Rule-Based Expert System to Infer Biological Pathways in *Rhodobacter sphaeroides*

By
Maritza Elizabeth Córdova Bermeo

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCES

in

COMPUTER ENGINEERING
(Bioinformatics)

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2005

Approved by:

_____	_____
Manuel Rodríguez Martínez, Ph.D. Member, Graduate Committee	Date
_____	_____
Pedro I. Rivera Vega, Ph.D. Member, Graduate Committee	Date
_____	_____
Jaime E. Ramírez Vick, Ph.D. President, Graduate Committee	Date
_____	_____
Genock Portela Gaudier, Ph.D. Representative of Graduate Studies	Date
_____	_____
Isidoro Couvertier, Ph.D. Chairperson of the Department	Date

Abstract

Rule-Based Expert System to infer Biological Pathways in Rhodobacter sphaeroides

By

Maritza Elizabeth Córdova Bermeo

A better method for functional analysis of the genes is the identification of biological pathways, since these provide information about the expression and regulation of genes. The search of these pathways is a complex task due to the intricate of its relations and data heterogeneity. The identified pathways in *Rhodobacter sphaeroides* are almost nonexistent, which is necessarily a method to help to this intention.

A biological pathway prediction rule-based expert system was developed in JESS, which is capable of inferring reactions using its inherent recursive capacity, between or around specific molecules (either proteins or xenobiotics) connecting individual reactions steps, which are stored in a database for the later validation by human experts. It is used a database called *BioPathDB*, which contains information on the bacterium *R. sphaeroides*, which is populated from biological databases existing on the Internet and/or in data published in peer-reviewed journals.

Resumen

Sistema Experto basado en reglas para Inferir Rutas Biológicas en "*Rhodobacter sphaeroides*"

Por

Maritza Elizabeth Córdova Bermeo

Un mejor método para el análisis funcional de genes es la identificación de rutas biológicas, dado que estas proveen información sobre la expresión y regulación de genes. La búsqueda de estas rutas es una tarea compleja debido a lo intrincado de sus relaciones y la heterogeneidad de la data, la identificación de rutas para R.sphaeroides, es casi inexistente, por lo que se es necesario un método para ayudar a este propósito.

Un sistema experto basado en reglas, para predicción de rutas biológicas fue desarrollado en JESS, es capaz de inferir reacciones usando su inherente capacidad recursiva, entre o alrededor de moléculas específicas (proteínas o xenobióticos) conectando pasos de reacciones individuales que serán almacenadas en una base de datos para la validación posterior por expertos humanos. Se usa una base de datos llamada "BioPathDB", conteniendo datos de la bacteria "Rhodobacter sphaeroides", que es poblada desde bases de datos biológicas que existen en Internet y/o en revistas científicas.

To God.

To my parents Máximo and Ursula for their love and dedication.

Acknowledgements

I express a sincere acknowledgement to my advisor, Dr. Jaime Ramírez-Vick because he gave me the opportunity to research under his guidance, for his imparted knowledge and for his friendship during theses years. I also want to thank to Dr. Manuel Rodriguez Martinez and Dr. Pedro I. Rivera Vega graduate committee members for reviewing my work and for giving me their valuable contributions. Special gratitude I owe to Dr. Carlos Rios Velazquez for transmitted knowledge for the completion of my work and to Dr. Nestor Rodriguez for give me the opportunity of work with his researching under his supervision in other area, his support and confidence in me.

I would also like to thank Dr. Timothy Donohue at the Pennsylvania State University, to Dr. Samuel Kaplan, Dr. Chris McKenzie, from the *Rhodobacter sphaeroides* Genome Project at the University Texas – Houston, for the invaluable help they provided during my search for new sources of information and especially to Dr. Larimer for giving me invaluable *R. sphaeroides* data. Also to the students Migcadeliz Oliveras, Myasotis Caro, Maria Perez, Nicole Ramos, Marcelo Quiles, Rafael Gonzalez, Joel Berrios and Cristian Sosa, for their help in the arduous and delicate work of populating the database with pathway data.

I also want to thank the Department of Electrical and Computer Engineering at University of Puerto Rico at Mayagüez for this opportunity to follow graduate studies and to my friends, which help me in the development of this thesis with his breath words and good advices.

At last, but the most important I would like to thank to my parents Máximo and Ursula and sisters Mari, Carolina, Marcela, Lyliana, Carmen and Ursula, for their unconditional support, inspiration and love, and specially to my husband, partner and friend Arturo for his love, dedication and support in all aspects of my life during all theses years, and to my son, Rodrigo, my major inspiration, breath and motivation.

Table of Contents

List of Tables

List of Figures

Chapter 1

1. Introduction

1.1. Overview

Biology has rapidly become a data-rich, information-hungry science because of recent massive data generation technologies. The biologists are designing more clever and informative experiments because of recent advances in science. These experiments and data hold the key to the deepest secrets of biology and medicine, but we cannot fully analyze this data due to the wealth and complexity of the information available. The result is a great need for intelligent systems in biology [24].

Genomic information integrated with biological data and presented from a pathway is rather than from the DNA sequence perspective, due that it does not reflect the context in which most gene act, i.e., functionally related genes are usually not physically clustered in DNA, but instead are distributed among distant sites. The protein products of these genes assemble at appropriate cellular locations to coordinate their biological functions. Thus an alternative to the DNA sequence for presenting and studying genomic information are the use of biological pathways, as mentioned [23].

The identification of realized pathways provides valuable information on gene expression and regulation. Detection of incomplete pathways helps to improve a constantly evolving genome annotation or discover alternative biological pathways [9] that help to identify novel genes and verify a current annotation.

These types of problems have been the focus of research in expert systems, and artificial intelligence in general. Specifically, on the development of techniques, which allow the modeling of information at higher levels of abstraction, as showed in [24], [23], [9], [36]. These techniques are embodied in languages or tools, as CLIPS and JESS, which allow programs to be built that, resemble human logic in their implementation and are therefore easier to develop and maintain.

In this research, we implemented a prototype rule-based expert system built on the Java Expert System Shell (JESS). This system is capable of inferring reactions between or around specific components (either proteins or xenobiotics) connecting individual reaction steps to find possible biological pathways. These pathways will be stored in the database until properly validated by a human expert using published data. The local database *BioPathDB* is managed by MySql database management system (DBMS), which stores *R. sphaeroides* data from published sources or from web-based external public biological databases.

1.2. Problem Formulation

The value of genomics and bioinformatics knowledge is enormous but will only be fully realized when knowledge and information are integrated at many levels. The idea of integrating all the biological and genomic knowledge is key for an efficiently reaching the understanding of how cells and organisms function. These depend on the current knowledge of biochemical reaction pathways which describe how molecules interact within cells to keep homeostasis (also known as health). For this reason, the study of biochemical reaction

pathways is the focus of much research and is the central theme for research in biopharmaceutical and genomic companies.

The problem to implement biochemical reaction pathways is that many of the chemical reactions involved in these pathways are unknown. In addition, there is no tabulated data existing in the databases, the sequence and expression of the genes associated with pathways are unknown, or the data stored of these is heterogeneous not allowing cross-referencing.

All these factors do not allow the integration of this knowledge into useful information for the users. The main limitation of current methods is that they still need a human expert to make the final integration in the form of a hypothesis, which will be either proved or disproved by the new knowledge acquired. The type of human reasoning required for this task can be artificially performed through computational means by Expert Systems.

1.3. Justification

To facilitate the use of genomic data, a new modeling perspective is needed to examine and study genome sequences in the context of many kinds of biological information. Pathways are the logical format for modeling and presenting such information in a manner that is familiar to biological researches [23].

Due, to the importance of biochemical reaction pathway knowledge for the understanding of biological and genomic information, we propose to implement a rule-based expert system capable of inferring reactions by connecting individual reaction steps using a set of rules over a relational database. This database will focus on the bacterium *R. sphaeroides* and will be managed using the MySQL Database Management System (DBMS).

This expert system builds *de novo* biochemical reaction pathways based solely on known single reactions in case that they have not being realized in the fragmented literature source from which all biological pathway maps are derived. The possible pathways found by the expert system will be stored in the *BioPathDB* database to be later validated by human experts. This will be a highly flexible database that will allow any modifications necessary to handle diverse and complex biological information.

The bacterium *R. sphaeroides* was chosen as a model system to implement this prototype expert system following fundamental criteria, mainly that there is complete knowledge on its sequence, partial knowledge on its biochemical reaction pathways, and because of the widespread interest in its study.

To populate the database data we will import the data from external biological databases such as the Swiss-Prot database, the Ligand and Pathway databases of Kyoto Encyclopedia of Genes and Genomes (KEGG), the Biocatalysis/Biodegradation Database, the Database of Integrated Microbial Genomes system (IMG), and data published by researchers in peer-reviewed journals.

1.4. Objectives

The main goal of this study is to design and develop a rule-based expert system using the JESS inference engine, which applies rules to find and connect the pathways in which a target molecule is involved or those connecting two target molecules (initial an end) in the our *BioPathDB* database.

The modeling and implementation of the *BioPathDB* relational database allows the storing of information about *R. sphaeroides,* which includes molecules (either proteins or xenobiotics), known reactions, and pathways including those inferred by the system but not corroborated experimentally. In addition, toolkits were developed to allow the population of the database using other web-based databases. These toolkits will consider the file format in the source database and will change its structure into the *BioPathDB* database model. The expert system and toolkits will be available on a web application.

1.5. Research Methodology

The following steps were performed to accomplish the objectives of the proposed research work:

- Literature review on the fundamental principles involved in the biological aspects such as construction of the biochemical reaction pathways; biological databases and software, expert systems, JESS embedded into applications written in Java, and the about on *R. sphaeroides*.

- Understanding and analysis of the existing biological databases such as Ligand at KEGG and specifically from databases storing *R. sphaeroides* data such as Swiss-Prot, The University of Minnesota Biocatalysis/Biodegradation Database and Integrated Microbial Genomes.

- Design and modeling of the *BioPathDB* relational database in MySQL DBMS, to support *R. sphaeroides* data from external databases and from other published sources, considering the information engineering notation and its file format.

- Create a knowledge base, which contains the rules used by the inference engine to find interconnections among the reactions stored in the database.

- Analysis the requirement, features, design and develop an internet application that allows the use of an expert system and imports data from external biological databases.

- Populate the *BioPathDB* database from biological databases which contain *R. sphaeroides* data (*e.g.,* Swiss-Prot, IMG, etc.) through a software tool developed on a Java-Servlet.

1.6. Contributions

This expert system will provide biologists with a prediction tool for new biochemical pathways in *R. sphaeroides*, which are generated from existing interconnections between

reactions of different pathways of other organisms, but in which *R. sphaeroides* genes have been identified. In addition, this expert system is flexible enough to be used with other organisms just by populating the *BioPathDB* database with the necessary data.

The database model used is capable of optimizing the space used for storing the information, in contrast to other biological databases that normally use flat format files.

The possibility of updating the data via web will allow users to improve the results obtained with the expert system.

1.7. Thesis's Structure

The remainder of this thesis is organized as follows. Chapter 2 provides information about related previous work in biological databases, software, and expert system applications. Chapter 3 provides a general introduction about biological aspects, biological databases, and expert systems and about *R. sphaeroides*. Chapter 4 provides a description of the rule-based expert system developed. We describe the prototype, the architecture behind it, and its special features. Likewise, we describe the *BioPathDB* database and the alternatives to populate it. Chapter 5 presents analysis of the result and finally, we present the conclusions, future work, references and the respective appendices.

Chapter 2

2. Related Previous Works

2.1. Overview

In this section we present and discuss some of the work related to biological databases and related software and expert system applications.

2.2. Biological Databases and related Software

Most of the efforts in bioinformatics have been to organize the data generated by worldwide genomics efforts including the Human Genome Project and other public and private genetic sequencing projects. Currently these efforts have moved beyond this to include areas such as gene expression, protein identification and structure, biochemical pathway data, pharmacogenomics, and chemical structure and activity.

There are currently integrated pathway-genome databases and systems which can be accessed over the internet, that allow access to the current knowledge on biochemical pathways (mainly metabolic), from sequence information to reaction networks. These integrated databases describe genes and genome of an organism using known metabolic pathways as the framework from which the current knowledge about their reactions, enzymes and metabolites can be accessed from other databases. These usually include visualization and analysis software along with the textual information. Examples are the EcoCyc *E. coli* database from Double Twist [5], [18], the Kyoto Encyclopedia of Genes and Genomes

(KEGG) from Kyoto's University Institute of Chemical Research [20], [11], the SoyBase at Iowa State University, the WIT (What Is There) project at Argonne National Laboratory [34], and GeneNet from the Russian Academy of Sciences [22]. All these examples, represent pathways that refer to their component reactions, which are encoded in binary relationships between reactants, products, and enzymes that catalyze the reactions [36]. The pathway visualization that is part of these system goes from being automatically generated (EcoCyc, SoyBase, and GeneNet) to pre-made maps (KEGG); even though in the latter the maps are generated using a deductive database method, in the former, the generation of these pathway maps is based on pre-specified interconnections between reaction nodes.

KEGG is a suite of databases and associated software, integrated current knowledge on molecular interaction networks in biological processes. Its data is distributed in different databases among them Ligand and Pathway which are modeled using object oriented methodology. The first database is a database of chemical compounds and reactions in biological pathways, which is designed to provide a nexus among the chemical and biological aspects in enzymatic reactions. This database consists of three sections: Molecule, Reaction and Enzyme [20]. We have used this data to populate our database since it is essential to be able to obtain the binary reactions that will be used for inferring biochemical pathways by our expert system. The second database stores pathways in the form of pre-made maps stored in graphical format. These maps have been inputted manually into our database.

EcoCyc Encyclopedia is a database that stores data for the bacterium *Escherichia coli* K-12 MG1655. This is included in the BioCyc Database Collection, which is a collection of pathway/genome databases for microorganisms and humans [5]. This database provides

several different mechanisms for querying Pathway/Genome Databases through of Pathway Tools software, which include three components, PathoLogic, Pathway/Genome Navigator, Pathway/Genome Editors. The first allows the creation of new Pathways using a Genbank entry as input, the second allows query, visualization and analysis of pathways and the last provide the capability of edit the pathway. This tool is valuable for organisms for which the whole genome is known since they require sequence information.

GeneNet is a system for description, visualization and modeling of gene networks. The database structure is in flat format to support an object oriented approach, and it is similar the structure the SwissProt database. This system cannot predict pathways since it only allows visualizing the data stored in its database [7].

In addition, different databases contain specialized information of certain features of one or more organisms. Among them the sequence databases (e.g., in GenBank or EMBL) contain information about DNA and protein sequences. The structure databases contain primary and secondary protein structures obtained from X-ray crystallographic and NMR data. An example is SCOP (Structural Classification of Proteins) of the University of California-Berkeley that stores structural and evolutionary relationships between all proteins whose structure is known and in addition, a list of protein of different species classified by class, folds, super families, and families [33]. The reference databases (e.g., PubMed, MedLine, etc.) store bibliographic data and biological knowledge of the organisms. Many of these databases are also integrated into other databases such as KEGG.

Public databases and projects provide information about of *R. sphaeroides*. These databases help us in the understanding of *R. sphaeroides* and allow us to populate the

BioPathDB database. Among these are the *R. sphaeroides* genome projects at the University of Texas – Houston that organizes the genomic information of *R. sphaeroides*, and describes the genomic features at the chromosome (it has two: CI and CII) and DNA sequences level. These data are related to both CI and CII previously sequenced and deposited into the GenBank [17]. This database is on web site: http://mmg.uth.tmc.edu/sphaeroides/.

The Genome Analysis and System Modeling Group of the Life Sciences Division at Oak Ridge National Laboratory also provides *R. sphaeroides* genomic sequence data, including computational genome and protein structure analysis tools [DOE03], protein coding genes and RNA genes, and also show those genes connected to KEGG pathways in graphical format [14].

In addition, the Biocatalysis/Biodegradation database (UM-BBD) of the University Minnesota provides information about microbial biocatalytic reactions and biodegradation pathways, but primarily for xenobiotics. This database has an *R. sphaeroides* pathway fully identified, available in graphical format, named the Dimethyl Sulfoxide and Organosulfide Cycle Map [UMBBD05].

Finally the Swiss-Prot database, which is a protein knowledgebase established in 1986 and maintained by the Swiss Institute for Bioinformatics and the European Bioinformatics Institute. It is a curated protein sequence database which provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases [35]. Among its organisms is *R. sphaeroides*, for which reason we import its data to our database.

In addition there are tools to retrieve pathways using a different perspective (e.g., enzyme, sequence, annotation, etc.) which have been previously identified and stored in a database in graphic or text format.

Among the tools available to generate pathways is ***PathAligner***, which is a tool to reconstruct/retrieve metabolic pathways from gene, sequence, enzyme, and metabolic data. It also provides an alignment to compare the similarity of pathways [3]. This tool only allows the retrieval of pathways from one molecule, but the results show limited information as shown in the figure 2.1 and figure 2.2, although they reference KEGG for more information.

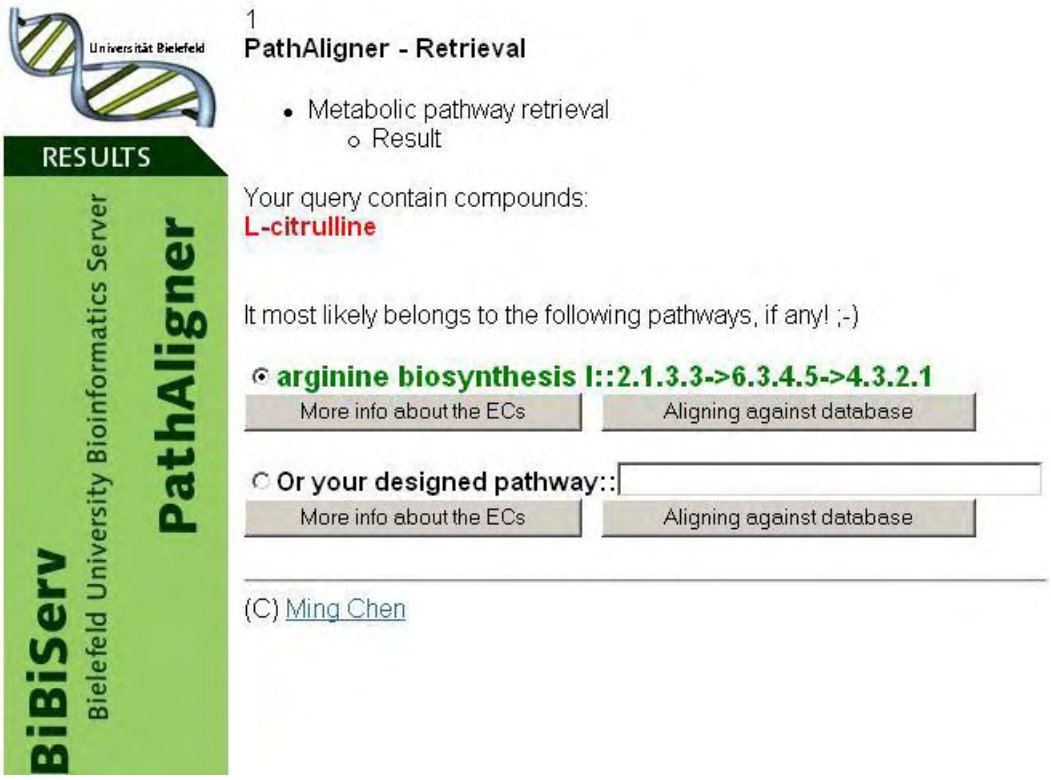

Figure **Error! No text of specified style in document.**.1. Result of query [3].

Figure **Error! No text of specified style in document.**.2. Result of query (Cont.) [3].

Another tool is ***PathFinder*****,** which is used for the dynamic visualization of metabolic pathways based on annotation data or EC-number set. It is an interactive web application that reads a list of EC-numbers or a given annotation in EMBL or GenBank format [13]. When the input data has high levels of detail about biochemical pathways this tool helps to identify genes and detect mistakes within the annotation. The tool Visualizing Metabolic Networks works in a three dimensional space producing a graphical representation of a metabolic network using the VRML (Virtual Reality Modeling Language). It only allows the pathway to be drawn manually using a graphics tool and then stored in a graphic format [32]. Finally, another tool called PRIAM uses a method for automated enzyme detection in a fully sequenced genome, based on all sequences available in the ENZYME database [2]. This tool only allows the prediction of metabolic pathways and it is valuable only if the enzyme coding genes has been identified since the prediction is based on enzymes. It is thus necessary to know the complete genome of the organism.

Other software available provides only general information of biological databases from integration of various biological databases such as the software developed by Riikonen et. Al, which allows to access to them through wireless protocol [31], in addition there is a system for extracting data from different databases at same time, while supporting fixed-form queries [28]. Another system is based on the integration of heterogeneous biological databases that supports the integration of data sets [25], and another with a federated architecture and mediator that integrates access to heterogeneous, distributed biological databases [12].

2.3. Expert System application

Currently there are some expert systems developed within bioinformatics, specifically for the prediction of biological pathways of any organism. Instead, there are many expert system oriented to geographic systems [42], automated documents [Mohamed90], diagnostics of fail and medical applications, in which the users interact directly with the system through question-answers. The majority of these systems have been developed in CLIPS.

An expert system related to bioinformatics is the expert system based on annotation strategies which is an open source genome annotation system that allows both manual and automated annotation of prokaryotic genomes. This system implements a rule-based auto annotation tool, intended to simulate the annotation process of a human expert. This auto annotator is capable of simulating the user's decision of how to annotate a specific gene [26].

Another expert system in this area and on which we have based it is the expert system developed in the cell signaling network database in human cell at the National Institute of Sciences of the Health in Japan [36]. This system is able to infer reactions between two molecules by connecting individual reaction steps using a set of rules stored in what is known as a knowledge base. This means that this system does not build the pathways based on pre-existing models but instead from the individual reaction steps. This is very important when considering that most information obtained from either databases or research articles is fragmented. Unfortunately, for reasons unknown this system is currently not available over the Internet.

Chapter 3

3. Background and Significance

3.1. Overview

In this section we discuss the biological and computational aspects necessary to understand the environment of the system.

3.2. Biological Aspect

According to the Medical Research Council, bioinformatics is the development and use of computational and mathematical methods for acquiring, storing, analyzing and interpreting biological data to solve biological question [27]. In addition, Bioinformatics can be defined as the informatics applied to molecular biological and genetics (networks and genome, microarrays, etc.) whose main goal is discovery new knowledge through the integration of biological knowledge. This new biological knowledge with the already existing knowledge will allow the understanding of biological processes that occur in organisms and is organized into genes and pathways.

Genes, gene structure and gene variations connect, through gene expression to regulatory and metabolic processes within cells. The fine control of these processes is modulated by spatial and temporal protein expression within both cells and tissues. Individual cells communicate through direct interactions as well as remotely through hormones and other secreted products, which connect the intracellular pathways, or networks,

of all cells to each other. The sum of these pathways defines the phenotype of the organism, which, when all are working normal limits, is at homeostasis. A shift away from this state leads to a disease phenotype, which might be represented by a new homeostatic set point or be progressive. This shift will be the result of changes in the networks that establish these set points. The changes in intracellular pathways would be the result of changes in gene expression or gene sequence.

Pathways are the sequential and cumulative action of genetically distinct but functionally related molecules. Each reaction in each pathway begins with specific substrates; use various combinations of molecules as cofactors, activators, and inhibitors, and ends with products that are chemically modified substrates. Thus pathways are an appropriate format for representing the functional role of most genes in the genome. [23].

In a pathway the product of one reaction is often the reactant (substrate) for the next, forming a lineal chain of reactions. Many pathways have branches that provide alternate methods for nutrient processing. Other takes a cyclic form, in which the starting molecule is regenerated to initiate other turn of the cycle. Pathways generally do not stand alone; they are interconnected and merge many sites [37].

Is important give a definition of pathway that can accommodate its use in science and our extension of its scope to expert systems: A pathway is an ordered set of R finite steps *{step1, step2,..., stepR}*, each of the form *reactants* $\Rightarrow$ *products*, such that the reactants of step *i* are a subset of the species *SM* $\cup$*products(1)* $\cup$ ... $\cup$ *products(i-1)*, where SM are "starting molecules" that are given, and products(k) are the products formed at step k. That

is, each reactant of a step must either be a starting material or have been formed as a product of a prior step [40].

For infer pathway is necessary know partial or completely the sequencing of genome as its assembly and annotation, which is explained in the figure 3.1.

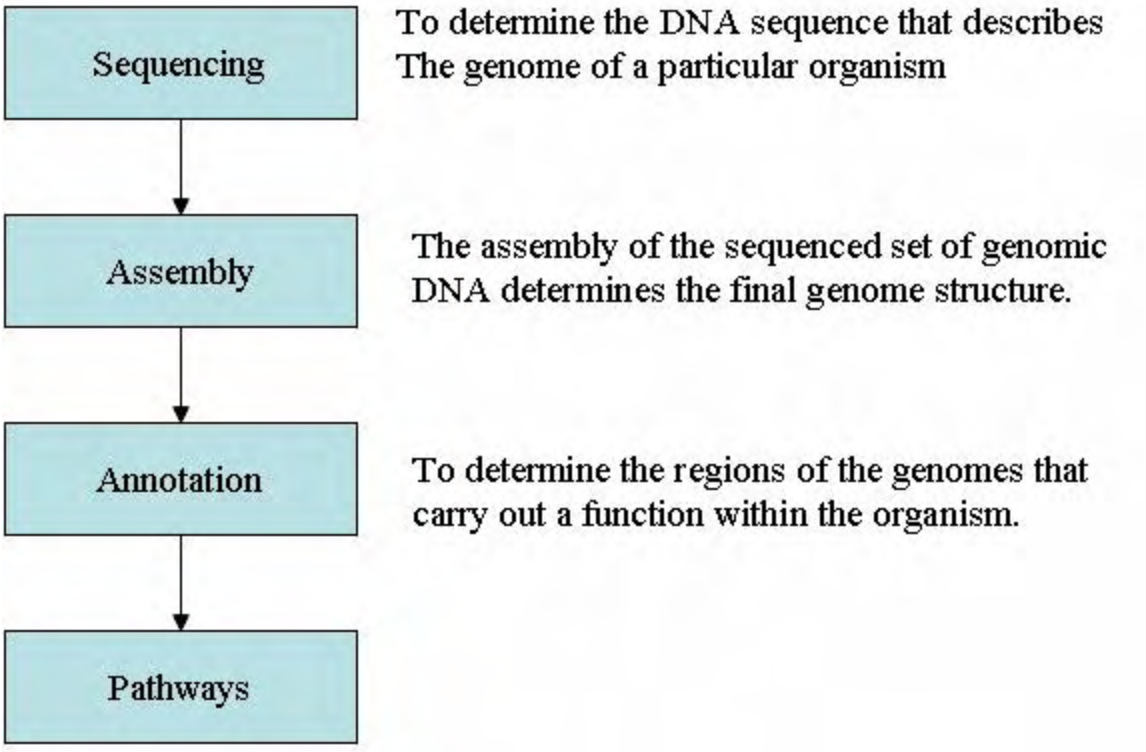

Figure **Error! No text of specified style in document.**.1. Process of inferring pathways [13].

There are three general classes of biological pathways: metabolic and biochemical; transcription, regulation and protein synthesis; and signal transduction [Krishnamurthy03]. The first consist of a linked series of individual chemical reactions that provide intermediary metabolites and lead to a final product. These pathways are catalyzed by an enzyme [37]. Its

reconstruction involves inferring the metabolism of an organism from its genetic sequence data supplemented by annotation data [13]. The second type is responsible for converting genetic information into proteins (genes products). The third, Signal transduction pathways are responsible for coordinating metabolic processes with transcription and protein synthesis [Krishnamurthy03].

3.3. Biological Databases

A database is a logically structured dataset and relationships which provides useful information. The organization of its data depends on the database management system, it can be: plane (which does not verify consistency of data), hierarchic, relational (tables, records, attributes and relations) and object oriented (class, object and message).

Beginning from the concept of a database, a biological database is a collection of raw data, theoretical or experimental, organized and stored in a logical structure.

The biological databases were born 20 years ago as simple repositories to store biological sequences, but with the increase of the technology efficiency in experiments applied to the biology, these repositories have turned into enormous mountains of results expecting to be analyzed, likewise the number of databases existing storing variety of information has increased, forcing to look for ways to treat the information in an automatic form and while different levels of abstraction [6].

The molecular biology databases available for public access constitute an emerging information medium, which is shared by medicine, biology, biotechnology, physiology, etc. [32]

Biologically, the databases are grouped in the primary, secondary and mixed databases. The first are a data repository derivates of experiment or scientific knowledge (Genbank, Swiss-Prot, PubMed, KEGG, etc.) while the second stores data from other sources such as primary databases (Refseq: data collection curate of GenBank in NCBI). The mixed database stores data from the primary and secondary databases, examples of these databases are those that store organism specific data [1].

These databases contain specialized information of certain features of one o more organisms, which we can classify them of the following way.

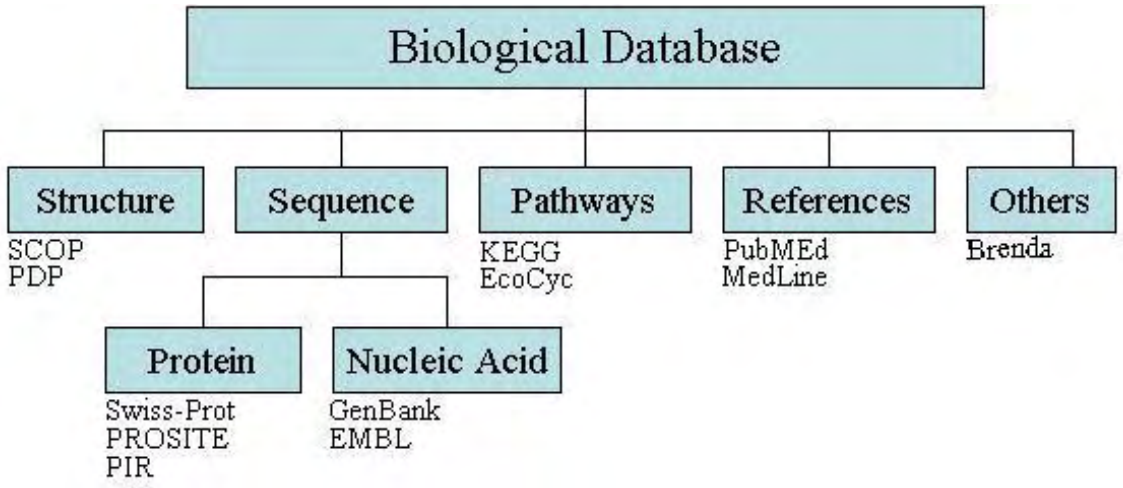

Figure **Error! No text of specified style in document.**.2. Classification of biological database.

Of all these, the pathway databases raise many important and challenging computational and bioinformatics issues, such as querying and visualizing graph structured database in multiples abstraction levels, seamless integration of data distributed in diverse sources; integrated, graph-based querying and navigation of data in multiple dimensions, i.e., from biological function to gene expression [23]

3.4. Knowledge-based expert system

Edward Feigenbaum, pioneer of expert systems technology, has defined an expert system or knowledge-based expert system as "an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solutions." That is, an expert system is a computer system that emulates the decision-making ability of a human expert. The terms emulate means that the expert system is intended to act in all respects like a human expert. [8].

Expert systems are a branch of Artificial Intelligence (AI) that makes extensive use of specialized knowledge to solve problems at the level of a human expert. An expert is a person who has expertise in a certain area.

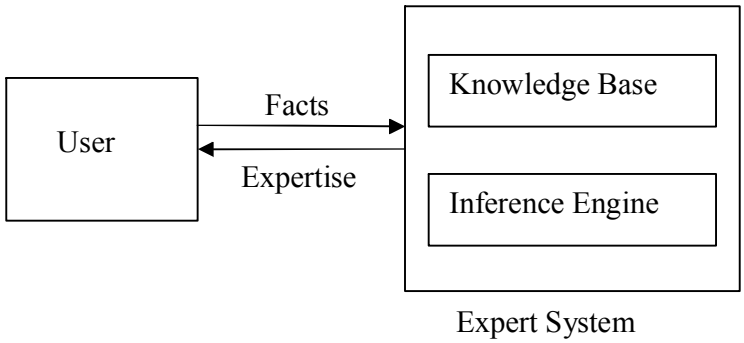

Figure **Error! No text of specified style in document.**.3. Components of expert system.

Internally, the expert systems consist of two main components: Knowledge base and the inference engine. The first is the text file that contains all the rules used by second, which controls the execution of rules [41].

The user supplies facts or other information to the expert system and receives expert advice or expertise in response. The Fact-list, and Instance-list is global memory for data, then facts are data that usually designate relations or information like: (is-animal duck) or (animals duck horse cow chicken). Rules can be applied on these facts in the form of *IF-THEN* rules [Van99].

Expert systems have a number of attractive features:

- Increased availability, expertise is available on any suitable computer hardware.

- Reduced cost; the cost of providing expertise per user is greatly lowered.

- Reduced danger because it can be used in environment that might be hazardous for a human.

- Permanence since the expertise is permanent. Unlike human experts, who may retire, quit or die, the expert system's knowledge will last indefinitely.

- Multiple expertise, the knowledge of multiple experts can be made available to work simultaneously and continuously on a problem at any time of day or night.

- Increased reliability due to increased confidence, should always agree with the expert, unless a mistake was made by the expert. However, this may happen if the human expert was tired or under stress.

The knowledge of an expert system may be represented in a number of ways (rules, semantic nets, frames, logic, etc.). It can be encapsulated in rules and object. One common method of representing knowledge is in the form of IF … THEN type rules.

3.4.1. Rule-based programming

Rule-based programming is one the most common used techniques for developing expert system. In this programming paradigm, rules are used to represent heuristics, or "rules of thumb", which specify a set of actions to be performed for a given situation. A rule is composed of an *if* portion and *then* portion. The *if* portion of a rule is a series of patterns which specify the facts (or data) which cause the rule to be applicable. The process of matching facts to patterns is called pattern matching. The expert system provides a

mechanism, called the inference engine, which automatically matches facts against patterns and determines which rules are applicable. The *if* portion of a rule can actually be thought of as the *whenever* portion of a rule since pattern matching always occurs whenever changes are made to facts. The *then* portion of a rule is the set of actions to be executed when the rule is applicable. The actions of applicable rules are executed when the inference engine is instructed to begin execution. The inference engine selects a rule and the actions of the selected rule are executed (which may affect the list of applicable rules by adding or removing facts), then selects another rule and executes its actions. This process continues until no applicable rule remains [15].

One of the most well-known tools for the development of rule-based expert systems over Internet is Java Expert System Shell (Jess) that is a rule engine and scripting environment written entirely in Sun's Java language by Ernest Friedman-Hill at Sandia National Laboratories in Livermore, CA [16]. This tool was inspired by the "C" Language Integrated Production System (CLIPS) environment developed by NASA in 1986 [19], but has grown into a complete, distinct, dynamic environment. Both provide a complete environment for the construction of rule and/or objects based expert systems. [4][16].

Like CLIPS, Jess's inference engine implements standard forward-chaining, modules, and uses the RETE pattern-matching algorithm to process rules, a very efficient mechanism for solving the difficult many-to-many matching problem. I addition it also has seven strategies (i.e., Depth, Breadth, Simplicity, Complexity, Lex, MEA and Random) in rule firing. Besides it adds many features such as working in memory queries, and the ability to create, manipulate and directly reason Java objects and call Java methods without compiling any Java code [16].

Jess has syntax similar to that of LIPS and it is compatible with CLIPS, in that many Jess scripts are valid CLIPS scripts and vice-versa. But it has the advantage that it can be embedded into other applications written in java, where java threads can be used to run in parallel multiples expert systems, especially into web applications, whereas CLIPS use the JClips library (http://www.cs.vu.nl/~mrmenken/jclips/) [Menken04], although some people prefer to use CLIPS with C++ in a CGI script. For this reason we chose Jess instead of CLIPS since our application is a web -based and requires java.

3.4.2. Rule-based expert system structure

A typical rule-based experts system consists of the following components:

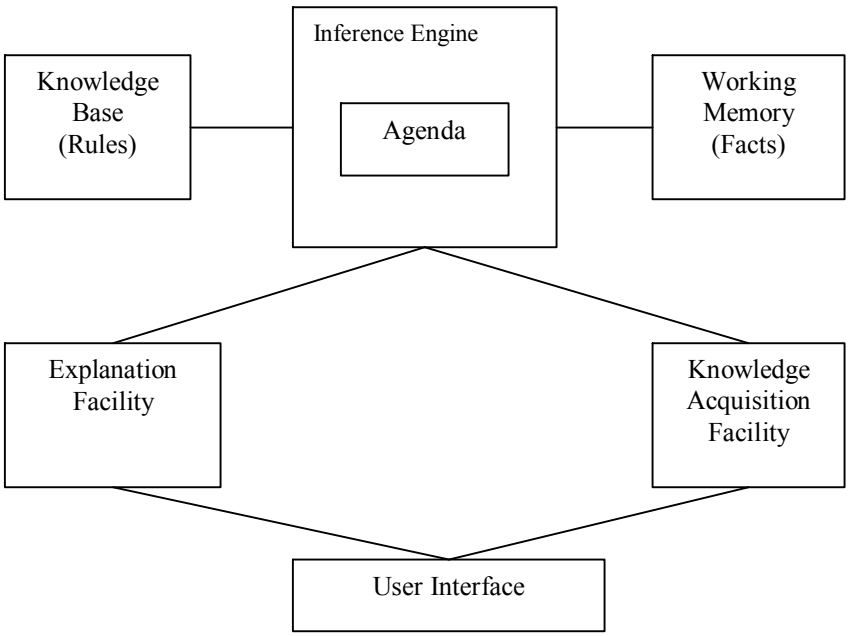

Figure **Error! No text of specified style in document.**.4. Structure of rule-base expert system [8].

User interface: The mechanism by which the user and the expert system communicate.

Explanation facility: Explains the reasoning of the system to a user.

Working memory: A global database of facts used by the rules.

Inference engine: Makes inferences by deciding which rules are satisfied by facts or objects, prioritizes the satisfied rules, and executes the rules with the highest priority.

Agenda: A prioritized list of rules created by the inference engine, whose patterns are satisfied by facts or objects in working memory.

Knowledge acquisition facility: An automatic way for the user to enter knowledge in the system instead of having the knowledge engineer explicitly codes the knowledge.

3.5. *Rhodobacter sphaeroides*

R. sphaeroides is a facultative photosynthetic member of the α-3 subdivision of *Proteobacteria.* Its genome size is about 4.4 megabases (b) and it consists of two circular chromosomes, chromosome I (CI, ~3.2 b), chromosome II (CII, ~0.9 b), and five other replicons (plasmid DNAs ~450 kb). [38].

Figure **Error! No text of specified style in document.**.5. Rhodobacter sphaeroides.

R. sphaeroides is a purple non-sulfur photosynthetic bacterium [21] [30], with the following characteristics:

Table 1. *Rhodobacter sphaeroides* Characteristic. Source: [29].

Motilty	**Unidirectional flagellum**
Modes of growth	**Chemoorganotroph, photoautotroph, photoorganotroph.**
Division	**Binary fission**
Oxygen relationship	**Facultative anaerobe**
CO_2 assimilation pathway	**Calvin Cycle**

R. sphaeroides is among the most metabolically diverse organism known, being capable of growing in a wide variety of growth conditions [38]. Its metabolic potential includes:

Chemoorganotroph growth for which organic compounds serve as both energy and carbon sources for cell synthesis in presence of O_2.

Photohetrotrophic growth using light as a source of energy and organic materials as carbon source in absence of O_2.

Photoautotroph growth able to use light as its sole source of energy and carbon dioxide as sole carbon source in absence of O_2.

This bacterium is the first free living bacterium known to utilize the regulatory systems and it is the subject of intensive investigations worldwide in structure, function and regulation of its photosynthetic membranes, its mechanism of CO_2 fixation, nitrogen fixation, cytochrome diversity and electron transport systems. In addition, *R. sphaeroides* ha been shown to detoxify a number of metal oxides and oxyanions and is the subject of ongoing studies on bioremediation.

Recent studies reveal that the methods of motility and environmental sensing in relation to bacteria, and movement in *R. sphaeroides*, are unique both genetically and physiologically [38].

The following table2 and table 3 shows some important details on the organism.

Table 2. Organism detail (General). Source: [14].

Taxon Name	**Rhodobacter sphaeroides 2.4.1**
Taxon ID	**272943**
Lineage	**Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Rhodobacter; sphaeroides**
Sequencing Status	**Finished**
Sequencing Center	**JGI**
Funding Agency	**DOE**
Finishing Group	**Collaborator**
IMG Release	**IMG/W 1.01, March 22, 2005**
Comment	**JGI's Genome Portal provides sequence files and annotation.**

Table 3. Organism details (Specific). Source: [14].

	Number	**% of Total**
DNA, total number of bases	4603060	100.00%
DNA coding number of bases	4063238	88.27%
DNA G+C number of bases	3166329	68.79% [1]
Genes total number	4364	100.00%
Protein coding genes	4304	98.63%
RNA genes	60	1.37%
rRNA genes	6	0.14%
tRNA genes	54	1.24%
Genes with function prediction	3038	69.62%
Genes without function prediction	1266	29.01%
Genes w/o function with similarity	1157	26.51%
Pseudo Genes	62	1.42%
Genes assigned to enzymes	603	13.82%
Genes connected to KEGG pathways	523	11.98%
Genes not connected to KEGG pathways	3781	86.64%
Sequencing Status	Finished	

Chapter 4

4. Overview

4.1. System Overview

In this chapter we present a general description about the system and all the parts involved. We consider first the architecture used and then we describe the system developed.

4.2. Architecture

The system includes three parts: inference engine & knowledge base (expert system), the *BioPathDB* relational database, and a flat file.

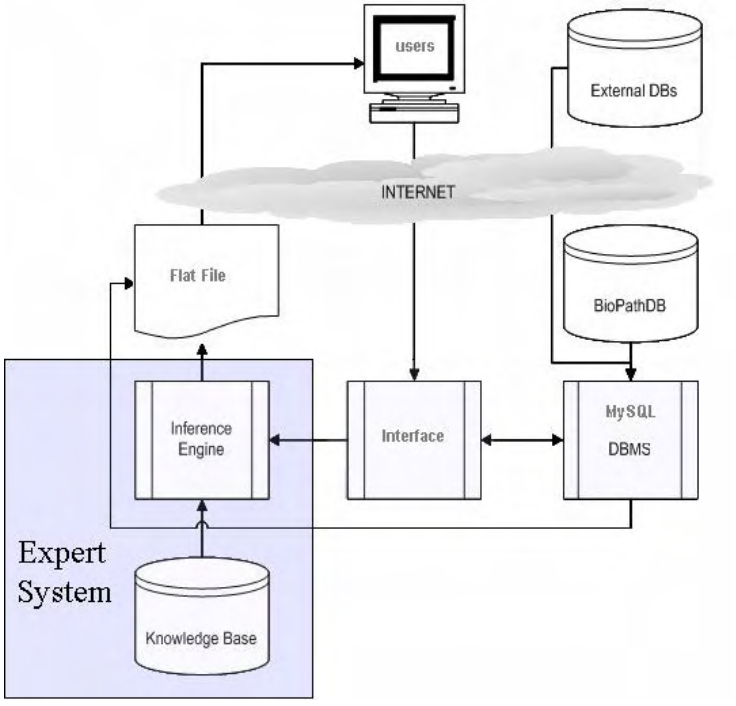

Figure **Error! No text of specified style in document.**.1. System Arquitecture.

For this application we use the client-server paradigm based on internet applications, where, multiples clients can access the same applications at any time. This application on internet was developed using a Java-Servlet on an Apache Tomcat 4.1 container. Where the clients have a user interface that allows inferring reactions between or around specific molecules by connecting individual reactions. In addition the system provides an interface that allows the direct population of the database.

The database is the called *BioPathDB* and it includes information on molecules, xenobiotics, in addition to synonyms, reactions, references, pathways with its nodes and edges. Data in *BioPathDB* are managed by the MySQL Database Management System (DBMS), and it is updated by users through a graphic user interface remotely over the internet and/or locally.

The knowledge base consists of rules stored in a text format file that are used by the inference engine.

The Interface is a bridge between the users and the system. When the user inputs a query, the interface receives it and reads the data from *BioPathDB* database using the MySQL DBMS, who then sorts the data, checks synonym and to returns the results to the interface. If there is not data to support the query then the interface returns an empty flat file to user. The user can add different restrictions as part of the query as number of connections or steps, and depending of option selected find the product in the reactions as individual components or as a set of components. This is then transferred to the Inference Engine, which

produces using individual sets of arguments who through the pattern matching concatenated string. Finally the Inference Engine returns a flat file with all possible pathways.

The Inference Engine used is Jess, a productive development and delivery expert system tool [16]. Jess is written in the Java programming language, and has the advantage that it can be embedded into other applications written in java over different platforms. The inference engine performs forward chaining of the rules and assertions contained in the knowledge base.

4.3. System Detail

4.3.1. Rule-Based Expert System

This rule-based expert system was developed using the Java Expert System Shell (Jess) embodied into Java-Servlet technology, which is included in the following package shown in the Figure 4.2.

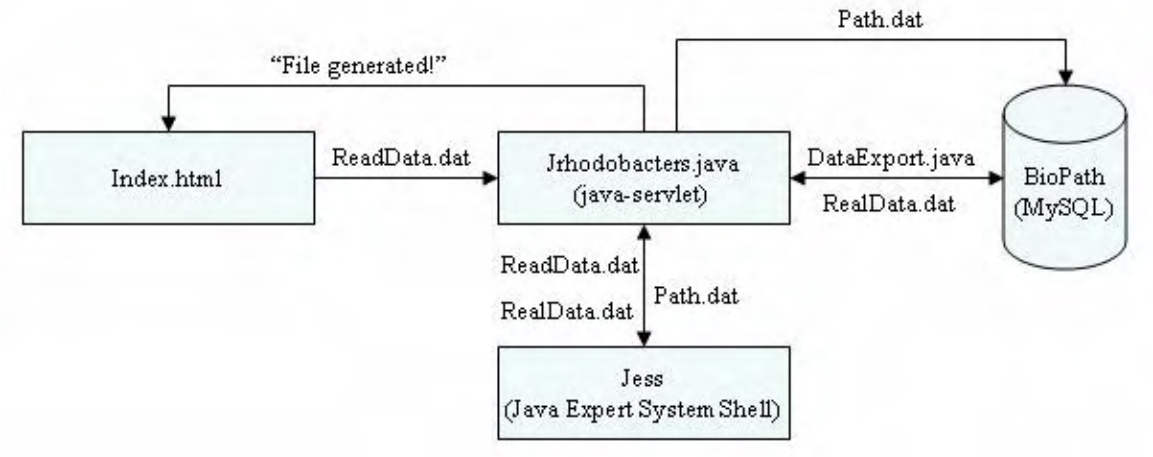

Figure **Error! No text of specified style in document.**.2. Data flow.

- ***Index.html*** is the graphic interface that allows users to input queries. From here we call the java servlet.

- ***JRhodobacters.java*** is a java-servlet, which receives the queries and transfers them as a file to the expert system.

- ***KBRhodobacters.clp,*** is a flat format file. It is the knowledge base where the set of rules are written for finding interconnections between the reactions. Here the assertions are reactions, which are the individual steps in the pathways. In general, these assertions take the form of a reaction or pathway. For example for a standard set of reactions assertions will look like

  Reaction1 (from "A" to "B") (in sequence "A→B")
  Reaction2 (from "B" to "C") (in sequence "B→C")
  Assert Pathway (Sequence "A→B" "B→C") (with "2" connecting steps).

- ***DataRead.dat,*** contains the facts that will be inferred by the knowledge base, these are extracted from *BioPath* by the *DataExport.java*, which is a program write in java.

- ***RealData.Dat,*** contain the petitions of the users, which will be read by expert system.

- ***Path.dat,*** contains the pathway found by expert system in form of nodes pairs. The newly found pathway will be temporally stored in the database, as the reaction that acts among them.

There are two procedures for connecting reactions:

(1) Finding pathways around a target molecule, and

(2) Finding pathways between two target molecules.

An example of assertions and rules for first procedure is as follows:

Assert target component "TC" to find_molecule_from
Assert step equal 0 and maximum number of connections "S"
Assert "TC" to molecule.

The system verifies if the target molecule(s) are in the database, if they are then the rules are executed as follows:

[RULE 1] *Initiate Pathway with From Target Molecule "TC" and a Maximum of "S" connections.*

IF "TC" is in reaction whit substrate then

Extract all reactions "R" and assert in find_reactions

[RULE 2] *Extract all the substrates "SC" of the reactions "R" extracted in Rule 3, to change them with components product "PC" and then Produce Succeeding Pathway.*

[RULE 3] *Stop the inference if the Maximum number of Connections is equal to "S" or there are not more reactions.*

[RULE 4] *Remove disused Assertions.*

An example of assertions and rules for the second procedure is as follows:

Assert initial target component "TC" and end target component EC to find_molecule_from
Assert step equal 0 and maximum number of connections "S"
Assert "TC" to molecule.

The second procedure is similar at first procedure; the difference is that the system verifies if initial target compound and end target compound are in the data and in the rule 3 where:

[RULE 3] *Stop the inference if the substrate component "CS" of the connected reaction is equal to the End Target component "CE" or if the Maximum number of Connections is equal to "S" or there are not more reactions.*

In both cases, the expert system generates a file named *path,txt*, that contains information about the molecules, the father node, which will be stored in our database for future validation.

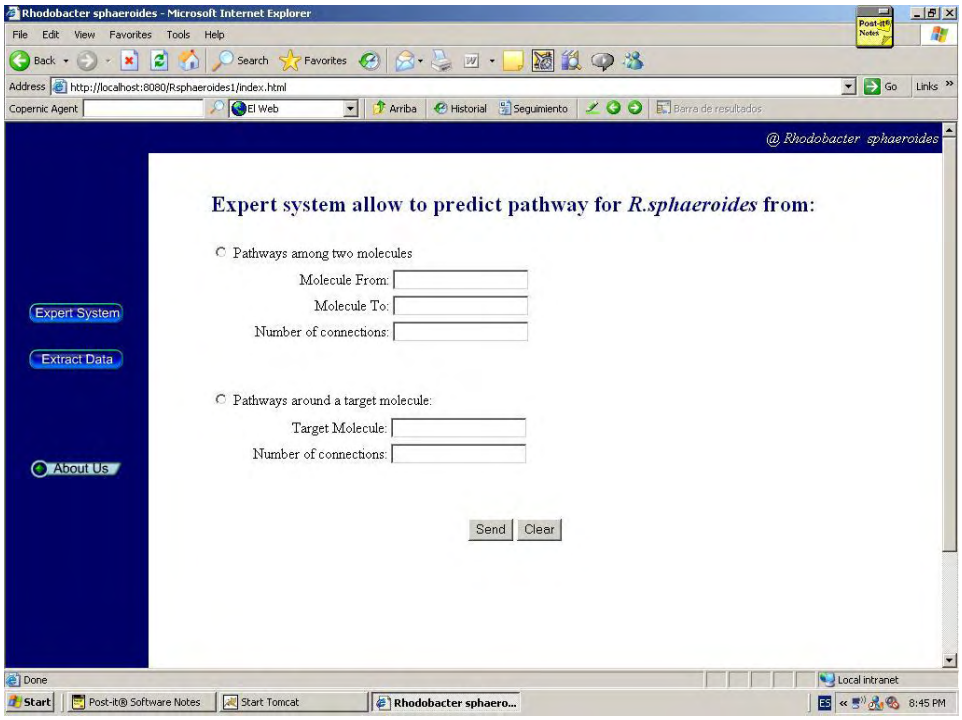

Figure **Error! No text of specified style in document.**.3. Input form: Expert System.

The procedure for using the expert system is as follows:

1. The user must select Pathways around a target molecule or a Pathway between two molecules.

2. The expert system might proceed in the following way:

   2.1. If user selects the first option, it will give the code of target molecule and the maximum number of connections.

   2.2. If user selects the second option, it will give the code of the initial and end molecule, and the maximum number of connections.

In both cases, the objects stored in the *Reaction* entity are retrieved from the database and translated into the form of our assertions, which are then transferred to the Inference Engine joint with the restrictions, and finally the answers are sent to a flat file to the user. If the target molecules there are not in the database then it will return the following message "Molecule does not exist." When a cycle occurs in a pathway, the inference engine stops to prevent an infinite loop from occurring.

4.3.2. BioPathDB Database

As seen above, this system works through the manipulation of one database that contains information on molecules and xenobiotics, biochemical reactions and pathways of *Rhodobacter sphaeroides*, although it contain all the chemical components and reactions knows that are stored in Ligand database of KEGG. The data of molecules that we have

stored correspond to *R. sphaeroides* and was imported from Swiss-Prot database and Integrated Microbial System.

This database is based on a relational technique, which offers highly flexible methods of data definition and modification necessary to handle this diverse and complex biological information. Although for the extraction and insertion of data we used an object-based paradigm. For its implementation we used the Erwin Platinum tool and it is constructed on the MySQL relational database management system (RDBMS) [10]. In addition, their structures consider the format of the external databases used.

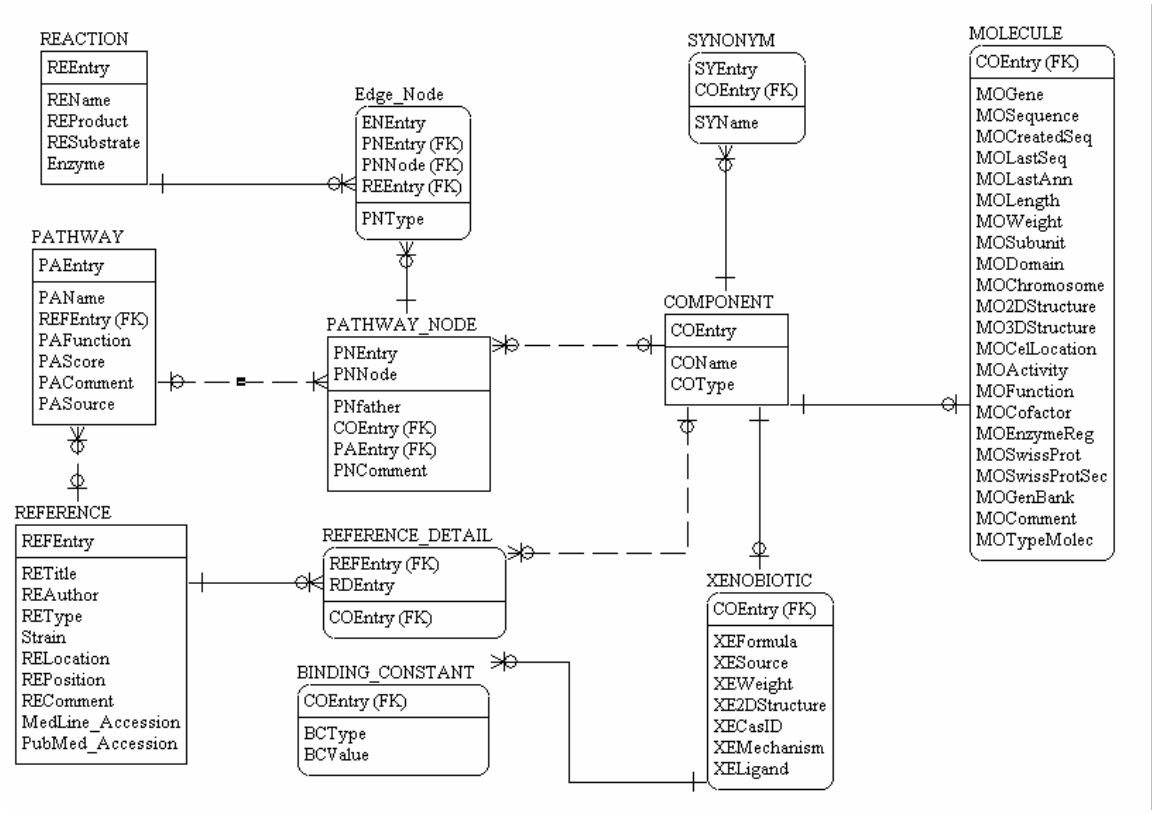

Figure **Error! No text of specified style in document.**.4. Database Model.

The Component entity stores common information of molecules and xenobiotics, which are identify for *COType* attribute. A molecule is really a node into a pathway, nevertheless this can belong to many pathways and vice versa, for this reason the pathway_node entity exists, to show the relationship of many to many, where attribute named *coentry* represent each node in the pathway, whereas *pnfather* indicate the ancestor of node and *pnnode* indicate the sequence of the nodes, the direction is from reactant to product. By every pair of nodes can occur one o more reaction, which we have created the edge_node entity. The above mentioned component can have one or many synonymous names in addition to its principal name, given to the heterogeneous nature of the information. These components are documented in publications and/or databases which references you find in the table reference. This way, it can have one or many references and vice versa, due to this bidirectional relation exists componente_reference table like appears in the figure 4.4

The Molecule entity has information on all molecules involved in biochemical reactions. This information is the gene name, DNA and protein sequence, molecular weight, location in the chromosome, structure, function (including other molecules they interact with), cofactor, accession number at GenBank and Swiss-Prot databases and other information related. This data is imported from Swiss-Prot database and Integrated Microbial Genome System and until now we have stored 4242 proteins and metabolic enzymes encoded by the *R. sphaeroides.*

The *Xenobiotic entity* stores chemicals compounds, which are foreign to the organism. It contains information about its structure, function, chemical formula, as well as their mechanism of activation or inhibition and its binding constant, which can be one o more

values. The data stored correspond to all xenobiotic stored in compound - Ligand database of KEGG, until now have stored 12289 xenobiotic in our database.

The Reaction entity stores information about all known reaction, its name, equation (as reactant and product), which is used by the expert system to infer biochemical reaction. A reaction can be in one o more pairs of nodes into of a pathway. The data stored correspond to 6568 reactions of the reaction – Ligand database of KEGG, which is used to produce assertions for the Inference Engine.

We have considered three types of reactions: standard, polymerization, and metabolic reactions. ***Standard reactions*** include reactions that consist of two elements: reactants and products. The signal transfer direction is explicit in the standard reactions, i.e., from reactants to products, which are stored in the attribute named *REReactant* and *REProduct,* respectively and it can be one or a set of values. ***Polymerization reactions*** are reactions with no directionality consisting only of pairing between two or more molecules [36]. These molecules can either be chemically the same or different. ***Metabolic reactions are reactions*** consisting of three elements, i.e., reactants, products and enzymes. This reaction produces intermediary metabolites and lead to a final product [37]. The object name is designed in the style “enzyme + reactant -> product”. In this case, the enzyme is stored in a separate attribute. In order to convert metabolic reactions into the same type of assertion as used with standard reactions, three elements need to be reduced two elements. The three elements are divided into two pairs, “reactant and products” and “enzyme and products”. This division is based on the consideration that adjacent predecessor reactions can point to either the reactant or the enzyme separately [36].

The *Pathway entity* stores information of the pathways in terms of name, function, source, and reference. The *pasource* indicates whether the pathway has been validated (i.e., experimentally proven). If *pascore* is equal to 0, then it has not been validated, *pascore* equal to 1 has some information and *pascore* equal to 2 has been completely validated by experimental data.

4.3.3. Data Source

To populate the database there are two alternatives; from data published by researchers within and without the UPR and/or from external biological databases. The first alternative is done through phpMyAdmin, which is a tool written in PHP intended to handle the administration of MySQL over the web. With this tool we can update our database in a graphical format. The second alternative is through the graphical user interface, but is necessary that before importing the data downloading the flat file from the external database using the following form:

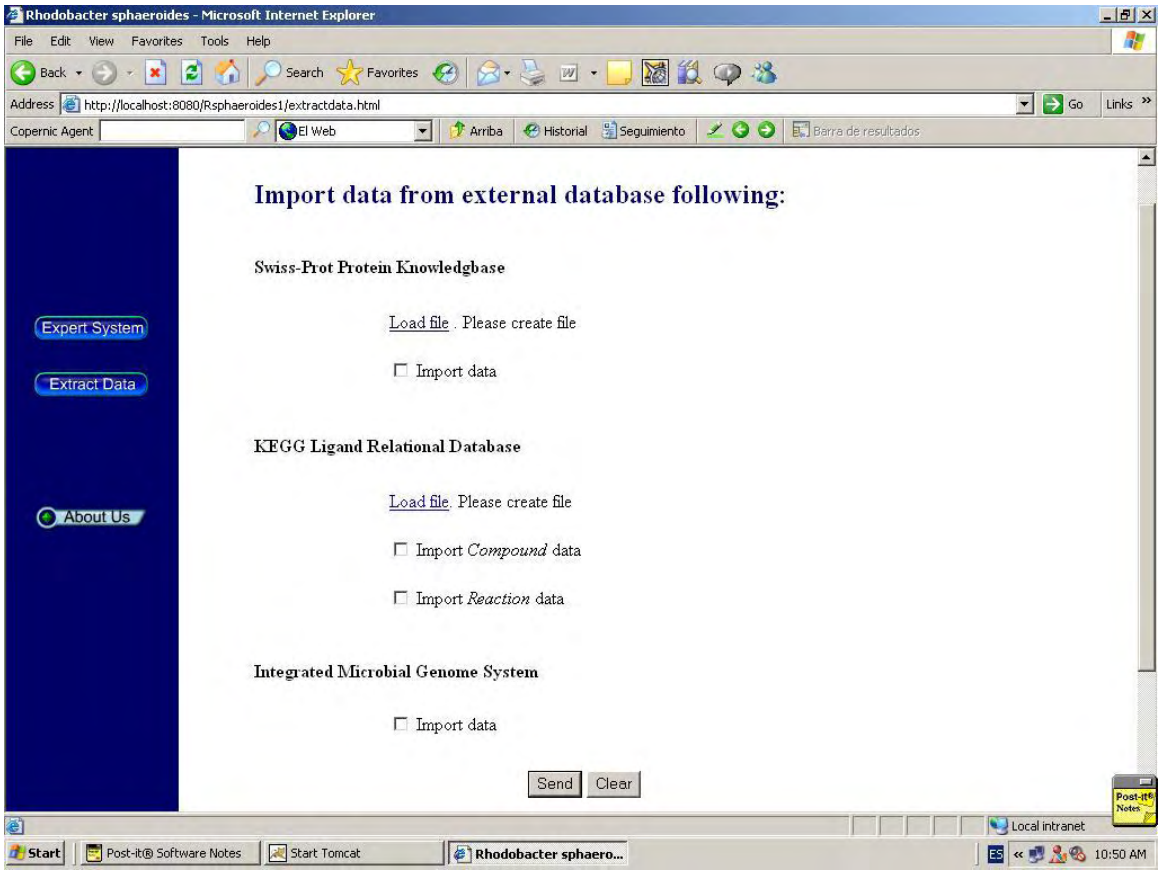

Figure **Error! No text of specified style in document.**.5. Input form: To import data from external database.

Using this interface, we will import data directly from external database in flat or fasta format files to our database; among them are Swiss-Prot knowledgebase, Ligand (compounds and reactions) of KEGG, UMBB Database and IMG system. But the data of the last two sources cannot be access directly from the web. The UMBB data was inputted manually and the other was obtained through of the author. Is very important that first we load file and then import the data to our database.

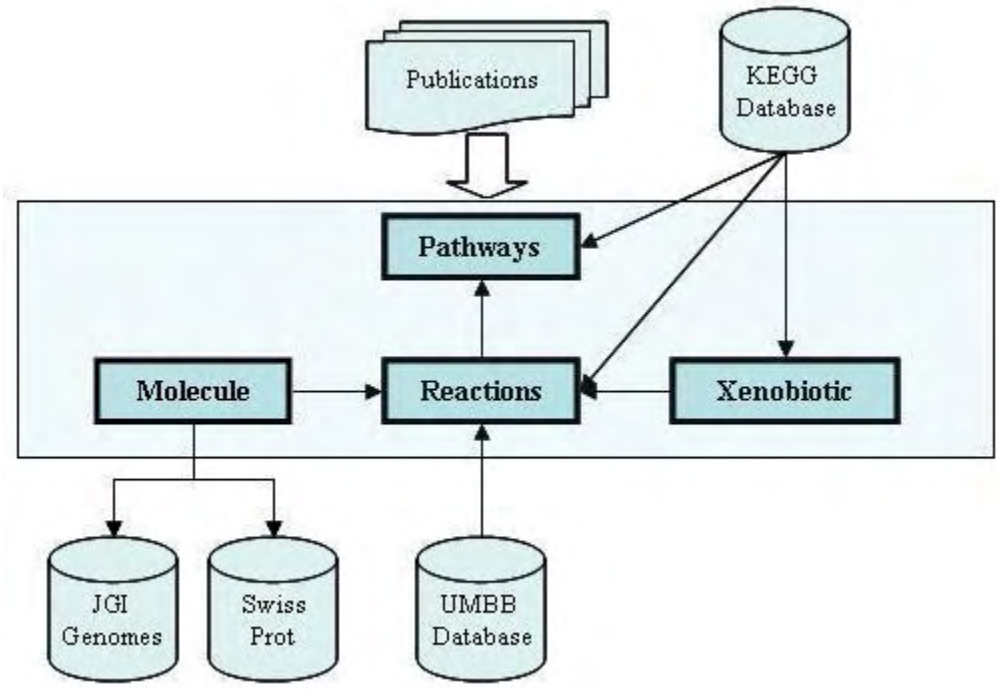

Figure **Error! No text of specified style in document.**.6. Data Source.

1. From the Swiss-Prot Protein knowledgebase, we import *R.spaheroides* data only related to proteins, and which is in the table 4.

Table 4. SwissProt Knowledgebase Data: Proteins.

ID 14KD_RHOSH STANDARD; PRT; 124 AA.
AC P16536;
DT 01-AUG-1990 (Rel. 15, Created)
DT 01-APR-1993 (Rel. 25, Last sequence update)
DT 28-FEB-2003 (Rel. 41, Last annotation update)
DE 14 kDa peptide of ubiquinol-cytochrome C2 oxidoreductase complex.
OS Rhodobacter sphaeroides (Rhodopseudomonas sphaeroides).
OC Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales;
OC Rhodobacteraceae; Rhodobacter.
OX NCBI_TaxID=1063;
RN [1]
RP SEQUENCE FROM N.A.
RX MEDLINE=91340695; PubMed=1651916;
RA Usui S., Yu L.;
RT "Subunit IV (Mr = 14,384) of the cytochrome b-c1 complex from
RT Rhodobacter sphaeroides. Cloning, DNA sequencing, and ubiquinone
RT binding domain.";
RL J. Biol. Chem. 266:15644-15649(1991).

RN [2]
RP SEQUENCE OF 61-108.
RX MEDLINE=90110107; PubMed=2153104;
RA Purvis D.J., Theiler R., Niederman R.A.;
RT "Chromatographic and protein chemical analysis of the ubiquinol-
RT cytochrome c2 oxidoreductase isolated from Rhodobacter sphaeroides.";
RL J. Biol. Chem. 265:1208-1215(1990).
CC -!- FUNCTION: Component of the ubiquinol-cytochrome c reductase
CC complex (complex III or cytochrome b-c1 complex), which is a
CC respiratory chain that generates an electrochemical potential
CC coupled to ATP synthesis.
CC ---
CC This SWISS-PROT entry is copyright. It is produced through a collaboration
CC between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC the European Bioinformatics Institute. There are no restrictions on its
CC use by non-profit institutions as long as its content is in no way
CC modified and this statement is not removed. Usage by and for commercial
CC entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC or send an email to license@isb-sib.ch).
CC ---
DR EMBL; M68939; AAA26107.1; -.
DR PIR; A40794; A40794.
KW Oxidoreductase; Electron transport; Inner membrane; Transmembrane;
KW Respiratory chain.
FT TRANSMEM 85 102 POTENTIAL.
SQ SEQUENCE 124 AA; 14393 MW; E390C856C1D752F3 CRC64;
MFSFIDDIPS FEQIKARVRD DLRKHGWEKR WNDSRLVQKS RELLNDEELK IDPATWIWKR

In according to manual of Swiss-Prot database, each sequence entry is composed of lines, where each line begins with a two-characters line code, which indicate the type of data contained in the line. Every entry being with an identification line (ID) and end with a terminator line (//)

The ID line indicate entry name, molecule type and sequence length; the AC line list the accession numbers associated with an entry, the DT line show the date of creation and last modification of the database entry; the DE line contain general description about the sequence which is generally sufficient to identify the protein; the GN line contain the name(s) of the gene(s) that code for the protein sequence; the OS line specifies the organism, the OG line indicate if the gene coding for a protein originated from the moitochondria, the

chloroplast, the cyanelle, the nucleomorph or a plasmid; the OC line contain the taxonomic classification of the source organism, the reference is identified in RN, RP, RC, RX, RA, RT and RL and finally the CC line are free text comments on the entry and are used to convey any useful information, since here we have retrieve information about function, catalytic activity, cofactor, domain, pathway, subcellular location, subunit and similarity; but additionally there is information about alternatives products, biotechnology, caution, etc.

2. From the Ligand database of the Kyoto Encyclopedia of Genes and Genomes [20] we import compounds and reactions data involved in all biochemical pathways known for organisms considered here in table 5.

Table 5. KEGG Knowledgebase Data: Compound.

```
ENTRY       C00116              Compound
NAME        Glycerol
            Glycerin
            1,2,3-Trihydroxypropane
            1,2,3-Propanetriol
FORMULA     C29H36O10
REACTION    R00841    R00847    R00850    R01034    R01036
            R01044    R01945
PATHWAY     PATH: MAP00904      Galactose metabolism
            PATH: MAP00561      Glycerolipid metabolism
ENZYME      1.1.1.1   1.1.1.2   1.1.1.6   1.1.1.21
            1.1.1.72  1.1.1.156
DBLINKS     CAS: 56-81-5
ATOM        3
            1  C1y C   6.2360  -5.9843 #R
            2  C1z C   6.8967  -6.3586 #R
            3  C1y C   5.7657  -6.5000 #R
BOND        3
            1   1  2 1
            2   1  3 1
            3   1  4 1
///
```

Where, every attribute is identified by the first word which appears on column 1 to12, the columns followings describe the item data. An entry being with the *Entry* item and

finishes with the end of entry (///) and, both and the name are obligatory. We have stored in compound the entry, name, synonyms names, formula, structure and CAS. The enzymes, reactions and pathways are being stored in the other tables.

Entry item contain the access number to database LIGAND, Name item contain the name and if exist a name alternative it is stored as synonym name. Formula item store the component chemical formula. The reaction, pathway and enzyme item contain access to the section correspondent. DBLINKS contain information to different databases, including the CAS (Chemical Abstract Service) record number.

Table 6. KEGG Knowledgebase Data: Reaction.

ENTRY	**R00005**
NAME	**Urea-1-carboxylate amidohydrolase**
DEFINITION	**Urea-1-carboxylate + H2O <=> 2 CO2 + 2 NH3**
EQUATION	**C01010 + C00001 <=> 2 C00011 + 2 C00014**
PATHWAY	**PATH: RN00220 Urea cycle and metabolism of amino groups**
	PATH: RN00910 Nitrogen metabolism
	PATH: RN00791 Atrazine degradation
ENZYME	**3.5.1.54**
COMMENT	**The yeast enzyme (but not that from green algae) also catalyses the reaction of EC 6.3.4.6 urea carboxylase, thus bringing about the hydrolysis of urea to CO2 and NH3 in the presence of ATP and bicarbonate. R00774 (6.3.4.6)**

As in the components, in the table 6 corresponding to Reaction table, the data item appear on columns 1 to 12 and every entry begin with *Entry* item and finishes with end of entry (///). The Entry, Definition, Equation and end of entry are obligatory for every entry. We store the Entry, Name, Definition, Equation and Enzyme. The equation is stored as reactant and product, where the reactant is left part and the product is right part.

The data about of the pathway will be input manually from KEGG, due that only is available in graphical format. The only *R. sphaeroides* pathway that we have inputted is from the Biocatalysis\Biodegradation database, which also is in graphic format.

Chapter V

5. Result Analysis

5.1. Overview

In this chapter, we analyze the results of our expert system through of three scenarios. In the first scenario a pathways were searched for around a target molecule. In the second scenario pathways among two molecules were searched for. Lastly in the third scenario a cyclic pathway was inferred to show the behavior of the expert system.

5.2. Method

The method that we have designed to corroborate the capacity of the expert system in inferring pathways in *Rhodobacter sphaeroides* is simple but effective; consist in crossing information with the individual components and reactions versus the information stored in the existing databases related to such molecules, reactions and pathways.

The results from the expert system were stored in our database with a *pascore* equal to 0 (not validate) and *psource* equal to S (source: Expert System). This was followed by a comparison with the stored pathway data (~20 pathways). Due to the fact that the data is currently incomplete we have crossed information between the results and data from the Ligand database in KEGG, which includes molecule, reaction and enzyme. This shows if every molecule and reaction is connected to a pathway stored in the Pathway database. To

simplify the work a molecule was chosen that had not been connected but was involved in one or more reactions, catalyzed by an enzyme which had been identified in *R. sphaeroides*.

5.3. Scenarios:

5.3.1. Scenario 1: Finding pathways around a target molecule

For this scenario we have entered the target molecule ATP:Adenosine 5'-triphosphate (C00002) with the maximum number of connections set to 10. The result is shown in the Figure 5.1. One of the reactions which is involved this molecule is catalyzed by the enzyme Nitrogenase (E.C. 1.18.6.1) identified in *R. sphaeroides*.

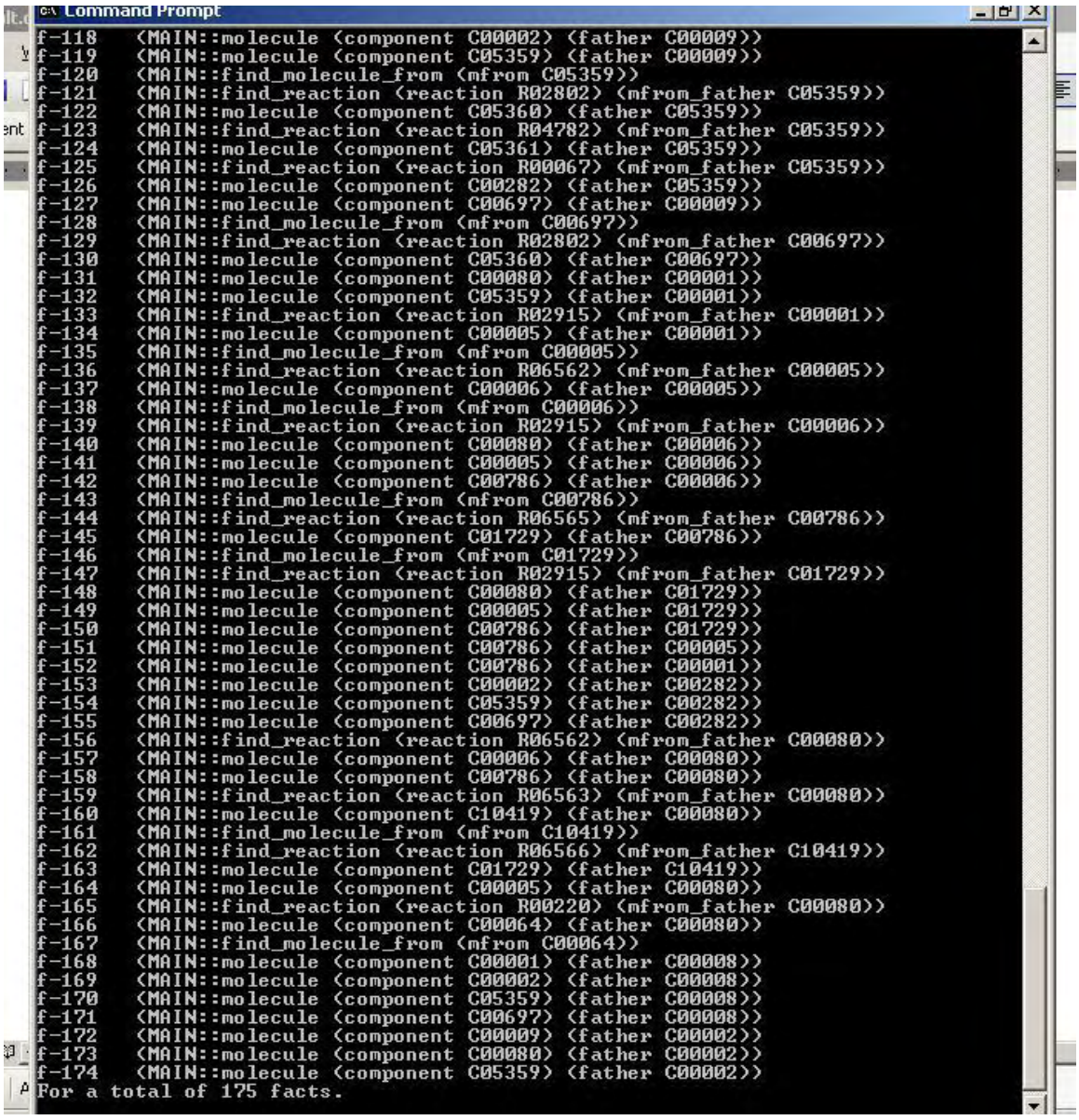

Figure **Error! No text of specified style in document.**.1. Result of expert system. Scenario 1.

This result (figure 5-1) is then sending to text file and the component pairs are stored in our database for validation by human experts.

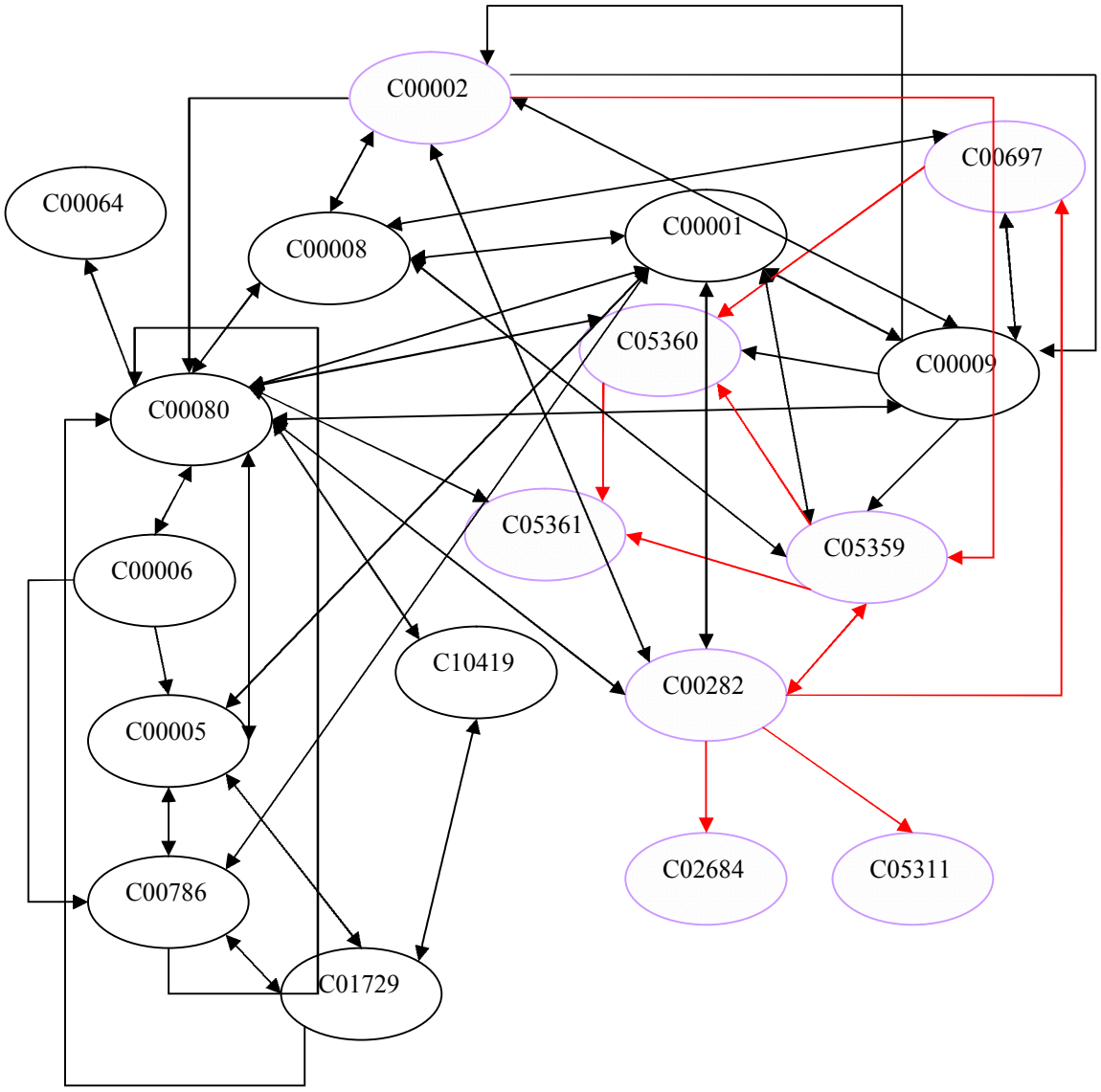

Figure **Error! No text of specified style in document.**.2. Possible pathway found by the expert system.

The Figure 5.2, shows the interconnections among pairs of the resulting molecules found by the expert system around the target molecule ***C00002*** (i.e., ATP or Adenosine 5'-triphosphate). This molecule is part of many known metabolic reactions and connects to the following pathways: Oxidative phosphorylation, Photosynthesis, Purine metabolism,

Puromycin biosynthesis, Cholera – Infection found in the Pathway database of KEGG. In this figure, the colored arrows are the new pathways, which will be described in Figure5.3.

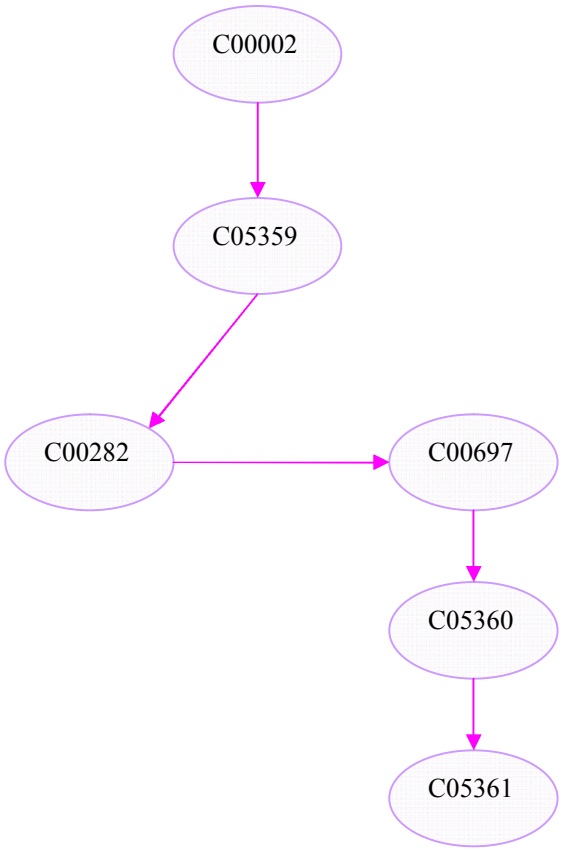

Figure **Error! No text of specified style in document.**.3. Pathway found by the expert system.

According the figure, there is one pathway, which is identified by the colored arrows and which occurs from Adenosine 5'-triphosphate (C00002) molecule to Hydrazine (C05361) molecule.

ATP => e- or electron
C00002 => C05359

e- or electron => Hydrogen
C05359 => C00282

Hydrogen => Nitrogen
C00282 => C00697

Nitrogen => Diimine
C00697 => C05360

Diimine => Hydrazine
C05360 => C05361

In the first step in the pathway, the reactant is ATP (C00002) and e- (C05359) is the product. It is a reaction known named Reduced ferredoxin:dinitrogen oxidoreductase (ATP-hydrolysing) and catalyzed by Nitrogenase (E.C. 1.18.6.1) enzyme, which has been identified in *R.sphaeroides* [14]. This reaction has not been involved in any pathway.

The second step In addition, the known reaction Reduced ferredoxin:dinitrogen oxidoreductase (ATP-hydrolysing) (R00067) have the e- or electron (C05359) as reactant and Hydrogen (C00282) as product. This reaction has been connected to the glyoxylate and dicarboxylate metabolism and methane metabolism pathways.

In the third step the reactant is the Hydrogen (C00282) and the product is the Nitrogen (C00697). It also is connect to the nitrogen metabolism pathway and the phosphotransferase system (PTS).

Other relation is the Nitrogen (C00697) as reactant and the Diimine (C05360) as prduct, which are involved in the known reaction Reduced ferredoxin:dinitrogen

oxidoreductase (ATP-hydrolysing) (R02802), but it are not involved in any pathway. For last is the Diimine (C05360) as reactant and Hydrazine (C05361) as product.

Here, the first and the two last reaction steps have not been connected to the pathway in KEGG as well as the molecules Diimine (C05360) and Hydrazine (C05361).

Another possible pathway generated in the same result is the following:

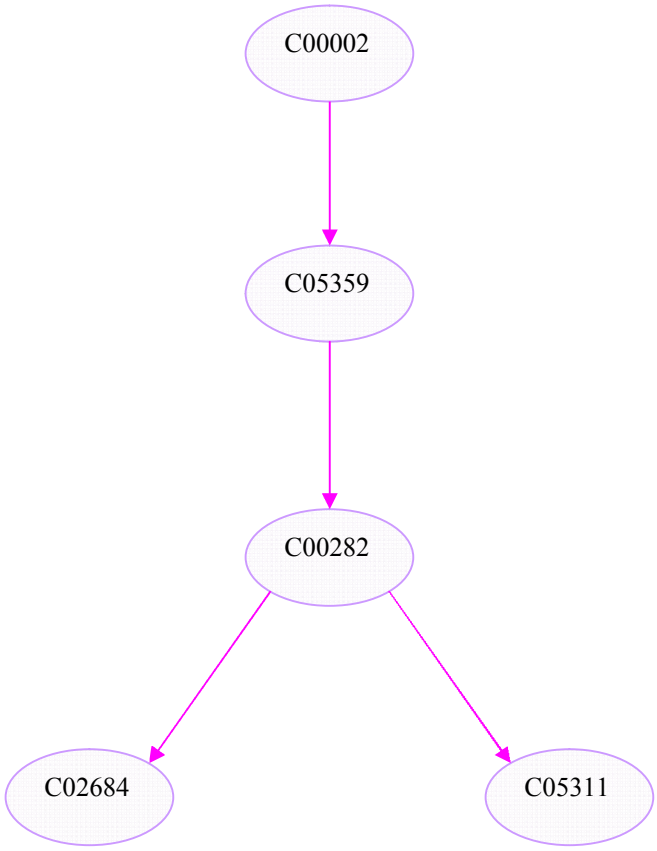

Figure **Error! No text of specified style in document.**.4. Pathway found by the expert system.

According the figure 5.4, this pathway occurs from ATP or Adenosine 5'-triphosphate (C00002) to Ferrocytochrome c3 (C02684) and Reduced Menaquinone (C05311).

ATP => e- or electron
C00002 => C05359

e- or electron => Hydrogen
C05359 => C00282

Hydrogen => Ferrocytochrome c3
C00282 => C02684

Ferrocytochrome c3=> Reduced Menaquinone
C00282 => C05311

The two first steps are equal to two first in the previous case. The difference is that the Hydrogen (C00282) act as reactant and Ferrocytochrome c3 (C02684) as product in the reaction R04015 named Hydrogen:ferricytochrome-c3 oxidoreductase, catalyzed by the enzyme 1.12.2.1. Likewise, occur with the component Hydrogen (C00282) as reactant and Reduced Menaquinone (C05311) in the reaction Hydrogen:quinone oxidoreductase (R02965), which yet have not been connecting to pathway of KEGG.

5.3.2. Scenario 2: Finding of pathways among two molecules

The expert system stops when it finds the target molecule or if it reaches the maximum number of connections allowed. In this case, it stopped when 4-Chlorobenzaldehyde (C06648) was found as the as product as is shown in the Figure 5.5 and Figure 5.7. For this case we was chose the 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) pathway.

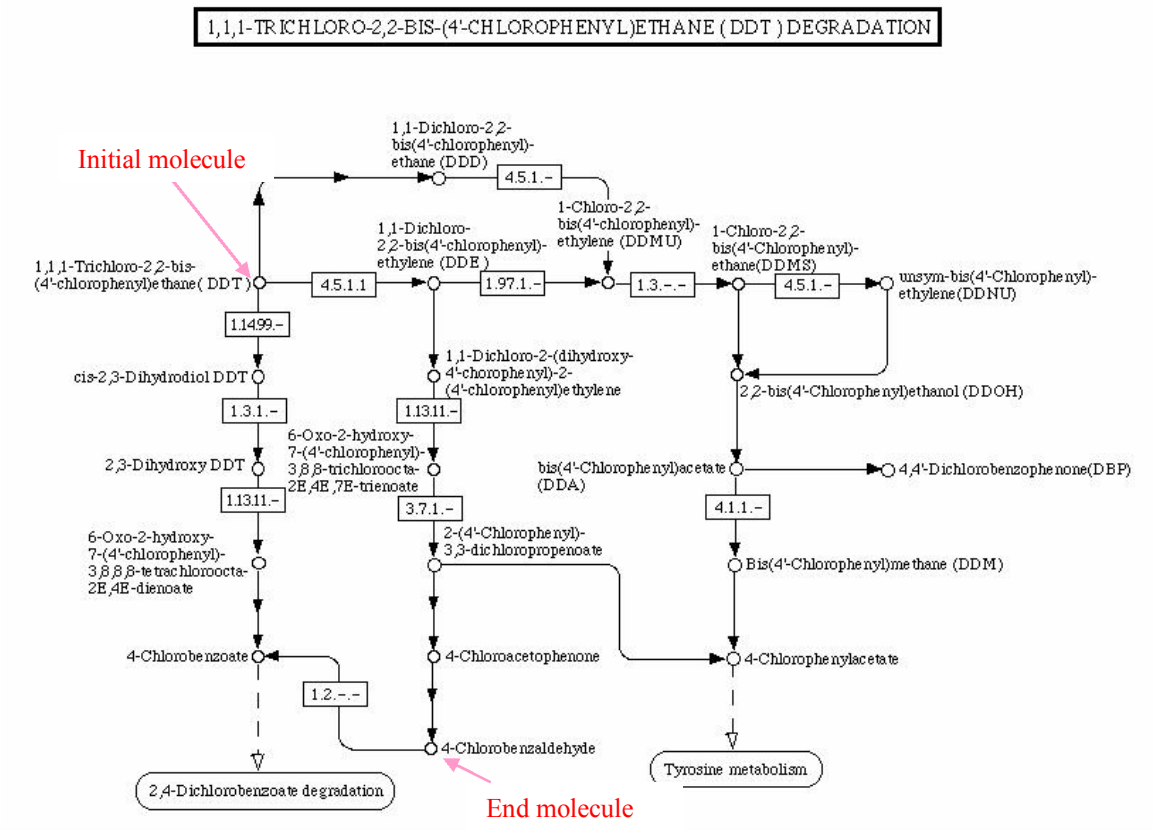

Figure **Error! No text of specified style in document.**.5. Reference to Pathway. Scenario 2.

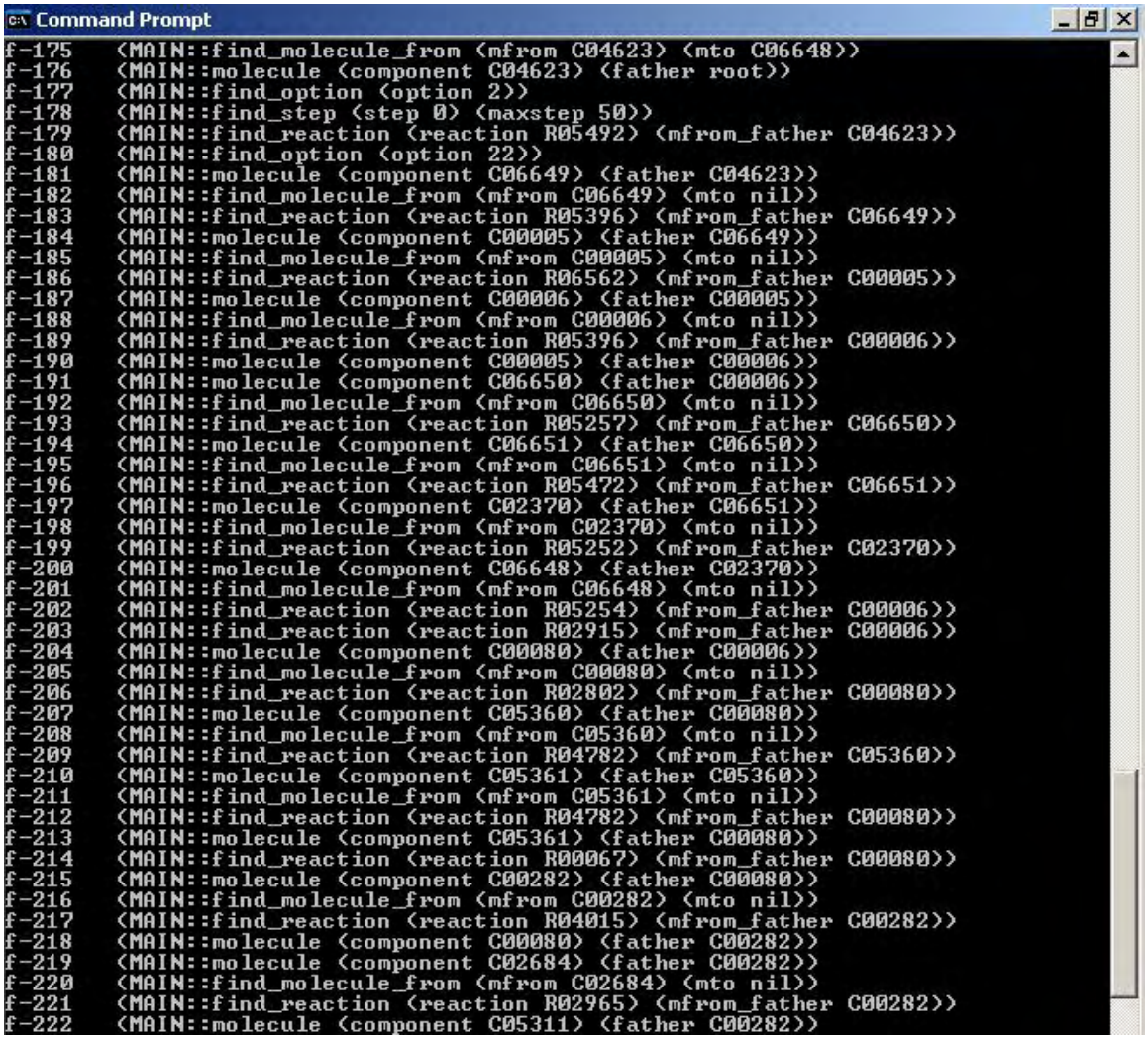

Figure **Error! No text of specified style in document.**.6. Result of expert system. Scenario 2.

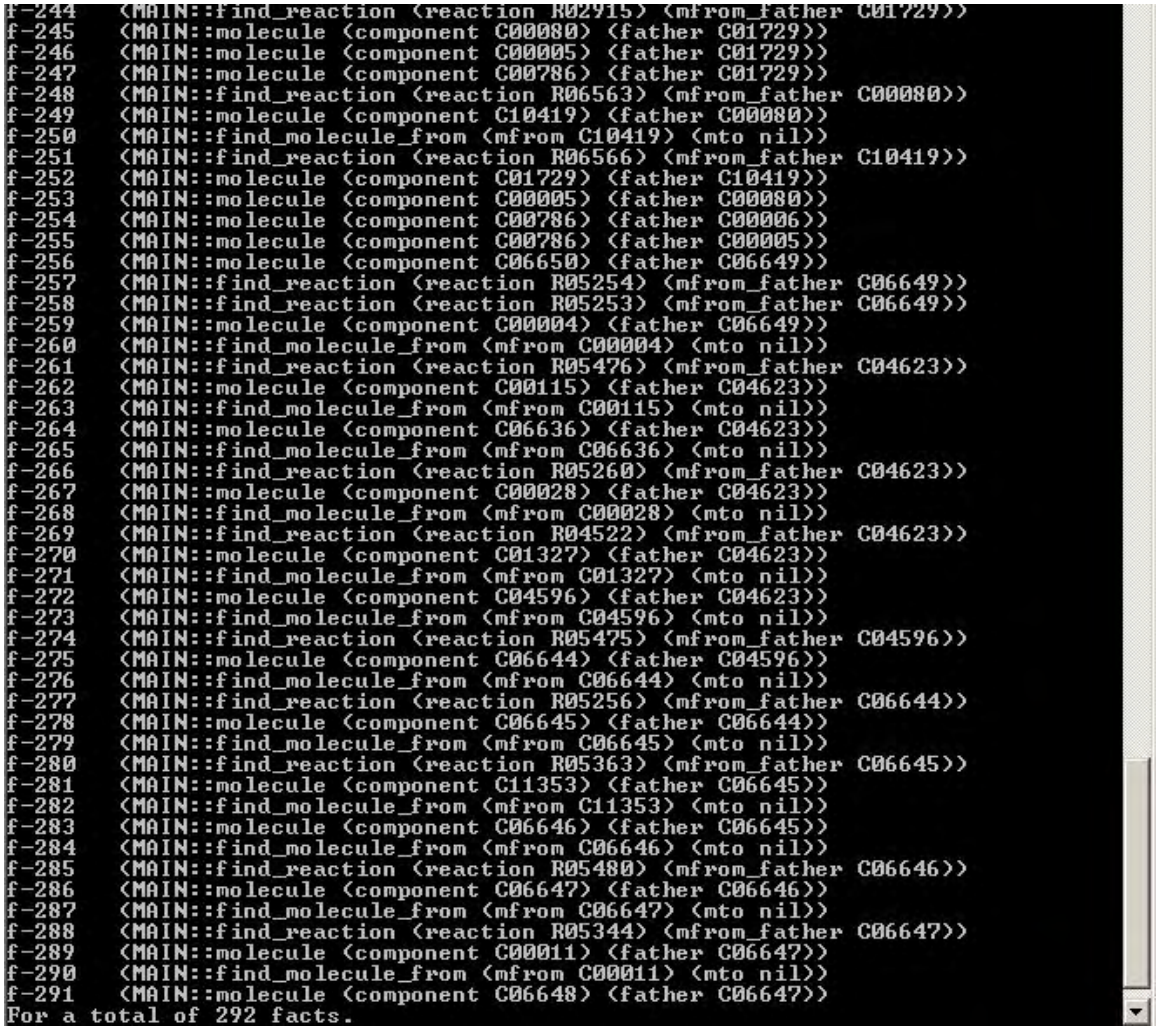

Figure **Error! No text of specified style in document.**.7. Result of expert system. Scenario 2 (cont.).

In Figure 5.7 we showed graphically the pathways found by the expert system. In this pathway the initial molecule is 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane (C04623).

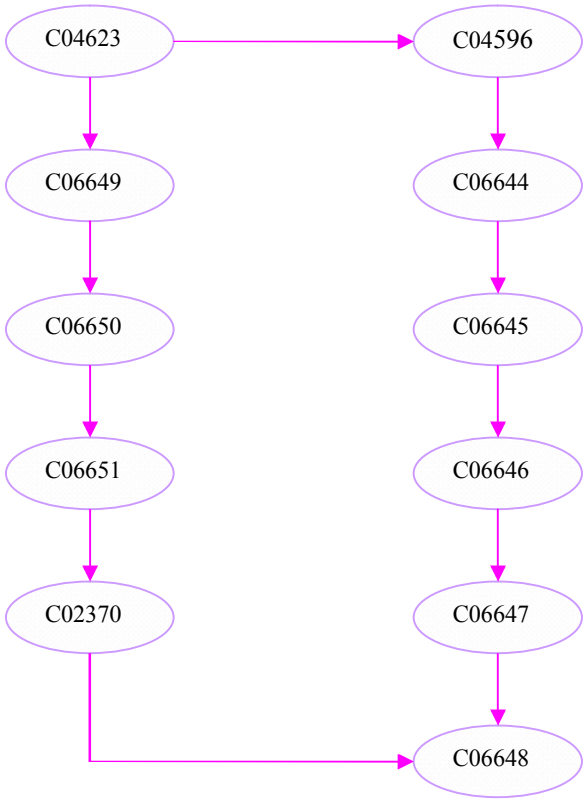

Figure **Error! No text of specified style in document.**.8. Pathways found by the expert system. Scenario 2.

As we can observe the system retrieved the branches formed by 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane (C04623) and 4-Chlorobenzaldehyde (C06648).

5.3.3. Scenario 3: Cyclic Pathway

In this scenario, we show that the expert system finishes when it comes to the number of connections or there are no any more interactions, but it always evade a loop.

The Dimethyl Sulfoxide & Organosulfide pathway is a cyclic pathway that is produced in nature in large amounts by biological and chemical oxidation of dimethyl sulfide. The first step in the pathway can also occur in many anaerobic and facultative anaerobic bacteria among them *R. sphaeroides*, which uses DMSO as an electron acceptor for anaerobic respiration, releasing dimethyl sulfide [39].

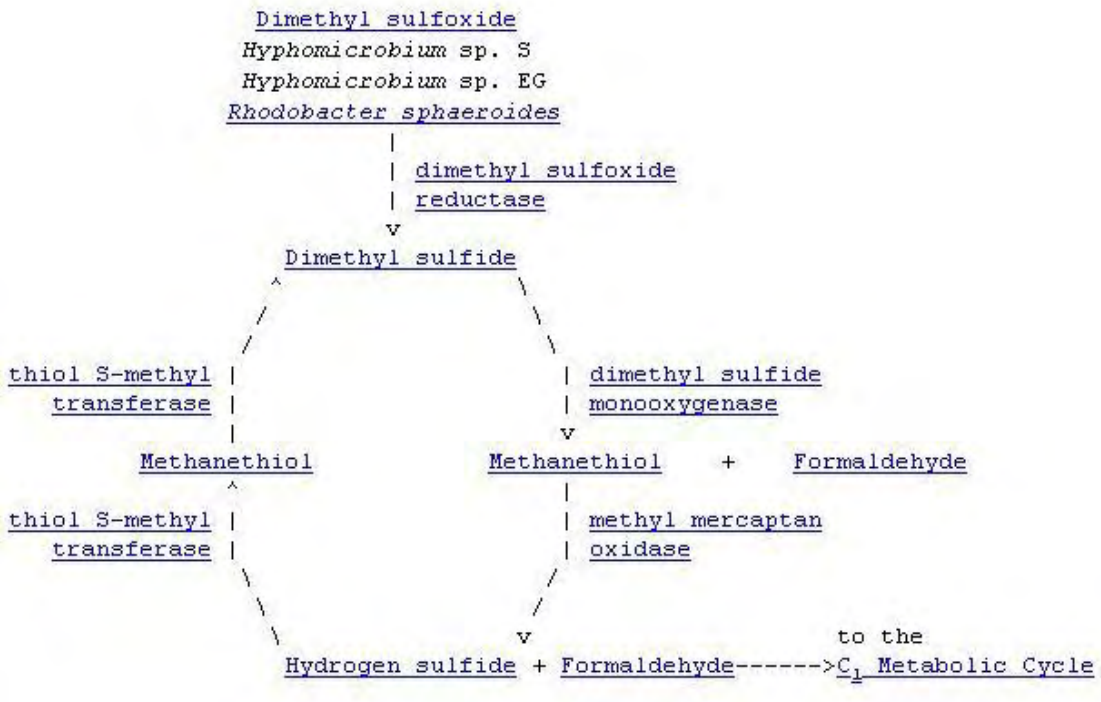

Figure **Error! No text of specified style in document.**.9. Dimethyl sulfide and Organosulfide cycle pathway [39].

```
f-92    (MAIN::find_molecule_from (mfrom C00580))
f-93    (MAIN::molecule (component C00580) (father root))
f-94    (MAIN::find_option (option 2))
f-95    (MAIN::find_step (step 0) (maxstep 7))
f-96    (MAIN::find_reaction (reaction R06570) (mfrom_father C00580))
f-97    (MAIN::find_option (option 22))
f-98    (MAIN::molecule (component C00067) (father C00580))
f-99    (MAIN::find_molecule_from (mfrom C00067))
f-100    (MAIN::find_reaction (reaction R06571) (mfrom_father C00067))
f-101    (MAIN::molecule (component C00067) (father C00067))
f-102    (MAIN::molecule (component C00283) (father C00067))
f-103    (MAIN::find_molecule_from (mfrom C00283))
f-104    (MAIN::find_reaction (reaction R06572) (mfrom_father C00283))
f-105    (MAIN::molecule (component C00409) (father C00283))
f-106    (MAIN::find_molecule_from (mfrom C00409))
f-107    (MAIN::find_reaction (reaction R06573) (mfrom_father C00409))
f-108    (MAIN::molecule (component C00580) (father C00409))
f-109    (MAIN::find_reaction (reaction R06571) (mfrom_father C00409))
f-110    (MAIN::molecule (component C00067) (father C00409))
f-111    (MAIN::molecule (component C00283) (father C00409))
f-112    (MAIN::molecule (component C00409) (father C00580))
For a total of 113 facts.
```

Figure **Error! No text of specified style in document.**.10. Result of the expert system. Scenario 3. Case 1.

In accordance with the Figure 5.8, the expert system ended when it found the initial molecule of the cycle since it could not find any more interactions, although it explored others branches as shown in the Figure 5.9.

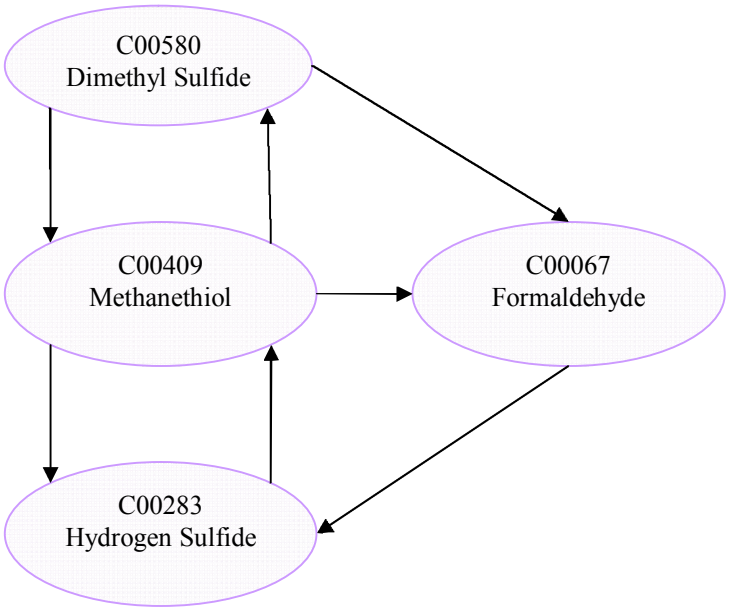

Figure **Error! No text of specified style in document.**.11. Pathway found by the expert system.

Where, the dimethyl sulfide (C00580) molecule acts as reactant and methanethiol (C00409) with formaldehyde (C00067) act as products in the dimethyl sulfide monooxigenase reaction (dimethyl sulfide => methanethiol + formaldehyde). Methanethiol (C00409) and formaldehyde (C00067) molecules as reactant and hydrogen sulfide (C00283) and formaldehyde (C00067) as product in the methyl mercaptan oxidase reaction catalyzed by the enzyme 1.8.3.4 (methanethiol + formaldehyde => Hydrogen sulfide + formaldehyde). Hydrogen sulfide (C00283) and formaldehyde (C00067) are reactants and methanethiol (C00409) is the product in the thiol S-methyl transferase reaction catalyzed by 2.1.1.9 (Hydrogen sulfide + formaldehyde => methanethiol). Notice that the reaction also occurs when methanethiol reacts to form dimethyl sulfide.

The dimethyl sulfide (C00580) and Methanethiol (C00409) molecules form part of the known reactions but are not part of the known biochemical pathways found on KEGG.

Another example is the Reductive carboxylate cycle (CO_2 fixation) pathway,in which we modified some reactions that do not form a cyclic pathway as it appears in Figure 5.10 as (S)-Malate:NADP+ oxidoreductase reaction (R00342) whose definition is L-Malate + NAD+ <=> Oxaloacetate + NADH + H+ (C00149 + C00003 <=> C00036 + C00004 + C00080) where L-Malate is the substrate and Oxalocetate is the product, different from the figure where the Oxalocetate is the substrate and L-Malate is the product. Other reactions are Succinate:(acceptor) oxidoreductase whose definition is Succinate + Acceptor <=> Fumarate + Reduced acceptor (C00042 + C00028 <=> C00122 + C00030), Isocitrate:NADP+ oxidoreductase (decarboxylating) reaction whose definition is Isocitrate + NADP+ <=> 2-

Oxoglutarate + CO_2 + NADPH + H+ (C00311 + C00006 <=> C00026 + C00011 + C00005 + C00080), Citrate hydro-lyase reaction whose definition is Citrate <=> cis-Aconitate + H_2O (C00158 <=> C00417 + C00001), Orthophosphate:oxaloacetate carboxyl-lyase (phosphorylating) reaction whose Orthophosphate + Oxaloacetate <=> H_2O + Phosphoenolpyruvate + CO_2 (C00009 + C00036 <=> C00001 + C00074 + C00011) and L-Alanine:NAD+ oxidoreductase (deaminating) reaction whose definition is L-Alanine + NAD+ + H_2O <=> Pyruvate + NH_3 + NADH + H+ (C00041 + C00003 + C00001 <=> C00022 + C00014 + C00004 + C00080).

The modifications have been to swap substrates by products and vice versa, but only to corroborate that the expert system evades the loops.

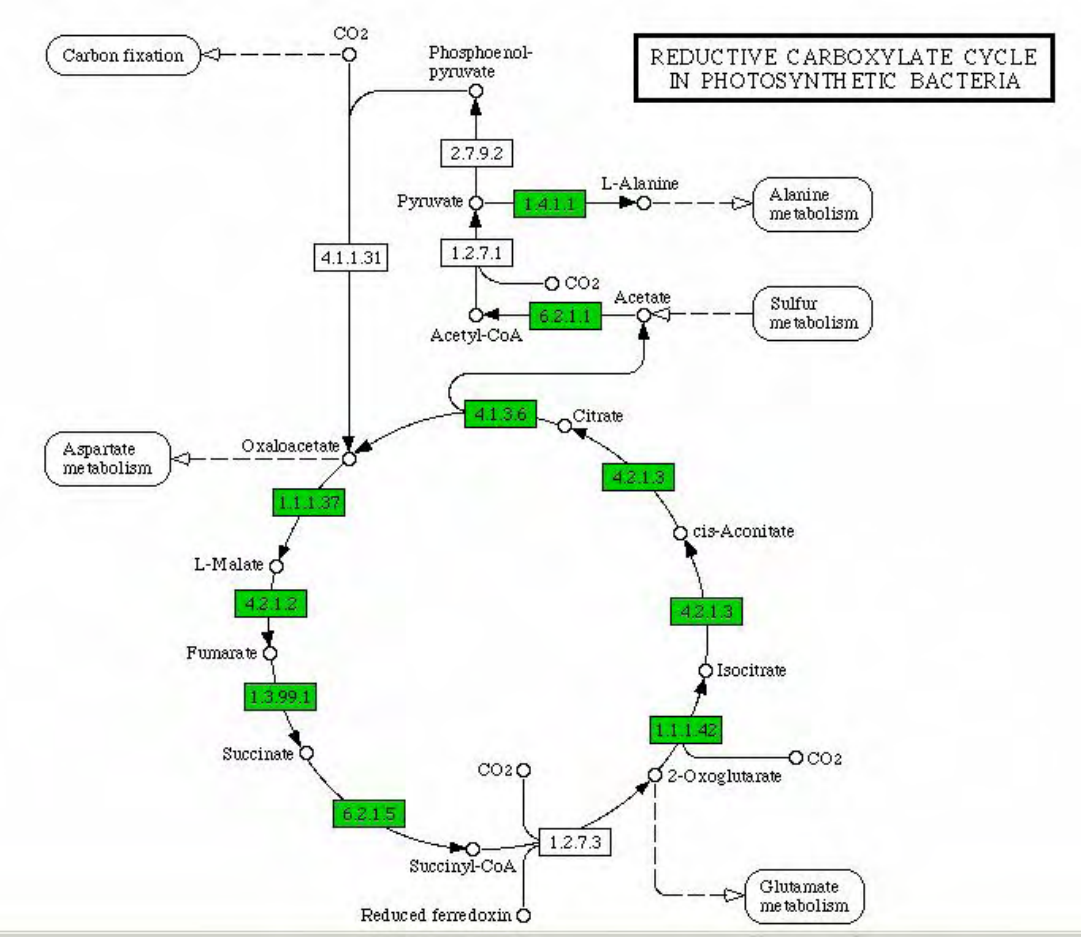

Figure **Error! No text of specified style in document.**.12. Reductive carboxylate cycle (CO2 fixation) pathway.

For this case, the target molecule was Oxaloacetate (C00036). In the first step it is a substrate. The system stops when it did not find any more interactions.

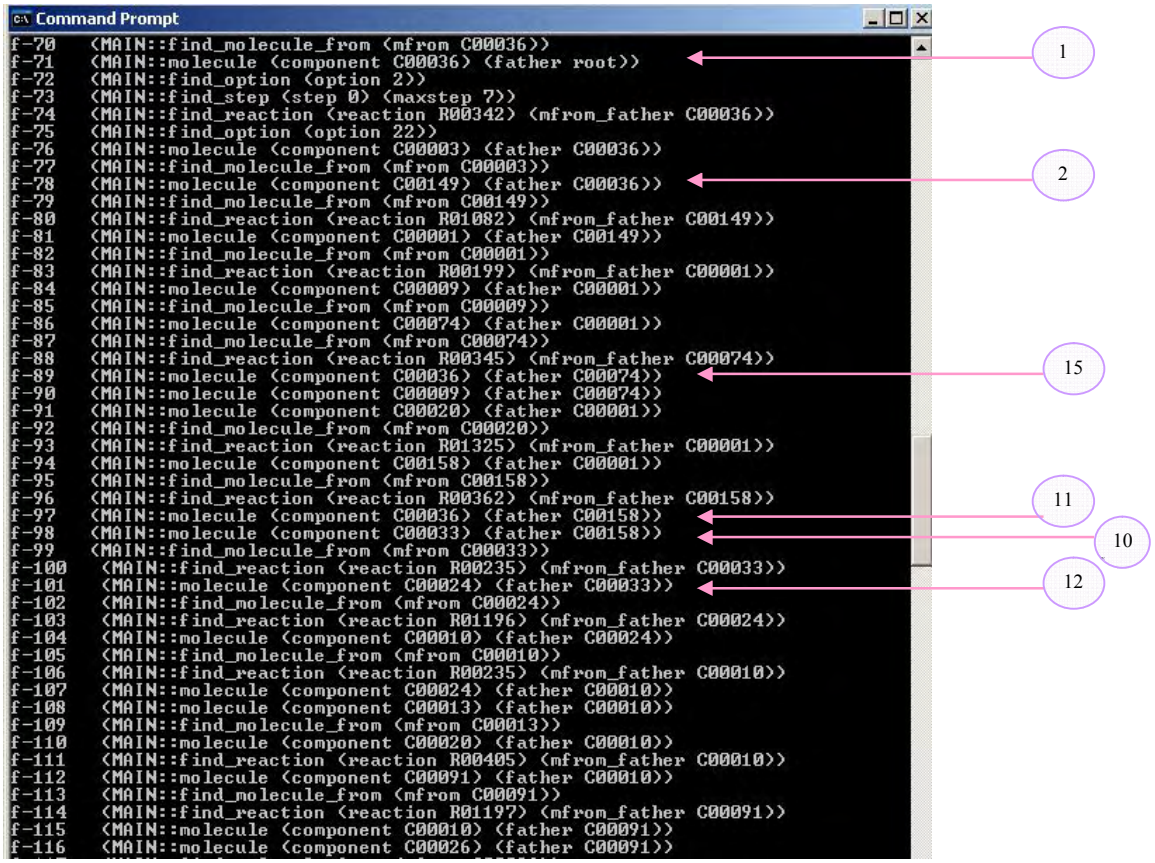

Figure **Error! No text of specified style in document.**.13. Result of the expert system. Scenario 3. Case 2.

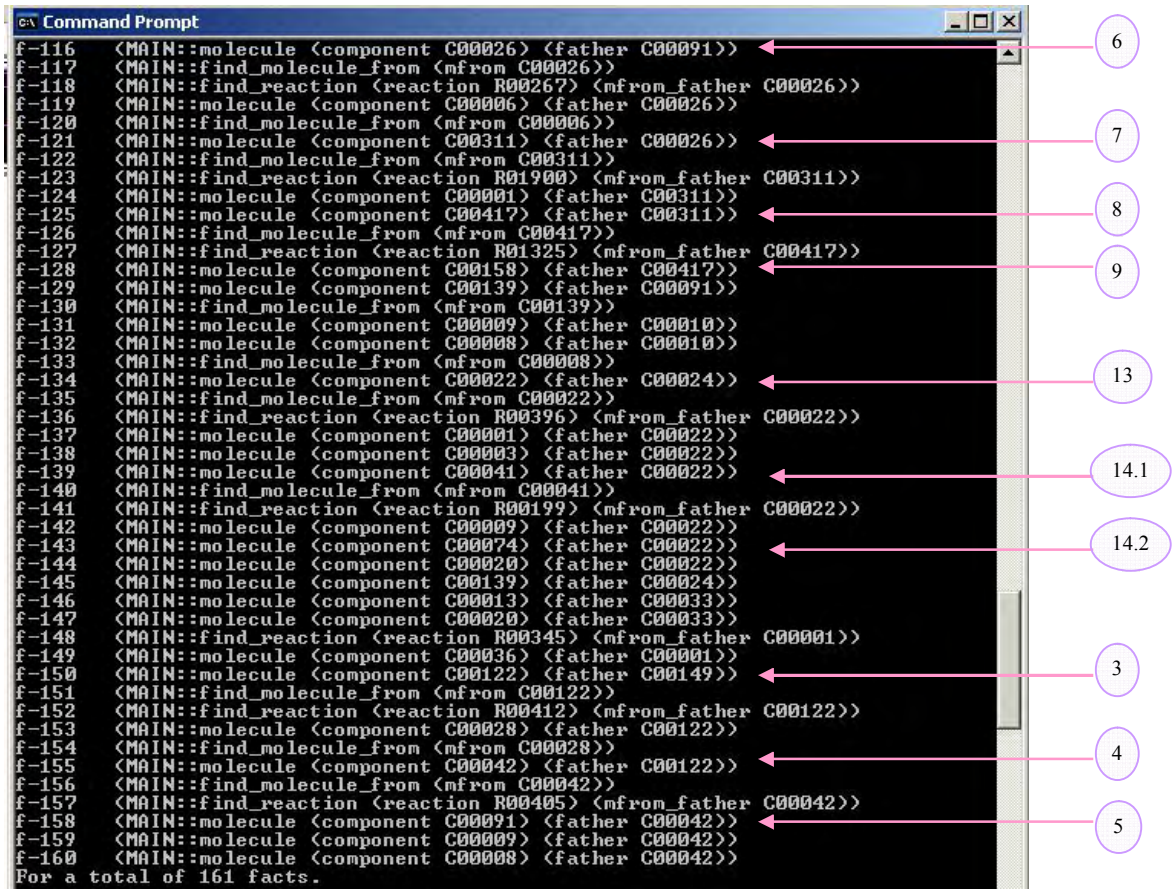

Figure **Error! No text of specified style in document.**.14. Result of the expert system. Scenario 3 Case 2 (cont.).

The edges and circles in the figures 5.11 and 5.12 indicate the sequence of nodes of the pathway in the Figure 5.13. In addition, the expert system finds other interactions between the reactions involved.

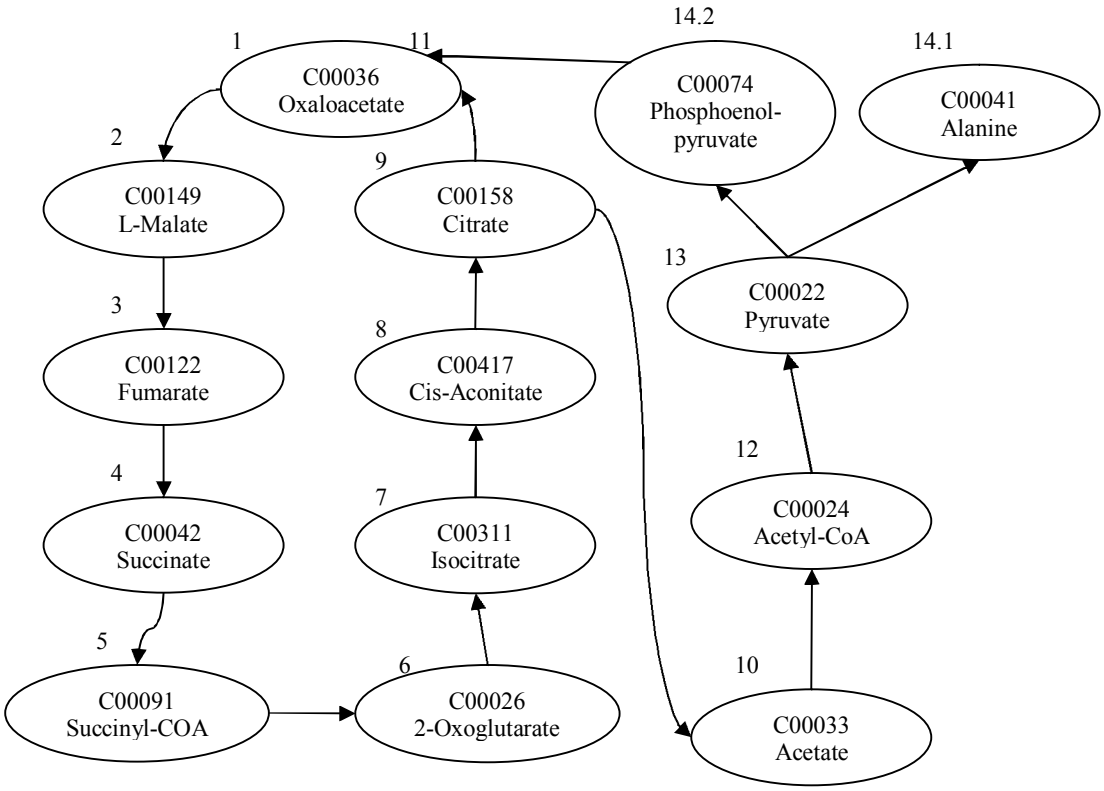

Figure **Error! No text of specified style in document.**.15. Pathway found by the expert system. Scenario 3. Case 2.

In this case, with Acetate (C00033) as the target molecule the expert system found it to be a branch of a cyclic pathway and ended when it did not find any more interactions. Here, the expert system first resolved the branch and then finished with the loop.

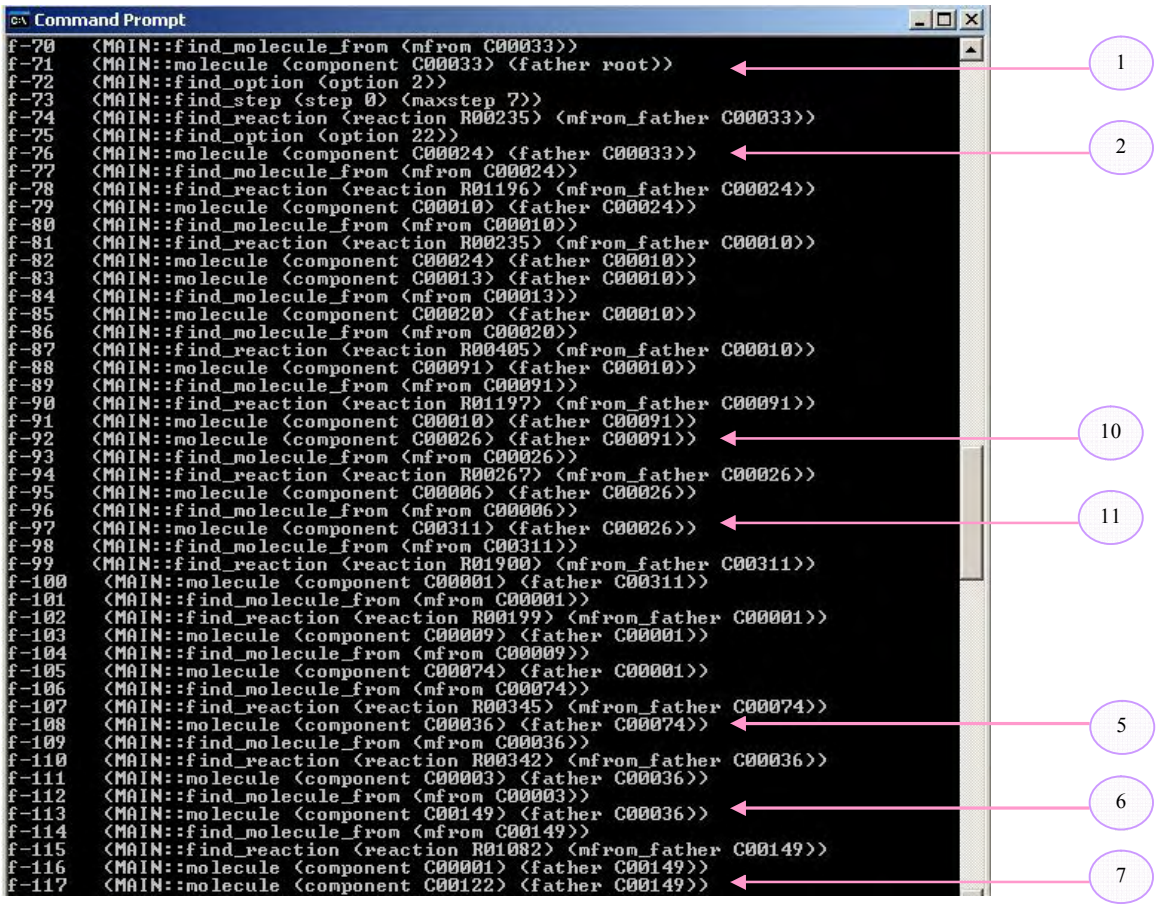

Figure **Error! No text of specified style in document.**.16. Result of the expert system. Scenario 3 Case 3.

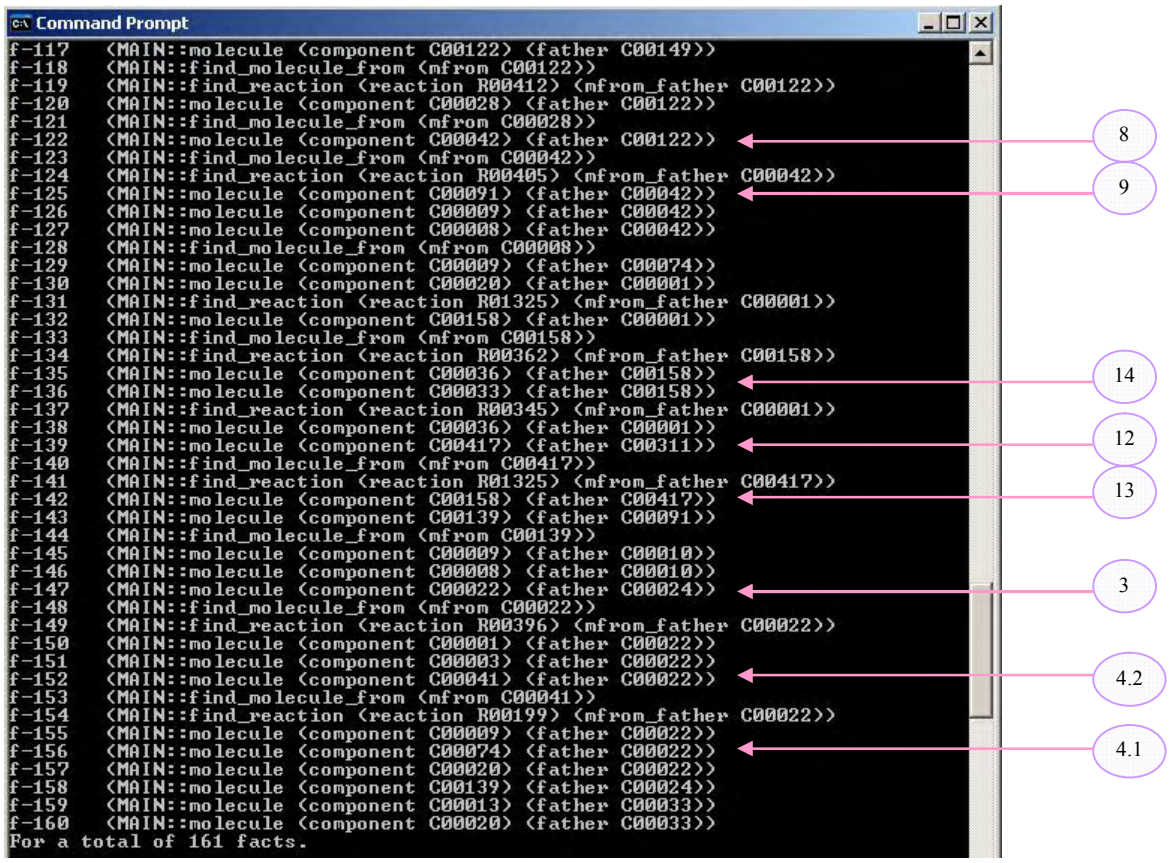

Figure **Error! No text of specified style in document.**.17. Result of the expert system. Scenario 3 Case 3 (cont.).

The edges and circles in figures 5.14 and 5.15 indicate the sequence of nodes of the pathway in Figure 5.16. Additionally the expert system finds others interactions from the reactions involved. This result is same that in case 2; the result not considered are of others interactions.

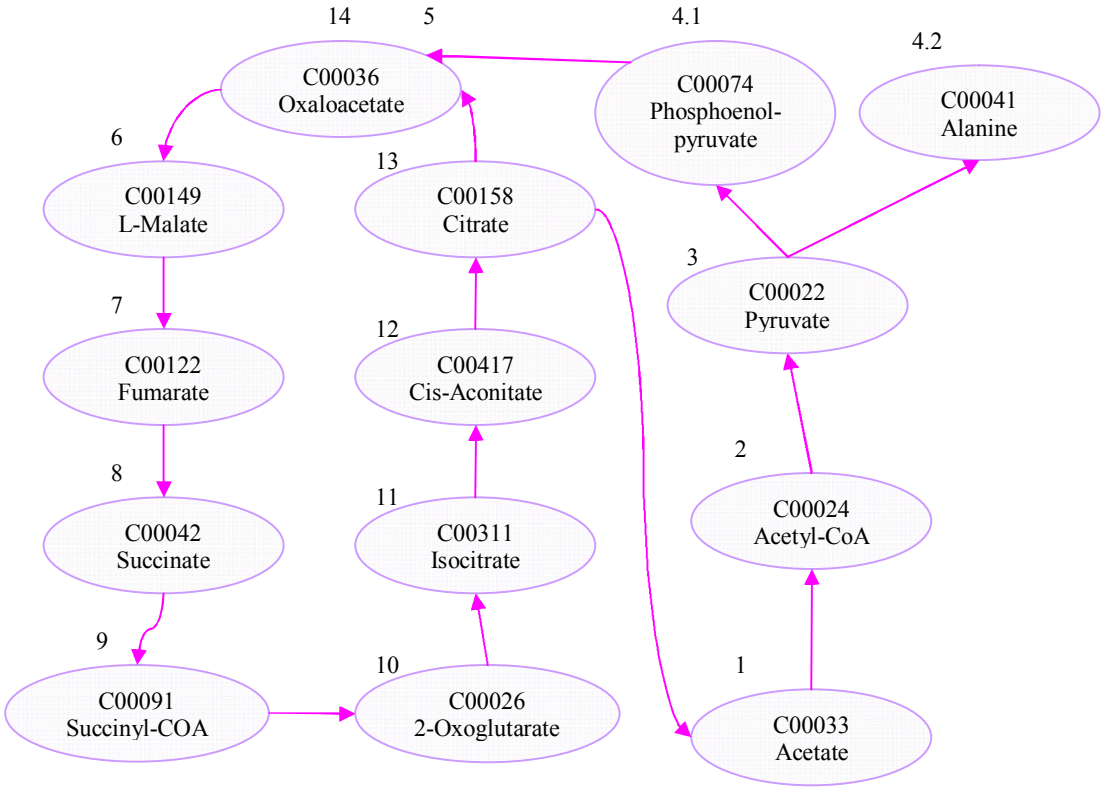

Figure **Error! No text of specified style in document.**.18. Pathway found by the expert system. Scenario 3. Case 3.

In accordance with Figure 5.16, the result is the same as in the previous case, this change in the order of the nodes but the pathway is the same.

As we can see that the expert system is capable of building a pathway similar to the known pathway in Figure 5.10. Notice that it can recover since any node of the pathway as in Case 3 where it began with the Acetate molecule.

Chapter VI

6. Concluding Remarks

6.1. Summary

In this thesis we have presented the design and development a based-rule expert system, the modeling and implementation of a relational database named BioPath and toolkit that allow populate the database. The result is an expert system prototype capable to predict new possible pathways by interconnecting pairs of molecules from known reactions, many of these molecules have not been identified into known pathways. We believe that this expert system will assist to the biologists in the search of new pathway for *R.sphaeroides*.

For access to the expert system was necessary creating a user interface as bridge between the users and the system, which was development using Java Servlet Technology. The expert system was development using Java Expert System Shell (JESS), and the toolkits were development using Java Programming Language.

We have conduced three scenarios to demonstrate that our expert system predict new possible pathway and is capable of retrieve pathways stored in our database. In the first scenario our system found one possible pathway among the Adenosine 5'-triphosphate to Hydrazine molecule, and another among the ATP or Adenosine 5'-triphosphate to Ferrocytochrome c3 and Reduced Menaquinone molecules. In the second scenario we have presented the 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) pathway for corroborate that our expert system stop when found the end target molecule, in this case was the 4-

Chlorobenzaldehyde molecule. In the last scenario we have discussed what happens when the expert system finds cyclical pathway. In this case the expert system stop when finds the end target molecule, reaches the maximum number of connections or not find any more interactions.

To populate our database we found four principal obstacles. The first was the task of different codification for the same attribute; the second was the heterogeneity and complexity of the biological information, the third was that data of *R. sphaeroides* was not available in a format easy to import to our database and finally the need to store data in a way that it yields useful knowledge to the experts.

Every source of information has its own structure that is a consequence of the domain (the biological focus) of the resource and the structure in which it is stored. This heterogeneity of structure makes it extremely difficult to design an interface generalized to multiple sources of information and to integrate the information inside that source of information.

To guarantee that the expert system is working correctly it is necessary to regularly update the data from the source databases and/or peer-reviewed publications.

6.2. Perspectives

We propose some future work to improve the quality of the system such as:

- Integrate an external open source Pathway Visualization Tool into the system to represent pathways as directed graphs. Where, in a pathway graph, nodes indicate biological molecules, arrows indicate the direction of reactions, and both nodes and arrows are labeled.

- To execute a study of complexity and time of response in a distributed environment to guarantee the suitable functioning of the system in terms of memory and performance using all the data on *R. sphaeroides*.

- To implement a automation mechanism that allows to update the information directly from source database.

Appendix A

Glossary of Terms

Biochemical Pathway - A network of interacting molecules that is responsible for a specific biochemical function, such as a metabolic pathway and a signal transduction pathway.

Biochemical Reaction - It is a chemical reaction that is carried out inside the alive organisms, on which the functioning of this one depends for his nutrition and survival.

Bioinformatics - A cross-disciplinary activity that includes aspects of computer science, software engineering, mathematics and molecular biology related to the handling and analysis of biological data.

Curated Database - A database that is maintained and updated regularly.

Chemical structure - Arrangement of atoms within a molecule, usually linked by covalent bonds.

DNA - Polynucleotide formed from covalently linked deoxyribonucleotide units; serves as the carrier of genetic information.

Functional Genomics - The study of obtaining an overall picture of genome functions, including the expression profiles at the mRNA level (transcriptome) and the protein level (proteome).

Gene - Region of DNA that controls a discrete hereditary characteristic, usually corresponding to a single protein or RNA. This definition includes the entire functional unit, encompassing coding DNA sequences, noncoding regulatory sequences, and introns.

Gene Expression - The multistep process in which a gene sequence is converted into a functional protein. The main steps in this process are transcription of a DNA sequence into RNA and translation of RNA into protein. Many times it also refers to the measurements of mRNA levels.

Genetic - Science of genes, heredity, and the variation of organisms. To describe the study of inheritance and the science of variation by English scientist William Bateson in a letter to Adam Sedgewick, dated April 18, 1905.

Genome - The total genetic material contained in the set of chromosomes.

Genomics - The study of the genome.

Genotype - Genetic constitution of an individual cell or organism.

Heuristics - A term in computer science that refers to "guesses" made by a program to obtain approximately accurate results.

Homeostasis - State of well being of an organism.

Messenger RNA (mRNA) - RNA molecule that specifies the amino acid sequence of a protein.

Molecular Biology - Study of biology at a molecular level; chiefly concerns itself with understanding the interactions between the various systems of a cell, including the interrelationship of DNA, RNA and protein synthesis and learning how these interactions are regulated.

Mutation - Heritable change in the nucleotide sequence of a chromosome.

Pharmacogenomics - The general study of all different genes that determine drug behavior.

Phenotype - Observable characteristics of an organism.

Polymorphisms - DNA variations among individuals that can either be single nucleotide changes or variations in tandem repeated sequences at a particular location in the genome.

Protein - The major macromolecular constituent of cells. A linear polymer of amino acids linked together by peptide bonds in a specific sequence.

RNA - Polynucleotide formed from covalently linked ribonucleotide units. Is has three classes, rRNA (catalyzes the binding of amino acids during protein synthesis), mRNA

(specifies the amino acid sequence of a protein), and tRNA (used as an interface between mRNA and amino acid during protein synthesis).

Xenobiotics - Compounds foreign to an organism.

References

[1] Aguero Fernan. “Bioinformatica y Genomica / IIB UNSAM (2002)”

[2] Claudel-Renard Clotilde, Claude Chevalet, Thomas Faraut and Daniel Kahn. PRIAM: Enzyme-Specific Profiles for metabolic pathway prediction. Nucleic Acid Research. (2003)

[3] Chen Ming and Ralf Hofestadt. “Web-Based Information Retrieval System for the Prediction of Metabolic Pathways”, Proceedings of the IEEE Transactions on Nanobioscience, Vol.3, No.3, September 2004.

[4] CLIPS Reference Manual: Volume 1: Basic Programming Guide. Version 6.23. And CLIPS, a Tool for Building Expert Systems. 2005

[5] Encyclopedia of Escherichia coli K12 Genes and Metabolism. 1996-2004. http://biocyc.org/ecocyc

[6] Fernandez G. Jose M., Maria del Mar Roldan, Jose Francisco Aldana and Alfonso Valencia. “Bases de Datos Biologicas y XML” 2001

[7] System for formalizad description, visualization, and modelling of gene networks. http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/

[8] Giarratano, Joseph “CLIPS User’s Guide”, Version 6.20, March 31st 20002

[9] Goesmann Alexander, Martin Haubrock, Folker Meyer, Jorn Kalinowski and Robert Giegerich. “PathFinder: reconstruction and dynamic visualization of metabolic pathways“. Bioinformatic.2002. Vol18 no 1 2002. Pages 124-129.

[10] Gonzalez Robledo Hugo F. “Manejador de Base de Datos MySQL”. 2000

[11] Goto, S, H. Bono, H. Ogata, W. Fujibuchi, T.Nishioka, K.Sato, and M.Kanehisa, “Organizing and computing metabolic pathway data in terms of binary relations,” Pacific Symp. Biocomp., 2:175-186 (1997)

[12] Graham J.L., Nicos Angelopoulos and Peter M.D.Gray., “Architecture of a Mediator for a Bioinformatics Database Federation”. IEEE Transactions on Information Technology in Biomedicine, Vol.6, No2, June 2002.

[13] Haubrock Martin and Alexander Goesmann. “Reconstruction and dynamic visualization of biochemical pathways.” Bioinformatics Vol.18 no.1 2002

[14] Integrated Microbial Genomes. The Regents of the University of California June 05. http://img.jgi.doe.gov/v1.1/main.cgi?page=taxonDetail&taxon_oid=400140000

[15] Jackson, Peter, Introduction to Expert Systems, 3rd Edition, Addison-Wesley, Harlow, UK, 1999

[16] Jess, the Rule Engine for the Java Plataform. Ernest Friedman-Hill at Sandia National Laboratories in Livermore, CA., March, 23 2005. http://herzberg.ca.sandia.gov/jess/

[17] Kaplan S., M.Choudhary, C.Mackenzie and N.J.Mouncey. "RsGDB, the Rhodobacter sphaeroides Genome Database". Nuclei Acids Res.1999 January 1;27(1):61-62

[18] Karp P.D., M.Krummenacker, S.Paley, and J.Wagg, "Integrated pathway-genome databases and their role in drug discovery," Trends Biotech., 17:275-281, (1999)

[19] Kazarov A. and Yu.Ryabov. "CLIPS Expert System tool: a candidate for the Diagnostic System engine.

[20] Bioinformatics Center, Institute for Chemical Research, Kyoto University with supports from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), and the Japan Science and Technology Corporation (JST)., "Kyoto Encyclopedia of Genes and Genomes", 2005, www.genome.ad.jp/kegg.

[21] Kiley Patricia and Samuel Kaplan. "Molecular genetics of photosynthetic membrane biosynthesis in Rhodobacter sphaeroides". Microbiological Reviews now published by the American Society for Microbiology. March 1988. 50-69

[22] Kolpakov, F.A., E.A. Ananko, G.B. Kolesov, and N.A. Kolchanov, "GeneNet: a gene network database and its automated visualization, "Bioninformatics, 14(6): 529-537 (1998)

[23] Krishnamurty L., J.Nadeau, G.Ozsoyoglu, G.Schaeffer, M.Tasan, W.Xu. "Pathways Database System: An integrated set of tools for biological pathways" 2003

[24] Lathrop Richard. "Intelligent Systems in Biology: Why the Excitement". IEEE Intelligent Systems 2001.

[25] Maibaum Michael, Galia Rimon, Christine Orengo, Nigel Martin and Alexandra Poulovassilis. "BioMap: Gene Family based Integration of Heterogenous Biological Databases using AutoMed Metadata". Proceedings of the 15th International Workshop on Database and Expert Systems Applications. (DEXA'04) IEEE

[26] Meyer Folker, Alexander Goesmann, Alice McHardy, Daniela Bartels, Thomas Bekel, Jorn Clausen, Jorn Kalinowski, Burkhard Linke, Oliver Rupp, Robert Giegerich and Alfred Puhler. "GenDB- an open source genome annotation system for prokaryote genomes". Nucleic Acids Research, 2003, Vol.31, No.8.

[27] Medical Research Council of Office of Science and Technology of UK Government. "What is bioinformatics?" London, December 01

[28] Parimala N. “Graphical User Interface to Multiple Biological Databases”. Proceedings of the 14th International Workshop on Database and Expert System Applications (DEXA’03) 2003.

[29] Paustian Timothy. and Robin S. Kurtz “Transposon Mutagenesis of Rhodobacter sphaeroides”. Department of Bacteriology. University of Wisconsin-Madison.

[30] Rajasekhar N., Ch. Sasikala and Ch. V.Ramana. “Photoproduction of L-tryptophan from indole and glycine by Rhodobacter sphaeroides OU5”. Biothecnology Appl. Biochem. (1999)

[31] Riikonen Pentti, Jorma Boberg, Tapio Salakoski, and Mauno Vihinen, “Mobile Access to Biological Databases on the Internet”, IEEE Transactions on Biomedical Engineering. Vol. 49, Nro.12, December 2002.

[32] Rojdestvenski Igor and Michael Cottam. “Visualizing metabolic networks in VRML”. Proceedings of the Sixth International Conference on Information Visualization. 1093-9547. (2002)

[33] Structural Classification of Proteins Database. 1.65 release (December 2003). http://scop.berkeley.edu/

[34] Selkov Jr E., et al., “MPV: The Metabolic Pathways Database”, “Nucleic Acids Res., 26(1):43-45 (1998)

[35] Swiss-Prot Protein Knowledgebase. Swiss-Prot Release 44.6 of 27-Sep-2004:159201 entries. http://us.expasy.org/sprot/

[36] Takako Takai-Igarashi and Tsuguchika Kaminuma. Division of Chem-Bio Informatics, National Institute of Health Sciences. Kamiyoga, Setagaya, Tokyo 158, Japan.

[37] Talaro Arthur and Kathleen Park Talaro, "Foundations in Microbiology", Fourth edition, (2002)

[38] Rhodobacter sphaeroides genome project. The University of Texas-Houston. Health Science Center, Department of Microbiology and Molecular Genetics. http://www.rhodobacter.org/

[39] The University of Minnesota Biocatalysis/Biodegradation Database. http://umbbd.ahc.umn.edu:8015/umbbd/servlet/search 2005

[40] Valdes-Perez, R.E., H.A. Simon, and R.F. Murphy, “Discovery of Pathways in Science,” Proc. Mach. Disc. Workshop, Intl. Conf. Mach. Learn.., Scotland, 1992.

[41] Van Laerhoven Kristof., Comparison of the CLIPS and JESS expert system shells. Project report for Industrial Applications of AI. June 4, 1999

[42] Yang Huiqing Maria A, Kevin B. Shaw. “A Clips-Based Implementation for Querying Binary Spatial Relationships”, 2001 IEEE. pag 2388 – 2393.