"The aMAZE LightBench: a web interface to a relational database of cellular processes"

Lemer, C ; Antezana, E ; Couche, F ; Fays, F ; Santolaria, X ; Janky, R ; Deville, Yves ; Richelle, J ; Wodak, SJ

ABSTRACT

The aMAZE LightBench (http://www.amaze.ulb.ac.be/) is a web interface to the aMAZE relational database, which contains information on gene expression, catalysed chemical reactions, regulatory interactions, protein assembly, as well as metabolic and signal transduction pathways. It allows the user to browse the information in an intuitive way, which also reflects the underlying data model. Moreover links are provided to literature references, and whenever appropriate, to external databases.

CITE THIS VERSION

The aMAZE LightBench: a web interface to a relational database of cellular processes

Christian Lemer, Erick Antezana, Fabian Couche, Frédéric Fays, Xavier Santoloria, Rekin’s Janky, Yves Deville¹, Jean Richelle* and Shoshana J. Wodak

The aMAZE Project, SCMBB, Université libre de Bruxelles, boulevard du Triomphe CP 263, B-1050 Bruxelles, Belgium and ¹Département d’Ingénierie Informatique, Université catholique de Louvain, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium

Received August 14, 2003; Revised September 30, 2003; Accepted October 28, 2003

ABSTRACT

The aMAZE LightBench (http://www.amaze.ulb.ac.be/) is a web interface to the aMAZE relational database, which contains information on gene expression, catalysed chemical reactions, regulatory interactions, protein assembly, as well as metabolic and signal transduction pathways. It allows the user to browse the information in an intuitive way, which also reflects the underlying data model. Moreover links are provided to literature references, and whenever appropriate, to external databases.

INTRODUCTION

With over 100 completely sequenced genomes available today the major challenge of present-day biology is to interpret this information in terms of the biological function of the constituent genes. This requires the characterization of not only the function of individual genes or proteins, but also the myriad physical and functional interactions in which they participate. These interactions can be viewed as forming a vast network of cellular processes, which include metabolic pathways, signal transduction cascades and regulatory networks.

Public databases such as Swiss-Prot (1) or InterPro (2) are rich sources of information on function, but they store it mostly in text form, not readily amenable to computer analysis. Efforts have therefore been undertaken to develop more specialized databases for representing information on cellular processes and interactions [see (3) for a recent review]. Some of these databases focus primarily on protein–protein interactions [DIP (4), BIND (5), MINT (6)], on gene regulations [TRANSFAC (7), RegulonDB (8)] or on signal transduction [CSNDDB (9) and TRANSPATH (10)]. Other widely used resources, such as KEGG/LIGANDS (11), EcoCyc (12) and WIT (13), originally specializing in metabolic pathways, now also provide information on gene regulation, transport and signal transduction. A more recent effort, the Genome KnowledgeBase (http://www.genomeknowledge.org) handles several different types of pathways.

The aMAZE database has been developed with the goal of providing a rich enough source of information on interactions and processes for the bench biologist, while at the same time enabling complex analyses, surveys and links to dynamic simulation packages [i.e. GEPASI (14)]. The aMAZE data model (3,15) embodies general rules for associating individual biological entities and interactions into large complex networks describing many different types of cellular processes.

Here we present the aMAZE LightBench (http://www.amaze.ulb.ac.be/). This is a web interface to the aMAZE relational database, which enables browsing of and navigation through the information currently stored in the database. This information pertains to chemical reactions, genes and enzymes involved in metabolic pathways and the transcriptional regulation of the corresponding genes (see Fig. 1). In addition, access is provided to information on protein–protein interactions and protein modification involved in a number of signal transduction pathways. The information on all the metabolic pathways, transcriptional regulation and signal transduction pathways has been expert curated from the scientific literature. Access to the aMAZE database is free for all and the information herein is available in downloadable form.

THE aMAZE DATA MODEL

The aMAZE LightBench enables browsing the information in an intuitive way, which reflects both the underlying data model and biological knowledge. The aMAZE model has been described in detail previously (3,15), and is therefore only briefly summarized here.

This model comprises two basic classes of objects: Biochemical Entities and Biochemical Interactions. The first represents physical entities (protein, gene, compound, etc.), with attributes pertaining to structural properties (e.g. gene position on the chromosome). The second represents molecular activities, which can be of several types. One type is Entity Processing, which are Biochemical Interactions having Biochemical Entities as input and as output (chemical reaction, protein–protein interaction are examples of Entity Processing). The other type is Control, which are Biochemical Interactions having a Biochemical Entity as input and another...
**Biochemical Interaction** as output (e.g. a catalysis is a *Control Biochemical Interaction* between a protein and a reaction).

A third important class in aMAZE is **Process**, which represents a collection of interconnected process elements. Using this representation, graphs of biochemical pathways can be built at various levels. They can be built at the enzyme and metabolite level by considering the chemical reactions as nodes and linking them through their inputs and outputs, or at the pathway level by considering whole pathways as nodes and linking those via common key metabolites. Currently, only the first option is available.

**THE LIGHTBENCH**

The aMAZE LightBench is implemented using Zope/Python technology (http://www.zope.org/) for management of the website and interrogation of the back-end relational database. This choice provides all the required functionalities and offers excellent possibilities for efficient integration with the aMAZE WorkBench. The latter is a modular stand-alone application, currently under development, which constitutes the main interface to the aMAZE database, offering annotation tools, query capabilities and graphical representation of pathways as well as programmatic access.

In contrast, the LightBench allows simple browsing of and navigation through the information stored in the aMAZE database using two modes of interaction: (i) a ‘query mode’, implemented with HTML forms, which allows the retrieval of lists of stored objects on the basis of their name; (ii) a ‘navigation mode’, which follows the relationship between the objects, implemented with HTML hyperlinks. To allow the widest and most flexible navigation possible, we attached hyperlinks to all the possible parts of the different pages presented to the user. Names of objects (*Gene*, *Polypeptide*, *Compound*, *Organism*, *Pathway*), shown in light blue (see Fig. 2), lead to a detailed description of these objects. Arrows (see Fig. 2C), denoting *Expression*, *Transcription Regulation*, *Reaction*, also have hyperlinks attached leading to the corresponding detailed description.

Collecting the detailed description of an object from the aMAZE relational tables, always requires formulating several SQL queries: one query for retrieving the single value attributes and one query per multi-valued attribute. SQL queries are also used to get information about the related objects displayed in the detailed description. Thus, the LightBench interface completely hides from the user the complexity of the database organization (schema) and provides a user-friendly means of displaying information and navigation.

**Accessing information on metabolic pathways and transcriptional regulations**

Access to the aMAZE database is provided through schematic diagrams summarizing the data organization. One such diagram is depicted in Figure 1. It shows the organization of the different types of objects stored in the database for describing metabolic pathways and transcriptional regulation. Clicking on any box gives access either to the ‘Query by Name’ page for the corresponding biochemical entity (square boxes) or to a page explaining how to query the given biochemical interaction (oval boxes). Only the biochemical entities can be queried by name. Biochemical interactions are retrieved through the entities participating the interaction. The special ‘EC Classification’ box gives access to a form accepting EC numbers, which are objects establishing the relations between *Polypeptides* and *Reactions* [in this version of the database, there is no direct relation between a polypeptide and the reaction(s) it catalyzes].

Clicking on a square box of the LightBench front page (see Fig. 1) gives access to a ‘Query by Name’ form, for example the *Genes* form (Fig. 2A). By filling in such a form, the user restricts the search in the database to objects whose names match the entered string. Partial string matching is handled by using ‘%’ as the wildcard character. The layout of the ‘Query
by Name’ form is customized for different types of objects. One can, for example, restrict the search of genes to a given organism. One can also require that the selected objects participate in a particular interaction, e.g. one can restrict the search to yeast genes that code for a polypeptide and are also regulated by a polypeptide. Upon entering a choice and pushing the ‘Submit’ button, the query is processed and a ‘Query Result’ page is displayed with the list of matching objects (see Fig. 2B).

Each of the elements in the table of a ‘Query Result’ page (Fig. 2B) gives access to the detailed description of the corresponding object. The layout of this detailed description is also object specific, and depends on the information stored in the database for that object.

For each object, key associated information is also retrieved and displayed. For example, if the retrieved object is a gene, the polypeptide it expresses or its transcriptional regulator(s) are displayed (Fig. 2C). This additional information not only provides a more complete description of the object but also allows navigation through the web of relations between the different objects stored in the database. Other examples of information displayed for particular object types are as follows. For a Compound, which can be a Substrate or a Product, the LightBench displays the Consuming- and Producing-reactions (i.e. the list of all the reactions in which the given compound participates as a substrate or as a product). For a Reaction, the LightBench displays, whenever available, all the Processes in which that reaction is involved,
the list of Polypeptides that catalyze it, and the lists of Reactions which, respectively, produce the substrate(s), or consume the product(s) of the considered reaction.

The detailed description of objects also provides links to external databases. Links to PubMed bibliographic references are provided for the stored metabolic pathways. For Polypeptides, links are provided to Swiss-Prot, whereas Compounds and Reactions are linked to the corresponding entries of the KEGG database.

For the Processes (or pathways) stored in aMAZE, the detailed description lists the set of biochemical reactions that are considered as part of the pathway by the biochemists. In addition, a diagrammatic representation of the pathway is displayed, showing how the source compound is transformed into the final product by a succession of reactions involving intermediate compounds. The names of displayed compounds and reactions are clickable, providing access to the detailed description of the corresponding objects.

Accessing information on signal transduction

The aMAZE LightBench provides access to information on signal transduction through a second schematic diagram summarizing the data organization (Fig. 3). The information is represented using the same two main aMAZE object classes, Biochemical Entity and Biochemical Interaction. The latter objects are all of the Transformation type and include Expression (of Genes), Assembly (of Biochemical Entities), Translocation (of Biochemical Entities between cellular localization) and Reaction (here mainly describing protein modifications such as phosphorylation, acetylation, ubiquitylation, etc.). An object of the Polypeptide class describes each of the unmodified or modified proteins. A Complex is a subclass of Biochemical Entity which is indirectly related to its component Entities (Polypeptides or other Complexes) through Assembly Transformation. To facilitate access, a Complex is also related to each of its Polypeptide components, and lastly, the detailed object description provides links to PubMed.

STORED DATA

The data currently stored in aMAZE are summarized in Table 1. The data on genes and proteins have been downloaded from other public resources such as KEGG and Swiss-Prot, whereas data on transcriptional regulation, protein interactions, assembly, translocation, as well as on pathways (metabolic and signal transduction), have all been curated in-house. We expect the amount of data stored in aMAZE to increase rapidly in the near future, as the aMAZE pipeline for automatic data loading and versioning, as well as the tools for custom annotation, become available.

CONCLUSIONS AND PERSPECTIVES

We present the first version of the aMAZE LightBench, which is a web interface to the aMAZE relational database. This interface makes use of very flexible technology and can therefore be readily adapted in the future to provide a
Table 1. Overview of the current contents of the aMAZE database

<table>
<thead>
<tr>
<th>Objects in the aMAZE database</th>
<th>Source of the data</th>
</tr>
</thead>
<tbody>
<tr>
<td>12689 genes</td>
<td>parsed from the KEGG files obtained at ftp://ftp.genome.ad.jp/pub/kegg/genes/genes/</td>
</tr>
<tr>
<td>4405 from Escherichia coli</td>
<td>aMAZE in-house curation</td>
</tr>
<tr>
<td>6719 from S.cerevisiae</td>
<td>parsed from Swiss-Prot files obtained at ftp://ftp.expasy.org/databases/sp_tr_nrdb/</td>
</tr>
<tr>
<td>1565 from Homo sapiens</td>
<td>automatic matching between genes and polypeptides</td>
</tr>
<tr>
<td>583 transcriptional regulations</td>
<td>parsed from the LIGAND files obtained at ftp://ftp.genome.ad.jp/pub/kegg/ligand</td>
</tr>
<tr>
<td>61249 polypeptides</td>
<td>parsed from the LIGAND files obtained at ftp://ftp.genome.ad.jp/pub/kegg/ligand</td>
</tr>
<tr>
<td>8717 from E.coli</td>
<td>parsed from the LIGAND files obtained at ftp://ftp.genome.ad.jp/pub/kegg/ligand</td>
</tr>
<tr>
<td>6913 from S.cerevisiae</td>
<td>aMAZE in-house curation</td>
</tr>
<tr>
<td>45619 from H.sapiens</td>
<td></td>
</tr>
<tr>
<td>12830 expressions</td>
<td></td>
</tr>
<tr>
<td>10464 compounds</td>
<td></td>
</tr>
<tr>
<td>4219 EC numbers</td>
<td></td>
</tr>
<tr>
<td>5281 reactions</td>
<td></td>
</tr>
<tr>
<td>106 metabolic pathways</td>
<td></td>
</tr>
<tr>
<td>76 from E.coli</td>
<td></td>
</tr>
<tr>
<td>338 signal transduction transformations</td>
<td>aMAZE in-house curation</td>
</tr>
<tr>
<td>65 assemblies</td>
<td></td>
</tr>
<tr>
<td>39 translocations</td>
<td></td>
</tr>
<tr>
<td>84 expressions</td>
<td></td>
</tr>
<tr>
<td>150 reactions</td>
<td></td>
</tr>
<tr>
<td>298 signal transduction controls</td>
<td>aMAZE in-house curation</td>
</tr>
<tr>
<td>18 signal transduction pathways</td>
<td>aMAZE in-house curation</td>
</tr>
<tr>
<td>all from S.cerevisiae</td>
<td></td>
</tr>
</tbody>
</table>

The first column lists the different types and the respective number of objects stored in the aMAZE database (as of October 2003), when applicable, the number of objects is also given according to the species currently covered by aMAZE. The second column lists the source from which the corresponding information was obtained.

*The information has been checked against data stored in the BREnda database (16), and modified in the case of inconsistencies.

A repertoire of `canned queries' representing requests of a wide range of complexity.

The LightBench, one component of the aMAZE environment, is aimed at providing simple ways of browsing and navigating through the stored information. The second component is the aMAZE Workbench, a Java-based application that features modules for data loading and extraction, data modification and annotation, data visualization and analysis. The first version of the Workbench will be released in the fall of 2003. It will enable the input of information on many more pathways and interactions, already compiled from the literature by the aMAZE team. It will furthermore allow this information to be updated at regular intervals, and collaborations with expert laboratories on the annotation of biological processes of particular interest to be undertaken. The Workbench will furthermore support state of the art tools for managing modifications of the data model (schema migration), making it possible to accommodate new data types or data models with increased efficiency. The technology adopted for the LightBench will allow this interface to be kept almost unchanged, since it will only require the scripts for generating the web pages and for querying the database to be adapted.

ACKNOWLEDGEMENTS

We thank Jacques van Helden for continuing discussions and help, David Gilbert for very useful interactions, Jesintha Maniraja for careful testing of the LightBench and constructive suggestions, and Anne Richelle for the layout reorganization of the metabolic pathway diagrams. This work was supported by a grant to the Université libre de Bruxelles from the Ministry of Research of the Brussels Capital Region (special thanks to M. F.-X. de Donné, Minister of State of Belgium). Y.D. acknowledges support from the UK Engineering and Physical Sciences Research Council (EPSRC, project GR/S07490/01).

REFERENCES


