Representing and analysing molecular and cellular function in the computer

Preprint version of the article published in Biol Chem 381(9-10), 921-35.

Jacques van Helden[1,2], Avi Naim[1], Renato Mancuso[1], Mattew Eldridge[1,&], Lorenz Wernisch[1,§], David Gilbert[1,3] and Shoshana J. Wodak[1,2*]

[1] European Bioinformatics Institute (EBI). Genome Campus - Hinxton Cambridge CB10 1SD - UK. Email: {jvanheld, drg, wernisch, shosh@ebi.ac.uk

[2] Unité de Conformation des Macromolécules Biologiques. Université Libre de Bruxelles.50 av. F.D. Roosevelt. B-1050 Bruxelles. Belgium. Email: shosh@ucmb.ulb.ac.be

[3] Department of Computing, City University, Northampton Square, London EC1V 0HB, UK. Email: drg@cs.city.ac.uk

[§] Present address: School of Crystallography. Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK. Email: l.wernisch@mail.cryst.bbk.ac.uk

[&] Present address: Synomics Ltd,Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9AD, Email: Eldridge@ synomics.com

* Corresponding Author: Shoshana J. Wodak email: shosh@ucmb.ulb.ac.be;Shosh@ebi.ac.uk

Abstract

Determining the biological function of a myriad of genes, and understanding how they interact to yield a living cell, is the major challenge of the post genome-sequencing era. The complexity of biological systems is such, that this cannot be envisaged without the help of powerful computer systems capable of representing and analysing the intricate networks of physical and functional interactions between the different cellular components. In this paper we try to provide the reader with an appreciation of where we stand in this regard. We discuss some of the inherent problems in describing the different facets of biological function, give an overview of how information on function is currently represented in the major biological databases, and describe different systems for organising and categorising the functions of gene products. In a second part, we present a new general data model, currently under development, which describes information on molecular function and cellular processes in a rigorous manner. The model is capable of representing a large variety of biochemical processes, including metabolic pathways, regulation of gene expression and signal transduction. It also incorporates taxonomies for categorising molecular entities and interactions, it offers means of viewing the information at different levels of resolution, and dealing with incomplete knowledge. The data model has been implemented in the database on protein function and cellular processes 'aMAZE' (http://www.ebi.ac.uk/research/pfbp/), which presently covers metabolic pathways and gene regulation. Several tools for querying, displaying, and performing analyses on such pathways are briefly described in order to illustrate the practical applications enabled by the model.

Keywords: Databases, metabolic, networks, pathways, function, genomics

Introduction

The availability of a growing number of completely sequenced genomes is changing the way in which research in biology is performed. So far, investigations focused mainly on the properties of single genes and proteins. With the imminent availability of the complete sequences from 50 genomes ranging from bacteria to human, it will become possible to study how the individual genes and gene products co-operate to give rise to complex cellular processes. This holds the promise of greatly facilitating the identification of targets for drugs that will cure human aliments, and of harnessing cellular processes to serve man's needs.

But before this vast potential can be exploited, the genome sequence information need to be decoded in terms of the biological function of the gene products. This as it turns out, is a much more ambitious and challenging task than the sequencing effort itself. This is due to several reasons, not the least of them being the complexity of the very notion of biological function.

First of all, biological function may be described at different levels. When we investigate the functional properties of the isolated protein, we are dealing with molecular function. For an enzyme such as aspartokinase, the molecular function is catalysis of a given reaction -the phosphorylation of aspartate- characterised by its substrate specificity, catalytic constants and mechanism. When studying this enzyme in the cellular context we are interested in the cellular process in which it is involved – its cellular function. The documented cellular function of aspartokinase in *bacteria*, is catalysis of the first step in the common biosynthetic pathway leading from asparagine to methionine, lysine, and threonine, respectively.

Second, it is not uncommon to find that a protein carrying out a given molecular function is involved in very different cellular processes. For example, the drospophila transcription factor Scute, is involved in both sex determination and positioning of neural precursors, two apparently unrelated processes. The picture is further complicated by the frequent observation that the cellular and molecular functions of a given protein can both vary with its cellular localisation, the cell type in which it is expressed, its oligomerization state, or with the change in the cellular concentration of effector molecules (Jeffery, 1999).

Characterising the biological function of a gene product thus requires combining information from different levels of the molecular and cellular organisation. Classically, this undertaking has been achieved by systematic biological and biochemical experimentation. But, with the increasing realisation that there probably is a limited repertoire of genes and proteins shared by most or possibly all organisms, and mounting evidence that shared proteins tend to carry out the same or similar functions in different organisms, predictive approaches based on inference have become widely used.

In their simplest form, they involve matching the newly determined sequence of each protein, taken individually, to the known sequences of proteins whose function has already been characterised and tentatively assigning the characterised function to the new protein (Hodgman, 2000). In more sophisticated versions functional links between proteins are also considered. In the so-called 'metabolic reconstruction' approach (Hodgman, 2000, Goryanin et al., 1999, Overbeek et al., 2000), one or more uncharacterised proteins are predicted to participate in a specific metabolic pathway if their homologues are known to do so from experiment (Marcotte et al., 1999b). In other methods, proteins whose homologues are fused into a single gene in another organism are predicted to be physically or functionally linked (Marcotte et al., 1999a, Marcotte et al., 1999b, Enright et al., 1999). Attempts have also been made to infer functional links between proteins on the basis of a common pattern of evolution in different organisms (Pellegrini et al., 1999), or based on the information that the corresponding genes may be co-regulated in a given organism (Hughes et al., 2000).

In parallel, efforts to develop new experimental procedures for large scale functional characterisation of gene products have also been undertaken. These include high-throughput techniques such as micro-array based gene expression analysis (Brown and Botstein, 1999, DeRisi et al., 1997), two-hybrid interaction screens (Transy and Legrain, 1995, Colas and Brent, 1998, Uetz et al., 2000) and the techniques for analysing expressed proteins (the 'proteome'), which combine 2D gel electrophoresis with automatic N-terminal sequence analysis by Edman degradation or mass spectroscopy (Williams, 1999). All these procedures generate enormous amounts of data, which in turn, need to be interpreted on the basis of available knowledge on the molecular and cellular function of genes in a continuous bootstrap approach within and across organisms.

These recent developments bring to the forefront the pressing need to use the full power of computers in order to develop better tools and systems for representing and analysing information on biological function.

The present paper focuses on this important problem. In a first part we review the ways in which molecular and cellular functions of genes and proteins are currently represented in the major biological databases. This includes a brief discussion of the various specialised databases representing information on metabolic pathways, gene regulation and signal transduction, and of several efforts to catalogue and categorise biological function and processes.

In a second part we describe a general data model for representing information on biological function in the computer, currently under development in our laboratories. This model describes the physical and functional interactions between genes and proteins as forming a large complex network. The network is constructed from two basic classes of objects, biological 'entities' (typically molecules) and 'interactions' (for example, a chemical reaction, catalysis or inhibition). The latter can link entities to other entities, entities to interactions, or interactions to other interactions. This allows the description of a large variety of biochemical processes. In addition, taxonomies are used to describe the relationships between generic and specific biological entities, as well as between generic and specific interactions, offering means of viewing the information at different levels of resolution, and dealing with incomplete knowledge.

This data model is being used to develop the database on protein function and cellular processes 'aMAZE' (http://www.ebi.ac.uk/research/pfbp/), which currently covers metabolic and regulatory pathways. Several tools for querying, displaying, and performing analyses on such pathways are briefly described in order to illustrate the practical applications enabled by the model.

Representation of function in biological databases

Information on function in sequence and structure databases

In the major biological databases, information is organised primarily on the basis of single biological entities. Typical examples are the EMBL (Baker et al., 2000) and GenBank databases (Benson et al., 1999), which contain one entry per gene, and SWISS-PROT (Bairoch and Apweiler, 2000), where each entry essentially corresponds an individual polypeptide. In the Protein Databank (PDB) (Berman et al., 2000), entries most commonly represent the structure of a single macromolecule, or that of a macromolecular complex (such as a multi-subunit protein, a protein-protein or protein-DNA complex).

Though these databases were created chiefly for storing information on primary, tertiary, and quaternary structure, a significant effort has also been made to include functional annotation. SWISS-PROT is certainly one of the database containing the most extensive and accurate information about protein function. This information pertains to both, molecular and cellular function. However, the organisation of these databases still reflects their primary focus, and information on function is stored in a few general description fields, essentially as free text. For example, the product of the *Escherichia coli* putA gene is described as a bifunctional enzyme in the "description" field of SWISS-PROT, whereas its third role as a repressor of transcription appears in the 'comment' field. This makes it rather difficult for any computer program to extract the information. For enzyme function, on the other hand, the situation is more favourable, with the hierarchic EC nomenclature system (EC number) being used to define and categorise catalysed reactions (Bairoch, 1993). But the relation between an EC number and a polypeptide is often ambiguous. Indeed, a given polypeptide can sometimes be characterised by several EC numbers, because it contains several domains, each catalysing a different reaction. Alternatively, the association of several different polypeptides may require to catalyse a given reaction, catalogued under a single EC number.

In parallel to the centralised sequence databases, independent resources have been developed for organism-specific genome projects. Some examples are listed in Table 1. These resources are also organised by entries corresponding to single genes, and contain information, which is sometimes redundant and sometimes complementary to the annotations in the centralised databases.

Two well-documented problems with the annotations in these databases need to be mentioned. Due to the mounting difficulty in generating experimental evidence for the rapidly increasing number of sequences, tentative functions assigned by inference are often used in turn to make further inference, leading to a rapid propagation of errors (Kyrpides and Ouzounis, 1999). The origin of such errors can however not be traced back easily, because no reference is given to how a given functional assignment was obtained

Another problem, recently amplified by the multiplicity of independent annotation efforts, is that different databases, and sometimes even different annotators of the same database, use different terms to describe the same biological molecule or process. This problem is being addressed in various ways, by the use of synonym tables, or by the definition of specialised vocabularies, hopefully agreed upon by the different databases.

Specialised databases

Driven by the need for a more detailed and precise description of molecular and cellular function, several specialised databases have been developed. Some have focused more specifically on molecular function, others on cellular function or on the integration of both levels of function. The most prominent specialised databases are listed in Table 2.

Molecular function

Several databases have been designed to represent in detail the function of a specific class of proteins. ENZYMES (Bairoch, 1993), BRENDA (http://www.brenda.uni-koeln.de), and EMP (Selkov et al., 1996, Overbeek et al., 2000) describe enzymes; YTPdb (André, 1999) describes yeast membrane transporters; CSNdb (Takai-Igarashi et al., 1998) is focussed on signalling pathways; Transfac (Wingender et al., 2000, Knuppel et al., 1994) and RegulonDB (Huerta et al., 1998) deal with transcriptional regulation.

Several of these databases are also organised around structural entities, but with a clearer separation between biological function and structure. For instance, Transfac defines proteins, genes and regulatory sites (cis-acting elements), and represents interactions as relationships between these different entities. In BRENDA and ENZYMES, each entry corresponds to a catalysed reaction, identified by its EC number, and groups information on all the enzymes known to catalyse this reaction. EMP is organised by literature citation.

In some more recent databases, such as that on cell signalling networks, CSNdb (Takai-Igarashi et al., 1998), information is explicitly split between two types of objects, describing molecules and interactions respectively. GIFdb (Genes Interacting in the Fly) (Jacq et al., 1997), focuses on protein-protein and protein-DNA interactions, and contains one record per interaction. The Database of Interacting Proteins (DIP) (Xenarios et al., 2000) stores experimentally determined protein-protein interactions, with each record describing one interaction between two proteins.

Cellular processes

In recent years several databases representing higher levels of functional organisation in the cell have been developed. These have primarily focused on metabolic pathways, which are among the best-documented cellular processes.

Three main databases on metabolic pathways must be mentioned. KEGG (Kyoto Encyclopaedia on Genes and Genomes) (Kanehisa and Goto, 2000), WIT/MPW (Overbeek et al., 2000), and EcoCyc/MetaCyc (Karp et al., 1996, Karp et al., 2000). In all of them, pathways are represented as collections of molecular functions. A typical example is KEGG, where generic metabolic pathways are stored as unordered collections of catalysed reactions. The order of the reactions is not stored within the pathway entry itself, but is provided by the graphical representations (maps). Organism-specific pathways are then defined by mapping into each generic pathway, the set of reactions catalysed by the relevant gene products of the selected organism. Information on the small molecule (compounds) involved in the pathways is stored in the sister database, LIGAND.

The same holds for MPW (Selkov et al., 1998), which consists primarily of a large collection of pathway graphs, associated with detailed annotations on individual pathway components (enzymes, substrates, products, etc...). EcoCyc is based on a more sophisticated hierarchic data organisation. It provides graphical representations of pathways, as well as literature-based annotation of gene and protein functions, and a detailed representation of small molecule compounds. Originally restricted to metabolic pathways in *Escherichia coli*, EcoCyc is currently being extended to other organisms and other types of pathways (MetaCyc),

The tools available for querying the information stored in these databases are in general rather rudimentary. They are essentially limited to browsing, visualising the pathway graphs, and retrieving information on genes, proteins, compounds and so on, on the basis of user specified criteria. KEGG however, allows to compute all the possible pathways between two specified compounds, and provides the output in a textual form. This is done using specialised tools operating directly on the unordered collections of stored enzymatic reactions. Analogously, in the cell signalling database CSNdb, which does not store pre-defined collections of molecules and functions, pathways can be calculated by interconnecting partners of successive interactions, and represented graphically using a hierarchical graph layout algorithm.

Categorising functions and processes

An important first step toward acquiring understanding of molecular and cellular function is to build systems for organising and categorising functions of gene products (for review see (Riley, 1998)). An attempt in this direction was made as early as 1983 for categorising the genes of E-coli. Further attempts to classify genes by their cellular function followed 10 and 13 years later (Riley, 1993, Riley, 1998). A similar functional catalogue has been defined by the Munich Information Center for Protein Sequences (MIPS) (Mewes et al., 2000) for the annotation of the yeast genome.

Further systems for describing information on molecular and cellular function have been built as part of the development of specialised databases such as EcoCyc and others. Some of these systems go a significant way towards describing the richness and complexity of biological information. Typically, they also include a vocabulary of terms and some specifications of their meaning, which can be used to ensure uniform annotation and to facilitate sharing information between different databases. In computer science similar systems are called 'ontologies' (Uschold et al., 1998), a term, adapted somewhat freely, from the field of philosophy. Reference to ontology in biology has been introduced only recently (for a recent review, see (Stevens et al., 2000), and different 'ontologies', developed for different purposes, can be mentioned as examples. These include the RiboWeb ontology (Altman et al., 1999) for the description of ribosome components and their function and the EcoCyc ontology, developed as part of the data structure underlying the EcoCyc database. The Ontology for Molecular Biology (OMB) (Schulze-Kremer, 1997, Schulze-Kremer, 1998) and the TAMBIS ontology (Baker et al., 1999), are attempts to standardise annotation and to enable asking questions over different molecular biology databases. The more recent Gene Ontology consortium (GO) (Ashburner et al., 2000) represents a coordinated effort of three organism-specific databases (drosophila, yeast and mouse). It defines three independent taxonomies to describe molecular function, cellular processes and cellular locations, respectively. Each gene can be annotated according to these three taxonomies.

Representing and analysing cellular processes in the computer: the aMAZE database

In order to achieve understanding of how living cells function, we need to go beyond defining and categorising function. Indeed, we must be able to represent the complex network of physical and functional interactions that take place in the living cell, in ways which enable us to manipulate and analyse it using the full power of computers. This requires rules, which define how the individual molecular functions and processes interconnect to form the network.

In what follows we describe a data model for representing and integrating different types of molecular functions and cellular processes in the computer, which has the ambition of adequately addressing several of the major requirements of the field. This data model has been recently implemented in the database on Protein Function and Biochemical Pathways (aMAZE), (http://www.ebi.ac.uk/research/pfbp/). As an illustration of the type of applications enabled by the database, we give examples of several software tools for analysing and manipulating metabolic and regulatory pathways stored in the database.

Goals and scope of the aMAZE database

The main remit of the aMAZE database has been as follows. It should include information on the molecular function of proteins and on organism- and tissue-specific biochemical processes, or pathways. The latter should cover metabolic pathways, gene regulation, transport, signal transduction and possibly more. The information should be organised in ways, which make possible the design of powerful and flexible software tools for querying the processes and analysing their properties.

Several issues were given special consideration in the aMAZE data model. One is the capacity to represent incomplete knowledge and subsequently add new knowledge in a flexible way. Another is the ability to distinguish between different types of evidence for the function assigned to a given gene product or its involvement in a given biological process. Of particular importance is the ability to distinguish between assertions based on experiments and those deduced by inference, and in case of information deduced by inference, the ability to trace back its origins is essential.

The aMAZE Data Model

Entity-relationship representation in an object-oriented approach

The entity-relationship model is a classical model for representing information in databases, and has been used in many different contexts for more than a decade (Moulin et al., 1976, Chen, 1976). Here, we illustrate its application to the description of biological function in ways that satisfy the requirements enumerated above. A particularity of the approach taken in aMAZE is that it exploits the full power of the entity-relation model, while using an Object Oriented representation to describe biological knowledge at all levels.

As shown in Figure 1, two main classes of objects are defined. The first class, *'biochemical entity'*, represents structural units. These can be complete molecules, such as metabolites or proteins. They can also correspond to parts of molecules, such as a gene or a regulatory sequence in a non-coding region (cis-acting element), or to supra-molecular assemblies (e.g. a protein complex, a ribosome). Theses objects have attributes, which describe their physical characteristics (sequence, compound formula, molecular weight and so on).

The second class represents relationships. The relationship objects considered here are not simple links between entities, as in many biological databases, but are objects in their own right. Each relationship object is characterised by a list of inputs and a list of outputs. In addition it may have a rich collection of attributes, which describe the properties of the relationship. Typical relationships defined in aMAZE are of the type 'interaction'. An 'interaction' can be a *'reaction'*, which converts a set of substrates (the input) into a set of products (the output). Another example is *'expression'*, which has a gene as input and a polypeptide as output. These interactions have in common the fact that both their input and output are sets of entities, and are denoted as 'transformations'. Both transformations are represented in the database as elementary interactions (they do not refer to other interactions), but can be used by the biologist to represent a series or interactions, for which only the first input and last output are described, as discussed below.

Complex cellular process (for example metabolic pathways), for which the intervening steps are described in the database, are represented using the 'process/pathway' objects. Each such object refers to the entities and interactions of all its intervening steps (Fig. 1a). A given process is thus described as a network of entities and interactions, whose connectivity is completely defined.

A particular feature of the aMAZE data model is that the 'interaction' objects have not only entities as input/output, but can have other interactions as output. This is the case for the object *'catalysis'*, which represents the action of a protein (enzyme) in accelerating a chemical reaction (Figure 2a). Similarly, *'inhibition'* is an interaction of a compound on another interaction, such as *'catalysis'* (Figure 2a). In all cases encountered so far, interactions having interactions as output, can be classified as *'control'* interactions, which either accelerate or slow down some other interactions.

Figures 2b and 2c illustrate how this simple model can be extended to gene regulation, and transport. *'Transcriptional regulation'* (Figure 2b) is represented as the positive (up-regulation) or negative (down-regulation) action of a protein (transcription factor) on the interaction *'expression'*. A *'translocation'* (Figure 2c) is an interaction characterised by the fact that it has as input and output a set of molecules, each associated with a location. Several molecules can be involved in the same translocation event (symport, antiport), just as several metabolites can be involved in the same chemical reaction. The input and output molecules are generally identical (except for catalytic transporters), but input and output locations differ. A transporter protein such as the lactose permease has an activity of *'transport facilitation'* on the *'translocation'* activity (Figure 2c). Similarly to *'catalysis'*, *'transcriptional regulation'* and *'transport facilitation'* can be activated or inhibited by metabolites.

As already alluded to above, our model enables the use of a shorthand notation at various levels in order to represent a complex process. Most obviously, the object '*process/pathway*' is a shorthand notation for an entire sub-section of the cellular network. At another level, the object *'reaction'* may represent a specific chemical reaction, or several successive reactions but for which only the first input and last output are specified. Similarly, *'expression'* is by definition a complex process, but which has specific types of molecules as input and output, respectively gene and polypeptide.

This has at least two important applications. One is to do away with irrelevant details, as in the case of *'expression'* where we are only interested in the input gene and output polypeptide whereas the detailed steps of RNA and protein synthesis are in most cases not relevant. There will however be some cases where one will be interested in a fine-grained description of the intermediate steps, with separated interactions for *'transcription'* and *'translation'*, in particular for genes that are spliced differentially, or to distinguish between transcriptional and translational regulation. Provision is made in the model to also represent the expanded information, whenever the need arises.

Another use of the shorthand notation is to represent incomplete knowledge. For example, when a molecule is known to be required for, or trigger, a certain cellular function, but the individual steps of the process are unknown. These can be filled in subsequently, when more data becomes available. A typical example is the transcriptional response to the presence/absence of some metabolite in the culture medium. This response can involve a transport system, a membrane sensor coupled to a signal transduction chain, and a transcriptional factor, all of which can be unknown at the time of the first experiment. It is however essential to represent the fact that a set of genes are known to respond to a given metabolite. We store this information in a specific class called *'indirect interaction'* (Figure 1b).

An important feature of the aMAZE data model, is that it defines the role of a particular structural entity (here compound or protein) within a functional context, rather than associating it with the entity itself. This makes it easy to represent multiple functions for the same molecule. Thus, a single protein can be a catalyst of different reactions (e.g. multifunctional enzymes) with each reaction having a different EC-number, or even play an entirely different role in another context. A compound acting as input for an *'inhibition'* activity (Figure 2a) is an inhibitor in that context, but could well play a different role in another context (substrate of another reaction).

Cellular processes as object graphs

In the graphical representation that illustrates the entity/relationship model (Figs 2), text labels are used to display entity objects such as compounds, genes and proteins, and labelled boxes are used to display interaction objects, such as expression, catalysis, inhibition etc. Interaction boxes are connected to their respective inputs and outputs through coloured lines. Cellular processes, such as metabolic and regulatory pathways, are presented by displaying the set of objects to which they refer, and their interconnections.

Since we can hope to ultimately store hundreds of thousands of entity and interaction objects in the computer, we should be able to describe all the processes that take place in a cell, and even between cells, as well as all the interconnections between these processes. This ensemble forms a vast network that we will be able to flexibly analyse and display.

Classical biochemical pathways represent defined portions of the network, usually considered as separate functional modules by the biochemists. The ability to represent the global network in the computer will allow us to represent the connections between different pathways, and to partition the network in any way we wish, offering the possibility exploring different definitions of functional modules.

Figure 3 shows a graphical representation of the methionine biosynthesis pathway using the objects presented above for describing entities and interactions. It shows an intricate network that links entities and interactions. This network includes not only the succession of chemical reactions that lead to the transformation of L-Aspartate to S-Adensyl-L-Methionine, but also the regulation of gene expression and enzymatic activities. It furthermore displays (in green) the links to other pathways, which are not detailed on the graph to preserve clarity.

It is important to realise that in aMAZE, the network exists not only as a picture, but is represented in the computer as an object graph, with entities and interactions forming the graph nodes. These nodes are connected to each other through their inputs and outputs, which form the graph arcs. The pathway diagram is thus merely a means for displaying the stored information in a manner familiar to the biologists. This requires navigating through the network while disabling certain connections which the biologist would consider as 'trivial' (f.e. those involving water molecules in metabolic pathways). Once this processing is done, layout programs can be used to draw the diagrams automatically (van Helden et al., 1999).

Generic objects and taxonomies

It is quite clear that the description of the molecular and cellular function must also include systems for defining and classifying biological entities and interactions. This is necessary in order to enable successive levels of generalisations. Those are useful in order to group objects for analysis purposes as well as for handling incomplete knowledge. Such hierarchical organisations must as much as possible incorporate the biologist's view, and hence reflect existing taxonomies of molecular and cellular function, as well as of cellular locations, tissues, organisms etc. There hence is an important intersection between data models such as that presented here and biological ontologies, such as the Gene Ontology (GO) (Ashburner et al., 2000). But the data model as a whole has a much wider scope, since it also described the interactions between the biological entities, activities and processes, which the so-called ontologies do not do.

To represent taxonomies, the aMAZE data model uses objects describing specific and generic biochemical entities, interactions and locations, respectively. A generic entity merely consists of a list of references to other entities. The referred entities can themselves be either specific or generic, extending the description to a full taxonomy.

Thus, each subclass of our *'location'* class (compartments, cell types, tissues, organs, organisms), shown in Fig. 1, can contain specific or generic objects. This hierarchy of objects can be used to represent taxonomies such as those of the Gene Ontology.

Documentation of the stored information

Providing means to trace back any piece of data to its source, is an essential requirement for any repository of biological information. In aMAZE a separate set of objects is defined, allowing a flexible description of different types of references (literature, Web, in-house references). This permits to associate multiple references with the same entity or interaction. Furthermore, references can refer to different lines of evidence to which reliability scores can be attached. The danger of treating at the same level information based on lines of evidence with different reliability scores, such as experimentation and theoretical predictions, is eliminated. Indeed, database users can select a subset of interactions and entities on the basis of the type of evidence and reliability score, according to their needs.

Querying and analysing networks of cellular function

The power of a data model lies in the query and analysis capabilities that it enables. In the following we give some examples of query methods currently implemented. These methods involve primarily path navigation routines.

A simple query is to get all the reactions catalysed by a gene product (Figure 4a). The answer is obtained by collecting the polypeptide(s) coded for by the gene, then the proteins that are assembled from the polypeptides, and finally the set of reactions that are catalysed by these proteins. This approach has the advantage of taking into consideration the multiple relationships that occur at each level. Indeed, a gene can produce several polypeptides if there is differential splicing, a polypeptide can be involved in the formation of distinct proteins by forming complexes with alternative partners, and a protein can catalyse several reactions.

More complex queries require the application of specialised graph analysis algorithms. These are for example:

- find all metabolic pathways that convert compound A into compound B in less than X steps (Fig. 4b).
- find all genes whose expression is directly or indirectly affected by a given compound.
- find all compounds that can be synthesised from a given precursor in less than X steps
- in the complete set of metabolic reactions, find all feedback loops including a given compound, or, in a defined biochemical pathway, find all feedback loops.

Another type of complex queries, is sub-graph extraction. Here the user specifies a set of nodes in the network -the 'seed' nodes (usually genes/proteins)- and prompts the system to extract the portions of the network or sub-graphs that interconnect each pair of seed nodes via the smallest number of individual links, as shown in Fig. 5. The user can specify the maximum number of individual links, or graph arcs, that can be inserted between any two seed nodes. The resulting sub-graph can then be displayed and analysed.

The algorithms for sub graph extraction and maximal path enumeration used in this context, have been described elsewhere (van Helden *et al.,* 1999 and 2000).

Interpreting results from DNA micro-array experiments.

One application of the sub graph extraction algorithms is to identify the interconnectivity of a cluster of functionally related genes, such as those identified by DNA micro array experiments (Brown and Botstein, 1999, DeRisi et al., 1997), or by theoretical predictions (Marcotte et al., 1999a, Marcotte et al., 1999b, Enright et al., 1999).

Enzymes and transporters participating in a common metabolic pathway are often co-regulated at the transcriptional level. Therefore, when the culture medium is modified by depleting (or adding) a given metabolite, it is expected that the genes that participate in the biosynthesis (or degradation) of the molecule will respond at the transcriptional level, and will appear as a cluster of co-regulated genes. When such a cluster is identified by DNA micro array experiments, the pathway or pathways in which the corresponding genes participate may not be immediately obvious. The graph/sub graph extraction algorithm can then be used to identify these pathways.

This is done by first using the type of simple queries illustrated in Fig. 4a in order to select the genes that code for enzymes and to find the reactions that they could catalyse. These reactions correspond to the seed nodes in the graph of all possible metabolic reactions (Fig. 5A). The method consists in trying to interconnect all these reactions in a meaningful way (Figure 5B), in order to extract a sub-graph (Figure 5C) corresponding to one or several putative metabolic pathways (Figure 5D).

The simplest way to interconnect reactions is to identify compounds that are produced by one reaction and consumed by another. In a second step, linking can be improved by intercalating reactions that were not part of the initial set. This has several justifications. Firstly, some genes could be involved a metabolic pathway without being regulated at the transcriptional level. Secondly, with the still limited reliability of DNA micro array experiments, the co-regulation of some gene might escape detection. Thirdly, a gene identified as co-regulated in a given experiment may not have been annotated as an enzyme yet

Once the sub graph corresponding to the putative pathway has been extracted, it can be compared to the set of known metabolic pathways stored in the database. In some cases it will correspond to a previously characterised pathway. For these, a simple matching of the set of reactions against a database of metabolic pathways would have yielded the same answer. In other cases, one might observe only a partial match with a known pathway. This may lead to the discovery of new variants of known pathways, and could hence be useful in mapping the metabolic pathways of newly sequenced organisms, whose metabolism has not been fully characterised. Finally, there is the attractive possibility that such an analysis may lead to the discovery of completely new pathways. The co-regulation of the enzyme-coding genes would provide a good support to indicate that this pathway is biologically relevant. An interesting field of application could be to discover metabolic pathways involved in largely unexplored processes, such as resistance to toxic compounds or extreme conditions, displayed by certain bacteria.

As a very preliminary test, we applied the above procedure to the 20 genes belonging to the MET cluster defined by Spellman *et al.,* (1998). Seven of these genes code for enzymes, which can catalyse 8 distinct reactions. Subgraph extraction and maximal path enumeration resulted in a linear pathway including 6 of the initial reactions. In this case, a linear path could be obtained without intercalating any reaction that was not part of the initial set.

The extracted pathway (Fig. 6A) shows partial matches with two distinct metabolic pathways: the 4 initial steps match the sulfur assimilation pathway (Fig. 6B), and perform a progressive reduction of sulfate into sulfide. The two last steps match the methionine biosynthesis pathway, and correspond to the incorporation of sulfur into homocysteine (Fig. 6C), and the transformation of the latter into methionine.

Concluding remarks

In this paper we present a general system, or data model, for describing information on molecular function and cellular processes in a rigorous manner. We argue that such data model goes beyond defining and categorising molecular and cellular function, an important task performed by the so-called biological 'ontologies', because it provides rules for describing the physical and functional interactions between genes and gene products. This is best illustrated by analogy with language (Bray, 1997). The biological ontologies, by defining controlled vocabularies for the molecular and cellular functions and the hierarchic relations between them, define the semantics of the language of biological function. But they provide no syntax or grammar rules, which define how words can be associated to form sentences. Such rules are however required to 'speak' the language of biological function, and all the power of this language will be necessary if we want computers to help us unravel the complexity of living cells. This is precisely the ambition of the data model presented here.

This data model is still evolving as our limited experience grows and the types of data we handle expands. The aMAZE database, which implements this model, currently handles information on metabolic pathways and gene regulation, and contains data on about 6000 compounds, and 20,000 genes, 15,000 polypeptides, and 5,200 reactions, primarily from *E.coli* and yeast. The database is implemented using a Java front end and the ObjectStore Object Oriented database management system, as the back end.

In the near future it will be extended to include information on other types of pathways, in particular T-cell signal transduction, and will contain information on tens of thousands of genes, proteins, small molecule compounds, and their interactions, describing hundreds of cellular processes.

Admittedly, databases of this complexity would have been inconceivable only a decade ago, but are becoming reality thanks to recent progress in computer science, software development and network communication. We furthermore believe that such databases and the systematic analysis methods that they are enabling, will become as essential and as useful a tool for biology, as molecular modelling software and computer simulation methods have become for the study of complex macromolecules.

A key contribution of our model is the general rules that it provides for associating the individual biological entities and interactions into large complex networks of cellular processes. A powerful features of these rules, is that information on molecular function is not tied directly to the biological entity (macromolecule or macromolecular complex), as in most other databases, but is derived from the network of relations between the entities and interactions. For example a protein is a catalyst only when it is connected by a 'catalysis' interaction to a reaction that transforms a set of substrates to a set of products. The same protein can be a repressor, when it is connected by a 'repression' interaction to a gene 'expression' interaction.

We thus foresee that in the future, the function, or more likely the functions, of a gene product will not be obtained by looking up catalogues, but will be computed on the fly from the stored network of interactions in which the gene products are known to take part. Then however, our ability to describe gene functions in a comprehensive manner will strongly depend on how efficient we are in collecting reliable data on the interactions and activities in which the genes and their products are involved. Clearly, the task that lies ahead in this respect is immense.

Acknowledgements

We thank Georges Cohen and Kirill Degtyarenko for valuable help in annotation of metabolic and regulatory pathways, and acknowledge Chris Lemer, Thure Etzold and Dietmar Schomburg, for useful discussion. The work described here has been sponsored by a consortium of industries, comprising, Aventis, Monsanto, Organon, Roche and Zeneca. We thank scientists from these companies for valuable input.

References

Altman R., Bada, M., Chai, X. J., Whirl Carillo, M., Chen, R. O. and Abernethy, N. F. (1999). RiboWeb: An Ontology-Based System for Collaborative Molecular Biology. IEEE Intelligent Systems *14*, 68-76.

André B. (1999). The yeast Transport Protein Database. Current Genetics *35*, 278-278.

Ashburner M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology [In Process Citation]. Nat. Genet. *25*, 25-29.

Bairoch A. (1993). The ENZYME data bank. Nucleic Acids Res. *21*, 3155-3156.

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

Baker P. G., Goble, C. A., Bechhofer, S., Paton, N. W., Stevens, R. and Brass, A. (1999). An Ongology for Bioinformatics Applications. Bioinformatics *15*, 510-520.

Baker W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M. A. (2000). The EMBL Nuccleotide sequence database: contributing and accessing data. Nucleic Acids Res. *28*, 19-23.

Benson D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A. and Wheeler, D. L. (1999). GenBank. Nucleic Acids Res. *27*, 12-17.

Berman H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235-242.

Bray D. (1997). Reductionism for biochemists: how to survive the protein jungle. Trends Biochem. Sci. *22*, 325-326.

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

Chen P. P. (1976). The Entity-Relationship Model: Towards a Unified View of Data. ACM Transactions on Database Systems *1*.

Colas P. and Brent, R. (1998). The impact of two-hybrid and related methods on biotechnology. Trends Biotechnol. *16*, 355-363.

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

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

Goryanin I., Hodgman, T. C. and Selkov, E. (1999). Mathematical simulation and analysis of cellular metabolism and regulation. Bioinformatics *15*, 749-758.

Hodgman, T. C. (2000). A historical perspective on gene/protein functional assignment. Bioinformatics *16*, 10-15.

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

Hughes J. D., Estep, P. W., Tavazoie, S. and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J. Mol. Biol. *296*, 1205-1214.

Jacq B., Horn, F., Janody, F., Gompel, N., Serralbo, O., Mohr, E., Leroy, C., Bellon, B., Fasano, L., Laurenti, P. and Roder, L. (1997). GIF-DB, a WWW database on gene interactions involved in Drosophila melanogaster development. Nucleic Acids Res. *25*, 67-71.

Jeffery C. J. (1999). Moonlighting proteins. Trends Biochem. Sci. *24*, 8-11.

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

Karp P. D., Riley, M., Paley, S. M. and Pelligrini-Toole, A. (1996). EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res. *24*, 32-39.

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

Knuppel R., Dietze, P., Lehnberg, W., Frech, K. and Wingender, E. (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol. *1*, 191-198.

Kyrpides N. C. and Ouzounis, C. A. (1999). Whole-genome sequence annotation: 'Going wrong with confidence'. Mol. Microbiol. *32*, 886-887.

Marcotte E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751-753.

Marcotte E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. Nature *402*, 83-86.

Mewes H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Sch#ller, C., Stocker, S. and Weil, B. (2000). MIPS: a database for genomes and protein sequences. Nucleic Acids Res. *28*, 37-40.

Moulin P., Randon, J., Savoysky, S., Spaccapietra, S., Tardieu, H. and Teboul, M. (1976). Conceptual model as database design tool. In *Proceedings of the IFIP Working conference on Modelling in Database Management Systems* (ed. G. M. Nijssen), North-Holland.

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

Pellegrini M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U S A *96*, 4285-4288.

Riley M. (1993). Functions of the gene products of Escherichia coli. Microbiol. Rev. *57*, 862-952.

Riley M. (1998). Systems for categorizing functions of gene products. Curr. Opin. Struct. Biol. *8*, 388-392.

Schulze-Kremer S. (1997). Adding semantics to genome databases: towards an ontology for molecular biology. ISMB *5*, 272-275.

Schulze-Kremer, S. (1998). Ontologies for molecular biology. Pac. Symp. Biocomput. *24*, 695-706.

Selkov E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr. and Yunus, I. (1996). The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. Nucleic Acids Res. *24*, 26-28.

Selkov E., Jr., Grechkin, Y., Mikhailova, N. and Selkov, E. (1998). MPW: the Metabolic Pathways Database. Nucleic Acids Res. *26*, 43-45.

Stevens R., Goble, C. A. and Bechhofer, S. (2000). Ontology-based Knoweldge representation for Bioinformatics. Bioinformatics *in press*.

Spellman P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) *Mol. Biol. Cell*, **9**(12), 3273-97.

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

Transy C. and Legrain, P. (1995). The two-hybrid: an in vivo protein-protein interaction assay. Mol. Biol. Rep. *21*, 119-127.

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. and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae [see comments]. Nature *403*, 623-627.

Uschold M., King, M., Moralee, S. and Zorgios, Y. (1998). The Enterprise Ontology. The Knowledge Engineering Review. Special Issue on Putting Ontologies to Use *13*.

van Helden J., Gilbert, D., Wernisch, L. and Wodak, S. Graph-based analysis of biochemical networks (submitted)

van Helden J., Gilbert, D. R., Wernisch, L. and Wodak, S. (1999) *ISMB*.

Williams K. L. (1999). Genomes and proteomes: towards a multidimensional view of biology. Electrophoresis *20***,** 678-688.

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

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. and Eisenberg, D. (2000). DIP: the Database of Interacting Proteins. Nucleic Acids Res. *28***,** 289-291.

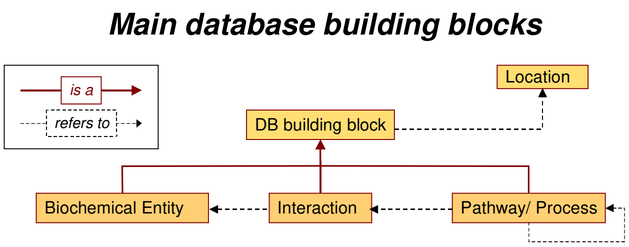
Figures

Figure 1: Schematic representation of the aMAZE data model

The aMAZE data model is based on the classical entity-relationship representation, and uses an Object Oriented representation to describe biological knowledge at all levels.

(a) Two main classes of objects are used. The ‘biochemical entity’ class, and the ‘interaction’ class. Both classes of objects refer to the ‘location’ class, which stores information on their intra- or extra- cellular location (see Fig. 1b). The pathway/process class, represents complex cellular process (for example metabolic pathways), for which the intervening steps are described in the database. Each such object refers to the entities and interactions of all its intervening steps, and each of those refer to a location, stored in the ‘location’ class (see also Fig. 1b).

(b) The ‘biochemcial entity’ class represents structural units, which can be complete molecules, such as metabolites or proteins, or parts of molecules such as a gene or a regulatory sequence. The ‘interaction’ class groups two main classes of interaction objects. The class ‘transformation’, and the class ‘control’ (see text for details). The ‘location’ class stores information on where the entities and interactions take place. The class ‘indirect interaction’ is used represent complex processes for which only the input and output are specified, but the intervening steps are not known (see text and Fig. 3).



(a)

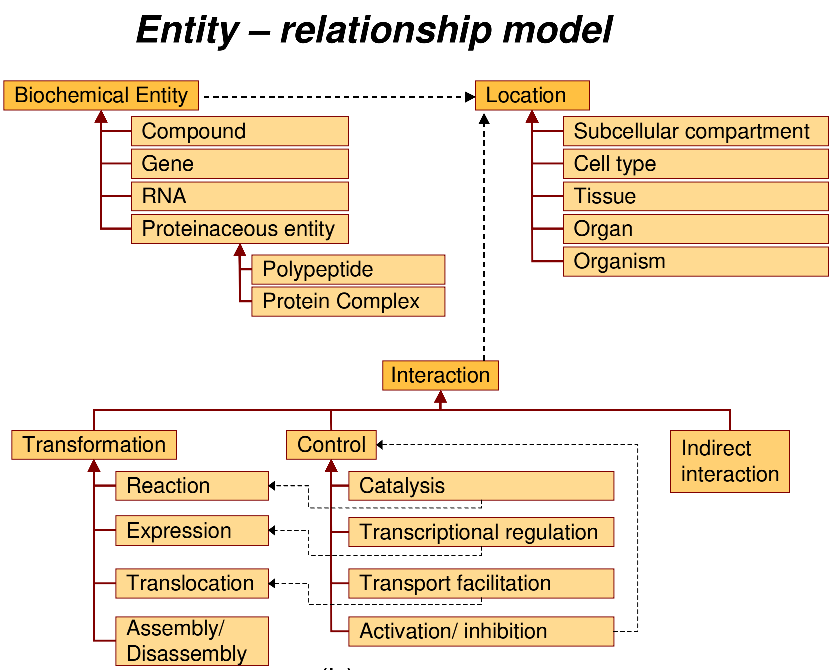


Figure 1 (b)

Figure 2: Using the aMAZE model to describe biological processes

(a) Enzymatic catalysis and its regulation. The example used is the catalysis of the phosphorylation of glutamate by gama-glutamyl kinase, regulated by proline. The displayed interaction objects, are ‘reaction’ , ‘catalysis’ and ‘inhibition’ . The ‘catalysis’ object has the EC number of the catalysed reaction (2.7.2.11) indicated. The compound names are marked as labels. The shown lines link each interaction object to their inputs and outputs. The graphical layout follows the convention of biologists.

(b) Transcriptional regulation. The example used is the regulation of the met6 and metB genes, by the combined actions of the products of other genes in the methionine biosynthesis pathway (the met4/met28/cbf1 protein complex, and the METJ protein respectively) and the end product of the methionine pathway S-adenosyl-Lmethionine). The depicted ‘interaction’ objects are, ‘up- and down-regulation’, ‘inhibition’, ‘activation’, and ‘expression’.

(c) Representation of the process of transport facilitation. The represented example is the transport of lactose from the extra-cellular space to the cytoplasm, which involves a translocation process, facilitated by a protein, the lactose permease. The information of the localisation of the different entities is stored in the ‘location’ objects (Fig. 1a,b).

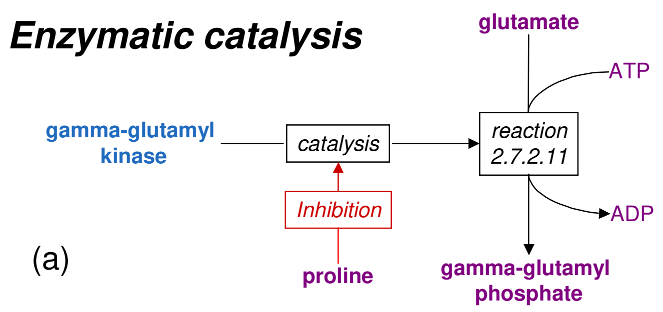


Figure 2

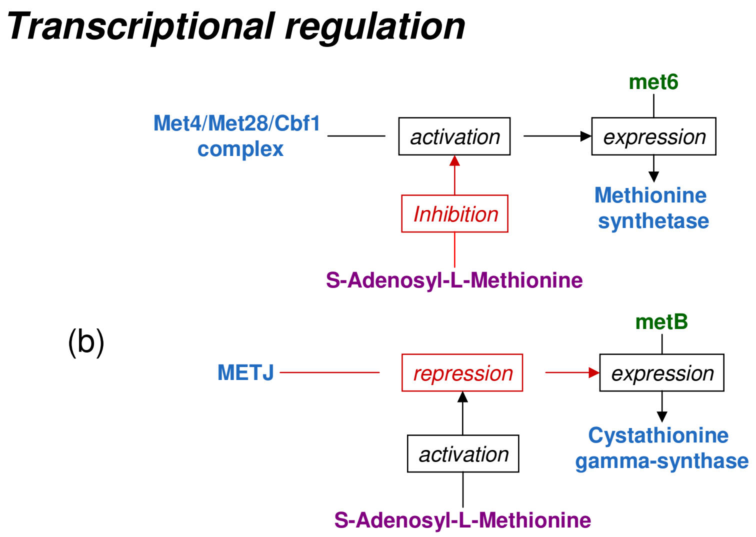


Figure 2

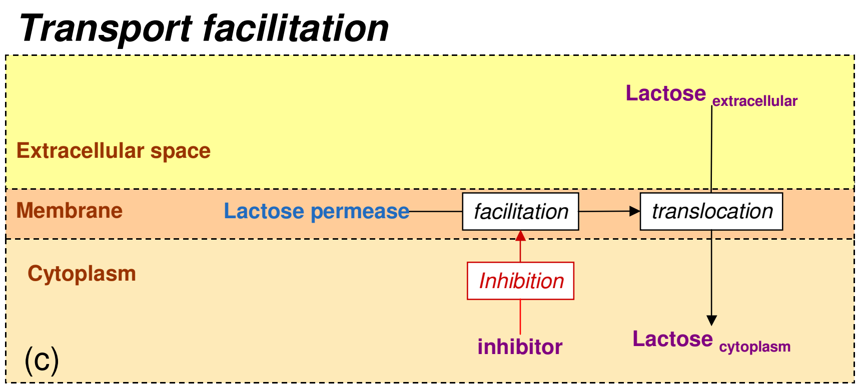


Figure 2

Figure 3: Graphic representation of the metabolic regulation network of methionine biosynthesis, using the aMAZE data model.

This figure illustrates how the various entity and interaction objects of the model are used to describe a complex biological process, such as methionine biosynthesis. In this graphic representation gene and proteins are displayed by their specific icons and labels. Green boxes on the upper right hand side of this figure indicate the names of other pathways, which use metabolites from this pathway, as inputs or outputs.

Methionine Biosynthesis in E.coli

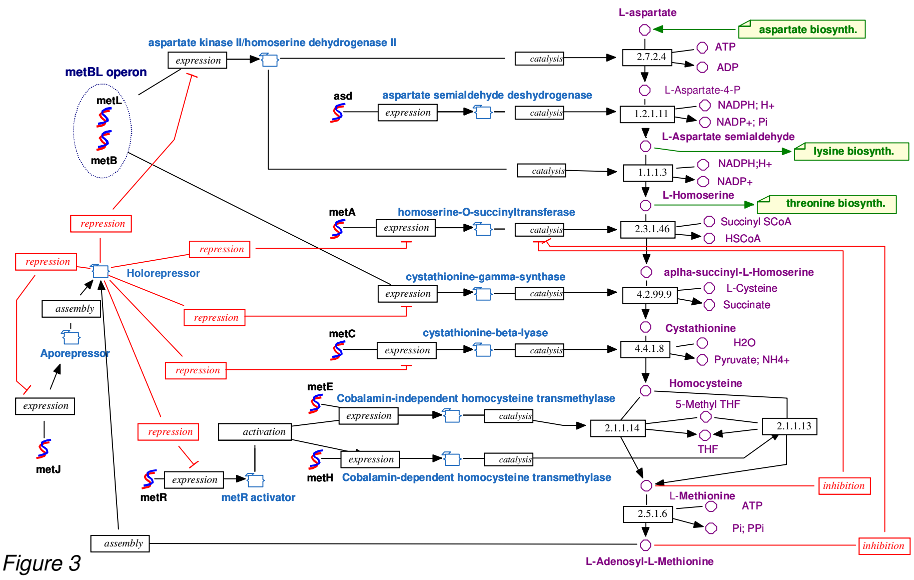
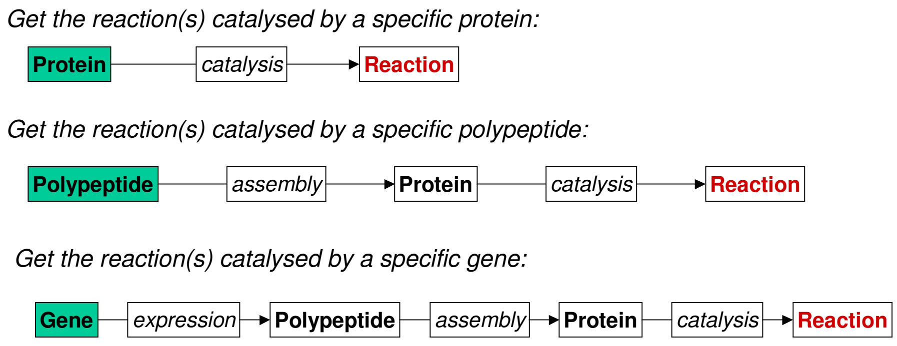


Figure 3

Figure 4: Schematic illustration of queries enabled by the aMAZE database

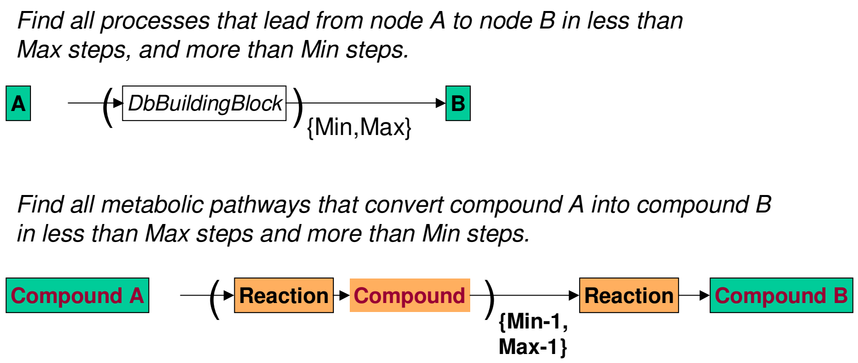
(a) Examples of simple queries: Get the reaction(s) catalysed by a specified protein; get the reactions catalysed by a specified polypeptide; get the reaction(s) catalysed by a given gene product, specified by its corresponding genes. The connectivity that needs to be established through the entities and interactions stored in the database, is displayed for each of the query.

(b) Examples of more complex queries that require the use of specialised graph analysis algorithms. The top query represents the following operation. Given two nodes, one specified as input and the other as output (with each node being either a biological entity or an interaction) find all the processes that lead from the input to the output in less than Max steps and more than Min steps. At the bottom, is illustrated the same query, but for which the search is constrained to consider only 'compound' objects as input and output nodes, and 'reaction' and 'compound' as intermediate nodes



(a)

Figure 4



(b)

Figure 4

Figure 5: Illustration of the procedure of sub-graph extraction from a complex network graph.

A, shows the position of the seed reactions in the graph. **B**, shows how the seed nodes are linked directly, via their inputs and outputs, or indirectly, via intercalated nodes. **C,** displays the resulting sub graph, and **D** the various linear paths that can be followed within the sub graph.

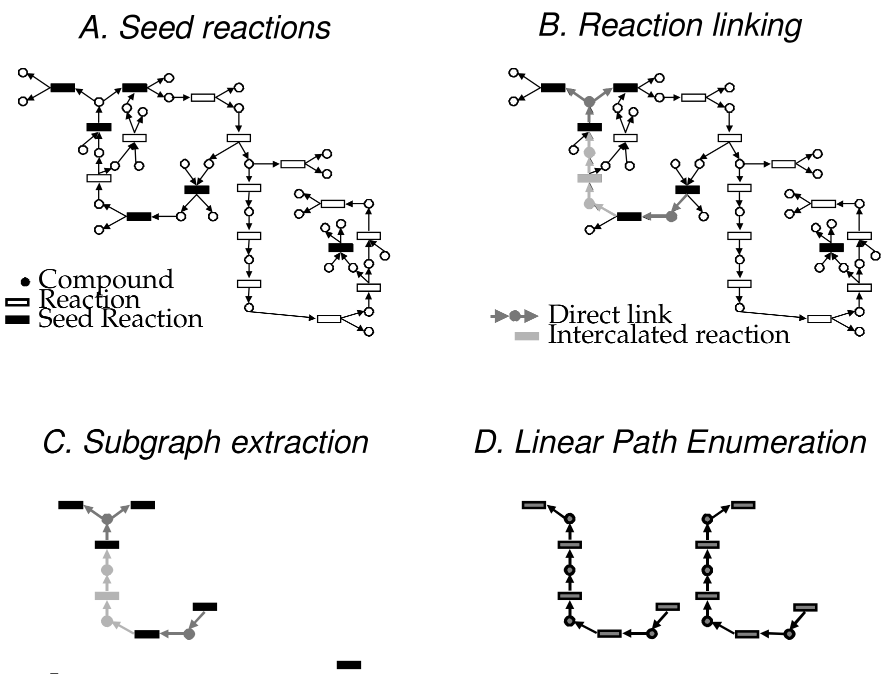


Figure 5

Figure 6: Result obtained by applying the sub-graph extraction analysis to the cell-cycle regulated gene cluster MET from Spellman *et al.* (1998) .

A. Pathway extracted by interconnecting the reactions catalysed by the 7 enzymes from the cluster. **B.** Sulfur assimilation pathway in yeast. **C.** Methionine biosynthesis in yeast.

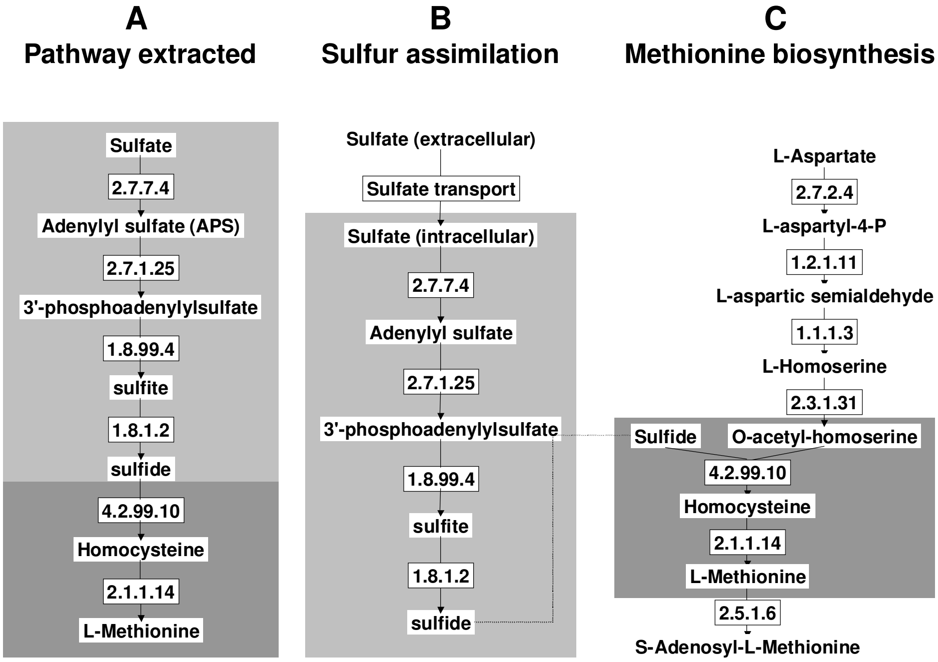


Figure 6

General sequence and structure databases

Acronym	Full Name	Scope	URL
EMBL	European Molecular Biology Laboratory	Nucleotide sequence database	http://www.ebi.ac.uk/embl/index.html
GenBank (NCBI)		Nucleotide sequence database	http://www.ncbi.nlm.nih.gov/Genbank/index.html
SWISS-PROT		Protein sequence database	http://www.expasy.ch/sprot/; http://www.ebi.ac.uk/swissprot/
PDB	Protein Data Bank	3D structures of macromolecules	http://www.rcsb.org/pdb/

Genome databases

Acronym	Full Name	Scope	URL
MGD	Mouse Genome Database	Mouse	http://www.jax.org/
Flybase		***Drosophila melanogaster***	http://fly.ebi.ac.uk:7081/
Sequencing Projects at Sanger Center		*Caenorhabditis elegans*, *Homo sapiens* + other organisms	http://www.sanger.ac.uk/Projects/
MIPS	Munich Information Center for Protein Sequences	*Saccharomyces cerevisiae*, *Arabidopsis thaliana*	http://www.mips.biochem.mpg.de/
SGD	Stanford Genome Database	*Saccharomyces cerevisiae*, *Arabidopsis thaliana*	http://genome-www.stanford.edu/
ECDC	E.coli database collection	***Escherichia coli***	http://susi.bio.uni-giessen.de/ecdc/ecdc.html

Integrated database

Acronym	Full Name	Scope	URL
SRS	Sequence Retrieval System	Integration of various databases (sequence, function, processes)	http://srs.ebi.ac.uk

Table 1: General and genome-specific sequence and structure databases

Molecular function

Acronym	**Full Name**	**Scope**	**URL**
BRENDA	The Comprehensive Enzyme Information System	Enzyme function	http://www.brenda.uni-koeln.de/
EMP	Enzymes and metabolic pathways	Enzyme function	http://wit.mcs.anl.gov/WIT2/
Transfac	Transcription Factor Database	Regulatory sites (all organisms)	http://transfac.gbf-braunschweig.de/TRANSFAC/index.html
RegulonDB		Regulatory sites (*E.coli*)	http://www.cifn.unam.mx/Computational_Biology/regulondb/
YTPdb	Yeast Transport Protein database		http://alize.ulb.ac.be/YTPdb/
DIP	Database of Interacting Proteins	Repertoire of pairs of interacting proteins	http://dip.doe-mbi.ucla.edu
GIFTS	Gene Interactions in the Fly Trans-World Server	Protein-DNA and Protein-Protein interactions in the fly	http://gifts.univ-mrs.fr/GIFTS_home_page.html

Cellular processes

Acronym	**Full Name**	**Scope**	**URL**
aMAZE	Database of protein function and biochemical processes	Metabolic pathways, regulation and signal transduction	http://www.ebi.ac.uk/research/pfbp/
KEGG/LIGAND	Kyoto Encyclopaedia of Genes and Genomes	Metabolic pathways	http://www.genome.ad.jp/kegg/
WIT - MPW – EMP	What Is There; Enzymes and Metabolic Pathways	Enzyme function, metabolic pathways, metabolic reconstruction	http://wit.mcs.anl.gov/WIT2/
EcoCyc/MetaCyc	Encycplopaedia of *Escherichia coli* metabolism	Metabolic pathways	http://ecocyc.panbio.com/ecocyc/
Biocatalysis-Biodegradation		Metabolic pathways	http://www.labmed.umn.edu/umbbd/index.html
Expasy (SWISS-PROT)		Metabolic Pathways (Electronic form of Boehringer-Mannheim map)	http://www.expasy.ch/cgi-bin/search-biochem-index
CSNdb	Cell Signalling Network database	Signal transduction and transcriptional regulation	http://geo.nihs.go.jp/csndb/

Table 2: Specialized databases