BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 21 2008, pages 2526–2533
doi:10.1093/bioinformatics/btn459

Databases and ontologies

PathCase: pathways database system

Brendan Elliott, Mustafa Kirac*, Ali Cakmak, Gokhan Yavas, Stephen Mayes, En Cheng, Yuan Wang, Chirag Gupta, Gultekin Ozsoyoglu and Zehra Meral Ozsoyoglu

Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106, USA

Received on March 6, 2008; revised on July 16, 2008; accepted on August 21, 2008

Advance Access publication August 26, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: As the blueprints of cellular actions, biological pathways characterize the roles of genomic entities in various cellular mechanisms, and as such, their availability, manipulation and queriability over the web is important to facilitate ongoing biological research.

Results: In this article, we present the new features of PathCase, a system to store, query, visualize and analyze metabolic pathways at different levels of genetic, molecular, biochemical and organismal detail. The new features include: (i) a web-based system with a new architecture, containing a server-side and a client-side, and promoting scalability, and flexible and easy adaptation of different pathway databases, (ii) an interactive client-side visualization tool for metabolic pathways, with powerful visualization capabilities, and with integrated gene and organism viewers, (iii) two distinct querying capabilities: an advanced querying interface for computer savvy users, and built-in queries for ease of use, that can be issued directly from pathway visualizations and (iv) a pathway functionality analysis tool. PathCase is now available for three different datasets, namely, KEGG pathways data, sample pathways from the literature and BioCyc pathways for humans.

Availability: Available online at http://nashua.case.edu/pathways

Contact: pathcase@case.edu

1 INTRODUCTION

Biochemical pathways are used for the presentation and modeling of genomic data, studying genome sequences in a particular biological context, and for functional studies that attribute functions to genes. Each reaction in a metabolic pathway is a biochemical step from specific substrates to specific products that are chemically modified substrates. Each step may also use various molecules as cofactors, activators, inhibitors and regulators, and usually involves at least one genetically unique gene product that catalyzes the reaction step. Pathways, in general, illustrate the functional relationships between molecules, which include, for example, the identity of the substrate(s), product(s), cofactors, activators, inhibitors, enzymes or other processing molecules, RNA and protein expression patterns, reaction kinetics, and associated phenotypic variation and diseases. Ultimately, many other kinds of knowledge can be incorporated. Such information forms a rich research resource that integrates genomic and biological information which can be managed, analyzed, queried and displayed in dynamic ways at various levels of biological and genetic detail to provide insight into diverse biological processes in health and disease.

In this article, we present the latest features of PathCase, an integrated software tool designed to

(1) Store pathways data in an extensible and flexible manner.
(2) Visualize and query metabolic pathways at different levels of genetic, molecular, biochemical and organismal detail.
(3) Perform pathways-related analysis. Since the last publication on PathCase (Krishnamurthy *et al*., 2003), two fundamental developments have taken place. First, at the interface level, a web version of PathCase has been developed to make PathCase available to a larger number of users in a convenient way and without requiring extensive software installation. The PathCase web site (http://nashua.case.edu/pathways) provides an online interface for each PathCase database, and a toolset including an interactive pathway visualization tool. Second, PathCase is redesigned to
(4) Easily adapt data from other open-data biochemical pathway data sources in general and metabolic pathways in particular.
(5) Provide extensibility, namely, extending pathways data in PathCase with other biological data such as systems biology models, or metabolomics data.

In terms of item (4), currently, PathCase runs on three different datasets: PathCase with (i) KEGG pathways (Kanehisa *et al*., 2006), (ii) sample pathways from the literature (Michal, 1999) and (iii) BioCyc pathways for humans. We have also tested and verified the suitability of Reactome metabolic pathways for PathCase. In terms of item (5), we are extending PathCase with systems biology models, and metabolomics analysis.

PathCase is continually evolving. At the present time, new and distinctive features of PathCase are as follows.

Powerful visualization capabilities: PathCase has an interactive and dynamic (i.e. generated at query execution time) client-side pathway visualization tool with (i) an integrated pathway viewer, gene viewer (to display locations of a set of genes that play a role in a pathway), organism viewer (allowing users to choose from organism groups) and ontology viewer, and (ii) a variety of automated and flexible capabilities for layout generation, manual layout editing and layout saving.

*To whom correspondence should be addressed.

Metabolic network querying capabilities: PathCase has two querying features. The first one, advanced querying interface (AQI), is a graphical query language (Mayes, 2007), and allows for custom queries for (i) pathways, processes, molecular entities and organisms, (ii) metabolic network neighborhoods, or pathway network neighborhoods and (iii) metabolic network paths or pathway network paths. The second querying capability is via powerful built-in queries with both tabular and graphical output presentation. Built-in queries can be issued in two ways, namely, from visualized objects (i.e. molecules, processes, pathways) by right-clicking on an object and selecting a relevant built-in query, or via a separate built-in query pane.
Pathway functionality analysis: The pathway annotation tool pathway annotation (PW-ANN) models each pathway as a network of Gene Ontology (GO)-based enzyme functions (Cakmak *et al*., 2007), and provides automated statistical enrichment analysis and visualization.
Scalability: PathCase software architecture includes high performance server-side software, web services and powerful client-side (visualization and querying) software. This pushes some of the computational load from the server-side to the client-side, and, hence scales to concurrent access by many users with an ease.
Pathway and process export and import capabilities: PathCase can export its pathways and processes as BioPAX- and image-formatted documents. In addition, PathCase can accept BioPAX-formatted pathways, and provides their visualizations.

This article is organized as follows. Section 2 summarizes the PathCase system architecture, the data model and the database. In Section 3, we briefly present the PathCase graph viewer. Section 4 is a summary of the querying capabilities of PathCase. In Section 5, we summarize the pathways annotation analysis tool. Section 6 briefly compares PathCase with the existing related systems.

2 METHODS

The main goal of PathCase is to provide life scientists with an integrated environment to study pathways, regardless of the source producing the corresponding data. More specifically, rather than becoming an ultimate and authoritative data source for pathways, PathCase's vision is to become a powerful, data source independent, one-stop computational environment encapsulating an extensive set of tools for systems-level research on cellular actions. To this end, PathCase emphasizes six distinct dimensions as the focus of the overall system: (i) storing, (ii) analyzing, (iii) visualizing, (iv) querying, (v) modeling and (vi) sharing pathways data.

- *Storage*: PathCase runs on a relational database, and the web content is dynamically created using the compiled data from the database. The main data objects are pathways, processes and molecular entities. To increase the efficiency of queries and web page content retrieval, the database employs a large number of indices defined on groups of database attributes. Some fields (e.g. links between pathways) are precomputed to improve the response time.
- *Analysis*: PathCase provides tools for the analysis of pathways at various levels of granularity in different dimensions. PW-ANN is one such tool. Another tool takes any user-provided set of genes, and retrieves a set of pathways that are tightly associated with the input genes. A third tool allows pathway analysis in terms of the locations of genes encoding the enzymes of its processes to see how closely they are located on a given genome.
- *Visualization*: PathCase has a powerful visualization tool with advanced, flexible, automated layout generation engine. Pathways can be visualized at different detail levels, such as the expanded form, which displays all the components of a pathway (from reactions to activators) and the collapsed form, which is used to visualize the connections between pathways in a compact way. In addition, PathCase also visualizes a large number of query results. All the visualizations are created dynamically (at the client side), and can be edited by users.
- *Querying and browsing*: PathCase puts an extensive emphasis on querying and browsing the underlying pathways data in an effective way. Users can either use powerful built-in queries to pose various queries from path finding between two molecules to locating neighbors of a pathway in a metabolic network. When predefined built-in queries are not sufficient to perform to the user's task at hand, the AQI allows users to build their own queries in a flexible and user-friendly way. Alternatively, for users who are browsing and do not yet know what they are looking for, PathCase provides an elegant browser for each biological entity type, e.g. organisms, proteins, reactions.
- *Modeling*: In PathCase, pathways are represented as hypergraphs (Berge, 1973) where nodes are molecular entities participating in a reaction as substrates or products, and edges are enzymes which catalyze reactions. Moreover, PathCase also models pathways as a graph of GO functional annotations, called *pathway functionality templates*, where each node corresponds to a GO annotation of an enzyme catalyzing a process. For a given pathway, different functionality templates can be created by using the generalization/specialization hierarchy of the GO.
- *Sharing*: PathCase provides three distinct mechanisms to share its data and visualizations with the scientific community: (i) all pathways can be exported to BioPAX-formatted documents (BioPAX, 2004), (ii) visualizations can be saved in an image format (JPEG) and (iii) PathCase web service methods can be directly accessed in order to query the PathCase database and obtain the results in XML format. PathCase can also import BioPAX-formatted pathway documents, and visualizes them.

2.1 System architecture

As a web-based system, PathCase is highly accessible and designed to be fully compatible with all major web browsers. The PathCase system architecture, shown in Figure 1, has four distinct layers. Starting with the server side, the first layer consists of the databases, which contain the actual pathways information and allow for efficient querying. The database server is Microsoft SQL Server 2005. The next layer is the *data object library*, which provides a programmer-friendly interface to the system content stored in the databases that allows for data to be accessed, created and updated. This library is written in Microsoft C# 2005. The next layer, the *web server*, includes the PathCase web site and the web service both of which are written in C# and ASP.NET, and generate standard HTML pages and XML data. This allows the site to be accessed by users from a standard web browser on any operating system. The final layer, on the client side, includes the components of PathCase that run on the user's web browser. This includes the basic HTML that renders the main site interface to the user, the JavaScript with AJAX that makes the site highly responsive to the user and enables the web-based AQI, and, finally, the graph viewer java applet used for interactive pathway graph visualization. Java is chosen for powerful and highly responsive dynamic graph drawings. The graph viewer applet makes use of the web service component in order to request additional data as needed, and enhance the graph visualization without requiring the user to reload the web page. All graph manipulations such as zooming in and out, panning and application of different layouts are carried out on the client side with no server side requests, which makes PathCase highly scalable.

2.2 Data model

In the PathCase data model, we represent a metabolic pathway (Michal, 1999) in the form of a graph where nodes represent molecules, and edges

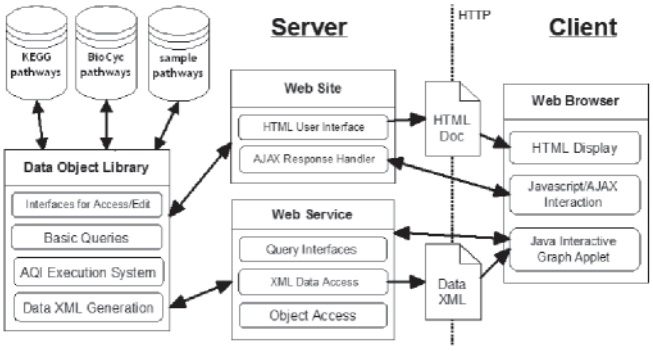

Fig. 1. PathCase system architecture.

represent reactions (or processes) connecting the molecules that take part in the reaction. As reactions may involve more than one substrate and more than one product, the reaction is actually a hyper-edge, and a pathway is a hyper-graph (Berge, 1973). We use the term *process* to denote reactions, which can include a catalyzing protein (which also identifies the reaction and may have an enzyme classification number), co-factors, inhibitors and activators. Reactions may be one-way or reversible. We refer to molecular objects as *molecular entities* which include basic molecules, proteins, enzymes, genes and amino acids. A *pathway* can be viewed as a set of interconnected processes, while a process can be viewed as being made up of molecular entities. More generally, an entire pathways database can be viewed as a single large graph of interconnected reactions, in which certain subgraphs are identified as specific pathways. This way, the system can dynamically visualize, query and analyze any subsection of the larger pathways network, or even show how pathways themselves are interconnected at a higher level. Connected pathways are precalculated for efficient querying and visualization. Pathways are also organized into *pathway groups*, representing a set of pathways with related functionality.

The PathCase model also maintains the *organism* in which a process occurs. This allows the user to switch from viewing a general pathway graph to viewing its *organism-specific* version, in which the components that do not apply are dynamically grayed out. Organisms can be organized into a hierarchy of *organism groups*, allowing for easy visualization of the pathway for an entire kingdom, phylum, etc. Organisms can also be associated with *chromosomes,* and store the location of *genes* that encode proteins. This enables the user to quickly find gene and chromosome location for a process' catalyzing protein and vice versa.

2.3 State of the database

Between the KEGG dataset and the sample dataset, PathCase now allows for searching and interactive visualization of more than 150 pathways, 7000 reactions, 5000 proteins, 25 000 molecules and 1.6 million genes in more than 450 organisms. Since the last major release of the system in 2003, this represents a 300% increase in the number of pathways, 700% more reactions and 2800% more molecules. PathCase is continually updated as new data becomes available.

As mentioned before, PathCase currently utilizes three different datasets. The first dataset contains a set of sample pathways from the literature (Michal, 1999). The second set contains the KEGG metabolic pathways (Kanehisa *et al*., 2006), licensed from KEGG. KEGG provides their data for academic users in the text format on their FTP site. On a periodical basis, we download and parse those text documents, extract and transform the information into our format, and insert the data into our relational database. This keeps our database up-to-date as KEGG pathways data are modified.

PathCase data model is designed to accommodate all essential entities that constitute structural and semantic aspects of metabolic pathways, such as enzymes, reactions, substrates, cofactors and so on. Consistent with PathCase's vision to become an integrated computational environment rather than an authoritative data source, all research and development decisions in PathCase are made with a special emphasis to stay as generic as possible in terms of the assumptions on the underlying pathways data. Hence, with no data source-specific engagement in its original data model (Krishnamurthy *et al*., 2003) as well as in its presentation layer, PathCase can effortlessly run on data from a variety of different data sources.

For each of the three pathways databases (KEGG, BioCyc human pathways and our curated sample pathways), the presentation layer remains the same, with a home page listing all available datasets that PathCase currently operates on. A promising future work on this aspect of PathCase is to establish automated or semi-automated means to establish crosslinks between different datasets in PathCase with the goal of side-by-side comparisons of similar pathway objects from distinct data sources.

In order to provide a better visualization of complex pathways, we simplify the reactions in the KEGG pathways while we insert them into the PathCase database as follows. If a metabolite in a reaction does not play an important role in any of the pathways that contain the reaction, then we hide the metabolite on the graph. This simplification prunes some metabolites, and provides a simpler view of these complex graphs.

3 PATHCASE GRAPH VIEWER

An important component of the PathCase system is the GraphViewer pathway visualization tool, which is a Java applet embedded into the PathCase interface. GraphViewer retrieves raw pathway information (as an XML document from the web services) including genes and organism taxonomy of a pathway, and translates it into an easy-to-understand graph visualization detailing the interconnections between pathways, processes and molecular entities at the client side. This architecture utilizes the server efficiently, and provides a responsive user interface on the client side. GraphViewer consists of three main components; (i) pathway viewer, (ii) organism taxonomy viewer and (iii) gene viewer.

3.1 Pathway viewer

The pathway viewer is available when viewing pathway or process details, as well as when viewing the results of one of PathCase built-in queries. The tool provides a large number of predefined and customizable layouts, and also has the ability to choose (i) the best layout among them, or (ii) the 'curated layout' that is manually laid-out to produce the textbook layouts (Michal, 1999) that biologists have become familiar over the years (see PathCase help pages for details).

Based on the entity type, the pathway viewer provides a number of advanced queries on graph elements (Fig. 2a). For example, for a process p, users can query its details, gene information, other processes sharing activators and inhibitors with p in the current pathway, as well as those processes which are within a given number of steps away from p in the current pathway or in the metabolic network (Fig. 2b and c).

Some molecules like H_2O, H_2O_2, NH_3, O_2, are present in most reactions and, hence, occupy a large part of the visualization. Such molecules are now separated in a group, and users are given the freedom to show or hide such molecules in the graph. And, users can also hide and display regulator and cofactors in the graph.

Pathway visualization component provides a number of tools for navigating graphs. An editing tool allows users to reposition graph entities, a magnifier enlarges a specific region of the visualization while looking at it from a bird-eye-perspective, a panning tool

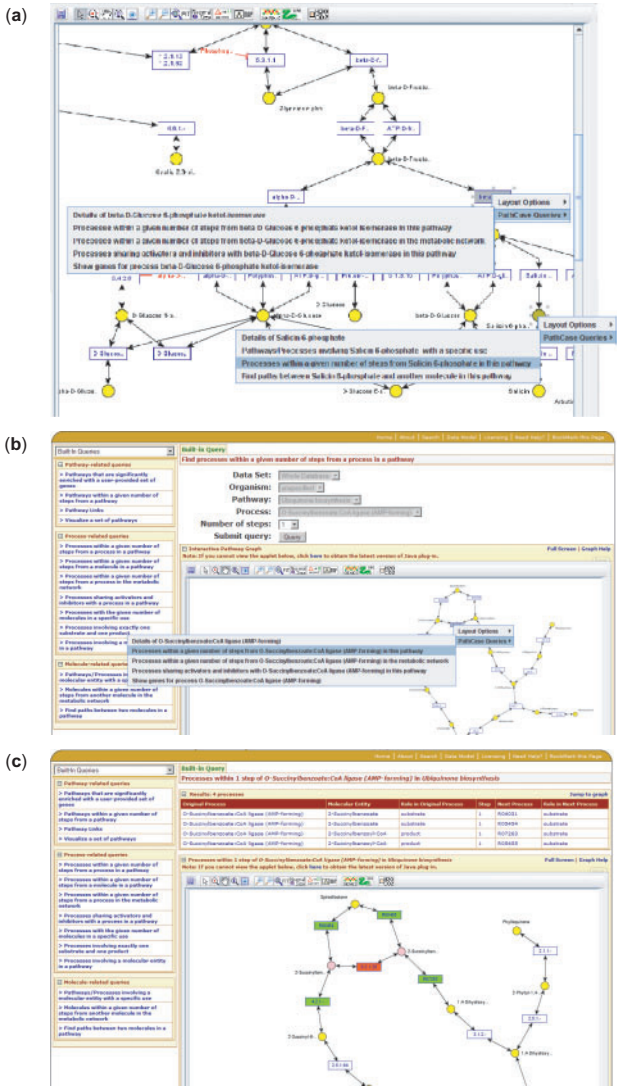

Fig. 2. (**a**) Molecule and process built-in queries superimposed together.(**b**) GraphViewer querying example. Right-click menu directs PathCase interface to a query page with automatically filled fields with respect to the entity being queried from the graph viewer. (**c**) Tabular and graph results of the query in (b).

relocates viewport of the visualization, and a number of zooming tools can be employed for zooming in and out of a specific region or a point. In addition, this component provides an additional panel called *minimap* that views a sketch of the graph at once, and allows user to quickly move to a particular position.

One important feature of the pathway viewer is its flexible and automated layout algorithms. We provide a large number of layout algorithms with detailed parameters, provided by yFiles (http://www.yworks.com/en/products_yfiles_about.htm) graph visualization toolkit. Since such algorithms are developed for general purpose graph viewing, we also provide a smaller set of layout options that are optimized for metabolic pathways. For instance, we relocate all regulator and cofactors to fixed positions around process entities, whereas automated layout algorithms would position them arbitrarily to reduce edge and node overlaps.

Once users of the system visualize the pathway or a query result of their interest, a menu item allows saving the visualization as a JPEG image for reuse in presentations, papers and lectures. In addition, PathCase system allows saving pathways in BioPAX-formatted files. Pathway visualization tool has the capability to load BioPAX files saved from the PathCase system, or posted on other pathway repositories, such as BioCyc (BioCyc, http://www.biocyc.org).

3.2 Organism taxonomy viewer

Organism (taxonomy) viewer is a simple yet powerful tool for listing in tree format the names of the organisms having the pathway (or a fragment of the pathway) displayed in the pathway visualization component. In default mode, i.e. when a pathway graph is not restricted to a subset of available organisms, entities in the visualization are displayed in their original colors. When organism viewer is employed to select/deselect a subset of organisms, the entities belonging only to the deselected organisms are grayed out in the graph, in order to visualize all elements of the pathway while contrasting selected and deselected organism groups (Fig. 3). Organism viewer is treated as the primary control point of organisms synchronized with the pathway visualization, and gene viewer. Queries posed from the visualization are parameterized for the organism selected from the organism viewer, and genomes for those organisms are listed in the gene viewer (Fig. 3b).

3.3 Gene viewer

The gene viewer allows users to view the genes that encode the enzymes of a given pathway. We visualize the genome of an organism as a set of chromosomes of lengths relative to the longest chromosome of that organism. PathCase users can highlight genes related to a process (i.e. genes that encodes enzymes catalyzing the corresponding reactions of the process) from a right-click query on a process graph entity. Also, processes related to genes can be highlighted by either selecting a gene from the gene viewer, or selecting all genes of a chromosome by selecting a chromosome. When a gene or chromosome is selected in the gene viewer, detailed information about the gene is shown at the bottom of the gene viewer (Fig. 4).

We precompute the positions of genes on the chromosomes. In order to specify the location of a gene on its chromosome (i.e. gene locus), various addressing schemes have been developed by geneticists. PathCase database currently accommodates three types of gene addresses: (i) molecular location, (ii) cytogenetic address and (iii) genetic linkage distance. The main distinction between different addressing schemes is in the means and methods employed by each scheme as well as the granularity (or the resolution) of the location on a chromosome that each type of addressing scheme can point to. By default, gene viewer uses the most precise address (i.e. molecular location), if available.

4 PATHCASE QUERYING

4.1 Built-in queries

Built-in queries are a set of predefined parameterized search interfaces designed for the frequently used queries. Currently, PathCase contains 14 built-in queries organized into three categories, namely, pathway-, process- and molecule-related queries. The provided queries include neighborhood queries (e.g. ‘Find processes

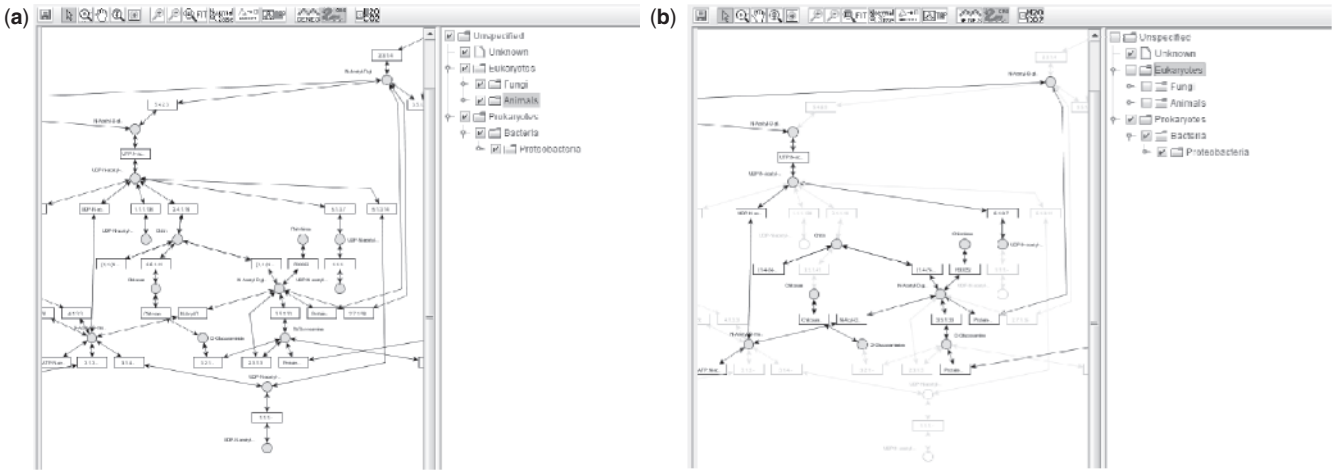

Fig. 3. (**a–b**) Choosing a set of organisms from the organism taxonomy (a) synchronizes pathway visualization by fading out the graph elements that belong only to the deselected organism (b).

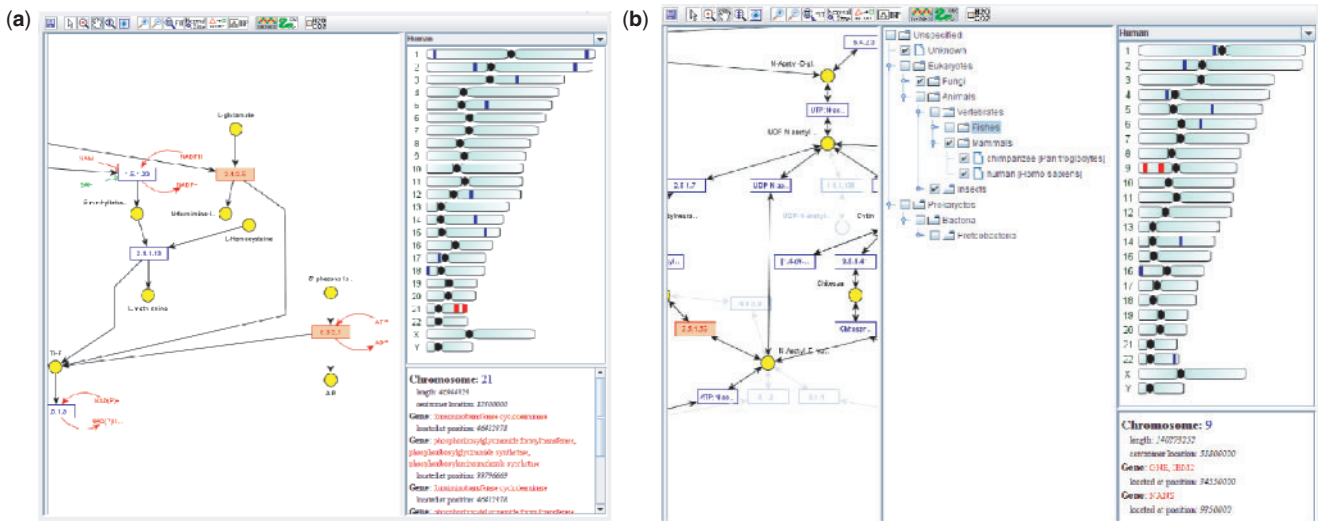

Fig. 4. (**a–b**) Gene viewer listing chromosomes of a selected organism (a and b). In (a), a chromosome is selected, and all processes catalyzed by an enzyme encoded by genes of the selected chromosome is highlighted. In addition, detailed information about the selected chromosome and genes on the selected chromosome are displayed in the information area at the bottom (a and b). In (b) collaboration between organism viewer, pathway visualization and gene viewer is illustrated. Organisms selected in the organism viewer gray out the processes belonging only to the deselected organisms, and transfers the list of selected organisms to the gene viewer. Gene viewer displays the genome of an organism selected in both the organism viewer and the gene viewer.

that are at most N steps from a given process'), path queries (e.g. 'Find the paths between molecule A and molecule B'), as well as queries asking for properties or components of a pathway, a process or a molecular entity. Queries can also be invoked from the Pathway Viewer. The query results are displayed both in the form of tabular outputs and graphical displays. Queries and the displayed graphs can be at multiple abstraction levels, e.g. process or molecular entity neighborhood versus pathway neighborhood. See PathCase web site's help section for more examples.

4.2 Advanced query interface

The AQI is a system that allows users of PathCase to dynamically construct and execute *ad hoc* queries on the database. The users are presented with a tree-like, hierarchical structure in which to formulate their queries, starting with a single root node and expanding hierarchically by adding other nodes, if desired (Fig. 5). JavaScript and AJAX technologies are used for a rich, dynamic user experience. For more details, see Elliott *et al*. (2008).

5 PATHWAY ANNOTATION ANALYSIS

The PW-ANN tool in PathCase provides a distinct functionality to perform pathway annotation analysis. Our approach is to model each pathway as a network of GO-based enzyme functions, which we call the (pathway) GO functionality template (PFT) (Cakmak *et al*., 2007). Users can view a pathway's GO annotations in PathCase by navigating to a pathway's detail page and clicking 'GO Pathway

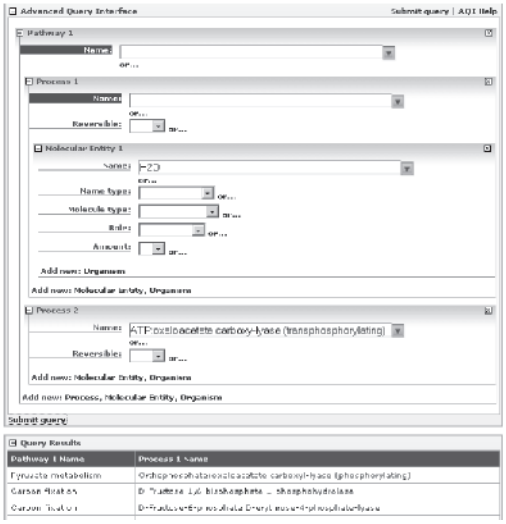

Fig. 5. Advanced query interface.

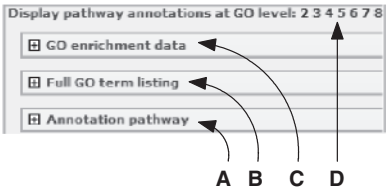

Fig. 6. PW-ANN menu.

Annotations (PW-ANN)'. Figure 6 presents partial PW-ANN menu. Annotation analysis can be performed with respect to different levels in GO (link D in Fig. 6). To gain a better understanding of which GO concepts exist at the selected level, all concepts can be listed by clicking the 'B' link. Link A in Figure 6 leads to PFT visualization of a given pathway. Figure 7a displays the regular drawing of pathway *metabolism of ether lipids,* and Figure 7b depicts the annotation pathway at the most specific level of GO. Processes with dashed lines are the ones whose enzymes do not have annotations. For the drawing of PFT graphs, the open-source QuickGraph library (QuickGraph) is used.

DEFINITION. (*GO Annotation Significance*): *Given a pathway set S, an annotation concept c and a pathway p in S, c is significant in p if the statistical significance of c in p is less than the threshold* γ*, namely,* $P(c, p, S) < \gamma$.

DEFINITION. (*GO Concept Enrichment/Deficiency*):*For a given GO concept c, assume K(c, S) out of N(S) processes in S are annotated by c. And, for a given pathway p with n processes, let k(c, p) be the number of processes annotated by c in p. We say that c enriches p if its annotation is significant in p, and the observed annotation count k(c, p) of c in p is greater than the expected annotation count n(p)*[K(c, S)/N(S)] of c in p, which is k(c,p) > n(p)*[K(c, S)/N(S)]. Likewise, we say that c is deficient in p if its annotation is significant in p, and the observed annotation count k(c, p) of c in p is less than the expected annotation count n(p)*[K(c, S)/N(S)] of c in p, that is, k(c,p) < n(p)*[K(c, S)/N(S)]. Furthermore, we say that c annotates p with the enrichment ratio R(c, p, S) = k(c, p)/ [n(p)*K(c, S)/N(S)].*

Example: Figure 8 shows a partial listing of enrichment statistics for the Proline and Hydroxyproline Metabolism (p). The listed annotations are from the fifth level of the GO hierarchy. The first column contains the name of the GO concept (g), followed by the number of processes in p which were expected to be annotated by g. The third column lists the enrichment ratios. Finally, the significance of annotation is listed as computed by the hypergeometric distribution. Concepts which enrich p are listed in green (enriching terms in Fig. 8). The GO concepts proline dehydrogenase activity and transaminase activity both enrich pathway p, but, in this case, the former enriches p more than the latter since the degree of enrichment is inversely proportional to the *P*-value. For each GO term, two lines of values are presented. The first line in each row provides the statistical analysis results with respect to the whole database, and the second line presents the results for statistical analysis with respect to p's pathway group (i.e. *Amino Acids and Derivatives* pathway group for this example).

In addition to enrichments, PathCase discovers annotation deficiencies. A pathway p is deficient in an annotation c, if c is significant in p (i.e. $P(c, p, S) < \gamma$) and the enrichment ratio R of c is less than 1; this means that c significantly under-annotates p. The 'missing' annotations are also included in the output data, defined as follows: a GO concept c annotates at least one pathway within p's pathway group, but does not annotate p, and the expected number of annotations with the concept c is at least 1. We only output missing annotations from a given pathway group because the number of concepts annotating all pathways is much larger than the number of concepts annotating pathways within a group.

6 COMPARISON WITH OTHER SYSTEMS

KEGG, BioCyc and Reactome are three major web-based metabolic pathways data sources. Contents of these data sources are put together, maintained and curated by large numbers of biologists. PathCase does not aim to compete with, but to complete, these systems by providing an alternative user interface with additional capabilities. Next, we compare features of PathCase with KEGG, BioCyc and Reactome as follows:

Browsing: in the KEGG website, metabolic pathway pages are separated from pathway browsing interfaces. The only way of accessing different pathway information is by returning back to pathways homepage. Reactome and BioCyc add a search form to pathway pages so that users can pose a query without leaving the actual page. PathCase provides a browsing section on the left-hand side of all pages that lists other pathways, pathway-related information (processes, molecular entities, genes), and links to other tools (e.g. built-in queries, AQI) in PathCase.

Visualization: KEGG, BioCyc and Reactome provide static visualization of pathway graphs, i.e. they are manually drawn/previsualized prior to query time, and are delivered to users. It is also possible to export visualizations of these databases to external applications (i.e. VisANT for KEGG and CytoScape for Reactome/BioCyc) for dynamic visualization and graph theoretical analysis of pathways (e.g. degree distributions, shortest paths). However, these applications are separated from the browser, thus it is not possible to query the complete database (but only the

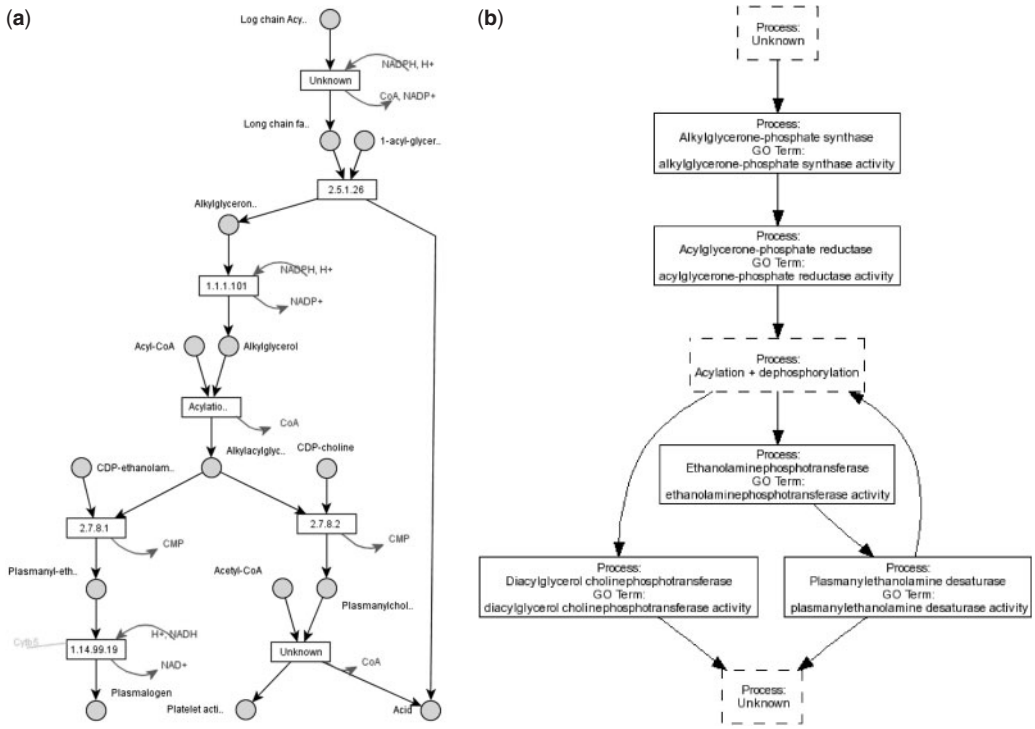

Fig. 7. (**a**) Regular pathway and (**b**) PFT view.

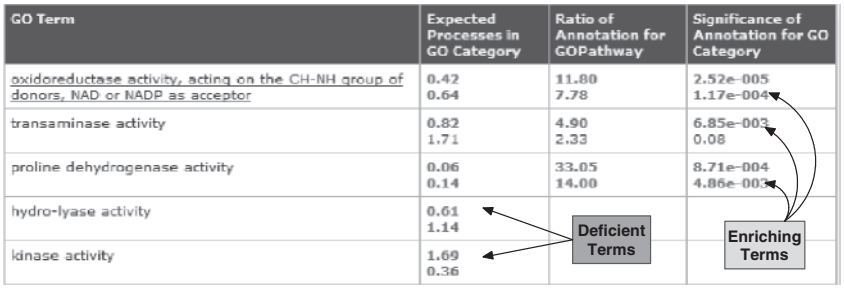

Fig. 8. Statistical annotation enrichment analysis.

visualized part) from the graph visualization. In comparison, PathCase integrates a dynamic (i.e. query-time) visualization applet that allows users to request changes to a pathway via queries, or to revise the pathway at hand via editing operations, and the system can then visualize the revised pathway or fragment of a network of pathways on the spot. In comparison, KEGG, BioCyc and Reactome provide a bird's-eye view of all pathways in a compact view; when an individual pathway is selected from the bird's-eye view, the pathway is displayed alone, isolated from the rest of the bird's-eye visualization. PathCase is able to increase the detail level of an individual pathway, while displaying the connections to all other pathways at a higher level.

The PathCase system supports visualization of any pathway that is provided in BioPAX format by the users. Our system has an interface for the users to upload the document. Next, it parses the document and extracts the entities which maps to the basic building blocks of our visualization system, such as metabolites, reactions, etc. Using this feature, users can visualize the metabolic networks which are exported from other database tools.

Querying: In addition to simple query forms that allow users to search for a keyword in database, BioCyc and Reactome provide advanced query interfaces that can add multiple predicates defined on different data object and fields. These interfaces can generate powerful queries; however, users need to understand the underlying data models, and specify table names, field names and field values in a query. PathCase AQI is more intuitive in that users are only exposed to common attributes of main data entities, such as the 'name of a pathway' or the 'role of a metabolite'. KEGG does not provide an advanced querying mechanism, and the KEGG database can only be queried by keyword searches. Query results in PathCase are displayed in both tabular and graphical outputs. KEGG, BioCyc and Reactome provide only tabular outputs (or only graphical output if an external visualization application is being employed), and they lack dynamic visualization mechanisms for query results.

Analysis: Reactome's SkyPainter provides users with pathways or reactions that are related most to a set of genes provided by the user. Given a set of genes, SkyPainter performs statistical analysis over events (i.e. pathways or reactions) via the hypergeometric test. PathCase provides a similar analysis tool where pathways that are significantly enriched by user-provided gene sets are listed in both graphical and tabular output. In addition, PathCase links enzymes to GO terms, and provides a GO-based enrichment/deficiency-based statistical analysis for functional annotations of pathways. KEGG and BioCyc do not provide significance analysis tools.

In addition to these systems, PATIKAWeb (Dogrusoz *et al*., 2006) is another well-known pathway database. PATIKAWeb employs a semi-dynamic system for displaying pathways. However, all navigation and editing operations in PATIKAWeb require retrieving the image of the resulting view from the server side. As a result, clients of PATIKAWeb put great overhead on the server, and the system easily becomes unresponsive, especially for slow/distant connections.

Other systems with pathways: There are many systems that, while targeted for other tasks, do provide limited pathway editing, storage and/or visualizations, and do not have the five goals of PathCase. Here, we briefly list two such systems. PathwayExplorer (Mlecnik *et al*., 2005) is a tool for mapping expression profiles on pathway data obtained from KEGG, BioCarta and GenMAPP. PathwayExplorer uses entities within static pathway images, and, lacks dynamic visualization and advanced querying capabilities.

BioMiner (Sirava *et al*., 2002) is a software library developed for modeling and visualizing biochemical data. BioMiner involves implementations of classes for essential biochemical entities, such as pathways, reactions and compounds. BioMiner contains a path finding tool, called PathFinder, which, given two molecules, computes possible paths over the metabolic network between them, and ranks the resulting paths.

ACKNOWLEDGEMENTS

The SQL Server software was donated by Microsoft.

Funding: National Science Foundation (DBI 0218061and CNS-0551603); Charles B. Wang Foundation to the Center for Computational Genomics, CWRU.

Conflict of Interest: none declared.

REFERENCES

Bader,G.D. *et al*. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34** (Database issue), D504–D506.

Berge,C. (1973) *Graphs and Hypergraphs*. North-Holland, Amsterdam.

BioPAX Working Group (2005) BioPAX—biological pathways exchange language. Level 2, Version 1.0 Documentation.

Cakmak,A. *et al*. (2007) Gene ontology-based annotation analysis and categorization of metabolic pathways. In *Proceedings of SSDBM*.

Dogrusoz,U. *et al*. (2006) PATIKAweb: a web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics*, **22**, 374–375.

Elliott,B. *et al*. (2008) Advanced querying interface for biochemical network databases, submitted for publication.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.

Krishnamurthy,L. *et al*. (2003) Pathways database system: an integrated system for biological pathways. *Bioinformatics*, **19**, 930–937.

Mayes,S. (2007) Advanced Interface for Querying Graph Data. Master's thesis, Case Western Reserve University, EECS Dept.

Michal,G. (1999) *Biochemical Pathways*. Spektrum Akademischer Verlag, Heidelberg.

Mlecnik,B. *et al*. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res*., **33**, W633–W637.

Sirava,M. *et al*. (2002) BioMiner - modeling, analysing, and visualizing biochemical pathways and networks. *Bioinformatics*, **19**, 219–230.