ARTICLE

OBC

www.rsc.org/obc

Enabling the exploration of biochemical pathways†

Martin Reitz,[a] Oliver Sacher,[b] Aleksey Tarkhov,[b] Dietrich Trümbach[a] and Johann Gasteiger*[a,b]

[a] *Computer-Chemie-Centrum and Institute of Organic Chemistry, University of Erlangen-Nuremberg, Naegelsbachstr. 25, 91052 Erlangen, Germany. http://www2.chemie.uni-erlangen.de*

[b] *Molecular Networks GmbH, Naegelsbachstr. 25, 91052 Erlangen, Germany. http://www.mol-net.de*

Received 19th July 2004, Accepted 17th September 2004
First published as an Advance Article on the web 22nd October 2004

The Biochemical Pathways Wall Chart (http://www.expasy.org/tools/pathways/ ref. 1) has been converted into a molecule and reaction database. Major features of this database are that each molecule is represented by lists of all atoms and bonds (as connection tables), and in the reactions the reaction centre, the atoms and bonds directly involved in the bond rearrangement process, are marked. The information in the database has been enriched by a set of diverse 3D structure conformations generated by the programs CORINA and ROTATE. The web-based structure and reaction retrieval system C@ROL provides a wide range of search methods to mine this rich database. The database is accessible at http://www2.chemie.uni-erlangen.de/services/biopath/index.html and http://www.mol-net.de/databases/biopath.html

1. Introduction

With the deciphering of the human genome, the blueprint of the human organism and its functions, interest has shifted towards the role of the gene product and the activity of the expressed proteins. In other words, the interest has shifted from genomics to proteomics. Increasingly, the attention is now focused on how some of these proteins, the enzymes, regulate the processes within the cell, how nutrients are metabolized and how energy is produced and transferred through metabolism: metabolomics has entered the stage.

Over many decades, a large body of information has been accumulated detailing the chemical species that occur and are processed within the cell and how these chemical species are interconverted by series of chemical reactions. Much of this knowledge has been beautifully assembled by G. Michal and colleagues in the Biochemical Pathways Wall Chart distributed initially by Boehringer Mannheim and now by Roche (see Table 1 for definitions of acronyms).[1]

The principal scientific questions have always been the interpretation of the pathway information functionally, temporally and spatially. The task has always been to connect, and indeed correlate, the information from enzyme regulation with the elucidation of the intracellular biochemical reactions. To this end, both bioinformatics and chemoinformatics specialists are now in a position to collaborate.

Impressive as the Biochemical Pathways Chart is, it has a number of drawbacks. Firstly, it has a high information density dictated by the complexity of known facts about metabolites and biochemical transformations. Secondly, it is difficult to locate specific compounds, particularly if they are contained at several places on the map (independent of their temporal and spatial locations). This is particularly true for such ubiquitous compounds, *e.g.* ATP or pyruvate, which are produced and consumed in many chemical reactions. This raises some essential questions: in which reactions does ATP participate? Or similarly, which reactions involve acetyl coenzyme A and in what way are they involved?

A critical inspection of the Biochemical Pathways Wall Chart shows the essence of the problem: relationships between many compounds through a large number of reactions have to be stored in a two-dimensional plane (see Fig. 1). This can only be achieved by large distortions and an awkward arrangement of reaction arrows, ostensibly for reasons of clarity.

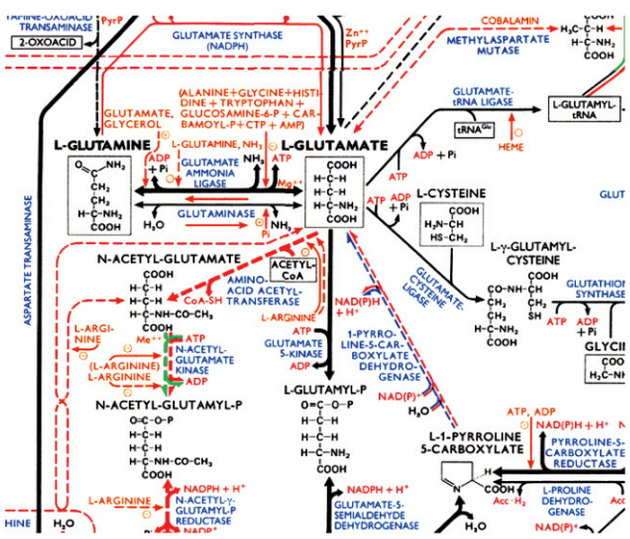


Fig. 1 View of a segment of the Biochemical Pathways Wall Chart;[1] from Michal, Biochemical Pathways 1999 © Elsevier GmbH, Spektrum Akademischer Verlag, Heidelberg.

Essentially, biochemical pathways form a high-dimensional space, a space interconnecting many compounds by a multitude of reactions, a high-dimensional space that had to be projected into two-dimensions in order to produce the Wall Chart. It is highly desirable to analyze this multifunctional nature of the space of biochemical pathways, a task that can be achieved by structure and reaction search methods that have been developed for searching in compound and reaction databases. In this way, the arsenal of chemoinformatics methods[2,3] can be brought in to analyze biochemical pathways. The task is then to store the structures and reactions of biochemical pathways in a structure and reaction database. It has to be stored in a way that is standard for chemical databases: store structures in the form of connection tables, providing access to each atom and bond of a chemical structure. Store reactions not only by giving information on the structures of starting materials and products but also by specifying the reaction centre, specifying which atoms and

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b410949j

Table 1 Table of acronyms

BioPath	http://www2.chemie.uni-erlangen.de/services/biopath/index.html http://www.mol-net.de/databases/biopath.html
Biochemical Pathways	http://www.expasy.org/tools/pathways/
KEGG	http://www.genome.jp/kegg/
BioCyc	http://www.biocyc.org/
CORINA	http://www2.chemie.uni-erlangen.de/software/corina/index.html http://www.mol-net.de/software/corina/index.html
ROTATE	http://www.mol-net.de/software/rotate/index.html
C@ROL	http://www.mol-net.de/software/carol/index.html
CACTVS	http://www2.chemie.uni-erlangen.de/software/cactvs/index.html
JME	http://www.molinspiration.com/jme/index.html
JMol	http://jmol.sourceforge.net/

bonds are directly involved in the reaction process. Only then can genuine reaction searching be performed by focusing on the changes in the arrangement of bonds in a chemical reaction.

2. The BioPath database

2.1 The data model

Six years ago, at the outset of our work on biochemical pathways no information on metabolic networks or biochemical pathways was available in a form, as was required to be able to employ the full power of structure and reaction search methods. KEGG, a database with a long history is in the meantime also providing chemical structures in the form of connection tables[4] as is done by the collection of databases on the BioCyc web page.[5a] These databases store chemical structures not only by name, but also as connection tables which are amenable to structure and substructure search methods. Recently, a database on metabolic reactions annotated for *Escherichia coli* was published that also contains a mapping of the atoms of the starting materials onto those of the products of a chemical reaction.[5b] Nevertheless, our original data model had a number of features that are not yet contained in all other metabolic or biochemical pathways databases. Thus, our BioPath database allows a depth of search on chemical structures and reactions not achievable with any other metabolic or biochemical pathways database. This is provided by a detailed data model for representing chemical structures, reactions, and enzymes.

2.1.1 Chemical structures. Chemical structures are represented by connection tables, *i.e.*, by lists of all atoms and all bonds, including those to hydrogen atoms. Stereochemistry at chiral centres and at double bonds is represented by stereochemical descriptors. Such a detailed representation is given to all small molecules, including starting materials, products, coenzymes, and regulators. Furthermore, the names of all molecular species including synonyms have been stored.

2.1.2 Chemical reactions. Chemical reactions are specified by the starting materials and products of a reaction, and the enzymes, coenzymes and regulators involved. Enzymes are characterized both by name and by their EC code number. Care was taken to ensure that each reaction equation was stoichiometrically balanced; even the involvement of a proton as a starting material or product was specified in a reaction equation.

A unique feature of our handling of reactions is that the reaction centre, the bonds broken and made in a reaction, have been marked. Furthermore, all atoms in the starting materials and products have been mapped against each other.

This feature is not available in any other metabolic or biochemical pathways database. On the other hand, this kind of information is essential for proper reaction searching. If one is interested in all reactions that reduce a carbonyl group to an alcohol and the only criteria are that the starting material should contain a carbonyl group and the product an alcohol group, one would also obtain the reaction in the top of Fig. 2 as a hit because the starting material has a carbonyl group and the product an alcohol group. The actual reaction, however, is a phosphorylation of an alcohol, of D-glyceraldehyde to D-glyceraldehyde-3-phosphate. Only if one specifies, in addition, that the atoms of the carbonyl group must map onto the atoms of the alcohol group, will this reaction not be perceived as a hit. Only the reaction on the bottom of Fig. 2, the reduction of D-glyceraldehyde to glycerol, will be perceived as satisfying the query.

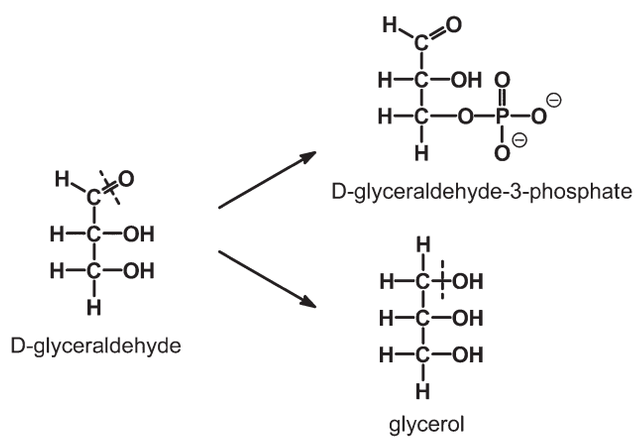


Fig. 2 Two reactions of glyceraldehyde, phosphorylation (top) and reduction (bottom).

Furthermore, the marking of the reaction centre also allows one to investigate intermediates and transition states of biochemical reactions. On this basis, the transition state hypothesis, formulated by Pauling[6] many decades ago, can be explored. This transition state hypothesis emphasizes that the role of an enzyme primarily lies in strongly binding the transition state of a biochemical reaction in order to decrease the activation energy. Studies probing this transition state hypothesis were made by superimposing CORINA-generated 3D molecular models[7,8] of transition states and intermediates of enzyme catalyzed reactions calculated from the information contained in the BioPath database onto the 3D structure of enzyme inhibitors. These studies are reported in a separate publication.[9] In addition, for each reaction, information is given on whether it is a general pathway, or whether it occurs in animals, in higher plants and yeasts, or in prokarya (bacteria and archae). Furthermore, it is indicated whether the reaction is reversible or irreversible, and whether it is a catabolic or anabolic reaction. Moreover, the compartment where a reaction occurs is specified. If a reaction belongs to a certain group of reactions, such as the citrate cycle, this is also specified.

2.1.3 Enzymes. Enzymes are represented by their names and the EC code number; a link to information on this enzyme in the BRENDA enzyme database[10] is provided.

2.2 Augmenting the contents

Having represented the chemical structures in the Biochemical Pathways database in such details as given by a connection table, allows the processing of chemical structures by chemo-

informatics software developed to provide additional information on chemical structures.

Thus, all chemical structures have been processed by the automatic 3D structure generator CORINA[7,8] providing for each small molecule a 3D molecular model. CORINA generates a single low energy conformation. In order to more fully explore the conformational space, an ensemble of conformations was generated by the program ROTATE.[11,12] For each chemical structure clearly, the conformational space can be quite large.

So as not to consider too many conformations, the number of conformations for each chemical compound was limited by exploring only the three central bonds in a molecule. In this process the generation of conformations was constrained in such a way as to obtain an ensemble of quite diverse conformations.

Other data and properties that were generated by computational methods are: molecular mass (weight), number of rotatable bonds, number of atoms, number of rings as well as number of hydrogen bond donor and acceptor atoms.

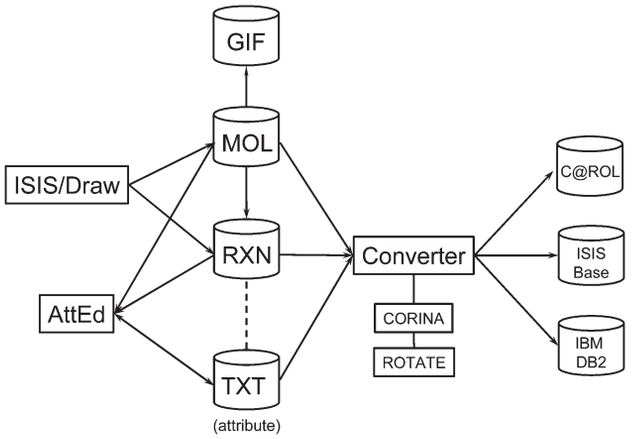


Fig. 3 Data processing from the raw data into the structure and reaction databases.

2.3 Data input

The chemical structures and reactions were input with ISIS/Draw, available from MDL Information Systems.[13] Additional information on chemical structures and reactions was input by an Attribute Editor, specifically developed for this purpose.

The basis for extracting the necessary data was the Biochemical Pathways Atlas[14] that contains information quite parallel to the Wall Chart as it has been produced by the same author. However, the Atlas provides more detailed and more up-to-date information than the Biochemical Pathways Wall Chart. In particular, the detailed reaction schemes in the Atlas were essential for marking the reaction centres. In order to emphasize the correspondence between Wall Chart and Atlas, the location of each structure and reaction on the Wall Chart has been indicated by giving the grid square of the Wall Chart where the structure, reaction, or enzyme is located. Furthermore, the implementation at the Computer-Chemie-Centrum[18a] also provides a link to the ExPASy server[15] that contains a scanned image of the Wall Chart. The grid square specification which is also maintained on the ExPASy server allows direct access to that part of the scanned image where the structure, reaction, or enzyme is contained. Clearly, the input of this detailed information was quite labour intensive and, in particular, the marking of the reaction centre required detailed chemical analysis. However, we believe that the quality of information that is now available was worth the effort.

2.4 Data processing and storing

The entire chemical information of structures is stored in MOL and ISIS Sketch files both generated with MDL ISIS/Draw. The reaction information is stored in RXNfiles also generated with MDL ISIS/Draw.

The 2D coordinates of the molecules were set while drawing the structures (preferably in the Fischer projection) and stored in the MOLfiles. In order to also get 3D structures of each molecule the program CORINA was then called for each file.

Besides the 2D and 3D structure information, stored in MOL and RXNfiles, the textual information (such as compound names, EC numbers, species, and compartments) is stored in so-called attribute files. These plain text files were generated by the program Attribute Editor (AttEd) based on the CACTVS system[16] (see Fig. 3).

The MOL, RXN, attribute, and GIF files are used to generate the various database versions of the Biochemical Pathways. Currently, the system is available for IBM DB2, MDL RDF for ISIS/Base, and in the C@ROL format.

In contrast to all other formats, the IBM version supports only the entire textual information of BioPath, but neither connection tables, nor reaction centres.

The textual content of the IBM DB/2 version is accessed through standard SQL statements.[17] The structure information of molecules or reactions is retrieved by displaying single or assembled GIF images which were generated during the data input by using CorelDraw.

The BioPath database was also implemented under ISIS/Base, a proprietary database management and retrieval system of MDL Information Systems.[13] For generating an ISIS/Base version of BioPath all RXNfiles were concatenated into one RD-File, whereas the additional information provided in the plain text files were also integrated for each reaction and molecule. This task was done by a script based on the CACTVS system.[16]

Finally, to take full advantage of the rich information input into the BioPath database it was integrated into the C@ROL retrieval system. This version was generated from the MOL, RXN, and attribute files. As this retrieval system also supports 3D searches in the conformational space, the program ROTATE was run and each calculated conformation of the unique molecules was stored as a property in the C@ROL database. This version is made available to the scientific community *via* the internet URLs http://www2.chemie.uni-erlangen.de/services/biopath/index.html and http://www.mol-net.de/databases/biopath.html.[18]

3. The C@ROL retrieval system

In order to take full advantage of the rich and diverse information stored in the BioPath database, we had to develop our own structure and reaction retrieval system. The C@ROL system (Compound Access and Retrieval OnLine)[19] will be briefly outlined here. It is a web-based retrieval system that allows searches either on chemical structure information or on chemical reaction information. Although here, the use of C@ROL for searching in the BioPath database will be detailed, C@ROL is a general retrieval system for searching chemical structure or reaction information on web based databases. Its use on a large structure database can also be freely explored.[19]

Figs. 4 and 5 show the main graphical user interface (GUI) for specifying a query in the structure, or in the reaction mode, respectively. A switch between these two operation modes can be made by clicking on the upper left-hand corner of the GUI. In both cases, the integrated applet of the JME molecule editor, developed by Peter Ertl at Novartis,[20] allows the graphical specification of a query of a chemical structure or of a reaction centre.

Fig. 4 shows this for the specification of the skeleton of L-glutamate for a stereochemical substructure search. If a substructure search is initiated, unspecified bonds needed to bring an atom to its standard valence state are assumed to be open sites, to carry any atom. If a full structure search is specified these open sites will be connected to hydrogen atoms.

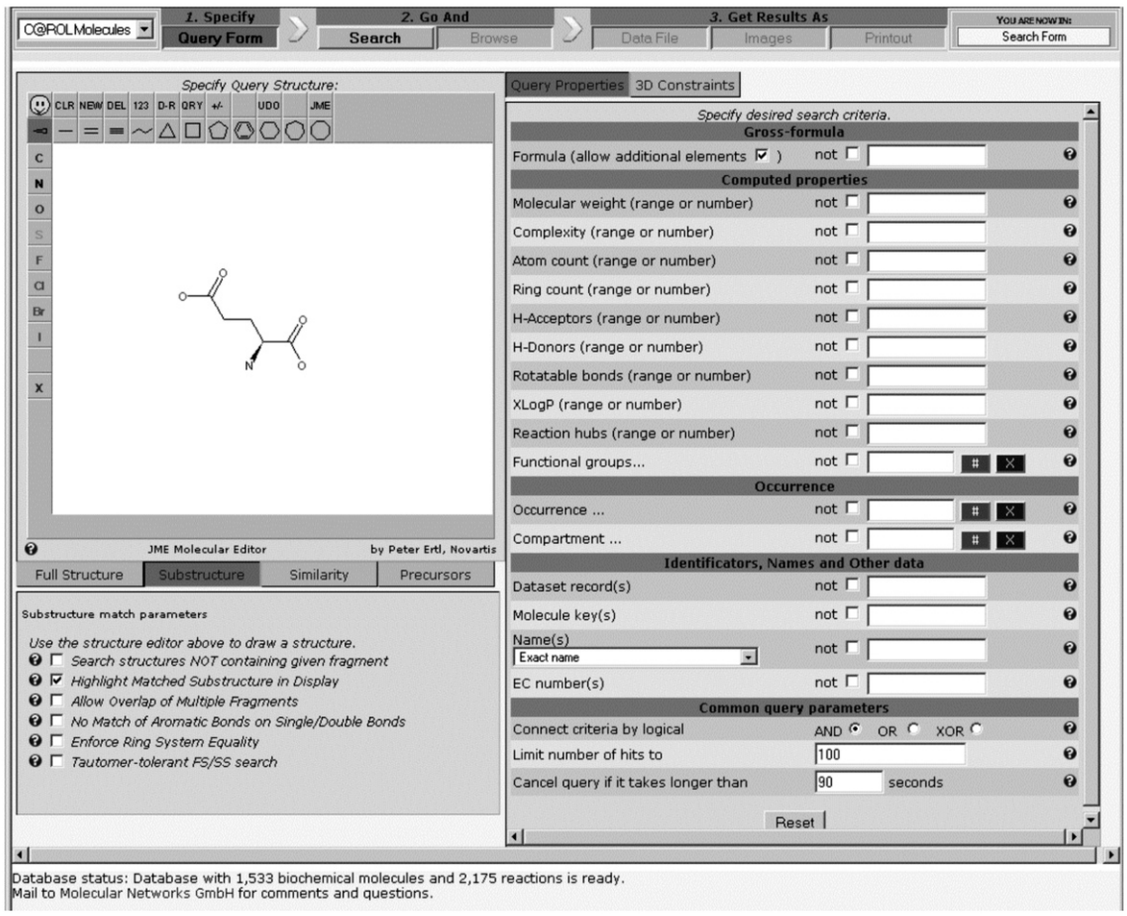


Fig. 4 C@ROL GUI for searching in a structure database with the substructure of L-glutamate specified as a query.

Fig. 5 shows the GUI for the specification of the reduction of a carbonyl group to an alcohol group.

In each case, whether it is a structure or a reaction search, a variety of additional query specifications can be made. Different search criteria can be combined by logical operators (AND, OR, XOR). Space prohibits a list of all search possibilities. A user manual of the C@ROL system details the full array and potential of search methods provided.[19]

Insight into some of the search methods incorporated into the C@ROL system can be obtained from the following examples of

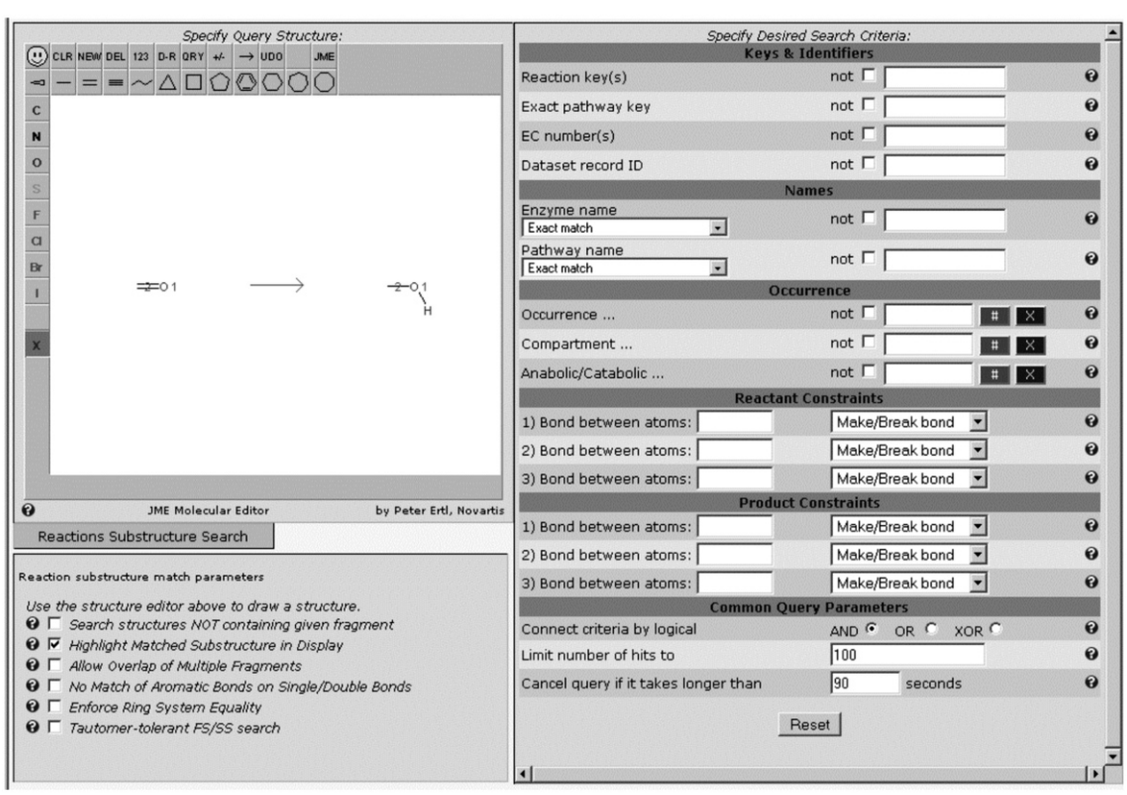


Fig. 5 C@ROL GUI for searching in a reaction database. The conversion of a carbonyl group to an alcohol group is specified as a query.

Inv.	Record	Formula	Name
☑	60	$C_3H_7O_6P$	sn-GLYCEROL
☑	73	$C_3H_5O_6P$	GLYCERONE-P
☑	92	$C_5H_{13}NO_6P$	GLYCERO-3-PHOSPHOETHANOLAMINE
☑	107	$C_3H_8O_3$	GLYCEROL
☑	148	$C_3H_5O_7P$	3-P-D-GLYCERATE
☑	151	$C_3H_6O_3$	D-GLYCERALDEHYDE
☑	152	$C_{10}H_{14}?O_8P$	DIACYLGLYCEROL-3
☑	232	$C_6H_{11}O_7P$	ACYL-GLYCEROL-3
☑	247	$C_{47}H_{77}N_3O_{15}P_2$	CDP-1,2-DIACYLGLYCEROL
☑	270	$C_{10}H_{15}?O_5$	D-1,2-DIACYLGLYCEROL
☑	314	$C_3H_5O_6P$	D-GLYCERALDEHYDE
☑	347	$C_3H_5O_7P$	2-P-D-GLYCERATE
☑	348	$C_3H_6O_4$	D-GLYCERATE
☑	349	$C_6H_{12}O_4$	MONOACYLGLYCEROL
☑	425	$C_{19}H_{26}?N_3O_{15}P_2$	CYTIDYL-5'-DIPHOSPHATE-1,2-DIACYLGLYCEROL
☑	428	$C_8H_{20}NO_6P$	L-1-GLYCERO-3-PHOSPHOCHOLINE
☑	430	$C_{13}H_{20}?O_{13}P_2$	3(3-PHOSPHATIDYL-)-L-GLYCEROL-1-PHOSPHATE
☑	431	$C_{13}H_{21}?O_{10}P$	3(3-PHOSPHATIDYL-)GLYCEROL
☑	432	$C_{13}H_{19}?O_6$	TRIACYLGLYCEROL
☑	438	$C_{38}H_{66}O_5$	1,2-DIACYLGLYCEROL
☑	457	$C_{23}H_{34}?_2O_{17}P_2$	DIPHOSPHATIDYL-GLYCEROL
☑	708	$C_3H_4O_{10}P_2$	2,3-P2-D-GLYCERATE
☑	774	$C_{15}H_{18}?N_3O_{18}P_2$	CDP-1,2-DIACYLGLYCEROL
☑	868	$C_3H_4O_{10}P_2$	1,3-P2-D-GLYCERATE
☑	938	$C_{11}H_{12}NO_6P$	(3-INDOLYL)-GLYCEROL-PHOSPHATE
☑	1206	$C_6H_9N_2O_6P$	D-ERYTHRO-IMIDAZOLE-GLYCEROLPHOSPHATE

Fig. 6 Hit list obtained by inputting the name fragment "glycer" as a query.

searches in the BioPath database as well as from a more detailed publication on the use of the BioPath database on enhancing our insight into biochemical pathways.[21]

4. Searching in the BioPath database

The BioPath database has been made accessible to the scientific community on the web.[18] With the following examples using the C@ROL retrieval system we want to assist the users in taking full advantage of the cornucopia of information contained in the BioPath database. A more detailed publication will show the application of the BioPath database to enhancing our insight into biochemical pathways.[21] An analysis of the metabolites of *Escherichia coli* contained in the EcoCyc database[22] has recently been published.[23] Here, we will first outline queries into the molecule database of the BioPath database and then make investigations into the reaction part of the BioPath database.

4.1 Searching in the molecule database

4.1.1 Name and name fragment searching. The C@ROL system allows a variety of search queries on names and name fragments. Thus, switching the entry in the drop-down list to "name fragment" and typing "glycer" provides 26 hits covering such compounds as glycerone-P, D-glyceraldehyde, cytidyl-5′-diphosphate-1,2-diacylglycerol, and D-*erythro*-imidazol-glycerol-phosphate (see Fig. 6).

By clicking on one of the record numbers, a C@ROL Detail Page will open giving the structure diagram of the compound and additional information such as molecular weight, elemental composition, in which organisms and pathways this compound occurs, on which grid spaces of the Biochemical Pathways Wall Chart the compound is contained, and which enzymes work on this compound. This is illustrated in Fig. 7 with compound no. 92 of the hit list in Fig. 6, glycero-3-phosphoethanolamine.

4.1.2 Gross-formula searches. C@ROL also allows searches based on the atomic compositions of molecules. A search with C3O3, also allowing the presence of any other atom, returns 10 hits. With C3O3N0 nine hits were obtained and with C3O3N0S0 seven compounds, containing, other than the 3 carbon and 3 oxygen atoms only hydrogen atoms. Fig. 8 gives the list of the compounds.

The structures of these compounds can be visualized as 2D or as 3D structures. The 3D structures were obtained from the automatic 3D molecule structure generator CORINA;[7,8] for visualization the 3D molecule viewer JMol[24] integrated into the C@ROL system was used. Fig. 9 shows the 3D structures of L- and of D-lactate, respectively. Rotation of the 3D molecular models allows one to obtain an excellent impression of the 3D structure of the corresponding molecule.

4.1.3 Full structure and substructure searches. The most powerful searches that provide detailed insights into the chemical nature of the metabolome are certainly structure and substructure searches. The full structure search is particularly helpful for locating chemical compounds on the Wall Chart. After input of a chemical structure by the molecule editor JME, and initiating a search, a result is obtained that also gives the grid squares of the Wall Chart where this compound is located. Fig. 10 shows the detailed information obtained through a full structure search on oxaloacetate.

Particular emphasis is given here to the list of enzymes that transform oxaloacetate, starting with EC number 1.1.1.37 malate dehydrogenase. By clicking on the EC number of a particular enzyme listed here, the reaction catalyzed by this enzyme will be shown. Thus, in effect, a switch from the molecule database to the reaction database can be performed. The detailed information also indicates the grid squares where oxaloacetate is contained on the Biochemical Pathways Wall Chart, and, by the same token, on the web site of the ExPASy server.[15]

By clicking on these specifications of grid squares, the Computer-Chemie-Centrum implementation[18a] establishes a link to the ExPASy server where the Biochemical Pathways Wall Chart is contained in scanned form. Thus, a click on grid square F5 directly leads to that part of the Wall Chart where the reduction of oxaloacetate by malate dehydrogenase is embedded (Fig. 11).

2D substructure searches can be initiated in much the same way as full structure searches: The 2D substructure is drawn with JME, the button "Substructure" is activated and a search is started by clicking on "Search" of Step 2 in the top most bar.

A search with the substructure indicated in Fig. 12 provides 60 hits of steroids. The substructure of the query is highlighted in the structures that are obtained as hits.

Fig. 7 Detailed information obtained for the compound with record no. 92, glycero-3-phosphoethanolamine, from the hit list in Fig. 6.

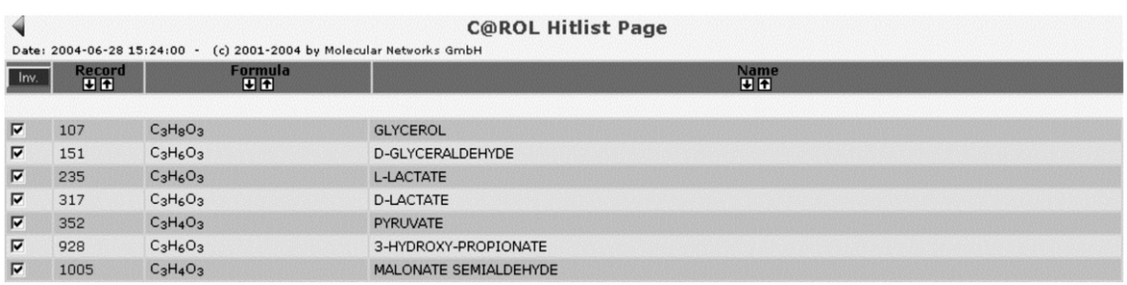


Fig. 8 Hit list obtained by inputting "C3O3N0S0" in a gross-formula search.

The same substructure additionally containing an OH group in position 3 gives 49 hits. Changing ring A of the steroid skeleton to an aromatic ring provides eight structures as hits.

4.1.4 Property retrieval. A variety of properties of chemical structures are offered for searching in the BioPath database. For example, searching for compounds with four rotatable bonds provides 131 structures.

4.1.5 3D-substructure searches. The BioPath database contains 3D structures generated by the combined application of the 3D structure generator CORINA[7,8] (giving a single low energy conformation) and ROTATE[11,12] (providing an ensemble of quite diverse conformations). On this basis, 3D substructure searches can be performed.

Diethylstilbestrol, DES (see Fig. 13a), is a synthetic estrogen not used any more as contraceptive because of its potential carcinogenicity. In order to search for structures with similar biological properties as DES the phenol substructure and the second oxygen atom were considered to be important for biological activity. The 3D structure of DES showed that the two oxygen atoms are at a distance of 11.92 Å. Accordingly, the 3D pharmacophore was specified as consisting of a benzene ring with an OH-group and an additional oxygen atom at a distance of 11.92 ± 1.5 Å from the oxygen atom of the OH-group (see Fig. 13b).

The 3D substructure search resulted in a hit list of 13 molecules. To further reduce the number of hits, the additional restriction was imposed that the hits should have only two hydrogen bonding acceptor atoms as DES has only two such atoms. This reduced the number of hits to two, estrone and estradiol (see Fig. 14). Thus, in this 3D search the natural ligands of the estrogen-receptor that also binds the query structure, diethylstilbestrol, were found.

4.2 Searching in the reaction database

4.2.1 Searching with chemical structures. When C@ROL is switched to reaction searching by choosing this feature in the upper left-hand corner of the GUI, the search window opens with the molecule editor already showing a reaction arrow (see Fig. 5).

Drawing a chemical structure on the left-hand side of this arrow allows one to search for all those reactions where this structure is a starting material. Fig. 15 shows this for a query for all reactions starting from chorismate. This query provided 10 reactions, as shown in Fig. 16. Six of these 10 reactions are superpathways, combining several single reactions.

4.2.2 Searches on the reaction centre. It has already been emphasized that the most important searches on chemical reactions have to go through the actual event occurring at the molecular level, investigating the bonds broken or made. In order to be able to perform such searches we had to go through the laborious task of marking the reaction centre, marking the bonds broken and made in a reaction and indicating how the atoms in the starting materials are mapped onto the atoms of the products.

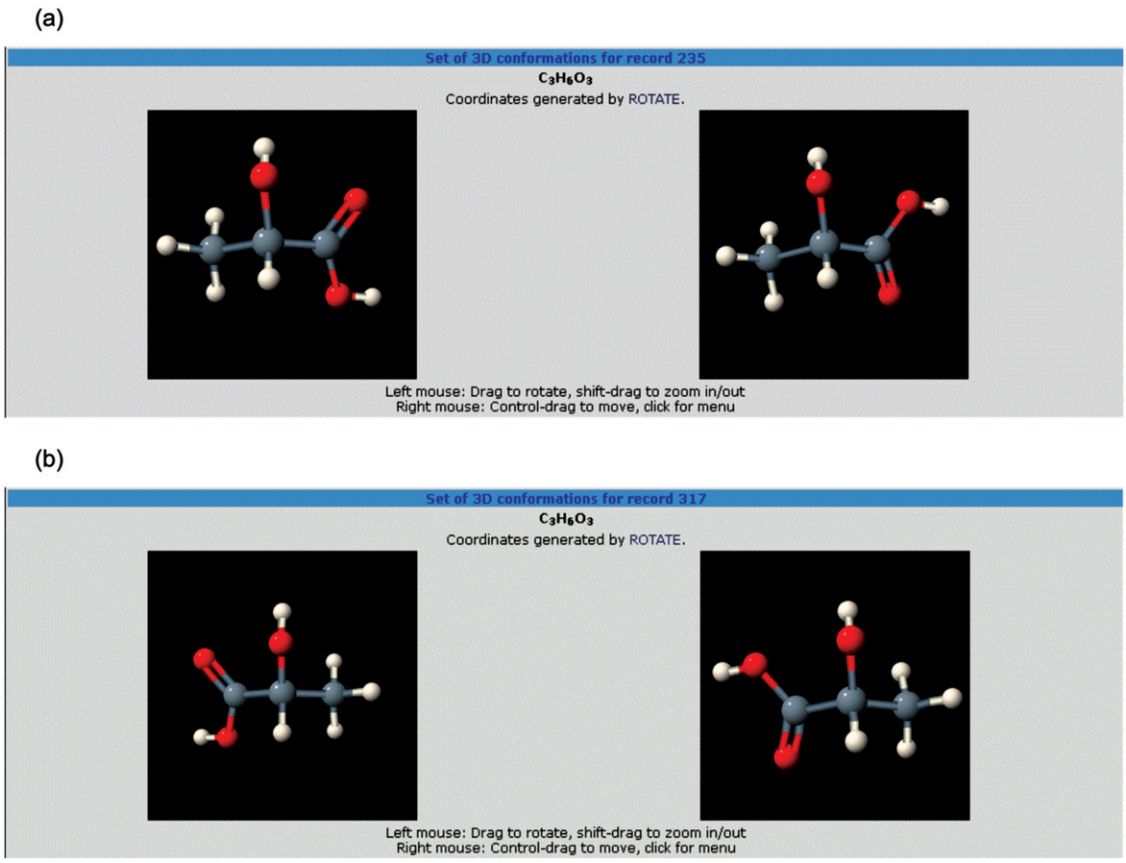


Fig. 9 3D molecular models of L-lactate and D-lactate obtained from the hit list in Fig. 8.

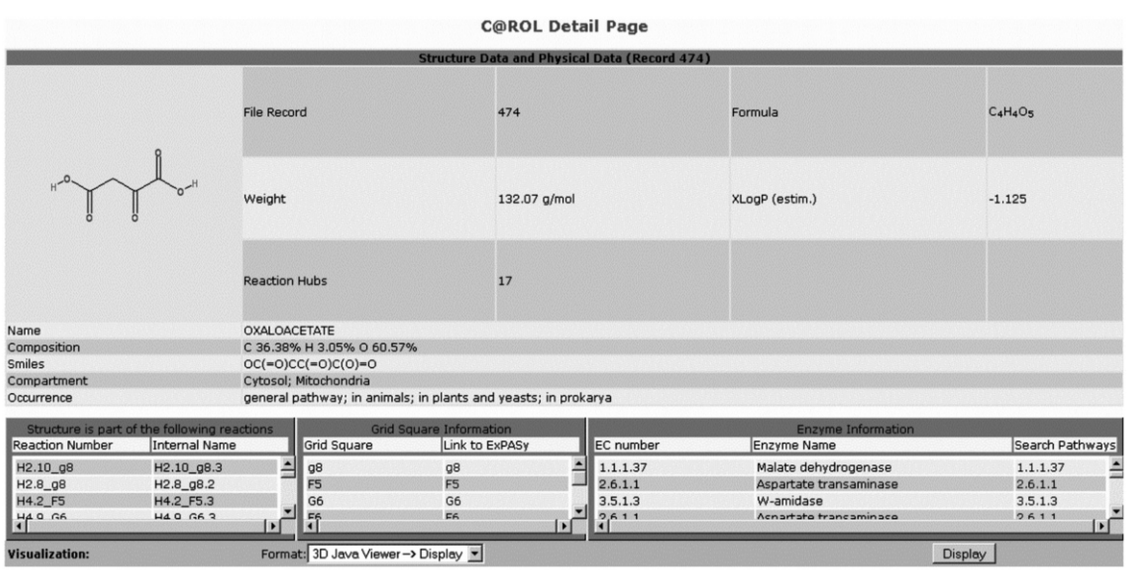


Fig. 10 Detail page obtained in a full structure search on oxaloacetate.

Having done so, allows us to ask questions on chemical reactions that would otherwise have remained unanswered. In particular, questions on reaction types, on reaction instances having common features can be asked. Thus, for example, specifying a C–C bond on the right-hand side of the arrow in the JME editor (Fig. 17) provides all those reactions that form a C–C single bond. In the case of the BioPath database, 119 reactions were obtained as hits. Among them such diverse reactions were contained as those listed in Fig. 18, encompassing reactions catalyzed by the enzymes prostaglandin synthase, choline kinase, or geranyl-*trans*-transferase, *etc.*

Furthermore, specific bond transformations can be asked, such as the conversion of a C–H bond into a C–O–H bond (Fig. 19). This provided 97 reactions, one is shown in Fig. 20.

4.2.3 Searching on enzymes. A variety of search methods for chemical reactions based on specifications on enzymes are offered. Enzymes can be searched by their name, by enzyme types through name fragments such as "oxidase", or by partial or complete EC-numbers such as 3.1.*.* or 3.1.3.3.

Searching with the EC number 3.1.3.3 provides two hits, one being the conversion of 3-phosphoserine to L-serine by the enzyme phosphoserine-phosphatase (Fig. 21).

Incidentally, this example again emphasizes the difficulty of representing the biochemical pathways on the 2D Wall Chart. Switching to the ExPASy server by following the link obtained in this query gives the grid square shown in Fig. 22. The starting material, the product and the name of the enzyme had to be highlighted in order to locate this reaction in this complicated

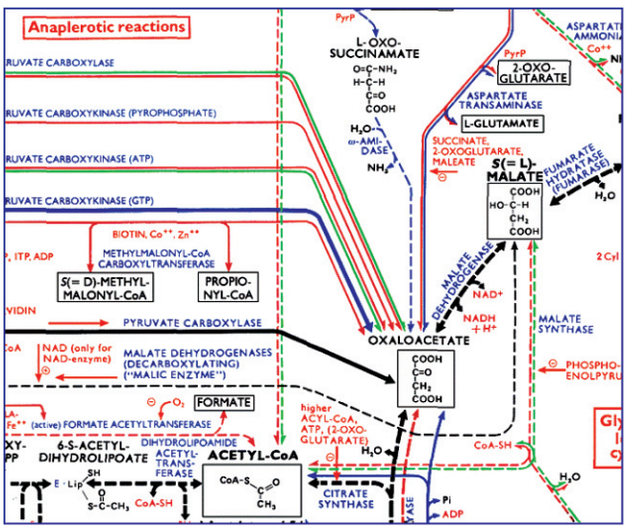
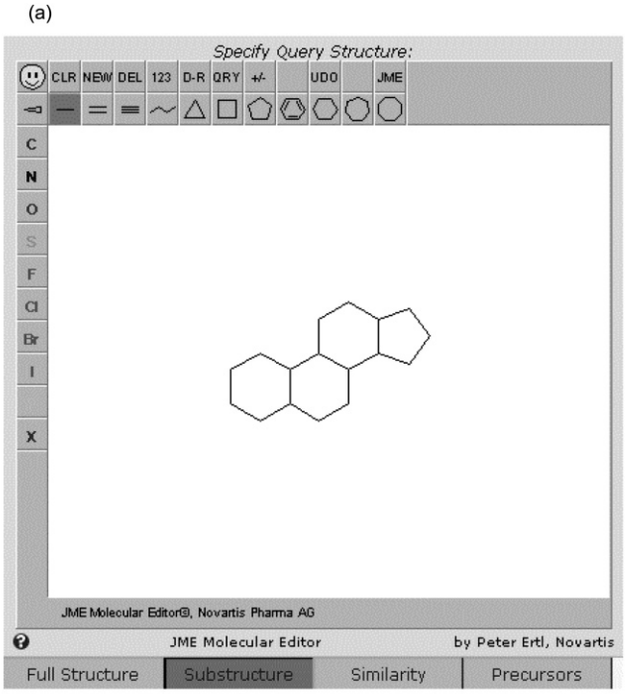


Fig. 11 View of the grid square F5 of the Biochemical Pathways Wall Chart as contained on the ExPASy server;[15] from Michal, Biochemical Pathways 1999 © Elsevier GmbH, Spektrum Akademischer Verlag, Heidelberg.

(a)



(b)

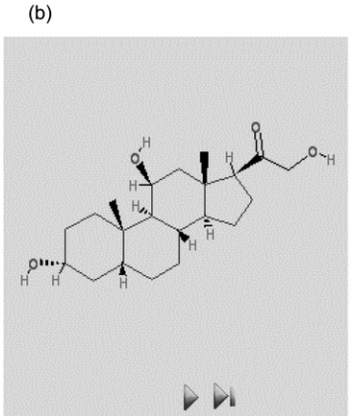


Fig. 12 Substructure specified with JME for a 2D substructure search (a) and one of the hits obtained in this query (b).

scheme, because of the strong distortion of the reaction scheme in this 2D plot.

Thus, the multi-dimensional nature of biochemical pathways is reiterated and the advantages of searching biochemical structures and reactions in a database highlighted.

Over and beyond containing links to the ExPASy server, the EC number also provides a link to the BRENDA database[10] where detailed information including kinetic data on the various enzymes can be obtained.

4.3 Combined searches

C@ROL allows the combination of several search queries.

As an example, the combination of a reaction search with the search for different organisms is given. L-Tryptophane is an essential amino acid that cannot be synthesized by animals. Inputting L-tryptophane as product of a reaction and selecting prokarya as organism provides a hit showing the synthesis of tryptophane from chorismate. If, however, animals are selected as organism, no hit is obtained, showing that this compound cannot be synthesized by animals and humans.

5. Conclusions

Storing the metabolites and the chemical reactions that interconvert them in a database opens biochemical pathways for a detailed inspection. The C@ROL retrieval system has been specifically extended for searching in the biochemical pathways database to extract information on these all-important structures and reactions.

We are now in a position to take advantage of the wealth of information stored in the BioPath database and contribute to the studies of metabolomics. Results of an investigation will be published in a subsequent paper.

Acknowledgements

We appreciate assistance in the construction of the BioPath database and the analysis of its contents through projects funded by the Bundesministerium fuer Bildung und Forschung (projects no. 08 C 5850 0, 08 C 5879, 031U112D, 031U212D, 031U112A, and 031U212A). We appreciate the initiation of the Biochemical Pathways project by Spektrum Akademischer Verlag and the collaboration with Professor Guido Moerkotte and Dr Carl-Christian Kanne, University of Mannheim, Germany in establishing our data scheme. Dr Wolf-Dietrich Ihlenfeldt, at that time at the CCC provided important contributions to this project, most notably the CACTVS system. Discussions with Dr Gerhard Michal were always stimulating. We are indebted to a number of students who carefully input the structures and reactions into the BioPath database. The BFAM project, initiated by Professor Hans-Werner Mewes allowed us to continue the work on the BioPath database.

References

1 *Biochemical Pathways Wall Chart*, ed. G. Michal, Boehringer Mannheim, Germany; now Roche also on the internet at: http://www.expasy.org/tools/pathways/.
2 *Chemoinformatics—A Textbook*, ed. J. Gasteiger and T. Engel, Wiley-VCH, Weinheim, Germany, 2003.
3 *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, Germany, 2003, 4 volumes.
4 S. Goto, Y. Okuno, M. Hattori, T. Nishioka and H. Kanekisa, *Nucleic Acids Res.*, 2002, **30**, 402–404. KEGG/LIGAND, Kyoto University, Japan, http://www.genome.ad.jp/kegg/.
5 (*a*) BioCyc Database Collection, http://www.biocyc.org; (*b*) M. Arita, *Proc. Natl. Acad. Sci USA*, 2004, **101**, 1543–1547. http://www.metabolome.jp.
6 L. Pauling, *Chem. Eng. News*, 1946, **24**, 1375–1377.
7 J. Sadowski, J. Gasteiger and G. Klebe, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1000–1008.
8 CORINA can be tested on the internet at http://www2.chemie.uni-erlangen.de/software/corina/free_struct.html and is available

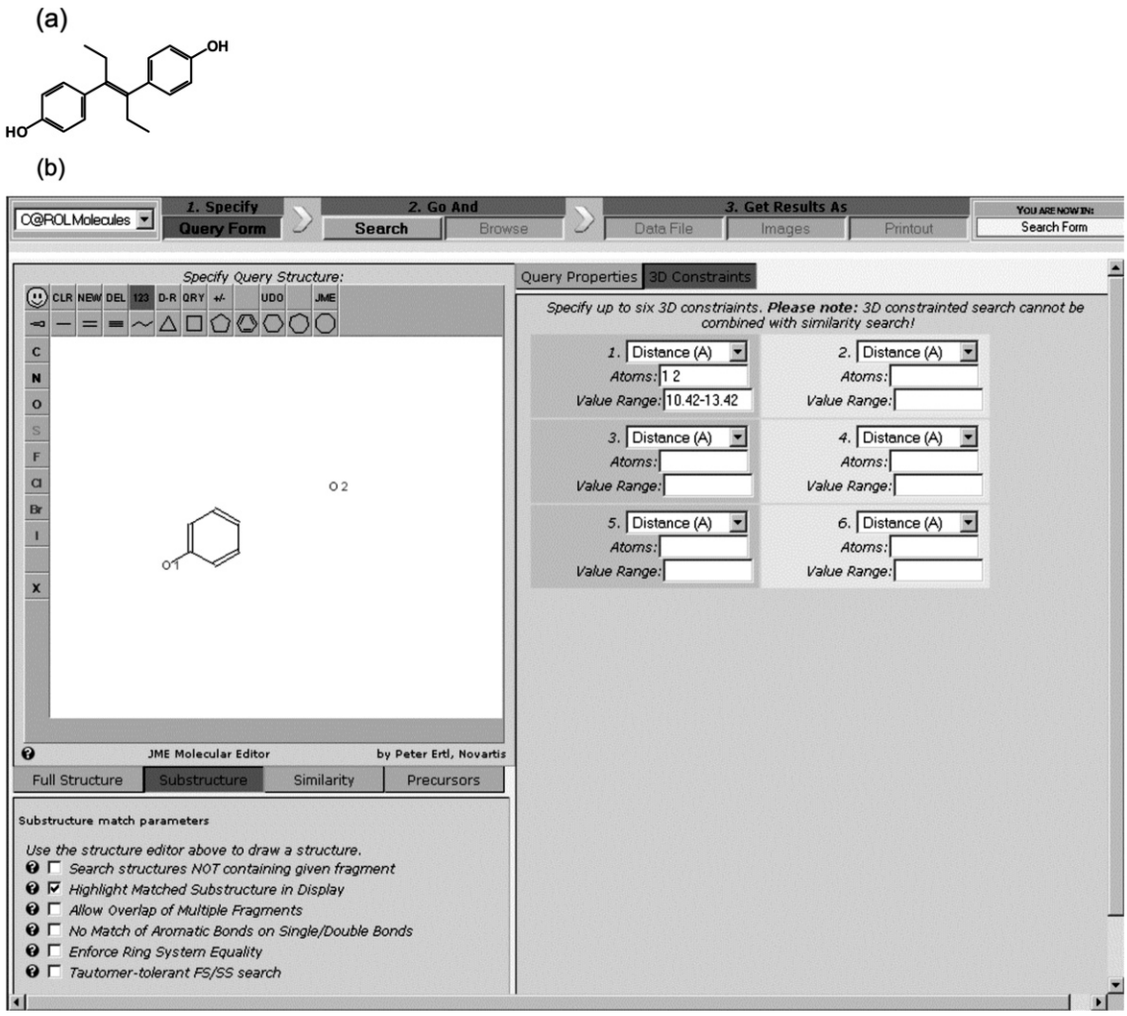


Fig. 13 Structure of diethylstilbestrol, DES (a) and a 3D substructure search query derived from DES (b).

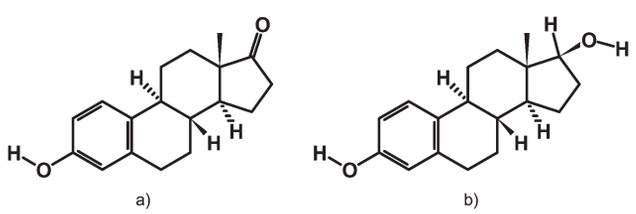


Fig. 14 Hits found with the query specified in Fig. 13b and only having two hydrogen bond acceptor atoms.

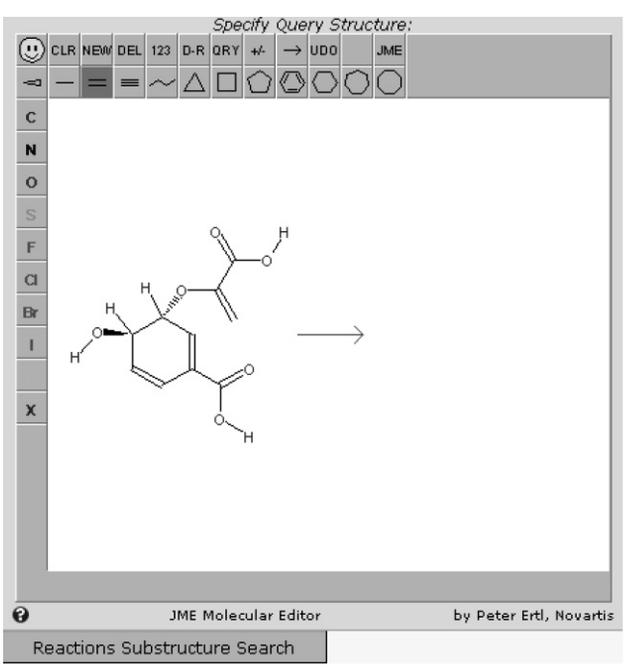


Fig. 15 Query for searching for all reactions that start with chorismate.

from Molecular Networks GmbH, Germany, info@mol-net.de, http://www.mol-net.de.

9 M. Reitz and J. Gasteiger, in preparation.

10 BRENDA—The Comprehensive Enzyme Information System http://www.brenda.uni-koeln.de.

11 C. H. Schwab, in *Handbook of Chemoinformatics—From Data to Knowledge*, ed. J. Gasteiger, Wiley-VCH, Weinheim, Germany, 2003, p. 262–301.

12 ROTATE is available from Molecular Networks GmbH, Germany, info@mol-net.de, http://www.mol-net.de.

13 MDL Information Systems, Inc., San Leandro, CA, USA, http://www.mdl.com.

14 *Biochemical Pathways, Biochemistry Atlas*, ed. G. Michal, Spektrum Akademischer Verlag, Heidelberg, Germany 1999.

15 *ExPASy Server*, University of Geneva, Switzerland, http://www.expasy.org/tools/pathways/.

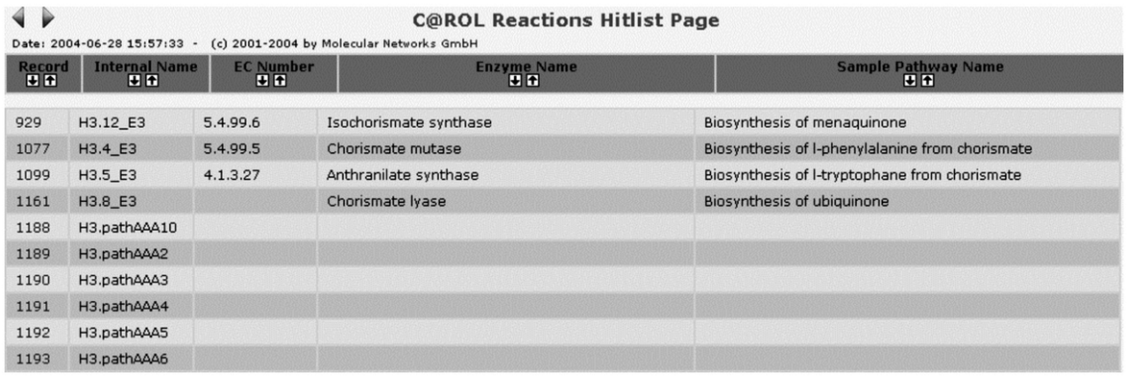
16 W. D. Ihlenfeldt, Y. Takahashi, H. Abe and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 109–116.

17 C.-C. Kanne, F. Schreiber and D. Trümbach, in *Proceedings of the 7th International Symposium on Graph Drawing (GD'99)—Lecture Notes in Computer Science*, ed. J. Kratochvil, Springer-Verlag, Berlin, 1999, vol. 1731, p. 418–419.

18 BioPath database available on the web at a) http://www2.chemie.uni-erlangen.de/services/biopath/index.html and b) http://www.mol-net.de/databases/biopath.html.

19 C@ROL is available from Molecular Networks GmbH, Germany, info@mol-net.de, http://www.mol-net.de.

20 JME molecule editor, developed by P. Ertl, Novartis, available from Molinspiration Cheminformatics, http://www.molinspiration.com.



C@ROL Reactions Hitlist Page

Date: 2004-06-28 15:57:33 - (c) 2001-2004 by Molecular Networks GmbH

Record	Internal Name	EC Number	Enzyme Name	Sample Pathway Name
929	H3.12_E3	5.4.99.6	Isochorismate synthase	Biosynthesis of menaquinone
1077	H3.4_E3	5.4.99.5	Chorismate mutase	Biosynthesis of l-phenylalanine from chorismate
1099	H3.5_E3	4.1.3.27	Anthranilate synthase	Biosynthesis of l-tryptophane from chorismate
1161	H3.8_E3		Chorismate lyase	Biosynthesis of ubiquinone
1188	H3.pathAAA10			
1189	H3.pathAAA2			
1190	H3.pathAAA3			
1191	H3.pathAAA4			
1192	H3.pathAAA5			
1193	H3.pathAAA6			

Fig. 16 Results for the query in Fig. 15, the reactions that chorismate undergoes.

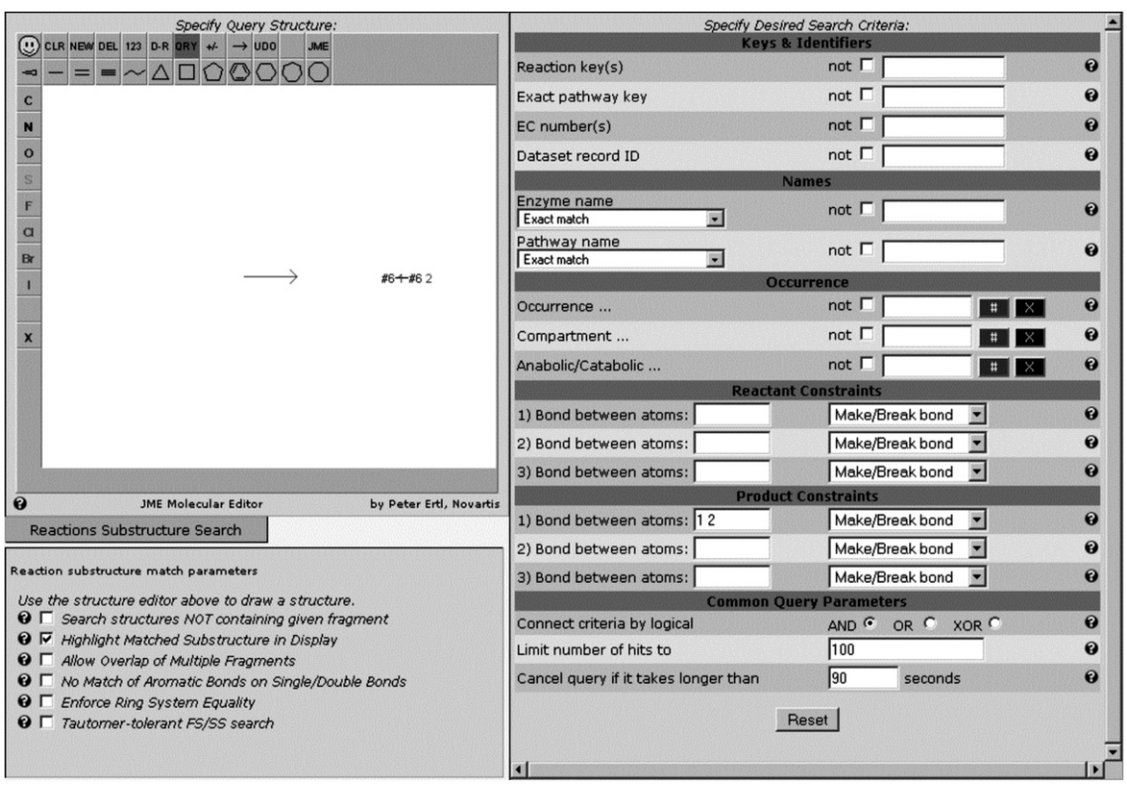
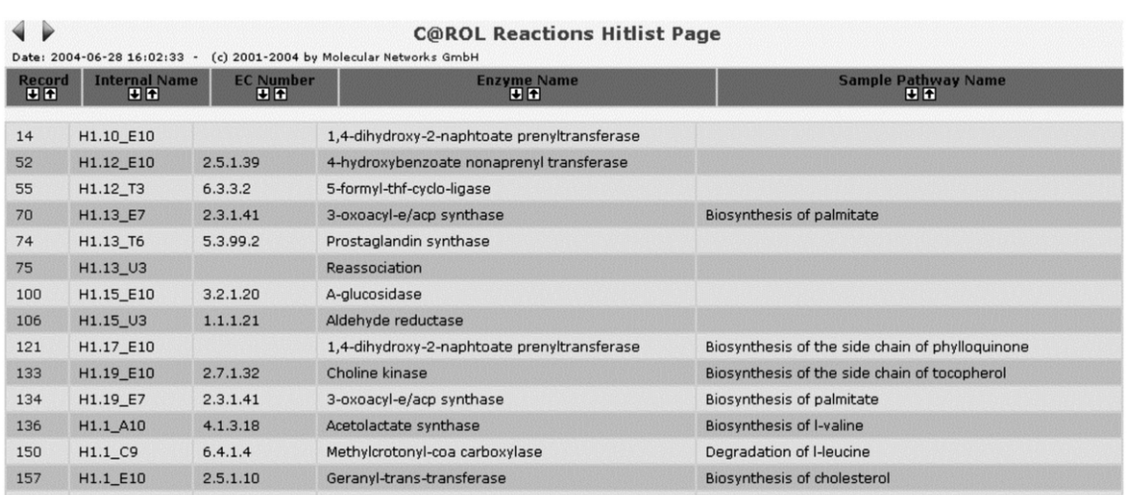


Fig. 17 Query for searching for all reactions that form a carbon–carbon atom bond.



C@ROL Reactions Hitlist Page

Date: 2004-06-28 16:02:33 - (c) 2001-2004 by Molecular Networks GmbH

Record	Internal Name	EC Number	Enzyme Name	Sample Pathway Name
14	H1.10_E10		1,4-dihydroxy-2-naphtoate prenyltransferase	
52	H1.12_E10	2.5.1.39	4-hydroxybenzoate nonaprenyl transferase	
55	H1.12_T3	6.3.3.2	5-formyl-thf-cyclo-ligase	
70	H1.13_E7	2.3.1.41	3-oxoacyl-e/acp synthase	Biosynthesis of palmitate
74	H1.13_T6	5.3.99.2	Prostaglandin synthase	
75	H1.13_U3		Reassociation	
100	H1.15_E10	3.2.1.20	A-glucosidase	
106	H1.15_U3	1.1.1.21	Aldehyde reductase	
121	H1.17_E10		1,4-dihydroxy-2-naphtoate prenyltransferase	Biosynthesis of the side chain of phylloquinone
133	H1.19_E10	2.7.1.32	Choline kinase	Biosynthesis of the side chain of tocopherol
134	H1.19_E7	2.3.1.41	3-oxoacyl-e/acp synthase	Biosynthesis of palmitate
136	H1.1_A10	4.1.3.18	Acetolactate synthase	Biosynthesis of l-valine
150	H1.1_C9	6.4.1.4	Methylcrotonyl-coa carboxylase	Degradation of l-leucine
157	H1.1_E10	2.5.1.10	Geranyl-trans-transferase	Biosynthesis of cholesterol
171	H1.1_T6	5.3.99.2	Prostaglandin synthase	

Fig. 18 Part of the list of the hits obtained with the query in Fig. 17.

21 O. Sacher, J. Marusczyk, Y. Han and J. Gasteiger, in preparation.

22 P. D. Karp, M. Riley, S. M. Paley and A. Pellegrini-Toole, *Nucleic Acids Res.*, 2002, **30**, 59–61.

23 I. Nobeli, H. Ponstingl, E. B. Krissinel and J. M. Thornton, *J. Mol. Biol.*, 2003, **334**, 697–719.

24 The free open source molecule viewer JMol is available at http://jmol.sourceforge.net.

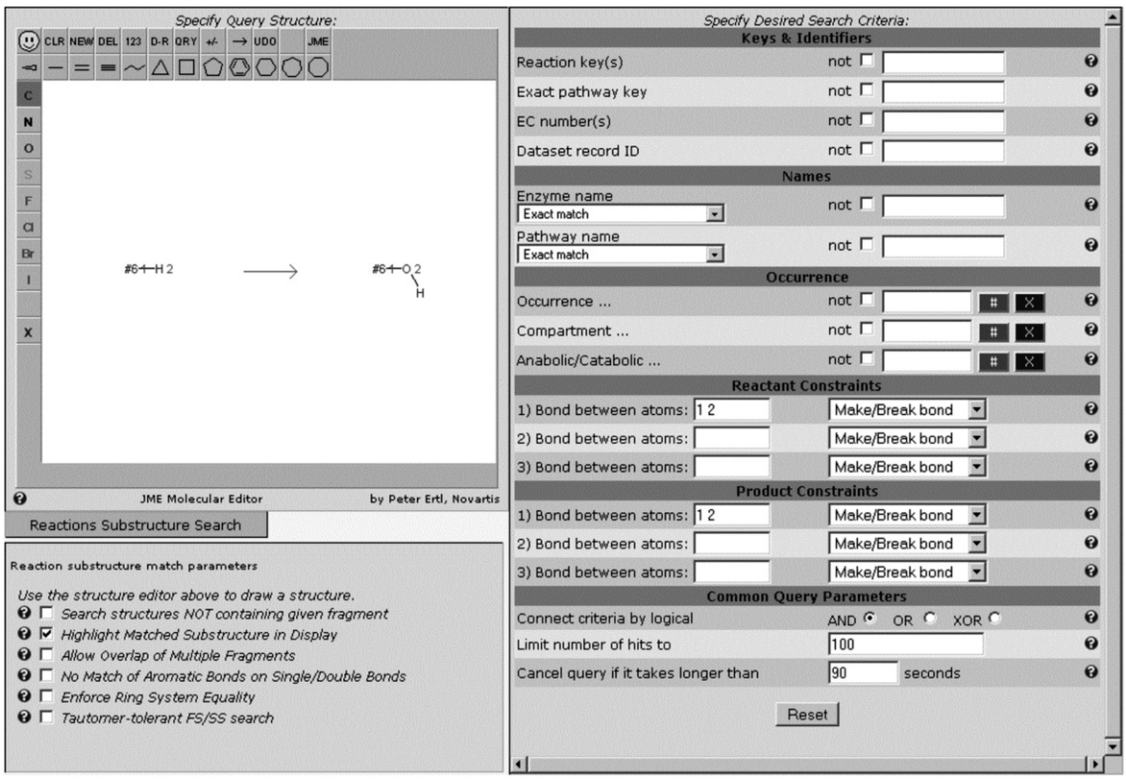


Fig. 19 Query for searching for all bonds that involve the oxidation of a C–H bond to a C–OH bond.

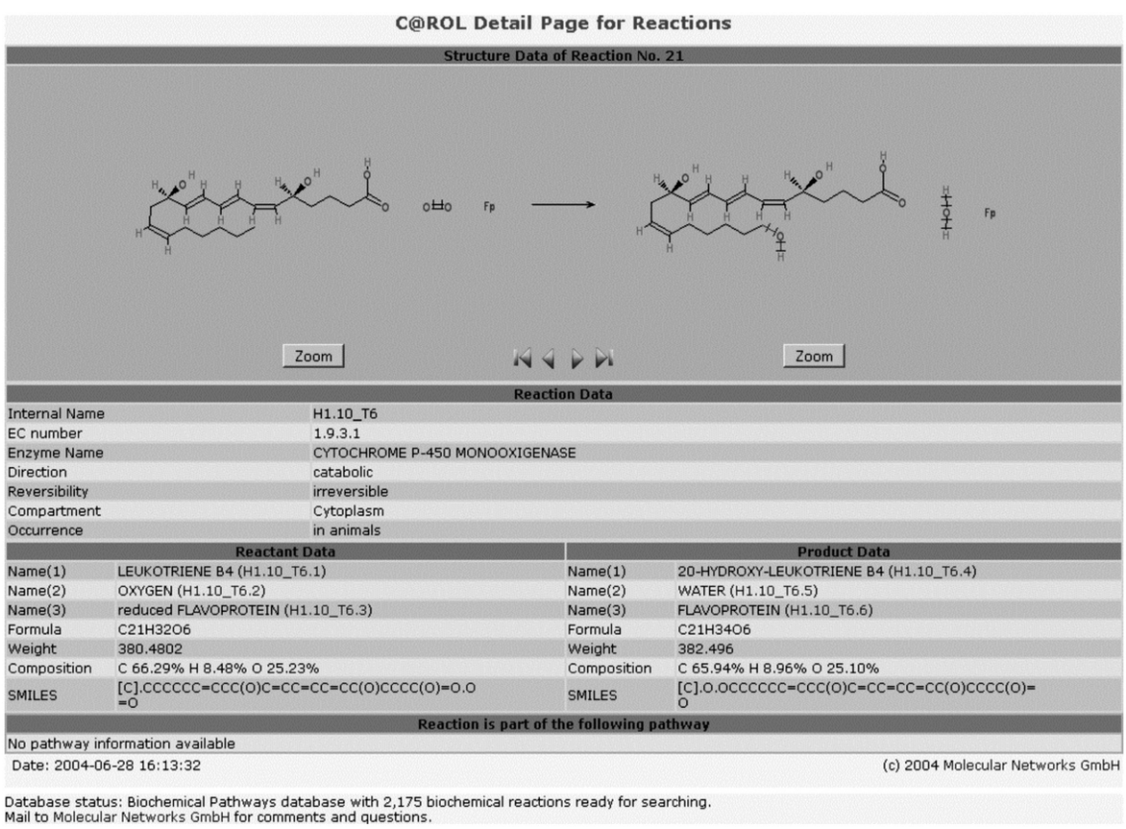
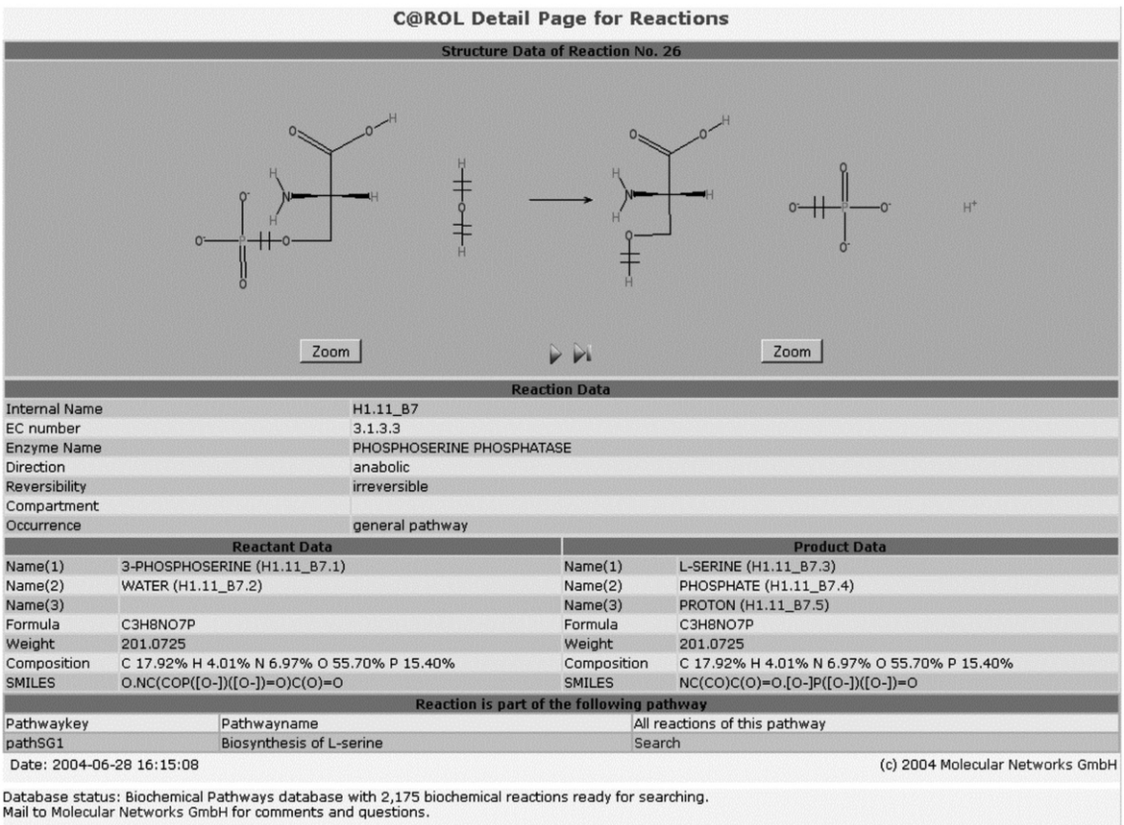


Fig. 20 One of the hits obtained in the query of Fig. 19.



C@ROL Detail Page for Reactions

Structure Data of Reaction No. 26

Zoom Zoom

Reaction Data

Internal Name	H1.11_B7
EC number	3.1.3.3
Enzyme Name	PHOSPHOSERINE PHOSPHATASE
Direction	anabolic
Reversibility	irreversible
Compartment	
Occurrence	general pathway

Reactant Data		Product Data	
Name(1)	3-PHOSPHOSERINE (H1.11_B7.1)	Name(1)	L-SERINE (H1.11_B7.3)
Name(2)	WATER (H1.11_B7.2)	Name(2)	PHOSPHATE (H1.11_B7.4)
Name(3)		Name(3)	PROTON (H1.11_B7.5)
Formula	C3H8NO7P	Formula	C3H8NO7P
Weight	201.0725	Weight	201.0725
Composition	C 17.92% H 4.01% N 6.97% O 55.70% P 15.40%	Composition	C 17.92% H 4.01% N 6.97% O 55.70% P 15.40%
SMILES	O.NC(COP([O-])([O-])=O)C(O)=O	SMILES	NC(CO)C(O)=O.[O-]P([O-])([O-])=O

Reaction is part of the following pathway

Pathwaykey	Pathwayname	All reactions of this pathway
pathSG1	Biosynthesis of L-serine	Search

Date: 2004-06-28 16:15:08 (c) 2004 Molecular Networks GmbH

Database status: Biochemical Pathways database with 2,175 biochemical reactions ready for searching.
Mail to Molecular Networks GmbH for comments and questions.

Fig. 21 Reaction obtained when searching with enzyme EC code 3.1.3.3.

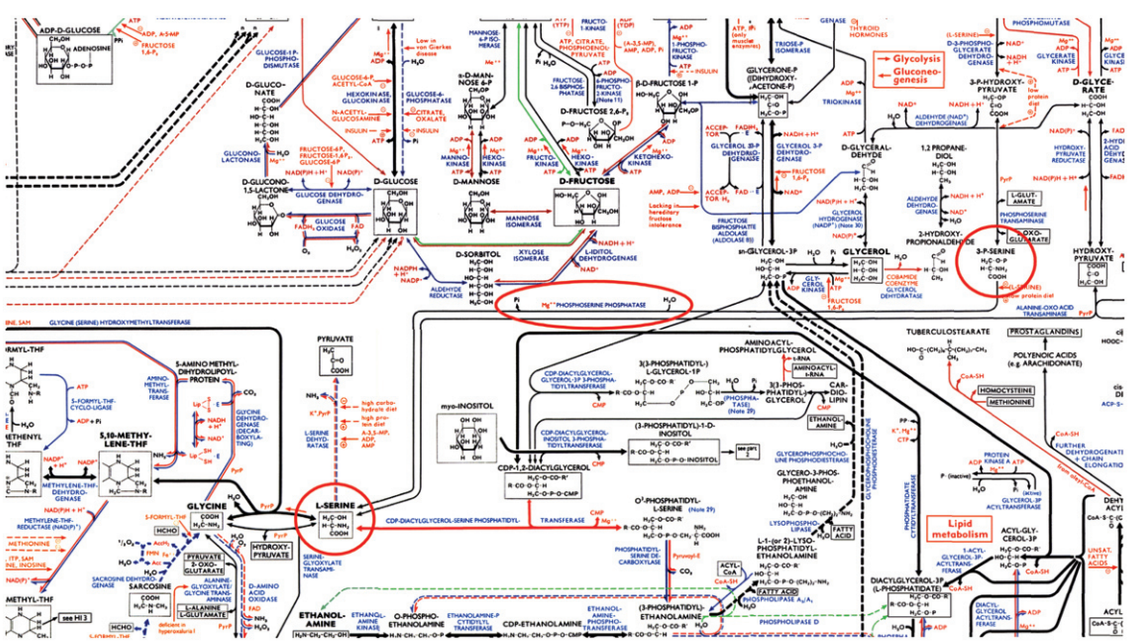


Fig. 22 The reaction of Fig. 21 as contained on the Boehringer Pathways Wall Chart and the ExPASy server; from Michal, Biochemical Pathways 1999 © Elsevier GmbH, Spektrum Akademischer Verlag, Heidelberg.