Expert System for Biochemical Pathway Inference

Maritza Elizabeth Córdova Bermeo

Advisor: Dr. Jaime E. Ramírez-Vick

Electrical and Computer Engineering Department
University of Puerto Rico – Mayagüez
Mayagüez, Puerto Rico 00681 – 5215
Elizabeth.Cordova@ece.uprm.edu, marelicorber@hotmail.com

Abstract

This paper describes a prototype rule-based expert system capable of inferring reactions between two molecules by connecting individual reaction steps using a set of rules stored in what is known as a knowledge base and molecules data stored in a local biological database named BioPathDB. This prototype focuses on the bacterium *Rhodobacter sphaeroides*.

1. Introduction

The value of genomics and bioinformatics in health care is enormous but will only be fully realized when knowledge and information are integrated at many levels.

The idea of integrating all the biochemical and genomic knowledge is important for efficiently reaching the understanding of how cells and organisms function. It is important to access the current knowledge on biochemical pathways, from the reaction networks to the sequence information on their constituents, into integrated pathway-genome databases. These integrated databases describe genes and genome of an organism using known biochemical pathways as the framework from which the current knowledge about their reactions, enzymes and metabolites can be accessed from other databases.

The main limitation of current biochemical pathway elucidation 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.

This paper proposes to implement a prototype rule-based expert system based on the C Language Integrated Production System (CLIPS). 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.

The bacterium *Rhodobacter 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 pathway reactions, and because of the widespread interest in its study.

In this paper we first give a brief biological background, followed by the system description.

2. Biological Background

Genes, gene structure and gene variations connect, through expression to regulatory and metabolic processes within cells. The fine control of these processes is modulated by spatial and temporal 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 can be quantified by measuring changes in gene expression or modifications in gene sequence, such as mutations.

Let's first give a definition of pathway that can accommodate its use in science and our extension of its scope to, e.g., expert systems:

> A pathway is an ordered set of R finite steps $\{step_1, step_2, \ldots, step_R\}$, each of the form reactants $\Rightarrow$ products, such that the reactants of step i are a subset of the species SM $\cup$ products(1) $\cup \ldots \cup$ 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 [5].

There are three reactions types: standard, metabolic and polymerization reactions; but we only include the two first; standard reaction includes reactions that consist of two elements, reactant and products and metabolic reactions consisting of three elements: reactants, products, and enzymes.

3. System Description

The rule-based expert system presented here allows us to find possible biochemical pathways reaction in *R. sphaeroides* from two target component. This system builds *de novo* biochemical reaction pathways based solely on known single reactions in the case that they have not being realized in the fragmented literature source from which all biochemical pathway maps are derived. Any new biochemical pathway found will be stored in BioPathDB with an initial validation, to be further analyzed by human experts.

3.1. Rhodobacter sphaeroides

R. sphaeroides is a facultative photoheterotroph belonging to the a-3 subdivision of the *Proteobacteria* [1]. Its genome consists of two circular chromosomes and five other replicons. The metabolic diversity of this group of organisms is unparalleled. The structure and function of the photosynthetic membrane protein complexes and their regulation, photolithography, nitrogen fixation, hydrogen metabolism, carbon dioxide fixation, taxis and tetrapyrrole biosynthesis, when considered *in toto* in *R. sphaeroides*, are remarkable. In addition, it has extremely facile methodologies for genetic manipulations, gene transfer, genetic analyses, and chromosomal mobilization [4].

3.2. Rule-Based Expert System

This system consists of two types of files (the local database BioPathDB and the knowledge base) and three main modules (the database management system, the extractor, and the inference engine).

BioPathDB is managed by PostgreSQL database management system (DBMS), which stores *R. sphaeroides* data extracted from existing databases (e.g., SWISSPROT, *R. sphaeroides* Genome Project, etc.). This database includes information about all known biochemical pathway reactions in *R. sphaeroides* (i.e., metabolic, signal transduction, etc.) and molecules involved in them (e.g., biomolecules and xenobiotics). [Figure 1]

The knowledge base is a text file that contains the rules used by the inference engine; here the assertions are reactions, which are the individual steps in the pathways. The rules find interconnections between the reactions. In general, these assertions take the form of a reaction or a pathway. For example for a standard set of reactions assertions will look like:

> *Define Reaction (from "A" to "B") (in sequence "A? B")*

Define Reaction (from "B" to "C") (in sequence "B? C")

Assert Pathway (Sequence "A? B" "B? C") (with "2" connecting steps).

In BioPathDB, the assertions are reactions, which are the individual steps in the pathways. The rules find interconnections between the reactions. There are two procedures for connecting reactions: (1) Finding pathways around a target molecule, and (2) Finding pathways between two target molecules. Here we will focus on Procedure 2. An example of assertions and rules for this procedure follows:

[RULE 1] *Verify if there are Target Molecule "S" and "E" are in the BioPathDB local database.*

[RULE 2] *Initiate Pathway with Target Molecule "S" and a Maximum of "M" connections*.

[RULE 3] *Produce Succeeding Pathway*

[RULE 4] *Stop the inference if the Receiver molecule of the connected Reaction is equal to the End Target "E", or if the Maximum number of Connections is equal to "M"*

[RULE 5] *Remove disused Assertions*

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 or the left side (LHS) and a *then* portion right side (RHS). The *if* portion of a rule is a series of patterns which specify the facts which cause the rule to be applicable. 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 [3].

The extractor accepts queries from the user interface, reads the data from the local database via the DBMS, to check for synonyms, sorts the data, to add restrictions, and transfers them to the inference engine.

The inference engine is CLIPS), which is written in the C programming language, and has the advantage that it can be embedded into other applications written in C. It is also a public domain production system shell highly productive and used easily. It provides a complete environment for the construction of rule and/or object-based expert systems. It can be called as a subroutine and information may be passed to and from it. Because of its portability, extensibility, capabilities, and low-cost, CLIPS has received widespread acceptance throughout the government, industry and academia [2].

The inference engine selects a rule and then the actions of the selected rule are executed (which may affect the list of applicable rules by adding or removing facts). Then, it selects another rule and executes its actions. This process continues until no applicable rules remain [3].

A typical session between a user and BioPathDB might proceed in the following way:

1. The user gives the name of target molecules and restrictions to BioPathDB.
2. The objects stored in the Reaction class are retrieved from the database and translated into the form of our assertions, which are then transferred to the Inference Engine. Restrictions specified by the user are processed.
3. The inference engine evaluates the rules for the assertions, including restrictions if they exist, and the answers are sent to the graph drawer utility.
4. The graph drawer utility converts the answers into a directed graph, and prints the graph for the user.

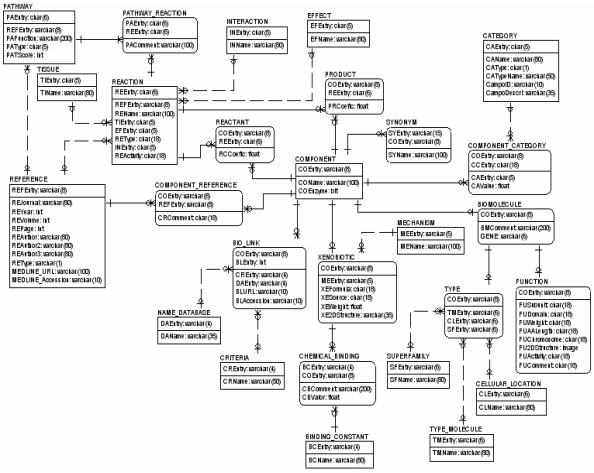

Fig 1. BioPathDB Database Structure

3.3. Architecture

When the extractor accepts a query from the user, it reads the data from BioPathDB via the DBMS, checks for synonyms, sorts the data, adds restrictions, and transfers them to the inference engine. The DBMS receives objects, and return them to the extractor. The communication between DBMS and inference engine is through a library in C.

Query strings specified by the user are checked to determine whether they are precise object names or synonyms, which the extractor converts into object names. This conversion is based on the synonym table included in BioPathDB. As mentioned above, restrictions are added by specifying the maximum number of connecting steps, narrowing the domain to be covered, or by eliminating sub trees. The total number of connecting steps is specified by the user and embedded in the knowledge base. Preceding or succeeding pathways can be selectively or disregarded. The Extractor regulated this by eliminating the corresponding rules from the Rule File and filtering out corresponding assertions in the process of transferring the assertions to the Inference Engine.

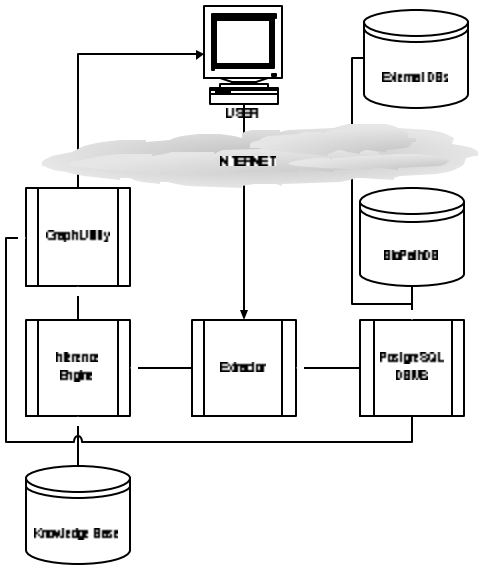

Fig 2. System Architecture

4. Conclusions

The rule-based expert system presented here can produce probable *R. sphaeroides* biochemical reaction pathways, which can be further validated by human experts.

The number biochemical reaction pathways produced will depend of maximum number of connections steps, narrowing the domain to be covered, or by eliminating sub trees.

References

1. http://spider.jgi-psf.org/JGI_microbial/html/rhodobacter/rhodob_homepage.html

2. Giarratano, J.C. "CLIPS User's Guide", Version 6.20, March 31st 20002.

3. Jackson, P., Introduction to Expert Systems, 3rd Edition, Addison-Wesley, Harlow, UK, 1999.

4. http://mmg.uth.tmc.edu/sphaeroides/

5. 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.