprises information from parsed SMBL (Systems Biology Markup Language) files obtained from KEGG, which were converted into the PathwayExplorer application client format. This was performed because of the lack of the SMBL format for encoding feasible graphical visualization, which is essential for the

graphical evaluation of mapped pathways. The EC (Enzyme Commission) defines the root identifier, which can hold several gene identifiers from all available organisms, i.e. the LocusLinks (they can contain again several gene identifiers, such as RefSeq or UniGene IDs) or the official gene identifiers for other organisms. To integrate BioCarta and GenMAPP into PathwayDB, the PathwayExplorer application client was once again used for automatically parsing the HTML pages holding the necessary pathway information. Since both of these pathway resources use many different gene identifiers, LocusLink was again used as root identifier. The LocusLinks are linked with the user-defined gene identifier groups (UniGene, GeneOntology, GenBank and/or RefSeq), which are used then to align the mapped gene IDs.

Accessibility

PathwayExplorer is a web-based service constantly available at https://pathwayexplorer.genome.tugraz.at, with a public, login-free data repository for uploading data sets.

The PathwayExplorer standalone client application can perform the same mapping operations on an independent, local-platform computer system. In this case, instead of uploading the expression data to the web server, the pathway information from PathwayDB is downloaded to the user's local computer system. The standalone client connects to PathwayDB through a SOAP interface. The standalone client is available at the PathwayExplorer homepage or at http://genome.tugraz.at/Software/PathwayExplorer/Setup.html.

Input

As input, PathwayExplorer receives a common tab-delimited text file containing expression profiles with the gene identifier as first column, the gene name as an optional second column and any experiment or time point data as further columns. Possible gene identifiers for organisms using LocusLinks are GenBank accession numbers, RefSeq IDs, UniGene IDs and Gene Ontology IDs. For all other organisms, systematic gene identifiers are possible. The RefSeq IDs are used as the initial default gene identifier group, and this can be changed later. The uploaded gene-expression data sets can be stored in either a public or a login-requiring repository where they can be modified or deleted again.

Calculations and visualization

An example for mapping a public data set from a yeast sporulation study (18) is given in Figure 1. In order to map data sets onto pathways, the user is requested to select the organism and the data set to be mapped. The loaded data set remains in the background as long as it has not been closed, and subsequently every pathway which becomes opened is then automatically mapped with this data set. To restrict the uploaded data set to certain criteria before mapping (e.g. to use only expression profiles of differentially expressed gene or proteins), filter options can be applied: (i) to filter out expression profiles with too many missing data points; (ii) to filter out weakly expressed profiles (based on a certain standard deviation threshold); and (iii) to filter out genes whose expression values do not meet a certain threshold.

PathwayExplorer example: a screenshot of a pathway mapped with expression data. (i) The toolbar frame (the row including the organism field) offers various setup and visualization options. (ii) Hierarchical tree frame (on the left) enables browsing through .After filtering the data set, PathwayExplorer provides two mapping options: (i) mapping the data set to a single pathway by choosing one in the hierarchical tree or (ii) Mapping the data set onto all available pathways at once. Option (ii) generates a list (Figure 3), which ranks all mapped pathways by their number of mapped genes and allows for sorting the list based on different criteria, such as (a) pathway name; (b) unique gene identifiers available in each pathway; (c) the number of gene identifiers which has passed the filter criteria and was mapped to the pathway; (d) the number of genes which would have been mapped to the pathway if they had passed the filter criteria; and (e) the right-tailed P-value of a Fisher's exact test (f) the false discovery rate (FDR) corrected Q-value.

(a) Shows the overall statistic of the filtered unique identifiers of the expression data set mapped on all pathways. (**b**) The ranking list of all mapped pathways is also displayed and can be sorted by different criteria, allowing easy navigation through **.**

With a right-tailed Fisher's exact test, we test whether the proportion of mapped genes within the set of differentially expressed genes is significantly larger than the proportion of mapped genes that are not differentially expressed. We use a Fisher's exact test because the number of counts might be smaller than five for any of the fields in the contingency table. Multiple hypotheses correction is needed to control the number of false positives, since many hypotheses are tested simultaneously.

The graphical visualization of the displayed pathway image can be changed to the SVG view using the freely available SVG Viewer plug-in from Adobe (http://www.adobe.com). This enables on-line zooming of the pathway graphic.

Output

PathwayExplorer provides graphical and textual output. It generates a gene cluster for each mapped pathway, which can be downloaded in the same tab-delimited text format as the uploaded data set. Each mapped expression profile can be displayed by selecting the corresponding box on the pathway image. The generated ranking list for all mapped pathways can also be downloaded as tab-delimited text file and can be used for statistical analyses.

By selecting one row (if there are more than one) of a mapped gene box the corresponding gene and expression profile information will be displayed. To obtain additional information links to GenBank, Entrez and OMIM are provided. Only one mapped **.**

A special feature is the PDF generator, which can be applied for every single pathway, irrespective of whether the genes were mapped to the pathway or not. This generator creates a PDF document of the currently loaded pathway, which can be downloaded or directly displayed in the user's web browser. If genes were mapped to the pathway, the PDF generator additionally adds the expression profiles to the document. In the case of human expression data sets, additional information about each gene is directly extracted from the OMIM (Online Mendelian Inheritance in Man) database. This feature offers the user a special opportunity to get a quick and comprehensive overview of the current pathway plus detailed information about each mapped gene. The graphical output of PathwayExplorer can be directly downloaded in PNG or SVG graphic format.

Pathway Browser

A brief overview of our data model and orthology–based electronic inference of non–human events will provide a basis for describing the potential uses for the Reactome. The user documentation also describes some of the technical aspects of the Reactome website and data. Instructions are also available on how to link to and reference Reactome.User Guide is an overview of the Reactome database of biological pathways and processes and its web site. This is not a comprehensive guide, but should provide you with enough information to browse the database and use its principal tools for data analysis.

Developer Guide provides access to all the documentation relating to the Reactome APIs, Analysis and Visualization web services.
Data Model in Reactome uses a frame-based knowledge representation. The data model consists of classes (frames) that describe the different concepts (e.g., reaction, molecule). These classes are hierarchically arranged into classes and parental superclasses. Orthology Prediction includes information about how computationally inferred pathways and reactions are created for 20 non–

human species, including Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, and Escherichia coli. These species represent more than 4,000 million years of evolution and span the main branches of life. Object/Relational Mapping explains the tables in the Reactome relational database.Wiki is part of our online documentation and curation resources. Linking to Reactome can be achieved by creating URLs containing the name of and an identifier from an "external" database.
Life on the cellular level is a network of molecular interactions. Molecules are synthesized and degraded, undergo a bewildering array of temporary and permanent modifications, are transported from one location to another, and form complexes with other molecules. Reactome represents all of this complexity as reactions in which input physical entities are converted to output entities. These reactions can occur spontaneously or be facilitated by physical entities acting as catalysts, and their progress can be modulated by regulatory effects of other physical entities. Reactions are linked together by shared physical entities: a product from one reaction may be a substrate in another reaction and may catalyze yet a third. It is often convenient, if sometimes arbitrary, to group such sets of interlinked reactions into pathways.

The functions of macromolecular entities such as proteins are often determined not only by their primary sequences, but by chemical modifications they have undergone. In Reactome, unmodified and modified forms of a protein are distinct physical entities and the modification process is treated as an explicit reaction. A macromolecule's function may depend on whether the molecule is free or complexed with specific other molecules. Reactome treats complexes as physical entities distinct from their components, and the multimerization events that build up complexes are modeled explicitly as reactions. Cellular compartments play a key role in biological processes. The segregation of molecules into different compartments often regulates the reactions in which those entities can participate, or can be responsible for driving a reaction forward. In Reactome, a molecule in one compartment is distinct from that molecule in another compartment. Thus, extracellular and cytosolic glucose are different Reactome entities and, e.g., the movement of glucose across the plasma membrane is a reaction that converts the extracellular glucose entity into the cytosolic one.

Many biochemical entities and processes appear redundant: there are two or more chemically distinct entities that can act more or less interchangeably. It is often useful to treat functionally equivalent protein isoforms, splice variants, and paralogues as a single entity, implying that any individual entity from the given set could fulfill the same role in a given situation. The Reactome data model allows this type of generalization, but does so explicitly in a way that allows us to trace specific functions back to the individual molecules covered by the generalization. The goal of the Reactome knowledgebase is to represent human biological processes, but many of these processes have not been directly studied in humans. Rather, a human event has been inferred from experiments on material from a model organism. In such cases, the model organism reaction is annotated in Reactome, the inferred human reaction is annotated as a separate event, and the inferential link between the two reactions is explicitly noted. Reactome uses a frame-based knowledge representation. The data model consists of classes (frames) that describe the different concepts (e.g., reaction, simple entity). Knowledge is captured as instances of these classes (e.g., "glucose transport across the plasma membrane", "cytosolic ATP"). Classes have

attributes (slots) which hold properties of the instances (e.g., the identities of the molecules that participate as inputs and outputs in a reaction).

Key data classes

PhysicalEntity

PhysicalEntities include individual molecules, multi-molecular complexes, and sets of molecules or complexes grouped together on the basis of shared characteristics. Molecules are further classified as genome encoded (DNA, RNA, and proteins) or not (all others). Attributes of a PhysicalEntity instance capture the chemical structure of an entity, including any covalent modifications in the case of a macromolecule, and its subcellular localization.

PhysicalEntity instances that represent, e.g., the same chemical in different compartments, or different post-translationally modified forms of a single protein, share numerous invariant features such as names, molecular structure and links to external databases like UniProt or ChEBI. To enable storage of this shared information in a single place, and to create an explicit link among all the variant forms of what can also be seen as a single chemical entity, Reactome creates instances of the separate ReferenceEntity class. A ReferenceEntity instance captures the invariant features of a molecule. A PhysicalEntity instance is then the combination of a ReferenceEntity attribute (e.g., Glycogen phosphorylase UniProt:P06737) and attributes giving specific conditional information (e.g., localization to the cytosol and phosphorylation on serine residue 14).
The PhysicalEntity class has subclasses to distinguish between different kinds of entity and to ensure data integrity while enabling different handling rules for different categories:

EntityWithAccessionedSequence – proteins and nucleic acids with known sequences.
GenomeEncodedEntity – a species-specific protein or nucleic acid whose sequence is unknown, such as an enzyme that has been characterized functionally but not yet purified and sequenced, e.g. cytosolic 15-HEDH enzyme
SimpleEntity – other fully characterized molecules, e.g. nucleoplasmic ATP or cytosolic glutathione
Complex – a complex of two or more PhysicalEntities, e.g. Trimerization of the FASL:FAS receptor complex
EntitySet – *a set of PhysicalEntities (molecules or complexes) which function interchangeably in a given situation, e.g.* Notch 3 heterodimer binds with a Notch ligand in the extracellular space. *This notation allows collective properties of multiple individual entities to be described explicitly.*
CatalystActivity

PhysicalEntities are paired with molecular functions taken from the Gene Ontology molecular function controlled vocabulary to describe instances of biological catalysis. An optional ActiveUnit attribute indicates the specific domain of a protein or subunit of a complex that mediates the catalysis. If a PhysicalEntity has multiple catalytic activities, a separate CatalystActivity is created for each. This strategy allows the association of specific activities

with specific variant forms of a protein or complex, and also enables easy retrieval of all activities of a protein, or all proteins capable of mediating a specific molecular function.

Event

Events – the conversion of input entities to output entities in one or more steps – are the building blocks used in Reactome to represent all biological processes. Two subclasses of Event are recognized, ReactionlikeEventand Pathway. A ReactionlikeEvent is an event that converts inputs into outputs. A Pathway is any grouping of related Events. An event may be a member of more than one Pathway. The Reaction like Event class is further divided into Reaction, BlackBoxEvent, Polymerisation and Depolymerisation. The Reaction class holds bona fide reactions with balanced inputs and outputs. The BlackBoxEvent class is used for 'unbalanced' reactions like protein synthesis or degradation, as well as 'shortcut' reactions for more complex processes that essentially convert inputs into outputs, e.g. the series of cyclical reactions involved in fatty acid biosynthesis. The De-/Polymerisation classes can hold reactions that describe the mechanics of a de-/polymerisation reaction, which is inherently 'unbalanced' due to the nature of a Polymer (that remains the 'same' entity even after adding or subtracting a unit).

Computational inferred events

We use the set of manually curated human reactions to electronically infer reactions in twenty evolutionarily divergent eukaryotic species for which high-quality whole-genome sequence data are available, and hence a comprehensive and high-quality set of protein predictions exists. These species include the laboratory mouse and rat, the nematode *C. elegans*, budding and fission yeasts, and two plants. The estimated success rates of our orthology inference strategy can be stated as 'the percentage of eligible reactions, defined in step 2 below, in the current human data set for which an event can be inferred in the model organism. By this measure, success rates range from 89.1% for the laboratory mouse to 6.7% for the protozoan *P. falciparum*.

Electronic inference proceeds in four steps.

1) Protein homology data were obtained from Ensembl Compara. Briefly, this method is based on the construction of gene trees, using the longest protein translation for every Ensembl gene, for all species included in the Compara database. Homologues are deduced from these trees. The method is described in more detail in EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates. Vilella et al., Genome Research, 2008. For the purpose of inferring homologous events in Reactome, we use the Core Compara database, containing vertebrate species, and the Pan Compara database, containing a wider spread of species, available from Ensembl Genomes.
2) All human reactions in the Reactome knowledgebase involving one or more proteins are eligible for electronic inference, with two exceptions. Reactions that were themselves inferred based on data from the model organism, and reactions involving species in addition to human (e.g., HIV infection of human cells) are excluded from electronic inference. Eligible reactions

are checked to determine whether each involved protein has at least one homologous protein (HP) in the model organism. If a human reaction involves a complex, at least 75% of the accessioned protein components of the human complex must have HPs in the model organism.

3) For each reaction that meets these criteria, an equivalent reaction is created for the model organism by replacing each human protein with its model organism HP. If a human protein corresponds to more than one model organism HP, a DefinedSet called 'Homologues of …' is created, with the model organism HPs as members. For human proteins that lack a model organism HP but that are included in complexes inferred due to the 75% threshold rule, placeholder model organism entities (called 'Ghost homologue of…') are created.
4) If this analysis generates reactions in the model organism corresponding to any of the steps of a human pathway, then the pathway event is also inferred for the model organism.

KDB Explorer

For powerful functional analyses using KEGG data

- Perform comparative and differential genomic analyses with multiple genomes & predict enzymatic activity. Explore, visualize and highlight the metabolic and non-metabolic pathways specific to your organism, Work interactively with KEGG data mirrored locally on your computer

Explore data quickly, extensively, and confidentiality, through KEGG ortholog (KO) relations as well as through connections calculated specifically by Genostar's unique IOGMA® technology.

Explore, query, and visualize relationships among genes, proteins, and metabolic and non-metabolic pathways:

- Import and mirror the KEGG data on the organisms of interest to your research
- Import other genome data in EMBL format
- Calculate homology relationships between the coding sequences of multiple organisms
- Compare genomes and identify genes specific to your strain or species
- Predict enzymatic activity and explore the metabolic pathways specific to your organism
- Use KEGG maps to visualize pathways
- Identify and display the reactions catalyzed by the product of a group of genes, for example, specific or co-regulated genes as known from expression experiments, and visualize the metabolic pathways in which these reactions are located

Browse and explore data connected through KEGG orthologs (KO) to investigate non-enzymatic functions

A set of add-on tools and functions are available for KDB Explorer, so you can choose to add in the analysis methods you need for your research. Search, navigate and highlight data with KDB Explorer.Consistent viewers and editors provide a clear view of all data and their relationships

Reference

The Pathway Tools cellular overview diagram and Omics Viewer by Suzanne M. Paley, Peter D. Karp

Tools for the functional interpretation of metabolomic experiments by Monica Chagoyen, Florencio Pazos -

Briefings in Bioinformatics, Volume 14, Issue 6, November 2013, Pages 737–744