



Integration of biological data using BioXM™ Knowledge Management Environment

Andrea Ramge, Sascha Losko and Klaus Heumann



Biomax Informatics AG
Lochhamer Str. 9
D-82152 Martinsried
Tel. +49 89 895574-0
Fax. +49 89 895574-825

www.biomax.com

BIOINFORMATICS SOLUTIONS ... designed with you in mind

Introduction

The vast quantities of information generated by high-throughput experimental methods are published in new articles every day. Processes in biological systems are interrelated on many levels and their regulation presents a complexity which needs to be understood in detail.

The BioXM™ Knowledge Management Environment efficiently models such complex research environments. This platform enables scientists to create knowledge networks with flexible workflows for handling experimental information and metadata, including the annotation of ontologies. Information from public databases can be incorporated using the embedded BioRST™ Integration and Retrieval System. Users can navigate and modify the information networks. Thus, research projects can be modeled and extended dynamically.

Modeling networks

The BioXM system is designed for the aggregation of information and the semantic modeling of scientific processes. A particular area of scientific interest can be modeled as a network of related elements. The user can define different *element types* and *relation classes*. For example, elements of type "gene" or "protein" can be linked using a relation of type "Gene regulation" or "Protein-Protein-interaction". Sub-networks, called *contexts*, which allow biological pathways and processes to be organized as parts of the overall network of knowledge, can be defined. Relationships between contexts and other "semantic objects", such as elements, can be established. This allows efficient modularization and abstraction of knowledge. All semantic objects (such as *elements, relations, context* or *ontology*) can be annotated. Annotations are form-based and support hierarchical organization of information. The BioXM system supports the conceptualization of entire areas of interest using arbitrary ontologies. The taxonomy of "is_a" relationships, which formally structure the ontology, can be used to infer facts and abstract queries in the BioXM system. The software provides graphical browsing through the network and an advanced query builder for guided construction of complex queries with natural-language-like syntax.

The BioXM system allows access to all public databases integrated by the BioRS Integration and Retrieval System. External databases can serve as either "virtual" semantic objects or "read-only" annotation of semantic objects. Although the information remains external, the databases entries used as "virtual semantic objects" can be organized in the project tree, become part of a network and be annotated like other semantic objects.

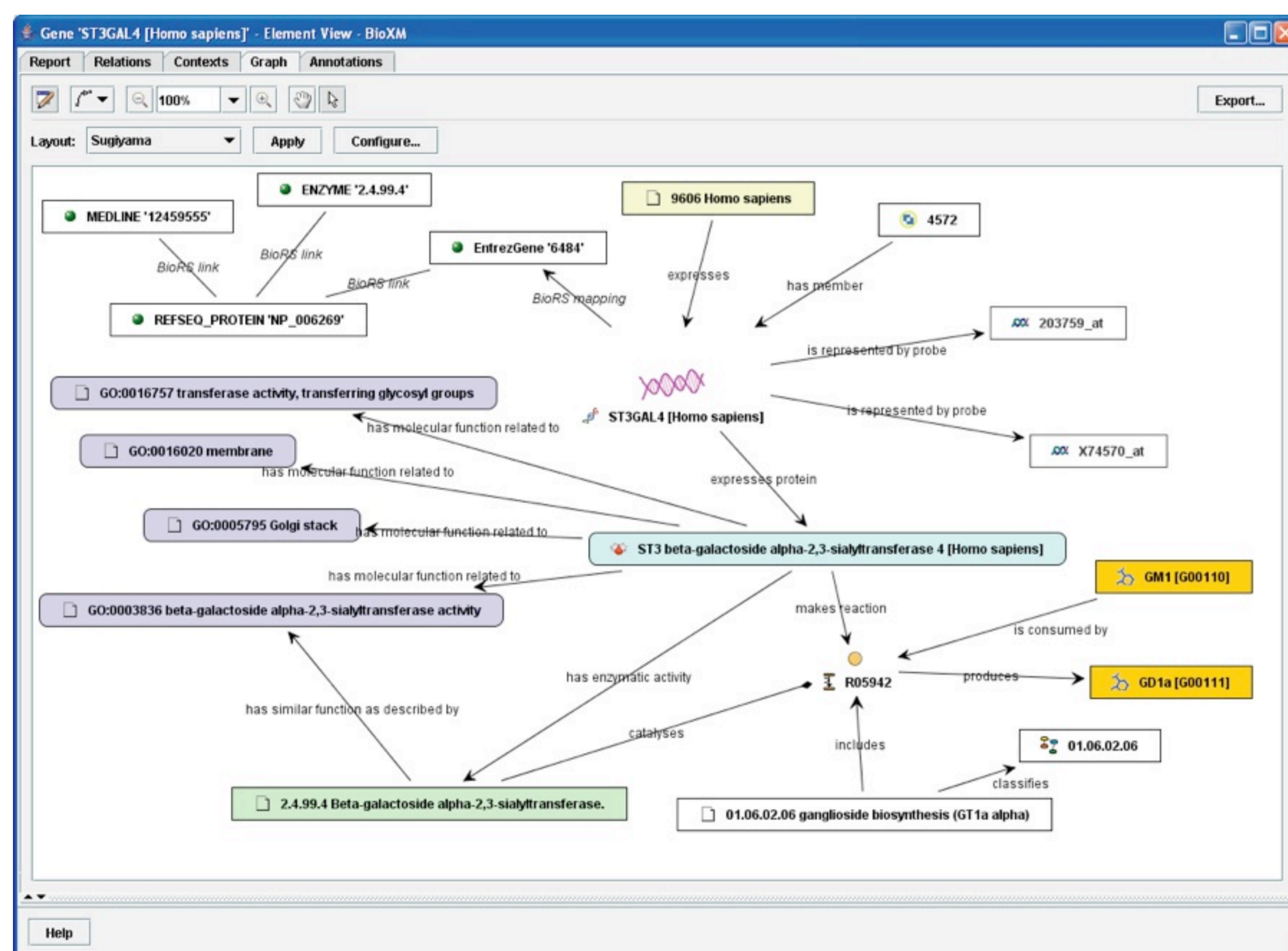


Figure 1: The BioXM graph view

The figure shows an example of a graphical view of objects and relations. Different types of relations are shown, for example the "Gene expression" relation between the *gene* ST3GAL4 and the *protein* ST3 beta-galactoside-2,3-sialyltransferase, the "Taxonomic classification" relation between the *ontology entry* GO:0005795 and the *protein* ST3 beta-galactoside-2,3-sialyltransferase, the "Functional mapping" relation between the *ontology entry* EC 2.4.99.2 and the *ontology entry* GO:0003836, the "Metabolite production" relation between the *reaction* R05942 and the *metabolite* GD1a, the "BioRS mapping" between the *gene* ST3GAL4 and the *BioRS entry* ENTREZGENE6484 and the "BioRS link" between the *BioRS entry* MEDLINE 12459555 and the *BioRS entry* REFSEQ_PROTEIN NP_006269.

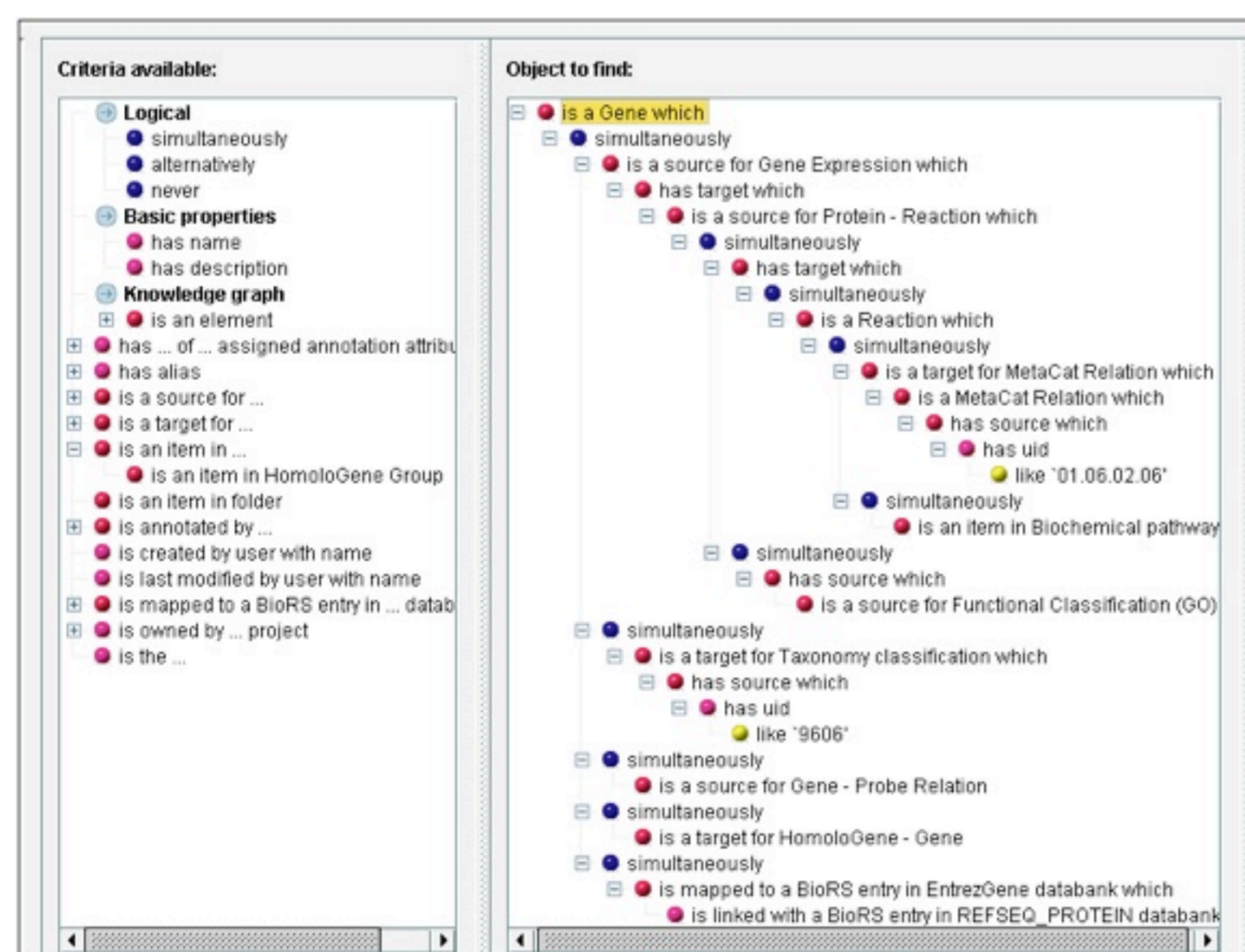


Figure 2: The BioXM query builder

The BioXM system has a powerful query builder, which facilitates querying the data, integrated in it. A query defines a search strategy (the kinds of objects which will be searched) and the search criteria (the values of the object attributes) for performing retrieval of BioXM objects. The query is a natural-language style representation of the search criteria for defining the query.

Complex queries can be stored for later reuse. A search strategy can be saved as a template which allows different values to be entered every time the template is used. Templates provide a convenient method for performing similar queries which are used frequently.

The query displayed in the screen shot searches for the gene shown in Figure 1.

MetaCat metabolic pathways ontology

For modeling biochemical pathways within the BioXM system, pathways from the MetaCat metabolic pathways ontology were used to create a corresponding *context*, called "Biochemical pathways". The MetaCat ontology is curated by professional experts in the field of biochemistry. Individual pathways in the MetaCat ontology are represented with respect to full operability in certain groups of organisms. They are arranged in a hierarchical structure consistent with scientific categorization of metabolic pathways (i.e., arranged in metabolic region and super-pathways, which can be easily represented as chapters and sub-chapters in biochemistry textbook).

The MetaCat ontology is a hierarchical ontology of metabolic pathways. The uppermost level of the ontology (01 METABOLISM and 02 ENERGY) creates a distinction to sort the lower levels. In 01 METABOLISM, lower levels are sorted by the chemical nature of the substrates and products. In 02 ENERGY, lower levels are sorted by their contribution to the energy turnover within the organisms. The second level defines a **metabolic region**, for example "01.06. Lipid, fatty-acid and isoprenoid metabolism". The third level defines a **super-pathway** (i.e., a metabolic network of interconnected, unbranched paths which relate to the same family of substances). Typically the members of such a super-pathway share a common precursor or intermediate. The fourth level defines a single unbranched **pathway** within the metabolic network of the super-pathway. There are often alternative paths from a substrate to product, which constitute a separate pathway. The fifth and lowest level of the hierarchy represents individual **reactions**. Note that an individual reaction can be carried out by different enzymes and individual enzymes are often capable of catalyzing different reactions. Each Biochemical Pathway *context* in the BioXM system refers to the fourth level of the MetaCat ontology, the pathway level. Such a Biochemical Pathway *context* is filled with all reactions of the corresponding fifth level, including the metabolites, cofactors and enzymes.

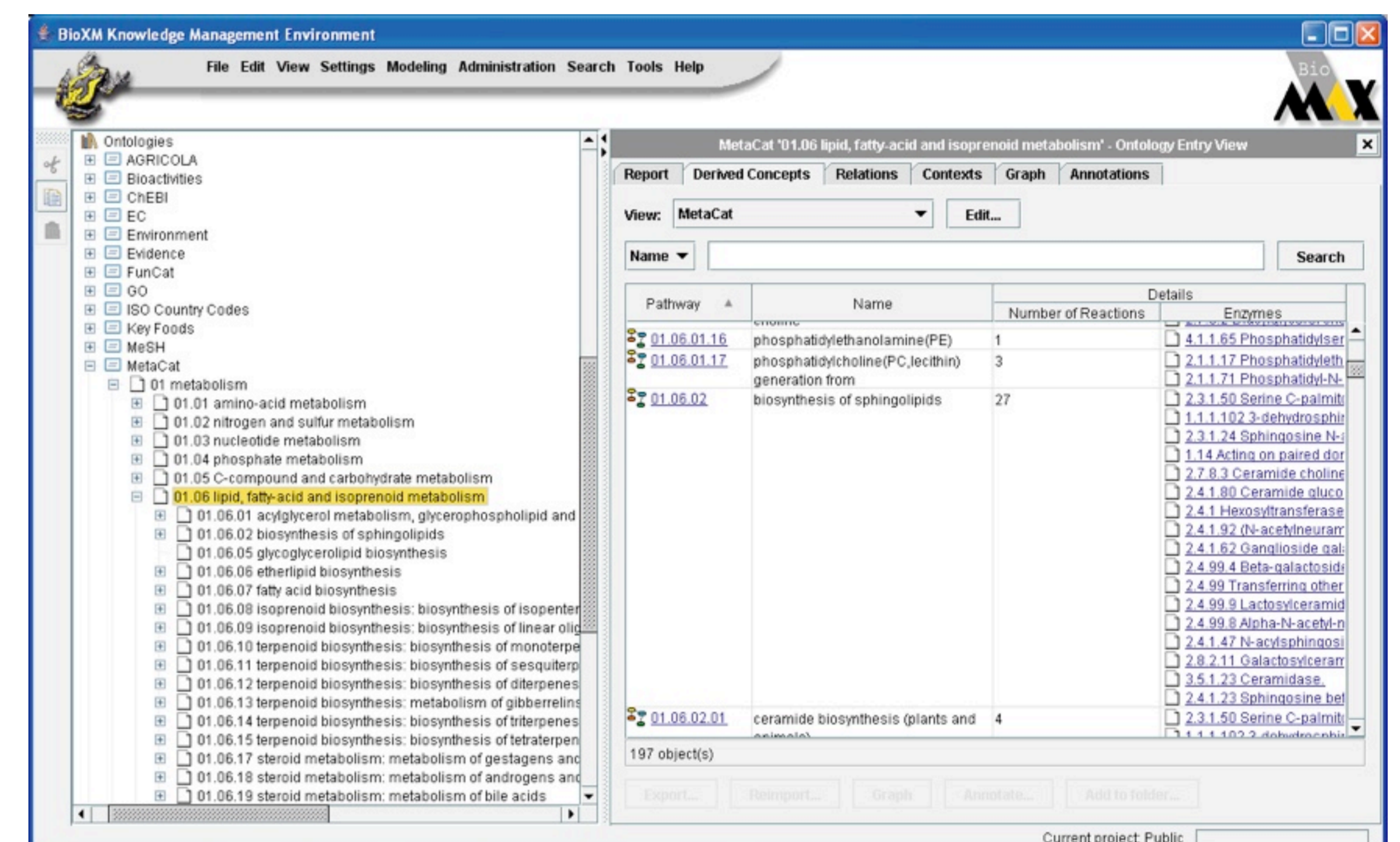


Figure 3: The MetaCat Ontology displayed in the BioXM system

On the left, a part of the project tree in the BioXM system is opened to show the the MetaCat ontology. The main window on the right displays information about the selected entry "01.06.02". The tab "Derived concepts" is selected and displays a list of all children ontology entries of the ontology entry "01.06.02". In the example a special view, which shows all pathways related to the selected MetaCat entry, was configured. For every pathway the description of the pathway, the number of reactions and the names of the enzymes participating in the particular pathway are given.

The Pathway Viewer

The BioXM system provides a generic viewer for the visualization of different kinds of relationships (see Figure 1). Additionally, a special viewer for displaying biochemical pathways in a more textbook-like manner has been developed. The Pathway Viewer is integrated in the BioXM system and gets data from the BioXM database. In addition to the visualization of pathways, the Pathway Viewer can visualize super-pathways and metabolic regions of the MetaCat ontology as well.

Additionally genome comparison on the pathway level is possible. After organisms are selected in a special Taxonomy Browser, the Pathway Viewer will display the absence or presence of all enzymes for the selected species in a selected pathway. If a particular enzyme is present, the gene and the protein information specific to the enzyme in the selected organisms is available.

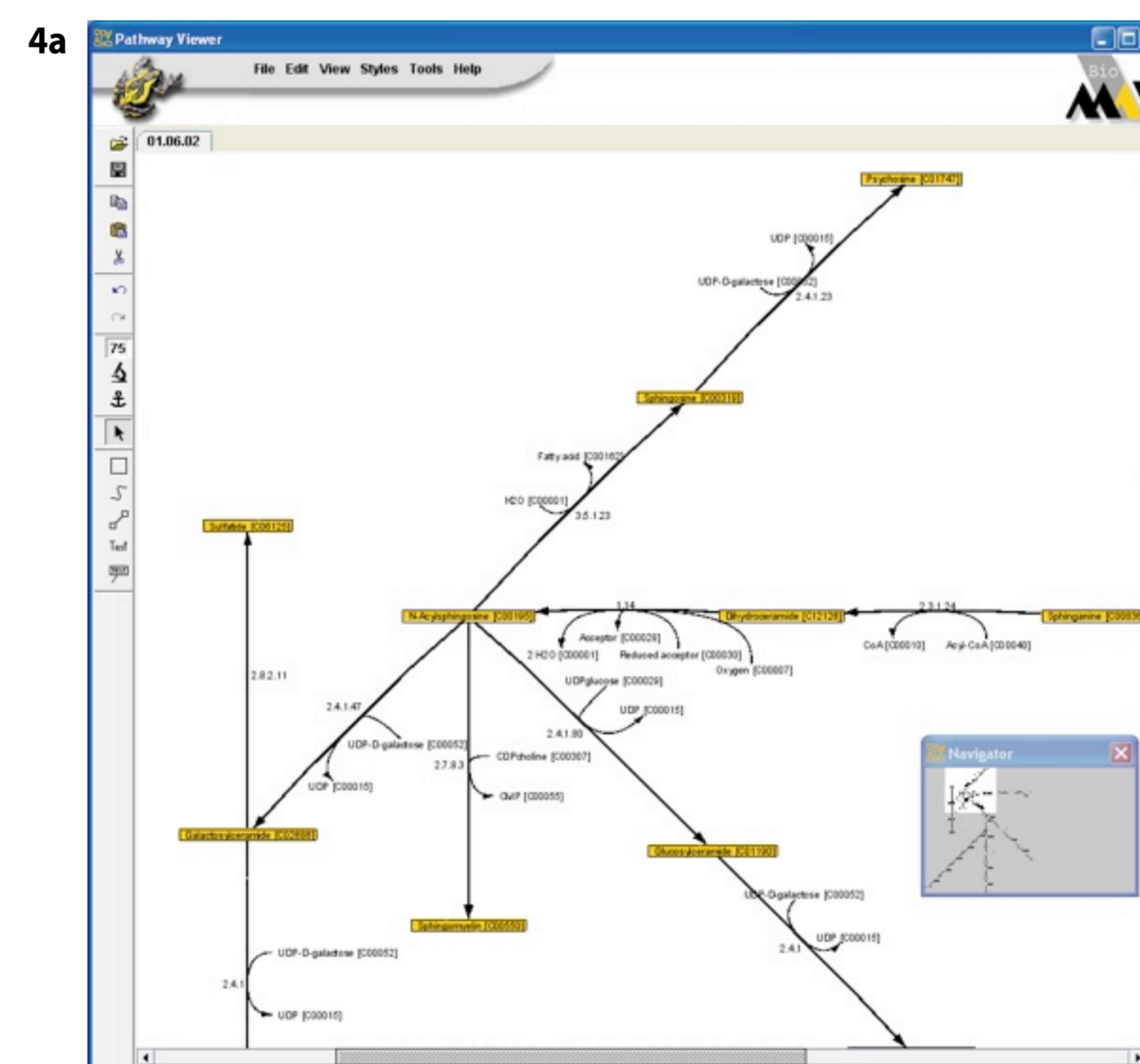


Figure 4a: The Pathway Viewer

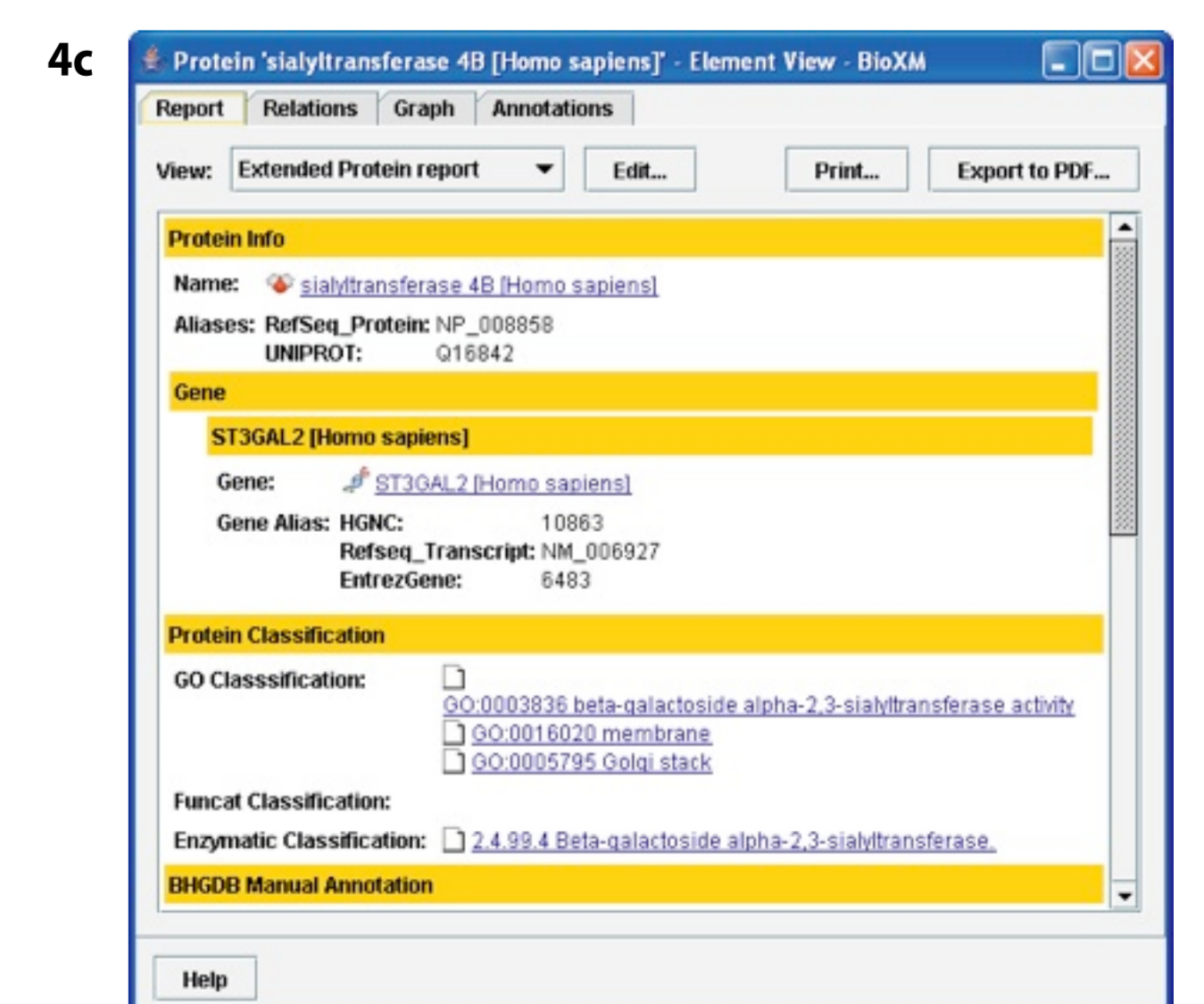
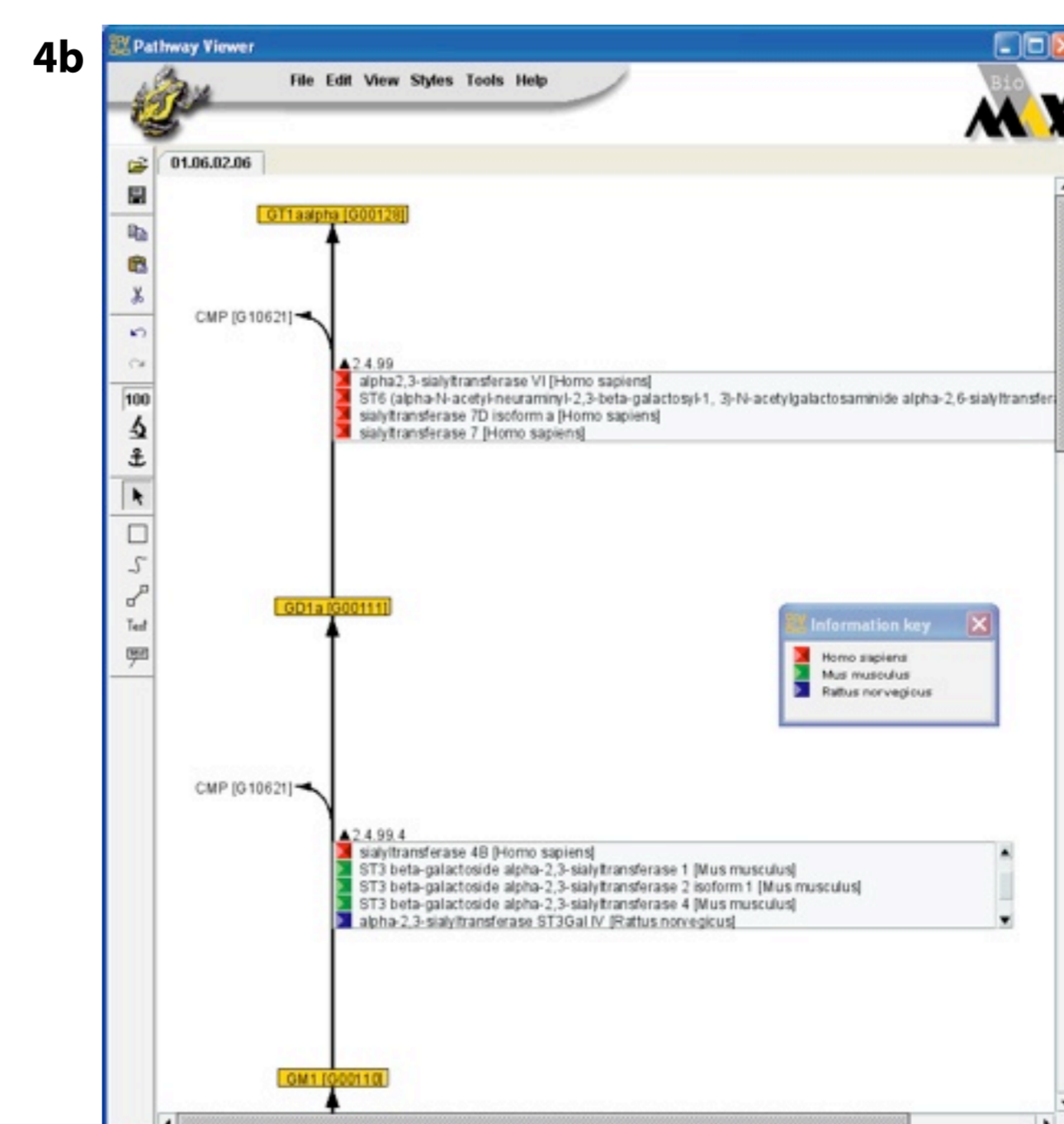
In this screen shot the super-pathway "01.06.02 biosynthesis of sphingolipids" is displayed. Metabolites are indicated by yellow boxes. Cofactors have no boxes. The EC number of the enzyme which catalyzes the particular reaction is written next to each arrow. The viewer provides the ability to export the pathway as SVG or PNG and the layout of the pathway can be changed, colored and commented.

Figure 4b: Genome comparison

After selecting the three organisms (*Homo sapiens*, *Mus musculus* and *Rattus norvegicus*) the Pathway Viewer displays the presence or absence of the enzymes in the species. As shown in the screen shot, the first enzyme of the pathway (2.4.99) is only present in *Homo sapiens*, the second enzyme (2.4.99.4) is present in all three selected organisms.

Figure 4c: Protein report

Double clicking a selected entry in the enzyme list (see figure 4b) opens the corresponding protein report. The figure shows the protein report for Sialyltransferase in *Homo sapiens*.



This work is supported by the German Ministry for Education, Science, Research and Technology (BMBF) under grant numbers 031U112G and 031U212G within the Bioinformatics for the Functional Analysis of Mammalian Genomes (BFAM) project.