aMAZE: A Database of Molecular Function, Interactions and Biochemical Processes

Christian Lemer[1], Avi Naim[2], Yong Zhang[2], Didier Croes[1], Georges N. Cohen[4], Gaurab Mukherjee[2], Lorenz Wernisch[2,3], Klaudia Walter[2], Jean Richelle[1], Jacques van Helden[1] and Shoshana J. Wodak [1,2*]

[1]Centre de Biologie Structurale et Bioinformatique, CP160-16, Université Libre de Bruxelles,
50 Av. F. Roosevelt, 1050 Bruxelles, Belgium

[2] European Bioinformatics Institute (EBI), Genome Campus –
Hinxton, Cambridge CB10 1SD, UK

[3] School of Crystallography. Birkbeck College. University of London.
Malet Street. London WC1E 7HX, UK

[4] Unité d' Expression des Gènes Eucaryotes,Institut Pasteur, 28,
rue du Docteur Roux, 75524 Paris Cedex 15, France

E-Mail:* *Shosh@ucmb.ulb.ac.be*

Received:15th July 2002 / Published: 15th May 2003

Abstract

The aMAZE database (http://www.amaze.ulb.ac.be) manages information on the molecular functions of genes and proteins, their interactions and the biochemical processes in which they participate. Its data model embodies general rules for associating molecules and interactions into large complex networks that can be analysed using graph theory methods. The processes represented include metabolic pathways, protein-protein interactions, gene regulation, transport and signal transduction. These processes are mapped into their spatial localisation. A distinct feature of aMAZE is its Object-Oriented, modular and open user interface. Queries are invoked through dedicated modules, data can be linked to external sources, interactively browsed and transferred between modules, and new modules can be readily added. Available modules also include, a custom-built Diagram Editor for the automatic layout, display, and interactive modification of pathway diagrams, and procedures for analysing network graphs.

INTRODUCTION

A major challenge of the post genomic era is to determine the biological function of all the genes and gene products in the growing number of newly sequenced genomes, and to understand how they interact to yield a living cell. To meet this challenge, experimental efforts of unprecedented magnitude are being undertaken for investigating gene expression patterns, analysing the entire protein complement of cells and characterising the full repertoire of protein-protein interactions. These efforts involve high-throughput techniques, such as micro-array based gene expression analysis (1,2), two-hybrid screens (3-6), and large scale protein characterisation (7). All these techniques generate very large amounts of data that must be interpreted in an iterative bootstrapping approach in light of the available information on the molecular and cellular function of genes.

This has brought to the forefront, the pressing need for more efficient systems for managing and analysing complex information on biological function. Public databases such as SWISS-PROT (8), are a rich source 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 specialised databases, for representing information on cellular processes and interactions (see reference 9 for a recent review). Some databases deal primarily with metabolic pathways (10-12). Others focus on protein-protein interactions (13,14), on gene regulations (15,16) or on signal transduction (16,17).

The aMAZE database system presented here, implements a data model, which embodies general rules for associating individual biological entities and interactions into large complex networks of cellular processes (9,18). It can deal with a large variety of cellular processes comprising metabolic pathways, protein-protein interactions, gene regulation, sub-cellular localisation, transport, and signal transduction. The major aim of this system is to provide a general open framework for: 1) combining information from different levels of cellular organisation, 2) flexibly querying and visualising this information, 3) custom-building of tools for advanced programmatic analyses, and 4) greatly facilitating annotation of data on complex cellular processes. The aMAZE system or others like it, should help the biologists in understanding, analysing, and ultimately modelling, complex cellular networks.

THE AMAZE DATA MODEL

In aMAZE data are organised using an Object Oriented model, as described previously (9,18). This model distinguishes between two fundamental classes of objects, *BiochemicalEntity* and *Interaction*. The first represents physical entities (protein, gene, compound, etc.), with attributes pertaining to structural properties (polypeptide sequence, gene position on the chromosome). The second represents molecular activities, which can be of several types. *EntityProcessing* and *Binding,* which are interactions having entities as input and as output (e.g. chemical reaction, protein-protein interaction), and *Control*, which are interactions having an entity as input and another interaction as output (e.g. a catalysis in a relationship between a protein and a reaction).

A third important class in aMAZE is *Process*, which represents a collection of interconnected process elements. These elements consist either of individual interactions or of entire processes. Using this representation, graphs of biochemical pathways can be reconstructed by linking the interactions through their inputs and outputs. Higher level views, for example, of the interconnections between different biochemical pathways (pathways of pathways), can also be generated.

As discussed previously (18), our model has the great advantage of defining the activities of a particular structural entity (compound, gene or protein) within a given context, rather than within the entity object itself, thereby allowing for a flexible description of multiple activities of individual genes and proteins.

The description of localisation, a central issue in representing biological function, is also handled in aMAZE. This is done using the class *Compartment*, which is further subdivided into the sub-classes, *SubcellularCompartment, CellType, Tissue, Organ,* and *Systematic group.* Any *Process* can be linked to a given combination of objects in the *Compartment* class in order to describe where it occurs (e.g. plasma membrane of T-cells in *Homo sapiens*).

The aMAZE data model also comprises various types of classification schemes. One type is the containment hierarchies, e.g. the nucleosome is contained in the nucleus, itself contained in the cell. Another type concerns various classifications of objects of the aMAZE model. Those include the systematic classification of organisms, sub-cellular compartments, compounds, and so on. The ability of representing independently and simultaneously one or more of the so-called biological Ontologies (19), which contain functional classifications, is also provided. Finally, the aMAZE database also includes its own meta description - classes used to represent the aMAZE data model itself - which can hence be queried and analysed like any other information stored in the database.

Implementation Aspects

The aMAZE system is implemented in a mixed Object-Oriented/Relational environment. The data are manipulated as objects, but stored in Tables using a relational DBMS (Oracle). The Object-Relational mapping is performed using an algorithm that converts the full object description into the relational schema, in ways that can be readily understood by humans, while preserving key properties of the object model, such as inheritance and polymorphism (Lemer et al., to be published). This conversion also enables the user to choose between the Object-Oriented or relational modes, according to need, without loss of user friendliness or query power.

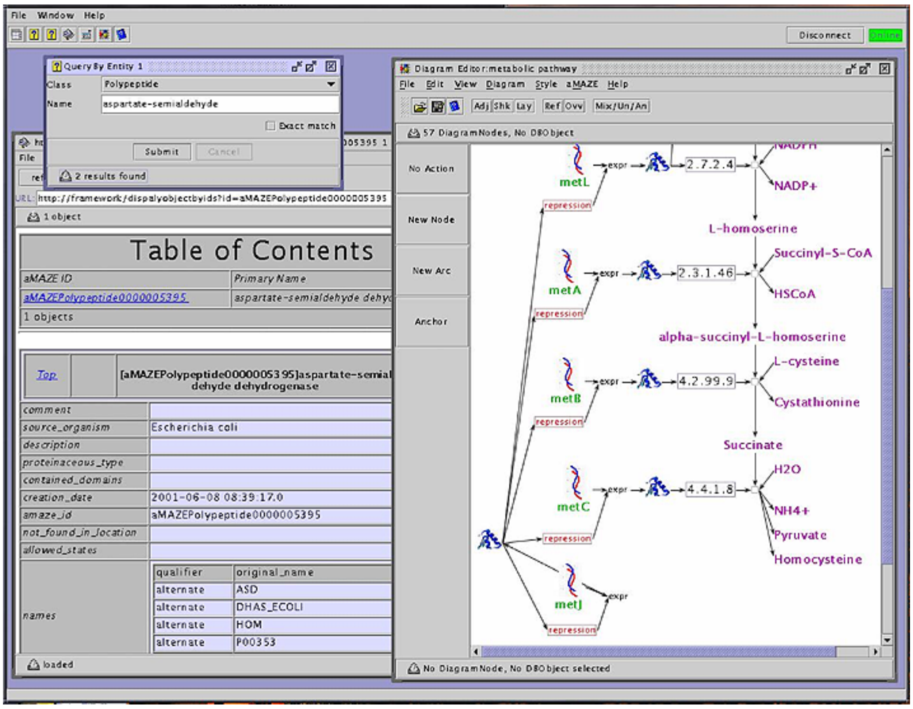
The server architecture comprises three layers: the Service, Access and Client layers. The Service layer manages multiple access to the relational database for query purposes and updates. The Access layer manages the users and the access rights, and is also in charge of load balancing between servers. The Client layer manages network communication and provides the API (Application Programmatic Interface) to the server layer. The different layers are developed in Java and are connected via the Remote Method Invocation system (RMI).

The aMAZE Framework:
A Modular Object-Oriented Open User Interface

User access to aMAZE is provided via the aMAZE_Framework. This is a multi-document application, where each type of query to the database is invoked through a dedicated module, as illustrated in figure 1, and data can be transferred between modules using *drag & drop* operations.

Query modules currently available comprise the retrieval of molecular entities (gene, protein small molecule) by name, and by requiring an exact or partial match. They also allow the retrieval of different types of interactions, each defined by one or more inputs, one of more outputs, or a list of both inputs and outputs. Another module allows the selection of objects on the basis of any feature (attribute) of interest. This module can be used iteratively to refine a selection by sequential filtering.

Upon query completion, the user can drag and drop the results into the Object-Holder, which enables viewing the results list. Individual items of the list can moreover be dropped into the Object-Viewer, in order to visualise their complete description (see Fig. 1). Importantly, retrieved entities or interactions can in turn be entered as input into a new query window via the drag and drop facility.



Class: Polypeptide

Name: aspartate-semialdehyde

□ Exact match

2 results found

URL: http://framework/dispalyobjectbyids?id=aMAZEPolypeptide0000005395

1 object

Table of Contents	
aMAZE ID	Primary Name
aMAZEPolypeptide0000005395	aspartate-semialdehyde dehyd
1 objects	

Top [aMAZEPolypeptide0000005395]aspartate-semialdehyde dehydrogenase

comment	
source_organism	Escherichia coli
description	
proteinaceous_type	
contained_domains	
creation_date	2001-06-08 08:39:17.0
amaze_id	aMAZEPolypeptide0000005395
not_found_in_location	
allowed_states	
names	qualifier / original_name alternate / ASD alternate / DHAS_ECOLI alternate / HOM alternate / P00353

loaded

Figure 1. Screen shot of the aMAZE-Framework in action. The upper left corner shows the query by entity module, where a polypeptide (aspartate-semialdehyde dehydrogenase) is retrieved using only part of its name (aspartate-semialdehyde) as input. The information on the retrieved polypeptide object is displayed using the Object-Viewer, which appears as the large application window visible directly beneath the query window. The Diagram Editor, window, displaying a pathway appears of the right half of the Figure. It shows part of the methionine biosynthesis pathway. The different functions available in the Diagram Editor appear can be activated by the icons displayed on the tool bars on the top and the left hand sides. Catalysis interactions are displayed by boxes containing the corresponding EC number. Small molecule compounds are displayed in purple, proteins appear as blue icons, all of which show the polypeptide backbone of a small protein. These can be replaced by picture of the individual proteins, once the link with the Protein Databank (PDB) is established. The 'exp' and 'repression' objects represent the expression and repression interactions, respectively. Gene names are in green.

SQL queries to the relational database can also be submitted directly by invoking a specialised module. Users unfamiliar with SQL, can run pre-canned custom queries, prepared in advance with help from the database administrator, and accessed via the SQL module. The results of the SQL query can in turn be dropped into any other display module (Object-Holder, Object-Viewer, Diagram-Editor). Thus, the framework combines the power of SQL queries with the user-friendliness of an object representation.

At any given time, multiple windows can remain open, facilitating data traffic between them. The Object-Viewer can furthermore be used as a general HTML WEB browser, thereby enabling direct links to external databases. The latter aspect is extremely useful for all operations that may benefit from analysing information from different sources. The aMAZE-Framework has been developed in ‘pure’ Java, and has already been tested on different platforms (Solaris, Linux, Windows, Macintosh).

The Diagram Editor

A key module of the Framework is the Diagram Editor. This is a custom built graphical editor, which allows to display diagrams of cellular processes retrieved from the database (e.g. pathway diagrams), and to interactively modify them to suit particular aesthetic preferences. Modified diagrams can be saved locally in different formats (jpg, png, text ..), printed, and (under some conditions) re-submitted to the database.

In addition to a number of useful standard display options, users can chose to collapse specific nodes or node types, with or without their complete set of anchored descendents. Modified diagrams can be stored locally, to be printed or displayed subsequently.

Options for an automatic layout of pathway diagrams are also provided. This is performed using a custom-built multi-directional hierarchical layout algorithm, tailored for the representation of biochemical pathways (van Helden, unpublished), as illustrated in figure 1.

When the Diagram Editor is integrated into the Framework, any set of displayed objects can be dragged and dropped into a query or Object-Viewer document. We expect this to be particularly helpful, as it enables the biologist to manipulate complex data in an intuitive fashion. It should likewise be very instrumental for entering new information on complex biological processes, as the newly entered information can not only be displayed graphically but also validated for consistency against information already stored in the database. In addition if a newly entered process involves entities or interactions already stored in aMAZE, work can be saved by using the stored information.

Graph Analysis Tools

Another group of modules features a set of tools for analysing cellular processes algorithms adapted from graph theory (20). They include for example, finding the shortest path or N-shortest paths, between a given pair of source and target nodes. In performing these operations a limit on path length (number of intervening nodes) can be imposed to limit the number of generated solutions, and thus reduce calculation time.

Path enumeration is not restricted to annotated pathways but can be performed on the global metabolic network built on the fly from all the chemical reactions and compounds stored in the database, irrespective of where (organism, compartment) the corresponding catalytic reactions were observed. This has many useful applications. A typical application is, given a cluster of co-expressed genes (identified in DNA micro-array experiments), find the shortest paths linking together, in a biologically meaningful way, most or all of the activities carried out by the cluster members (9,18). This can help in assigning gene function, in identifying alternatives to classical pathways and in discovering new pathways. Figure 2a, b illustrates the method, and Figure 3 its application to the reconstruction of the path of chemical reactions linking the activities carried out by the genes from the so-called Met cluster in the experiments of Spellman et al. (1998) (19).

A detailed description of the path-finding approach and some more general issues in representing networks of metabolic processes can be found in reference 20.

Another application is to characterise the functional interactions between pairs of enzyme coding genes found to be involved in fusion/fission analyses (22-25). To that end, the path enumeration algorithms are used to compute the length of the shortest path between to corresponding pair of catalyzed reactions. The shorter the path, the greater the likelihood that the two enzymes functionally interact (20).

The performance of theses algorithms is currently being tested by applying them to rebuild known metabolic pathways, starting from different subsets of catalysed reactions (20). Results suggest that the algorithms should also be very useful for the on-going work on pathway annotation, as pathways for which all reactions and compounds are already stored in the database can be rebuilt and displayed automatically before examination by the annotator.

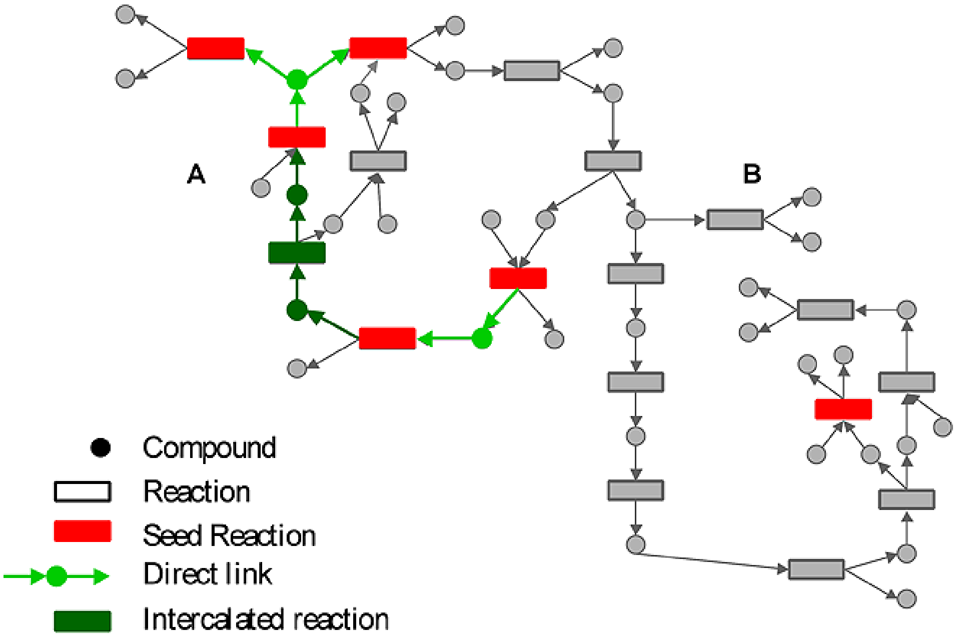


Figure 2. Reconstruction of a metabolic pathway from a set of seed reactions.

A) This figure schematises a putative network of chemical reactions linked via their substrates and products (filled circles), which are small molecule compounds being transformed into one another by the chemical reactions (filled rectangles). Information on all the reactions of the network is stored in the database. Red rectangles indicate chemical reactions given as seeds to the graph analysis program. The program then attempts to links these seed reactions via their substrates and products. When a direct link (light greed arrows and circles) cannot be made, the program is allowed to intercalate 1 or more reactions, which were not part of the initial set of seeds (dark green arrows, circles and rectangles).

B) Sub-graph extracted using the procedure described in **A)**.This sub-graph represents the pathway built from the set of initial seed reactions. Note that one seed reaction (right hand side of Figure) could not be connected, and is therefore discarded from the analysis. Comparing the constructed pathway with known pathways stored in the database can then be then used to determine if the pathway is already known, is a variant of a known pathway, or a completely new one.

Data content, and Annotation Efforts

The main sources of the data in aMAZE are, the BRENDA database (27), containing information on the majority of classified enzyme reactions and the proteins that catalyse them, the KEGG database (12), and SWISS-PROT (8). For data on pathways, we completely rely on in-house annotations.

Presently, aMAZE contains information on more than 150,000 genes, about 200,000 polypeptides, over 8000 chemical reactions, a similar number of small molecule compounds, and about 60 pathways of metabolic regulation in *E.coli* and *S. cerevisiae*. Information on over 150 additional pathways has been collected and will be entered into the aMAZE system by the fall of 2002. These will include all the known metabolic regulation and signal transduction pathways of *E.coli* and most such pathways in *S. cerevisiae*.

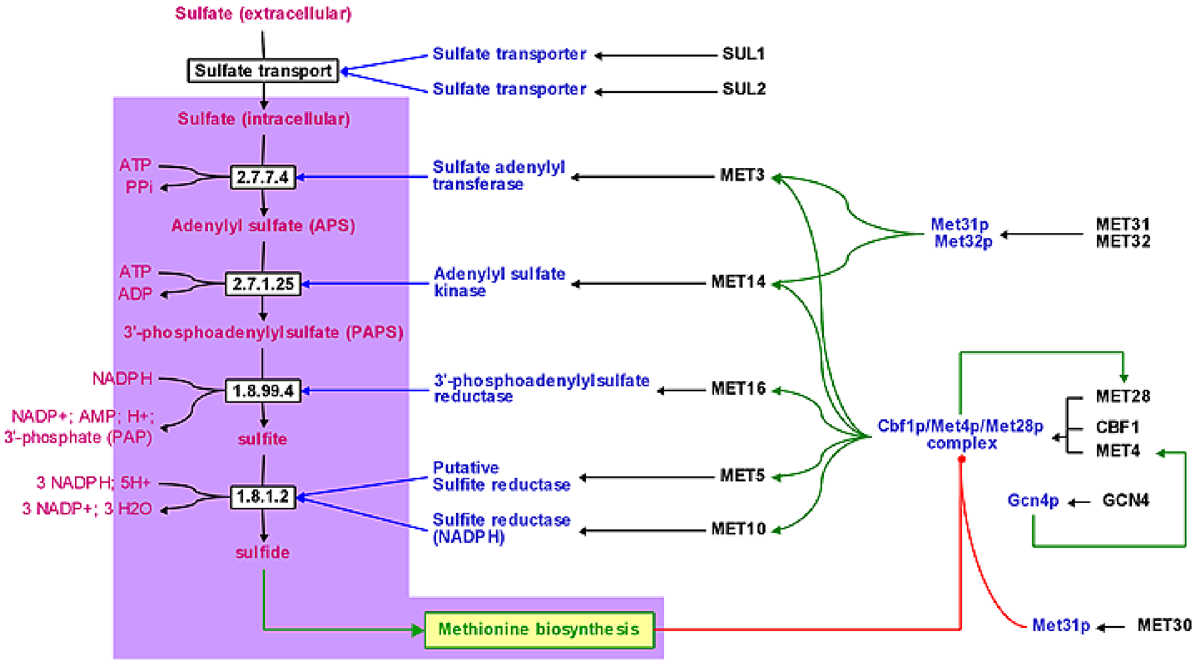
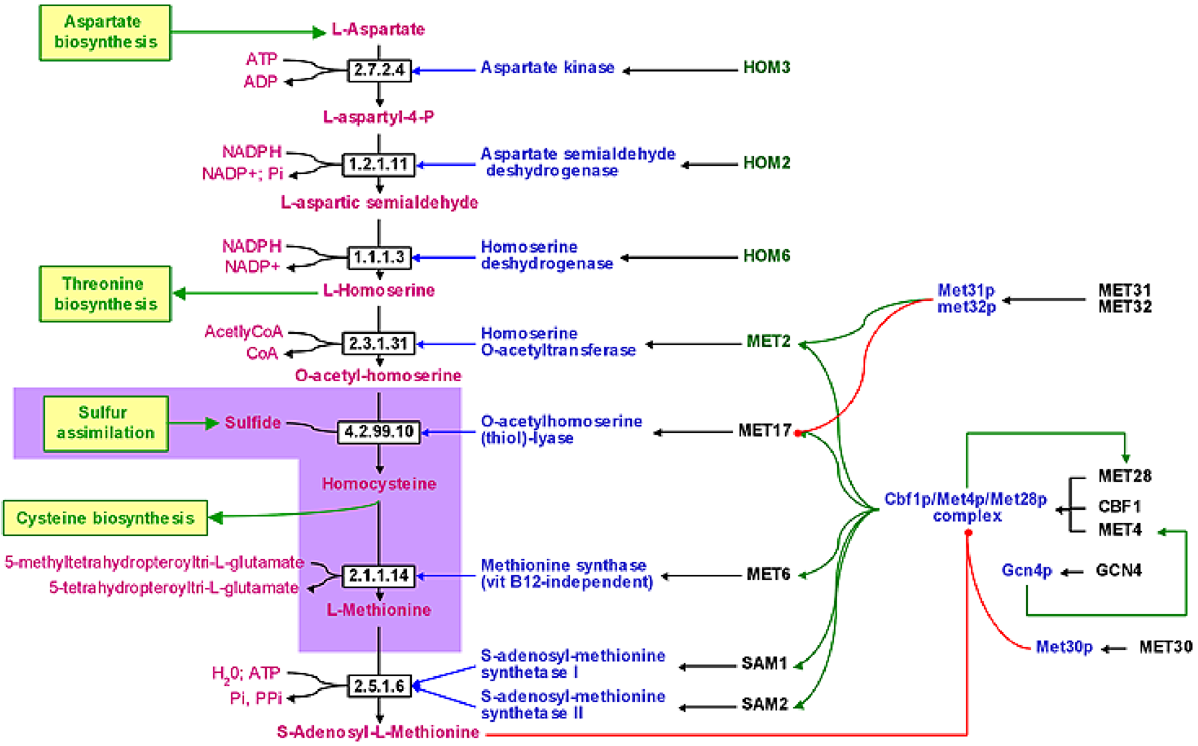


Figure 3. Pathway reconstruction from the cluster of cell-cycle regulated genes (19), involved in sulfur assimilation in *S. Cerevisiae***.**

A) Part of the reconstructed pathway that involves the enzyme coding genes marked in black (Met3, Met5, Met10, Met14, Met16). This part deals with sulfate to sulfite transformation.



B) Part of the reconstructed pathway, which belongs to the classical pathways of methionine biosynthesis. It involved the enzyme coding genes marked in black (Met6, Met17). The method used to reconstruct the pathway is the one illustrated in Fig. 2. The genes used as seeds for the pathway reconstruction belong to those from the so-called Met cluster in Spellman's experiments (19).

The flexible query and graphical tools available in aMAZE, together with the specialised annotation modules that are currently being developed (see below), should greatly facilitate the annotation efforts in the very near future. Groups interested in annotating pathways will be provided with access to aMAZE, as soon as these modules become functional, and we very much hope that such groups will contact us. As with all information in aMAZE, contributed data are processed and assigned unique references, enabling the identification of the annotator for the record, as well as for query purposes.

FUTURE DIRECTIONS

A major priority is to further populate the database with available information on cellular processes and data on protein-protein interactions in the yeast *S. cerevisiae*, which is among the best characterised eukaryotic systems so far. This will allow evaluating the full potential of aMAZE on a consistent dataset, and should form the basis for extending the annotation efforts to higher eukaryotes, or less well characterised prokaryotes. In the near future, this effort will focus on processes pertaining to cell division and cancer, in mouse and human.

A second priority is the expansion of the query capabilities. In particular we will focus on the application of graph analysis tools to the interpretation of gene expression and protein-protein interaction data on terms of cellular pathways, and tools for visual and quantitative comparison of pathways and networks.

Much work also remains to be done on the representation of data on the small molecule compounds, so that queries can be formulated on selecting and comparing compounds and reaction on the basis of the compound chemical structure and sub-structure. This information should also allow the introduction of criteria based on chemistry for the analysis and construction of metabolic pathways.

AVAILABILITY AND ACCESS

The aMAZE system will be freely accessible to academic research groups over the Web starting the fall of 2002. The available functionalities will include all query modules, and protocols for custom development of new query methods. All the data in aMAZE, except for those belonging to third parties or declared as confidential, are freely available. It is also our intention to make the entire aMAZE system publicly available starting April 2003, date at which the commitments towards the consortium of Industries supporting this project is ending. Groups interested in developing their own applications are most welcome to contact us. We would also be happy to provide specialised groups, or individuals, interested in contributing annotations on cellular pathways, with privileged access to the custom annotation modules of aMAZE, as soon as those are available.

ACKNOWLEDGEMENTS

We thank Thure Etzold, and David Gilbert for useful discussions. Many of the ideas for the aMAZE data model are due to Renato Mancuso, and the first prototype version of the database was implemented by Matthew Eldridge. Kirill Degtyarenko is thanked for valuable help with data annotation. This work was sponsored in part by a consortium of industries, comprising Astra-Zeneca, Aventis, Monsanto, Organon, and Roche. We thank scientists from these companies for valuable input. Support from the European Commission (contracts QLRI-CT-1999-01333 and BIO4-CT98-0226), is also gratefully acknowledged.

REFERENCES

[1] Brown P. O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat. Genet*. **21**:33-37.

[2] DeRisi J. L., Iyer, V. R., Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680-686.

[3] Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340** (6230):245-6.

[4] Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., Legrain, P. (2001). The protein-protein interaction map of Helicobacter.

[5] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403** (6770):623-7.

[6] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U S A* **98** (8):4569-74.

[7] Fields, S. (2001). Proteomics. Proteomics in genomeland. *Science* **291** (5507):1221-4.

[8] Bairoch A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. **28**:45-48.

[9] van Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D., Wodak, S. J. (2000). Representing and analysing molecular and cellular function using the computer. *J. biol. Chem.* **381** (9-10): 921-35.

[10] Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Res*. **28**:56-59.

[11] Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M. Jr. E. S., Kyrpides, N., Fonstein, M., Maltsev, N., Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res*. **28**:123-125.

[12] Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28**:27-30.

[13] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res*. **28** (1):289-91.

[14] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., Hogue, C. W. (2001). BIND : The Biomolecular Interaction Network Database. *Nucleic Acids Res*. **29** (1):242-5.

[15] Huerta A. M., Salgado, H., Thieffry, D., Collado-Vides, J. (1998). RegulonDB: a database on transcriptional regulation in *Escherichia coli. Nucleic Acids Res*. **26**:55-59.

[16] Takai-Igarashi, T., Nadaoka, Y., Kaminuma, T. (1998). A database for cell signaling networks. *J. Comput. Biol*. **5**:747-754.

[17] Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pr, M., Reuter, I., Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*. **28**:316-319.

[18] van Helden, J., Naim, A., Lemer, C., Mancuso, R., Eldridge, M., Wodak, S. (2001). From molecular activities and processes to biological function. *Briefings in Bioinformatics* **2** (1):98-93.

[19] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9** (12):3273-97.

[20] van Helden, J., Wernisch, L., Gilbert, D., Wodak, S. J. (2002), Graph-based analysis of metabolic networks. Ernst Schering Research Foundation Workshop 39, Bioinformatics and Genome Analysis.

[21] Mewes, H. W., Seidel, H., Weiss, B. Editors, *Springer* 245-274.

[22] Rison, S. C. G., Hodgman, T. C., Thornton, J. M. (2000). Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics* **1**:56-69.

[23] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285** (5428):751-3.

[24] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature* **402** (6757):83-6.

[25] Enright, A. J., Iliopoulos, I., Kyrpides, N. C., Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*. **402** (6757):86-90.

[26] Tsoka, S. & Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat. Genet*. **26** (2):141-2.

[27] Schomburg, D., Salzmann, D., Stephan, D. (1990-1995). *Enzyme handbook*. 13 vols, Springer.