In Silico Systems Analysis of Biopathways

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
der Universität Bielefeld.
Vorgelegt von

Ming Chen

Bielefeld, im März 2004

MSc. Ming Chen
AG Bioinformatik und Medizinische Informatik
Technische Fakultät
Universität Bielefeld
Email: mchen@Techfak.Uni-Bielefeld.DE

Genehmigte Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr.rer.nat.)
Von Ming Chen am 18 März 2004
Der Technischen Fakultät an der Universität Bielefeld vorgelegt
Am 10 August 2004 verteidigt und genehmiget

Prüfungsausschuß:

Prof. Dr. Ralf Hofestädt, Universität Bielefeld
Prof. Dr. Thomas Dandekar, Universität Würzburg
Prof. Dr. Robert Giegerich, Universität Bielefeld
PD Dr. Klaus Prank, Universität Bielefeld
Dr. Dieter Lorenz / Dr. Dirk Evers, Universität Bielefeld

Abstract

In the past decade with the advent of high-throughput technologies, biology has migrated from a descriptive science to a predictive one. A vast amount of information on the metabolism have been produced; a number of specific genetic/metabolic databases and computational systems have been developed, which makes it possible for biologists to perform in silico analysis of metabolism. With experimental data from laboratory, biologists wish to systematically conduct their analysis with an easy-to-use computational system. One major task is to implement molecular information systems that will allow to integrate different molecular database systems, and to design analysis tools (e.g. simulators of complex metabolic reactions). Three key problems are involved: 1) Modeling and simulation of biological processes; 2) Reconstruction of metabolic pathways, leading to predictions about the integrated function of the network; and 3) Comparison of metabolism, providing an important way to reveal the functional relationship between a set of metabolic pathways.

This dissertation addresses these problems of in silico systems analysis of biopathways. We developed a software system to integrate the access to different databases, and exploited the Petri net methodology to model and simulate metabolic networks in cells. It develops a computer modeling and simulation technique based on Petri net methodology; investigates metabolic networks at a system level; proposes a markup language for biological data interchange among diverse biological simulators and Petri net tools; establishes a web-based information retrieval system for metabolic pathway prediction; presents an algorithm for metabolic pathway alignment; recommends a nomenclature of cellular signal transduction; and attempts to standardize the representation of biological pathways.

Hybrid Petri net methodology is exploited to model metabolic networks. Kinetic modeling strategy and Petri net modeling algorithm are applied to perform the processes of elements functioning and model analysis. The proposed methodology can be used for all other

metabolic networks or the virtual cell metabolism. Moreover, perspectives of Petri net modeling and simulation of metabolic networks are outlined.

A proposal for the Biology Petri Net Markup Language (BioPNML) is presented. The concepts and terminology of the interchange format, as well as its syntax (which is based on XML) are introduced. BioPNML is designed to provide a starting point for the development of a standard interchange format for Bioinformatics and Petri nets. The language makes it possible to exchange biology Petri net diagrams between all supported hardware platforms and versions. It is also designed to associate Petri net models and other known metabolic simulators.

A web-based metabolic information retrieval system, PathAligner, is developed in order to predict metabolic pathways from rudimentary elements of pathways. It extracts metabolic information from biological databases via the Internet, and builds metabolic pathways with data sources of genes, sequences, enzymes, metabolites, etc. The system also provides a navigation platform to investigate metabolic related information, and transforms the output data into XML files for further modeling and simulation of the reconstructed pathway.

An alignment algorithm to compare the similarity between metabolic pathways is presented. A new definition of the metabolic pathway is proposed. The pathway defined as a linear event sequence is practical for our alignment algorithm. The algorithm is based on strip scoring the similarity of 4-heirachical EC numbers involved in the pathways. The algorithm described has been implemented and is in current use in the context of the PathAligner system.

Furthermore, new methods for the classification and nomenclature of cellular signal transductions are recommended. For each type of characterized signal transduction, a unique ST number is provided. The Signal Transduction Classification Database (STCDB), based on the proposed classification and nomenclature, has been established. By merging the ST numbers with EC numbers, alignments of biopathways are possible.

Finally, a detailed model of urea cycle that includes gene regulatory networks, metabolic pathways and signal transduction is demonstrated by using our approaches. A system biological interpretation of the observed behavior of the urea cycle and its related transcriptomics information is proposed to provide new insights for metabolic engineering and medical care.

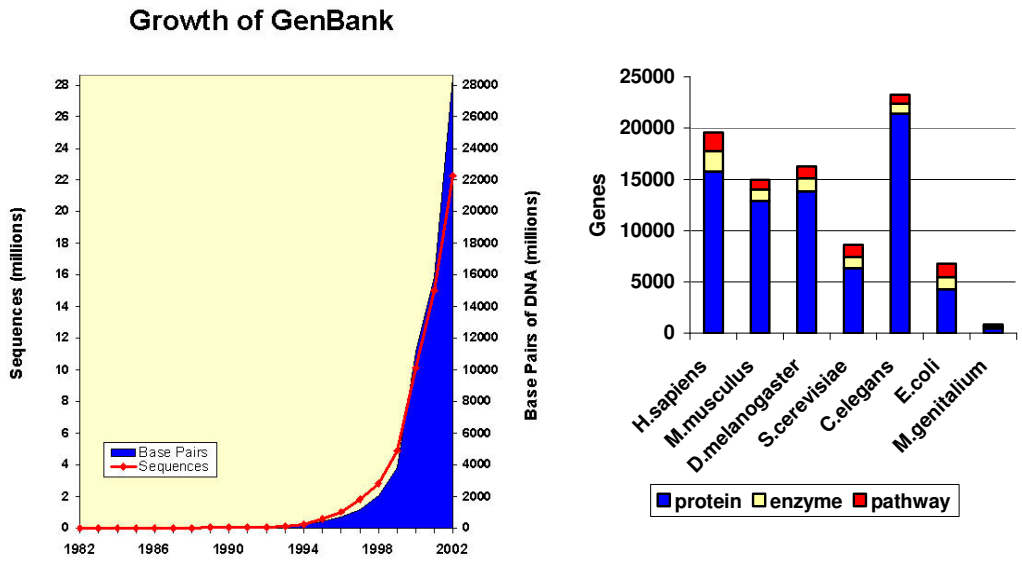
Table of Contents

Chapter 1
Introduction

1.1 The Problem

In the past, much of biological research has focused on data collection. The main reason for this is that gathering data was by itself much work. However biology is changing, especially because of the availability of large amounts of data that is easily accessible via the Internet [Col02]. Genome projects generate enormous amounts of information. The amount of sequence data is increasing exponentially over time (Figure 1.1a), and this growth will likely continue for the foreseeable future. The diversity and accumulating of biological data both on genomic and metabolic levels from different species (Figure 1.1b) bring a new challenge for revealing what life really is. Extraordinary successes of the genome projects push the need for the development of more sophisticated and powerful computational techniques.

We are in a "post-genomic" era. Although sequence analysis have been and still are the most common topics in the bioinformatics studies, we are looking for computational methods and tools to predict functional details. It takes bioinformatics beyond its original boundaries. It is certainly not data acquisition for molecular biology, but it is about the application of computer techniques, such as data abstraction, data manipulation, modeling, simulation, and functional analysis. The data generated by the experimental scientists requires annotation and detailed analysis in order to turn it into knowledge that can then be applied to, for example, healthcare, agriculture, industry and environment, to improve health care via gene prediction, drug design, gene therapy, and much more.



(a) (b)

Figure 1.1 Ever growing biological data and their complexity. (a) The exponential growth of DNA sequences in GenBank over time (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). (b) This chart indicates the known genes and their functional contribution in different species. The source data is taken from KEGG, 12.04.2003 (http://www.genome.ad.jp/kegg/docs/upd_genes.html).

Moving from sequence to structure to function to application, bioinformatics developments are occurring in genome modeling and annotation, comparative protein modeling and folding assignment, algorithmic development, in silico drug design, mechanistic enzymology and modeling of cellular processes. Biological data functional analysis is a major topic beyond genome research. Computational metabolics focuses on the computational interpretation of cellular phenomena that involve not only nucleic acid and protein sequences, but also metabolic pathways, gene regulatory networks and signaling pathways. In this sense systems analysis of metabolic network is becoming a promising field.

The development of computer science makes it possible to represent the complex metabolic network of physical and functional interactions, which take place in living cells, in ways which enable us to manipulate, analyze and achieve understanding of how cells function.

In order to understand the logic of cells, methods of systems modeling and simulation are needed to find the interrelationships among different molecules and reactions. Fortunately the data and knowledge of genes, proteins and pathways are available, and various biology database systems are accessible. The *status quo* in modern biology, especially in molecular biology, is that exponential growing biological data are produced. For instance, with powerful computers and robot sequencers, small genomes, such as bacterial genomes, can be completely sequenced in a matter of weeks or days. Protein substances can be promptly analyzed by

automatic amino acid analyzer and latest development of high technology in GS-MS, HPLC, NMR and others accelerates the data accumulation. In the meantime, some of these experimental data are collected and stored into a well-formatted way and provide an easy access for the public.

Now, suppose a patient gets a diagnosis of metabolic disease (a disorder caused by malfunction of normal enzyme reactions), what is the metabolic mechanism of it? Most diseases are related to some kind of enzyme insufficiency and the malfunction of signal transduction pathways which regulate the expression of the genes that encode the desired enzymes. How does the gene or enzyme defect, that leads to a blocked biochemical reaction function? A good model of the metabolic reactions is appreciated to see the detail information about the essential proteins or enzymes and their regulations to the disease. With such a model, we might easily figure out the real causes, further development of the disorder, and possible alternative pathways to overcome the blockades. Unfortunately, due to the complex interconnection among metabolic reactions, current models are only present parts of the whole metabolic network in a living cell. It is necessary to develop a large-scale network model automatically.

Although a lot of biological data are available today, some other data, especially those on metabolic regulation, are still insufficient. Given a set of rudimentary biological data, such as DNA and protein sequences, some enzymes and chemicals, can we predict the complete gene controlled metabolic pathway, understand the complexity of networks (cross interaction and regulation both on biochemical reactions and gene regulation, transcription factors, etc.), and try to model and simulate it? Considering the availability of metabolic databases we try to find relations of the rudimentary data. Is it possible to develop a web-based information system for biopathway retrieval and functional analysis with the emphasis on analysis rather than storage? Can this information system ensure that analyzed data remains up-to-date in the light of new data, as well as reporting new information as it becomes available?

If the data is still rough, can we make a comparative analysis of pathways between human and mouse or some other model animals that have more detailed pathway information? In order to find function-related pathways, to interpret evolution processes on metabolic level and to determine alternative pathways, pathway alignment is needed. That is, given two biopathways (metabolic pathways or signaling pathways), can we calculate the similarity of them?

1.2 Three Profiles

The problem mentioned above contains three main profiles of in silico systems biology research: 1) Modeling and simulation of biological processes, i.e. computer modeling of metabolism, based on experimental data. 2) Prediction of metabolic pathways based on

annotated genome (or transcriptome) sequences and metabolic data, leading to predictions about the integrated function of the network. 3) Comparison of metabolism based on the analysis of presence/absence of sequence and/or metabolite patterns, providing an important way to reveal the functional relationship between a set of metabolic pathways.

Various groups of academic scientists and researches from biotechnology, informatics and pharmaceutical companies are coming together to try to solve these problems. A little more description of the motivation of the three problem profiles is presented in the following three sections.

1.2.1 Metabolic Networks Modeling, Simulation and Analysis

New high-throughput technologies in genomics, transcriptomics, proteomics and metabonomics enable us to estimate the metabolism on a system-wide level and decipher the biological regulatory processes in a quantitative manner. Modeling and simulation is a fundamental and quantitative way to understand complex systems, which is complementary to the traditional approaches of theory and experiment. In some cases, simulating increasingly complex networks will help us to understand the impact of various factors (e.g. enzyme insufficiency, metabolic blockade, drugs effects, etc.) on metabolic systems. This is particularly useful in the pharmaceutical industry for designing site-directed drugs to target mutant enzymes.

Nevertheless, it is very difficult and challenging to model metabolic systems and to perform computer simulations on them, as metabolic systems are inherently complex information processing systems that are governed by numerous biological and natural processes. The availability of high performance computers, coupled with mathematical modeling, has contributed to the development of increasingly accurate models of metabolic systems. This makes it possible to represent the complex metabolic network of physical and functional interactions in ways which enable us to manipulate, analyze and understand cell functioning. Several well-known biological simulation software packages such as Gepasi [Men93], Dbsolve [Gor99] and DynaFit [Kuz96] for quantitative simulation of biochemical metabolic pathways, based on numerical integration of rate equations, have been developed. Those studying biochemical system simulations usually limit their models to focus on only one of the several levels of time-scale hierarchy in cellular processes. Linking the gaps between the various levels of this hierarchy is an extremely challenging problem that needs to be addressed. Several approaches such as E-Cell project [Tom01] are attempting to achieve a systems cell modeling. We propose to model and simulation integrated metabolic networks by using Petri net methodology [Pet62] and hope to give a highlight on the field. We want to use Petri net methodology to explore the cellular processes on a system-wide level. The aim is to

understand not only the functions of individual genes, proteins and smaller molecules like hormones, but also to learn how all of these molecules interact within a cell. We hope to use this information to generate more accurate computer models that will help unravel the complexities of cellular functions and the underlying mechanisms of metabolic disorders.

In the past years, different Petri net tools were used to model and simulate metabolic pathways and gene regulatory networks. However, most of Petri net tools import and/or export Petri net diagrams in a binary file format, which poorly supports the possibility of making diagrams distributed in multiple format files or constructing a net by a text format file. That means it was impossible to extract data from biology databases and construct a Petri net model automatically. Although several Petri net tools such as PNK [Jue00], Renew [Kum00] and CPN [Lyn98] have been equipped with an XML based file format, they have their different definitions and ontology because of the differences of design destinations. As a result it is difficult to exchange models between different analysis and simulation tools and take advantage of them. One cannot adapt ones Petri net XML file to fit without any modification. Moreover, every user has to write an XML file from the original data source, which is time-consuming. With regard to applying the Petri net methodology to metabolic networks, a new standard would be helpful. With so many software tools, but few common exchange formats, even with XML format, we are motivated to propose a common exchange language – Biology Petri Net Markup Language (BioPNML) for metabolic networks Petri net modeling. The aim is to enable exchange of models between metabolic data and Petri net tools, as well as other bio-simulators. It uses a simple, well-supported textual substrate (XML) and adds components that reflect the natural conceptual constructs.

1.2.2 Biopathway Prediction

More than 500 database systems are available which represent molecular data. Therefore, experimental data and experimental results of fundamental metabolic processes like gene regulation, metabolic pathway control, signal pathway control and cell differentiation processes are available via the internet [Col02].

In order to improve our understanding of cells and organisms as physiological, biochemical, and genetic systems, we have to study them as an integrated metabolic system. It is clear that the next step of implementing these databases is to integrate them under a specific biological perspective. Retrieving metabolic pathways from current biological data, reconstructing metabolic pathways from some rudimentary components such as genes, gene sequences, proteins, protein sequences and other biological molecules are one of the major tasks in bioinformatics. Broadly speaking, there are two senses in which reconstruction of metabolic pathways is being done: (1) Completing metabolic pathways by mining genomic databases to ‘discover’ enzymes and proteins that are not cloned and that may not have been

suspected to exist. (2) Integrating available genome, transcriptome, and proteome information into useful computer models of pathways and cellular processes without the global, quantitative comprehension, that at first sight seems necessary.

Attempts have been made to reconstruct metabolic pathways either via genome sequence comparison [Mus96] [Bon98] [Sun02], enzyme assignment [van00] and enzyme EC numbering [Oga98]. However, they have a number of limitations. Predicting each gene function based solely on sequence similarity often fails to reconstruct cellular functions with all the necessary components. They do not contain comprehensive information about metabolic pathways, such as physical and chemical properties of the enzymes that are involved. Some approaches are not fully computer-aided. The individual database search process requires too much human intervention, and the quality of annotation largely depends on the knowledge and work behavior of human experts. The future of metabolic pathway analysis may depend upon its ability to capitalize on the wealth of genetic and biochemical information currently being generated from genomic and proteomic technologies. An ideal system for metabolic pathway prediction would include a web-based architecture to allow remote and local access to the different biological databases. It would offer a proven approach that can perform complex queries, data transformations, and data integration under one simple interface, without requiring extensive programming. We are motivated to develop such a web-based information retrieval system that will help the prediction of metabolic pathways.

1.2.3 Biopathway Alignment

Nucleic acid and protein sequence comparison is an important tool in genome informatics. Initial clues to understanding the structure or function of a macromolecular sequence arise from homologies to other macromolecules that have been previously studied. Many applications and tools, such as BLAST [http://www.ncbi.nlm.nih.gov/BLAST] and FASTA [http://www.ebi.ac.uk/fasta3], are developed to further understand the biological homology and estimate evolutionary distance.

Recently the emphasis of research efforts begins to turn back from gene sequences to metabolic pathways. It is therefore not surprising that the development of computational algorithms to predict metabolism function from gene, amino acid sequences and metabolic networks is now a core aim of bioinformatics. As more genomes are sequenced and the metabolic pathways reconstructed, it becomes possible to perform biological comparison from a biochemical-physiological perspective. Alignments represent one of the most powerful tools for comparative analysis of metabolism. Metabolic pathway alignment is of importance to study biology evolution, pharmacological targets and other biotechnological applications [Dan99], such as metabolic engineering and metabolism computation. A metabolic pathway alignment is a mapping of the coordinates of one pathway onto the coordinates of one or more

other pathways. For example, the same metabolic pathway from two organisms may have diverged if the organisms evolved from a common ancestor, where individual metabolites and enzymes may have been changed, added or lost in one pathway. It involves recognition of metabolites that are common to a set of function-related metabolic pathways, interpretation of biological evolution processes and determination of alternative metabolic pathways. Moreover, it is of assistance in function prediction and metabolism modeling. Although researches on genomic sequence alignment have been intensively conducted, so far the metabolic pathway alignment is less studied. Several approaches of metabolic pathway alignment have already been made by Dandekar et al. [Dan99], Forst C.V. [For99] and Yukako T. [Toh00a] [Toh00b]. However, their definitions of pathways are traditional biochemical pathways such as glycolysis, the pentose phosphate pathway, and the citric acid cycle. Less effort is made on analysis of gene regulatory networks as well as signaling pathways.

In this thesis, we try to give out basics of common definitions of metabolic pathway, gene regulatory pathway and signaling pathway. We also present a biopathway alignment to characterize comparatively the metabolic pathways and signaling pathways in cells.

1.3 Content of Dissertation

This thesis is primarily concerned with systems analysis of biopathways. It provides a toolbox to predict metabolic pathways from rudimentary data, to automatically construct a Petri net model for modeling and simulation, and to comparatively analyze biopathways. In Chapter 2 a brief overview of systems metabolic analysis and literature review is provided on the biological complexity, Petri net based modeling and simulation, and prediction of metabolic pathways, as well as algorithms for biopathway alignment with particular emphasis on metabolic pathways. Chapters 3, 4 and 5 are the core of the thesis. Chapter 3 presents the Petri net methodology for metabolic network modeling and simulation. An explicit example is explained. A proposed standard for biological data interchange, BioPNML, is presented. Chapter 4 presents the theoretical and practical approaches for the retrieval and reconstruction of metabolic pathways from rudimentary components. In Chapter 5 we present a new algorithm for metabolic pathway alignment. A classification of signal transductions is recommended and we discuss biopathway alignment in Chapter 6. Chapter 7 presents a case study of urea cycle biopathway. The conclusion of the study is presented in Chapter 8.

Chapter 2
State of the Art

2.1 Biology Basics

2.1.1 Biological Complexity

Life is the process of metabolism that transforms compounds such as carbohydrates, amino acids and lipids, and the energy required and all the other components that take up living systems to synthesize them and to use them in creating proteins and cellular structures as well as sustaining life. A cell contains a great numbers of organelles, specific proteins, and much more (Figure 2.1.1a). There are thousands of biochemical reactions taking place per second in a living cell. In *Escherichia coli*, for instance, there are 225,000 proteins, 15,000 ribosomes, 170,000 tRNA-molecules, 15,000,000 small organic molecules and 25,000,000 ions inside the a few µm cell [Goo93]. There are estimated 10^{14}-10^{16} biochemical reactions in a cell [End01]. These reactions are interconnected by the metabolic molecules. Many molecules involved in one reaction can also be found in other reactions where the molecules act as substrate or activator or repressor, the activities of enzymes are enhanced or inhibited by some molecules. Proteins and enzymes are synthesized from encoding genes which can also be switched on or off by some other molecules. Thus a densely connected, intricate and precisely regulated reaction network is built (Figure 2.1.1b). These connected biochemical reaction is normally called a metabolic network. Obviously, the more interconnections exist, the harder it gets to predict how the system will react. When systems reach a certain size, they will be become unmanageable and difficult to understand without the help of computational support. It also gets harder to change any part of the system without influencing other parts.

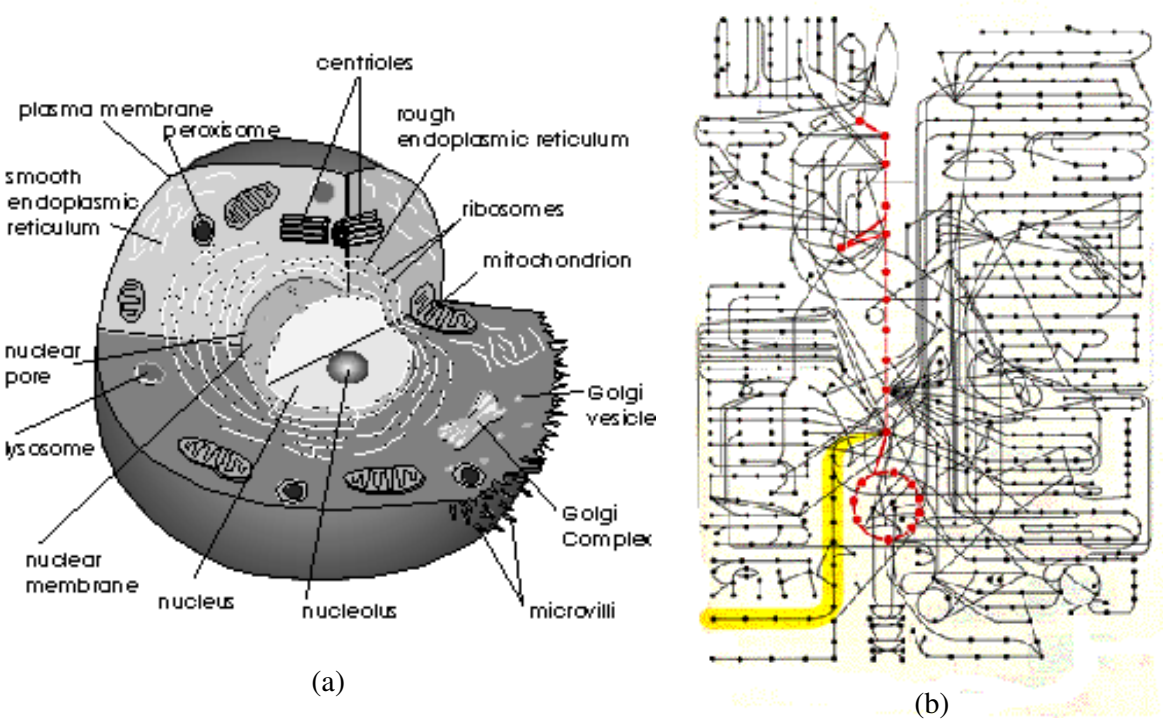


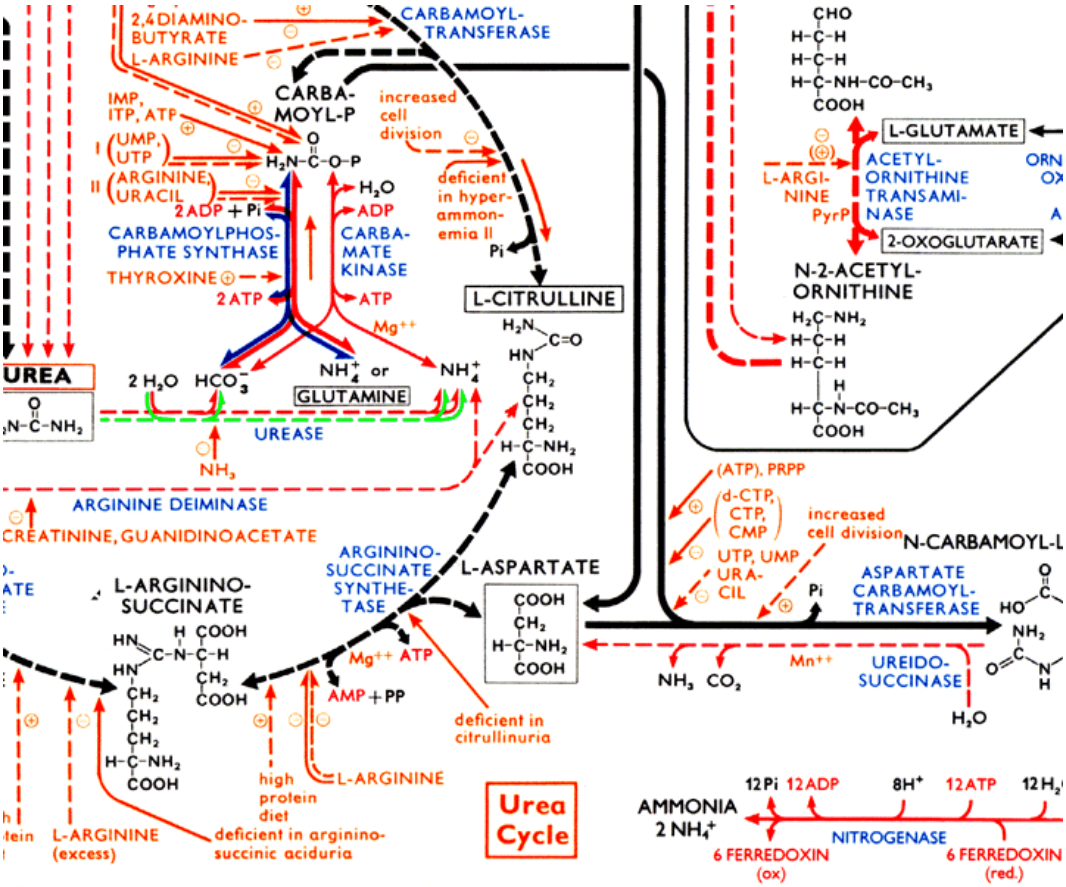
Figure 2.1.1 Illustration of complexity of cell. (a) A typical illustration of a mammalian cell structure [Rud97]. (b) Representation of about 500 common reactions of the basic metabolic network [Alb94]. Each point (node) represents a distinct chemical substance and each line (edge) represents a simple chemical transformation, catalyzed by a separate enzyme. A typical mammalian cell synthesizes more than 10,000 different proteins, a major proportion of which are enzymes. The central vertical line and circle represent what biochemists call the "glycolytic pathway" and the "citric acid cycle", the bases of cellular energetics.

2.1.2 Biopathways

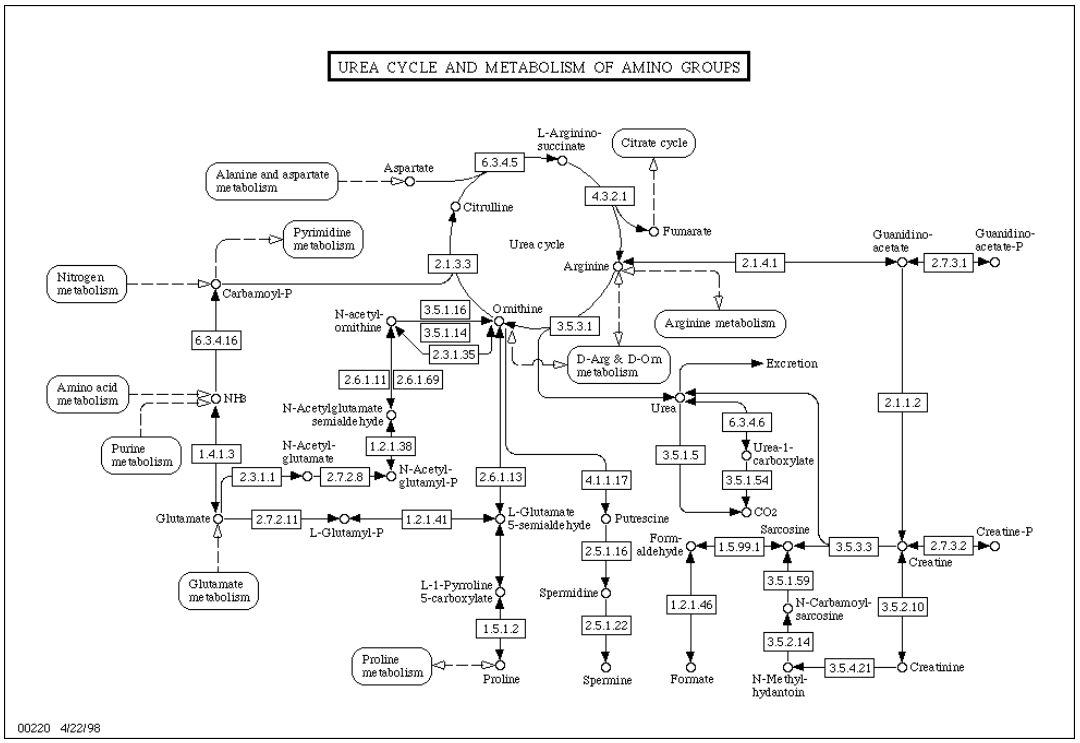
Although many of the interconnected systems of biochemical reaction pathways have been known for some time, knowledge of integrated functioning of metabolic systems remains elusive. That is, the functional definition of metabolic networks and their role in the context of the whole cell is lacking. On the other hand, it is usually impossible to evaluate and analyze the huge amount of interactions as a whole due to its extreme complexity. People divide those biological processes into three levels: gene regulation, biochemical reaction and signal transduction. This classification is helpful when we look into a specific biological process. Typically metabolic networks deal with the flow of mass and energy; in gene regulation, process involved in the transforming gene to encoded protein is the essential purpose; while in signaling networks the purpose is the regulation of other processes, and the use of energy and mass flow is a requirement, but not really the point. In general, biopathways are those biological processes taking place in metabolic systems.

2.1.2.1 Metabolic pathways

Metabolic pathways are so far the most intensively studied by biologists and bioinformatists. It is not surprising that metabolic pathways are often misunderstood as the whole set of biochemical reactions that sustain life. But, a metabolic pathway is a subset of these reactions that describes the biochemical conversion of a given reactant to its desired end product. In other words, a metabolic pathway is a special case of a metabolic network with distinct start and end-points, initial and terminal vertices, respectively, and a unique path between them [For99]. Traditionally, metabolic pathways can be interpreted as relational graphs. Typical metabolic pathways are given by the wall chart of Boehringer Mannheim [Mic82] [Mic99] and KEGG [Kan00], which are available via a number of printed and on-line sources (Figure 2.1.2.1A).



(a)



(b)

Figure 2.1.2.1A Typical diagram representation of metabolic pathways (urea cycle). (a) ExPASy, http://tw.expasy.org/cgi-bin/show_image?G8. (b) KEGG, http://www.genome.ad.jp/kegg/pathway/map/map00220.gif

The prevailing definition of a metabolic pathway is a graph (***V***, ***E***), where ***V*** is a finite vertex set, whose elements are called vertices and ***E*** is collection of edges, where an edge is a pair (*u*, *v*) with *u*, *v* in ***V***, that is, the edge is adjacent to *u* and *v* and connects these two vertices. Each vertex represents a metabolite and each edge represents a biochemical reaction that is catalyzed by specific enzyme. In an undirected graph edges are unordered pairs and connect the two vertices in both directions, hence in an undirected graph (*u*, *v*) and (*v*, *u*) are two ways of writing the same edge (Figure 2.1.2.1Ba). In a directed graph, edges are also called arcs, connecting a source vertex to a target vertex. In this case, a directed graph is a pair (***V***, ***A***), where ***V*** is a finite set and ***A*** is a set of ordered pairs of elements in ***V***. ***V*** will be called the set of vertices and ***A*** will be called the set of arcs (Figure 2.1.2.1Bb). All chemical reactions, including enzyme-catalyzed reaction, are to some extent reversible. Within living cells, however, reversibility may not occur, because reaction products are promptly removed by additional enzyme-catalyzed reactions. Metabolite flow in living cells is largely unidirectional. Thus an irreversible directed graph is often used to model a metabolic pathway (Figure 2.1.2.1Bc). A weighted graph is a graph, in which each edge has been assigned a number (usually positive) called its weight (Figure 2.1.2.1Bd).

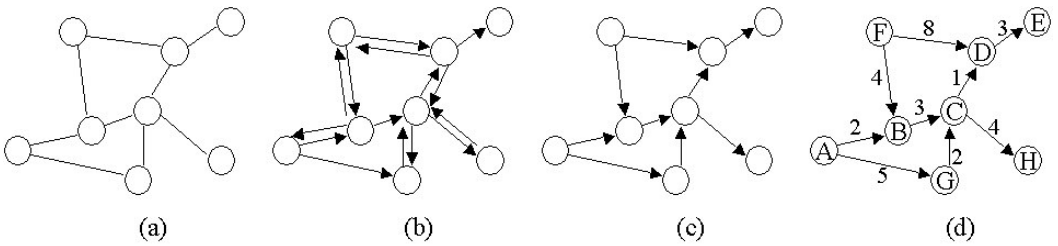


Figure 2.1.2.1B Examples of graphs.

2.1.2.2 Gene regulatory networks

Gene regulatory networks have become a significant field of research for biologists and Bioinformatists. Gene regulatory networks are most often described and interpreted as the on-off switches and rheostats of a cell operating at the gene level. They dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into RNA. Each RNA transcript then functions as the template for synthesis of a specific protein by the process of translation. Process of gene regulatory networks is not restricted to the level of transcription, but also may be carried out at the levels of translation, splicing, posttranslational protein degradation, active membrane transport, and other processes [Ana00]. In addition, such networks often include dynamic feedback loops that provide for further regulation of network architecture and output.

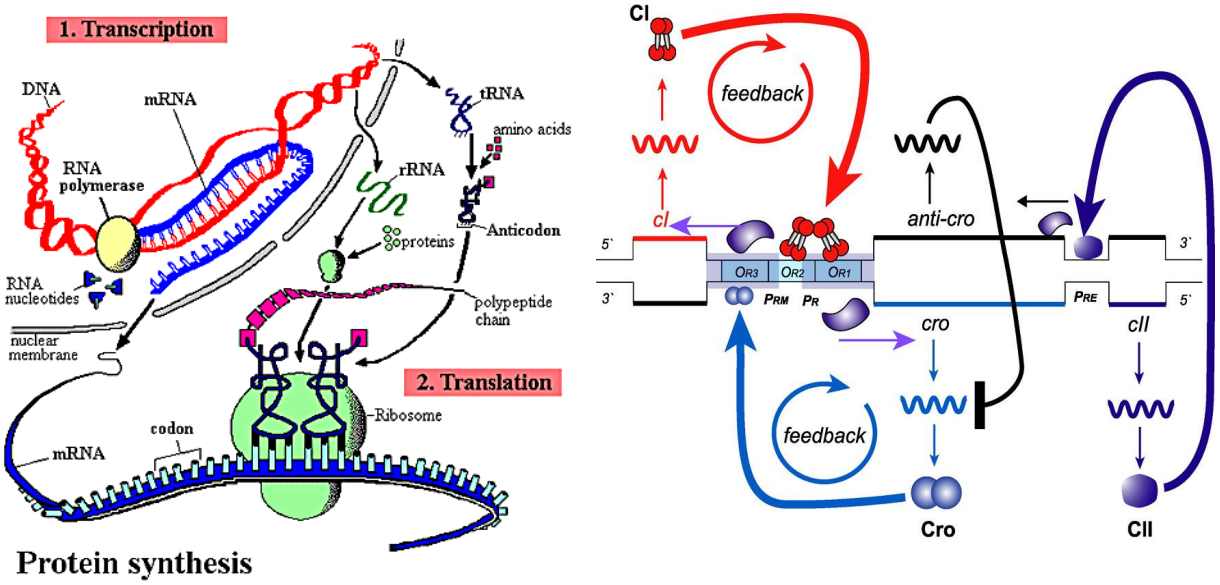


Figure 2.1.2.2 (a) Central dogma (from the City University of New York); (b) A gene regulatory network (from http://doegenomestolife.org/gallery/REGNET.jpg) Transcription of the genes *cro*, *cII* and genes followed by *cII* gene from the promoter *PR* begin, when neither CI protein nor Cro protein does not bind to the operator sites *OR3*, *OR2*, and *OR1*. The genes cro, cII and the genes followed by cII

will be transcripted from the promoter *PR*, when neither CI protein nor Cro protein does not bind to the operator sites *OR3*, *OR2*, and *OR1*. The condition of *E. coli* gives an effect to the concentration of CII protein. If the concentration of CII protein is low, the transcription from *PR* continues and keeps the concentration of Cro protein at some level by the feedback control of the Cro protein itself. On the other hand, if the concentration of CII protein is high, the CII protein binds to the promoter *PRE* as a positive transcription factor, then the transcription from *PRE* begins. Then, anti-sense RNA of the gene cro is produced, which helps to degrade the concentration of Cro protein more rapidly. Transcription of cI gene is followed and concentration of CI protein keeps at some level by the feedback control of the CI protein itself. (http://www.genomicobject.net/member3/GONET/img/lambda_switch.jpg)

As indicated in the schematic presentation (Figure 2.1.2.2a), Genes (DNAs) are transcribed into RNAs by the enzyme RNA polymerases. RNA acts as a go-between from DNA to proteins. Only a single copy of DNA is present, but multiple copies of the same piece of RNA may be present, allowing cells to make huge amounts of proteins. RNA transcripts are subjected to post-transcriptional modification and control: rRNA transcript cut into appropriate size classes and initial assembly in nuclear organizer; tRNA transcript folds into shape; mRNA transcripts are modified, noncoding sequences (introns) removed from interior of transcript; in eukaryotes, all RNA types are transported to the cytoplasm via the nuclear membrane pores. Then mRNA molecules are translated by ribosomes (rRNA + ribosomal proteins) that match the 3-base codons of the mRNA to the 3-base anticodons of the appropriate tRNA molecules. Finally, newly synthesized proteins are often modified after translation (post-translation) before carrying out its function, which may be transporting oxygen, catalyzing reactions or responding to extracellular signals, or even directly or indirectly binding to DNA to perform transcriptional regulation and thus forming a closed feedback loop of gene regulation. Figure 2.1.2.2b shows a gene regulatory network. The interaction among different parts makes cellular regulations extremely complex.

2.1.2.3 Signal transduction pathways

Researchers have known for decades that for cells to grow and function in a complex environment they must communicate with each other. Cell communication, or signal transduction, is simply the means by which cells in the body respond to signals coming from outside those cells. A “biological signal” could be defined as a molecule that acts as a pre-arranged sign, indicating either the commencement and/or the termination of (one or more) intracellular processes. In other words, the nature of the signaling molecule decides it's effects, just as pre-arranged signals have pre-arranged effects [Cla96]. Virtually cell behavior is regulated by a complex network of intracellular and extracellular signal transduction pathways. Signal transduction, in general, is the mechanism by which a signal encountered at a cell's surface (i.e. an extracellular signal) is transformed into an intracellular signal that in turn invokes physiological changes within a cell.

Research on signal transduction or cell communication, at present, is still basic. Important biotechnological advances in recent years have allowed increasingly detailed studies of a variety of signaling pathways. These advances include production of recombinant DNA, the PCR [Alb94], gel electrophoresis [Vin88], microarrays [DeR99], and the serial analysis of gene expression (SAGE) technique [Vel95]. Development of such techniques is ongoing, and large-scale assays of peptides and protein-DNA binding activity are becoming more feasible [Abb02]. It has a wide range of therapeutic possibilities including novel treatments of cancer or other abnormal cell growth.

A simple schematic presentation of signal transduction is shown below.

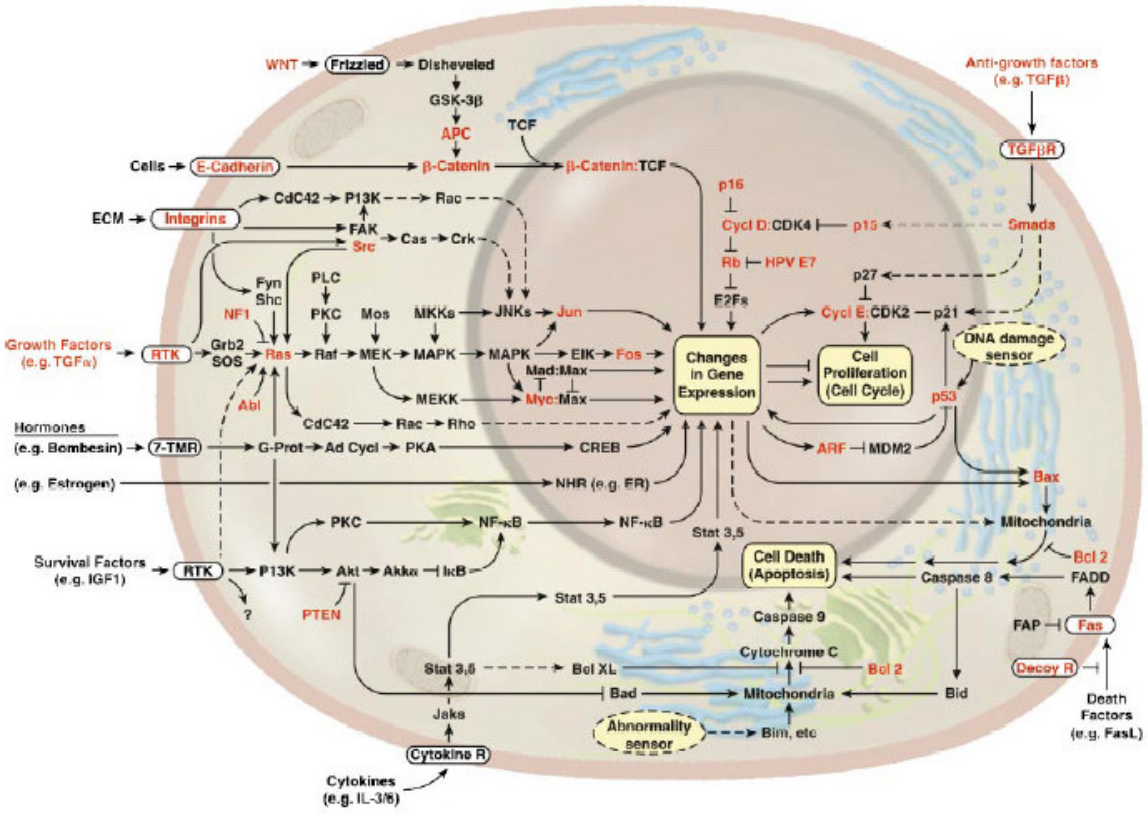


Figure 2.1.2.3 Graphical representation of signal transduction pathways. Cited from the Emergent Integrated Circuit of the Cell [Han00]. Dashed arrows are activating reactions, bar-ended arrows are inhibiting ones; Inhibiting arrows in some cases are shown to act on molecules, in other cases they act on reactions.

The cascade of processes by which an extracellular signal (typically a hormone or neurotransmitter) interacts with a receptor at the cell surface, causing a change in the level of a second messenger (for example calcium or cyclic AMP) and ultimately effects a change in the

cell's functioning (for example, triggering glucose uptake, or initiating cell division). It can also be applied to sensory signal transduction, e.g. of light at photoreceptors [Dow99].

2.2 Molecular Databases and Integration

Modern biology has produced enormous biological data that have been accumulating and systematically stored in specific databases. For the last 10 years *Nucleic Acids Research* (*NAR*) [http://nar.oupjournals.org] has been devoting a special issue to the molecular biology database compilation [Gal04]. Figure 2.2 shows the growth of molecular biological databases. The collection is listed in annual specific database issue of *NAR*.

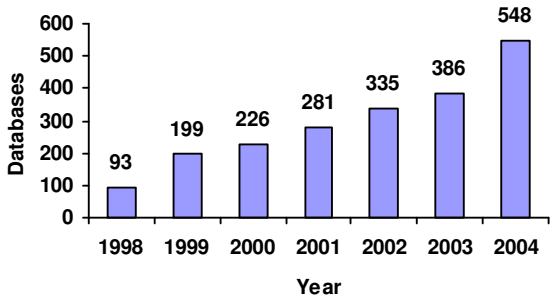


Figure 2.2 Growth of molecular biological databases that are collected by NAR.

The database list and the database descriptions can be accessed online via http://www3.oup.co.uk/nar/database/a. All databases fall into the following categories:

- Nucleotide Sequence Databases,
- RNA sequence databases,
- Protein sequence databases,
- Structure Databases,
- Genomics Databases (non-vertebrate),
- Metabolic and Signaling Pathways,
- Human and other Vertebrate Genomes,
- Human Genes and Diseases,
- Microarray Data and other Gene Expression Databases,
- Proteomics Resources,
- Other Molecular Biology Databases.

All databases included in this *Collection* are freely available to the public. Computational analysis of metabolic pathways based on the information of genes, enzymes and metabolites, which requires access to suitable databases. Table 2.2 lists those major databases that make the integrative information retrieval of metabolic pathways possible. URLs of these databases are appended at the end of the “References” section.

Table 2.2 A list of biological information sources for biopathway analysis accessible via the Internet.

Database	Domain	Access Possibility	Query Interface	Data Format	Identification	Data Update	Access Cost
GenBank	Nucleotide sequences	Local flat files, WWW	WWW	ASN.1 HTML	Unique accession numbers	Daily	Free
EMBL	Nucleotide sequences	Local flat files, WWW	WWW	ASCII, HTML	Unique accession numbers	Daily	Free
KEGG	Metabolism	Local flat files and WWW	WWW	ASCII, HTML	Unique identification numbers	Monthly	Free
TRANSFAC	Transcription factors	Local flat files and WWW	WWW	ASCII, HTML	Unique Identification numbers	Daily for commercial users, 4 times per year for non-commercial users	Commercial
Swiss-Prot	Proteins	Flat files and WWW	WWW	ASCII, HTML	Unique accession numbers	Daily	Free
ENZYME	Enzyme	Local flat files and WWW	WWW	ASCII, HTML	Unique EC numbers	Daily	Free
BRENDA	Enzyme	WWW	WWW	HTML	EC numbers and Organisms	Monthly	Free for non-commercial users
WIT/EMP	Enzyme	WWW and PostgreSQL Client	WWW	HTML, SQL	Unique identification numbers	Unknown	Free
OMIM	Metabolic diseases	WWW	WWW	HTML	Unique MIM numbers	Unknown	Free
GeneCards	Human genes	WWW	WWW	HTML	Unique GC numbers	Unknown	Free for academic or non-profit institutions
GeneNet	Gene networks	WWW	WWW (Java applet)	HTML	Unique identification number	Unknown	Free
Klotho	Compounds	WWW	WWW	HTML	Compounds	Unknown	Free
BioCyc	Pathway/Genome	Local flat files WWW	WWW	HTML, Attribute-Value	Pathways and Organisms	Unknown	Free

The presence of numerous informational and programming resources on gene networks, metabolic processes, gene expression regulation, etc., described above, raises an acute problem of data integration and suitable access. The idea of data integration in molecular biology is not a new one. There are several previous and underlying projects that focus on the challenging problem of interoperability among biological databases. P. Karp first addressed the biological database integration in early nineties [Kar95]. At the same time the requirements for these integration approaches were formulated [Dav95]. Many integration approaches for molecular biological data sources are currently available. These systems are based on different data integration techniques, e.g. federated database systems (ISYS [Sie01] and DiscoveryLink [Haa01]), multi database systems (TAMBIS [Ste00]) and data warehouses (SRS [Etz96] and Entrez [Tat99]).

Different approaches have different advantages and disadvantages [Fre02]. ISYS provides a dynamic and flexible platform for integration of molecular biological data sources. This system is developed as a Java application and must be installed on a local computer. One main feature is the global view onto the integrated data sources with the help of a global scheme. DiscoveryLink system is based on federated database techniques. A federated system requires the development of a global scheme. Thereby, the degree of integration must be rated as tight. DiscoveryLink accesses its original data sources through views. Read-only SQL is supported as query language. TAMBIS integration system is based on multi-database techniques. It is used through a Java applet. Due to the use of a multi-database query language, it is not necessary to built an integrated global scheme. But the interfaces and the number of input formats are disadvantageous. SRS is based on local copies of each integrated data source. SRS runs on a web-server and is accessible via any web-browser. An HTML-interface for data queries is provided. Various output formats are possible (HTML or ACSII-text). One problem with the result presentation in SRS is the necessity to parse the outputs for a further computer-based processing. The absence of any scheme integration is also disadvantageous for the use of the SRS system. Similar to SRS is the Entrez system. This system integrates only data sources of NCBI. HTML is the only interface provided. Another Entrez feature is the manual construction of special URLs. Various output formats prove to be useful. These include HTML or ASCII-text, as well as XML and ASN.1 files. The biggest disadvantage of the Entrez approach is the restricted number of integrated data sources (only NCBI internal data sources).

Although these integration systems are available to realize the data query process, the process still requires much human intervention and the quality of annotation largely depends on the knowledge and skills of human experts. Moreover, scientists have to invest extensive efforts to learn how to use all different database interfaces, query languages, and parameter specifications for specific analytical programs. On the other side, biologists wish to perform

metabolic pathway analysis with easy-to-use local or Internet-based tools with friendly user interfaces. Simple text mining approaches such as a web-based biological information retrieval and integration system could be one solution.

2.3 Modeling and Simulation of Metabolic Networks

The vast complexity of biological systems requires modeling to design and interpret biological experiments. Computer based models of genes and metabolic networks are the first step to a complete understanding of the cell. We are at the point where the level of computer technology and biological knowledge are sufficient to experiment with different approaches to model cellular metabolism.

Before designing a model, careful consideration must be given to the questions being asked, and the nature of the biological system. For example, some types of models require a larger quantity of data, or more accurate data, than others. In other words, a model should be able to raise additional questions, giving directions to experimental work.

2.3.1 Model Classification

A model is built in order to capture the nature of objects. Models can be divided into many different types. Not all scientific models are precise, numerical, or quantitative. Neelamkavil [Nee87] classified models into physical, symbolic and mental ones. To model biological systems four forms were introduced [Hae96]:

1. Conceptual or verbal - descriptions in a natural language.
2. Diagrammatic - graphical representations of the objects and relations (e.g., physiological diagrams of metabolic pathways such as the Krebs cycle).
3. Physical - a real, physical mock-up of a real system or object (a "tinker-toy" model of DNA or 3D structure of protein).
4. Formal - mathematical (usually using algebraic or differential equations).

Our primary interest here will be in diagrammatic and mathematical.

For many reasons mathematical models are the most important and most widely used category of models. They are concise, unambiguous and uniquely interpretable, while their manipulation and the evaluation of alternatives are relatively inexpensive [Mat92]. To show the scope of the range of mathematical models that are potentially applicable to biological systems, a simple classification of mathematical models is illustrated in Table. 2.3.1.

Table 2.3.1 A classification of mathematical models.

Criteria of classification	Yes	No
Having an explicit representation of mechanistic processes?	Process-oriented or mechanistic models: (e.g., mathematical equations)	Descriptive or phenomenological models: (e.g., graphics, rule based systems)
Having an explicit representation of future system states or conditions?	Dynamic models: include the transient as well as the steady state behavior of a system.	Static models: (e.g., linear regression equation relating variables x and y) given for the steady state only, are described with algebraic equations.
Representing time continuously?	Continuous time models, time may take on any values: described with differential equations	Discrete time models, time is an integer (time invariant) only: those where the shapes of their outputs are independent of the moment of onset of their inputs or disturbances.
Having an explicit representation of space?	Spatially heterogeneous models (e.g. objects have a position in space, or occupy a finite region of space). A. Discrete: space is represented as cells or blocks, and each cell is represented as spatially homogeneous. B. Continuous: every point in space is different (e.g. diffusion equations)	Spatially homogeneous models: (e.g., simple equations of enzyme kinetics) in many cases only one most important spatial coordinate is taken into account.
Allowing random events?	Stochastic models: the relations between variables are given in terms of statistical values.	Deterministic models: are those in which the probability of events does not feature.

The basis of the classification is whether the mathematics incorporates (or not) a particular mathematical structure. In some cases it is subjective whether the mathematics has its characteristics or not.

Historically, mathematical modeling in biology has been of only minor importance. Unlike physics or chemistry, biology has not discovered underlying principles, such as an equivalent of Newton's Laws to build upon. The unit of life, the cell, is an enormously complicated structure and the behavior and properties of living things are not easily reduced to equations.

2.3.2 Modeling Metabolism

The complexity of biological systems is in part due to the large number of interactions among components at different levels of the organizational hierarchies. A mathematical modeler must combine qualitative knowledge of relationships and desired quantitative data to construct a model of the system. With numerous data on the structure and processes, it is possible to construct mathematical computer models that allow the formalization of the knowledge on complex metabolic systems.

Current applications of metabolic pathways modeling include [Bow01]:

- Finding pathways of maximum yield, for example in the area of biotechnology, where foreign genes are spliced into a host genome to mass-produce a desired molecule.
- Finding non-redundant pathways, important in drug design.
- Testing whether a set of enzymes can produce a desired product.
- Genome comparisons, by aligning metabolic pathways, missing genes can be identified and new pathways identified.
- Detecting the medical significance of enzyme deficiencies.

To model metabolism requires the concept of a state. A state is a snapshot of the system in time, and with the knowledge of one state, the future state can be calculated. Depending on the kind of model used, the state is represented in different ways. Two broad categories of modeling and simulation exist: deterministic modeling and analytic simulation based on differential equations, and stochastic modeling and discrete event simulation.

An analytic simulation uses mathematical analysis to represent the temporal behaviors of components, often in closed form. Analytic simulations capture aggregate system behavior by modeling small and relatively similar entities. A discrete-event (discrete-state) simulation is used when the system's overall behavior is not understood well enough to permit formal mathematical analysis.

For example, the analytic approach to metabolic simulation typically requires the determination of steady-state rate equations for constituent reactions, followed by numerical integration of a set of differential equations describing fluxes in the metabolism. The feasibility of the analytic approach is however limited by the extent to which the metabolic processes of interest have been characterized. For most metabolic pathways, either we are unaware of all the steps involved, or we lack rate constants for each step. This lack of information precludes the use of the mathematical approach in describing the process. Even when reaction rates are known, differential equations incur great computational costs. Analytic representations, such as differential equations, lack the robustness required to handle partial and uncertain knowledge. In addition, because analytic simulations model relatively similar structures over relatively similar temporal intervals, interleave simulations are highly constrained.

The discrete-event approach to simulation, on the other hand, can use all available data, both quantitative and qualitative, and can even incorporate analytic methods where applicable; semi quantitative models, which couple symbolic and numeric computing techniques, have been developed for a number of domains, including the human cardiovascular system [Sir96] and gene regulation in bacteria [Bru92]. Most importantly, discrete-event simulations provide natural support for qualitative representation and reasoning techniques, which offer explicit treatment of causality. The discrete-event approach can provide declarative representations for both the structures in the domain and the processes that act on these structures.

1. Structural knowledge

Structural knowledge of a system is the foundation of a simulation. Most analytic and discrete-event simulations employ state-variable representations of physical entities. State variables describe the relevant qualitative or quantitative attributes of the system, but the structure of the system is expressed in terms of mathematical relationships among the state variables. For example, enzyme and substrate concentrations are state variables in a simulation of Michaelis-Menten enzyme kinetics.

2. Process knowledge

Structural knowledge alone captures the state of a system at a fixed point in a time-independent way, but it does not capture the relationships and interactions among structural components over time. Process knowledge is functional knowledge of dynamic change. A declarative process representation is critical to the success of a simulation. Process knowledge can be represented declaratively in several forms. A rule-based representation specifies the preconditions for change and the effects of the change in a unit known as a rule. For example, the effect of tetracycline on the mechanism of protein synthesis can be expressed in the following form:

(IF tetracycline is present

THEN tetracycline will inhibit the binding of aminoacyl-tRNA to ribosome)

Rules are the predominant declarative representation of processes. Processes can also be represented with constraints. For example, a chemical reaction can be represented as a set of reactants, a set of products, and a set of stoichiometric constraints.

3. Declarative device models

A declarative device model allows different computational agents to reason about the model by accessing its structural and functional components. We should acknowledge that biological devices are less well understood than manufactured devices; consequently, biological simulations often yield highly uncertain results. A goal of simulation researchers is to develop robust methods for quantifying the uncertainty in device models and in simulation predictions.

In a differential equation based model, the concentrations of enzymes and substrates is the state. By making assumptions one can transform almost any set of biochemical reactions into a system of ordinary, non-linear differential equations (ODEs). The equations specify reaction rates between molecules. If the number of states becomes too large, then coarser approximations can be used. The equations can be solved numerically, and the trajectory of the state can be analyzed for dependence on initial parameters. However, not all systems can be meaningfully modeled by differential equations. One also has to make several assumptions: that the solution is well mixed, that the number of molecules is sufficiently high, that discrete changes of a single molecule can be approximated as a change in the concentration, and that fluctuations around the mean are small compared to the mean itself. For systems consisting of small number of molecules a stochastic framework is a more realistic choice.

Most stochastic methods consider the exact number of molecules. The state indicates exactly how many molecules of each type are present in the system. Even though the state changes discretely, how and when it changes is probabilistic. For example, a simple chemical equation $X \rightarrow Y$, i.e. a molecule of X turns into Y, is governed by a probability. This probability, multiplied by the time step, is the chance of this molecule changing over the specified time. Because the outcome is probabilistic, it is possible to get different successor states. A method to deal with this is to use a Monte Carlo simulation, where a series of random numbers are generated and decide the next state. For a given set of numbers, the next state is deterministic, however new random numbers are used for each new state calculation. There are efficient algorithms for Monte Carlo calculations [Gil77] [Gib00].

Several variations on the stochastic simulation exist, such as Petri nets, which have an intuitive analogy to biological systems [Red93] (see Section 2.3.4 and Chapter 3).

2.3.3 Related Simulation Environments

Many attempts have been made to simulate molecular processes in both cellular and viral systems. Several software packages for quantitative simulation of biochemical metabolic pathways, based on numerical integration of rate equations, have been developed. A list of biological simulators can found at http://www.techfak.uni-bielefeld.de/~mchen/BioSim/BioSim.xml. Table 2.3.3 shows a comparison of the most well-known metabolic simulation systems.

Table 2.3.3 A comparison of metabolic simulators.

Tools	**Gepasi**[a]	**Jarnac**[b]	**DBsolve**[c]	**E-Cell**[d]	**VON++/GON**[e]
Stoichiometry matrix presentation	+	+	+	+	-
Core algorithm and method	MCA	MCA	MCA	SRM, MCA	Petri net
Pathway DB retrievable	-	-	WIT/MPW, EMP	KEGG, EcoCyc	KEGG
Pathways graphic editor	-	++++	++++	-	+++++
Kinetic types	++++	+++	+++	++	++++
Virtual cell model	-	-	-	+	+
Simulation graphic display	++++	+++	++	++	+++
Mathematical model accessible and modifiable	+	+	+	+	+
Data XML export	SBML	SBML	SBML	SBML	Biopathway XML
User interface	++++	+++	++++	+++	++++
Programming language	C++	Delphi 5	C++	C++	Delphi /Java

a. Gepasi [http://www.gepasi.org/]

b. Jarnac [http://members.lycos.co.uk/sauro/biotech.htm]

c. Dbsolve [http://homepage.ntlworld.com/igor.goryanin/]

d. E-Cell [http://www.e-cell.org/]

e. VON++ [http://www.systemtechnik.tu-ilmenau.de/~drath/visual.htm] is further developed to GON, later Cell Illustrator^{tm} [http://www.gene-networks.com/ci/]

SBML (Systems Biology Markup Language) [http://www.cds.caltech.edu/erato/] is a description language for simulations in systems biology. It is oriented towards representing biochemical networks that are common in research on a number of topics, including cell signaling pathways, metabolic pathways, biochemical reactions, gene regulation, and many others. SBML is the product of close collaboration between the teams developing BioSpice [http://biospice.lbl.gov/], Gepasi, DBSolve, E-Cell, Jarnac, StochSim [http://www.zoo.cam.ac.uk/comp-cell/StochSim.html] and Virtual Cell [http://www.nrcam.uchc.edu/].

A plus symbol "+" indicates a feature which has been implementated/enhanced in the tool. A minus symbol "-" indicates no such feature available in the tool.

Each tool possess some prominent features which others have only a little or not at all. After a decade's development, Gepasi is widely applied both for research and education purposes to simulate the dynamics and steady state of biochemical systems due to its powerful simulation engine and user-friendly interface. Jarnac, as a replacement of SCAMP, has a nice pathway

graphic editor, called Jdesigner, which enable users to draw interactively a biochemical network and export the network in an XML format. Dbsolve is good at model analysis and optimization. Dbsolve uses numerical procedures for integration of ODEs or NAEs (non-linear algebra equations) to describe the dynamics of these models and offers explicit solver, implicit solver and bifurcation analyzer. The primary focus of E-Cell is to develop a framework for constructing simulatable cell models based on gene sets derived from completed genomes. Contrast to other computer models that are being developed to reproduce individual cellular processes in detail, E-Cell is designed to paint a broad-brush picture of the cell as a whole. There is another program, DynaFit [http://www.biokin.com/dynafit/], which is also useful in the analysis of complex reaction mechanism for which traditional (algebraic) kinetic equations cannot be derived.

In predicting cell behavior, the simulation of a single or a few interconnected pathways can be useful when the pathways being studied are relatively isolated from other biochemical processes. However, in reality, even the simplest and most well studied pathways, such as glycolysis, can exhibit complex behavior due to connectivity. In fact, the more interconnections exist between different parts of a system, the harder it gets to predict how the system will react. Moreover, simulations of metabolic pathways alone cannot account for the longer time-scale effects of processes such as gene regulation, cell division cycle and signal transduction. When systems reach a certain size they will become unmanageable and hard to understand without decomposition into modules (hierarchical models) or presentation of graphs. In this sense the tools mentioned above appear week. In comparison, Petri nets capture the basic aspects of concurrent systems of metabolism both conceptually and mathematically. The major advantages of Petri nets comprise graphical modeling representation with sound mathematical background, which make it possible to analyze and validate the qualitative and quantitative behavior of a Petri net system. Petri nets also provide the ability for clear description of concurrency and long experience in both specification and analysis of parallel systems and the ability to describe a Petri net model on different levels of abstraction (hierarchical models). In addition, the development of computer technology enables Petri net tools to have more friendly interfaces and the possibility of standard data import/export supporting. We are motivated to exploit Petri net methodology to model and simulate gene regulated metabolic networks.

2.3.4 Petri Net Modeling and Simulation

Since 1960's Petri net was first introduced and formally defined by Prof. Dr. Carl Adam Petri [Pet62], Petri net and its concepts have been extended and developed. Both the theory and the applications of this model have been flourishing. The properties, concepts, and techniques of Petri nets are being developed in a search for natural, simple, and powerful methods for

describing and analyzing the flow of information and control in systems, particularly systems that may exhibit asynchronous and concurrent activities. The major use of Petri nets has been the modeling of systems of events in which it is possible for some events to occur concurrently but there are constraints on the concurrence, precedence, or frequency of these occurrences.

Petri nets are conceptually simple: they consist of places, transitions and arcs. Each place has a non-negative number of tokens. A transition is enabled if the number of tokens exceeds the weights of the arcs connecting the places. For metabolic pathways, places could represent biomolecules and transitions could represent the individual reactions. Arc weights represent the proportion of a reaction during each discrete step. A definition for the ordinary Petri net is given in the following [Rei82] [Dav92]:

Definition 2.1 *An ordinary Petri net is a 3-tuple, PN=(P,T;F) with*

$P = \{p_1,p_2,\ldots,p_m\}$ *is a non-empty, finite set of places, drawn as circles;*

$T = \{t_1,t_2,\ldots,t_n\}$ *is a non-empty, finite set of transitions, drawn as bars;*

$P \cap T = \emptyset$ *and* $P \cup T \neq \emptyset$*;*

$F \subseteq (P \times T) \cup (T \times P)$ *is a non-empty, finite set of arcs, connecting places to transitions or transitions to places but never two places or two transitions.*

The ordinary Petri net given in Definition 2.1 contains only structural elements. To define dynamic Petri nets and their firing rules, we need some terminology to identify special sets of places and transitions and the concept of markings.

Definition 2.2 *Pre- and Post-Sets*

The pre-set ${}^\circ t_i$ *of a transition* $t_i \in T$ *contains all places that are connected to* t_i *via a directed arc from the place to the transition:* ${}^\circ t_i = \{p \in P: (p, t_i) \in F\}$. *The elements of* ${}^\circ t_i$ *are often called input places.*

The post-set t_i° *of a transition* $t_i \in T$ *contains all places that are connected to* t_i *via a directed arc from the transition to the place:* $t_i^\circ = \{ p \in P: (t_i, p) \in F\}$. *The elements of* t_i° *are often called output places.*

The pre-set p_i *and post-set* p_i *of a place* $p_i \in P$ *are defined in the same way*:

${}^\circ p_i = \{t \in T: (t, p_i) \in \mathrm{F}\}$

$p_i^\circ = \{t \in T: (p_i, t) \in \mathrm{F}\}$

Definition 2.3 *Marking*

A marking of a Petri net is a mapping $M: P \rightarrow \mathbf{N}$ *that assigns a finite non-negative integer number of tokens to each place of the ordinary Petri net.* $M_0: P \rightarrow \mathbf{N}$ *is the initial marking.*

Definition 2.4 *Enabled transition*

Let PN be a Petri net with marking M, M(p) be the number of tokens contained in p ∈ P and t ∈ T be a transition. The transition t is enabled if and only if p ∈ P:M (p) ≥ Pre(p, t).

The dynamic behavior of a Petri net is expressed by changing markings. A marking changes when a transition fires, and a transition may fire when it is enabled.

Definition 2.5 *Firing a transition*

Let PN be a Petri net with marking M, M(p) be the number of tokens contained in p ∈ P and t ∈ T be an enabled transition. Firing the transition t results in a new marking M', written as M|t > M', given by M'(p_i) = M(p_i) - Pre(p_i, t) + Post(p_i, t). Hence Pre(p_i, t) denotes the number of tokens needed in p_i for the firing of transition t and Post(p_i, t) denotes the number of tokens added to place p_i when transition t has fired.

State changes are carried out by firing enabled transitions. In an ordinary Petri net a transition is enabled when all its input places have at least one token. When an enabled transition *t* is fired, a token is removed from each input place of *t* and a token is added to each output place of *t*. This gives a new state.

Graph theory has been exploited in metabolic process modeling [Koh83]. In contrast to naive graph, Petri net is a graph oriented design, specification, simulation and verification language. It offers a formal way to represent the structure of a discrete and/or event system, simulate its behavior, and draw certain types of general conclusions on the properties of the system. Because of their good properties in theoretical analysis, practical modeling, and graphical visualization of concurrent systems, Petri nets especially high-level Petri nets are widely used in work-flows, flexible manufacturing, operations research, railway networks, defense systems, telecommunications, Internet, commerce and trading, and even biological systems. The Petri net world web site has been set up at http://www.daimi.au.dk/PetriNets/, where a large amount of investigations on Petri nets have been compiled in the literature, and various applications have chosen Petri nets as their control models due to the intuitively understandable graphical notation of Petri nets.

The application of Petri nets for the simulation of biochemical reactions was firstly formally introduced by Reddy et al. [Red93]. Nevertheless, the model used was a qualitative one. Ordinary Petri nets models do not have such functions as quantitative aspects, so there are some extension of Petri nets that can support dynamic change, task migration, superimposition of various levels of activities and the notion of mode of operations. Various extensions of Petri nets, such as (Stochastic) Timed PNs [Wan98], Colored PNs [Jen97], Predicate/Transition Nets [Gen87] and Hybrid PN [Dav92], allow for qualitative and/or quantitative analysis of resource utilization, effect of failures, and throughput rate. Using suitable Petri nets, we can extend Petri nets to support flexible modeling of kinetic effects of biochemical reactions [Hof98]. The desirable Petri nets should allow the modeling of biochemical processes using

actual concentrations. It should make sense to model this biocatalytic reaction using functions, which allow each transition to simulate kinetic effects. Moreover, complex relations and conditions can be combined which will activate transitions.

Following its early application of modeling metabolic pathways [Red93] [Hof94], Petri nets as a new tool and terms of modeling and simulating biological information system are investigated more and more. Later in 1996 [Red96], an example of the combined glycolytic and pentose phosphate pathway of the erythrocyte cell was presented to illustrate the concepts of the methodology. However, the reactions and other biological processes were modeled as discrete events and not possible to simulate the kinetic effect. Hofestädt [Hof98] investigated a formalization showing that different classes of conditions can be interpreted as gene, proteins, or enzymes and cell communication; and also presented the formalization of self-modified Petri nets, which allows the quantitative modeling of regulatory biochemical networks. Chen [Che00] introduced the usage of hybrid Petri nets (HPNs) for expressing glycolysis metabolic pathways. Using this approach, the quantitative modeling of metabolic networks is also possible. Koch I. et al. [Koc99] extended the model proposed by Reddy by taking into account reversible reactions and time dependencies. Kueffener [Kue00] exploited the knowledge available in current metabolic databases for the functional predictions and the interpretation of expression data on the level of complete genomes, described the compilation of BRENDA, ENZYME and KEGG into individual Petri nets and unified Petri nets. Goss [Gos99] and Matsuno [Mat00] applied Petri nets to model gene regulatory networks by using stochastic Petri nets (SPNs) and HPNs respectively. In the DFG workshop "Modeling and Simulation Metabolic Network" 2000 participants also discussed the applications and perspective of Petri nets [Hof00]. Genrich et al. [Gen01] discussed executable Petri net models for the analysis of metabolic pathways. Heiner et al. [Hei01] studied the analysis and simulation of steady states in metabolic pathways with Petri nets. R. Srivastava et al. [Sri01] also exploited a SPN model to simulate the σ32 stress circuit in *E. coli*. Oliveira J.S. et al. [Oli01] developed the mathematical machinery for the construction of an algebraic-combinatorial model to construct an oriented matroid representation of biochemical pathways. Recently a special issue on “Petri nets for metabolic network” appeared at http://www.bioinfo.de/isb/toc_vol_03.html.

Table 2.3.4 presents a summary of Petri net tools that were used to model biological systems. Most publications present their models only based on a general Petri net tool utilization. These publications are not listed in the table. More Petri net tools can be found at http://www.daimi.au.dk/PetriNets/tools/quick.html. The intuitively understandable graphical notation and the representation of multiple independent dynamic entities within a system makes Petri nets the model of choice since it is highly suitable for modeling and simulation of metabolic networks.

Table 2.3.4 Summary of Petri net that were used for modeling and simulation of biological systems.

Petri nets type	Petri net tool	Tool brief description	Application	Reference
High level	Stella	The STELLA software is based on a feedback control framework. The basic self-regulatory, or homeostatic, mechanisms that govern the way living systems operate, are reinforced by the way the software itself operates, it enables users to make their hypotheses explicit using simple iconic building blocks, and then to test these hypotheses via simulation.	Modeling dynamic biological systems, especially ecological system.	[Rut97] [Ela95]
Hybrid	VON++	Visual Object Net ++ is an innovative Petri net CAE Tool for PC that supports mixed continuous and discrete event Petri nets. Beside the new continuous net elements, the whole well tried concept of the traditional Petri nets is available. The goal of Visual Object Net ++ is to study the behavior and characteristics of a class of hybrid Petri nets.	Gene regulatory; Metabolic pathways; Bioprocess	[Doi99] [Mat00] [Che00] [Che02a] [Mat01]
Stochastic	UltraSAN	UltraSAN employs stochastic activity networks (SANs), a variation of Petri nets, to model and analyze the performance and dependability of software, hardware and network system designs. UltraSAN provides analytic solvers as well as discrete-event simulators.	Protein synthesis from mRNA; Plasmid Replication; Prion Propagation	[Gos98] [Gos99] [Sri01]
Hierarchical	PED	PED supports basically the construction of hierarchical place/transition nets with the specification of different types of places, transitions, and arcs, including their marking.	Pentose phosphate pathway	[Koc99]
High level	THORNs	THORNs is a general-purpose, graphical, discrete-event simulation tool based on a special class of high-level Petri Nets called Timed Hierarchical Object-Related Nets. THORNs allows the specification of individual tokens, they provide delay times and firing durations for transitions, and THORN models can be hierarchically structured with respect to transition refinement and subnet invocation.	Ecological system	[Gro97] [Gro98]
High level	Design/CPN	Design/CPN supports CPN models with complex data types (color sets) and complex data manipulations (arc expressions and guards). The functional programming language Standard ML enable the software package support hierarchical CP-nets and generate a model from the data extracted from databases.	Glycolysis	[Vos00] [Kue00] [Gen01]
Functional	GON/Cell Illustrator	Genomic Object Net is an environment for simulating and representing biological systems.	Biopathways; Cell development	[Mat03a] [Nag03]

Among these Petri net tools, VON++ is more suitable for biopathway modeling and simulation under the following considerations:

1. VON++ is a small, quick, uncomplicated and intuitive Petri net tool that supports both discrete event Petri nets and timed event/condition Petri nets;
2. VON++ has a user-friendly graphical interface, which consists of object oriented Integrated Developing Environment, file management tool; it also represents a class hierarchy in the Factory window;
3. There are animation features to make it easier to observe the dynamic behavior of the nets.

So its object oriented user interface allows the easy design, simulation, visualization and documentation of hybrid Petri nets. However, VON++ does not support ASCII file (text format) import, but its text format file export ability is available in VON2.6 version. VON++ has further developed to GON. GON has been commercialized under the name of Cell Illustrator[tm]. Both GON and Cell Illustrator are developed in Java and support XML import/export.

2.4 In Silico Prediction of Metabolic Pathways

With the achievement of biological data collection and the development of useful biological tools, metabolic pathway prediction becomes possible. The *in silico* prediction of metabolic pathways is already an essential tool for the functional assignment of predicted genes, for which almost no data exist by biochemical experiments [Mus96] [Dan97]. It is also essential to do research from genotype to phenotype.

2.4.1 Existing Resources

In the past decade hundreds of biological databases have been set up. Among them several have became indispensable resources for the development of metabolic databases Such databases typically describe collections of enzymes, reactions and biochemical pathways and are used in conjunction with software that allows querying and visualizing metabolic information [Kar98]. They are used in various contexts and have gained recognition in the context of functional genome annotation and metabolic pathway reconstruction [Gal98] [Bon98].

One approach is being achieved by KEGG [Kan02], which contains an EC numbering scheme for enzymatic functions that integrates different gene names in different organisms. Under the KEGG project, all known metabolic pathways are computerized as graphical diagrams. The LIGAND database [Got98] has been organized to fill in the gap between genomic information and chemical information, and applied to actual reconstruction of metabolic pathways in the completely sequenced organisms in the KEGG [Kan97]. If the set

of open reading frames (ORFs) is complete for an organism, the organism-specific pathways should be reconstructed, which can be visualized by marking the assigned enzymes on diagrams [Oga98].

WIT/EMP [Ove00] also provides a pathway retrieval web-interface, which can be queried with initial substrate, coenzyme, enzyme, intermediate, and end product. WIT is a system designed and implemented to support the curation of functional assignments and metabolic models for sequenced genomes. It generates metabolic reconstructions based on chromosomal sequences and metabolic modules from the EMP/MPW family of databases.

EcoCyc/MetaCyc [Kar02] is a comprehensive reconstruction of metabolic pathways of organisms for which the complete genomes are known. It became a commercial database one year ago, but it is available to academic users for free.

However, the existing metabolic pathway databases have a number of limitations in metabolic pathway reconstruction. They do not contain comprehensive information about metabolic pathways, such as physical and chemical properties of the enzymes that are involved. None of these databases provides methods for solving the whole complex of tasks necessary for a gene network effective study, which demands analysis of the large bulk of heterogeneous experimental data. Some collect information only about metabolism of single organism and/or attain only special pathways. Moreover, metabolic pathways may not easily be reconstructed by simple collection of enzymatic reactions, thus assigned solely on sequence similarity. It often finds missing enzymes and leads to an incomplete set of metabolic pathways.

2.4.2 Reconstruction Algorithms

The existing resources for metabolic pathway reconstruction use a variety of methods to predict which enzymes are present in an organism and hence which pathways may be inferred.

KEGG presents a method that utilizes higher-level information of molecular pathways to reconstruct a complete functional unit from a set of genes. Specifically, a genome-by-genome comparison is first made for identifying enzyme genes and assigning EC numbers, which is followed by the reconstruction of selected portions of the metabolic pathways by use of the reference biochemical knowledge. Then the KEGG's pathway diagram is utilized as a reference for the functional metabolic pathway reconstruction [Bon98].

The WIT reconstruction starts with an organism's whole genomic DNA sequence [Ove99]. First, a program called CRITICA [Bad99] searches the genome for ORFs using a combination of comparative and non-comparative methods. Then a "bi-directional best hit" approach is used to assign a function to each of these predicted genes. WIT also defines a pathway as a set of reactions that have been observed to form a metabolic unit in some organisms. Each pathway is evaluated for the new genome on the basis of the proportion of its

enzymes identified by the above procedure. A prototype called PUMA2 [Dso99] has been developed to provide a framework for the automated reconstruction of the metabolism of microbial consortia and individual species, and to be able to comparatively analyze the metabolic subsystems in different organisms.

The bacterial reconstructions derived from EcoCyc are combined by a program called PathoLogic [Pal02]. The basic PathoLogic algorithm asserts the presence of an EcoCyc pathway in the new genome if at least one of its enzymes has been identified. It requires a fully annotated genome as input, where the function of each gene has been assigned manually. PathoLogic then uses a text-based approach to link the annotated genes with enzymes in EcoCyc. Functions are matched either by EC number or by enzyme name. Once the gene-enzyme matching is complete, the pathways in EcoCyc are evaluated with respect to the enzymes found in the annotated genome.

Aside of approaches of genome sequence comparison [Mus96] [Bon98], genome annotation data parsing [Geo02], annotated whole genome sequence assembly [Gaa95] [Ove97] [Sel97] [Nak99] [Cov01] and enzyme assignment [van00], Arvind K.Bansal [Ban00] describes a framework of automated reconstruction of metabolic pathways using the information about orthologous and homologous gene groups archived in the GenBank. Ma H.-W. et al. [Ma03a] conducts further analysis of their global structure for various organisms. David Allen [All01] presents a reconstruction method by the exploration of gene expression data with factor analysis. Factor Analysis is shown to identify and group genes according to membership within independent metabolic pathways for steady state microarray gene expression data. F. Boyer et al. [Boy03] proposes a new formulation for the problem of *ab initio* metabolic pathway reconstruction. They use the similar idea of Arita M.'s [Ari00] to consider chemical compounds as sets of individual atoms and reactions as transfers (partial injections) of atoms between compounds. Given a source and sink compound, the reconstruction problem consists in finding all the successions of reactions that result in a minimum number of transferred atoms from the source to the sink.

Moreover, several software tools have been developed to assistant reconstruction of pathways. For instance, PathoLogic [Pal02] is used by Sophia T. et al. [Sop03] and PathMiner by McShan et al. [McS03].

However, these approaches of predicting each gene function based on sequence similarity searches often fail to reconstruct cellular functions with all the necessary components. Knowledge of the genome sequence alone is really only the start of the work. The future of metabolic pathway analysis may depend greatly upon its ability to capitalize on the wealth of genetic and biochemical information currently being generated from the fields of genomics and proteomics. The challenge is how to automate and simplify the process of information retrieval and integration in order to turn this growing deluge of data into

knowledge. Under these considerations, we are motivated to combine sequence information with information about the underlying biochemical reactions. We attempted to present a new framework of metabolic pathway reconstruction and developed a system to integrate the access to different biological databases, and determine metabolic pathways in the cell.

The ideal system for metabolic pathway reconstruction would at least include a web-based architecture to allow remote and local access to the different biological databases. It would offer a proven approach that can perform complex queries, data transformations, and data integration in one powerful biological tool, without requiring extensive programming. An automated primary and secondary database update and report system would enable the internal data to remain consistent, accurate and reliable, with the ability to incorporate information flowing from experimental validation, such as gene expression, enzyme catalyzation, protein interaction and pathways. An essential feature would include a quality assurance process, to allow quick distribute queries and retrieve primary results. In light of these desirable features, we have designed a prototype system which has a single common data representation to handle the diverse range of rudimentary data, such as enzymes, proteins, metabolites as well as incomplete or fragments of gene sequences of metabolic pathways.

2.5 Analysis and Alignment of Biopathways

Beyond the scope of modeling, analysis of metabolic pathways has received an increasing amount of attention over the past few years. Progress has been made in many aspects such as the metabolic control analysis, stoichiometric analysis and comparative analysis.

2.5.1 Functional Analysis

The functional analysis of metabolic systems based on the information of genes, signals, enzymes, etc. will undoubtedly have an impact on our views of metabolism from its capabilities to its regulation and potentially also on its evolution.

Fell D.A. et al. [Fel00a] studied the structural characteristics of the metabolic network that is stable and in operation yet evolvable. Relevant characteristics are the number and size of the modules of metabolism and the number of interconnections between them. As an example, the analysis of *E. coli* metabolism may reveal aspects of the evolution of metabolism. In [Wag01] a graph theoretical analysis of the *E. coli* metabolic network was done and found that this network is a small-world graph. Moreover, the connectivity of the metabolites follows a power law, another unusual but by no means rare statistical distribution. The small world architecture may serve to minimize transition times between metabolic states, and contains evidence about the evolutionary history of metabolism. Ma H.-W et al. [Ma03b] also analyzed the connectivity of metabolic pathways and find that the metabolites in a

metabolic network are far from fully connected. Elementary modes for the analysis of metabolism and metabolic engineering are being exploited.

Elementary modes are the smallest functioning subunits of a metabolic network. They are also genetically regulated as a unit. A formal definition of such elementary flux modes requires that an elementary mode has a maximal number of vanishing fluxes and cannot be decomposed into smaller pathways [Sch00a]. Elementary mode analysis has been shown to be useful for investigating the metabolic capacity and pathway structure of metabolic networks. They have several promising applications in metabolic design, drug development or functional genomics [Sch00a]. Even though the kinetic characteristics of enzymes play an important role, many aspects of metabolism are actually constrained by the nature and connectivity of the metabolic reaction network (e.g. [Fel86]). The existence and number of feasible routes from nutrients to metabolites, and theoretical maximal yields, can be calculated without recourse to computer simulation [Sch99]. Schuster et al. [Sch96] [Pfe99] also devised a particularly efficient algorithm for determining all the available routes in a metabolic network.

Flux Analysis addresses the problems of actually determining the flow of all metabolites in the pathways through a limited set of measurements in the pathway. Researchers focusing on metabolic control analysis (MCA) emphasize the importance of profiling analyses in understanding the effects on metabolic networks when changing the activity of specific enzymes [Kel02]. Metabolic control theory is a formalism that describes the control of flux through the network as a function of enzyme and substrate quantities. In MCA one studies the relative control exerted by each step (enzyme) on the system's variables (fluxes and metabolite concentrations). This control is measured by applying a perturbation to the step being studied and by measuring the effect on the variable of interest after the system has settled to a new steady state. Stoichiometry-based metabolic control analysis complements the stoichiometric relations by measured fluxes. The interconnectivity of metabolites within a network of biochemical reactions is given by reaction equations defining the stoichiometric conversion of substrates into products for every reaction. From a methodological viewpoint stoichiometry-based metabolic flux analysis is a mature tool for metabolic network analysis. Several achievements are Mendes' Gepasi software that enables steady state analysis [Men99] [Men01]; works of Palsson, Schilling, and Schuster et al. [Sch00a] [Sch00b] explore pathways with constrained fluxes and optimal phenotypes. However, without energy balancing the flux balances are usually underdetermined. On the other hand, the stoichiometric network model is suitable for metabolic flux analysis but it contains no information about regulatory mechanisms. Thus it has little predictive power with respect to pathway alterations.

Combining metabolic control analysis with experimental observations on systems exhibiting large changes in metabolic flux led us to propose that these changes can only be explained if control mechanisms act at a number of points along the length of the metabolic

pathway [Tho96] [Tho98]. However, activation of single enzymes is less effective and has different characteristics [Tho97a] [Tho97b] [Fel00b]

Analysis of metabolic networks based on Petri net theory is also conducted. Kueffener et al. [Kue00] present an algorithm to systematically generate all pathways satisfying additional constraints in Petri nets. Based on the set of valid pathways, so-called differential metabolic displays (DMDs) are introduced to exhibit specific differences between biological systems, i.e. different developmental states, disease states, or different organisms, on the level of paths and pathways. Schuster et al. [Sch02a] presented a decomposition algorithm for metabolic networks based on the local connectivity of metabolites. The interrelations of pathway analysis of biochemical networks with Petri net theory are outlined.

2.5.2 Comparative Analysis

Comparative analysis of metabolic pathways in different organisms can give insights for understanding evolutionary and organizational relationships among species. This type of analysis allows one to measure the evolution of complete processes (with different functional roles) rather than the individual elements of a conventional analysis. Comparative analysis includes pathway clustering where the distances between pathway pairs are calculated by aligning enzymes, and pathways are classified based on distance measures [For99]. Pathway comparison can be conducted by comparing assigned genes on the genomes, by comparing assigned enzymes to specific pathways [Bon98] [Dan99], and by finding similarity of catalyzed enzymes that are classified according to the EC (Enzyme Commission) numbering system [Toh00a].

Metabolic pathway alignment represents one of the most powerful tools for comparative analysis. To align sequences, to measure distances, and to use similarity matrices in multiple sequence alignment algorithms, is a common approach to compare individual enzymes. Either by direct usage of molecular sequence data with, e.g. parsimony or maximum likelihood methods, or by a two-step approach via (1) multiple sequence alignment and calculation of a corresponding distance matrix, and (2) visualization of the distance data as graphs in that way a phylogenetic graph can be constructed. In Forst's paper [For01] these methods are extended to define distances between metabolic pathways. They combine sequence information of involved genes with information of the corresponding network. Metabolic pathways are considered as reaction graphs (networks) with specific graph-topological information, such as connectivity. For each functional role of a pathway, all genes in the genomes that code for this functional role are used. The sequences corresponding to the functional roles are combined into a set of sequences.

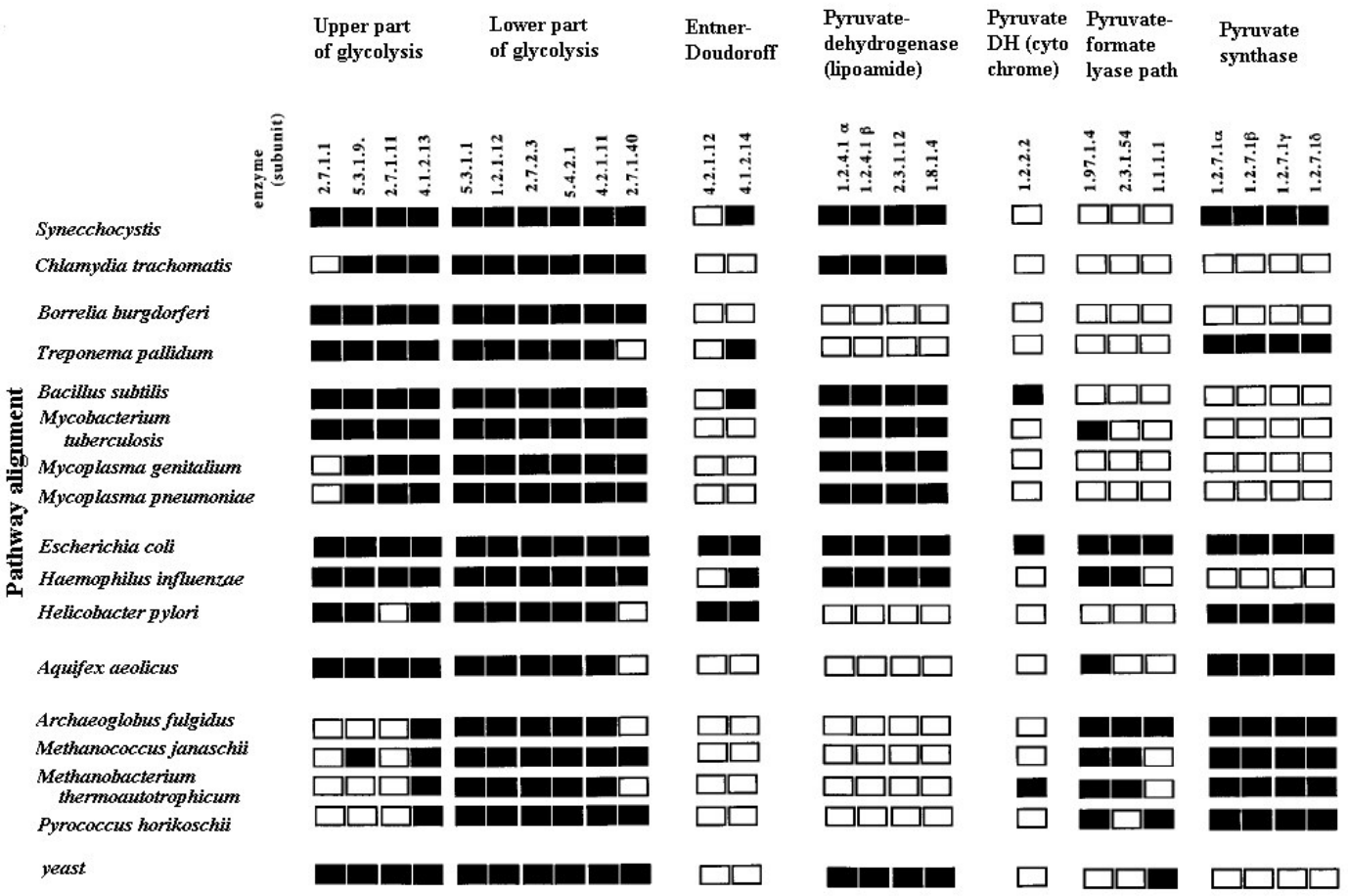


Figure 2.5.2 Pathway alignment for glycolysis, Entner-Doudoroff pathway and pyruvate processing [Dan99]. Enzyme for each pathway part are compared in 17 organisms and represented as small rectangles. Filled and empty rectangles indicate the presence or absence of enzyme-encoding genes in the different species.

Dandekar et al. [Dan99] studied three alternative ways of comparing biochemical pathways: (1) analysis and comparison of biochemical data, (2) pathway analysis based on the concept of elementary modes, and (3) a comparative genome analysis of 17 completely sequenced genomes. An example is given (Figure 2.5.2) that reveals a surprising plasticity of the glycolytic pathway. Liao et al. [lia02] have developed a computational method to compare organisms based on whole metabolic pathway analysis. The presence and absence of metabolic pathways in organisms is presented as a Boolean vector. Based on this methodology and by using some specific distance measures on these profiles, pairwise comparison of a set of completed genomes are performed. Tohsato Y. et al. [Toh00a] [Toh00b] presented a method for the alignment of reaction similarity of EC numbers. They use a dynamic programming based technique to align two or more pathways. They also proposed a multiple (local) alignment algorithm by utilizing information content that was extended to symbols that have a hierarchical structure like EC numbers. They considered that reaction similarities can be expressed by the similarities between EC numbers of the respective enzymes and applied their method to pathway analysis of sugar, DNA and amino acid metabolisms. Maureen Heymans et al. [Hey03] present a technique for the phylogenetic analysis of metabolic pathways based on the topology of the underlying graphs. A distance measure between graphs is defined using the similarity between nodes (enzymes) of the graphs and the structural relationship between them. This distance measure is applied to enzyme-enzyme relational graphs (two enzymes are related if they activate reactions which share at least one chemical compound) derived from metabolic pathways. Using this approach, pathways and groups of pathways of different organisms are compared to each other and the resulting distance matrix is used to obtain a phylogenetic tree.

In this thesis, a new algorithm for metabolic pathway alignment to reveal the similarities between metabolic pathways will be developed.

2.6 Summary

In this chapter we have introduced the complexity of cellular biology and explained the basics of biopathways and three traditional classifications: metabolic pathway, gene regulatory network and signaling pathway. The rapid accumulation of biological data makes it possible to compile detailed schemes of the bioprocesses within a cell. We have outlined several major molecular biological databases and addressed data integration problems. Concerning systems analysis of biopathways, the research *status quo* in modeling and simulation of metabolic networks, metabolic pathway prediction as well as metabolism comparison has been shown. Compared with different modeling and simulation approaches, the Petri net methodology is found to be a promising one. We have briefly given an overview of metabolic pathway

reconstruction, and discussed the existing approaches and their applicability. We have also described previous work on functional analysis and alignment of metabolic pathways.

In the next chapter, we will continue the discussion of Petri net-based modeling and simulation of biopathways with more details. A hybrid Petri net is to be introduced, strategies on cellular modeling will be suggested, the problem of large scale modeling and simulation is to be addressed, and finally we will propose a standard language for biological data interchange and modeling.

Chapter 3

Hybrid Petri Net Based Modelling and Simulation of Biopathways*

During the last decade applications of Petri nets to modeling and simulation of metabolic pathways appeared (§.2.3.4). However, these studies focused on either metabolic pathways or gene regulation separately. Moreover, some Petri net models only present initially qualitative aspects of a system. We attempt to use the Hybrid Petri nets to model an integrated metabolic network. Aside from handling discrete events, the hybrid Petri nets also allow the modeling of metabolic networks using actual concentrations. They are able to model biological processes with functions, which allow each transition to simulate kinetic effects.

3.1 Hybrid Petri Nets

Let us give a brief description of hybrid Petri nets as follows:

Definition 3.1 *A hybrid Petri net is a six tuple* $Q = (P, T, Pre, Post, h, M)$ *such that:*

$P = \{P_1, P_2,...,P_n\}$ *is a non-empty, finite set of places;*

$T = \{T_1, T_2,...,T_m\}$ *is a non-empty, finite set of transitions;*

$P \cap T = \varnothing$, *i. e. the sets P and T are disjointed;*

$h : P \cup T \rightarrow \{D, C\}$, *called "hybrid function", indicates for every node whether it is a discrete node* (*sets* P^D *and* T^D) *or a continuous node* (*sets* P^C *and* T^C);

$Pre : P \times T \rightarrow \boldsymbol{R}^+$ *or* $\boldsymbol{N}$, *is the input incidence mapping* ($\boldsymbol{R}^+$ *denotes the set of positive real numbers, including zero, and* $\boldsymbol{N}$ *denotes the set of natural numbers*);

$Post : T \times P \rightarrow \boldsymbol{R}^+$ *or* $\boldsymbol{N}$ *is the output incidence mapping*;

* Parts of Chapter 3 have been published in *ESM'02* [Che02a], *ISB* [Che03] and *Lecture Notes in Informatics* [Che02b].

$M : P \rightarrow \boldsymbol{R}^+$ *or* $\boldsymbol{N}$ *is the marking*.

We denote by $M_{(t)} = (m_1^t, m_2^t,...,m_n^t)$ the vector which associates with each place of P its marking at the instant t. $M_0 = M_{(t0)} = (m_1^0, m_2^0,...,m_n^0)$ is the initial marking. At any time the present marking M is the sum of two markings M^r and M^n, where M^r is the reserved marking and M^n is the non-reserved marking. If $h(P_i) =$ D or C then $m_i(t) = m_i^r(t) + m_i^n(t)$.

When a variable dT_j (called the delay time of T_j) is assigned to each discrete transition $T_j(h(T_j) = D)$ and T_j is fired at time $t + dT_j$, then

$\forall P_i \in {}^\circ T_j$ ($^\circ T_j$ denotes the set of input places of transition T_j), $m_i(t) \geq Pre(P_i,T_j)$,

$m_i(t + dT_j) = m_i(t) - Pre(P_i,T_j)$.

$\forall P_i \in T_j^\circ$ (T_j° denotes the set of output places of transition T_j),

$m_i(t + dT_j) = m_i(t) + Post(P_i,T_j)$.

When a variable vT_j (called the speed of T_j) is assigned to each continuous transition $T_j(h(T_j) = C)$ and T_j is fired at time t during a delay dt, then

$\forall P_i \in {}^\circ T_j$, $m_i^n(t) \geq Pre(P_i,T_j)$,

$m_i(t + d_t) = m_i(t) - v_j(t) \times Pre(P_i,T_j) \times d_t$;

$\forall P_i \in T_j^\circ$,

$m_i(t + d_t) = m_i(t) + v_j(t) \times Post(P_i,T_j) \times d_t$;

where $v_j(t)$ is the instantaneous firing flow of T_j at time t.

The concept of an inhibitor arc of weight r from a place P_i to a transition T_j allows the firing of T_j only if the marking of P_i is less than r. When this is used in a hybrid Petri net, we can extend the above-defined hybrid Petri net. If the inhibitor arc has its origin at a discrete place and has a weight $r = 1$, the corresponding transition can be fired only if $m_i > 1$, actually, only if $m_i = 0$, since m_i is an integer. If the origin place is continuous, then a conventional value 0^+ is introduced to represent a weight infinitely small but not zero. The new definition of an extended hybrid Petri net is similar to the definition of a hybrid Petri net (**Definition 3.1**), except that:

One can have, in addition, inhibitor arcs;

The weight of an arc (*inhibitor or ordinary*) *whose origin is a continuous place has its value in* $\boldsymbol{R}^+ \cup \{0^+\}$ *instead of* $\boldsymbol{R}^+$;

The marking of a continuous place has its value in $\boldsymbol{R}^+ \cup \{0^+\}$ *instead of* $\boldsymbol{R}^+$.

So far, the defined hybrid Petri net turns to be a flexible modeling process that makes sense to model biological processes, by allowing places using actual concentrations and transitions using functions.

The hybrid Petri net tool, VON++, is exploited to model and simulate gene-regulated network. Documentations for this tool can be downloaded via its website at http://www.systemtechnik.tu-ilmenau.de/~drath/visual.htm. Figure 3.1A shows the basis

elements of VON++, discrete place, continuous transition, continuous place and discrete transition connected with test arc, normal arc and inhibitor arc, respectively. There are no real input and output within test arcs, but the value of the places linked are exploited by the transition firing speed.

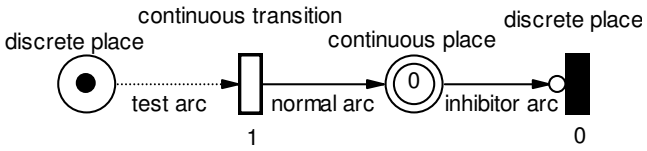


Figure 3.1A Elements of VON++.

It is clear that the discrete transition is the active element of discrete event Petri nets. A transition can fire if it is enabled by a sufficient number of tokens at its input places. It can be assigned a delay time. A continuous transition differs from the traditional discrete transition; its activity is not comparable with the abrupt firing of discrete transition. The firing speed assigned to a continuous transition describes the firing behavior of it and can be constant or a function, i.e. transport of tokens according to $v(t)$, where in Figure 3.1A $v(t) = 1$.

The rate of bioprocesses is not defined within a Petri net, it should be specified separately. In automated control systems represented by Petri nets, execution of transitions usually depends on the presence of specific number of tokens in all staring places. However, in most chemical and biological systems the rate of processes (transitions) is defined by the mass action law. The rate of change in the number of tokens (or concentration) is proportional to the number of tokens (or concentration) in all starting places as expressed in the Figure 3.1B below. V is the rate of firing of the transition; k is a constant (called a rate coefficient in chemical kinetics); m_3, m_4 are the concentration of place S_1 and place S_2. Coefficient k varies with temperature, pressure, solvent, and other factors. As a result, k will become a function of several variables.

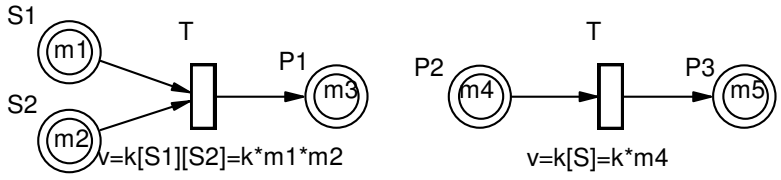


Figure 3.1B Presentation of transition rate in continuous systems.

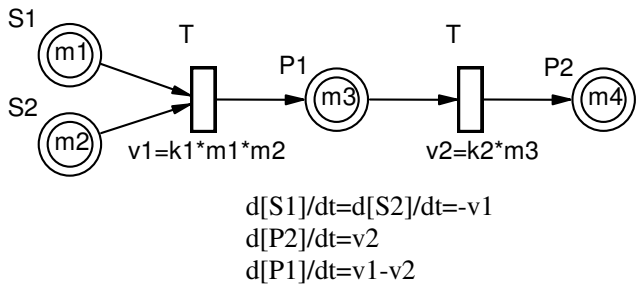


Figure 3.1C Presentation of intermediate reaction rate.

Figure 3.1C indicates the change of the number of tokens (or concentration) of reaction intermediate P1.

Normal discrete systems are easy to understand, so we emphasize here on continuous, which are very useful for modeling and simulating dynamic systems. We will describe some mathematical formulations that occur frequently in biology models, a general differential equation for a single state variable is $\frac{dx}{dt} = \sum flowin - \sum flowout$, while the expressions for the inflow and outflow can be quite complex, as every bioprocess gives rise to its own system of differential equation involving many dependent variables (species concentrations) and many free parameters (reaction rate constants). Mass action law assumes that particles move incessantly. However, cells are not like gas molecules.

Biochemical reactions are very complex, and interaction delay or saturation effect often exists in biological system. In these cases, mass action law becomes violated and should be replaced by equations that better describe the biological interaction while the rest of the algorithm remains the same.

3.2 Cellular Model Development

To understand the behavior of metabolic networks, modeling and simulation are of importance. Kinetic models of biochemical networks are becoming very important not only full genome sequences and biochemical reaction pathways are becoming available but also kinetic models of biochemical pathways are extremely complex and there is a strict requirement of software for their simulation, as in general these models do not have a known analytical solution.

3.2.1 Petri Net Model Construction of Metabolic Networks

The interpretation of Petri nets as metabolic system will be (Table 3.2.1):

Table 3.2.1 Mapping Metabolism to Petri nets.

Metabolism terms	Petri net terms
S, P, E, metabolites, genes, promoters, signals...	Places
Bioreaction, interaction, other bioprocesses, ...	Transitions
Defines reagent of bioprocess	Input arcs
Defines product of bioprocess	Output arcs
Initial token or state of system	Initial marking
State of reaction system	Marking
Rate of reaction system	Weight function or differential function
Enough of all the reagents must be present for the reaction to complete	A transition enabled
A single reaction	A firing transition
...	...

According to the interpretation, mapping from general metabolic terms to Petri net terms is reasonable. Places of Petri nets can represent all possible compounds, metabolites, enzymes, genes and so on; transitions can be interpreted as all possible bio-events such as biochemical reactions, transcription, transport and so on.

In figure 3.2.1A, an example of gene regulated metabolic network is displayed. The network consists of gene regulatory, metabolic reactions and signal transductions. The gene network contains five genes ***a***, ***b***, ***c***, ***d*** and ***e*** that encode molecules ***A***, ***B***, ***C***, ***D*** and ***E*** respectively. Gene ***a*** encodes a protein ***A*** which when binding with a metabolic ***G*** to the genomic site where gene ***c*** and ***d*** are triggered off. With the availability of binding complex of ***A*** and ***C*** at the transcription initiation site, Gene ***b*** regulates its own expression by encoding a repressor protein ***B*** which inhibits the transcription of the gene ***a*** and ***e*** that it regulates. Gene ***e*** regulates its own expression by encoding an enzyme that catalyses a reaction step in the metabolic pathway that consists of 8 metabolites (***D*** to ***K***). All the metabolites are connected with straight arrows. The gene e encoded protein ***E*** acts as an enzyme in the reaction that catalyzes ***F*** into ***H*** and ***I***; while the accumulation of ***I*** may result in a feedback on the transformation of ***D***. The resulting metabolic ***K*** represses / induces the regulatory action of protein ***B*** on a by modifying ***B***'s conformation. Moreover, an outer signal ***S***'s approaching to the complex of ***B*** and ***K*** may re-modify the conformation and bring more molecular functions.

The place/transition Petri net structure of metabolic system (Figure 3.2.1B), without any specific initial marking, is shown below:

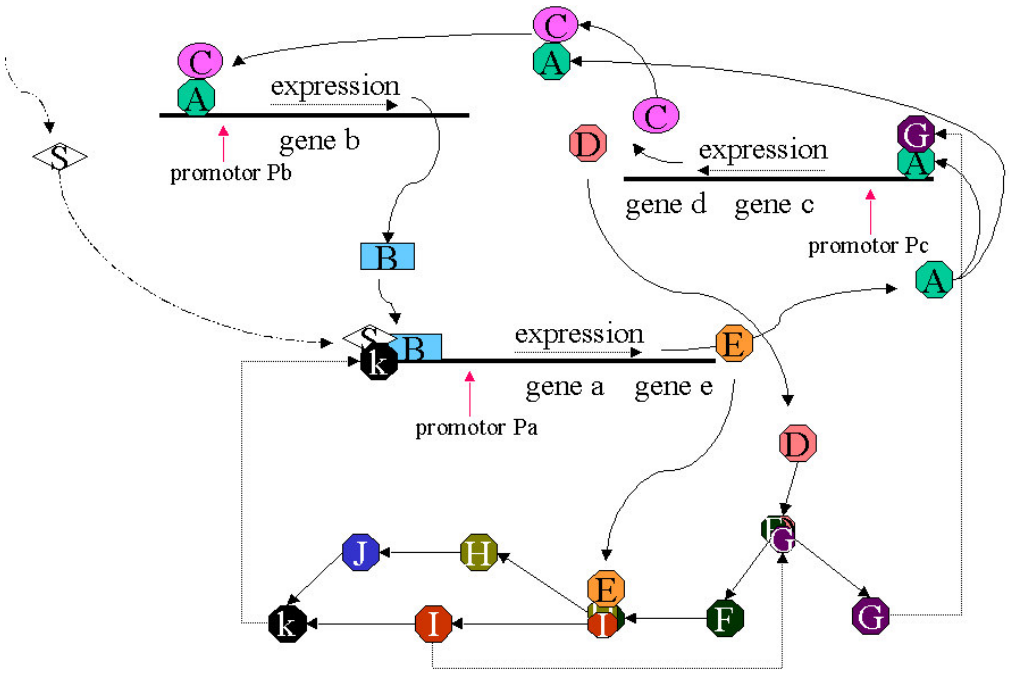


Figure 3.2.1A An example of a gene regulated metabolic network.

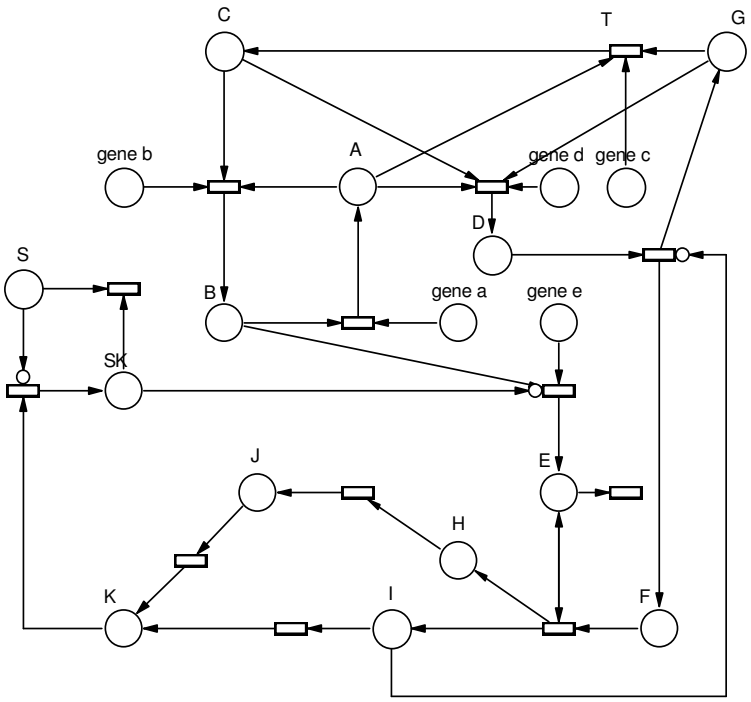


Figure 3.2.1B Petri net model of the generalized metabolic network.

Obviously, Petri nets have a simple graphical representation. Moreover, basic aspects of concurrent systems are captured conceptually as well as mathematically; the ability to visualize structure and behavior of a Petri net promotes the understanding of the modeled system; and Petri nets are executed and dynamic behavior observed graphically by various software tools which also support graphical construction and visualization.

3.2.2 Petri Net Model of Metabolic Reactions

In biochemistry, the most commonly used expression that relates the enzyme catalyzed formation rate of the product to the substrate concentration is the Michaelis-Menten equation, which is given as $v = \frac{v_{\max} \cdot S}{K_m + S}$. An example of its Petri net model and simulation result is shown as below.

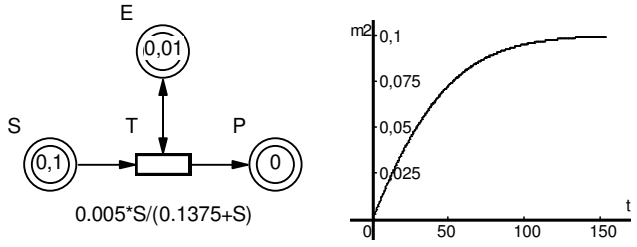


Figure 3.2.2 Petri net model of a simple enzyme catalyzed biochemical reaction (Michaelis-Menten reaction).

It is clear that such enzyme reactions are characterized by these two parameters: $V_{\max}$ and K_m, and biochemists are interested in determining these parameters from experiments. Fortunately, there are several biological databases available for public access, such as BRENDA, that provides enzyme reaction parameters. However, only for a subset of the well known pathways, those parameters are complete, and moreover an enzyme reaction can be affected by the presence of other compounds, *i.e*., the simplest form of the Michaelis-Menten equation does not account for the higher than first order substrate concentration dependence found in many allosteric enzymes. In the first case, we can introduce a general function $v=K_{app} \cdot S$ to meet the lack of unknown parameters, where K_{app} is the apparent rate constant. As we know the Michaelis-Menten equation is only valid when the concentrations of substrate and enzyme meet the precondition $[E]$ is not less than $0.001[S]$. When we consider the effect of enzyme concentration on the reaction rate in case the enzyme is regulated, *i.e*. the enzyme concentration is a variable of the model, the Michaelis-Menten equation can be written as $v = \frac{v_{\max} \cdot S}{K_m + S} = \frac{k_{cat} \cdot E \cdot S}{K_m + S}$, where k_{cat} is known as turnover number. When there are more than

two substrates and/or products involved in one enzymatic reaction, and its kinetic type is unknown, one then gets processes more complicated than discussed in the previous section. As the Michaelis-Menten equation is obviously invalid at this time, we simply apply the following function: $v = v_{\max} \cdot \prod_{i=1}^{n} \frac{S_i}{K_{mi} + S_i}$. In fact it is also in Michaelis-Menten form, e.g. for a two-substrate biochemical reaction, $v = \frac{v_{\max} \cdot S_1 \cdot S_2}{(K_{m1} + S_1) \cdot (K_{m2} + S_2)}$. Fortunately, if a two or more substrate biochemical reaction is already determined as one of the kinetic types list on the appendix A, the corresponded function should be applied.

3.2.3 Models of Gene Regulatory Networks

Gene regulatory networks are the on-off switches and rheostats of a cell operating at the gene level. The regulation of gene expression determines whether a protein is present to carry out its particular metabolic reaction and reaction specific kinetics. Based on interactions between genes and proteins, and reactions of genes and proteins, they dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into RNA. Each RNA transcript then functions as the template for synthesis of a specific protein by the process of translation. Process of gene regulatory networks is not restricted to the level of transcription, but may also be carried out at the levels of translation [Pyr96], splicing [Yao96], posttranslational protein degradation [Hoc96], active membrane transport [Wei93], and other processes. In addition, such networks often include dynamic feedback loops that provide for further regulation of network architecture and output.

Building complete kinetic models of gene regulatory systems requires detailed knowledge on reaction mechanisms. Often the following steps are considered:

1. The gene (DNA) is transcribed into RNA by the enzyme RNA polymerase.
2. RNA transcripts are subjected to post-transcriptional modification and control: rRNA transcript cut into appropriate size classes and initial assembly in nuclear organizer; tRNA transcript folds into shape; mRNA transcripts are modified, noncoding sequences (introns) removed from interior of transcript; in eukaryotes, all RNA types must move to the cytoplasm via the nuclear membrane pores.
3. Then mRNA molecules are translated by ribosomes (rRNA + ribosomal proteins) that match the 3-base codons of the mRNA to the 3-base anticodons of the appropriate tRNA molecules.
4. Finally, newly synthesized proteins are often modified after translation (post-translation) before carrying out its function, which may be transporting oxygen, catalyzing reactions or responding to extracellular signals, or even directly or

indirectly binding to DNA to perform transcriptional regulation and thus forming a closed feedback loop of gene regulation.

However, at the present time, the information of the bioprocesses from genes to the gene-encoded products is often unclear or unavailable. In such cases, we can regard the unknown part as a black box (one transition that stands for several other transitions) and simplify the whole procedure of a higher level of abstraction (Fig 3.2.3):

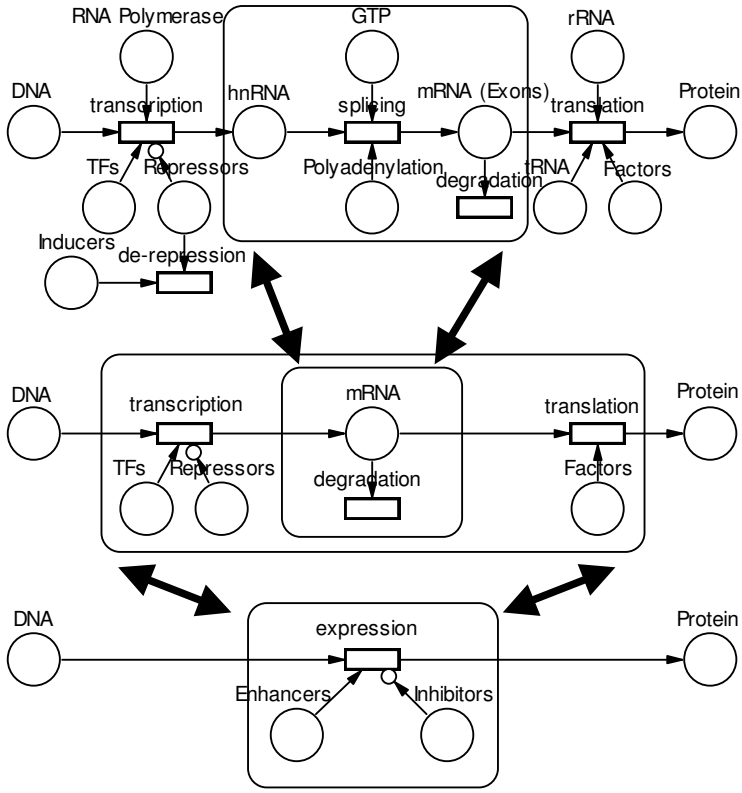


Figure 3.2.3 Petri net model simplification. In case of insufficiency of modeling values, the splicing of RNAs (upper block) can be abstract simplified modeled as mRNA (middle-inner block), while the whole process of transcription and translation (middle-outer block) can be simplified as expression process (bottom block).

Such simplifications do not require a change of the structure of the complete net and any modification to this subnet should be reflected in the behavior of the transition. Therefore, Petri net models are extensible and can be extended without significant deviation from the existing structure.

As to model gene regulatory networks quantitatively, we use the state equations of the following form to model bioprocesses such as activation of proteins, binding of proteins to genes, binding of RNA polymerase and so on.

If $state_{[i]}$(condition), **then** $\frac{dstate_{[i]}}{dt} = state_{[i]}(consequence)$

For example, the concentration of a gene product is $state_{[i]}$. The condition contains regulatory terms for this gene and describes whether the gene is being expressed or not. It depends on the state of the cell, and may contain models for promoters, enhancers, other proteins, nucleic acid, etc. The consequence is the result of the changed condition, here, the rate of gene expression.

So that the differential mass balances describing the concentration of mRNA and gene encode protein can be given as:

If (∃ (Gene, transcriptional factor(s), RNA nucleotides, binding of RNA polymerase, etc.) **not** ∃ (Repressors, etc.)),

then (transcription is initiated and mRNA is produced,

$$\frac{d[mRNA]}{dt} = [mRNA](GPC, mRNA) = k_{ts}[GPC] - k_d[mRNA]) ;$$

If ∃ (Modified mRNA, tRNA, initiation factor(s), amino acid, binding of ribosome, etc.),

then (the gene-encoded protein is synthesized,

$$\frac{d[P]}{dt} = [P](P, mRNA) = k_{tl}[mRNA] - k_d[P] - k_r[P]) ,$$

where k_{ts} and k_{tl} are the rates of transcription and translation respectively, k_d is the rate of degradation and k_r is the rate of consumption of biochemical reaction. GPC denotes the concentration of the binding complex of gene, TFs, RNA ploymerase, etc. DNA is a stable molecule, but mRNA and proteins are constantly being degraded by cellular machinery and recycled. Specifically, mRNA is degraded by a ribonuclease (RNase), which competes with ribosomes to bind to mRNA. If a ribosome binds, the mRNA will be translated, if the RNase binds, the mRNA will be degraded. Proteins are degraded by cellular machinery including proteasomes signaled by ubiquitin tagging. Protein degradation is regulated by a variety of more specific enzymes (which may differ from one protein target to another). In practice, the first-order rate constant of degradation k_d often is replaced by a half life H, and the degradation rate is expressed as $\frac{dC}{dt} = -\frac{0.693}{H}C$, where H=0.693/$k_d$. mRNAs have specific half-lives ranging from hours to days.

Regarding to the model of binding procedures which also are common phenomena in signal transduction, say

- converting inactive proteins into active proteins, and vice versa;
- binding of proteins to genes, proteins;
- binding of RNA polymerase to genes and gene-protein complex;

- binding of receptors to transcription factors.

A general model $[Complex] = K_b \cdot \prod_{i=1}^{n} [A_i]$, where K_b is the binding constant, is presented for systems consisting of one subject A_i binding with other subjects.

As in many situations, the information of gene regulatory pathway and mechanism is not available and one needs to take recourse to more approximate models. In this sense, the discrete model will be favorable.

3.2.4 Diffusion Transportation

Most of the models deal with the amount of metabolites in a cell. In the simplest case, we may be able to assume that the cell is a "well-mixed pool", i.e. the amount of metabolites is uniform across the cell. In many situations, however, concentration gradients exist which will affect the local rate of biochemical reactions, in particular for large systems and different compartments, we have to consider the effect of diffusion or of transport explicitly.

In general, if concentration gradients exist within the spatial scale of interest, it is very likely that diffusion will have an impact on the modeling results, unless the gradients change so slowly that they can be considered stationary compared to the timescale of interest. A growing number of modeling studies [Nar97] [Mar98] have emphasized the important effects diffusion can have on molecular interactions. Moreover, many bioprocesses take place in different compartments in a cell, e.g. glycolysis conducts in cyotoplasma while TCA in mitochondria. Membranes play an important role to separate these bioprocesses and meanwhile maintain the normal transportation of metabolites inside and outside of them. In addition, signal transduction also occurs across membranes.

So far, in order to model a metabolic network, not only all effects of metabolites and reaction behaviors, but different compartments should be considered. Diffusion will be the most important physical effect in the models we consider, but in other systems active transport could be as important, or even more important. We will focus on membrane transportation. The rate of penetration of a metabolite across a membrane is related to the concentration gradient by Fick's Law of Diffusion:

Rate of penetration $J = D \cdot A \cdot \beta \cdot \frac{d[S]}{dx} = \frac{D\beta}{\Delta x} \cdot A \cdot ([S]_{out} - [S]_{in})$,

where $[S]_{out}$ and $[S]_{in}$ are concentrations of metabolite outside and inside the membrane respectively; D denotes the diffusion coefficient (D decrease with the size of the metabolite); A is the area of membrane (the greater it is, the more metabolite that can pass); β is the partition coefficient (β increases with increasing solubility), and d_x is the membrane thickness

(the greater the thickness, the slower the rate). Usually $\frac{D\beta}{\Delta x}$ is called the permeability constant (P), a constant for a given substance moving through a given membrane.

Diffusion will be most important physical effect in the models we consider, but in other systems such as facilitated diffusion and active transport could be as important or more important. In carried systems, the carrier exhibit saturation kinetics, so this "Michaelis-Menten equation" formula might be used to describe such process. Low K_m means a high affinity and transport rate, and high K_m means a low affinity and transport rate. Some metabolites and/or signals (hormones) may modify carriers and change K_m. V_m is related to "carrier mobility", the total number of carriers present.

3.3 Petri Net Modeling Strategy

Modeling algorithm and analysis of hybrid Petri nets can be done by the following procedures:

1. Draft network construction

Normally, a Petri net model is built manually by drawing places, transitions and arcs with mouse events. Fortunately, the XML based Petri net interchange format standardization, which consists of a Petri Net Markup Language (PNML) [Web03] and a set of document type definitions (DTD) or XSL Schema is coming into being and intended to be applied. Several Petri net tools such as PNK, Renew and CPN have been equipped with an XML-based file format exchange. We have developed an environment to extract data of metabolic networks from KEGG, BRENDA and RegulonDB and transform them into XML-based files that can be used by PNK and Renew to display the Petri net models automatically.

2. Data searching

The main feature of metabolic processes is that the concentration of metabolites will influence the reaction activity of bioprocesses. Therefore, the actual concentration of any metabolite is an important component of the quantitative model. Although some data nowadays are available to the public via the Internet, some other data may not be complete. It requires time-consuming literature searches. Assignment of initial value of places is made after data gathering.

3. Defining the kinetics of each reaction

We have collected a series of predefined kinetic types that are the types most often used in biochemical reaction models (see Appendix A). However there are some circumstances in which the kinetic types are not yet defined. New kinetic type self-definitions are handled by the mass law. A certain kinetic type is only presented as a choice for reactions that have the same number of substrates as marked in that type. If it is a reversible reaction, then it also needs to match the number of products. Once you have defined all the reactions of the model, you must assign a kinetic type to each of them. You will also need to provide

values for all the kinetic constants and information about variables (that is, if the kinetic type you selected has any).

4. Self-defined kinetics

Many metabolic pathway schemes contain mass conservation relations that must be taken into account in order to carry out the simulation. To check the mass conservation relations of a model we can go to the original reaction data from databases. In fact, we construct the model with the identification of reaction stoichiometry. Otherwise, it will lose something when the simulation is carried out because in continuous Petri nets, the weight of arcs is disabled, so that all components involved in the reaction are changed with same rate, which is defined by the transition function. However, in the reaction $2A+B \rightarrow C$, the change of A should be twice that of the reaction rate. In VON++, unfortunately, we have to add more transition from *A* with the same function in order to obey the mass conservation law.

5. Parameter tuning and simulation

To build a model precisely requires as many variables as possible and parameters involved in a metabolic network. The values of variables and parameters are determined either by experimental methods or deduced from other related values. However, it is impossible or sometimes unnecessary to put all variables and parameters into a model. The model is plausible when main influences are included. On the other hand, because of different purposes and situations, most data from laboratory do not fit the model very well, and vice versa. We have to compare and tune the differences in order to find suitable ones. Then the effects of various parameters on the gene regulated metabolic networks and their relations can be determined. The key enzymes/proteins, as well as intermediates related in the metabolic pathway, can be determined, which can provide the necessary information to identify and solve metabolic bottlenecks.

3.4 Large Scale Network Modeling and Simulation

One of the ultimate goals of computational metabolism is the modeling and simulation of the whole cell - virtual cell development. There are currently several ambitious attempts to build whole cell models of cellular biochemistry [Gib01], including the virtual cell project [Sch00c] [Sch01a], E-Cell project [Tom99] [Tom01], and other works [Oli01] [Res01] [Nob02a] [Nob02b] [Fel01] [Voi00a] [Voi00b] [Hei02a] [Hei02b].

3.4.1 Problems and Methods

3.4.1.1. Constitutive model development

In order to perform a virtual cell modeling and simulation, the following problems have to be considered:

1. Cellular process analysis

Metabolisms in living organisms are not at the equilibrium state. Cellular processes involve material, signal and energy transfer. Metabolic processes involve a high number of interconnected biochemical reactions. The product of one reaction serves as the reactant for the next. The same compound may serve as a reactant for several parallel reactions that produce different products. The mathematical relations between the thermodynamic properties of a metabolite and its activity in the living environment allow calculating their thermodynamic properties in the stationary metabolic state. All chemical reactions, including enzyme-catalyzed reactions, are to some extent reversible; a readily reversible reaction has a small numerical value of ΔG. We can then combine this information algebraically to describe the thermodynamics of metabolism.

A biochemical reaction with a large negative value for ΔG might be termed "effectively irreversible" in most biochemical situations. Within living cells, however, reversibility may not occur, because reaction products are promptly removed by additional enzyme-catalyzed reactions. Metabolite flow in living cells is largely unidirectional. True equilibrium, far from being characteristic of life, is approached only when cells die. The living cell is a dynamic steady-state system, maintained by a unidirectional flow of metabolites. In mature, the mean concentrations of metabolites in cells remain relatively constant over considerable periods of time. Short-term oscillations of metabolite concentrations and of enzyme levels do occur, however, and are of profound physiologic importance. The flexibility of this steady-state system is illustrated by the delicate shifts and balances by which organisms maintain the constancy of the internal environment despite wide variations in food, water, and mineral intake, work output, or external temperature.

2. Regulation mechanism

Metabolic regulation is exerted primarily at branch points, where a metabolic intermediate is partitioned between two pathways. The branch point metabolite is the substrate for two or more enzymes, and the relative amount of the metabolite that enters each pathway depends on competition between the two enzymes. The outcome of such competition depends largely on the relative affinities of the two enzymes for their common substrates. This modulation of the affinities of competing enzymes must lead to a kind of interaction between pathways.

Net metabolic flow of any enzyme-catalyzed reaction may be influenced (1) by changing the absolute quantity of enzyme present, (2) by altering the catalytic efficiency of the enzyme, and (3) by reversibly modifying the catalytic activity of the enzymes. All three options are exploited in most forms of life.

(1) The absolute quantity of an enzyme present is determined by its rate of synthesis (k_s) and rate of degradation (k_d). The presence or absence of substrates, coenzymes, or metal ions alters proteolytic susceptibility, which can convert an inactive proenzyme to a catalytically active form. The concentrations of substrates, coenzymes, and possibly ions in cells may also influence the rates at which specific enzymes are degraded. Arginase and tryptophan oxygenase (tryptophan pyrrolase) illustrate these concepts. Regulation of liver arginase levels can involve a change either in k_s or in k_d. After a protein-rich diet is ingested, liver arginase levels rise owing to an increased rate of arginase synthesis. Liver arginase levels also rise in starved animals. Here however, it is arginase degradation that is decreased, while k_s remains unchanged.

(2) The control of enzyme activity could be allosteric effects. The catalytic activity of certain regulatory enzymes is modulated by low-molecular-weight allosteric effectors that generally have little or no structural similarity to the substrates or coenzymes for the regulated enzyme. Notice that allosteric and catalytic sites are spatially distinct. Allosteric effects may be on K_m or V_{max}. Reference to the kinetics of allosteric inhibition as "competitive" or "noncompetitive" with substrate carries mechanistic implications that are misleading. The kinetics of feedback inhibition may be competitive, noncompetitive, partially competitive, uncoupled, or mixed.

(3) Reversible, covalent modification of the catalytic activity of enzymes can occur by covalent attachment of a phosphate group to one or more Ser, Thr, Tyr, or His residues. Enzymes that undergo covalent modification with attendant modulation of their activity are termed "interconvertible enzymes." Interconvertible enzymes exist in two activity states, one of high and the other of low catalytic efficiency. They play important roles in signaling events, though some precise details by which these enzymes act are in most instances still far from clear.

3.4.1.2. Model simplification

For a number of practical and esthetic reasons, we wish our models and explanations of biological phenomena to be as simple as possible. On the other hand, biological systems are complex, having many processes and variables that interact in complicated, non-linear ways. There are a few principles for simplifying models:

1. **Eliminate state variables**

Every state variable must have a dynamic equation (differential equation or finite difference equation) as well as parameters and initial conditions. There are two ways to reduce model complexity arising from state variables: (1) convert a state variable into a constant, and (2) aggregate state variables.

2. Make "stronger" assumptions

Two methods are exploited: (1) convert functions of state variables into constants, and (2) convert nonlinear relationships into linear relationships.

3. Remove temporal complexity

(1) convert random models into deterministic models, and (2) convert driving variables to constants.

4. Remove spatial complexity

An assumption of the Michaelis-Menten kinetic approach is that the concentration of total substrate is essentially equal to the concentration of free substrate. This assumption may be valid when modeling small volumes but should be carefully evaluated in all other contexts, especially within membrane transportation. The simulation of models with more than one compartment is not hard to implement in a generic simulation program, but the inclusion of diffusion effects is more problematic. Studies of realistic reaction-diffusion metabolic models would greatly increase our understanding of cellular processes.

3.4.2 Prospect of Petri Net Tools

In the following requirements for a biology specific Petri net tool are discussed:

3.4.2.1 Cell modeling theory

1. Metabolic pathway layout

Structural knowledge of a physical system is the foundation of a simulation. As we know a Petri net representation is a type of object-oriented representation in which metabolites are grouped together into objects that correspond to real-world entities. Thus, it can present a structure of metabolism in a natural way. The metabolic pathway editor tends to be based on the Petri net methodology. Because Petri nets are mathematically well defined and have a mature theoretical background, so that many static and dynamic properties of a Petri net (and hence a system specified using the technique) may be mathematically proven. In addition, Petri nets may be executed and the dynamic behavior observed graphically. As the matrix format of biochemical reactions are commonly used throughout many tools, our software should be capable to handle matrix data format. Which on the other hand can simplify the transfer from other tools.

2. Metabolic data connection

XML is a standard for storing and transferring data and many biological databases such as TRANSPATH and MPW have already or are being considering using the technique. The intended software should support XML import & export, link to internal & external related databases on gene, enzyme, reactions, kinetics, organisms, compartment and initial values of (ODE-NAE) system. BioPNML (§ 3.5) could be used as a standard.

Moreover, with the complete annotated sequence of a genome we can generate drafts of the organism's metabolic networks. For the next generation of metabolic models, which will probably be integrated with genome databases, it would be possible to include fields containing information on the evidence for particular values, e.g. Evidence Codes in Gene Ontology [http://www.geneontology.org/doc/GO.evidence.html].

3. Hierarchical concept

Hierarchical biochemical systems are biochemical systems that consist of multiple modules that are not connected by a common mass flux, but communicate only through regulatory interactions. The models of a virtual cell should contain metabolic pathways and the levels of transcription and translation, and so on. Reactions in different compartments require a hierarchical model representation. The translation rates are increased by the concentration of mRNAs, the metabolic rates are increased by the concentration of enzyme and the transcription rates are affected by the concentrations of metabolites. E-Cell models fulfill this technique very well. With the mature mathematical support, Petri nets also can handle it and at the same time it make the Petri net structural reduction possible as usually the state space of Petri net structure will be very large in graphs.

4. Metabolic kinetics

Structural knowledge alone captures the state of a system at a fixed point in time, but does not capture the relationships and interactions among structural components over time. Process knowledge is functional knowledge of dynamic change. A dynamic process representation is critical to the success of a simulation. Traditionally, ODE and NAE models, such as Michaelis-Menten kinetic models, are used to simulate metabolic reactions. In order to build a quantitative model, kinetic properties of enzyme-catalyzed reactions involved in pathways should be outlined.

5. Determining kinetic mechanisms

Different types of enzyme kinetics (Michaelis-Menten equation, Reversible mass action kinetics equation, Allosteric inhibition equation, etc.) and initial parameter values can partly be obtained via certain databases and literature. Otherwise, a user-defined model should be built as E-Cell and Gepasi are. Nevertheless, a Petri net based simulation system still can deal with it as a discrete-event or semi-quantitative model when the required data are unavailable. The well-known Michaelis-Menten equation is not a complete description of the

behavior of single-substrate enzymes in vivo because of product inhibition and reversibility, so that a new kinetic algorithm should be worked out.

6. Thermodynamic control of reactions

Biochemical reactions are thermodynamically feasible. A link between thermodynamics and kinetics can show the relation between the kinetic constants, K_{eq}, ΔG^0 and ΔG, and the driving forces for the reaction. Factors limiting the rate of an enzyme-catalyzed reaction include temperature, pH, competitors, ionic, allosteric molecules, substrate concentration, substrate location and so on, which should be considered for user defined models.

3.4.2.2 Computation method

In metabolic systems, detailed quantitative knowledge is unavailable today. Models of this system are to be constructed by combining qualitative knowledge of relationships and partial quantitative data. In this case, simulating such a model may be the only means for generating predictions. Initially, Petri nets are developed as a discrete-event modeling and simulation systems. Traditionally, kinetics has been taught in biochemistry courses in terms of enzyme steady-state kinetics. This corresponds to a detailed study of the local properties of the individual enzymes. However, one can go further and create kinetic models of whole pathways. Such models are composed of coupled ordinary differential (for time courses) or algebraic (for steady states) equations. These equations are non-linear and most often without analytical solution. This means that they can only be studied through numerical algorithms, such as the Newton method for solving non-linear equations and numerical integrators. With many years of development, quantitative modeling is now possible to be handled by Petri nets. They have a mature mathematical algorithm and can solve NAE and ODE and stoichiometric matrices. But biochemical systems are also rich in time scales and thus require sophisticated methods for the numerical solution of the differential equations that describe them.

Parallel treatment of these equations during simulation is of importance, yet difficult to achieve. Moreover, when we consider other functions of the metabolism, such as MCA methodology and bifurcation analysis, it is necessary for the tool to be powered by a more efficient algorithm. MatLab is one of the most popular software systems in the area of applied mathematics, so that integrating MatLab [http://www.mathworks.com/] in Petri net models is probably a good solution. In addition, MatLab itself can be applied as an attractive Petri net tool builder. Now it is possible to analyze and visualize Petri net models by transferring them to convenient graphical design tools. Export to matrix representation in MatLab is possible, and M. Svádová [Svá00] reported an approach to use the MatLab standard libraries and built a Petri nets toolbox that enabled Petri net modeling, analysis and visualization of simulation results.

As metabolism is far less well understood than a manufacture system. Consequently, biological simulations often yield highly uncertain results. So far, a bifurcation analyzer or fuzzy analyzer should be included in the software as Dbsolve does. And concentration values of metabolites within a cell fluctuate in a normal range; pre-arrangement of such data in the software is necessary.

Furthermore, the pathway simulator is able to predict pathways, from several known biochemical reactions; we can predict the whole connected network of them based on the tool algorithm. It can also calculate thermodynamical characteristics. Each reaction is thermodynamically feasible, that is, ΔG be equal to or less than zero. Otherwise, the requirements for coupling of reactions (combined with ATP utilization) should be checked and any two-coupled reactions must proceed via a common intermediate. The reversibility of one reaction is determined and displayed in case abnormal situations occur, though the metabolite flow tends to be unidirectional.

3.5 Biology Petri Net Markup Language

As previously mentioned, there are many biological simulators and Petri net tools available, but few common exchange formats, even with XML format. As a result it is difficult to exchange models between different analysis and simulation tools, and take advantage of different tools. In this section, a proposal for a common exchange language - Biology Petri Net Markup Language (BioPNML) is presented.

3.5.1 Introduction

In the post-genomic era new methods are proposed to store these data and retrieve them and analyze and reanalyze. XML, as an emerging standard for data interchanging, is more and more adopted to structure data exchange in bioinformatics. The following sections briefly discuss the relationship between XML, bioinformatics and PN.

3.5.1.1 Bioinformatics & XML

There are already two good review papers on this topic by V.H.Guerrini [Gue00] and F.Achard [Ach01]. We would like to highlight a few of their points and supplement them with a few fresh examples for biopathway applications.

XML is derived from the Standard Generalized Markup Language (SGML), the international standard for defining descriptions of the structure and content of different types of electronic documents. XML is a web-dedicated data exchange language, which omits the complex and less used parts of SGML. The World Wide Web Consortium (W3C) has supervised the specifications of XML since its inception in 1996. More documentation can be found at http://www.w3.org/XML/.

In bioinformatics, XML was widely used within the last few years, and several XML based data formats have been developed. BSML (Bioinformatic Sequence Markup Language) [http://www.bsml.org/] uses XML to provide genomic information and a graphical BSML browser was developed. BioML (Biopolymer Markup Language) [http://www.bioml.com/BIOML/] integrates nucleotide and protein sequence data. The XML based RDF format [http://www.w3.org/RDF/] is also adopted by the Gene Ontology Consortium [http://www.geneontology.org] to provide controlled vocabularies for the description of molecular functions, biological processes and cellular locations of gene products. Moreover, major biology databases such as NCBI, WIT and ExPASy also provide XML output after users' database queries. Obviously, XML is widely adopted as a standard for the exchange of biological data.

Both CellML (Cell Markup Language) [http://www.cellml.org/] and SBML present description languages for cellular simulation. CellML is intended to be used to represent many different types of models, for instance biochemical pathway models. Aside from specifying a model purely in terms of mathematics, CellML can use some additional elements to fully capture the information in biochemical pathway models. SBML is oriented towards representing biochemical networks common in research on a number of topics, including cell signaling pathways, metabolic pathways, biochemical reactions, genomic interactions, and many others. The main difference is that CellML has a very general and flexible syntax, while SBML's syntax is specific to metabolic pathway modeling. Currently, SBML is closely collaborated among several teams that develop metabolic simulators.

Although many biological databases and bioinformatics research groups use XML, it is however so flexible that anyone can create his/her own versions in entirely different ways. XML enables advancements in application integration, but they are difficult to achieve without a consistent framework for XML implementations.

3.5.1.2 Petri nets & XML

At present most Petri net tools import and/or export Petri nets in proprietary file formats and poorly support other data formats. In these proprietary file formats it is difficult to add and remove features to the language and to make modularization of diagrams as easy as it might be in an ASCII based text format such as XML.

In order to solve the problems caused by the use of different file formats, many Petri net tools are currently being equipped with XML support. R.B. Lyngsø et al. [Lyn98] presented a text format based on SGML for Design/CPN diagrams and proved that the framework is indeed possible to use SGML to represent High-level Petri Nets. Renew [Kum00], from its version 1.3, supports XML import and display Petri net automatically. Matthias Jüngel et al. [Jue00] presented the concepts and terminology of PNML (Petri Net

Markup Language), and thus provided a starting point for the development of a standard interchange format for Petri nets.

Although the above-mentioned Petri net XML standards are available, they are incompatible due to different design destinations. Actually, one cannot adapt ones Petri net XML file to fit them without any modification. Moreover, a common problem to implement it is, that every user has to write an XML file from the original data source. For instance, in order to construct a model out of our database, we have to transform the original data into the desired XML file. By using the W3C recommended Extensible Stylesheet Language Transformations (XSLT), new structured data formats can be created from existing XML documents. That is, XSLT is a language for transforming XML documents into other XML documents. An XSLT file (appendix B) is developed to convert the original XML source file from our metabolic pathway data stored in an Oracle system into the desired XML format that can be executed by the Renew XML parser. Figure 3.5.1.2A shows the automatic layout of Petri net model with Renew. The Petri net model layout with PNK is shown in Figure 3.5.1.2B.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet
href="http://sanfrancisco/xml/db2xml/test.xsl"
type="text/xsl"?>
<database URL="jdbc:oracle:thin:@edradour.cs.uni-
magdeburg.de:1521:orcl">
<table0 QUERY="select * from enzyme where ec =
'3.5.3.1' or ec= '4.3.2.1' or ec='6.3.4.5' or ec='2.1.3.3' or
ec='6.3.4.16'"
>
<record0>
<EC><![CDATA[6.3.4.16]]></EC>
<PRODUCT><![CDATA[ADP]]></PRODUCT>
<SUBSTRATE><![CDATA[NH3]]></SUBSTRATE>
</record0>
<record0>
<EC><![CDATA[6.3.4.16]]></EC>
<PRODUCT><![CDATA[ADP]]></PRODUCT>
<SUBSTRATE><![CDATA[CO2]]></SUBSTRATE>
</record0>
...
</record0>
...
</table0>
</database>
```

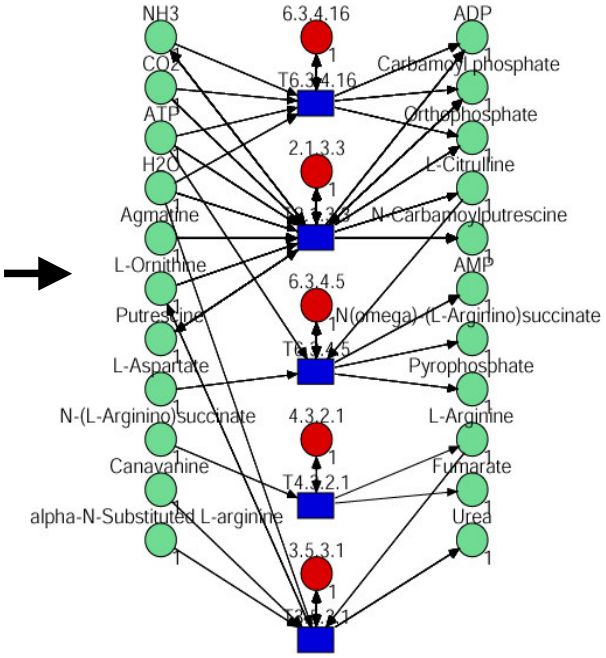


Figure 3.5.1.2A Petri net model layout based on XML (Renew).

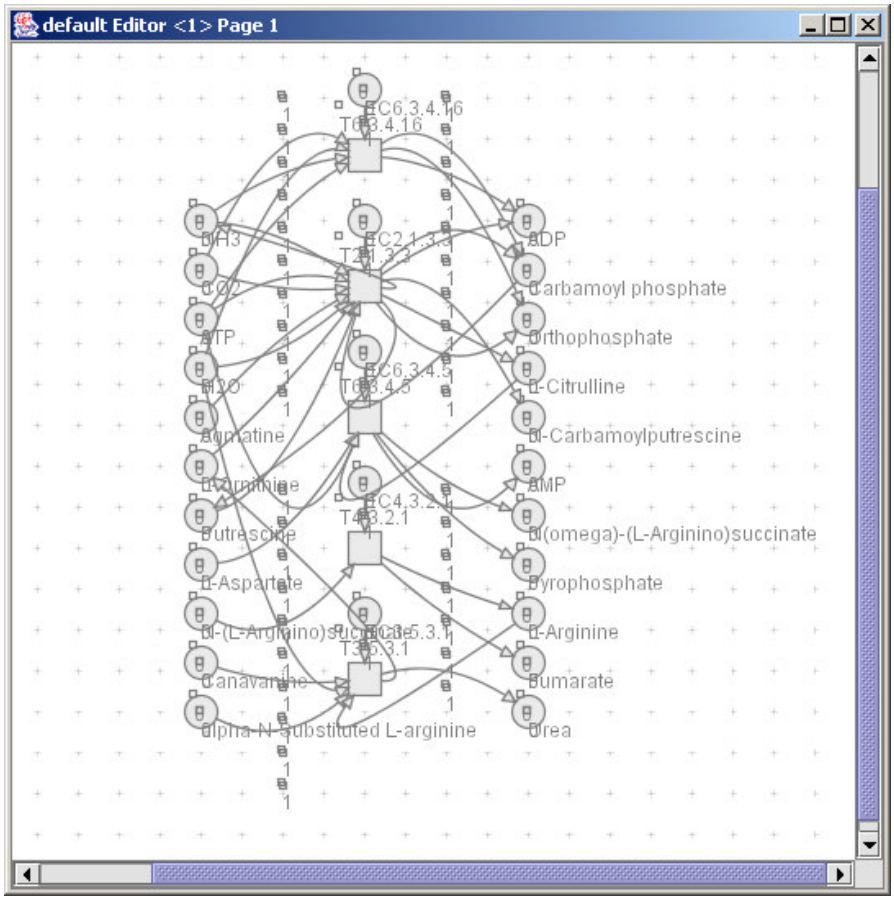


Figure 3.5.1.2B Petri net model layout with PNK2.

As we know PNML is generic and can be extended according to the user specific needs. So that a special "Bio-PNTD" for PNML can be defined when a simple biological system is modeled. However, a metabolic network model can contain a large number of named components representing different parts of a model. In this case, SBML model definitions are more suitable. Therefore, with regard to the application of Petri net methodology to bioinformatics, particularly for modeling and simulation of metabolic networks, a new interchange format is what is really needed.

The PNML code for the model is translated by an XSLT file and outlined as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<!--pnml for metabolic reaction petri nets - mchen@techfak.uni-
bielefeld.de-->
<pnml>
      <net id="N" type="PTNet">
            <name>metabolic petri nets</name>
            <place id="6.3.4.16">
                  <marking>
                        <graphics>
                              <offset page="1" x="-4" y="-10"/>
                        </graphics>
                        <value>0</value>
                  </marking>
                  <name>
                        <graphics page="1" x="-15" y="-30"/>
                        <value>EC6.3.4.16</value>
                  </name>
                  <initialmarking>
                        <graphics>
                              <offset page="1" x="-4" y="-10"/>
                        </graphics>
                        <value>0</value>
                  </initialmarking>
                  <graphics>
                        <position page="1" x="300" y="120"/>
                  </graphics>
            </place>
            <transition id="T6.3.4.16">
                  <name>
                        <graphics>
                              <offset page="1" x="-15" y="-30"/>
                        </graphics>
                        <value>T6.3.4.16</value>
                  </name>
                  <graphics>
                        <position page="1" x="300" y="160"/>
                  </graphics>
            </transition>
            <arc id="ARC6.3.4.16" source="6.3.4.16"
target="T6.3.4.16">
                  <inscription>
                        <graphics>
                              <offset page="1" x="0" y="-1"/>
                        </graphics>
                        <value>1</value>
                  </inscription>
                  <graphics>
                        <position page="1" x="300" y="140"/>
                  </graphics>
            </arc>
            <place id="NH3">
                  ...
            </place>
      </net>
</pnml>
```

In this following section, the concepts and terminology of BioPNML interchange formats, as well as its syntax is to be presented.

3.5.2 Concepts and Terminology of BioPNML

The intended BioPNML is a XML-based description language that allows the representation of metabolic networks as Petri nets. Before introducing the syntax of the interchange format, we briefly discuss its basic concepts and terminology, which is independent of the XML representation. Previous approaches proved, that by using hybrid Petri net methodology, it is feasible to model and simulate metabolic systems [Mat00] [Mat01] [Che00] [Che02a]. Therefore the first version of BioPNML supports the hybrid Petri net type. BioPNML contains Petri net objects as well as data needed for the exchange and graphical representation of metabolic networks. An XML schema defines the labels for a Petri net and its objects and metabolic models.

3.5.2.1 Petri net objects and labels

From Table 3.2.1 we know that places can be used for the representation of biological subjects such as genes, metabolites, proteins, enzymes, compounds and other molecules, while transitions represent biochemical reactions and interactions. The value of tokens in places can represent the actual concentrations of biological subjects. Transitions can be classified into two types: discrete and continuous. A discrete transition fires, if it has concession, and a delay time can be assigned to it. Continuous transitions are not comparable to the abrupt firing of discrete transition. The firing speed assigned to a continuous transition is defined by a constant or a function. Arcs between places and transitions fall into three categories: normal arcs, inhibitor arcs and test arcs. In metabolic pathways, arc weights of continuous transitions are assigned according to the stoichiometric coefficients of the biochemical reactions.

3.5.2.2 Petri net graphics

Every object is equipped with some graphical information. For a place and transition, the information is its shape, size and position; for an arc, it is a list of positions that defines start and end points of this arc. In the Biology Petri net, the main properties are, that the arc weights are described by the stoichiometric coefficient of the biochemical reaction, and the transition condition is described by using functions or by assigning a delay time. Figure 3.5.2.2 shows the Petri net representation of a biochemical reaction. ***S***, ***E***, ***P*** and ***ES*** denote substrate, enzyme, product and the enzyme-substrate complex respectively. The biochemical reaction indicates that a substrate is enzymatically catalyzed into a product with a transformation rate v. Three places represent substrate, enzyme and product with ***S***, ***E*** and ***P*** as the label of places. The tokens (real concentrations) of each place in the Petri net can be used as variables, m_1, m_2 and m_3, while the transition rate is assigned with a known function, the Miachlis-Menten equation.

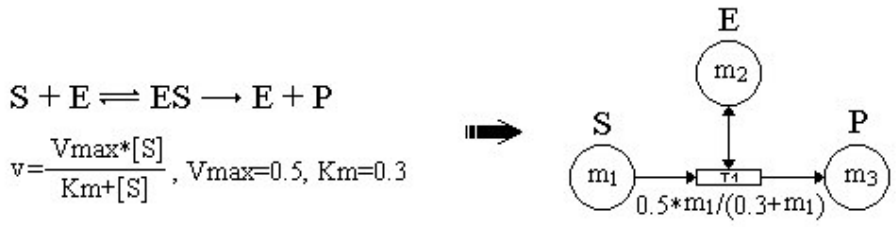


Figure 3.5.2.2 Petri net presentation in BioPNML.

3.5.2.3 Classes

Figure 3.5.2.3 shows the class diagram of BioPNML. The left side shows the Petri net part of BioPNML that was derived by extending PNML slightly. The right part shows the Biological part that is based on SBML. The Petri net part contains only very few extensions to PNML. More changes to SBML were made in the biological part, although those changes are open for discussion. This is due to the fact that Petri nets have a comparatively long history and well-defined, generally accepted syntax and semantics, whereas molecular biology is still evolving at a rapid pace. The purpose of the schema is to give a rough idea of how Petri nets and biological systems are related. This diagram can also serve as a conceptual guidance to researchers who are designing databases to store networks and reaction data.

Main classes include metabolic pathway, gene regulation and signal transduction, being consistent with the traditional classification of metabolic networks.

1. Metabolic pathway

In BioPNML, the metabolic reaction class is defined as biochemical reactions and related objects such as: enzyme, substrate(s), product(s), their stoichiometries, and parametric values for separately defined kinetic laws. In figure 3.5.2.3, the metabolic reaction class structure that was derived by extending SBML's biochemical reaction class [Huc01] is shown.

The metabolic reaction class contains mandatory fields (enzyme, substrate, product, and *KineticLaw*), as well as optional fields (enhancer and inhibitor). Enzyme is a reference to the gene that encodes the enzyme. Both substrate and product are references to molecules implemented using lists of *SpecieReference* structures. The *SpecieReference* structure contains fields for recording the names of molecules, the types of molecules that are references to lists of *TypeRef* structure; the stoichiometry filed indicates the proportions of substrate and product within a reaction. The *KineticLaw* structure is an optional field of the type *KineticLaw*, used to provide a mathematical formula for the reaction rate. The Boolean field, reversibility, indicates whether the reaction is reversible. The field is optional, and has default "true" when it is not specified. Information about reversibility is useful in certain kinds of structural analysis such as elementary mode analysis [Sch99].

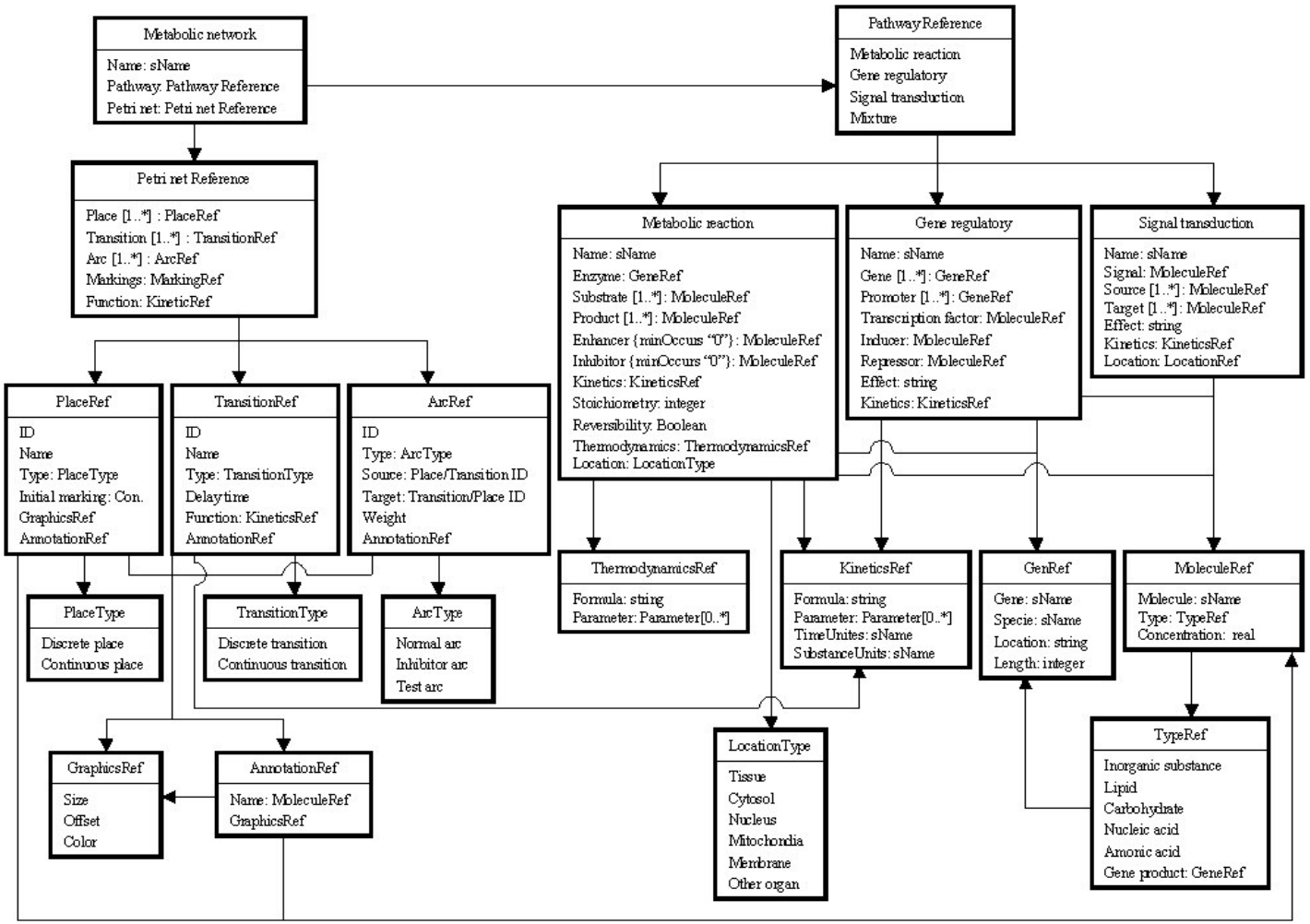


Figure 3.5.2.3 Diagrammatic class relations of BioPNML.

In addition to these fields, the reaction structure also has a Thermodynamics field as a reference to *ThermodynamicsRef*. The *ThermodynamicsRef* structure is an optional field that is used to provide the Gibbs energy that indicates the favorability of the reaction.

2. Gene regulatory

In BioPNML, the gene regulation class is defined as a set of objects such as genes, promoters, transcription factors, inducers, repressors, the gene encoding proteins, other metabolites and the effect of interaction and kinetics.

3. Signal transduction

The Signal transduction class in the BioPNML is defined as a set of signal instances through the message passing between source and target. It references molecular interaction motifs, effects of the signals, components of the transductions, and properties of signal transduction.

4. Other bioprocesses

Biological cells are highly complex systems. Some biological systems, such as membrane transportation, do not fit in one of the above-mentioned three basic categories, but should also be taken into account when required. Many models assume that the amount of metabolites in a cell is uniform across the cell, i.e. it is assumed that the cell is a "well-mixed pool". In many situations, however, concentration gradients exist which will affect the local rate of biochemical reactions. In particular for large systems with different compartments, we must consider explicitly the effect of diffusion or transportation.

In BioPNML other bioprocesses classes can be defined. This concerns not only all effects of metabolites, but also different compartments and properties of biological processes.

3.5.3 An Example

In this section, we present some concrete XML syntax in order to exemplify the concepts discussed in the previous section by using a simple enzymatically catalyzed reaction (Figure 3.5.3). The model defines the single biochemical reaction from L-arginine to L-ornithine catalyzed with the enzyme arginase. We assume the reaction kinetics complies with the Michaelis-Menten equation, and the values of K_m and V_{max} are 0.5mM and 0.3mM respectively.

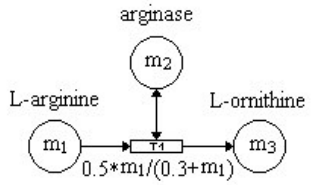


Figure 3.5.3 An example for biology Petri net model, where m_1, m_2 and m_3 are variables for the concentrations of the substances involved.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE net SYSTEM "BioPNML.dtd">
<BioPNML>
       <Petrinet id="pn1" type="Hybrid">
              <!--place-->
              <place id="p1" type="continuous">
                     <name id="C00062">
                            <text>L-arginine</text>
                            <value>m1</value>
                     </name>
                     <graphics>
                            <size>10</size>
                            <position x="-20" y="10"/>
                            <color>red</color>
                     </graphics>
                     <initialMarking>
                            <value>1</value>
                     </initialMarking>
                     <annotation>
                            <name>L-arginine</name>
                            <TypeRef>Amino acid</TypeRef>
                            <species>human</species>
                            <location>plasma</location>
                            <concentration>0.1mM</concentration>
                            <comment/>
                     </annotation>
              </place>
              <!--transition-->
              <transition id="t1" type="continuous">
                     <!--reaction-->
                     <PathRef>metabolic reaction</PathRef>
                     <reaction name="reaction_1" reversible="false">
                            <enzyme>arginase</enzyme>
                            <substrate stoichiometry="1">L-
arginine</substrate>
                            <product stoichiometry="1">L-ornithine</product>
                            <KineticsRef>
                                   <formula>0.5*m1/(0.3+m1)</formula>
                            </KineticsRef>
                            <thermodynamics/>
                     </reaction>
                     <graphics>
                            <size>10</size>
                            <position x="-30" y="0"/>
                            <color>yellow</color>
                     </graphics>
                     <annotation/>
              </transition>
              <!--arc-->
              <arc id="a1" source="p1" target="t1" type="normal">
                     <graphics>
                            <size>1</size>
                            <offset x="0" y="0"/>
                            <color>blue</color>
                     </graphics>
                     <weight>
                            <value>1</value>
                     </weight>
                     <annotation/>
              </arc>
<!--more places and arcs-->
…
       </Petrinet>
</BioPNML>
```

The first part of the XML example contains the biological information, whereas the second part is mainly PNML. idref tags are used to link the PNML 'place' and 'transition' tag to the respective SBML based 'species' tag. Since it was tried to develop BioPNML in a way that it should be readable both by existing PNML and SBML tools, some redundancies could not be avoided, i.e. the names of the compounds and the initial concentrations appear in both parts of the file. Properties which are not part of the present PNML standard, such as the formula used to calculate the changes in the concentrations of the substrates and the product, are only stored in the SBML part of the file. The example shows the basics of idea of BioPNML. In real applications, the PNML part may contain many reactions.

3.5.4 Discussion

BioPNML is a XML framework for the exchange and unification of molecular biological Petri net models. By formalizing the process of expressing bioprocess interchanges in a consistent and extendible way, BioPNML makes it easier for users and developers of biological software to map data in different formats. Easier mapping enables developers of biological software who are using open standards, such as XML, to adopt changes in biological data formats faster.

BioPNML defines a core set of XML elements, attributes, and tags that enable researchers to develop technologies that are optimized for data exchange. This XML based core data model is important because it eliminates the need to find a common application programming interface or implementation platform. Currently, its XML schema is based on the SBML and PNML standard. However, BioPNML is not static; we continue to develop it. BioPNML will be updated in line with future changes of SBML and PNML.

Extensions to Petri nets have been developed which transform Petri nets into a powerful tool for modeling biological systems. These enhancements include timing, token typing, non-homogeneous places, priorities and resources. It is possible to extend our BioPNML classes to these requirements by using additional tag sets.

BioPNML files can be generated computationally from existing data sources. Users can extract XML data from molecular biological databases via the Internet and transform them into BioPNML files via XSLT (Figure 3.5.4).

There are many approaches that address the challenging problem of interoperability among biological databases. They are based on different data integration techniques, e.g. federated database systems, multi database systems and data warehouses. In order to model and simulate gene controlled metabolic networks, we focus on a flexible and thin, but universally applicable solution with powerful query and retrieval capabilities. The architecture of our system MARGBench [http://cweb.uni-bielefeld.de/agbi/home/index.html?id=104] is a mediator-based approach for database integration. The aim of MARGBench is to support the

seamless integration of multiple heterogeneous molecular biology databases and to allow the development and the execution of global applications that extend beyond the boundaries of individual databases [Fre02a].

The general principle of BioPNML data integration is shown in Figure 3.5.4. Integration of heterogeneous and physically distributed databases is implemented by the BioDataServer (BDS) system, which provides a homogeneous database view. IIUDB (Individually Integrated User Database) accesses JDBC (Java Database Connectivity) interfaces followed up by an object network. Provided with the JDBC driver, the IIUDB is developed for users to define their own specific integrated schemes, i.e. the system is adaptive by connecting to heterogeneous databases and integrates the information retrieved into user-defined persistent databases and analyses the networks that can be found in these databases. The structure of metabolic networks and the molecular information contained is changing, and depending on the user view. Then based on the Object Management (OMG) architecture, we can do SQL queries and build up a metabolic network. IIUDB also includes several interfaces to export the resulting networks into common formats, e.g., CORBA, GML and XML as well as BioPNML.

So far, the IIUDB offers integrated access to biological databases, currently mainly to KEGG, BRENDA and RegulonDB, which cover considerable features including details on the enzymatic reactions, substrates and products, binding parameters, catalytic constants and gene regulations. Based on these techniques, bio-Petri net tools could be provided with models of metabolic pathways, gene regulation and signaling pathways.

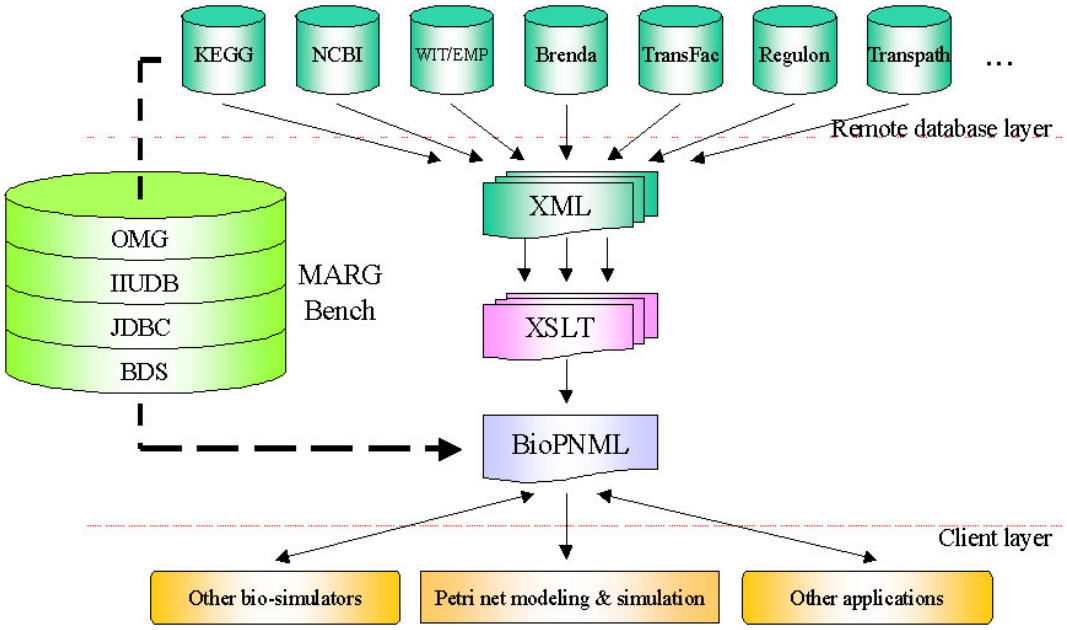


Figure 3.5.4 BioPNML data integration schema.

3.6 Summary

This chapter shows that the Petri net allows easy incorporation of qualitative insights into a pure mathematical model and adaptive identification and optimization of key parameters to fit system behaviors observed in gene regulated metabolic pathways. The advantages of the hybrid Petri nets applying to model and simulate are: The HPN model has a user-friendly graphical interface that allows an easy design, simulation and visualization. With the discrete and continuous events, the HPN can easily handle gene regulatory and metabolic reactions. The inhibitor arcs are useful for mechanistic studies to learn about how enzymes interact with their substrates, to know the role of inhibitors in enzyme regulation and gene expression. Moreover, powered with mathematical equations, simulation is executed and dynamic results are visualized.

As in the cell, there are usually hundreds of interconnected metabolic pathways and gene regulatory networks and control of these presents more complex features. It is feasible to extend the Petri net model with a plug-in way. A large complex network model can be handled with the same set of structural and behavioral properties. When applying to such a large one, the HPN model will be very complex and the hierarchical concept makes it possible to develop a generalized variant of HPN at a global level. On the other hand, the subnet of Petri net model provides us the basic model that we already know its inner behavior and functions. Then we can construct a system by plugging together sub-models and can understand the working of the higher-level system and are able to predict its behavior.

Building integrative models of the whole cell (virtual cell modeling) that incorporate gene regulation, metabolism and signaling is becoming a promising field during the post-genomic era. Several projects have been established under way. The challenge created with Petri nets is to understand how all the cellular proteins work collectively as a living system. Using powerful Petri nets and computer techniques, data of metabolic pathways, gene regulation, signaling pathways can be converted for Petri net destination application. Thus, a virtual cell Petri net model can be implemented; the attempt to understand the behavior of cell activity could be accomplished.

The aim of the BioPNML is to present a common data exchange format and to enable exchange of models between metabolic data and Petri net tools as well as other bio-simulators. It uses a simple, well-supported, textual substrate (XML) and can add components that reflect the natural conceptual constructs used by modelers in the domain. The ultimate purpose is to serve as a common framework for exchanging data about metabolic networks, and to provide guidance to researchers who are designing databases to store pathway and reaction data.

Obviously, in order to model a biopathway, we need a structural knowledge of the system. Nevertheless in reality, most often only parts of a system are known. Rudimentary

knowledge, such as sequences, involved metabolites or enzymes, can be examined by experimental work. The gap between rudimentary experimental data, and a satisfied model, should be overcomed. In the next chapter we are going to present a metabolic pathway prediction approach. It is developed as a web-based metabolic information retrieval and pathway reconstruction system. A predicted metabolic pathway can be assigned with kinetic values and automatically translated into XML data format that can be parsed by some Petri net tools for further modeling and simulation.

Chapter 4

In Silico Prediction of Metabolic Pathways*

In order to model and simulate a metabolic network, the more information available the better. Fortunately, there are more than 500 database systems available that represent molecular data. However, for the analysis of complex metabolic networks only rudimentary data and knowledge are available today. Therefore, we have to develop and implement special algorithms for the analysis and synthesis of complex metabolic networks, which are able to complete the rudimentary data. Previous approaches and existing metabolic pathway databases have a number of limitations in metabolic pathway reconstruction. Some present knowledge of the genome alone does not contain comprehensive information about metabolic pathways, such as physical and chemical properties of the enzymes that are involved. Some are not fully computer-aided. The individual database search process requires too much human intervention and the quality of annotation, largely depends on the knowledge and work behavior of human experts. The aim of this chapter is to develop such a web-based information retrieval system that will help in the prediction of metabolic pathways

4.1 Introduction

In silico retrieval/reconstruction of metabolic pathways based on the information of genes, enzymes and metabolites requires access to suitable databases ranging from genomics to metabolics. In Table 2.2 we have listed some major databases that make the integrative information retrieval of metabolic pathways possible. Here we describe some of them in more

* Part of Chapter 4 is to be published in *IEEE Transactions on Nano-Bioscience* [Che04a].

detail. GenBank [Ben03] is a database that contains an annotated collection of all publicly available DNA sequences. Internet access is provided through several interfaces directly from the NCBI web pages. Each sequence is linked to other sequences that are similar based on sequence alignments. Swiss-Prot [Boe03] is a curated protein database that provides a protein retrieval interface that can be searched by AC, ID, description, gene name, organism and more. Several mirror sites of Swiss-Prot are distributed in Europe, America and Asia. BRENDA [Sch02b] systematically collects enzyme data. It is essential for both interpretation of the kinetic aspects of enzymatic reactions and retrieval of enzymes by various query terms.

Metabolic pathway databases such as KEGG [Kan02], WIT/EMP [Ove00] and EcoCyc/MetaCyc [Kar02] have been developed to present diagrams depicting metabolic pathways. KEGG is composed of three interconnected sections: genes, molecules, and pathways. It represents the data of interacting molecules or genes by using the simplest form of representation: binary relations that correspond to pairwise interactions. It provides both an online map of metabolic pathways and the ability to focus on metabolic reactions in specific organisms. WIT/EMP includes some 3000-pathway diagrams covering primary and secondary metabolism, membrane transport, signal transduction pathways, intracellular traffic, translation, and transcription. Initially, EcoCyc/MetaCyc described only metabolic pathways. Now it is extended towards an integrative information system that represents genes (sequences, function), enzyme (amino acids, function and structure), and metabolic pathways of *E. coli* [Kar99]. Figure 4.1 shows the databases that make the integrative information retrieval of metabolic pathways possible.

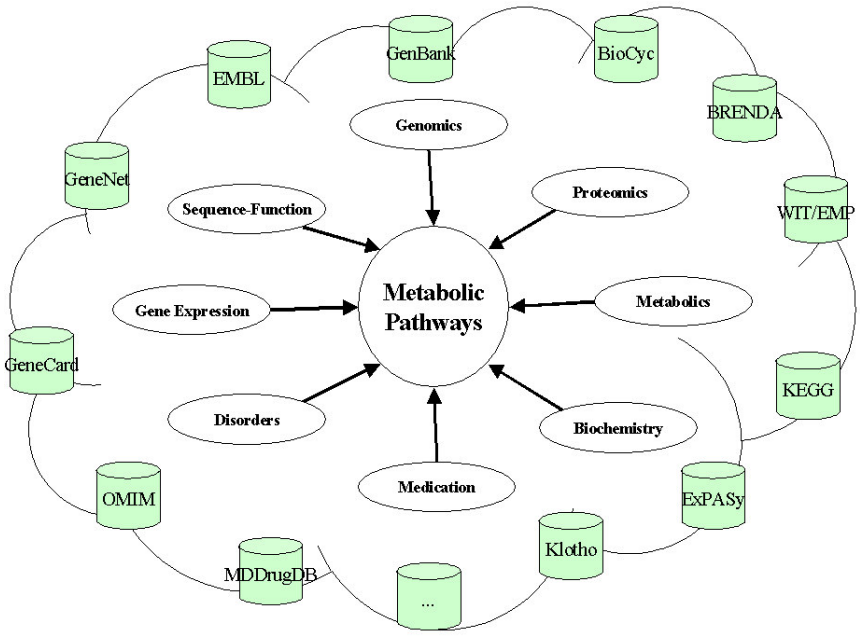


Figure 4.1 A schematic diagram of information sources for metabolic pathways prediction.

We exploited these databases for the construction of our pool of metabolic pathway datasets that are at present mainly based on KEGG and MetaCyc. Other databases, such as PIR [McG00], PDB [Ber02] and TRANSPATH [Kru03] / TRANSFAC [Mat03b] can also be potentially utilized in our future research for protein information, molecular structure and gene regulation and signal transduction. The pool database consists of 623 metabolic pathways of *E.coli*. Data servers handle the access (storage and retrieval). While the use of an up-to-date metabolic pathway database is essential to any similarity search. The pool database is constantly being updated. In the future, a more powerful integrated metabolic pathway database system, BioDataSever [Fre02b], which contains all metabolic pathway data from KEGG, WIT and MetaCyc will support our system.

With the achievement of biological data collection, in silico metabolic pathway retrieval/reconstruction and sophisticate analysis becomes possible. Here we limit our discussion to sequence analysis. Suppose that we have a set of sequences $S=\{s_1,s_2,\ldots s_n\}$. It is most widely used to search for each sequence similarity against the sequence databases such as GenBank. If there is a strong evidence in terms of sequence similarity, we may conclude that s_i belongs to a certain protein family or other similar genes with a known function. Obviously, one of the main problems with the database search strategy is that the search result needs to be evaluated manually by human experts. Although there are several integration systems, such as SRS [Etz96], available to realize the data query process, it still requires much human intervention, and the quality of annotation largely depends on the knowledge and skills of human experts. Moreover, scientists have to invest extensive efforts to learn how to use all different database interfaces, query languages, and parameter specifications for specific analytical programs. On the other side, for the prediction of metabolic pathways from rudimentary data, powerful tools are still missing. Biologists wish to perform metabolic pathway prediction and analysis with local or Internet-based tools.

4.2 Methods and System

In an attempt to answer these questions, a web-based information retrieval system is proposed. The system would at least include an Internet-based client/server architecture that allows remote and local access to the system. The main benefit of building such a web-based system is that it exploits the results of the existing databases on the web, and meanwhile acts as a virtual environment that allows the access to remote databases using Internet resources. Internet mechanisms support and maintain communication between web-browsers and database shells. The system is not transparent to the users. They do not need to known anything about how the system processed their problems.

We discuss relevant issues for conducting sophisticated metabolic pathway reconstruction and metabolic information retrieval. The basic methodology used to reconstruct

metabolic pathways is retrieving all related EC numbers and searching or aligning against our pool database. This process selects all metabolic pathways from our database that currently embraces all pathways from MetaCyc.

4.2.1 Pathway Reconstruction Method

We consider a metabolic pathway as a special case of a metabolic network with distinct start and end-points, initial and terminal vertices, respectively, and a unique path between them [For99].

Let $M=\{m_1,m_2,...,m_n\}$ be a set of metabolites that are involved in the enzymatic reactions acting as substrates and products; $E=\{e_1,e_2,...e_m\}$ be a set of enzymes. Normally a 3-tuple (M,E,A) is called linear metabolic pathway, where $A=\{(m_i,e_j) \cup (e_j,m_{i+1}) \mid 1\leq i\leq n,\ 1\leq j\leq m\}$ is the subset of the successive relationship between M and E.

The enzymes normally are separate enzymes. For those enzymes that can form a multienzyme complex (noncovalent aggregates of enzymes) or may be a membrane-bound system, we can choose a representative enzyme unless there is a unique term for it. However, in many of the metabolic reactions in living cells, enzymes act as catalysts in the conversion of certain metabolites (substrates) into other metabolites (products). So enzymes are the cores of metabolism and make the whole cellular processes connected, and the metabolic network can be interpreted as sets of enzyme catalyzed biochemical reactions. The representation of a metabolic pathway might be given as a set of successive related E. In addition, E can be a set of enzyme names or the corresponding 4-hierarchical-level EC numbers.

The problem studied here can be stated formally as follows (Figure 4.2.1):

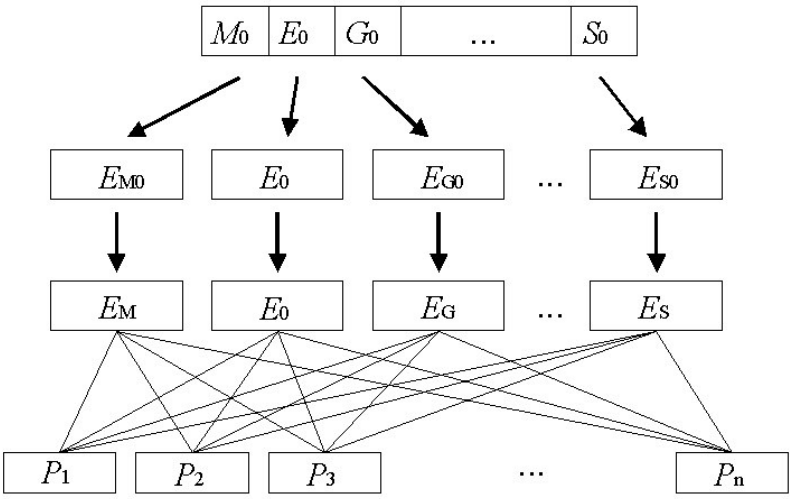


Figure 4.2.1 The concept of functional pathway prediction.

The aim is to use a rudimentary metabolic pathway, e.g. a list of metabolites ($\boldsymbol{M_0}$), enzymes ($\boldsymbol{E_0}$), genes ($\boldsymbol{G_0}$) and sequences ($\boldsymbol{S_0}$), either nucleotide or amino acid sequences or both, in order to retrieve all relevant EC numbers. When they are queried to the remote data resources, then sets of associated EC numbers are extracted:

$$\boldsymbol{M_0} \Rightarrow \boldsymbol{E_{M0}} = \{E_{m1}, E_{m2}, E_{m3}, ..., E_{mi}\};$$

$$\boldsymbol{G_0} \Rightarrow \boldsymbol{E_{G0}} = \{E_{g1}, E_{g2}, E_{g3}, ..., E_{gj}\};$$

$$\boldsymbol{S_0} \Rightarrow \boldsymbol{E_{S0}} = \{E_{s1}, E_{s2}, E_{s3}, ..., E_{sk}\};$$

where E_{mi} is a set of EC numbers related to the m_i ($m_i \in M_0$); E_{gj} is a set of EC numbers related to the g_j ($g_j \in G_0$); E_{sk} is a set of EC numbers related to the s_k ($s_k \in S_0$).

Example 4.1 Given ($\boldsymbol{M_0}$,$\boldsymbol{E_0}$,$\boldsymbol{G_0}$), where $\boldsymbol{M_0}$={L-arginine, L-ornithine}, $\boldsymbol{E_0}$={6.3.4.5}, $\boldsymbol{G_0}$={ASL, OTC}, then $\boldsymbol{E_{M0}}$={(1.5.1.11, 1.5.1.19, 1.13.12.1, 1.14.13.39, 2.1.4.1, 2.1.4.2, 2.3.1.109, 2.4.2.31, 2.7.3.3, 3.2.2.19, 3.4.17.3, 3.4.17.10, 3.5.3.1, 3.5.3.6, 4.1.1.19, 4.3.2.1, 5.1.1.9, 6.1.1.19, 6.3.2.24), (1.5.1.24, 2.1.3.3, 2.1.4.1, 2.1.4.2, 2.3.1.35, 2.3.1.127, 2.6.1.13, 2.6.1.68, 3.5.1.16, 3.5.1.20, 3.5.3.1, 4.1.1.17, 4.3.1.12, 5.1.1.12)} and $\boldsymbol{E_{G0}}$={(4.3.2.1), (2.1.3.3)}.

Then the sets of associated EC numbers are combined to produce a new list of sets of EC numbers. That is

$$\boldsymbol{E_M} = \{E_{m1} \times E_{m2} \times E_{m3} \times ... \times E_{mi}\};$$

$$\boldsymbol{E_G} = \{E_{g1} \times E_{g2} \times E_{g3} \times ... \times E_{gj}\};$$

$$\boldsymbol{E_S} = \{E_{s1} \times E_{s2} \times E_{s3} \times ... \times E_{sk}\};$$

In Example 4.1 we get:

$\boldsymbol{E_M}$={(1.5.1.11, 1.5.1.24), (1.5.1.11, 2.1.3.3), …, (1.5.1.11, 5.1.1.12), (1.5.1.19, 1.5.1.24), (1.5.1.19, 2.1.3.3), …, (1.13.12.1, 1.5.1.24), …, (1.14.13.39, 1.5.1.24), …, (6.3.2.24, 1.5.1.24), …, (6.3.2.24, 5.1.1.12)};

$$\boldsymbol{E_G} = \{(4.3.2.1, 2.1.3.3)\};$$

Now we select the set elements of $\boldsymbol{E_M}$, $\boldsymbol{E_G}$ and $\boldsymbol{E_S}$ to perform a combinatorial operation. The results are a set of possible pathways $\boldsymbol{P}$,

$$\boldsymbol{P} = \boldsymbol{E_M} + \boldsymbol{E_0} + \boldsymbol{E_G} + \boldsymbol{E_S} = \{E_{mi} \cup E_{0j} \cup E_{gk} \cup E_{sl}\},$$

where $E_{mi} \in \boldsymbol{E_M}$; $E_{oj} \in \boldsymbol{E_0}$; $E_{gk} \in \boldsymbol{E_G}$; $E_{sl} \in \boldsymbol{E_S}$.

Finally we have:

$$\boldsymbol{P} = \{P_1, P_2, ... P_n\},$$

where $P_i = \{e_{i1}, e_{i2}, ..., e_{ik} \mid 1 \le i \le n\}$ is a set of EC number.

When we continue Example 1, we have a set of pathways $\boldsymbol{P}$,

P={(1.5.1.11, 1.5.1.24, 6.3.4.5, 4.3.2.1, 2.1.3.3), (1.5.1.11, 2.1.3.3, 6.3.4.5, 4.3.2.1, 2.1.3.3), …, (6.3.2.24, 5.1.1.12, 6.3.4.5, 4.3.2.1, 2.1.3.3)}.

P_i is then searched against a pool database to find the metabolic pathway with the highest similarity score.

To execute such data retrieval and combinatorial problems, an algorithm can be specified, which takes online data query and calculates an integrated relation over all specified data sources, using the already defined combinatorial operations. The algorithm outlined above can now be expanded into the following pseudo-code:

```
Begin: a query with rudimentary element set of M0, E0, S0, G0
      While (all requests are not processed) loop
            while (the Metabolite queue is not empty) loop
                  process M0 request: M0=>EM0
            end loop
            while (the Sequence queue is not empty) loop
                  process S0 request: S0=>(M0')=>ES0
            end loop
            while (the Gene queue is not empty) loop
                  process G0 request: G0=>EG0
            end loop
            recombine EM0, ES0, EG0 => EM, ES, EG
      end loop
      recombine EM, ES, EG, E0 => P
      search P against a pool database
End: a metabolic pathway predicted
```

In the algorithm, all the three different types of requests are processed in batches. After processing the queued metabolite requests, batches of sequence requests are processed, before processing the demands of gene requests. Here, once the control has passed over to the sequence loop, all retrieved metabolite requests (***M*₀'**) will get processed with the metabolite loop before EC number finding is resumed. The main drawback is that all requests are web-communication depended. However, there is no problem with the data being out of date because it queries databases remotely instead of locally. The accessibility and update are guaranteed since the databases are global oriented and maintained by reputed institutes. The computing time cost is largely depended on the Internet communication. While in the combinatorial part, suppose there are k rudemantory elements, each element retrieve n EC numbers, then the combination costs n^k. The combination of all assoiated EC numbers cost n^k times n^k. Therefor the total complexity is of order $O(n^{2k})$.

4.2.2 Web-based Metabolic Data Retrieval

4.2.2.1 PathAligner system architecture

PathAligner is a web-based biological information retrieval system designed with one main purpose: the retrieval and alignment of metabolic pathways. The PathAligner system contains a PathModeler and a pathway alignment tool (Figure 4.2.2.1). PathModeler consists of four parts. The first part is a database mining tool that pulls out potential metabolic relationships from various databases, based on the queried rudimentary components such as metabolites, genes, sequences, etc. It allows easy access to distributed heterogeneous biological resources through a simple interface. The relationships are then organized and recombined, and queried against metabolic pathway database to retrieve a metabolic pathway result. Genetic and metabolic information involved in the retrieved pathway are extracted and displayed in the second part. In the third part, the retrieved metabolic information is visualized using an interactive graph display module. Finally, a XML data file that contains the basic information of metabolic and regulatory network as well as their kinetic values is formed for further modeling and analysis.

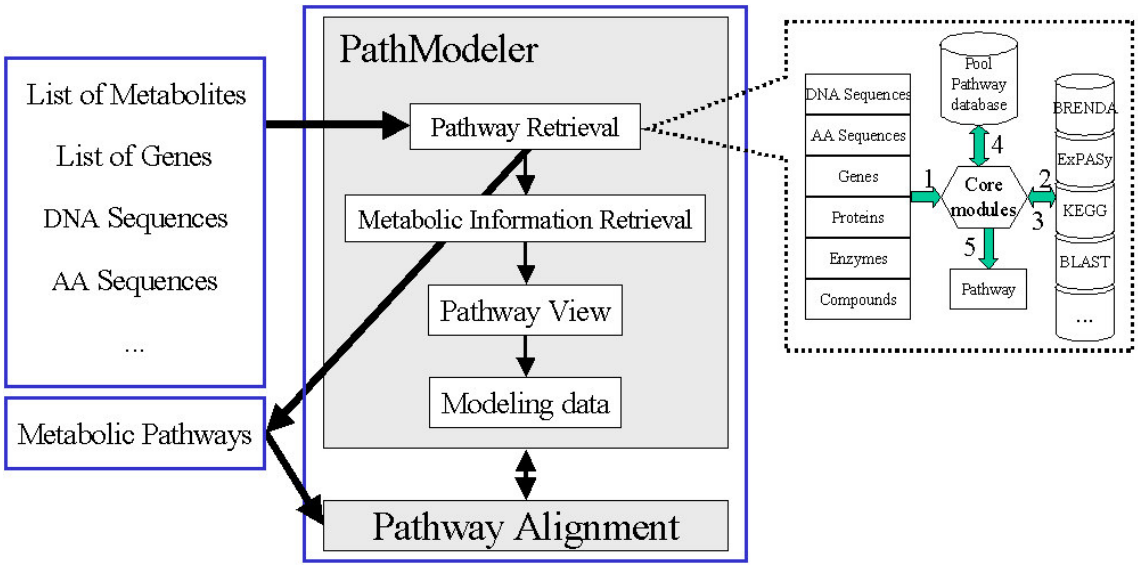


Figure 4.2.2.1 The concept of PathAligner system. An initial set of rudimentary components such as metabolites, genes and sequences (left side) are submitted to the "Pathway Retrieval". The core modules will recognize all components and make queries against relevant remote databases such as BLAST, KEGG and ExPASy. Then they pull out potential metabolic relationships from these databases, and search through a pool metabolic pathway database to predict a metabolic pathway. The result pathway can be further analyzed to retrieve other metabolic information, such as kinetic values of enzymatic reactions. A graphical model can be constructed based on the retrieved information afterwards. The result pathway can also be aligned with other pathways by using a "Pathway Alignment" tool.

By introducing a remote access communication architecture, the system allows different distributed heterogeneous biological resources communicating through the same common easy-to-use web-interface and enables researchers to perform efficient and effective biological data retrieval and metabolic pathway reconstruction. The processes of the core modules are for distributing and locating the responding database system to answer user's local queries via the web-interface. It has a single common data representation to handle the diverse range of biological data formats, and the ever increasing amount of bioinformatics data can be accessed and analyzed in a synchronous, integrated manner, allowing biologists to concentrate on gathering and analysis of data and relieving them of the burden of learning and utilizing individual stand alone tools. The prototype system is designed to handle enzymes, proteins, metabolites, as well as incomplete or fragments of gene/protein sequences via the Internet. We have also constructed a pool pathway database of known metabolic reactions from several online databases such as EcoCyc/MetaCyc, regarding the metabolism of *E. coli* and other organisms. It contains metabolic pathways with EC numbers.

4.2.2.2 System workflow

The PathAligner work differs from previous attempts due to a combination of system design decisions. PathAligner is oriented toward assisting biologists in retrieving and reconstructing metabolic pathways rather than fully automatic construction and storage, thus avoiding information retrieval precision limitations. The procedures of PathAligner toward reconstructing the metabolic network are:

Step 1: User input. Keyboard input of rudimentary pathway components of interest by the user. Components range from gene names, genomic sequences, enzymes, EC numbers, other compounds and more.

Step 2: Component identification. Classify all components and query their responding databases. For example, nucleotide and protein sequences are queried by BLAST, proteins and other compounds are searched against Swiss-Port, and so on.

Step 3: Data retrieval. The nucleotide sequences and protein sequences are aligned against the BLAST, with the aim of identifying the aligned encoded proteins. Other rudimentary metabolic components such as compounds and proteins are searched against Swiss-Port that provides the richest information on enzymatic reactions. The problem of synonymy and polysemy is solvable by using Swiss-Port search engineering to obtain all enzyme EC numbers by remote retrievals.

Step 4: Pathway building. After the relevant EC numbers to all components of the rudimentary metabolic pathway are retrieved, a set of rough pathways expressed as sequences of EC numbers are combined.

Step 5: Pathway identification. Using our pool pathway database, map the assigned pathways (sequence of EC numbers) to find the one with highest similarity.

The pathways can be modified afterwards. Finally, the reconstructed metabolic pathways (network) with kinetic values, which might be obtained from BRENDA or from literatures if any, are ready for modeling and simulation.

4.3 System Implementation

4.3.1 Perl Scripts

Perl originally was designed to be able to easily process text files. Powered by the Internet connection ability, it becomes the leading language in everything regarding text data mining. The process can range from simple rearranging of the information to heavy statistical analysis. The way Perl scripts are capable to grab data from the web and manipulate it, is a Perl module, which is effectively an optional, very specialized set of Perl commands. One of the particular Perl modules is called LWP, short for "Library for WWW access for Perl". LWP is a collection of programs and programming tools to allow surfing the web from inside your programs. In general, a "request" is created for data from the web, give it to a "user agent" which will actually make the request, coordinate the transfer of data, etc., and in return receive a "response" (called a response object by LWP) is received. An example of using LWP to retrieve a single protein entry from NCBI web site is shown below.

```
# This is a Perl program to retrieve a single protein entry from the
entrez web site.
#!/usr/bin/perl -w

use LWP;
# This tells Perl you want to use the web access modules
use strict;

my $url = "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=search&db=
nucleotide&dispmax=100&term=OTC&doptcmdl=FASTA";
# Here we set a variable with the class URL

my $agent = LWP::UserAgent->new;
# This initializes the LWP system

my $request = HTTP::Request->new(GET => $url);
# Here we create an HTTP GET request

my $response = $agent->request($request);
# Give it to the user agent

$response->is_success or die "failed";
# Get request back & check if agent did the job correctly. Then, print
# the result:

print $response->content;
```

Users can specify the parameters "cmd", "db" and so on according to their purposes. The precise form of the query to Medline or any other NCBI database, including GenBank, is detailed at: http://www.ncbi.nlm.nih.gov/entrez/query/static/linking.html. In this program, we

just print out the whole big HTML file, although we could choose to parse and extract out information and analyze it. However, we can often figure out the form of the URL just by looking something up in a database, then noting the address of the web page with the data. A brief tour of some of the biological data on the web available for our program is listed below.

Table 4.3.1 A collection of biological database for the construction of metabolic pathways*.

Database	Records	The URL and its parameters (example)
KEGG	Most known pathways, in 151 graphical diagrams and 78 ortholog group tables	http://www.genome.ad.jp/dbget-bin/www_bfind?reaction or ?compound or ?ligand
SWISS-PROT	129,768 sequence entries from 8,202 species	http://www.expasy.org/cgi-bin/enzyme-search-ec or /enzyme-search-ca
EcoCyc/MetaCyc	173 pathways/ 150 species	http://biocyc.org/META/substring-search?type=ENZYME&object=
Genbank	18,197,000 sequence records	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&cmd=search&term= or http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=
TRANSFAC®/ TRANSPATH®	5,241 factors; 12,976 sites / 12,262 molecules; 2,604 genes	http://www.biobase.de/cgi-bin/biobase/TRANSFAC/8.1/bin/getTFProf.cgi?
BRENDA	3635 EC numbers	http://www.brenda.uni-koeln.de/php/result_flat.php4?ecno=

* Last Statistic: July 1, 2003

This very simple approach could easily be the basis for a program to consult biological databases and to map the Internet in real time. Moreover, some of these computational works are provided by Bioperl association (http://www.bioperl.org). Bioperl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatics applications. It is the leading open source project. It contains modules for representing biological sequences, protein structure, sequence alignments, BLAST and FASTA reports, biological maps, sequence features and their locations including complex locations, annotations & bibliographic references, phylogenetic trees, and gene structures. Provided by various Bioperl module, it is becoming more and more easier to automate your request results with a desirable format.

4.3.2 Web Interface

PathAligner provides an easy-to-use interface environment to access the related heterogeneous databases, analysis and display results. A web-based interface of PathAligner system has been established to implement the retrieval of metabolic pathways. The web-interface is responsible for the communication with the client queries. It receives the query in terms of an HTTP request. After parsing the request, it triggers the corresponding functionality of the query engine that processes the query and returns the result. Then the result for each query and the protocols between them are returned as HTML data to be displayed in a browser. The result includes the responding metabolic pathway, and some URL links to the original databases and

a graph of the metabolic pathway. There is no problem with the data being out of date because it queries databases remotely instead of locally. Although the procedure might take a little while to retrieve the data, the accessibility and update are guaranteed since those databases are global oriented and maintained by reputed institutes.

Users access the PathAligner system via the web-interface, while standard, platform independent Perl applications and modules are used to connect the applications to the central database and external data sources. Using web-browsers, users will not need special hardware or software to consult these services. The PathAligner home page is located at http://bibiserv.techfak.uni-bielefeld.de/pathaligner.

4.4 Applications

An example of PathAligner usage is retrieval of a metabolic pathway using several initial rudimentary components: Metabolite (L-citrulline), Enzyme (arginase//4.3.2.1), DNA sequence (ctgtgttcactg…), protein sequence (mtkdfrqnvfq…) and gene symbol (OTC). PathAligner retrieves all relevant EC numbers from various public databases and searches pathways against the pool pathway database. The example query and its query result is shown in Figure 4.4A.

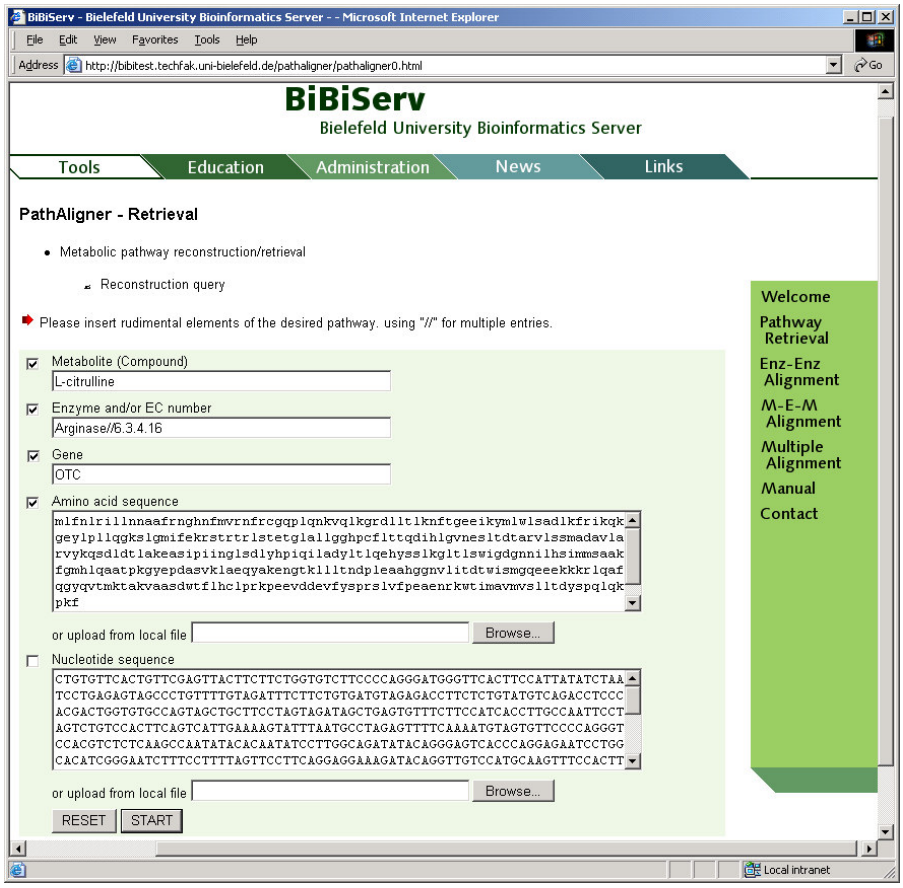


Figure 4.4A A pathway retrieval example in PathAligner.

The highest scoring metabolic pathway "urea cycle pathway": {6.3.4.16, 2.1.3.3, 6.3.4.5, 4.3.2.1, 3.5.3.1} is retrieved and displayed (Figure 4.4B).

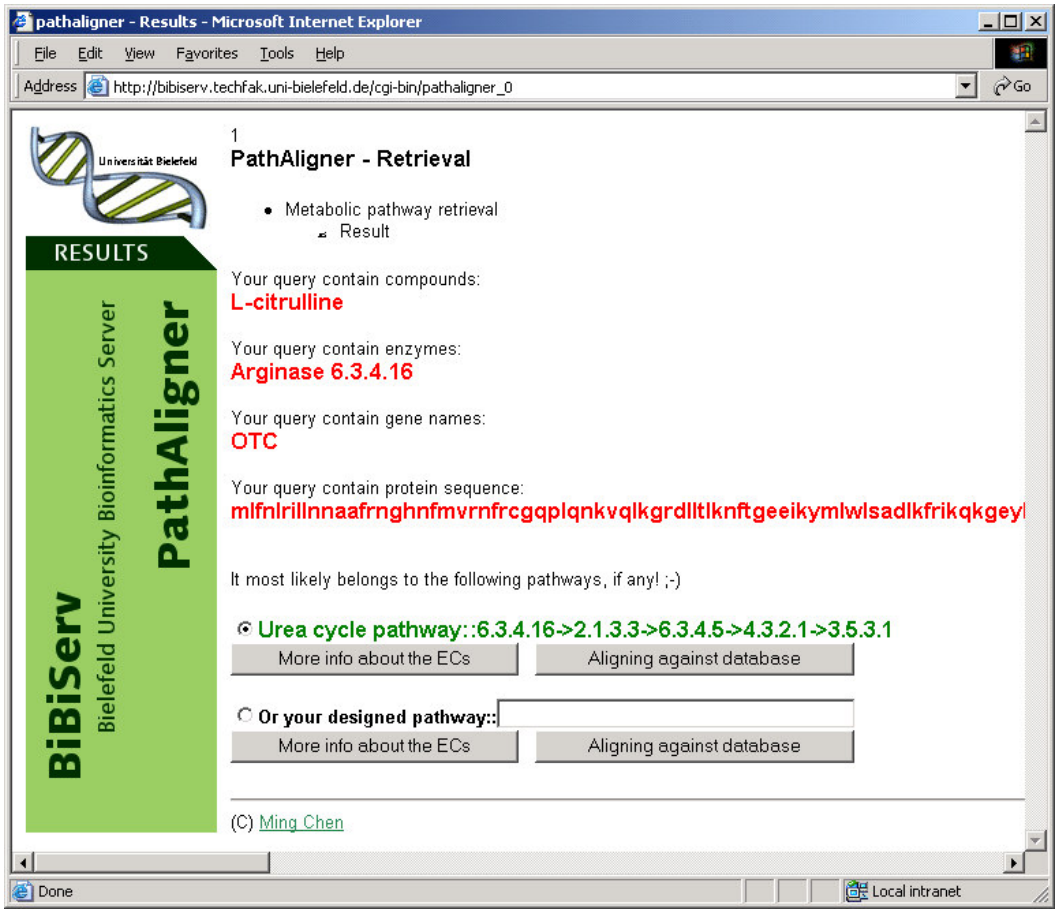


Figure 4.4B An example query of the PathAligner for pathway retrieval.

By clicking the button "More info about the ECs" under the retrieved pathway, additional information about the involved enzymes and enzyme-associated pathways are displayed (Figure 4.4C). The table lists not only the enzymes that are involved in the query, but also more related metabolic and genetic information. Clicking the corresponding hyperlinks can retrieve additional information about the enzymatic reactions and K_m values. The K_m values and the reaction data are retrieved from BRENDA. The encoding genes and their transcription factors are also displayed. The genes involved, as well as the pathways associated, are obtained from KEGG; while the factors and GeNetView are extracted from BioBase. For instance, for EC 2.1.3.3, is encoded by gene OTC and a number of transcription factors which are shown in the column Factor. Moreover, the interactions between the genes and the transcription factors are also available by clicking the hyperlink OTC in the column GeNetView. However, not all data is available due to incompleteness of the source database. In the current version of PathAligner, this additional information is restricted to *Homo sapiens*.

Future work will be extended to various species. The enzyme associated metabolic pathways out KEGG are also presented.

Figure 4.4C Screenshot of the representation of metabolic/genetic information of the retrieved enzymes.

PathAligner provides a graphic representation that illustrates the interrelationships between the retrieved enzymes and their metabolic reactions and gene regulations. The graph visualization is based on visualizing and interacting with dynamic information spaces. The graph layout program is *dot*, which as a part of the *Graphviz* [Gan00] program, developed at AT&T (http://www.research.att.com/sw/tools/graphviz/). The web-interface uses the layout and graphics engine to transform the graph into a picture and delivers it to the client as a PNG image-file (Figure 4.4D). The graph may just be too large to be viewed as a whole on the screen. The user can resize the graph to examine different parts of the graph in varying levels of detail.

PathAligner models the initial rudimentary pathway so that important relationships can be retrieved and illustrated. The retrieved functional data provide a basis for further

analysis. For instance, the graph can be used as a blueprint for modeling and simulation with some biological simulators such as Petri net tools. A hybrid Petri net model that contains qualitative and quantitative aspects can be used as a predictive tool. As quantitative model requires kinetic values, assignment of initial value of metabolites and kinetics are to be made after data retrieval. The table in the Figure 4.4D requires the user to fill in the blanks with actual concentrations of substrates and products, and K_m values involved in responding enzymatic reactions. Although some of such data nowadays are available, some other data may not be complete. A series of predefined kinetic types that are most often used in the biochemical reaction models are available in the literatures. However, there are some circumstances in which the kinetic types are not yet defined. Then a new kinetic type is to be self-defined by the mass law. In that case, mass conservation relations must be taken into account in order to carry out the simulation. In principle, we construct the model with the identification of the reaction stoichiometry.

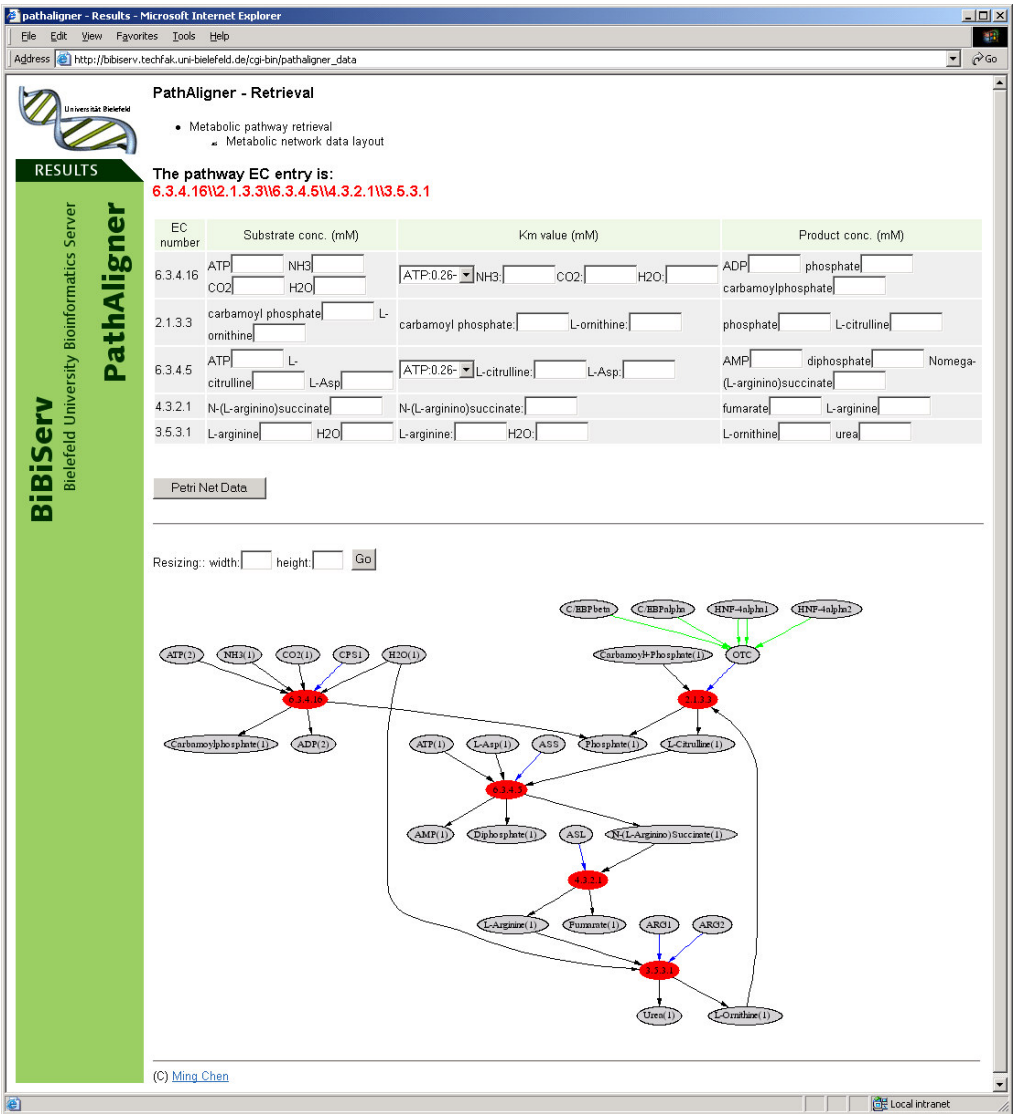


Figure 4.4D Screenshot of a network graph. The bold arrows represent catalyst links. The green arrows are gene regulation links. The blue arrows are gene-encoding links. Enzymes are shown as red ellipses. The graph is resizable. The table above the graph indicates an initial kinetic value assignment. These data are intended for modeling. Some K_m values are extracted from BRENDA.

After the assignment of concentration and kinetic values to the reconstructed metabolic pathway, a data file is generated for storage and interchange. We propose it to be specified in an XML format, BioPNML (see §3.5 in Chaper 3). Figure 4.4E shows the web-layout of BioPNML data for the retrieved and value-assigned metabolic pathway. The BioPNML is designed to provide a starting point for the development of a standard interchange format for Bioinformatics and Petri nets. The language will make it possible to present biology Petri net diagrams between all supported hardware platforms and versions. It is also designed to associate Petri net models and other known metabolic simulators. PathAligner provides a translation tool to transform BioPNML into other XMLs.

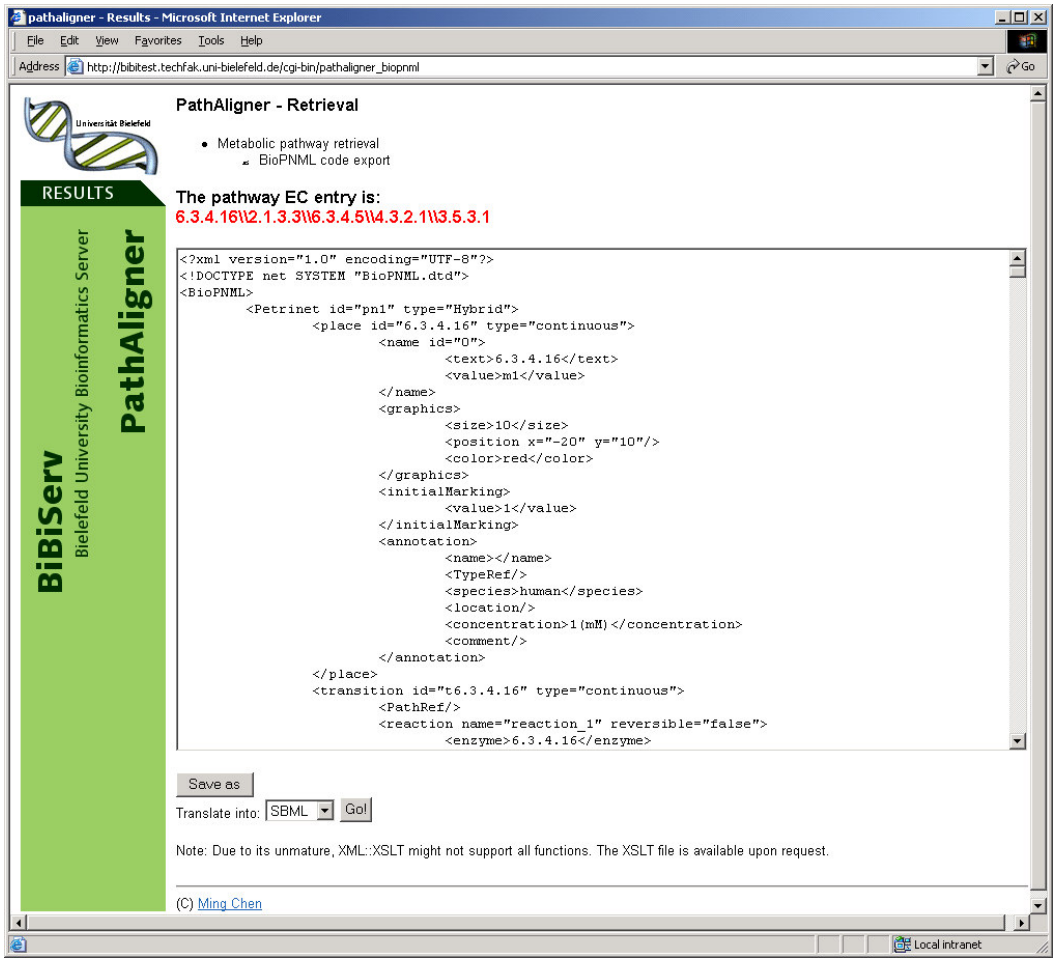


Figure 4.4E BioPNML presentation of the retrieved metabolic pathway.

4.5 Evaluation

The object of the evaluation is to verify the usability of the system. Some evaluation exercises were undertaken and the results are summarized:

- To determine the availability of information retrieval, evaluation is done programmatically by assessing the accessibility of URL sources. We examined the source code for URLs of different remote databases, allowing a more complete assessment.
- To evaluate the accuracy of the prediction, we compared automatic searches with manual searches against various data sources, and then checked back the predicted results to find the related genomic and metabolic information. Experiment results show that the predicted pathway consistently contains the known pathway.
- To evaluate the comprehensiveness of our approach, we chose fragments of sequences, genes and metabolites to perform metabolic pathway prediction. We observed that our approach is quite versatile in the sense that it can handle a variety of rudimentary elements. Most of pervious approaches can either only accept queries of metabolites/enzymes, or only require annotated sequences.
- In terms of compatibility, several different web-browsers were tested. There are no significant differences between Internet Explorer, Netscape and Mozilla. The user interface of the systems is quite simple and very user-friendly. It starts with the main query page; users can follow the web annotation to perform further steps. The design is kept simple for clarity.

PathAligner is a web-based information retrieval tool and an alignment tool. The following table is created to compare the features of PathAligner with other databases and tools.

Table 4.5 Comparison of PathAligner with related approaches.

Features	KEGG/WIT	PathoLogic	PathFinder	PathMiner	PathAligner
Algorithm	EC numbering, genome annotation	Genome annotation	Annotation data parsing	Heuristic search	Web-based information retrieval
Access	WWW	Local installation	WWW	WWW (Java applet)	WWW
Input	Molecules, enzymes	Specific files	Sequences, enzymes	Specific files	Rudimentary data
Output	Pathway	PGDB form	Pathway	Unknown	Pathway
Pathway visualization	+	Database	+	Unknown	+
Extra linkage	+	+	-	-	+
Alignment possibility	-	-	-	-	+
User interface	Dialog	Complex	Dialog	Menu	Dialog

4.6 Summary

Modern biology requires rapid development of new methodologies and algorithms in order to make an optimal use of intelligent computational tools. Our work has lead to the development of a web-based biological information retrieval system that exhibits an ability to reconstruct metabolic pathways.

This chapter demonstrated how the PathAligner system implements metabolic pathway reconstruction problems in a simple way, and significantly reduces the effort and difficulty involved in data integration and analysis. PathAligner is designed to handle metabolites, enzymes, proteins as well as incomplete or fragments of gene sequences, it handles nucleotide and protein sequences from the GenBank, proteins and metabolites from the ExPASy. By introducing a remote Internet access communications architecture, the ever increasing amount of metabolic biological related data can be accessed and analyzed in a synchronous, integrated manner, allowing biologists to concentrate on gathering and analysis of metabolic data, and relieving them of the burden of learning and utilizing individual stand alone tools.

In silico metabolic pathway reconstruction from rudimentary components requires combining information from a large number of sources: classical biochemistry, genomics, functional genomics (e.g. microarray experiments). As ever more experimental biological data are generated and analysis tools are developed and accessible to us, the expansibility of PathAligner system, via simple addition of modules that would allow the system to incorporate new technologies such as molecules' physical and chemical properties and microarray data analysis, become possible.

In summary, PathAligner is such a web-based tool for metabolic pathway retrieval. It possess the following operation features:

- It has a simple user interface.
- The web-based system requires no additional hardware or software. Users do not need to known anything about how the system processed their problems by using the system itself, or remote system through remote accessing techniques or communication protocols.
- No problem with data updating. With distributed biological and biomedical data sources supporting online, PathAligner facilitates and retrieves active participation of all data sources.
- Visualization. PathAligner presents the retrieved pathway using a graph visualization tool. Results are directly processed as web page layout. PathAligner implements the alignment algorithm and provides a graphical representation.

PathAligner focuses efforts on reconstructing metabolic pathways from diverse rudimentary components. It also provides a method to analysis the similarity and distance of different metabolic pathways. Assuming that the two pathways are related in some biologically meaningful way, whether from different organisms or the same, it is capable to discover their regulation and their evolution (to be discussed in Chapter 5).

Chapter 5
Metabolic Pathway Alignment*

In this Chapter a formal definition of metabolic pathway is given. Metabolic pathway alignment algorithms are presented and discussed. Alignment examples are demonstrated by the PathAligner system.

5.1 Metabolic Pathway Definitions

Traditionally biochemical pathways have been defined in the context of their historical discovery, often named after key molecules (e.g. “glycolysis”, “urea cycle”, “pentose phosphate pathway” and “citric acid cycle” and so on). In that context, pathway refers to a path from substrate to product or receptor to transcription factor, even though usually some of the molecules involved are also found in other pathways. The classification of molecules into pathways has historical reasons, and is not explicitly based on qualitative differences of the interactions. One reason for this is that it is easier to think about the network as pathways which are connected densely. But we should emphasize that the definition of a metabolic pathway is not exact, there are always interactions among pathways. A pathway's substrates are usually the products of another pathway, and there are junctions where pathways meet or cross. Now, the question is how to define a boundary for a pathway under these circumstances that the biology processes are so interacted and actually there is no such clear boundary between two pathways (glycolysis and urea cycle pathway, or MAPKinase signaling pathway and p38 MAPK signaling pathway).

Normally, the basic strategy to represent and compute pathways is the reactant-product binary relation. Properties of the pathway that rely upon the integration of two or more

* Part of Chapter 5 is to be published in *Applied Bioinformatics* [Che04b] and *BGRS’04* [Hof04]

input molecules and unrelated output molecules, and feedback effects are ignored (Figure 5.1A).

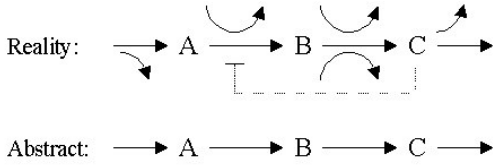


Figure 5.1A Abstract metabolic pathway with binary relation compare to that in reality.

Obviously, a metabolic pathway is a special part of complex network of reactants, products and enzymes with multiple interconnections representing reactions and regulation. Metabolic pathways are defined in the literature [Voe95] [Har97] [Sel98] in various ways with varying degrees of formality. A biochemical pathway is defined by Mavrovouniotis M.L. [Mav95] as an abstraction of a subset of intricate networks in the soup of interacting biomolecules. A prevailing definition is, that a metabolic pathway is a special case of a metabolic network, with distinct start and end-points, initial and terminal vertices, respectively, and a unique path between them, i.e. a directed reaction graph with substrates as vertices and arcs denoting enzymatic reactions [For99]. Some databases such as KEGG, WIT represented metabolic pathway graphs with labeled arcs indicating the involved enzymes.

Schuster et al. [Sch00a] provided a general definition of metabolic pathways based on the concept of elementary flux modes. It allows one to test whether sets of enzymes form a consistent pathway allowing mass balancing for each intermediate and complying with the directionality of reactions (irreversibility). However, it represents modes under idealized situations without regulation and feedback and from simulation point of view it is impossible to analyze the whole metabolism of organisms. Moreover, in order to determine a single mode, the large metabolic network has to be decomposed into smaller ones based on graphical algorithm.

One is called a pathway only if they are linear and unbranched. A metabolic pathway is an unrepeatable, irreversible sequence of a series of vertices and arcs leading from a molecule vertex, labeled as substrate via a molecule vertex, labeled as enzyme, to a molecule vertex, labeled as product. When the end vertex meets the previous vertex of the sequence, they make a complete close pathway, called a (partial) cyclic metabolic pathway.

There are four linear pathways in the example of Figure 5.1Ba. In practice, representation of enzymatic catalysization is omitted (Figure 5.1Bb). However, in many of the biochemical reactions in living cells, enzymes act as catalysts in the conversion of certain compounds (substrates) into other compounds (products). So enzymes are the cores of metabolism and make the whole cellular processes connected, and the metabolic network can be interpreted as sets of enzyme catalyzed biochemical reactions. That is, pathways are

abstractions of sets of enzymatic reactions; they are substructures that are partitions of the metabolism. Then the representation of metabolic pathways might be given as graph G(***E,A***), Figure 5.1Bc shows the graph example, which is exploited for our Enzyme-Enzyme relationship metabolic pathway alignment.

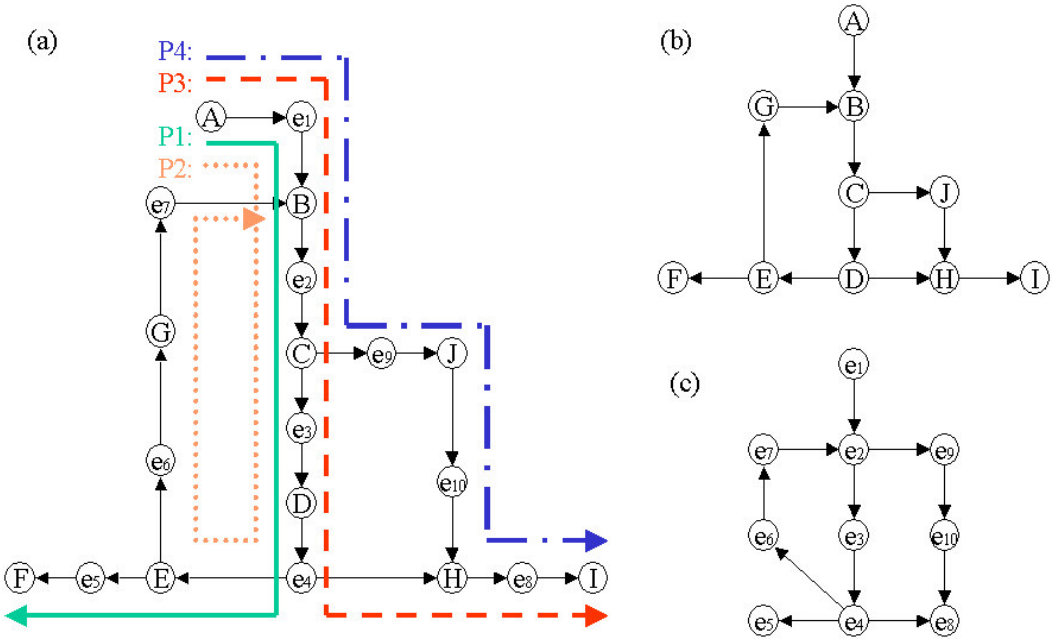


Figure 5.1B An example of metabolic pathways; A, B, …, I are metabolites, e1, e2, …, e10 enzymes.

As a result the traditional well-known pathways such as glycolysis or TCA will not be considered as a well-defined pathway, because they often contain some branches or alternative pathways. In fact there are several pathways inside them. Obviously there are always interactions among pathways.

We consider that a metabolic pathway is a subset of these reactions that describe the biochemical conversion of a given reactant to its desired end product. Rather to say, several biochemical reactions act together in a pathway to transform a set of initial substrates into products with very different structures, a new proposed definition of the metabolic pathway is presented and discussed in the following paragraphs.

Let $M = \{m_1,...,m_n\}$ be a set of metabolites in cells. Let $f_i : M \to M$ be a function for bioprocess events taking place in the cells. Bioprocess events are any kinds of biological actions among metabolites in cells.

The fact that f_i is a bioprocess function from a set of reactants R ($R \subseteq M$) into a set of products P ($P \subseteq M$) is written as follows:

$$f_i : R \to P$$

for all $m_1, m_2, m_3 \in M$, the following property holds:

$$f_1(m_1) = m_2 \text{ and } f_2(m_2) = m_3 \Rightarrow f_2(f_1(m_1)) = m_3 .$$

Let $f_1(m_1) = m_2, f_2(m_2) = m_3, \ldots, f_k(m_k) = m_{k+1}$, we define $f_1 f_2 \ldots f_k(m_1) = f_k(f_{k-1} \ldots f_1(m_1)) = m_{k+1}$.

Definition 5.1 *Given $f : M \rightarrow M$, a bioprocess pathway is defined as a subset of successive bioprocess events $\mathcal{P} = f_1 f_2 \ldots f_k$.*

For each f_i $(1 \leq i \leq k)$, there exist a pair of metabolites (m_i, m_{i+1}), $m_i \in M$, $m_{i+1} \in M$, involved in the bioprocess event as reactant and product. $\mathcal{P}$ is a composition function of the functions f_i, and we have $\mathcal{P}(m) = f_k f_{k-1} \ldots f_1(m)$. Given an initial substrate m_1, then the ending product m_{k+1} is predictable. If $m_{k+1} = m_i$ $(1 \leq i \leq k)$, i.e. the product of f_k is the one of the metabolites involved in the previous steps, then the pathway is called as a cyclic pathway. Otherwise, if $m_{k+1} \neq m_i$ $(1 \leq i \leq k)$, then it is a linear (non-cyclic) pathway.

Note that the function f is a genetic term for all bioprocess events in a cell, including biochemical reactions, membrane transportations, signal transductions, and so on. In case of enzymatic reaction, the function "f" can be written as "e" in order to distinguish the enzymatic reaction function from the genetic reaction function.

Definition 5.2 *A metabolic pathway is defined as a subset of successive enzymatic reaction events $P = e_1 e_2 \ldots e_k$.*

Each enzymatic reaction e_i $(1 \leq i \leq k)$ is catalyzed by a certain enzyme that is denoted as a unique EC number. The EC number is expressed with a 4-level hierarchical scheme that has been developing by the International Union of Biochemistry and Molecular Biology (IUBMB) [Web92]. The 4-digit EC number, $d_1.d_2.d_3.d_4$ represents a sub-sub-subclass indication of biochemical reaction. For instance, arginase is numbered by EC 3.5.3.1, which indicates that the enzyme is a hydrolase (EC 3.*.*.*), acts on the "carbon-nitrogen bonds, other than peptide bonds" (sub-class EC 3.5.*.*) in linear amidines (sub-sub-class EC 3.5.3.*). The enzymes normally are separated enzymes. For those enzymes may form a multienzyme complex (noncovalent aggregates of enzymes) or may be a membrane-bound system, we can choose the representative one of the enzymes unless there is a unique term for it. Thus we can adapt the EC number as a unique name for the responding enzyme catalyzed reaction.

Example 5.1 $e_{3.5.3.1}$ means the biochemical reaction that is catalyzed by the enzyme 3.5.3.1, which catalyze arginine into urea.

$$e_{3.5.3.1}(\text{arginine}) = \text{urea}$$

and

$$e_{2.1.3.3} e_{6.3.4.5} e_{4.3.2.1} e_{3.5.3.1}(\text{carbamoyl-P}) = \text{urea},$$

indicates that a metabolic pathway $e_{2.1.3.3}e_{6.3.4.5}e_{4.3.2.1}e_{3.5.3.1}$ starts from the enzymatic reaction 2.1.3.3 with carbamoyl-P as reactant, and after a series of reactions (2.1.3.3, 6.3.4.5, 3.5.3.1) results in urea as product.

5.2 Metabolic Pathway Alignment

Alignment as one of the most powerful methods to comparatively analyze the relationship between two sequences has been widely investigated in the field of bioinformatics to further understand the biological homology and estimate evolutionary distance. A common approach is to align sequences to each other, and measure distances by direct usage of molecular sequence data with, e.g. parsimony or maximum likelihood methods, or to calculate a corresponding similarity/distance matrix in multiple sequence alignment algorithms. Recently the emphasis of research efforts begins to turn back from gene sequences to cell functions as the completion of a long series of genomes and the accumulation knowledge of metabolism have made the comparison of complete metabolic pathways possible. Some approaches emphasized on either comparisons of gene sequence of involved enzymes [Dan99] [For01] or maximum likelihood mapping of enzyme using EC numbers [Toh00a] [Toh00b] have been made.

Our metabolic pathway alignment is a mapping of one pathway onto another by calculating the similarity of them at a metabolic level instead of genomic level. The basic concept is to measure the similarity.

5.2.1 Theory Basics

In order to score the similarity (percent identity) between two metabolic pathways, we define the similarity function. The notion of similarity function is the key to the pathway alignment.

Definition 5.3 *Let E be a finite set of e functions, an edit operation is an ordered pair* $(\alpha,\beta) \in (E\cup\{\varepsilon\})\times(E\cup\{\varepsilon\})\backslash\{(\varepsilon,\varepsilon)\}$.

α and β denote 4-digit EC strings of enzymatic reaction function, e.g. $\alpha=e_{1.1.1.1}$ $\beta=e_{2.3.4.5}$, ε denotes the empty string for null function. However, if $\alpha\neq\varepsilon$ and $\beta\neq\varepsilon$, then the edit operation (α,β) is identified with a pair of enzymatic reaction function.

An edit operation (α,β) is written as $\alpha\rightarrow\beta$ (we can simply written α, β as EC numbers). There are three kinds of edit operations:

$\alpha\rightarrow\varepsilon$ denotes the deletion of the enzymatic reaction function α,

$\varepsilon\rightarrow\beta$ denotes the insertion of the enzymatic reaction function β, and

$\alpha\rightarrow\beta$ denotes the replacement of the enzymatic reaction function α by the enzymatic reaction function β.

Notice that $\varepsilon\rightarrow\varepsilon$ never happens.

Definition 5.4 *Let $E_1=e_1e_2\ldots e_m$ and $E_2=e_1'e_2'\ldots e_n'$ be two metabolic pathways, an alignment of E_1 and E_2 is a pair sequence*

$$(\alpha_1\rightarrow\beta_1,\ldots,\alpha_h\rightarrow\beta_h)$$

of edit operations such that $E_1'=\alpha_1,\ldots,\alpha_h$ and $E_2' = \beta_1,\ldots, \beta_h$.

Note that the unique alignment of ε and ε is the empty alignment, that is, the empty sequence of edit operations. Empty element ε can be inserted at any position, i.e. also at the beginning or end. An alignment is usually written by placing the EC numbers of the two aligned pathways on different lines.

Example 5.2 The alignment A = (2.4.2.3→2.4.2.4, 3.5.4.5→ε, 3.1.3.5→3.1.3.5, ε→2.7.4.9) of the pathways $e_{2.4.2.3}e_{3.5.4.5}e_{3.1.3.5}$ and $e_{2.4.2.4}e_{3.1.3.5}e_{2.7.4.9}$ is written as follows, one over the other:

$$\begin{pmatrix} 2.4.2.3 & 3.5.4.5 & 3.1.3.5 & \varepsilon \\ 2.4.2.4 & \varepsilon & 3.1.3.5 & 2.7.4.9 \end{pmatrix}$$

Example 5.3 Five alignments of $E_1=e_{2.7.4.14}e_{3.1.3.5}e_{3.5.4.5}e_{2.4.2.3}$ and $E_2=e_{2.7.4.14}e_{3.2.2.10}e_{3.5.4.1}e_{2.4.2.3}e_{3.5.4.5}$

$$A_1=\begin{pmatrix} 2.7.4.14 & \varepsilon & 3.1.3.5 & 3.5.4.5 & \varepsilon & 2.4.2.3 & \varepsilon \\ 2.7.4.14 & 3.2.2.10 & \varepsilon & \varepsilon & 3.5.4.1 & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

$$A_2=\begin{pmatrix} 2.7.4.14 & \varepsilon & 3.1.3.5 & \varepsilon & 3.5.4.5 & 2.4.2.3 & \varepsilon \\ 2.7.4.14 & 3.2.2.10 & \varepsilon & 3.5.4.1 & \varepsilon & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

$$A_3=\begin{pmatrix} 2.7.4.14 & \varepsilon & \varepsilon & 3.1.3.5 & 3.5.4.5 & 2.4.2.3 & \varepsilon \\ 2.7.4.14 & 3.2.2.10 & 3.5.4.1 & \varepsilon & \varepsilon & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

$$A_4=\begin{pmatrix} 2.7.4.14 & 3.1.3.5 & \varepsilon & \varepsilon & 3.5.4.5 & 2.4.2.3 \\ 2.7.4.14 & 3.2.2.10 & 3.5.4.1 & 2.4.2.3 & 3.5.4.5 & \varepsilon \end{pmatrix}$$

$$A_5=\begin{pmatrix} 2.7.4.14 & 3.1.3.5 & 3.5.4.5 & 2.4.2.3 & \varepsilon \\ 2.7.4.14 & 3.2.2.10 & 3.5.4.1 & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

Lemma 5.1 *Let $A=(\alpha_1\rightarrow\beta_1,\ldots,\alpha_h\rightarrow\beta_h)$ be an alignment of $E_1=e_1e_2\ldots e_m$ and $E_2=e_1'e_2'\ldots e_n'$. Then $m+n \geq h \geq max\{m,n\}$.*

Proof. 1.) The alignment

$$\begin{pmatrix} e_1 & e_2 & \ldots & e_m & \varepsilon & \varepsilon & \ldots & \varepsilon \\ \varepsilon & \varepsilon & \ldots & \varepsilon & e_1' & e_2' & \ldots & e_n' \end{pmatrix}$$

of E_1 and E_2 is of maximal length. Its length is $m+n$, hence $m+n \geq h$.

2.) Let $m \geq n$. then

$$\begin{pmatrix} e_1 & e_2 & \ldots & e_n & e_{n+1} & e_{n+2} & \ldots & e_m \\ e_1' & e_2' & \ldots & e_n' & \varepsilon & \varepsilon & \ldots & \varepsilon \end{pmatrix}$$

3.) The case $m < n$ is similar to case 2. Hence $h \geq max\{m,n\}$.

To estimate the number of pathway alignments, we define that *Aligns*(m,n) is the number of alignments of one pathway E_1 of m EC numbers with another pathway E_2 of n EC numbers.

Lemma 5.2 *For all* $m,n \geq 0$, *Aligns*$(0,0) = 1$, *Aligns*$(m,0) = 1$ *and Aligns*$(0,n) = 1$, *then*

$$Aligns(m,n) = Aligns(m\text{-}1,n) + Aligns(m, n\text{-}1) + Aligns(m\text{-}1,n\text{-}1)$$

Proof. The idea is to focus on the end of the alignment. If e_m is deleted, then there exist *Aligns*$(m\text{-}1,n)$ alignments of the earlier part of the pathway. If e_n' is deleted, then *Aligns*$(m,n\text{-}1)$ alignments result. If e_m and e_n' are aligned, *Aligns*$(m\text{-}1,n\text{-}1)$ alignments result. Therefore,

$$Aligns(m,n) = Aligns(m\text{-}1,n) + Aligns(m, n\text{-}1) + Aligns(m\text{-}1,n\text{-}1)$$

If not to count $\begin{matrix} e_i & \varepsilon \\ \varepsilon & e_j \end{matrix}$ and $\begin{matrix} \varepsilon & e_i \\ e_j & \varepsilon \end{matrix}$ as distinct, the new way of counting alignments is to identify aligned pairs $\begin{matrix} e_1 \\ e_2 \end{matrix}$ and to ignore permutations of $\begin{matrix} e_1 & e_2 & \varepsilon \\ \varepsilon & \varepsilon & e_3 \end{matrix} \cdots$. The notation of index alignment is introduced.

Definition 5.5 *A index alignment of* E_1 *and* E_2 *is a set of index pairs*, $(i_1,j_1),(i_2,j_2),\ldots,(i_r,j_r)$ *satisfying*:

$$1 \leq i_1 < i_2 < \ldots < i_r \leq m$$

and

$$1 \leq j_1 < j_2 < \ldots < j_r \leq n.$$

For $1 \leq h \leq r$, the index pair (i_h,j_h) stands for the replacement $e_{ih} \rightarrow e'_{jh}$. We say that e_{ih} is matched or aligned with e'_{jh}. All EC numbers in E_1 and E_2 not occurring in an index alignment are considered to be deleted in E_1 or E_2. In a graphical representation, the index pairs of the index alignment appear as lines connecting the EC numbers (Example 3).

Example 5.4 The following index alignment of $E_1 = e_{2.7.4.14}e_{3.1.3.5}e_{3.5.4.5}e_{2.4.2.3}$ and $E_2 = e_{2.7.4.14}e_{3.2.2.10}e_{3.5.4.1}e_{2.4.2.3}e_{3.5.4.5}$ represent the alignments of Example 7.2. In particular, P_1 represents A_1, A_2 and A_3, while P_2 represents A_4, and P_3 represents A_5.

$$P_1 = \begin{pmatrix} 2.7.4.14 & 3.1.3.5 & 3.5.4.5 & 2.4.2.3 & \\ | & & & | & \\ 2.7.4.14 & 3.2.2.10 & 3.5.4.1 & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 2.7.4.14 & 3.1.3.5 & 3.5.4.5 & 2.4.2.3 & \\ | & | & & & \\ 2.7.4.14 & 3.2.2.10 & 3.5.4.1 & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

$$P_3 = \begin{pmatrix} 2.7.4.14 & 3.1.3.5 & 3.5.4.5 & 2.4.2.3 & \\ | & | & | & | & \\ 2.7.4.14 & 3.2.2.10 & 3.5.4.1 & 2.4.2.3 & 3.5.4.5 \end{pmatrix}$$

Lemma 5.3 *Let Indexaligns*(m,n) *be the number of index alignment of two fixed pathways of length m and n. Then*

$$Indexaligns(m,n) = \sum_{r\geq 0}^{\min(m,n)} \binom{m}{r} \cdot \binom{n}{r} = \binom{m+n}{n}.$$

Proof. 1.) For each $r \in [0, min\{m,n\}]$ we have: for the ordered selection of the indices $i_1,\ldots,i_r$ there are $\binom{m}{r}$ possibilities; for the ordered selection of the indices $j_1,\ldots,j_r$ there are $\binom{m}{r}$ possibilities. All these possibilities have to be combined: $Indexaligns(m,n) = \sum_{r\geq 0}^{\min(m,n)} \binom{m}{r} \cdot \binom{n}{r}$.

2.) The key to the last equality is to consider the binomial expansion $(x+y)^n$. For details see Appendix C.

Obviously, *Indexaligns*(m,n) is a sepcial case of *Aligns*(m,n), and it is possible to further reduce the number of alignments by requiring conditional matches. There are at least $\binom{m+n}{n}$ different alignments between E_1 and E_2.

5.2.2 Similarity Function

The notion of alignment requires some scoring or optimization criterion. A variety of different similarity measures can be used to calculate the similarity. A scoring scheme must account for replacements, insertions and deletions. Scores are measures of sequence similarity (similar sequences have high scores); this is given by a similarity function. A characteristic of a similarity function is that the results of the function increase as the comparing item become more similar. The value is zero if the items are totally dissimilar. The similarity function is measured by the following definition:

Definition 5.6 *A similarity function* σ *assigns to each edit operation* (α,β) *a nonnegative real number. The similarity* $\sigma(\alpha,\varepsilon)$ *and* $\sigma(\varepsilon,\beta)$ *of the deletion operation* (α,ε) *and insertion operation* (ε,β) *is* 0. *For all replacement operations* (α,β) $\alpha\neq\varepsilon$, $\beta\neq\varepsilon$, *say*, $\alpha=d_1.d_2.d_3.d_4$ *and* $\beta=d_1'.d_2'.d_3'.d_4'$, *then the similarity function* $\sigma(\alpha,\beta)$ *is defined by*:

$$\sigma(\alpha,\beta) = \begin{cases} 0, \text{ if } (d_1 \neq d_1'); \\ 0.25, \text{ if } (d_1 = d_1' \text{ and } d_2 \neq d_2'); \\ 0.5, \text{ if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 \neq d_3'); \\ 0.75, \text{ if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 = d_3' \text{ and } d_4 \neq d_4'); \\ 1, \text{ if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 = d_3' \text{ and } d_4 = d_4' \text{ i.e. } \alpha=\beta). \end{cases}$$

The definition does not exclude the possibility that d_4, $d_3.d_4$, and $d_2.d_3.d_4$ can be respectively expressed as wide card symbols *, *.* and *.*.* which means no clear classification of the enzyme.

According to the Enzyme Nomenclature (IUBMB) [Web92], the EC number is function-based (the substrate-product conversion) instead of structure-based (the physical nature of the catalyst). So it is possible that two structurally dissimilar enzymes could catalyze a single reaction. In this case, the similarity score of $\sigma(\alpha,\beta)$ is unrelated to the physical nature of enzymes but dependent on theire catalytic reactions. The higher similarity score, the closer the classes of the two reactions.

Single pair of EC string comparison just means to measure how different EC strings are. Often it is additionally of interest to analyze the total difference between two strings into a collection of individual elementary differences. The most important mode of such analyses is an alignment of the pathways. The function σ can be extended to alignments in a straightforward way: the similarity $\sigma(A)$ of an alignment $A=(\alpha_1\rightarrow\beta_1,\ldots,\alpha_h\rightarrow\beta_h)$ is the sum of the similarities of the edit operations A consists of.

$$\sigma(A) = \sum_{i=1}^{h} \sigma(\alpha_i \rightarrow \beta_i)$$

Example 5.5 The similarity of the alignment A_5 in the example 3 is:

$$\sigma(A_5) = \sigma(2.7.4.14 \rightarrow 2.7.4.14) + \sigma(3.1.3.5 \rightarrow 3.2.2.10) + \sigma(3.5.4.5 \rightarrow 3.5.4.1) + \sigma(2.4.2.3 \rightarrow 2.4.2.3) + \sigma(\varepsilon \rightarrow 3.5.4.5)$$
$$= 1+0.25+0.75+1+0$$
$$= 3.0$$

When considering the lengths of pathways, an alignment scoring scheme is given.

Definition 5.7 *An alignment scoring scheme*, *Score*(E_1,E_2) *of two metabolic pathways is the average degree of their similarity of the alignment*

$$Score(E_1, E_2) = \frac{1}{\max(m,n)}\sigma(A)$$

Obviously, the worst case of scoring is that there is no index pair exists in the alignment, i.e. all edit operations are deletions and/or insertions. Hence similarity $\sigma(A)$ is zero. The best case is that the similarities of all edit operation are 1, i.e. $\alpha=\beta$, there is neither deletion nor insertion. Hence similarity $\sigma(A)$ is n, and E_1 and E_2 share the same length, $m=n$. Therefore $Score(E_1,E_2) = \frac{n}{n} = 1$. They are actually the same.

In order to achieve a maximum possible score of the alignment, the edit distance could be adapted to measure the similarity between two pathways by calculating the minimal cost of the edit operations [Lev66] [Wag74] [Sel80]. However, when taking the biological aspects of metabolic pathways into account, especially when we considering that two evolutional related pathways are diverged for some certain biological purpose, the alignment with the edit distance is arbitrary and sometimes biological meaningless. For example, the same metabolic pathway from two organisms may have diverged since the organisms evolved from their common ancestor, and individual metabolites and enzymes may have been changed or added or lost in one pathway. There are two theories exist. The "retrograde evolution" theory [Hor45] states that sequential disappearance of key intermediary metabolites induces the recruitment of similar available substrates via new enzymes. The "substrate ambiguity" theory [Jen76] indicates that enzyme recruitment from a pool of ancestral enzymes with basic functions and substrate ambiguity. So the intended function similarities of metabolic pathways are taken into account, i.e., two pathways are supposed to be function related. They performed some similar biological purposes from certain starting substrates to the ending products. Based on these considerations, we define a new function to perform the removal of unmatched elements from both ends of the pathway.

5.2.3 Strip and Index Function

Definition 5.8 *Strip function* δ *and index function* λ *of* $E_1=e_1e_2\ldots e_m$ *and* $E_2= e_1'e_2'\ldots e_n'$ *are defined as*

$$\delta(e_1e_2\ldots e_m, e_1'e_2'\ldots e_n') = (e_{i+1}e_{i+2}\ldots e_{k-1}, e_{j+1}'e_{j+2}'\ldots e_{l-1}')$$

$$\lambda(e_1e_2\ldots e_m, e_1'e_2'\ldots e_n') = \{(i,j),(k,l)\}$$

where: $1 \le i \le k \le m$ and $1 \le j \le l \le n$,

$e_i, e_k \in E_2$,

$e_1, e_2, \ldots e_{i-1}, e_{k+1}', \ldots e_m' \notin E_2$,

e_j' is the first element matching from left to right such that $e_i = e_j'$,

and e_l' is the first element matching from right to left such that $e_k = e_l'$.

Also set

$$\delta(e_1e_2\ldots e_m, e_1'e_2'\ldots e_n') = (e_1e_2\ldots e_m, e_1'e_2'\ldots e_n')$$

when $e_i \ne e_j$ $(1 \le i \le m$ and $1 \le j \le n) \mid e_1e_2\ldots e_m = \varphi \mid e_1'e_2'\ldots e_n' = \varphi \mid e_1e_2\ldots e_m = e_1'e_2'\ldots e_n' = \varphi$.

The result of $\delta(e_1e_2\ldots e_m, e_1'e_2'\ldots e_n')$ can be further performed with δ until $\delta(E_1, E_2) = (E_1, E_2)$ (Figure 5.2.3A).

$$\delta^k(E_1, E_2) = \delta(\delta^{k-1}(E_1, E_2))$$

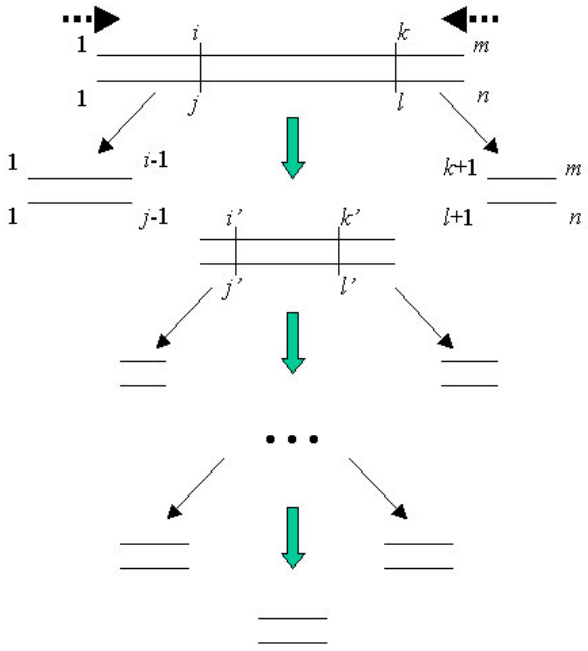


Figure 5.2.3A A schematic alignment with strip function.

The set of all matched points will be denoted as $\lambda^*(E_1, E_2)$. Such that:

$$\lambda^*(E_1, E_2) = \lambda(E_1, E_2) \cup \lambda(\delta(E_1, E_2)) \cup \cdots \lambda(\delta^k(E_1, E_2))$$

$$= \bigcup_{i=0}^{k} \lambda(\delta^r(E_1, E_2)) = \{(i, j), (i', j'), \ldots, (k', l'), (k, l)\}$$

where $\lambda(\delta^0(E_1, E_2)) = \lambda(E_1, E_2)$; $1 \le i < i' < \ldots < k' < k \le m$, $1 \le j < j' < \ldots < l' < l \le n$.

According to Definition 5.5, we know that $(i, j),(i', j'),\ldots, (k', l'),(k, l)$ is a index alignment of E_1 and E_2.

Example 5.6 Let two pathway $E_1=e_{2.7.7.6}e_{3.6.1.5}e_{2.7.4.14}e_{3.1.3.5}e_{3.5.4.5}e_{2.4.2.3}e_{1.3.1.1}e_{3.5.2.2}e_{3.5.1.6}$ and $E_2=e_{2.7.7.6}e_{2.7.4.6}e_{2.7.4.14}e_{3.2.2.10}e_{3.5.4.1}e_{1.3.1.2}e_{3.5.2.2}e_{3.5.1.6}$, then $\delta^2(E_1, E_2)= \delta(\delta(E_1, E_2))= (e_{3.1.3.5}e_{3.5.4.5}e_{2.4.2.3}e_{1.3.1.1}, e_{3.2.2.10}e_{3.5.4.1}e_{1.3.1.2})$.

Lemma 5.4 *Let Indexaligns($\delta^r(E_1, E_2)$) be the number of index alignment of two fixed pathways of length m and n with r matched points. Then*

$$1 \le Indexaligns(\delta^r(E_1, E_2)) \le \binom{m+n}{n}$$

Proof. See Appendix D.

Lemma 5.5 *Let E_1 and E_2 be two pathways of length m and n, there exists a minimum r that enables* $\overbrace{\delta(\delta(\cdots\delta}^{r}(E_1, E_2))) = (E_1, E_2)$, *then Score*$(E_1,E_2)=\frac{r}{\max(m,n)}\sum_{k=0}^{r}\sigma(\delta^k(E_1, E_2))$.

Algorithms for optimal alignment can seek either to minimize a dissimilarity measure or maximize a scoring function. However our scheme is based on the index pair matching with strip functions. An alignment scores is the maximum over all possible alignments $s=\max\{\sum_{i=1}^{h}\sigma(\alpha_i, \beta_i) : \text{all alignments}\}$.

Definition 5.9 *Given two pathway E_1 and E_2, a mapping $M(E_1,E_2)$ is defined as a set of position correspondences* (i, j) *satisfying* $1 \le i \le m$ *and* $1 \le j \le n$ *such that* e_i :::: e_j'. *The notation* "::::" *denotes that e_i and e_j' are compared to be identical in turn according to their 4-digit hierarchical patterns. A mapping is maximal if there does not exist another pair* (l, k) *such that* e_l :::: e_k'.

Obviously, each map site has two characteristics, site position (location) i and site feature name e_i. The map $E_1=e_1e_2\ldots e_m$ consists of a sequence of pairs $e_i = (a_i, r_i)$, where a_i = the location of the i-th site in number of pathways and r_i = the feature name at the i-th site. Similarly, $E_2= e_1'e_2'\ldots e_n'$ is a map where $e_j' = (b_j, s_j)$. For further analysis, let the symbol $\overrightarrow{E_1}$ denote natural order sequence of a_i, and $\overrightarrow{E_2}$ denote the sequence of corresponding position b_j. The number of E_1, E_2 maps is defined as $\left|\overrightarrow{E_1}\right|$ and $\left|\overrightarrow{E_2}\right|$. Clearly, $\left|\overrightarrow{E_1}\right| = \left|\overrightarrow{E_2}\right| \le m, n$.

Example 5.7 The maximal mapping of two pathways E_1= $e_{3.6.1.5}e_{2.7.4.14}e_{3.1.3.5}e_{2.7.1.48}$ and E_2= $e_{3.6.1.8}e_{3.1.3.5}e_{2.7.1.48}$ $e_{2.7.4.14}$ is {(1,1),(2,4),(3,2),(4,3)}. $\overrightarrow{E_1}$ = 1, 2, 3, 4 and $\overrightarrow{E_2}$ = 1, 4, 2, 3. The mapping can be represented as line $\overrightarrow{E_1}$ over line $\overrightarrow{E_2}$: $\begin{matrix}1 & 2 & 3 & 4\\ 1 & 4 & 2 & 3\end{matrix}$.

Suppose a mapping $M = \dfrac{\overrightarrow{E_1}}{\overrightarrow{E_2}} = \dfrac{a_{i1}a_{i2}\cdots a_{id}}{b_{j1}b_{j2}\cdots b_{jd}}$ (Figure 5.2.3B), the following properties hold:

1. $\left|\overrightarrow{E_1}\right| = \left|\overrightarrow{E_2}\right| = d$,
2. $a_{i1} < a_{i2} < \cdots < a_{id-1} < a_{id}$,
3. $b_{j1} < b_{j2} < \cdots < b_{jd-1} < b_{jd}$ is not always true.

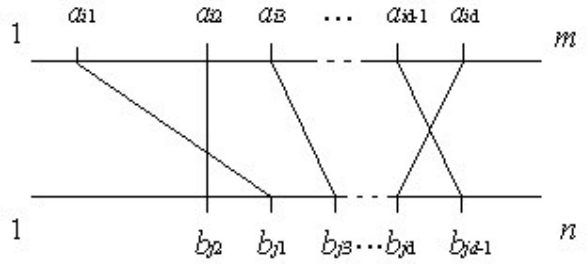


Figure 5.2.3B An illustrative map of two pathways.

As emphasized above, locations of $\overrightarrow{E_2}$ is not necessary in natural order. We define a map alignment as a mapping where $\overrightarrow{E_2}$ is a sequence in natural order.

Lemma 5.6 *Given the maximal mapping M of two pathways* $E_1=e_1e_2\ldots e_m$ *and* $E_2= e_1'e_2'\ldots e_n'$, *and let* $\overrightarrow{E_1} = a_{i1}a_{i2}\ldots a_{id-1}a_{id}$ *and* $\overrightarrow{E_2} = b_{j1}b_{j2}\ldots b_{jd-1}b_{jd}$ *be two maps of M. Then we have*:

A sub-sequence of $\overrightarrow{E_2}$ *in natural order with the longest length is the maximal map alignment of* E_1 *and* E_2.

A sub-sequence of $\overrightarrow{E_2}$ *in natural order between* b_{j1} *and* b_{jd} *is the index alignment*.

Proof. 1.) Suppose $b_{jk0}b_{jk1}\ldots b_{jkt}$ is a sub-sequence of $\overrightarrow{E_2}$ in natural order with the longest length, we can obtain the responding positions of this sub-sequence: $a_{ik0}a_{ik1}\ldots a_{ikt}$, so that they are a map alignment of E_1 and E_2. For contradiction, let us assume that there exist another mapping pair (a_{ik}', b_{jk}') excluded from the maximal map alignment, then $b_{jk0}b_{jk1}\ldots b_{jkt}$ is not the longest subsequence, which is a contradiction.

2.) According to the definition of index alignment, the first step is to map from both ends of E_1, b_{j1} and b_{jd} are found. Next step is to map a_{i2} and a_{id-1}, b_{j2} is one map of $\overrightarrow{E_2}$ only

when b_{j2} is greater than b_{j1}, and b_{jd-1} is one map only when b_{jd-1} is less than b_{jd} and greater than b_{j2}. Repeat the mapping till no more b_{jk} is satisfied.

Example 5.8 The map of two pathways in the Example 5.7 is:

$$\begin{matrix}1 & 2 & 3 & 4\\ 1 & 4 & 2 & 3\end{matrix} \Rightarrow \begin{pmatrix}3.6.1.5 & 2.7.4.14 & 3.1.3.5 & 2.7.1.48\\ | & & & \\ 3.6.1.8 & 3.1.3.5 & 2.7.1.48 & 2.7.4.14\end{pmatrix}$$

The maximal map alignment

$$\begin{matrix}1 & 3 & 4\\ 1 & 2 & 3\end{matrix} \Rightarrow \begin{pmatrix}3.6.1.5 & 2.7.4.14 & 3.1.3.5 & 2.7.1.48\\ | & & & \\ 3.6.1.8 & 3.1.3.5 & 2.7.1.48 & 2.7.4.14\end{pmatrix}$$

The index alignment is

$$\begin{matrix}1 & 3 & 4\\ 1 & 2 & 3\end{matrix} \Rightarrow \begin{pmatrix}3.6.1.5 & 2.7.4.14 & 3.1.3.5 & 2.7.1.48\\ | & & & \\ 3.6.1.8 & 3.1.3.5 & 2.7.1.48 & 2.7.4.14\end{pmatrix}$$

Lemma 5.7 *Pair number of index alignment is less or equal than that of maximal map alignment*, *which is less or equal than that of maximal matching*.

Proof. Given two pathways E_1= $e_{3.6.1.5}e_{2.7.4.14}e_{3.1.3.5}e_{2.7.1.48}$ and E_2= $e_{3.6.1.8}e_{3.1.3.5}e_{2.7.1.48}$ $e_{2.7.4.14}$, then $Score(E_1,E_2)$= ¼*σ(2.7.4.14->2.7.4.14)+σ(3.6.1.5->3.6.1.8)=1/4*(1+0.75)=0.43.

While the score of the maximum alignment will be

$$S_{maxa}=1/4*(0.75+1+1)=0.69.$$

The score of the maximal matching is

$$S_{maxm}=1/4*(0.75+1+1+1)=0.94.$$

5.2.4 Algorithms

Before describing our algorithms, we introduce four different subscript notations of the Strip function δ and the Index function λ. From its definition we know that the Strip function δ is able to strip two pathways if there are two pairs of elements that are matched, e.g. e_i=e_j' and e_k=e_l'. Here, e_i, e_j', e_k and e_l'are 4-hirarchical numbers. We count all four numbers by default. Now if we count only the first three numbers for matching, then the Strip function δ can be written as δ_3 in order to distinct the default setting δ that can also be written as δ_4 in this case. Similarly for δ_2 and δ_1. Accordingly λ_4 is defined as the default 4-number indexing, and λ_3, and so on.

5.2.4.1 Pairwise alignment

In general, we align metabolic pathways one above the other. The alignment algorithm is based on likelihood calculations of index pairs. Given two metabolic pathways P_1 and P_2 (Figure 5.2.4.1), the implemented algorithm is given by the following pseudo-code:

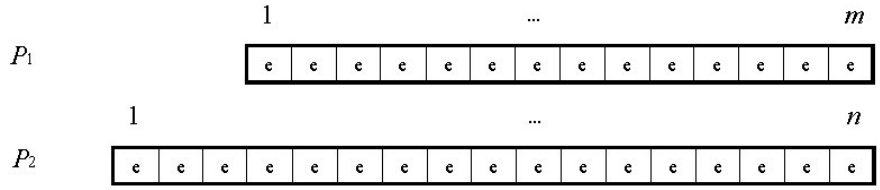


Figure 5.2.4.1 Two aligned metabolic pathways.

Begin

Input P1: $E_1=e_1e_2\ldots e_m$

P2: $E_2=e_1'e_2'\ldots e_n'$

Initial set Score=0.0

Strip $\delta_4^*(E_1, E_2)$

Compute $\lambda_4^*(E_1, E_2)$

Foreach stripped sub-pathway, **do**

$E_1'=e_1e_2\ldots e_{mi}$

$E_2'=e_1'e_2'\ldots e_{nj}'$

Strip $\delta_3^*(E_1', E_2')$

Compute $\lambda_3^*(E_1', E_2')$

Foreach stripped sub-sub-pathway, **do**

$E_1''=e_1e_2\ldots e_{mii}$

$E_2''=e_1'e_2'\ldots e_{njj}'$

Strip $\delta_2^*(E_1'', E_2'')$

Compute $\lambda_2^*(E_1'', E_2'')$

Foreach stripped sub-sub-sub-pathway, **do**

$E_1'''=e_1e_2\ldots e_{miii}$

$E_2'''=e_1'e_2'\ldots e_{njjj}'$

Strip $\delta_1^*(E_1''', E_2''')$

Compute $\lambda_1^*(E_1''', E_2''')$

Foreach stripped sub-pathway, **do**

Count score

Next

Count score

Next

Count score

Next

Count score

Next

Output λ_4^*, λ_3^*, λ_2^*, λ_1^* and Score

End

Line 1-3: Initialize the set of unaligned EC number sequences, their lengths and score value.

Line 4-29: Starting from both ends towards the middle, align one sequence to another and attempt to find all EC numbers with same 4-level hierarchical numbers. Score the similarities. Recall the alignment positions, where EC number are identical, and cut the sequences into more sub-sequences by removing the identical EC numbers.

Line 7-27: Each pair of sub-sequences is initialized to begin a new round of 3-level hierarchical EC number matching. Till all pairs of sub-sequences are aligned. A similarity score is calculated afterwards.

12-25: Apply the same rule again, find the similarities of rest unaligned sub-sub-sequences based on 2-level hierarchical EC number matching.

17-23: Then sub-sub-sub-sequences on 1-level are matched.

5.2.4.2 Time complexity analysis

The best case to strip two pathways and obtain their index pairs is that they are identical, which costs $O(m)$ computing time. While the worst case is that they are un-matchable, it needs $O(mn)$ to cover all elements. In general, we define $O(m \bullet n)$ to show the "average" time complexity. Therefore, to create $\lambda_4^*(E_1, E_2)$, it takes $O(m \bullet n)$. $\lambda_3^*(E_1', E_2')$ takes $O(\sum_{l=1}^{K+1} m^l \bullet n^l)$, $\lambda_2^*(E_1'', E_2'')$ takes $O(\sum_{l=1}^{K+1}\sum_{i=1}^{R_l+1} m_i^l \bullet n_i^l)$ and $\lambda_1^*(E_1''', E_2''')$ takes $O(\sum_{l=1}^{K+1}\sum_{i=1}^{R_l+1}\sum_{j=1}^{S_i+1} {}_j m_i^l \bullet {}_j n_i^l)$, where K denotes numbers of index alignment of (E_1, E_2), R_l denotes each striped sub-pathway of (E_1', E_2'), S_i denotes each striped sub-sub-pathway of (E_1'', E_2''), and ${}_j m_i^l$ and ${}_j n_i^l$ represent the length of each striped sub-sub-sub-pathway (E1''', E2'''). Therefore, the total time complexity is $O(m \bullet n + \sum_{l=1}^{K+1} m^l \bullet n^l + \sum_{l}^{K+1}\sum_{i=1}^{R_l+1} m_i^l \bullet n_i^l + \sum_{l=1}^{K+1}\sum_{i=1}^{R_l+1}\sum_{j=1}^{S_i+1} {}_j m_i^l \bullet {}_j n_i^l)$.

The best-case complexity of the algorithm is the minimum number of steps taken on any instance of size m and n. It represents the alignment of two identical pathways, which takes $O(n)$.

The worst-case complexity of the algorithm is the maximum number of steps taken on any instance of size m and n. It represents the alignment of two pathways with no first 2-number is same, which takes at least $O(3mn)$.

5.2.4.3 Multiple alignment

Multiple alignment is useful for finding the phylogenetic analysis. Multiple sequence alignment is important for the recognition of patterns or motifs common to a set of function-related DNA sequences and is of assistance in structure prediction and molecular modeling. Multiple sequence alignment algorithms use variations of the dynamic programming method. Dynamic programming methods use an explicit measure of alignment quality, consisting of defined costs for aligned pairs of residues or residues with gaps and use an algorithm for finding an alignment with minimum total cost.

The multiple metabolic pathway alignment allows us to extract and represent biologically important but faintly/ widely dispersed pathway similarities, which, for instance, makes it possible to identify pathways preserved by evolution that play an important role in the cellular function and can give us hints about the evolutionary history of certain pathways.

By allowing the alignment of more than two metabolic pathways, the pairwise alignment algorithm can be extended. A multiple alignment of metabolic pathways E_1, E_2, …, E_k, can be seen as a generalization of pairwise metabolic pathway alignment - instead of aligning two pathways, k pathways are aligned simultaneously, where k is any number greater than two. A heuristic algorithm is used to perform the multiple metabolic pathway alignment. The general idea of the method is to construct a succession of pairwise pathway alignments:

Step 1: Choose one pathway E_1 from $\boldsymbol{E}=\{E_1,E_2,\ldots,E_k\}$ and align with $\{E_2,\ldots,E_k\}$ one after another to find the most similar pathway E_i $(2 \le i \le k)$.

Step 2: Choose the pathway E_i and align with $E\backslash\{E_1,E_i\}$ to get the most similar pathway E_i'.

Step 3: Iterate step 2 until all pathways are aligned.

The time complexity of multiple alignment of k pathways is $O(\frac{1}{2}k(k-1)({}^{p}m\bullet^{p+1}n+\sum_{l=1}^{K+1}{}^{p}m^{l}\bullet^{p+1}n^{l}+\sum_{l}^{K+1}\sum_{i=1}^{R_l+1}{}^{p}m_i^{l}\bullet^{p+1}n_i^{l}+\sum_{l=1}^{K+1}\sum_{i=1}^{R_l+1}\sum_{j=1}^{S_i+1}{}^{p}{}_{j}m_i^{l}\bullet^{p+1}{}_{j}n_i^{l}))$, where $p\in(1,k\text{-}1)$, the lengths of pathway p and p+1 are ${}^{p}m$ and ${}^{p+1}n$.

The multiple alignment is a complicated problem and there are some technical difficulties such as the choice of the pathways. The method proposed only makes sense if they

are assumed to be dealing with a set of homologous pathways i.e., pathways sharing a common ancestor. Given inappropriate (unrelated) pathways, the multiple alignment method will nonetheless produce an alignment. It will be the responsibility of the biologist to realize that this alignment is meaningless

5.3 PathAligner Implementation and Examples

5.3.1 Implementation

The algorithm has been implemented in the PathAligner system (http://bibiserv.techfak.uni-bielefeld.de/pathaligner). It is written in Perl and runs under UNIX. The graphical representation of alignment is done with the help of a simple graphical Perl module. Three web-based alignment interfaces are implemented in the current version. They are "*E-E Alignment*", "*M-E-M Alignment*" and "*Multiple Alignment*". "*E-E Alignment*" uses the basic algorithm to align two linear metabolic pathways (represented as EC number sequences). Users can also align any such a metabolic pathway against our pool database to find a list of hits. "*M-E-M Alignment*" considers the differences of metabolites in two pathways, which are presented as "Metabolite-EC number-Metabolite" patterns of sequence. It is possible to pick up two such pathways and align them to identify whether they are alternative pathways or partial ones. "*Multiple Alignment*" allows the alignment of more than two metabolic pathways. Some examples are illustrated in the following sections.

5.3.2 E-E Pairwise Alignment

The retrieved metabolic pathway can be aligned with other functionally similar metabolic pathways from other species. Based on the KEGG's pathway database, two metabolic pathways related to the urea cycle are selected and aligned. User can align the metabolic pathway with all pathways in the pool database (Figure 5.3.2). The upper left window shows the web-interface of pathway alignment. Users can align one pathway either with another pathway, or with all pathways deposited in the database. The lower left screenshot is an example of pairwise pathway alignment. All paired enzymes are highlighted in color. Blue color indicates that two EC numbers are exactly the same. EC numbers with green color share same d_1,d_2 and d_3 of EC 4-digit hierarchy. Pink EC numbers have the same d_1 and d_2. While red colored EC numbers only belong to the same main class of enzyme nomenclature. The similarity score is calculated after the comparison of two pathways. The right window is the alignment result of aligning against the database.

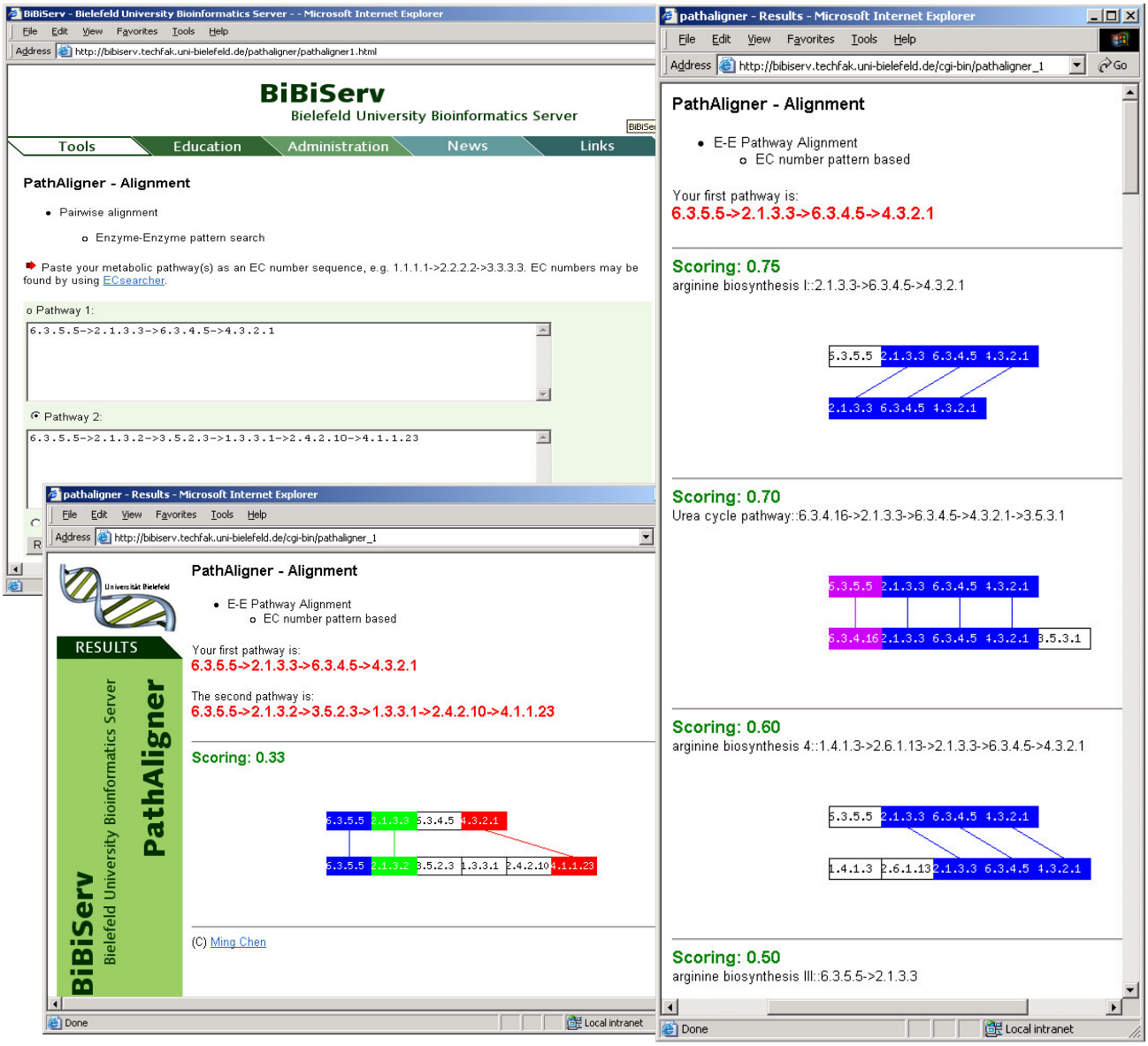


Figure 5.3.2 Screenshots of metabolic pathway alignment.

5.3.3 M–E–M Pairwise Alignment

In addition, the current version of PathAligner can also align metabolic pathways that are presented as (M–E–M). It is possible to pick up two such pathways and align them to identify whether they are alternative pathways or partial ones. This method differs from classic alternative pathway finding, based on Dijkstra/Floyd's algorithm [Dij59] [Flo62], that is used by some well known metabolic pathway databases such as KEGG [Oga96].

Pathway alignment is considered as a substrate-enzyme-product unit alignment. It again can be possibly analyzed and investigated in terms of gene sequences and evolution. The metabolic

pathways can be different from organisms to organisms. A screenshot example of M-E-M pairwise alignment is shown in Figure 5.3.3.

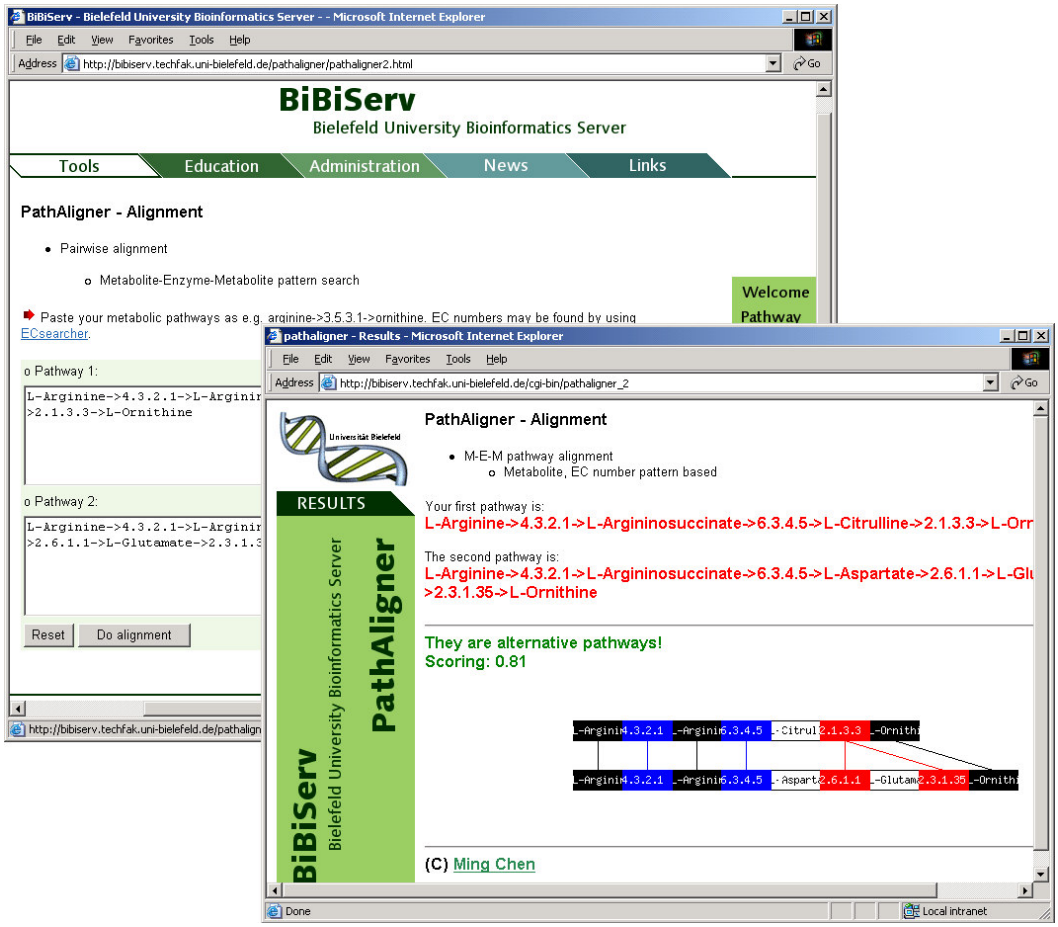


Figure 5.3.3 Screenshot of M-E-M alignment.

The algorithm takes two pathways and determines the similarity between them. We must define what the similarity is, in order to understand this problem. Alignment score of alternative pathways is equal and greater than 50%, 50% is the case that only the both ends of pathways are 100% identical. However, inconsistent naming conventions, synonymy and open, growing vocabulary for many classes lead to task difficulties in molecular biology ontologies. For instance, L-Arginine may have synonyms: 2-Amino-5-guanidinovaleric acid; Arg; Arginine; L-(+)-Arginine; 2-Amino-5-guanidinopentanoic acid. So, the entry numbers for compounds from KEGG are used as the unique IDs to the molecular elements in BioPNML, e.g. C00062 for L-Arginine.

5.3.4 Multiple Alignment

Multiple metabolic pathway alignment can be seen as a generalization of pairwise metabolic pathway alignment. A heuristic algorithm is used to perform the multiple metabolic pathway alignment. An example of multiple alignment of methionine pathways from different species as well as alternative methionine pathways from the same specie is performed (Figure 5.3.4)

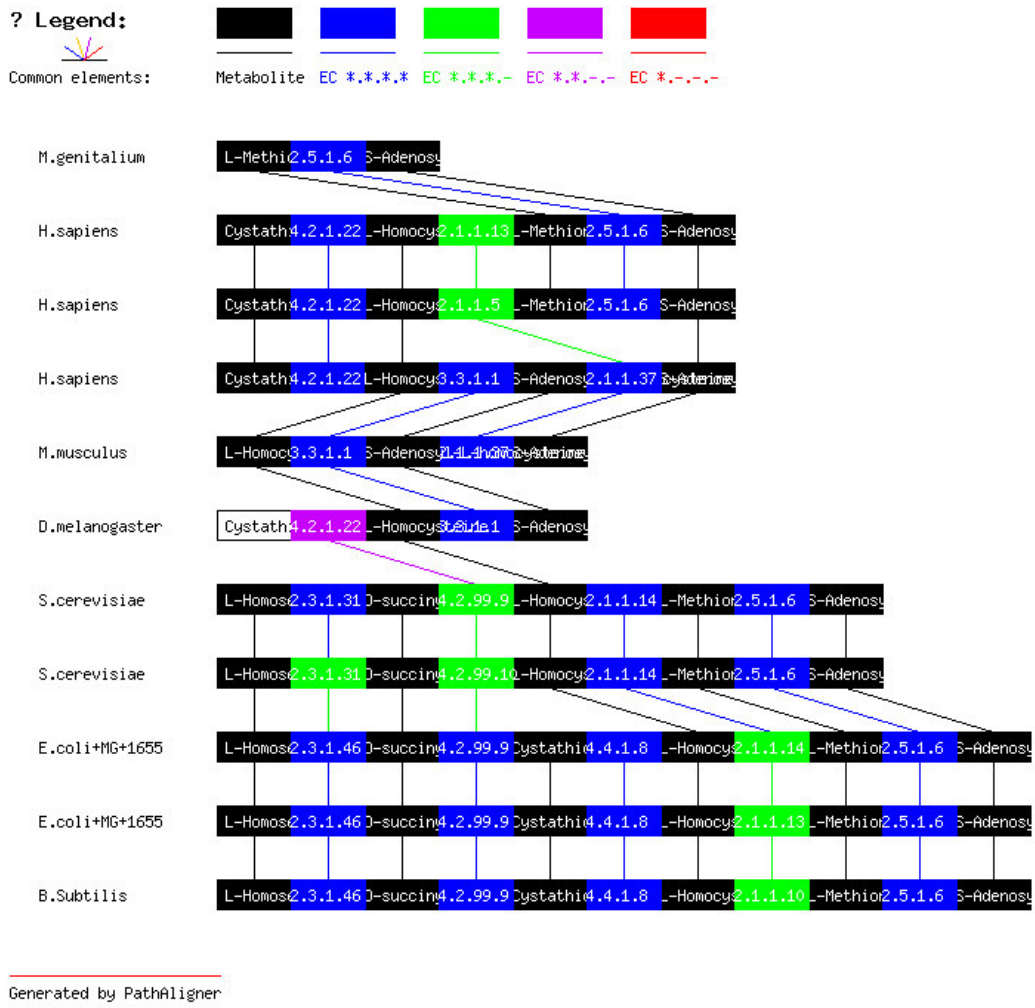


Figure 5.3.4 Graphic representation of multiple alignment of metabolic pathways. The same metabolites between two aligned pathways are colored in black and linked by a black line. Enzymes with the same 4-digits are colored in blue, with the same first 3-digits are colored in green, with the same first 2-digits are colored in purple, and with the same 1-digits are colored in red.

5.4 Summary

We have presented an algorithm to study the problem of metabolic pathway alignment. The entire processing for pairwise E-E alignment takes the time of order

$O(m \bullet n + \sum_{l=1}^{K+1} m^l \bullet n^l + \sum_{l} \sum_{i=1}^{R_l+1} m_i^l \bullet n_i^l + \sum_{l=1}^{K+1} \sum_{i=1}^{R_l+1} \sum_{j=1}^{S_i+1} {}_j m_i^l \bullet {}_j n_i^l)$, where *m*, *n* are the lengths of two aligned metabolic (sub)pathways. The algorithm described here has been successfully implemented and is in current use in the context of the PathAligner system.

The identification and analysis of metabolic networks is a complex task due to the complexity of the metabolic system. Abstract pathway defined as a linear molecule sequence, is practical for our alignment algorithm. However, when the topology of network is concerned, more information related to the components of the pathways, like their length, size, number of feedback cycles, number of crosstalks between pathways and area reachable from any point in the network, should be considered as much as possible. In this case, the pathway comparison will be the comparison of sub-network, or as a tree, rather as a single linear path sequence. As a result, this leads to another type of pathway computation, which can be categorized as the comparison of biological networks. By comparing such type of networks from different biological system, it is possible to identify similarities and variations among different species.

It is important to note that we are assuming that the two pathways are related in some biologically meaningful way, whether from different organisms or the same. Because high conservation of identity between two pathways is a strong indicator of their biologically significant relationship, we model every comparison as an experiment that seeks to quantify the related-ness of two putatively previously related pathways.

We can adopt our alignment algorithms to comparatively analyze other kinds of biopathways, such as signaling pathways. However, there is no nomenclature system for signal contradictions at the moment. We are going to present our classification of signal transductions in the next chapter. We will construct a database to host the classification system and perform signaling pathway alignment based on the classification and the algorithms we have discussed.

Chapter 6
Signaling Pathway Alignment*

Signal transduction has been of great interest to many academic and pharmaceutical scientists as it is becoming increasingly clear that the regulation of signal transduction is critical for understanding both basic biological processes, as well as how they may go awry leading to disease. The widespread use of modern biological technique to a unit operation in various fields of cellular technology has led to a proliferation of terminology. However, no reference has been made to the classification and definitions of transductions involving the signal reception, transportation and function. In this chapter, a classification and nomenclature of signal transduction is proposed. A systematic classification scheme is given for the various types of signal transduction and related reactions currently available.

Based on the nomenclature, each type of signal transduction processes a unique ST number. The alignment algorithms of metabolic pathways alignment are used to compare the similarity of signaling pathways. It makes the biopathway alignment possible.

6.1 STCDB: Signal Transduction Classification Database

6.1.1 Introduction

Signal transduction, at the cellular level in general, is the mechanism by which a signal encountered at a cell's surface (i.e. an extracellular signal) is transformed into an intracellular signal, that in turn invokes cellular responses such as proliferation, differentiation, secretion and apoptosis within the responding cells. Signal transduction refers to the movement of

* Part of Chapter 6 has been published in *NAR* [Che04c].

signals from outside the cell to the inside. Schematic representation of the signal transduction within a eukaryotic cell includes (Figure 6.1.1):

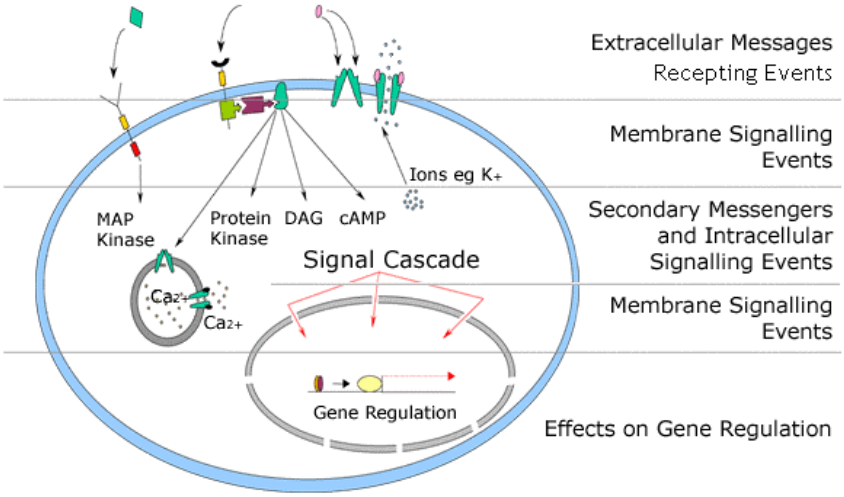


Figure 6.1.1 Basic schematic presentation of signal transduction within a eukaryotic cell.

A systematic classification scheme is given for the various types of signal transduction and related reactions currently available:

- Starting with the arrival of a signaling molecule, typically a hormone or a neurotransmitter on the cell surface.
- The signaling molecules bind to specific membrane proteins, the receptors, which are activated.
- These receptors activate proteins, which themselves stimulate other proteins in the cytosol.
- The active proteins bind to the transcription factors which when activated regulates gene expression.
- Finally, changes in gene expression initiate the biological answer of the cell to the original signal.

The movement of signals can be simple, like that associated with receptor molecules of the acetylcholine class: receptors that constitute channels which, upon ligand interaction, allow signals to be passed in the form of small ion movement, either into or out of the cell. These ion movements result in changes in the electrical potential of the cells that, in turn, propagates the signal along the cell. More complex signal transduction involves a complex network of interwoven signaling cascades (e.g. Phosphorylations by tyrosine kinases and/or serine/threonine kinases). These cascades change enzyme activities and protein conformations and cause a change in the level of a second messenger (for example calcium or cyclic AMP)

and ultimately regulate such cellular responses as proliferation, differentiation, secretion and apoptosis in the responding cells.

Extracellular signals (typically a hormone or neurotransmitter), perceived at the surface of a cell, must be translated into an intracellular response that involves a complex network of interwoven signaling cascades (e.g. phosphorylation). Signal transduction cascades cause a change in the level of a second messenger (for example calcium or cyclic AMP) and ultimately regulate such cellular responses as proliferation, differentiation, secretion and apoptosis in the cell.

With the widespread use of modern biological techniques in various fields of cellular technology, more and more cellular data are accumulated which has led to a proliferation of knowledge and its terminology. The complexity created by the crosstalk among signal transduction network makes it virtually impossible to infer by hand all the consequences that follow after the modification of one part of the network. Fortunately, a number of databases such as SPAD [Tat95], CSNDB [Iga98] and TRANSPATH [Sch01b] have been constructed to bring the signal transduction knowledge into a well-organized format, providing simple and fast access to the signal transduction system. Moreover, signal molecules and pathways are classified and illustrated by graphs [Bio01]. At present, other major databases are known describing different aspects of gene network organization, e.g. CSNDB contains and information about signal transduction mechanisms in the human cells; TRANSFAC (The Transcription Factor Database) compiles data about gene regulatory DNA sequences and protein factors binding to and acting through them; TRANSPATH is an information system on gene-regulatory pathways. It focuses on pathways involved in the regulation of transcription factors in different species, mainly human, mice and rats. Elements of the relevant signal transduction pathways like hormones, receptors, enzymes and transcription factors are stored together with information about their interaction and references in an object-oriented database. SPAD contains the structure-functional data on the mechanisms of signal transduction; EPD (the Eukaryotic Promoter Database) contains general information about promoters, as they are defined by an experimentally proven transcription start site, and their tissue-specificity. The Transcription Regulatory Regions Database (TRRD) is designed for accumulation of experimental data on extended regulatory regions of eukaryotic genes. However, none of these databases provides the solving of the whole complex of tasks necessary for a gene network effective studying, which demands analysis of the large bulk of heterogeneous experimental data. Some integrative databases or models, e.g. Genenet, E-Cell and MARG are attempting to fulfill this task, but there is still a long way to go. However, no reference has been made to the classification of transductions involving the signal reception, transportation and function.

This section presents classifications concerned with signal transductions and brings order into a nomenclature recommendation of them.

6.1.2 Classification

An important first step toward acquiring understanding of molecular and cellular function is to build systems for organizing and categorizing functions of bioprocesses. Biochemical reactions that are normally catalyzed by enzymes can be easily inferred from the enzymes involved. For example, the transformation of L-arginine to L-ornithine is normally catalyzed by arginase, 3.5.3.1. According to the Classification and Nomenclature of Enzymes (IUBMB Recommendation), it is clear that the reaction belongs to the hydrolyzation (EC 3.*.*.*). It acts on "carbon-nitrogen bonds, other than peptide bonds" (EC 3.5.*.*), and so on. A similar strategy is employed to classify signal transductions. Below is the overview listing of recommended classification, whereas the expanded "full" listing can be found at the web page: http://bibiserv.techfak.uni-bielefeld.de/STCDB.

A four-digit ST number $d_1.d_2.d_3.d_4$ denotes a particular signal transduction, with classes defined as:

d_1 := location of transduction

d_2 := type of interaction

d_3 := signal molecule's nature

d_4 := ID

The sub-class notations are briefly described as

d_1 = 1: Extracellular signal reception events

- d_2 = 1: Physical stimulation of receptors
- d_2 = 2: Binding with hormones
- d_2 = 3: Binding with non-GF cytokines
- d_2 = 4: Binding with Growth Factors
- d_2 = 5: Binding with neuronal receptors
- d_2 = 6: Binding with other ligands

d_1 = 2: Plasma membrane transduction events

- d_2 = 1: Channels operation
- d_2 = 2: Ion channel transduction
- d_2 = 3: G-proteins transduction
- d_2 = 4: Other Ser/Thr phosphorylation
- d_2 = 5: Tyr phosphorylation
- d_2 = 6: Cleavage
- d_2 = 7: Others

d_1 = 3: Plasma membrane to cytoplasmatic transduction events

- d_2 = 1: Membrane receptor releasing
- d_2 = 2: Protein-protein interaction
- d_2 = 3: Others

d_1 = 4: Intracellular signal transduction events

d_2 = 1: Ser/Thr phosphorylation

d_2 = 2: Tyr phosphorylation

d_2 = 3: Other phosphorylation

d_2 = 4: Dephosphorylation

d_2 = 5: Ubiquitination

d_2 = 6: Methylation

d_2 = 7: Deamination

d_2 = 8: Nitrosylation

d_2 = 9: GDT/GTP exchange

d_2 = 10: Dimerization

d_2 = 11: Protein-protein interaction

d_2 = 12: Others

d_1 = 5: Cytoplama to nucleoplasma transduction events

d_2 = 1: Ungrouped

d_1 = 6: Nucleoplasma to nucleoplasma transduction events

d_2 = 1: Nuclear receptor binding

d_2 = 2: Transcription factor binding

d_2 = 3: Acetylation

d_2 = 4: Histone deacetylation

d_2 = 5: Others

6.1.3 STCDB Description

6.1.3.1 Data source

The main source for the data in the STCDB database comes from the CSNDB. A minor part of the data has been extracted from TRANSPATH and Biocarta as well as the literature. Additionally, a web-based submission form is available for the users' contribution.

6.1.3.2 Database structure

The STCDB database contains data for each type of characterized signal transduction for which an ST number has been provided. The entries in the database data file (ST number.html) are structured so as to be usable by human readers, as well as by computer programs. Each entry in the database is composed of lines. Different types of lines, each with its own format, are used to record the various types of data that make up the entry. The general structure of a line is the following:

- `ST number`
- `Recommended name`

- Alternative names (if any)
- Reference
- Pointers to the CSNDB entrie(s) that correspond to the signal transduction (if any)
- Pointers to the BioCarta entrie(s) that correspond to the signal transduction (if any)

A search interface for the Internet service provides two kinds of direct search: by keyword and by ST number. Searching by keyword allows the user to input free text that might be found in the content of each data file. A ST number can be chosen which will restrict the search to the specific entry. A wild card (*) search is available and more than one word in the text field will find a match to either word. Screenshots of STCDB are shown in Figure 6.1.3.2.

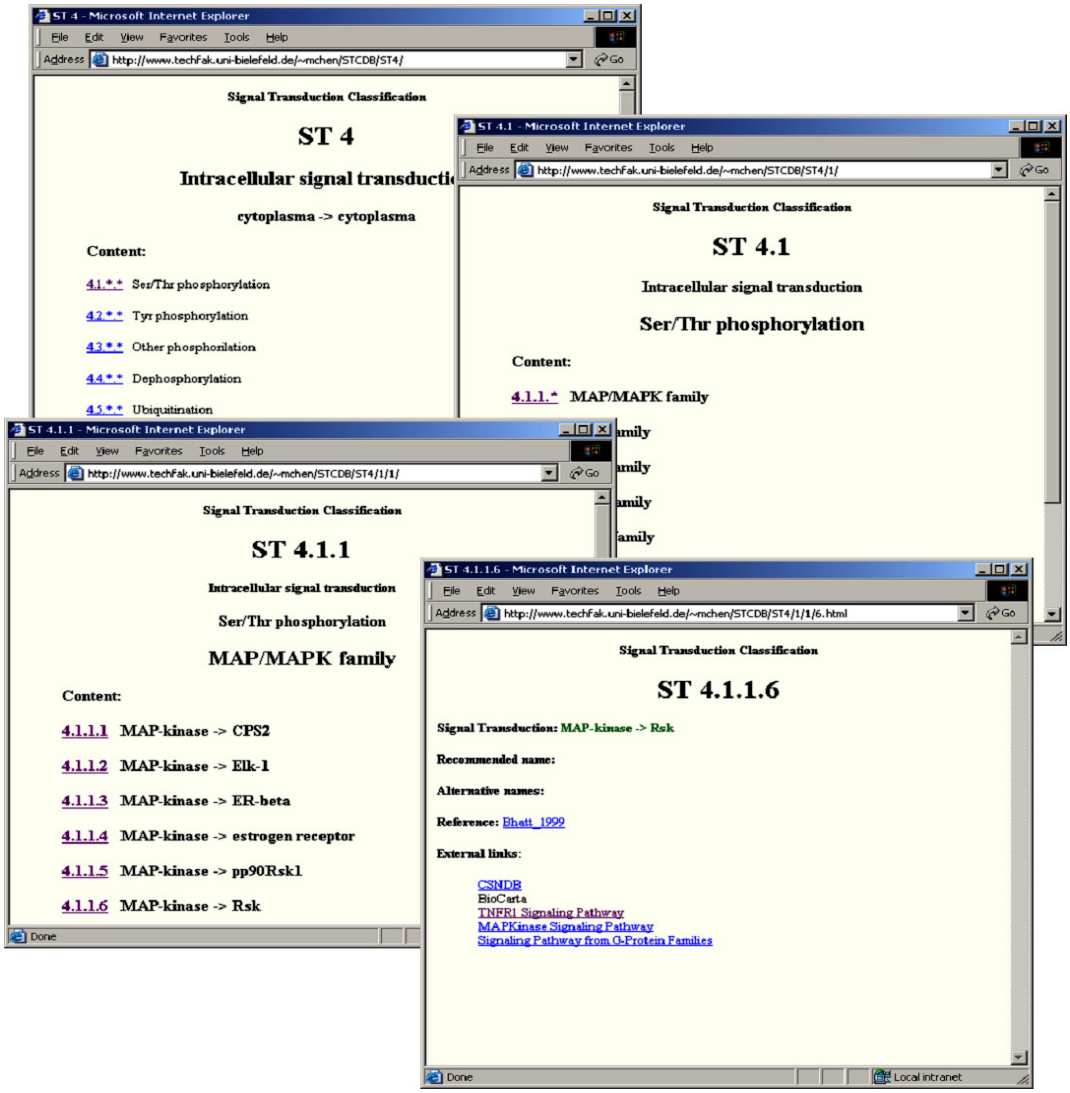


Figure 6.1.3.2 An example of ST entries.

6.1.3.3 Latest data update

STCDB is regularly updated to reflect updates and additions to the classification. We also update the CSNDB and BioCarta pointers, correct eventual errors, and complete the information concerning synonyms using the literature. We welcome and encourage any type of feedback.

The latest data release of STCDB was given out on Dec. 2003. Sequence corrections, mainly frame shift errors, led in most cases to the modification of classification. In a few cases, more dramatic changes, such as merging several entries or adding/removing entries, were required. Furthermore, additional corrections of signal transduction classifications resulted from a revised analysis of sources data. Currently STCDB contains over 486 entries/pages, 400 cited references and 700 external hyperlinks. The numbers of entries of main class of signal transduction is shown in Table 6.1.3.3.

Table 6.1.3.3 Summary of signal transduction classification entries in the latest release (Dec. 2003).

Signal transduction classification	Entries
ST 1.*.*.*	176
ST 2.*.*.*	53
ST 3.*.*.*	22
ST 4.*.*.*	201
ST 5.*.*.*	4
ST 6.*.*.*	31

We would like to encourage users to submit their request for a new classification via the web-based submission form (http://www.techfak.uni-bielefeld.de/~mchen/STCDB/submit.html or to contact us directly by e-mail if they have large data sets. Further analyses and database searching would validate every record that is entered in this way. We would also reply on assistance from a number of specialist advisors and communication with the scientific community in general to maintain accuracy.

6.2 Signaling Pathway Alignment

Signaling pathways are not only interconnected to other pathways in the cell. One of additional perspectives to analyze their interactions and regulations is signaling pathway alignment. Signaling pathway alignment reveals differences in signal transduction flux, conversion and regulation in different species. In this section, we present signaling pathways as ST number sequences and exploit the PathAligner system to align them.

6.2.1 ST Representation of Signalling Pathways

To enhance the exploration of signal transduction in a pathway context, one requires an application that allows the visualization of the signaling process in a pathway map. The

current graphic representation of signaling pathway is based on a concept description and lack of a fully understanding mechanism or a common taxonomy. However, according to our classification of signal transduction, it is possible to commonly represent signaling pathways by a "metabolic pathway"-like structure in which the proposed nomenclature of signal transductions as the ST classification system.

Figure 6.2.1A shows a ST representation of signaling pathways of IFN-gamma pathways.

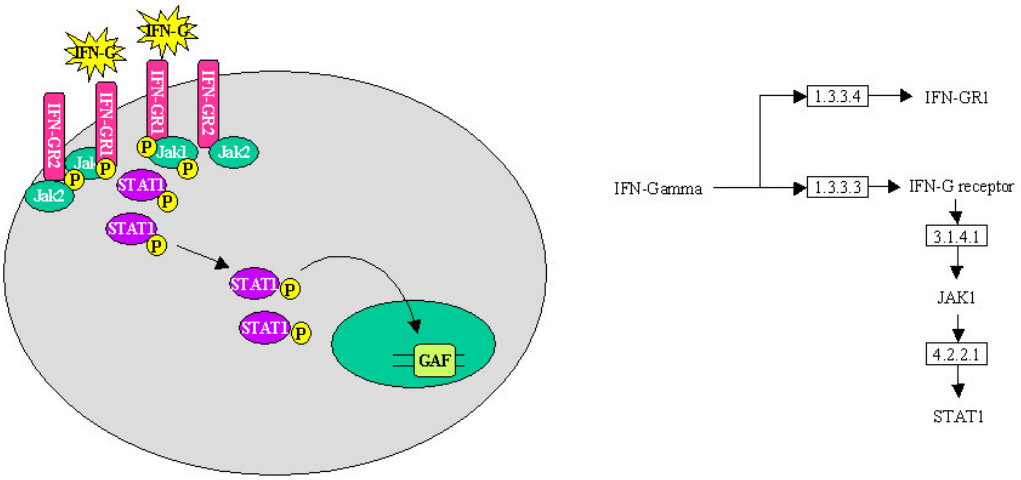


Figure 6.2.1A ST representation of interferon gamma signaling pathway. The ST number consists of four distinct numbers that classify the signal transduction based on our classification. For example, the signal transduction IFN-Gamma -> IFN-GRJ has the ST number 1.3.3.4, which designates an external event (class 1), binding with non-GF cytokines (subclass 1.3) with IFN GR as receptor (sub subclass 1.3.3). The number 4 designates the fourth transduction in this class.

A strong advantage of ST numbers is that they provide unique identifiers for signal transductions. For example, when comparing the signal transductions of two signaling pathway, the ST number is used to determine if two signal transductions have the same function without the need to understand the multiple and confusing names that can be used for the same signal molecule. Similarly, the multiple names used for signal molecules cause confusion when we try to determine if two signal transductions refer to the same molecule. Extensive synonym lists for signal molecules are essential in a signaling database.

Based on the STCDB classification and nomenclature, graphic signaling pathways are produced (Figure 6.2.1B).

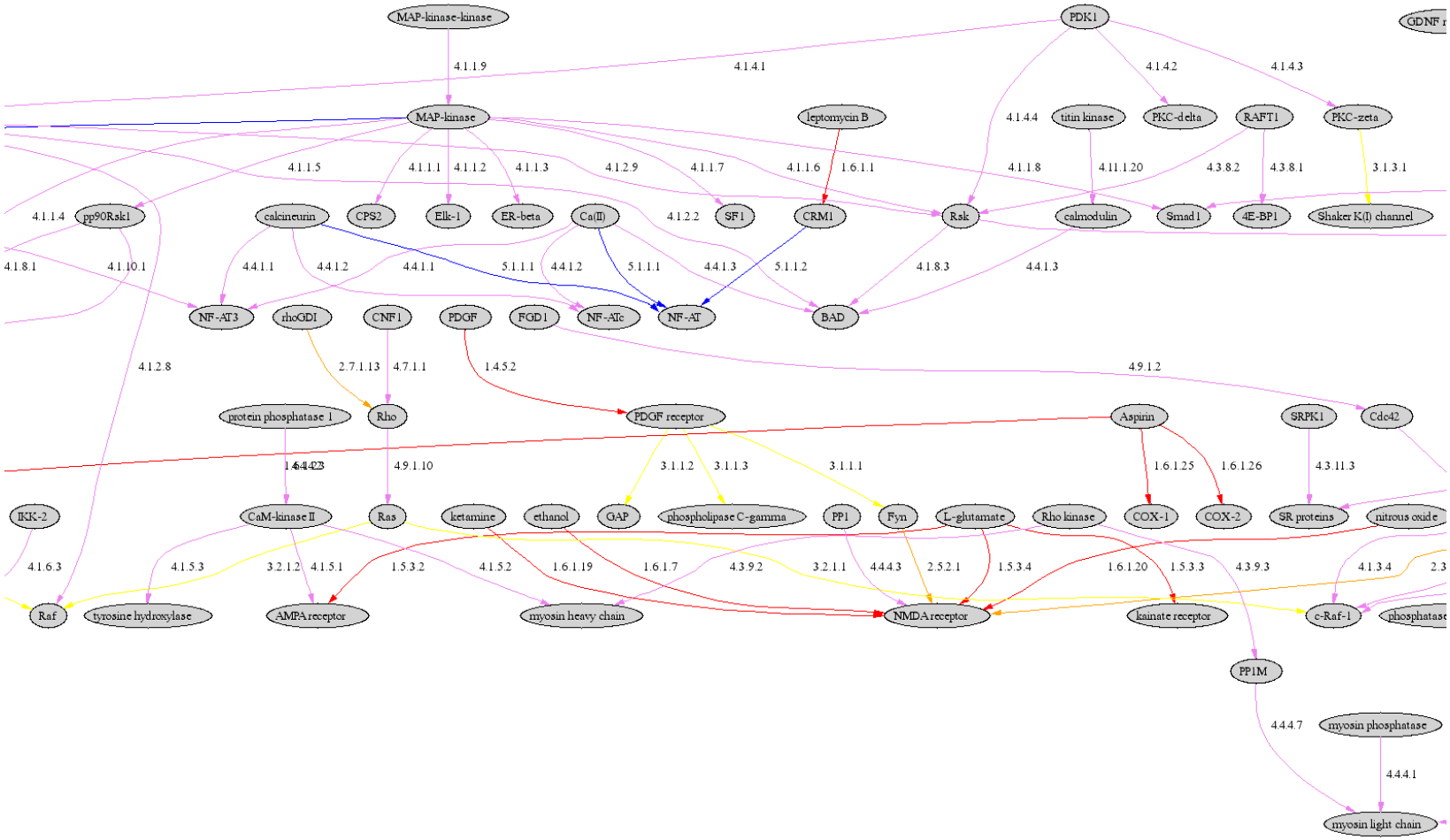


Figure 6.2.1B A part of signal transduction pathways based on the STCDB entries. The whole network graphic representation is available at STCDB web site.

6.2.2 An Alignment Example

By applying the metabolic pathway alignment algorithm, it is possible to align signaling pathways. We extract partial signaling pathways from STCDB and list them below.

PKA::PKA->2.3.3.1->phosphorylase kinase->4.3.11.13->glycogen phosphorylase
Apaf1::Apaf-1->3.3.1.1->caspase-9->4.12.8.13->caspase-3->4.12.8.3->Acinus
Cam-KK2::CaM-KK->4.1.5.5->PKB->4.1.2.3->caspase-9->4.12.8.13->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.1->NF-kB
Cam-KK3::CaM-KK->4.1.5.5->PKB->4.1.2.3->caspase-9->4.12.8.13->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.2->I-kB-alpha
CDK5_1::CDK5->4.3.3.3->PAK1->4.1.3.2->LIMK-1->4.3.11.10->cofilin
PDK2_2::PDK2->4.1.4.5->PKB->4.1.2.3->caspase-9->4.12.8.13->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.2->I-kB-alpha
Lck1::Lck->2.5.1.1->VAV->4.12.2.1->Rac->4.9.1.6->PAK3->4.1.3.4->c-Raf-1
Lck2::Lck->2.5.1.1->VAV->4.12.2.1->Rac->4.9.1.5->p35->4.3.3.3->PAK1->4.1.3.2->LIMK-1->4.3.11.10->cofilin
cytochrom2::cytochrome c->3.3.1.1->caspase-9->4.12.8.13->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.2->I-kB-alpha
Cam-KK1::CaM-KK->4.1.5.5->PKB->4.1.2.3->caspase-9->4.12.8.13->caspase-3->4.12.8.6->presenilin1
Lck3::Lck->2.5.1.1->VAV->4.12.2.1->Rac->4.9.1.5->p35->4.3.3.3->PAK1->4.1.3.3->MLCK->4.3.11.11->myosin light chain
cytochrom1::cytochrome c->3.3.1.1->caspase-9->4.12.8.13->caspase-3->4.12.8.3->Acinus
CDK5_2::CDK5->4.3.3.3->PAK1->4.1.3.3->MLCK->4.3.11.11->myosin light chain
procaspase-3::procaspase-3->4.12.8.2->caspase-3->4.12.8.3->Acinus
caspase-8_1::caspase-8->4.12.8.11->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.1->NF-kB
CDK5_3::Cdr2->4.3.4.2->Wee1->4.3.10.1->CDK1->4.3.10.2->cyclin B
survivin::survivin->4.12.7.1->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.2->I-kB-alpha
caspase-8_2::caspase-8->4.12.8.11->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.2->I-kB-alpha
Apaf2::Apaf-1->3.3.1.1->caspase-9->4.12.8.13->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.2->I-kB-alpha
PDK2_1::PDK2->4.1.4.5->PKB->4.1.2.3->caspase-9->4.12.8.13->caspase-3->4.12.8.10->MEKK1->4.1.11.3->IKK->4.1.6.1->NF-kB

In order to distinguish these pathways, a temporary name is labeled. That is, each pathway begins with “name::”. When we input these pathways to the PathAligner’s interface and perform multiple pathway alignment, the alignment result and screenshot, are shown in Figure 6.2.2.

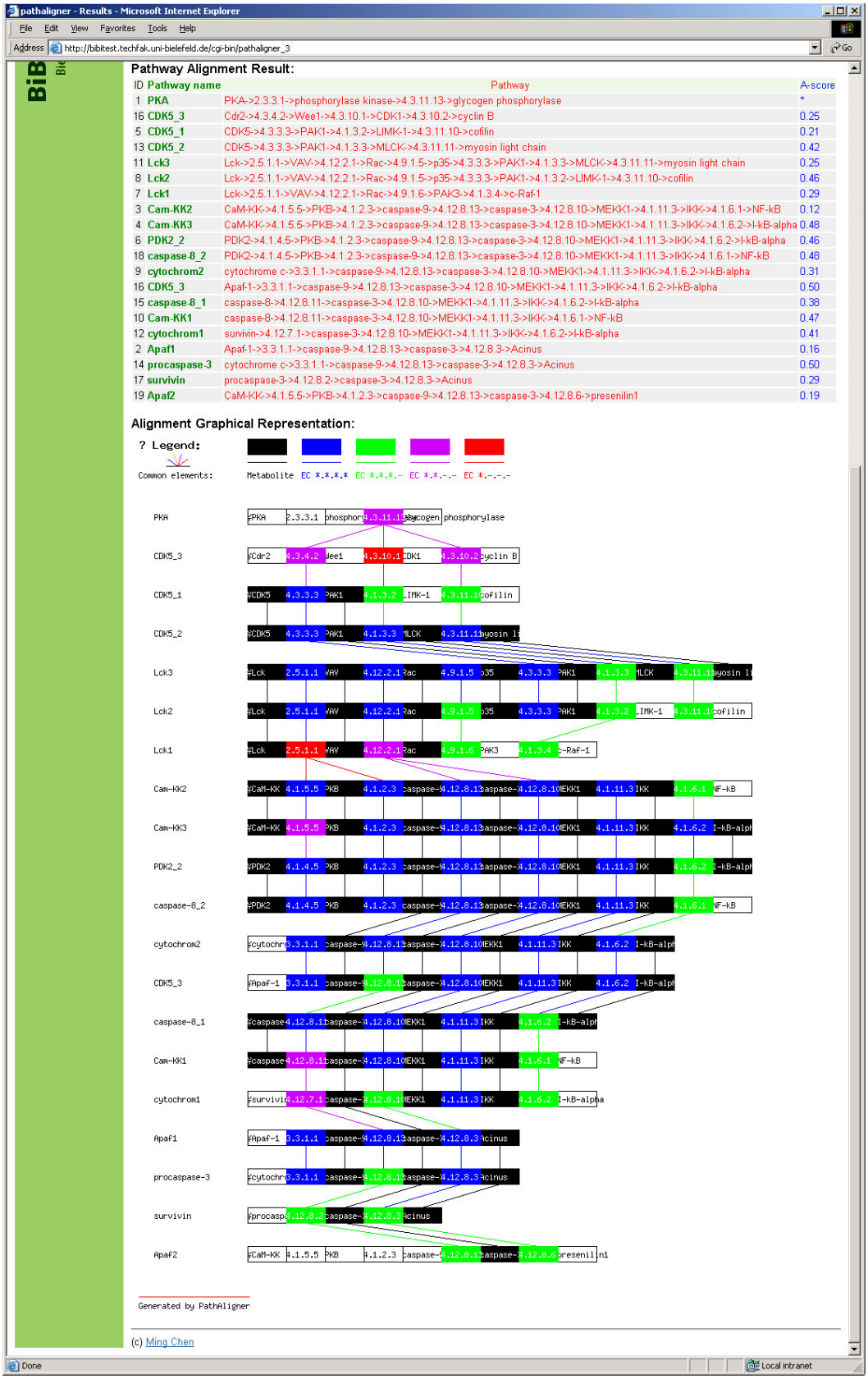


Figure 6.2.2 The multiple alignment result of the listed signaling pathways.

6.3 Biopathway Alignment

So far, biochemical reactions that are normally catalyzed by enzymes can be easily inferred from the enzymes involved. For example, the transformation of L-arginine to L-ornthine is normally catalyzed by arginase, 3.5.3.1, then it is clear that the reaction is catalyzed by a enzyme that is a hydrolase, acting on the "carbon-nitrogen bonds, other than peptide bonds" in "linear amidines". Signal transductions can also be inferred from its ST numbers. The ST 4.1.1.6 signal transduction indicates that the process takes place in the intracellular compartment, it is a kind of Ser/Thr phosphorylation, acting from the MAP/MAPK family to the specific molecule Rsk (MAP-kinase -> Rsk).

However, this partial understanding is very artificial. Cells respond in this interconnected fashion, involving several pathways. We need integrated information and generalized biopathway representation and analysis. In fact, by combining the EC numbers and ST numbers, the integrative biopathway alignment is reliable.

6.4 Summary

Signal Transduction Classification Database (STCDB) is a database of information relative to the classification of signal transduction. It is primarily based on a proposed classification of signal transduction and it describes each type of characterized signal transduction for which a unique ST number has been provided. This document presents, in a first version, the classification and nomenclature of signal transduction. Approved classifications are available for browsing and querying at http://bibiserv.techfak.uni-bielefeld.de/STCDB.

The ST number is a 4-level hierarchical structure, which makes it possible to exploit our metabolic pathway alignment algorithm to perform signaling pathway alignment. We have presented a graphical representation of signaling pathways with ST numbers indicating every signal transduction. An example of signaling pathway alignment has been presented. By combining the EC numbers and ST numbers, alignment of biopathways is reliable.

In the next chapter, we will present a concrete example of biopathway, the urea cycle, for the systems analysis. Some examples of urea cycle have been discussed in the prediction and alignment parts. We will mainly focus on the modeling and simulation and further analysis of urea cycle disorders.

Chapter 7
A Biological Application

In this Chapter a case study is presented to describe our approaches that have been discussed in the previous chapters, mainly in the Petri net chapter. Several applications of PathAligner have been described in Chapter 4 and 5.

For the past century, studies of urea cycle disorders have focused research efforts to improve clinical diagnosis and management. The availability of human genome sequences and other metabolic data provides us with a challenging opportunity to develop computational tools for systematically analyzing urea cycle disorders. We exploit the current data available, and integrate these from genomics and poteomics at novel levels of understanding urea cycle disorders. We also systematically analyze transcription factors and signaling pathways involved in the urea cycle biopathways.

7.1 Urea Cycle and its Regulation

In human cells, excess nitrogen is removed either by excretion of NH_4^+ (of which only a little happens) or by excretion of urea. Urea is largely produced in the liver by the urea cycle, a series of biochemical reactions that are distributed between the mitochondrial matrix and the cytosol (Figure 7.1A). The cycle centers around the formation of carbamoyl phosphate in hepatocyte mitochondria to pick up NH_4^+ incorporate it into ornithine to make citrulline that is transported to the cytosol where aspartate is added. As urea is removed it is converted back to ornithine that goes back into the mitochondria to start over again. Deficiencies in the urea cycle enzymes lead to excessive NH_4^+ and its intermediates accumulation, which results in neurological disorders. Any of the five enzymes of the urea cycle may be deficient: carbamoyl phosphate synthetase (CPS) deficiency, ornithine transcarbamylase (OTC) deficiency, citrullinemia, argininosuccinic aciduria and argininemia.

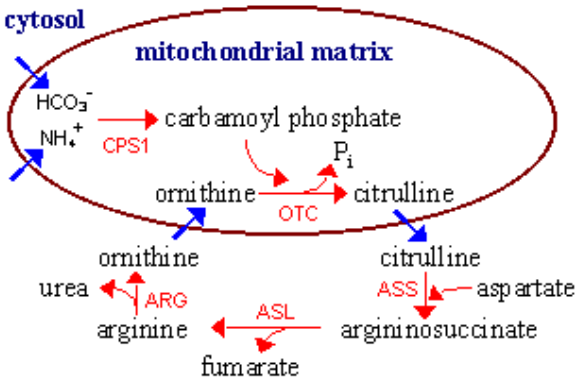


Figure 7.1A Key enzymes in regulation of urea cycle in cells. CPS1: Carbamyl phosphate synthetase, EC 6.3.4.16; OTC: Ornithine transcarbamylase, EC 2.1.3.3; ASS: Argininosuccinate synthetase, EC 6.3.4.5; ASL: Argininosuccinate lyase, EC 4.3.2.1; ARG: Arginase, EC 3.5.3.1.

Although the urea cycle was discovered by Dr. Hans A. Krebs early in 1930's, analysis of the urea cycle so far has never been systematically explored. This chapter therefore will also focus on the possibility of integrative analysis of the urea cycle within the scope of systems biology. An integrative model is built. A Petri net model is constructed in order to estimate the regulation both on genomic and metabolic levels. Simulations can be used to test the physico-chemical limitations and feasibility of certain proposed reactions. We are also going to analyze the genetic variations and figure out the regulation of signaling pathways. One of the aims is to highlight at large in the identification and treatment of urea cycle disorders, and give some hints on the systems analysis of inborn errors of metabolism.

Figure 7.1B shows a graphical representation of the urea cycle using the objects presented above for describing entities and interactions. It shows an intricate network that links entities and interactions. This network includes not only the succession of chemical reactions that lead to the transformation of CO_2 and NH_4^+ to urea, but also the regulation of gene expression and enzymatic activities. It furthermore displays (e.g. asparate, fumarate) the links to other pathways, which are, to preserve clarity, not detailed in the graph.

In some cases if the complete interrelationships of the biopathway is unclear, or only a rudimentary pathway is provided, we can use PathAligner to retrieve metabolic information and reconstruct the complete network, as discussed in Chapter 4.

7.2 Petri Net Model

Based on the proposed modeling strategy in Chapter 3, a hybrid Petri net model of the urea cycle and its transcriptional regulation is presented (Figure 7.2). The model of intracellular urea cycle is made of the composition of the gene regulatory network and the metabolic pathways. It comprises 152 Petri net elements, 14 kinetic blocks, 39 dynamic variables, and 22 reaction constants.

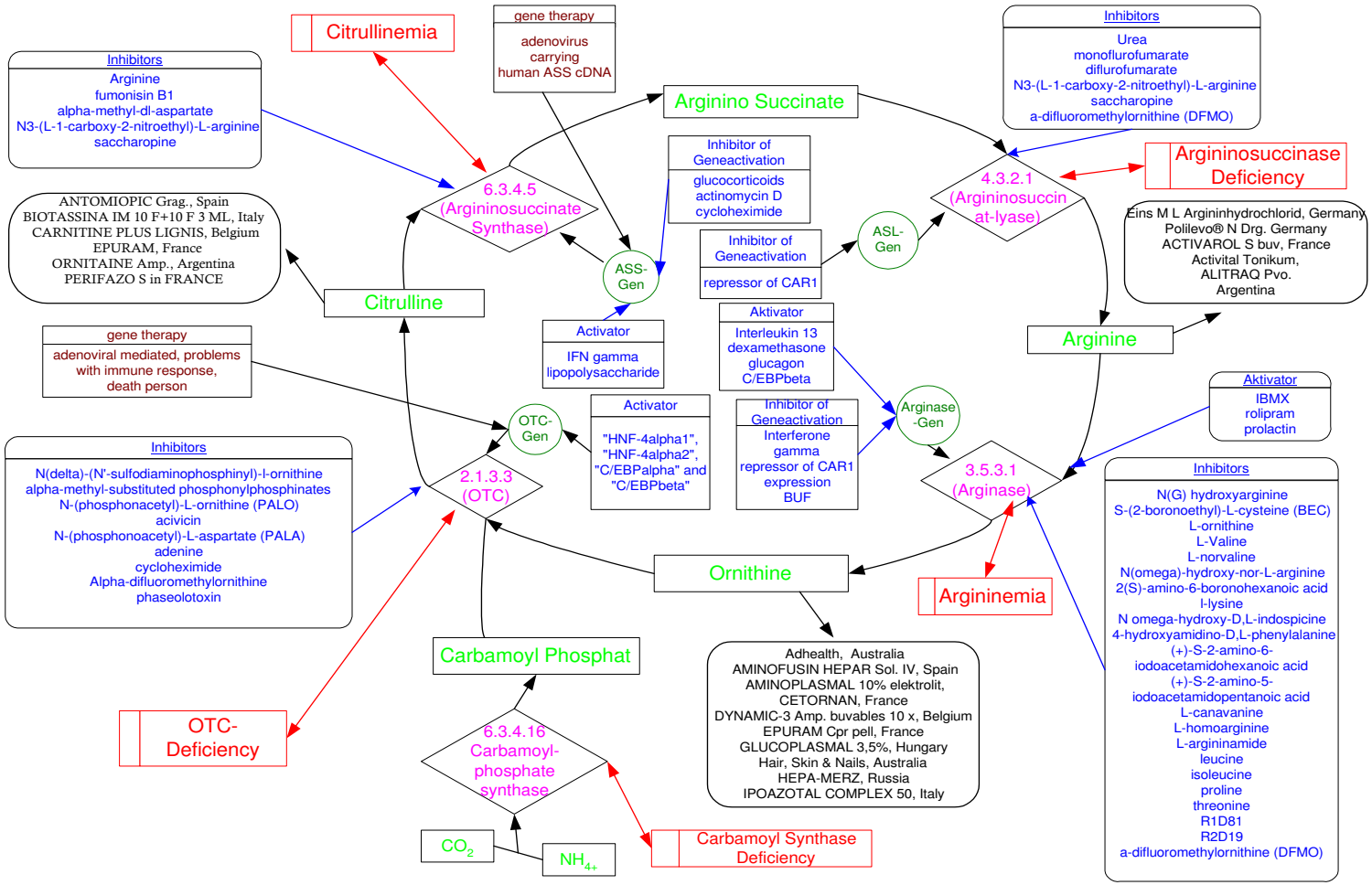


Figure 7.1B Schematic diagram of urea cycle. Data sources: Metabolic pathway (enzyme reactions) from KEGG and BRENDA; Gene regulation: TRANSFAC; Drug information: MDDrugDB (http://edradour.cs.uni-magdeburg.de/~rkauert/MDDrugDB/Main.htm; drawing by Dr. Ralf Kauert).

Experimental data, partially listed in Table 7.2, are used for the initial evaluation of certain parameters of enzymatic reactions with the system. The value of model parameters lacking in the literature are verified through numerical experiments or modifed from several references.

Table 7.2 Some kinetic parameters of enzyme reactions in human cells.

Enzyme	Substrate (mean concentration, mM)	Compartment	K_m, mM	K_{cat}, S^{-1}	Reference
CPS1	HCO_3^- (0.05), NH_4^+ (0.025), ATP	Mitochondia	HCO_3^-, 6.7 NH_4^+, 0.8 Mg ATP, 1.1	17	[Pie80]
OTC	Carbamoyl phosphate (0.001), L-ornithine (0.05)	Mitochondia	CP, 0.16 L-ornithine, 0.40	180	[Scr97]
ASS	L-citrulline (0.02), L-aspartate (0.325), ATP	Cytoplasm	L-citrulline, 0.03 L-aspartate, 0.03	400	[Scr97]
ASL	Argininosuccinate (0.034)	Cytoplasm	Argininosuccinate 0.017	3	[Pie80] [Scr97]
ARG	L-arginine (0.06)	Cytoplasm	L-arginine, 10	2200	[Scr97]

The dynamic behavior of the model system, such as metabolite fluxes, NH_4^+ input and urea output are well described with continuous elements, while control of gene expression is modeled using discrete ones due to the insufficiency of explicit expression data. Nevertheless, when explicate knowledge about expression levels of the enzymes are available; it is possible to exploit our model of gene regulatory network to handle realistic gene expression data with state equations. The initial values of variables are assigned and tuned so that the model system behavior would comply maximally with available experimental data on the dynamic characteristics of the system's behavior, based on the following considerations:

The availability of ammonia or amino acids (denoted as NH_3) is ingested continuously from plasma into mitochondria with a stable speed, i.e. the changes of ammonia concentration due to the rate of protein metabolism are not taken into account. The concentration of nitrogen excreted (urea) in plasma ranges from 3mmol/L to 8mmol/L and is then discharged. The degradation of all enzymes is 0.001 times of their concentration.

In the model, inhibitor arcs are also used to present negative effects of repressors and/or inhibitors to gene expression. In order to get a better understanding of these relationships, several test arcs are used, e.g. the test arc between asparate and transition of ASS. On the biochemical reaction level, negative effects of metabolites are expressed as enzyme inhibitions that include competitive inhibition, noncompetitive inhibition, irreversible inhibition and feedback inhibition. Sequentially, the regulation of the urea cycle enzyme activities can be modeled in these two ways. First, gene expressions that are regulated by activators and inhibitors control enzyme synthesis, while enzyme synthesis and degradation determine the amount of enzymes. Second, the activities of these enzymes can be altered during metabolic catalyzations.

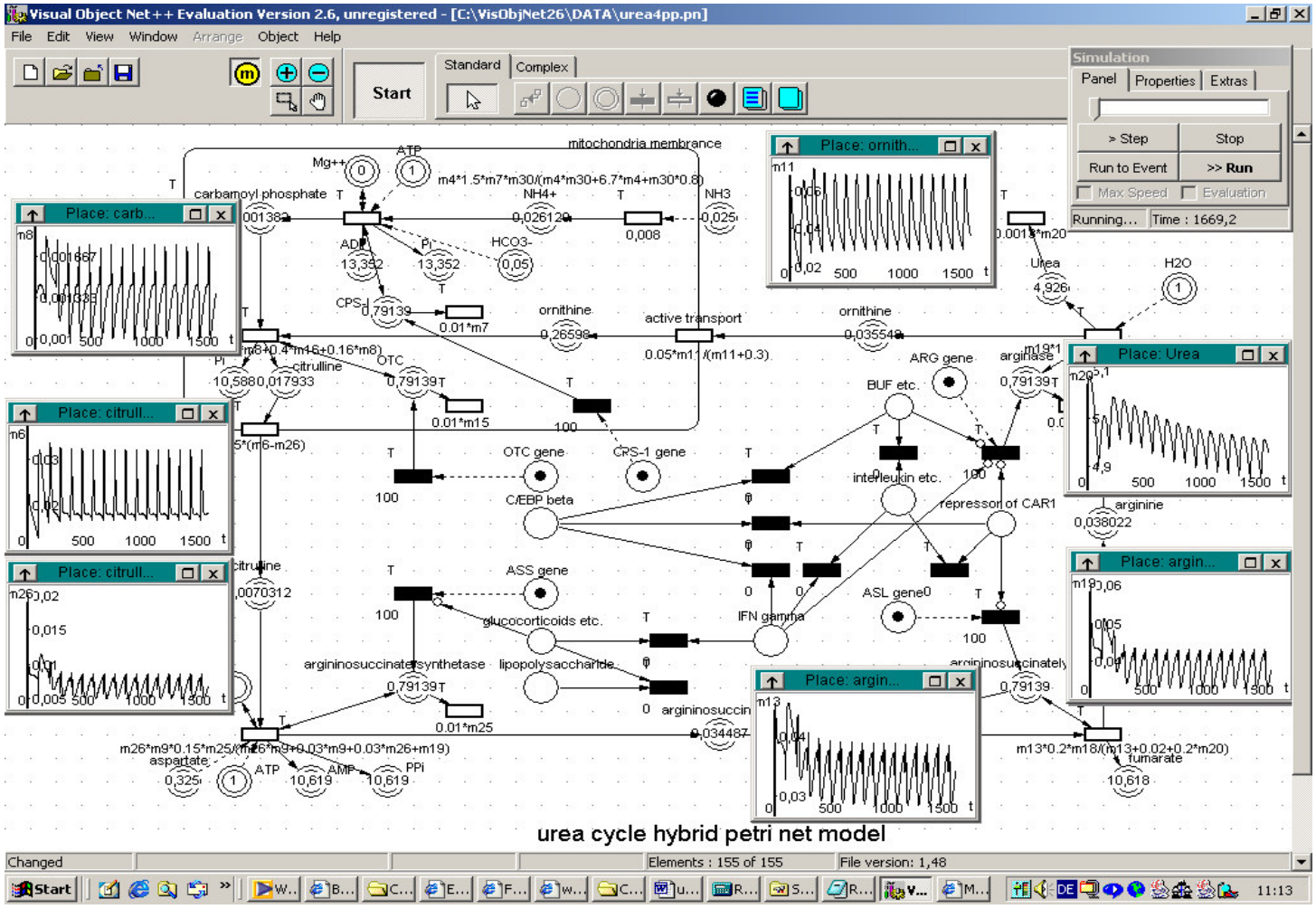


Figure 7.2 Hybrid Petri net model and simulation results of the gene regulated urea cycle metabolic network.

7.3 Investigation of the Behaviors of the Model

The formalization of the urea cycle model allows the quantitative simulation of this metabolic pathway. Dynamics of the main components on the model regulating the urea cycle were also shown in Figure 7.2. Moreover, several tests on interfering the fluxes intentionally are conducted and results are observed in Table 7.3.

Table 7.3 Interfering tests on the urea cycle Petri net model.

Interfering test	Value of metabolites					Urea cycle defect
	NH4+	Citrulline (plasma)	Argininosuccinate (plasma)	Ornithine (plasma)	Arginine	
CPS1 blockade	↑	↓	↓	↓	↓	Carbamylphosphate synthase deficiency
OTC blockade	↑	↓	↓	↑	↑	Ornithine transcarbamylase deficiency
ASS blockade	↑	↑	↓	↑	↑	Argininosuccinate synthase deficiency
ASL blockade	↑	↑	↑	↑	↑	Argininosuccinase deficiency
ARG blockade	↑	↑	↓	↓	↑↑	Arginase deficiency
Membrane transportation blockade	↑	↑	-	↑↑	↑	HHH syndrome

Note: the symbol "↑" indicates an increment of the concentration, "↑↑" indicates a quick increment, while "↓" indicates a decrement. "-" indicates no dramatic changes of the concentration.

The urea cycle eliminates excess nitrogen. A high concentration level of ammonia in the cell results in hyperammonemia that leads to coma and even death. Laboratory studies can reveal elevated arginine levels, mild hyperammonemia, and a mild increases in urine orotic acid. The diagnosis now can be confirmed by enzymatic analysis in the model. On high-protein diets or under starvation, proteins are degraded and amino acid carbon skeletons are used to provide energy. Thus the quantity of nitrogen that must be excreted is increased, but the amino nitrogen must be excreted. To facilitate this process, enzymes of the urea cycle are controlled by regulating the expansion of their genes to enhance the concentration of enzymes. As the urea cycle takes place both in mitochondria and cytoplasma, these effects can also be caused by membrane transportation deficiencies. Some mitochondial membrane diseases, e.g. ornithine transporter deficiency, surely effect the transportation of ornithine into matrix and result in high concentration of ornithine accumulation in plasma, which creates a feedback regulation to the transition of arginine into urea and finally hyperammonnemia. From the model we know the treatment for defects in urea cycle enzymes could be either to limit input of ammonia (limit protein intake) or to replace missing intermediates from cycle (supplement with arginine or citrulline). Patients with OTC deficiency benefit from citrulline supplementation, because citrulline can accept ammonia to form arginine.

7.4 Treatment of Urea Cycle Disorders

It is important to understand the mechanism of urea cycle disorders in order to properly treat the disorder. Once the disease is diagnosed and the model is presented, several treatments are proposed. Limit the toxic ammonia by placing the patient on a diet with limited amount of food protein, is generally the first course of treatment which is considered. Another treatment is to remove the toxic ammonia through alternative pathways. Scientists developed methods to exploit other vehicles of waste nitrogen synthesis and excretion to substitute for the defective urea pathway. Batshaw and Brusilow et al. [Bat82] devise several ways of allowing people to remove ammonia without having to make urea. One way was to give the patient large doses of the preservative sodium benzoate. An enzyme in our livers couples the benzoate molecule with a molecule of the amino acid glycine. The resulting compound, hippurate, is rapidly removed by the kidney and is excreted in the urine. The liver can produce glycine from ammonia, carbon dioxide, and a folic acid compound. Each glycine produced in this way removes one ammonia from the body. Other compounds such as arginine, sodium phenylacetate, and sodium phenylbutyrate can remove ammonia by similar mechanisms. Buphenyl® (sodium phenylbutyrate, Ucyclyd Pharma, Hunt Valley, MD, 1996) has been developed and approved by the FDA. A prospective treatment trial of this drug for neonatal onset urea cycle disorders showed that cognitive function is improved. Phenylbutyrate also has a dramatic effect on the survival of patients with arginosuccinate synthetase deficiency, another urea cycle disorder. Despite treatment and dietary manipulation, it is not possible to restore patients with neonatal urea cycle disorders to a state of normal or near normal health. The greatest impact of phenylbutyrate is its efficacy in treating late onset disease.

Considering the insufficiency of five enzymes involved, we might also be able to activate enzymes with cofactors, as some enzymes require non-protein cofactors for their activity. Another interest involves two enzymes of the urea cycle, argininosuccinate synthetase (AS) and argininosuccinate lyase (AL), and their role in the arginine-citrulline cycle. The primary physiological role of AS and AL is in the urea cycle, but along with nitric oxide synthase (NOS), these enzymes form the arginine-citrulline cycle which is found in all mammalian tissues. The significance of the arginine-citrulline cycle was only recently realized with the discovery that arginine-derived nitric oxide (NO). The key cell signaling molecule, was responsible for the hypotension in septic and cytokine-induced circulatory shock. The rate-limiting step in the production of NO is the availability of arginine. Since AS is the rate-limiting step in the de novo production of arginine, AS, but also AL, are attractive drug targets. Inhibitors of these enzymes have the potential not only to be useful in the treatment of septic shock, but could also increase the usefulness of a number of anticancer agents (e.g. IL-2), as co-administration of an inhibitor would suppress the dose-limiting hypotension caused

by these drugs. Catalytic mechanisms for the proteins have been proposed and are currently being tested. The design of novel inhibitors for AS has been initiated (Quote as reported by The Hospital for Sick Children at the University of Toronto).

After 20 years of experience, it must be acknowledged that alternative pathway therapy has limited effectiveness in preventing hyperammonemia and must be combined with effective dietary management. Therefore in children with neonatal-onset disease or in those with very poor metabolic control, liver transplantation should be considered. There should also be the continued search for innovative therapies that may offer a more permanent and complete correction, such as gene therapy [Bat01].

7.5 Gene Therapy and Expression

Because the basis of the disorders is a defect in a gene, researches have been working on ways of getting a working gene into cells. Scientists have established that in the animal model, sparse fur (spf/Y) mouse, partial correction with gene therapy may be sufficient to normalize urea synthesis. Because the hepatotropic properties of human adenoviruses make them suitable vectors if injected parenterally, and because the hepatocyte is so easily accessible via the circulation, in-vivo approaches to gene therapy have been developed [Ye96] [Ye97] [Ye00]. However, the current therapy is unsatisfactory for humans. Optionally, we would like to target the working gene to the right cells and have it regulated and expressed just as well as the normal gene would be. Single nucleotide polymorphism (SNP) and transcription factor binding sites are two aspects that have to be considered.

In the progress of Human Genome Project, scientists recognized that the existence of SNP in genome is helpful to explain the rich diversity of individuals, and the difference of susceptibility to diseases. A single base variation may cause gene function abnormities. Therefore, searching and studying SNPs has become an important objective of biomedical informatics. Appendix E shows the computationally annotated mutations of genes in the urea cycle. Further information can be obtained by browsing the related web-pages at http://mutdb.org/cgi-bin/Search.py?GOCODE=0000050. However, these mutations are meant to be used for basic research and not to make clinical decisions. In this section, we focus on the discovery of transcription factor binding sites by computational searching the 1kb upstream promoter region sequences.

A TRANSFAC database search for the transcription factor binding sites, using the human promoter sequences that are provided by UCSC [http://genome.ucsc.edu/], are shown in Figure 7.5 and in Appendix F. All potential binding sites of the urea cycle genes are summarized in Table 7.5. The search found 23 functional binding sites. ARG and OTC share 4 binding sites, which means that the expression of ARG and OTC might be simultaneously regulated by Cdx-2, Cdc5, Nkx2-5 and POU1F1. Nkx2-5 also affects ASL regulation.

ARG promoter/upstream 1kb
>hg16_refGene_NM_000045 range=chr6:131873935-131874934 5'pad=0 3'pad=0
revComp=FALSE strand=+ repeatMasking=none
ataattttaaagtcggaaggatctttaaggtgcctttattttaaattcat
acttttgtatggtgacaaatggtagctcaggggcatagaggttgacacct
tcccagcatttagactataagctcgacggttaagtggattcagaatggca
gagactaaatcccgacttttcttctacagcctatgttggcaacgggtctg
agcttcagtttattcatcagtataatggcaccaatgatgagacttcacat
aaaattggtataaatataaatatggtattttgaaacagaactgcatcgga
cacatggtaaaaactcaatgttagctatttttatttctatactttgatta
tgatatgattctacaattattttcctgtacaccatacttcaaaaatggta
acctctctgggttaccaatcaagtaactaatttttaaagtaatcatcaa
aaaaggaagttatatactctttattatattatacccctaaaagtttatgaa
atgtgtctcatggattaaccatttaccctcatgtgtgaaatctcaactca
ggatttttagggctggaagggatgtgacagacgatcttgccaagcccggcc
cttcttctacaaggacgtcttcagagatctggaggaggaaagggccttgc
cctgagttcgctgagccagaacaataggacttcttctgtagttgtgaaac
ttgtcagttgttgaagtcaggttaatgtcatctggctggctttttaaaag
ggtgtgaagtgagaacatgaataattgtcacttgattagagacctagact
cagagttaggttactccatgtatgaagtaaccccatatagttacttcata
catggagtaaccatatagttactccatgtatgaaaaattgcaagactgtt
gactgtcattctttggtttagtgggtggagccagctgtcctcattagata
aaggttgtttattcaacccaagtataaatggaaaaaaaagatgcgccctc

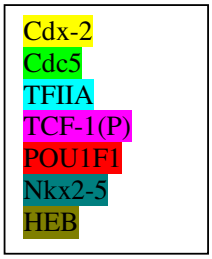


Figure 7.5 Computational prediction of transcription factor binding sites of the human ARG genes.

Table 7.5 List of potential transcript factor binding site of the urea cycle genes.

	ARG	ASL	ASS	CAD	OTC
AML1		+	+		
Cdx-2	+				+
Cdc5	+				+
E2F-1			+		
GATA-4			+		+
HEB	+				
HNF-3alpha					+
HNF-4alpha2					+
Lentiviral Poly A					+
MAZ			+		
NF-1			+		
NF-kappaB			+		
Nkx2-5	+	+			+
Pax-2					+
Pax-4a		+			+
POU1F1	+				+
POU2F1					+
RFX1			+		

SREBP-1				+	
TCF-1(P)	+				
TFIIA	+				
USF			+		
Xvent-1				+	

7.6 Signaling Pathway and Assoiated Diseases

Further analysis on the gene and their transcript factors are conducted. We obtain a list of signaling events that effect the gene expression of urea cycle by browsing the BIOBASE database. A graphical layout of the signaling pathways is constructed by using the Biolayout tool [Enr01] (Figure 7.6A).

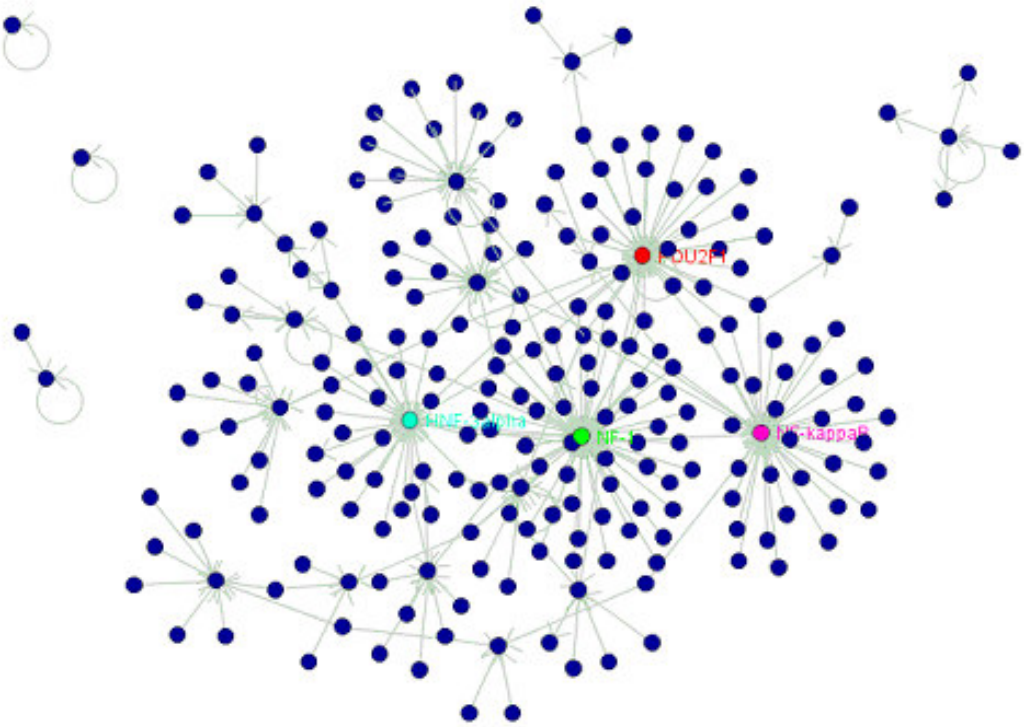


Figure 7.6A A graphical layout of the signaling pathways involved in the urea cycle. Colored nodes are those molecules with intensive divergence interconnection with others. They are NF-1 (57), NF-kappaB (38), POU2F1 (38) and HNF-3alpha (29).

Molecules with high degree of convergences are calculated. They are Cdx2 (5), SREBP-1 (4), NF-1 (2) and POU1F1 (2). Obviously, the degree of divergence is large, greater than that of convergence, which seems to be a common phenomenon in cellular signaling pathways. A possible explanation could be that cells have to reserve much more regulation mechanisms. On

one hand, activation/inactivation of important molecules are regulated by many other molecules. On the other hand, most molecules regulate only a small amount of specific molecules. This mechanism can enable cellular functions to be of robustness, sustain cells in face of various environment changes.

We also investigate the associated diseases of these signaling pathways. By querying against biological databases, such as Swiss-Prot and KEGG, all related enzymes can be retrieved. Then, searching the BRENDA database helps to determine the involved diseases. All diseases that are regulated by these signaling pathways are listed in Appendix G. On the left column, we do not consider the redundancy of enzymes that encoded by different genes. For example, there are 10 hits of the enzyme protein kinase (EC 2.7.1.37) that is involved in various diseases, such as "acromegaly", "adhesions", "amyotrophic lateral sclerosis", "anemia, sickle cell", and so on. While on the right column, these 10 hits are regarded as 1 hit. Under this treatment, some diseases with high hits on the left column may show low hit score on the right column. On both lists, there are some already known diseases related to the urea cycle diseases, including "chronic liver disease", "ornithine carbamoyltransferase deficiency", "citrullinemia", etc. We are more interested in those with high hit scores. Common diseases with association degree (hits ≥ 3) are shown in Figure 7.6B.

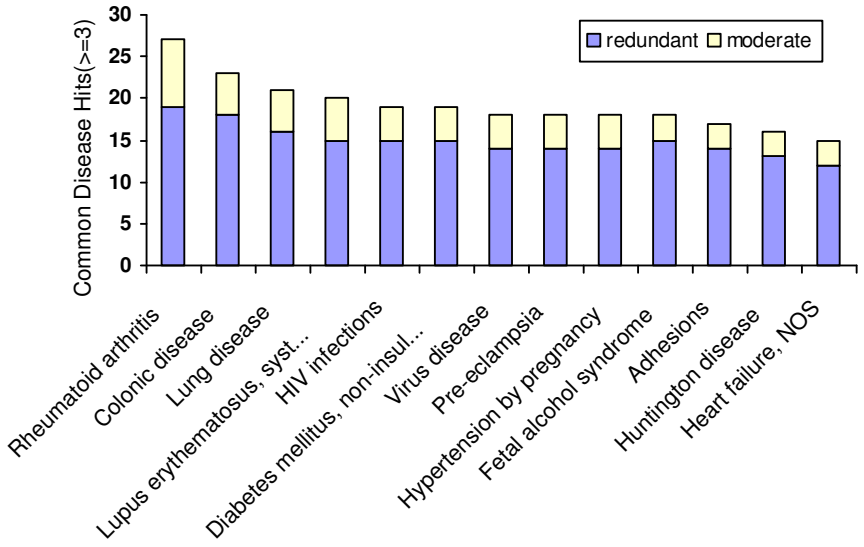


Figure 7.6B Diseases related to the list of signaling pathways.

We surprisingly find that Rheumatoid arthritis is highly related. This is consistent with a recent research by Nissinen R, et al. [Nis03]. They studied whether the enzyme peptidylarginine deiminase (PAD; EC 3.5.3.15), responsible for the post-translational modification of peptide-bound arginine residues to citrulline, constitutes an antigen for patients with rheumatoid arthritis (RA). The study shows that the arginine-citrulline

converting enzyme PAD was recognized as a new antigen against which patients with inflammatory rheumatic diseases frequently show IgG class antibodies. From Figure 7.6B, we can see that systemic lupus erythematosus (SLE) also shows a significant involvement. Both RA and SLE are due to disorders of the musculoskeletal system and connective tissue, which is intensively related to immune systems. It is interesting that three decades ago researches have observed the altered immunoglobulin metabolism between SLE and RA [Lev70]. Later, the prevalence clinical and laboratory associations of SLE and RA were determined by many researches [New93] [Wit00] [Car03]. Other latest observations of the association between RA and urea cycle relevance were achieved by Yonekura Y. et al. [Yon03] and Iwashige K. et al. [Iwa04].

7.7 Summary

We have presented an analysis of urea cycle in a systematical way. Regarding the development of methods and concepts of bioinformatics to analyze metabolic disorders, the integrative aspect stands in the center. By exploiting the existing large amounts of data available in the various databases, we described metabolic mechanisms and pathways, structural genomic organization, patterns of regulatory regions, proteomics, transcriptomics, and metabolomics data of urea cycle.

We also presented a Petri net model to reveal the mechanism of urea cycle disorders. Petri net allows easy incorporation of qualitative insights into a pure mathematical model and adaptive identification and optimization of key parameters to fit system behaviors observed in gene regulated metabolic networks. The study of modeling and simulation plays an important role in detecting genetic/metabolic defects, as well as drug research.

Currently the main urea cycle disorders' management is dietary manipulation by reducing the protein intake. It is possible to increase residual enzyme activity by supplying cofactors. The alternative pathway therapy [Bat82], by intake of chemicals to remove NH3 via other pathways, are practiced, but have limited effectiveness in preventing hyperammonemia, and must be combined with effective dietary management [Bat01]. The future therapy will focus on gene repair, or genetic counseling. This needs more knowledge about cellular functions. The systems analysis approach will also represent the backbone of the concept of disorders management in the post-genomic era. We hope our approach can give a highlight in this direction.

Chapter 8
Conclusions

The rapid development of molecular biology and achievements of modern technology have raised many questions of great bioinformatics interest. Analysis of biopathways is one of the key topics in the post-genomic era. In order to understand the cellular mechanisms, to automatically retrieve metabolic information and predict metabolic pathways, and also to perform comparison of biopathways, we have to develop and implement useful methodologies, algorithms and tools for the analysis of complex biopathways. In this thesis we have investigated several problems of biopathway analysis based on the above considerations.

1) Modeling and simulation of biopathways

The hybrid Petri net has been exploited for modeling and simulation of gene regulated metabolic networks. A global Petri net modeling and simulation strategy and technique is described to systematically investigate metabolic networks. The methodology of this model can be used to all other metabolic networks or the virtual cell metabolism. Moreover we discussed the perspective of Petri nets on modeling and simulation metabolic networks.

A Biology Petri Net Markup Language (BioPNML) for biological data interchange among diverse biological simulators and Petri net tools has been proposed. The BioPNML is designed to provide a starting point for the development of a standard interchange format for Bioinformatics and Petri nets. The language makes it possible to present biology Petri net diagrams between all supported hardware platforms and versions. It is also designed to associate Petri net models and other known metabolic simulators.

2) Prediction of metabolic pathways

A web-based system for prediction of metabolic pathways has been developed. The system, PathAligner, allows to reconstruct metabolic pathways from rudimentary elements such as genes, sequences, enzymes, metabolites, etc., and to extract metabolic information

from biological databases via the Internet. PathAligner also provides a navigation platform to investigate more related metabolic information, and transforms the output data into XML-files for further modeling and simulation. Using the PathAligner system, it is possible to construct a complete Petri net model of biopathway from a rudimentary dataset.

3) Alignment of biopathways

A global definition of bioprocess pathways has been presented. A new method to align metabolic pathway has been described and implemented into the PathAligner system. The algorithm is based on strip scoring the similarity of 4-heirachical EC numbers involved in the pathways.

We have set up the STCDB database. STCDB is an information system on cellular signal transductions. It recommends a classification of cellular signal transduction, and attempts to standardize the representation of signaling pathways. Every characterized signal transduction is assigned a unique 4-heirachical ST number. Our alignment algorithm can be applied to both metabolic pathways and signaling pathways. The general representation of alignment of biopathways is possible by using the recommended signal transduction classification system and the introduced alignment algorithm.

In addition, a concrete biological example has been studied. A detailed model of the urea cycle has been modeled and systematically analyzed. The discoveries of transcription factors and their associated diseases are useful for the treatment of the urea cycle disorders.

The process of “from sequence to structure to function to application” will dominate bioinformatics in the next decades. Biopathways presents many questions and problems worthy to focus on. Some are well studied while others are entirely open problems. We hope that our work has brought us a small step forward in applying computational methods to handle the complexity of metabolic data and that it may some day bring us closer to understand life itself.

Acknowledgments

My supervisor Prof. Dr. Ralf Hofestädt deserves special thanks for supervising me all these years. I have learned a lot from his enthusiastic attitude towards life and work. My sincere gratitude also goes to Prof. Dr. Robert Giegerich, head of our GK-Bioinformatics at the University of Bielefeld, for his guidance and valuable working environment I had the possibility to enjoy.

I have spent very special four years in Germany studying for this degree. There are many friends whom I would like to thank for making my stay in Magdeburg and Bielefeld so valuable. Andreas Freier shared with me not only the same office both at the University of Magdeburg and the University of Bielefeld but also some scientific ideas and joy of daily life. Dr. Jacob Köhler proofread parts of this dissertation and gave very helpful suggestions. Philipp Fahr corrected the English very carefully. I would also express my sincere appreciation to Dr. Ralf Kaurt, Matthias Lange, Susana Lin, Dr. Dieter Lorenz, Mark Niemann, Alex Rüegg, Dr. Uwe Scholz, Andreas Stephanik and other colleagues and friends for supporting my work.

In addition to those people who have been explicitly cited in this thesis, there are others who have helped, encouraged, and made the project more enjoyable than it otherwise would have been, and to whom I therefore wish to record my thanks.

I would like to thank Tanja Möller, Britta Quisbrock and Brigitte Schweer for being so patient and helpful.

Most of all, I want to thank my wife, Yan Li, for all her love and encouragement.

The German Research Foundation (DFG) graduate program "GK-Bioinformatics" at the University of Bielefeld awarded the scholarship that supported me for the past three years. From Feb. 2000 to June 2001, the research was supported by the Ministry of Science and Art of the Government of Sachsen-Anhalt.

Bibliography

[Abb02] Abbot A. (2002) Betting on tomorrow's chips, *Nature* **415**: 112-114.

[Ach01] Achard F., Vaysseix G. and Barillot E. (2001) XML, Bioinformatics and Data Integration, *Bioinformatics*, **17(2)**: 115-125.

[Alb94] Alberts A., Bray D., Lewis J., et al. (1994) *Molecular Biology of the Cell*, Garland, New York, p83.

[All01] Allen H.D. (2001) Reconstruction of metabolic pathways by the exploration of gene expression data with factor analysis, *Dissertation*, Virginia Polytechnic Institute and State University, Blacksburg, VA.

[Ana00] Ananko E.A., Kolpakov F.A. and Kolchanov N.A. (2000) GeneNet database: a technology for a formalized description of gene networks, *Proc. of the Second International Conference on Bioinformatics of Genome Regulation and Structure. BGRS'2000*, August 7-11, Novosibirsk, Russia, p174-177.

[Ari00] Arita M. (2000) Metabolic reconstruction using shortest paths, *Simulat. Pract. Theory*, **8**: 109-125.

[Bad99] Badger J.H. and Olsen G.J. (1999) CRITICA: Coding Region Identification Tool Invoking Comparative Analysis, *Mol. Biol. Evol.*, **16(4)**: 512-524.

[Ban00] Bansal A.K. (2000) A Framework of Automated Reconstruction of Microbrial Metabolic Pathways, in Proceedings of the *IEEE International Symposium on Bio-informatics and Biomedical Engineering*, Arlington, VA, November 8-11, p184-190.

[Bat82] Batshaw M.L., Brusilow S., Waber L., et al. (1982) Treatment of inborn errors of urea synthesis: activation of alternative pathways of waste nitrogen synthesis and excretion, *New Eng. J. Med.*, **306**: 1387-1392.

[Bat01] Batshaw M.L., MacArthur R.B. and Tuchman M. (2001) Alternative pathway therapy for urea cycle disorders: twenty years later, *J Pediatr.*, **138(1 Suppl)**: S46-55.

[Ben03] Benson A.D., Karsch-Mizrachi I., Lipman J.D., et al. (2003) GenBank, *Nucleic Acids Res.*, **31(1)**, 23–27.

[Ber02] Berman M.H., Battistuz T., Bhat N.T., et al. (2002) The Protein Data Bank, *Acta Cryst.*, **D58**: 899–907.

[Bio01] BioCarta (2001) *Biocarta: Charting pathways of life*, http://www.biocarta.com.

[Boe03] Boeckmann B., Bairoch A. and Apweiler R. (2003) The SWISSPROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, **31(1)**: 365–370.

[Bon98] Bono H., Ogata H. and Goto S. (1998) Resconstuction of amino acid biosynthesis pathways from the complete genome sequence, *Genome Research*, **8**: 203–210.

[Bow01] Bower J.M. and Bolouri H. (2001) *Computational modeling of genetic and biochemical networks*, Massachusetts Institute of Technology.

[Boy03] Boyer F. and Viari A. (2003) Ab initio reconstruction of metabolic pathways, *ECCB'2003* (European Conference on Computational Biology), September 27-30th, Paris.

[Bru92] Brutlag D.L., Galper A.R., Millis D.H. (1992) Knowledge-Based Simulation of DNA Metabolism: Prediction of Action and Envisionment of Pathways, In: Hunter L. eds., *Artificial Intelligence & Molecular Biology*, MIT Press, Cambridge, MA.

[Car03] Carlson E. and Rothfield N. (2003) Etanercept-induced lupus-like syndrome in a patient with rheumatoid arthritis, *Arthritis Rheum.*, **48(4)**: 1165-1166.

[Che00] Chen M. (2000) Modelling the glycolysis metabolism using hybrid petri nets, *DFG-Workshop - Modellierung und simulation metabolischer netzwerke*, Mai 19-20, Magdeburg, Germany, p25-26.

[Che02a] Chen M. (2002) Modelling and Simulation of Metabolic Networks: Petri Nets Approach and Perspective, in the proceeding of "*ESM 2002, 16th European Simulation Multiconference*", 2002, June 3-5, Darmstadt, Germany, p441-444.

[Che02b] Chen M., Freier A., Köhler J., et al. (2002) The Biology Petri Net Markup Language. In: Jörg Desel, Mathias Weske (Hrsg.), *Lecture Notes in Informatics - proceedings of Promise'2002*, Oct. 9-11, Potsdam, Germany, Vol. **21**: 150-161.

[Che03] Chen M. and Hofestädt R. (2003) Quantitative Petri net model of gene regulated metabolic networks in the cell. *In Silico Biol.*, **3(3)**: 347-65.

[Che04a] Chen M. and Hofestädt R. (2004) Web-based Information Retrieval System for the Prediction of Metabolic Pathways. *IEEE Transactions on Nano-Bioscience*, accepted.

[Che04b] Chen M. And Hofestädt R. (2004) PathAligner: Metabolic Pathway Retrieval and Alignment. *Applied Bioinformatics*, accepted.

[Che04c] Chen M., Lin S. and Hofestädt R. (2004) STCDB: Signal Transduction Classification Database. *Nucleic Acids Research*, **32(1)**: D456-D458

[Cla96] Clark L. (1996) Information Processing Perspectives Of Cellular Communication, *Web page*, Available via http://www.csc.liv.ac.uk/~laurence/research/report.html.

[Col02] Collado-Vides J., Hofestädt R. (2002) *Gene regulation and Metabolism – Post genomic Computational Approaches*, MIT Press, Cambridge, MA

[Cov01] Covert M., Schilling W. and Famili C.H. (2001) Metabolic Modeling of Microbial Strains in silico, *Trends in Biochemical Sciences*, **27**: 179–186.

[Dan97] Danchin A. (1997) Comparison between the Escherichia coli and Bacillus subtilis genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP, *DNA Res.*, **28**: 9-18.

[Dan99] Dandekar T., Schuster S., Snel B., et al. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes, *Biochem J.*, **1**:115-24.

[Dav92] David R. and Alla H. (1992) *Petri Nets and Grafcet -- Tools for Modeling Discrete Event Systems*, Prentice Hall.

[Dav95] Davidson S.B., Overton C. and buneman P. (1995) Challenges in Integrating Biological Data Sources, *J. Comput. Biol.*, **2**: 557-572.

[DeR99] DeRisi J.L. and Iyer V.R. (1999) Genomics and array technology, *Curr Opin Oncol*, **11(1)**: 76-9.

[Dij59] Dijkstra E.W. (1959) A note on two problems in connexion with graphs, *Numerische Math.*, **1**: 269-271.

[Doi99] Doi A., Drath R., Nagasaki M., et al. (1999) Protein Dynamics Observations of Lambda Phage by Hybrid Petri Net, *Genome Informatics*, **10**: 217-218.

[Dow99] Dow J. (1999) *The dictionary of Cell & Molecular Biology (3rd ed.)*, Academic Press, London.

[Dso99] DSouza M.G., Huan J., Sutton S., et al. (1999) PUMA2: An Environment for Comparative Analysis of Metabolic Subsystems and Automated Reconstruction of Metabolism of Microbial Consortia and Individual Organisms from Sequence Data, *Technical Memorandum ANL/MCS-TM-240*, Mathematics and Computer Science Division, Argonne National Laboratory.

[Ela95] Elaine G., http://mvhs1.mbhs.edu/mvhsproj/CellResp/cell_table.html

[End01] Endy D. and Brent R. (2001) Modelling cellular behavour, *Nature*, **409**: 391-395.

[Enr01] Enright A.J. and Ouzounis C.A. (2001) BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17(9)**: 853-4.

[Etz96] Etzold T., Ulyanow A. and Argos P. (1996) SRS: InformationRetrieval System for Molecular Biology Data Banks, *MethodsEnzymol.*, **266**: 114–128.

[Fel86] Fell D.A. and Small J.R. (1986) Fat synthesis in adipose tissue: an examination of stoichiometric constraints, *Biochem. J.*, **238**: 781-786.

[Fel00a] Fell D.A. and Wagner A. (2000) Structural properties of metabolic networks: implications for evolution and modelling of metabolism, In: Hofmeyr J.H.S., Rohwer J.M. and Snoep J.L. (eds), *Animating the cellular map*, Stellenbosch University Press, Stellenbosch, p79-85.

[Fel00b] Fell D.A. and Thomas S. (2000) Exercising control when control is distributed, In: Cornish-Bowden A. and Cárdenas M. L. (eds), *Technological and medical implications of metabolic control analysis*, Kluwer Academic Publishers, Dordrecht p267-274.

[Fel01] Fell D.A. and Wagner A. (2001) The Small world inside large metabolic networks, In *2nd Workshop on Computation of Biochemical Pathways and Genetic Networks*, Logos Verlag, Berlin, p11-19.

[Flo62] Floyd R.W. (1962) Alogrithm 97: Shortest path, *Comm. ACM*, **5**: 345.

[For99] Forst C.V. and Schulten K. (1999) Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information, *J. Comput. Biol*. **6**: 343-360.

[For01] Forst C.V. and Schulten K. (2001) Phylogenetic Analysis of Metabolic Pathways, *J. Mol. Evol.*, **52**: 471-489.

[Fre02a] Freier A., Hofestädt R. and Lange M. (2002) IIUDB: an Objectiv-oritented System for Modelling, Integration and Analysis of Gene Controlled Metabolic Networks, In *proceeding of the Third International Conference on Bioinfomatics of Genome Regulation and Structure (BGRS'2002),* July 14-20, Novosibirsk, Russia.

[Fre02b] Freier A., Hofestädt R., Lange M., et al. (2002) BioDataServer: A SQL-based service for the online integration of life science data, *In Silico Biology*, **2**: 0005.

[Gaa95] Gaasterland T. and Selkov E. (1995) Reconstruction of Metabolic Networks Using Incomplete Information, In *3rd Int'l Conf. On Intelligent Systems for Molecular Biology*, Cambridge, England, p127–135.

[Gal98] Galperin M.Y. and Brenner S.E. (1998) Using metabolic pathway databases for functional annotation, *Trends Genet.*, **14**: 332-333.

[Gal04] Galperin M.Y. (2004) The Molecular Biology Database Collection: 2004 update, *Nucleic Acids Res.*, **32(1)**: D3–D22.

[Gan00] Gansner R.E. and North C.S. (2000) An open graph visualization system and itsapplications to software engineering, *Discrete Algorithm Engineering*, **30(11)**: 1203–1233.

[Gen87] Genrich H.J. (1987) Predicate/Transition Nets, In: Brauer W., Reisig W., Rozenberg G., eds. *Lecture Notes in Computer Science, Vol. 254*, Springer-Verlag, p207-247.

[Gen01] Genrich H., Kueffner R. and Voss K. (2001) Executable Petri Net Models for the Analysis of Metabolic Pathways, *Intl. J. STTT*, **3(4)**: 394-404.

[Gib00] Gibson M.A. and Bruck J. (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels, *Journal Physical Chemistry*, **104**: 1876-1889.

[Gib01] Gibbs W.W. (2002) Cybernetic Cells, *Scientific American*, 285(2): 53-7.

[Gil77] Gillespie D.T. (1977) Exact stochastic simulation of coupled chemical reactions, *Journal Physical Chemistry*, **81**: 2340-2361.

[Goe02] Goesmann A., Haubrock M. and Meyer F. (2002) PathFinder: Reconstruction and dynamic visualization of metabolic pathways, *Bioinformatics*, **18**: 124–129.

[Goo93] Goodsell D.S. (1993) *The Machinery of Life*, Springer-Verlag, Berlin.

[Gor99] Goryanin I., Hodgman T.C., Selkov E. (1999) Mathematical simulation and analysis of cellular metabolism and regulation, *Bioinformatics*, **15(9)**: 749-58.

[Gos98] Goss P.J.E. and Peccoud J. (1998) Quantitative Modeling of Stochastic Systems in Molecular Biology by Using Stochastic Petri Nets, *Proc.Natl.Acad.Sci. USA*, p6750-6755.

[Gos99] Goss P.J.E. and Peccoud J. (1999) Analysis of the stabilizing effect of Rom on the genetic network controlling ColE1 plasmid replication, In *Pacific Symposium on Biocomputing'99*, p 65-76.

[Got98] Goto S., Nishioka T. and Kanehisa M. (1998) LIGAND: chemical database for enzyme reactions, *Bioinformatics*, **14**: 591-599.

[Gro97] Gronewold A. and Sonnenschein M. (1997) Asynchronous Layered Cellular Automata for the Structured Modeling of Ecological Systems. In: *9th European Simulation Symposium (ESS `97), SCS*, p286-290.

[Gro98] Gronewold A. and Sonnenschein M. (1998) Event-based Modelling of Ecological Systems with Asynchronous Cellular Automata, *Ecological Modeling*, **108**: 37-52.

[Gue00] Guerrini V.H. and Jackson D. (2000) Bioinformatics and extended markup language (XML), *Online Journal of Bioinformatics*, **1**:1-13.

[Haa01] Haas L.M., Schwarz P.M., Kodali P., et al. (2001) DiscoveryLink: A system for integrated access to life sciences data sources, *IBM Systems Journal*, **40**: 489-511.

[Hae96] Haefner J.W. (1996) *Modeling biological systems: principles and applications*, Chapman & Hall, New York.

[Han00] Hanahan D. and Weinberg R.A. (2000) The hallmarks of cancer, *Cell*, **100(1)**: 57-70.

[Har97] Hartwell L. (1997) A robust view of biochemical pathways, *Nature*, **387**: 855-857.

[Hei01] Heiner M., Koch I. and Voss K. (2001) Analysis and simulation of steady states in metabolic pathways with Petri nets, In *CPN'01 – Third Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools*, University of Aarhus, Denmark, p15-34.

[Hei02a] Heinrich R., Meléndez-Hevia E. and Cabezas H. (2002) Optimization of kinetic parameters of enzymes, *Biochem. Mol. Biol. Educ.*, **30**: 184-188.

[Hei02b] Heinrich R., Neel B.G. and Rapoport T.A. (2002) Mathematical models of protein kinase signal transduction, *Molecular Cell*, **9**: 957-970.

[Hey03] Heymans M. and Singh A.K. (2003) Deriving phylogenetic trees from the similarity analysis of metabolic pathways, *Bioinformatics*, **19(Suppl 1)**: I138-I146.

[Hoc96] Hochstrasser M. (1996) Protein degradation or regulation: Ub the judge, *Cell*, **84**: 813-815.

[Hof94] Hofestädt R. (1994) A Petri Net Application of Metabolic Processes, *Journal of System Analysis, Modelling and Simulation*, **16**: 113-122.

[Hof98] Hofestädt R. and Thelen S. (1998) Quantitative Modeling of Biochemical Networks, *In silico Biol.*, **1**: 980006.

[Hof00] Hofestädt R., Lautenbach K. and Lange M. (2000) *Proceeding of "Modellierung und Simulation Metabolischer Netzwerke"*, DFG-Workshop, Mai 2000, Magdeburg, Germany.

[Hof04] Hofestädt R. and Chen M. (2004) Metabolic Pathway Prediction/Alignment, *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004)*, July 25-30, 2004, Novosibirsk, Russia.

[Hor45] Horowitz N.H. (1945) On the evolution of biochemical syntheses, *Proc.Natl.Acad.Sci. USA*, **31**: 153-157.

[Huc01] Hucka M., Finney A., Sauro H., et al. (2001) *Systems Biology Markup Language (SBML)* Level 1, Version 1 (Final).

[Iga98] Takai-Igarashi T., Nadaoka Y. and Kaminuma T. (1998) A database for cell signaling networks, *J Comput Biol.*, **5(4)**:747-54.

[Iwa04] Iwashige K., Kouda K., Kouda M., et al. (2004) Calorie restricted diet and urinary pentosidine in patients with rheumatoid arthritis, *J Physiol Anthropol Appl Human Sci.*, **23(1)**: 19-24.

[Jen76] Jensen R.A. (1976) Enzyme recruitment in evolution of new function, *Annu Rev Microbiol*, **30**: 409-425.

[Jen97] Jensen K. (1997) Coloured Petri Nets - Basic Concepts, Analysis Methods and Practical Use, In: *EATCS Monographs on Theoretical Computer Science*, 2nd edition, Springer-Verlag, Berlin.

[Jue00] Jüngel M., Kindler E. and Weber M. (2000) Towards a Generic Interchange Format for Petri Nets, *Workshop on XML/SGML based Interchange Formats for Petri Nets*, June 2000, Aarhus, Denmark.

[Kan97] Kanehisa M. (1997) A database for post-genome analysis, *Trends Genet.*, **13**: 375-376.

[Kan00] Kanehisa M. and Goto S. (2000) KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, **28(1)**: 27-30.

[Kan02] Kanehisa M., Goto S. and Kawashima S. (2002) The KEGG databases at GenomeNet, *Nucleic Acids Res.*, **30(1)**: 42–46.

[Kar95] Karp P.D. (1995) A Strategy for Database Interoperation, *J. Comput. Biol.*, **2**: 573-586.

[Kar98] Karp P.D. (1998) Metabolic databases, *Trends in Biochemical Sciences*, **23(3)**: 114-116.

[Kar99] Karp P.D., Riley M., Paley S.M., et al. (1999) EcoCyc: Encyclopedia of Escherichia coli Genes and Metabolism, *Nucl. Acids Res.*, **27(1)**: 55.

[Kar02] Karp D.P., Riley M. and Paley S.M. (2002) The MetaCyc Database, *Nucleic Acids Res.*, **30(1)**: 59–61.

[Kel02] Kell D.B. and Medes P. (2002) Snapshots of systems: Metabolic control analysis and biotechnology in the post-genomic era, In: Cornish-Bowden A.J. and Cardenas M.L. (eds), *Technological and Medical Implications of Metabolic Control Analysis*, Kluwer Academic, Dordrecht, p. 3-25.

[Kim01] Kim S.J. and Lee Y.S. (2001) In Silico Metabolic Pathway Modeling and Analysis of Mycoplasma pneumoniae, *Genome Informatics*, **12**: 298–299.

[Koc99] Koch I., Schuster S. and Heiner M. (1999) Simulation and analysis of metabolic networks by time-dependent Petri nets, In *Proceedings of the German Conference on Bioinformatics GCB'99*, Oct 4-6, Hannover, Germany.

[Koh83] Kohn M. and Letzkus W. (1983) A Graph-theoretical Analysis of Metabolic Regulation, *J. Theor. Biol.*, **100(2)**: 293-304.

[Kru03] Krull M., Voss N. and Choi V. (2003) TRANSPATHr: an integrated database on signal transduction and a tool for array, *Nucl. Acids. Res.*, **31(1)**: 97–100.

[Kue00] Kueffner R., Zimmer R. and Lengauer T. (2000) Pathway Analysis in Metabolic Databases via Differential Metabolic Display (DMD), *Bioinformatics*, **16(9)**: 825-836.

[Kum00] Kummer O. and Wienberg F. (2000) The XML File Format of Renew, In *Workshop on XML/SGML based Interchange Formats for Petri Nets*, June 2000, Aarhus, Denmark.

[Kuz96] Kuzmic P. (1996) Program DYNAFIT for the Analysis of Enzyme Kinetic Data: Application to HIV Proteinase, *Anal Biochem.*, **237(2)**: 260-73.

[Lev66] Levenshtein V. (1966) Binary codes capable of correcting deletions, insertions and reversals, *Soviet Physics doklady*, **6**: 707-710.

[Lev70] Levy J., Barnett E.V., MacDonald N.S., et al. (1970) Altered Immunoglobulin Metabolism in Systemic Lupus Erythematosus and Rheumatoid Arthritis, *J Clin Invest.*, **49(4)**: 708–715.

[Lia02] Liao L., Kim S. and Tomb J.F. (2002) Genome comparisons based on profiles of metabolic pathways, In *Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2002)*, Crema, Italy, p469-476.

[Lyn98] Lyngsø R.B. and Mailund T. (1998) Textual Interchange Format for High-Level Petri Nets, In Proceedings of *First Workshop on Practical Use of Coloured Petri Nets and Design/CPN*, Aarhus, Denmark, p47-64.

[Ma03a] Ma H. and Zeng A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*, **19**: 270-277.

[Ma03b] Ma H. and Zeng A.-P. (2003) The connectivity structure, giant strong component and centrality of metabolic networks, *Bioinformatics*, **19**: 1423-1430.

[Mar98] Markram H., Roth A. and Helmchen F. (1998) Competitive calcium binding: implications for dendritic calcium signaling, *J Comput Neurosci*, **5**: 331-348.

[Mat92] Matko D., Rihard Karba R. and Zupanèiè B. (1992) *Simulation and Modelling of Continuous Systems*, Prentice Hall International Ltd.

[Mat00] Matsuno H., Doi A., Nagasaki M., et al. (2000) Hybrid Petri net Representation of Gene Regulatory Network, *Pacific Symposium on Biocomputing*, **5**: 338-349.

[Mat01] Matsuno H., Doi A., Drath R., et al. (2001) Genomic Object Net: Basic Architecture for Representing and Simulating Biopathways, In *RECOMB'2001*, April 20, Montréal, Canada.

[Mat03a] Matsuno H., Murakami R., Yamane R., et al. (2003) Boundary formation by notch signaling in Drosophila multicellular systems: experimental observations and a gene network modeling by Genomic Object Net, *Pacific Symposium on Biocomputing*, p152-163.

[Mat03b] Matys V., Fricke E. and Geffers, R. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucl. Acids. Res.*, **31(1)**: 374–378.

[Mav95] Mavrovouniotis M.L. (1995) Computational methods for complex metabolic systems: Representation of multiple levels of detail, In: Lim H.A. and Cantor C.R. (eds.): *Bioinformatics & Genome Research*, World Scientific, p265-273.

[McG00] McGarvey B.P., Huang H. and Barker W.C. (2000) PIR: a newresource for bioinformatics, *Bioinformatics*, **16(3)**: 290–291.

[McS03] McShan D.C., Rao S. and Shah I. (2003) PathMiner: predicting metabolic pathways by heuristic search, *Bioinformatics*, **19(13)**: 1692-1698.

[Men93] Mendes P. (1993) GEPASI: A software package for modelling the dynamics, steady states and control of biochemical and other systems, *Comput. Applic. Biosci.*, **9**: 563-571.

[Men99] Mendes P. and Kell D. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parametaer estimation, *Bioinfomratics*, **14(10)**: 869-883.

[Men01] Mendes P. (2001) Modeling large scale biological systems from functional genomic data, parameter estimation, In: Kitano H. (eds.), *Foundations of Systems Biology*, MIT Press, Cambridge, MA, p163-186.

[Mic82] Michal G. (1982) *Biochemical Pathways Wall Chart*, Boehringer Mannheim GmbH Biochemica.

[Mic99] Michal G. (1999) *Biochemical Pathways: An atlas of Biochemistry and Molecular Biology*, Wiley.

[Mus96] Mushegian R.A. and Koonin V.E. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes, *Proc. Natl. Acad*, USA 93, p10268–10273.

[Nag03] Nagasaki M., Doi A., Matsuno H., et al. (2003) Genomic Object Net: a platform for modeling and simulating biopathways, *Applied Bioinformatics*, **2(3)**: 181-184.

[Nak99] Nakao M., Bono H. and Kawashima S. (1999) Genome-scalegene expression analysis and pathway reconstruction in KEGG, *Genome Informatics*, **10**: 94-103.

[Nar97] Naraghi M. and Neher E. (1997) Linearized buffered Ca2+ diffusion in microdomains and its implications for calculation of [Ca2+] at the mouth of a calcium channel, *J Neurosci*, **17**: 6961-6973.

[Nee87] Neelamkavil F. (1987) *Computer Simulation and Modelling*, John Wiley, NY.

[New93] Newkirk M.M., Rauch J., Mageed R.A., et al. (1993) Restricted immunoglobulin variable region gene usage by hybridoma rheumatoid factors from patients with systemic lupus erythematosus and rheumatoid arthritis, *Mol Immunol.*, **30(3)**: 255-263.

[Nis03] Nissinen R., Paimela L., Julkunen H., et al. (2003) Peptidylarginine deiminase, the arginine to citrulline converting enzyme, is frequently recognized by sera of patients with rheumatoid arthritis, systemic lupus erythematosus and primary Sjogren syndrome, *Scand J Rheumatol.*, **32(6)**: 337-42.

[Nob02a] Noble D. (2002) The Rise of Computational Biology, *Nature Reviews*, **3**: 460-463.

[Nob02b] Noble D. (2002) Modeling the Heart-from Genes to Cells to the Whole Organ, *Science*, **295**: 1678-1682.

[Oga98] Ogata H., Goto S. and Fujibuchi W. (1998) Computationwith the KEGG pathway database, *BioSystems*, **47**: 119–128.

[Oga96] Ogata H., Bono H. and Kanehisa M. (1996) Analysis of binary relations and hierarchies of enzymes in the metabolic pathways, *Genome Informatics*, **7**: 128-136.

[Oli01] Oliveira J.S., Bailey C.G., Jones-Oliveira J.B., et al. (2001) An Algebraic-Combinatorial Model for the Identification and Mapping of Biochemical Pathways, *Bulletin of Mathematical Biology*, **63(6)**: 1163-1196.

[Ove97] Overbeek R., Larsen N. and Smith W. (1997) Representation of function: the next step, *Gene*, **191**: GC1–9.

[Ove99] Overbeek R., Fonstein M., D'Souza M., et al. (1999) The use of gene clusters to infer functional coupling, *Proc Natl Acad Sci USA*, **96(6)**: 2896-2901.

[Ove00] Overbeek R., Larsen N. and Pusch G.D. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction, *Nucleic Acids Res.*, **28(1)**: 123–125.

[Pal02] Paley S. and Karp P.D. (2002) Evaluation of computational metabolic pathway predictions for *Helicobacter pylori*, *Bioinformatics*, 18(5): 715-724.

[Pet62] Petri C.A. (1962) Kommunikation mit Automaten, *Dissertation*, Institut für Instrumentelle Mathematik, Schriften des IIM Nr. 2, Bonn.

[Pfe99] Pfeiffer T., Sanchez-Valdenebro I., Nuno J., et al. (1999) METATOOL: For studying metabolic networks, *Bioinformatics*, **15**: 168-176.

[Pie80] Pierson D.L., Brien J.M. (1980) Human carbamylphosphate synthetase I. Stabilization, purification, and partial characterization of the enzyme from human liver, *J. Biol. Chem.*, **255**: 7891-7895.

[Pyr96] Pyronnet S., Vagner S., Bouisson M., et al. (1996) Relief of ornithine decarboxylase messenger RNA translational repression induced by alternative splicing of its 5' untranslated region, *Cancer Res.*, **56**: 1742-1745.

[Red93] Reddy V.N., Mavrovouniotis M.L. and Liebman M.N. (1993) Petri Net Representation in Metabolic Pathways, In *Proceedings First International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, p328-336.

[Red96] Reddy V.N., Liebman M.N. and Mavrovouniotis M.L. (1996) Qualitative analysis of biochemical reaction systems, *Comput Biol. Med.*, **26(1)**: 9-24.

[Rei82] Reisig W. (1982) *A Primer in Petri Net Design*, Springer, Berlin.

[Res01] Resat H., Wiley H.S. and Dixon D.A. (2001) Probability Weighted Dynamic Monte Carlo Method for Reaction Kinetics Simulations, *Journal of Physical Chemistry*, **105(44)**: 11026-11034.

[Rud97] Rudin N. (1997) *Dictionary of Modern Biology*, Barrons Educational Series, http://www.forensicdna.com/Bookstore/%20bookstore_images/cell.gif.

[Rut97] Ruth M. and Hannon B. (1997) *Modeling dynamic biological systems*, New York, Springer.

[Sch96] Schuster S., Hilgetag C., Woods J.H., et al. (1996) Elementary modes of functioning in biochemical reaction networks, In: Cuthbertson, R., et al. (eds.), *Computation in Cellular and Molecular Biological Systems*, World Scientific, Singapore, p151-165.

[Sch99] Schuster S., Dandekar T. and Fell D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, *Trends. Biotechnol.*, **17**: 53-60.

[Sch00a] Schuster S., Fell D. and Dandekar T. (2000) A General Definition of Metabolic Pathways Useful for Systematic Organization and Analysis of Complex Metabolic Networks, *Nature Biotechnol*, **18**: 326-332.

[Sch00b] Schilling C.H., Letscher D. and Palsson B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective, *Journal of Theoretical Biology*, **203**: 229-248.

[Sch00c] Schaff J.C., Slepchenko B.M. and Loew L.M. (2000) Physiological modeling with the virtual cell framework, In: Johnson M. and Brand L., (eds.), *Methods in Enzymology*, Academic Press, San Diego, p1-23.

[Sch01a] Schaff J.C., Slepchenko B.M., Choi Y.S., et al. (2001) Analysis of nonlinear dynamics on arbitrary geometries with the Virtual Cell, *Chaos*, **11(1)**:115-131.

[Sch01b] Schacherer F. (2001) An object-oriented database for the compilation of signal transduction pathways, *PhD thesis*, Technical Universitaet Braunschweig, Braunschweig.

[Sch02a] Schuster S., Pfeiffer T., Moldenhauer F., et al. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae, *Bioinformatics*, **18(2)**: 351-61.

[Sch02b] Schomburg I., Chang A. and Schomburg D. (2002) BRENDA, enzyme data and metabolic information, *Nucleic Acids Res.*, **30(1)**: 47–49.

[Scr97] Scriver C.R., Sly W.S., Childs B., et al. (1997) *The metabolic and Molecular Bases of Inherited Disease*, McGraw-Hill Companies, Inc.

[Sel97] Selkov E., Maltsev N. and Olsen G. (1997) A reconstruction of the metabolism of Methanococcus jannaschii from sequence data, *Gene*, **197**: GC11–26.

[Sel98] Selkov E.Jr., Grechkin Y., Mikhailova N., et al. (1998) MPW: the Metabolic Pathways Database, *Nucleic Acids Research*, **26**: 43-45.

[Sel80] Sellers P. (1980) The theory and computation of evolutionary distances: Pattern recognition, *J. of Algorithms*, **1(4)**: 359-373.

[Sie01] Siepel A., Farmer A., Tolopko A., et al. (2001) ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources, *Bioinformatics*, **17**: 83-94

[Sir96] Siregar P., Sinteff J.P., Chahine M., et al. (1996) A cellular automata model of the heart and its coupling with a qualitative model, *Comput Biomed Res.*, **29(3)**:222-46.

[Sop03] Sophia T., David S. and Christos A.O. (2003) Automated metabolic reconstruction for *Methanococcus jannaschii*, *Arachaea*, **1**, Heron Publishing, Canada.

[Sri01] Srivastava R., Peterson M.S. and Bentley W.E. (2001) Stochastic kineticanalysis of the Escherichia coli stress circuit using σ32-targeted antisense, *Biotechnology Bioengineering*, **75(1)**: 120-129.

[Ste00] Stevens R., baker P., Bechhofer S. et al. (2000) TAMBIS: Transparent Access to Multible Bioinformatics Information Sources, *Bioinformatics*, **16**: 184-185.

[Sun02] Sun J. and Zeng A.-P. (2002) In silico reconstruction of metabolic network fromunannotated raw genome sequences, In *ISMB'02 poster*, Edmonton, Canada.

[Svá00] Svádová M. and Hanzalek Z. (2000) Matlab Toolbox for Petri nets, In *INOVACE 2000*, Praha: Asociace inovačního podnikání ČR, p15-20.

[Tat95] Tateishi N., Shiotari H., Kuhara S., et al. (1995) An integrated database SPAD (Signaling PAthway Database) for signal transduction and genetic information, In *Genome Informatics Workshop, GIW95*, Dec. 11th-12th, Yokohama, Japan.

[Tat99] Tatusova T.A., Karsch-mizrachi L. and Ostell J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis, *Bioinformatics*, **15**: 536-543.

[Tho96] Thomas S. and Fell D.A. (1996) Design of metabolic control for large flux changes, *J. Theor. Biol.*, **182**: 285-298.

[Tho97a] Thomas S., Mooney P.J.F., Burrell M.M., et al. (1997) Finite change analysis of lines of transgenic potato (*Solanum tuberosum*) over expressing phosphofructokinase, *Biochem. J.*, **322**: 111-117.

[Tho97b] Thomas S., Mooney P.J.F., Burrell M.M., et al. (1997) Metabolic control analysis of glycolysis in tuber tissue of potato (*Solanum tuberosum*): explanation for the low control coefficient of phosphofructokinase over respiratory flux, *Biochem. J.*, **322**: 119-127.

[Tho98] Thomas S. and Fell D.A. (1998) The role of multiple enzyme activation in metabolic flux control, *Adv. Enzyme Reg.*, **38**: 65-85.

[Toh00a] Tohsato Y., Matsuda H. and Hashimot A. (2000) A Multiple Alignment Algorithm for Metabolic Pathway Analysis using Enzyme Hierarchy, In *ISMB'2000*, Reception, San Diego, p376-383.

[Toh00b] Tohsato Y., Matsuda H. and Hashimoto A. (2000) An Application of a Pathway Alignment Method to the Analysis of Amino Acid Biosynthesis, *Genome Informatics*, **11**: 284-285.

[Tom99] Tomita M., Hashimoto K., Takahashi K., et al. (1999) E-CELL: Software environment for whole cell simulation, *Bioinformatics*, **15(1)**: 72-84.

[Tom01] Tomita, M. (2001) Whole cell simulation: A grand challenge of the 21st century, *Trends in Biotechnology*, **19(6)**: 205-210.

[van00] van Helden J., Naim A. and Mancuso R. (2000) Representing and analysing molecular and cellular function using the computer. *Biol. Chem.*, **381(9-10)**: 921–935.

[Vel95] Velculescu V. E., Zhang L., Vogelstein B., et al. (1995) Serial analysis of gene expression, *Science*, **270**: 484-487.

[Vin88] Vincens P. and Tarroux P. (1988) Two-dimensional electrophoresis computerized processing, *Int J Biochem*, **20(5)**: 499-509.

[Voe95] Voet D. and Voet J. (1995) *Biochemistry*, 2nd ed., J.Wiley&Sons.

[Voi00a] Voit E.O. (2000) Tomas Radivoyevitch: Biochemical systems analysis of genome-wide expression data, *Bioinformatics*, **16(11)**: 1023-1037.

[Voi00b] Voit E.O. (2000) *Computational analysis of biochemical systems*, Cambridge University Press, Cambridge, UK.

[Vos00] Voss K. (2000) Ausführbare Petrinetz Modelle zur Simulation Metabolischer Pfade, *DFG-Workshop, Modellierung und Simulation Metabolischer Netzwerke*, 2000, Mai 19-20, Magdeburg, Germany, p11-13.

[Wag74] Wagner R. and Fisher M. (1974) The string-to-string correction problem, *J. ACM* **21(1)**: 168-173.

[Wag01] Wagner A. and Fell D. (2001) The small world inside large metabolic networks, In *Proc. Roy. Soc. London Ser. B*, **280**: 1803-1810.

[Wan98] Wang J. (1998) *Timed Petri Nets: Theory and Application*. Kluwer Academic Publishers, Boston.

[Web92] Webb E.C. (1992) *Enzyme nomenclature 1992: Recommendations of the Nomenclature Committee of the International union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*, Academic Press, New York, NJ.

[Web03] Weber M. and Kindler E. (2003) The Petri Net Markup Language, In: *Lecture Notes in Computer Science*, 2472: 124-144.

[Wei93] Weissmuller G. and Bisch P.M. (1993) Autocatalytic cooperativity and self-regulation of ATPase pumps in membrane active transport, *Eur. Biophys. J.*, **22**: 63-70.

[Wit00] Witte T., Hartung K., Sachse C., et al. (2000) Rheumatoid factors in systemic lupus erythematosus: association with clinical and laboratory parameters, *Rheumatol Int.*, **19(3)**: 107-111.

[Yao96] Yao K.S., Godwin A.K., Johnson C., et al. (1996) Alternative splicing and differential expression of DT-diaphorase transcripts in human colon tumors and in peripheral mononuclear cells in response to mitomycin C treatment, *Cancer Res.*, **56**: 1731-1736.

[Ye96] Ye X., Robinson M.B., Batshaw M.L., et al. (1996) Prolonged Metabolic Correction in Adult Ornithine Transcarbamylase-deficient Mice with Adenoviral Vectors, *Journal of Biological Chemistry*, **271(7)**: 3639-3646.

[Ye97] Ye X., Robinson M.B., Pabin C., et al. (1997) Adenovirus-mediated in vivo gene transfer rapidly protects ornithine transcarbamylase-deficient mice from an ammonium challenge, *Pediatric Research*, **41**: 527-534.

[Ye00] Ye X., Robinson M.B., Pabin C., et. al. (2000) Transient depletion of CD4 lymphocyte improves efficacy of repeated administration of recombinant adenovirus in the ornithine transcarbamylase deficient sparse fur mouse, *Gene Therapy*, **7(20)**: 1761-1767.

[Yon03] Yonekura Y., Koshiishi I., Yamada K., et al. (2003) Association between the expression of inducible nitric oxide synthase by chondrocytes and its nitric oxide-generating activity in adjuvant arthritis in rats, *Nitric Oxide*, **8(3)**: 164-169.

Main Internet-References (all valid at the time of writing, March 2004):

BioCyc: http://biocyc.org/

Biological Simulators: http://www.techfak.uni-bielefeld.de/~mchen/BioSim/BioSim.xml

BioML: http://www.bioml.com/BIOML/

BioPerl: http://www.bioperl.org

BioSpice: http://biospice.lbl.gov/

BLAST: http://www.ncbi.nlm.nih.gov/BLAST/

Boehringer Mannheim: http://us.expasy.org/tools/pathways

BRENDA: http://www.brenda.uni-koeln.de/

BSML: http://www.bsml.org/

CellML: http://www.cellml.org/

CSNDB: http://geo.nihs.go.jp/csndb/

Dbsolve: http://homepage.ntlworld.com/igor.goryanin/

DDBJ: http://www.ddbj.nig.ac.jp/

DynaFit: http://www.biokin.com/dynafit/

E-Cell: http://www.e-cell.org/

EcoCyc: http://ecocyc.org/

EMBL: http://www1.embl-heidelberg.de/

Entrez: http://www.ncbi.nlm.nih.gov/Entrez/

ExPASy: http://www.expasy.ch/

FASTA: http://www.ebi.ac.uk/fasta3/

GeneCard: http://bioinfo.weizmann.ac.il/cards/

GeneNet: http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/

Gene Ontology: http://www.geneontology.org/

Gepasi: http://www.gepasi.org/

GraphViz: http://www.research.att.com/sw/tools/graphviz/

IUBMB: http://www.chem.qmul.ac.uk/iubmb/

Jarnac: http://members.lycos.co.uk/sauro/biotech.htm

KEGG: http://www.genome.ad.jp/kegg/

Klotho: http://www.biocheminfo.org/klotho/

MARG: http://cweb.uni-bielefeld.de/agbi/home/index.html?id=104

MetaCyc: http://metacyc.org/

MutDB: http://mutdb.org/

NCBI: http://www.ncbi.nlm.nih.gov/

GenBank: http://www.ncbi.nlm.nih.gov/Genbank/

OMIM: http://www.ncbi.nlm.nih.gov/omim/

PathAligner: http://bibiserv.techfak.uni-bielefeld.de/pathaligner

PDB: http://www.rcsb.org/pdb/

Petri Nets Tools: http://www.daimi.au.dk/PetriNets/tools/quick.html

Petri Nets World: http://www.daimi.au.dk/PetriNets

PIR: http://pir.georgetown.edu

PNML : http://www.informatik.hu-berlin.de/top/pnml/

RDF: http://www.w3.org/RDF/

SBML: http://www.cds.caltech.edu/erato/

SRS: http://srs.ebi.ac.uk/

STCDB: http://www.techfak.uni-bielefeld.de/~mchen/STCDB

StochSim: http://www.zoo.cam.ac.uk/comp-cell/StochSim.html

Swiss-Prot: http://us.expasy.org/sprot/

TRANSFAC & TRANSPATH: http://www.biobase.de/

UCSC: http://genome.ucsc.edu/

Virtual Cell: http://www.nrcam.uchc.edu/

VON++: http://www.systemtechnik.tu-ilmenau.de/~drath/visual.htm

XML: http://www.w3.org/XML/

XSLT: http://www.w3.org/TR/xslt/

Appendix A.

Predefined functions of biochemical reaction kinetic types (irreversible).

Michaelis-Menten	$v = \frac{V\max \cdot S}{Km + S}$
Hill Kinetics	$v = \frac{V \cdot S^h}{S_{0.5}^h + S^h}$
Substrate Inhibition Kinetics	$v = \frac{V \cdot S / Km}{1 + S / Km + S^2 / Ki}$
Substrate Activation	$v = \frac{V \cdot (S / Ksa)^2}{1 + S / Ksc + (S / Ksa)^2 + S / Ksa}$
Competitive Inhibition	$v = \frac{V \cdot S / Km}{1 + S / Km + I / Ki}$
Noncompetitive Inhibition	$v = \frac{V \cdot S / Km}{1 + I / Ki + S / Km \cdot (1 + I / Ki)}$
Uncompetitive Inhibition	$v = \frac{V \cdot S / Km}{1 + S / Km \cdot (1 + I / Ki)}$
Allosteric inhibition	$v = \frac{V \cdot (1 + S / Ks)^{n-1}}{L \cdot (1 + I / Ki)^n + (1 + S / Ks)^n}$
Ordered BiBi kinetics	$v = \frac{Vf \cdot (AB - PQ / Keq)}{AB \cdot (1 + P / Kip) + Kmb \cdot (A + Kia) + KmaB + K1}$ where $K1 = \frac{Vf / VrKeq}{KmqP \cdot (1 + A / Kia) + QK2}$ $K2 = Kmp(1 + KmaB / KiaKmb + P(1 + B / Kib))$
Ping Pong BiBi kinetics	$v = \frac{Vf \cdot (AB - PQ / Keq)}{AB + KmbA + KmaB \cdot (1 + Q / Kiq) + K1}$ where $K1 = \frac{Vf / VrKeq}{KmqP \cdot (1 + A / Kia) + Q \cdot (Kmp + P)}$

S_i: substrate

P_i: Product

V_{max}: forward maximm velocity

K_m: forward Miachaelis-Menten Constant

K_i: Inhibition constant for the substrate

Appendix B.

A XSLT source code for Petri net XML transformation.

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:fo="http://www.w3.org/1999/XSL/Format">
    <xsl:output method="xml" doctype-system="http://apogonidae.techfak.uni-
bielefeld.de/BioPNML/xrn1.dtd"/>
    <xsl:template match="*|/">
        <xsl:apply-templates/>
    </xsl:template>
    <xsl:template match="text()|@*">
        <xsl:value-of select="."/>
    </xsl:template>
    <xsl:template match="/">
        <xsl:comment>xml for metabolic reaction petri nets - mchen@techfak.uni-
bielefeld.de</xsl:comment>
        <net id="N" type="hlnet">
            <xsl:variable name="enur">
                <xsl:value-of select="0"/>
            </xsl:variable>
            <xsl:variable name="snur">
                <xsl:value-of select="0"/>
            </xsl:variable>
            <xsl:variable name="pnur">
                <xsl:value-of select="0"/>
            </xsl:variable>
            <xsl:for-each select="//record0">
                <xsl:for-each select="EC">
                    <xsl:if test="not(.=preceding::*)">
                        <xsl:variable name="enur">
                            <xsl:value-of select="($enur)+1"/>
                        </xsl:variable>
                        <EC_order>
                            <xsl:value-of select="position()"/>
                        </EC_order>
                        <place>
                            <xsl:attribute name="id"><xsl:value-of select="."/></xsl:attribute>
                            <graphics>
                                <size>
                                    <xsl:attribute name="w"><xsl:value-of
select="$round"/></xsl:attribute>
                                    <xsl:attribute name="h"><xsl:value-of
select="$round"/></xsl:attribute>
                                </size>
                                <offset>
                                    <xsl:attribute name="x"><xsl:value-of select="ceiling(($enur)div
10)*300-100"/></xsl:attribute>
                                    <xsl:attribute name="y"><xsl:value-of select="80*(($enur)-
floor(($enur)div 10)*10)+40"/></xsl:attribute>
                                </offset>
                                <xsl:call-template name="ECgraph"/>
                            </graphics>
                            <annotation>
                                <xsl:attribute name="id">EC<xsl:value-of
select="."/></xsl:attribute>
                                <xsl:attribute name="type">name</xsl:attribute>
                                <text>
                                    <xsl:value-of select="."/>
                                </text>
```

```
                        <xsl:call-template name="Namegraph"/>
                    </annotation>
                    <annotation>
                        <xsl:attribute name="id">initialmarking<xsl:value-of
select="."/></xsl:attribute>
                        <xsl:attribute name="type">initialmarking</xsl:attribute>
                        <text>
                            <xsl:value-of select="1"/>
                        </text>
                        <xsl:call-template name="Textgraph"/>
                    </annotation>
                </place>
                <transition>
                    <xsl:attribute name="id">T<xsl:value-of select="."/></xsl:attribute>
                    <graphics>
                        <size>
                            <xsl:attribute name="w"><xsl:value-of
select="$Twidth"/></xsl:attribute>
                            <xsl:attribute name="h"><xsl:value-of
select="$Theight"/></xsl:attribute>
                        </size>
                        <offset>
                            <xsl:attribute name="x"><xsl:value-of select="ceiling(($enur) div
10)*300-100"/></xsl:attribute>
                            <xsl:attribute name="y"><xsl:value-of select="80*(($enur)-
floor(($enur) div 10)*10)+80"/></xsl:attribute>
                        </offset>
                        <xsl:call-template name="Tgraph"/>
                    </graphics>
                    <annotation>
                        <xsl:attribute name="id">TT<xsl:value-of
select="."/></xsl:attribute>
                        <xsl:attribute name="type">expression</xsl:attribute>
                        <text>T<xsl:value-of select="."/>
                        </text>
                        <xsl:call-template name="Namegraph"/>
                    </annotation>
                </transition>
            </xsl:if>
        </xsl:for-each>
        <xsl:for-each select="SUBSTRATE">
            <SUB_nm>
                <xsl:value-of select="."/>
            </SUB_nm>
            <xsl:if test="not(.=preceding::*)">
                <xsl:variable name="snur">
                    <xsl:value-of select="($snur)+1"/>
                </xsl:variable>
    ……

</xsl:stylesheet>
```

The complete source code is available at:
http://www.techfak.uni-bielefeld.de/~mchen/BioPNML/XML2PN.html

Appendix C.

Proof of Lemma 5.3 2.) The key to the last equality is to consider the binomial expansion $(x+y)^n$. We have

$(x+y)^{m+n} \Rightarrow$ the r+1th term, $T_{r+1} = \binom{m+n}{r} x^{m+n-r} y^r$

$\Rightarrow$ the n+1th term, $T_{n+1} = \binom{m+n}{n} x^m y^n$, i.e. the binomial coefficient of $x^m y^n$ is $\binom{m+n}{n}$.

$(x+y)^m \Rightarrow$ the r+1th term, $T_{r+1} = \binom{m}{r} x^{m-r} y^r$,

$(x+y)^n \Rightarrow$ the r+1th term, $T_{r+1} = \binom{n}{n-r} x^{n-(n-r)} y^{n-r} = \binom{n}{r} x^r y^{n-r}$

For $0 \le r \le min(m,n)$, then the term of $x^m y^n$ in $(x+y)^m (x+y)^n$ is $\sum_{r\ge 0}^{\min(m,n)} \binom{m}{r}\binom{n}{r} x^m y^n$.

Since $(x+y)^{m+n} = (x+y)^m (x+y)^n$, we compare their binomial coefficients of $x^m y^n$, and obtain

$$\sum_{r\ge 0}^{\min(m,n)} \binom{m}{r}\binom{n}{r} = \binom{m+n}{n}.$$

Appendix D.

Proof of Lemma 5.4 With no loss of generality assume that the r matched point $(i, j),(i', j'),\ldots, (k', l'),(k, l)$ (Figure 5.2.2) then $Indexaligns(\delta^{r}(E_1, E_2)) = \binom{i+j}{j}\binom{i'-i+j'-j}{j'-j}\cdots\binom{k-k'+l-l'}{l-l'}\binom{m-k+n-l}{n-l}$.

Since $(x+y)^{m+n} = (x+y)^{i+j}(x+y)^{i'+j'-i-j}\cdots(x+y)^{k+l-k'-l'}(x+y)^{m+n-k-l}$, and

$(x+y)^{n} = \sum_{k=0}^{n}\binom{n}{k}x^{k}y^{n-k}$, we have

$$\sum_{r=0}^{n}\binom{m+n}{r}x^{r}y^{m+n-r} = \sum_{r=0}^{n}\binom{i+j}{r}x^{r}y^{i+j-r}\sum_{r=0}^{n}\binom{i'+j'-i-j}{r}x^{r}y^{i'+j'-i-j-r}\cdots\sum_{r=0}^{n}\binom{k+l-k'-l'}{r}x^{r}y^{k+l-k'-l'-r}\sum_{r=0}^{n}\binom{m+n-k-l}{r}x^{r}y^{m+n-k-l-r}$$

$$\geq \binom{i+j}{i}x^{i}y^{j}\binom{i+j-i'-j'}{i'-i}x^{i'-i}y^{j'-j}\cdots\binom{k+l-k'-l'}{k-k'}x^{k-k'}y^{l-l'}\binom{m+n-k-l}{m-k}x^{m-k}y^{n-l}$$

$$= \binom{i+j}{i}\binom{i+j-i'-j'}{i'-i}\cdots\binom{k+l-k'-l'}{k-k'}\binom{m+n-k-l}{m-k}x^{m}y^{n}$$

Compare the coefficient of $x^m y^n$, obtain

$$\binom{m+n}{m} \geq \binom{i+j}{i}\binom{i+j-i'-j'}{i'-i}\cdots\binom{k+l-k'-l'}{k-k'}\binom{m+n-k-l}{m-k}$$

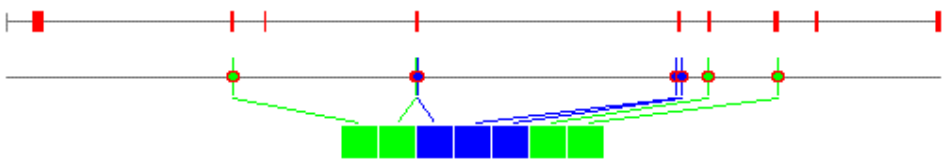
The least number is the case there is no more subpath to align, so that

$$Indexaligns(\delta^{r}(E_1, E_2)) = \binom{0}{0}\binom{0}{0}\cdots\binom{0}{0}\binom{0}{0} = 1.$$

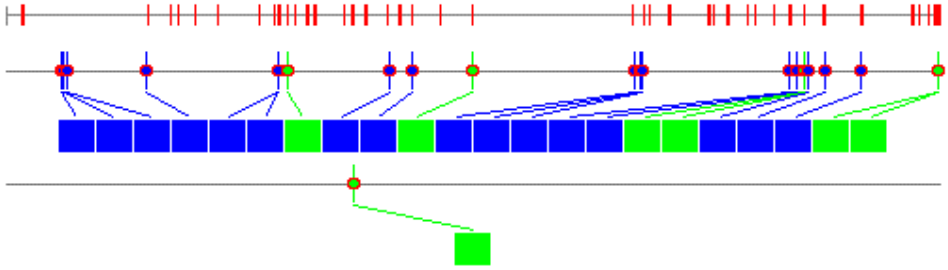
Appendix E.

Computationally annotated mutations of genes in the uear cycle.

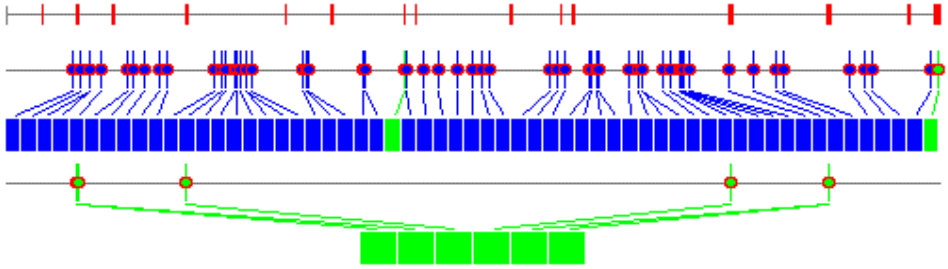
OTC



CPS1

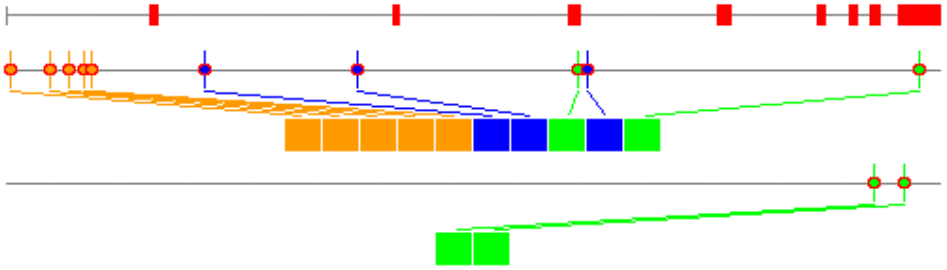


ASS

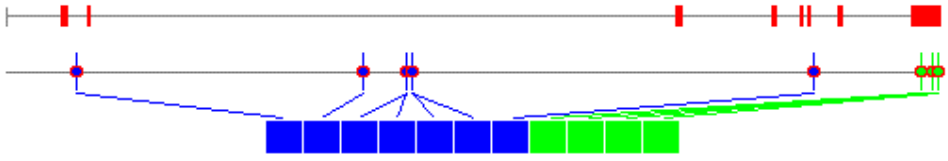


ASL

ARG1



ARG2



Legend: Promoter Exon Intron Unknown

Appendix F.

Computer research for transcription factor binding sites of the human ASL, ASS, CAD and OTC genes.

ASL promoter/upstream 1kb

>hg16_knownGene_BC008195 range=chr7:64951430-64952429 5'pad=0 3'pad=0
revComp=FALSE strand=+ repeatMasking=none
gggttcaagagattctcctgcctcagcctccggagtcgctgggattgcag
gcacccgccaccgtgcccggctaatttttgtattcttagtagagacgggg
tttcaccaccttggccaggctggtctcaaactcctgacctcgtgatccac
ccgcctcggccttccaaagtgctgggattacaggcgtgagccaccgcgcc
cggcctccggccgcgtttctttttcttttttagaggagtcaggctggagtg
ccgtggcacaatagctcactgcagcttcgaactcctgggctcaagtgctc
ctcccgcctcagcctcctgagtagctaggactgcaggcgtgcaccaccac
gcccggctttttttttattattattaatttttttgtagagacgggatctt
gttgtattgcccaggctggtctccaactcgtagcctcaagcgatcctccc
acctcggcctcccaaagtgctgggattacagctgtgagccaccgcgcctg
gcacaggcttcatttctgatggtccttccttttttccttgatgcctttctc
tgtacctggcacatagaggtgcctggtacgtgtttgttgaatgaatgaat
gaatgagtgaatgagcgaacatgccatttcaccttatatatcttgtgaac
ctgccaggcccgggcctgatgtcatagcctctacccctggcccgagtctc
cagtcccctgcgtgtctgctgaccacagcacgaacgccagcgcactaccc
tcctcaaccccagcccaggccccttccccgtcggggtcccccccaaccctt
tccccgcccgcttccccgccccggggccgcttagcctccagctcagcggg
aggtatccccgcccacggccaggattggaggatggaggcaacgcccaccc
cgccgggcggcctcctattggcgcggccgtcgccaggggtggggacagga
ccggcggctgctgacgccatcccggccagaaaagccctggccagtggcgg

Pax-4a
Nkx2-5
AML1

ASS promoter/upstream 1kb

>hg16_knownGene_BC009243 range=chr9:128595361-128596360 5'pad=0 3'pad=0
revComp=FALSE strand=+ repeatMasking=none
taaaaggagcactgactccagggtagcggctgggcaaggcttggggctcc
cacctcccccaggttcagagccggctgaggaccgggagtcctcccttctg
gggtcagtgctcactatggagcaactgccttggatggggtcccaggactt
gctcttttactggggctgtggttgcagtgatcagggctttggagccggca
gaccaagctgggaaactcctgaggtagagaagctgttgaaggcgggcctg
gggtcacaactcccagctgctttttacaagcaagagacttctctctgaac
ctcaaccttccctcctgtctagtgggttcgcagccagacagtcttttact
cactgcttactgggtgccctctggagtctgggcaggtgccaggctctgag
aagacaggccagcaatcagccctggcctaagggatgaaagccgggccttc
ccgcgccggctcacctcggttttctcatccttactcggctaccagaggct
atggttggggagggaggggctctgggggctgcagaaggcccaggctgcc
ggcaccggatagaagtgagcacgaagctccctgcgccagtggaacttta
tcccggctcccaccgcgaagcgtttaaattgcttccccagggccaggagg
caagtctctcgaaggacggctgcggccaccctccgcccctgagttacatg
ggtcgcagccactgccgccctccttggcgcctccagcccgcgggccaggg
ccaggaaccgcgagccgcctgcgccccgccggcgcgccccctgggagggt
gagccggcgccgggcccaggcccggacctggtgggaggcgggggggaggtg
gggacgaggcctggggaggcgggccccgcccatctgcaggtggctgtgaa
cgctgagcggctccaggcgggggccgggcccgggggcggggtctgtggcg
cgcgtcccccccacgtgtcccggtcaccggccctgcccccgggccctgt

AML1
NF-kappaB
GATA-4
NF-1
RFX1
MAZ
E2F-1
USF

CAD promoter/upstream 1kb

>hg16_knownGene_D78586 range=chr2:27413929-27414928 5'pad=0 3'pad=0
revComp=FALSE strand=+ repeatMasking=none
tgtgtcctgagaatggatcttgtgtacctgatggtccaggtcttttgcag
tgtgtttgtgctgatggtttccatggatacaagtgtatgcgccaggtgag
gaattaggccgtctaactagggatacaaggaatgcatagagcaagtcttc
tcagaaaggagagccacaagaccaggagctgatacaaatttcctataggt
ggaaaactatagaattgccctagacaaagtgataggtatattaggaaaga
actaggtgttaggatgtggccctccgtgtaacgttgtatgggggtgtttt
ttggtctgttgttcacagggctcgttctcactgcttatgttcttcgggat
tctgggagccaccactctatccgtctccattctgctttgggcgacccagc
gccgaaaagccaagacttcatgaactacataggtcttaccattgacctaa
gatcaatctgaactatcttagcccagtcagggagctctgcttcctagaaa
ggcatctttcgccagtggattcgcctcaaggttgaggccgccattggaag
atgaaaaattgcactcccttggtgtagacaaataccagttcccattggtg
ttgttgcctataataaacacttttttctttttttttcctctctttctttt
taaggaaaggcgccctgccttacgtgtttcttgcttggggtgggaggga
ctgctttagggcagccggtttttcaggtttccccggttttgcagtgcg
gagaccacgggggcccactcccgtggctccgcggaccccgcccccttacg
tgcccggccccgccccctcacgccgcctgtgtccgcgccgccgcagtctct
gctgctgccgccaagcgcgcccgaggctcctacgctgccgcgcccggctt
ctctccagcgccccgcgccgttagccacgtggaccgactccggcgcgccg
tcctcacgtggttccagtggagtttgcagtccttcccgcttctccgtact

Xvent-1
SREBP-1

OTC promoter/upstream 1kb

>hg16_knownGene_K02100 range=chrX:37241716-37242715 5'pad=0 3'pad=0
revComp=FALSE strand=+ repeatMasking=none
ctcggtatctgatacagaattgactttgaatcacctgatttctaactgag
gataaatgaataaatgtgaagttgcagatggccccttagtgatctgaata
ggctgctaggggaagagcatatggtatccccacttcccacttgtactgac
tgtcaggtgctgttagaatcaataggcaactatttctttttcttttttctttt
ctttctttctttttttgagacagtgtctctctctgtcacccaggctgga
gtacagtggtgcaatctgggctcactgcaacctctgtctcccgggttcaa
gcgactctcatgcctcagcctcccaaatagctgggattacaggtgtgcac
caccacgtctagctaatttttgtatttttagtggagacggggattcacca
tgttggccaggctggtctcgaactcctgggctcaagtgatccgcccgcct
cagcctcccaaagtgctgggattacaggcgtgagccaccgtgcccggcca
gcaattatttctttattgaagacttatgtgcaaggcacaaagggagctcc
aggactgagatatttttactatacctctctatcatcttgcaccccaaa
atagcttccagggcacttctttctatttgtttttgtggaaagactggcaa
ttagaggtagaaaagtgaaataaatggaaatagtactactcaggactgtc
acatctacatctgtgtttttgcagtgccaatttgcattttctgagtgagt
tacttctactcaccttcacagcagccggtaccgcagtgccttgcatatat
tatatcctcaatgagtacttgtcaattgattttgtacatgcgtgtgacag
tataaatatattatgaaaaatgaggaggccaggcaataaaagagtcagga
tttcttccaaaaaaaatacacagcggtggagcttggcataaagttcaaat
gctcctacaccctgccctgcagtatctctaaccaggggactttgataagg

POU1F1
Pax-4a
Nkx2-5
Lentiviral Poly A
Pax-2
GATA-4
HNF-3alpha
POU2F1
Cdc5
Cdx-2
HNF-4alpha2

Appendix G.

Assoiated Diseases involved in the urea cycle disorders
(Numbers show the degree of association, right column consider the redundancy, listed only those hits greater than 2).

2 Acanthosis Nigricans	2 Aortic Aneurysm, Abdominal
2 Acidemia	2 Arthritis
2 Aortic Aneurysm, Abdominal	2 Bronchopulmonary Dysplasia
2 Bronchopulmonary Dysplasia	2 Chronic heart failure
2 Cerebral atrophy	2 Coma
2 Chronic heart failure	2 Coronary Arteriosclerosis
2 Coma	2 Deep vein thrombosis of lower limb
2 Coronary Arteriosclerosis	2 Diabetes Mellitus, Insulin-Dependent
2 Deep vein thrombosis of lower limb	2 Duodenal Ulcer
2 Developmental delay	2 Epilepsy
2 Epilepsy	2 Essential hypertension, NOS
2 Epstein-Barr Virus Infections	2 HELLP Syndrome
2 Gastritis	2 Hamman-Rich Syndrome
2 Hamman-Rich Syndrome	2 Heart Diseases
2 Heart Diseases	2 Heart failure, NOS
2 Helicobacter Infections	2 Hepatitis, Chronic
2 Hematologic Diseases	2 Hepatitis, Chronic Active
2 Hepatitis, Chronic Active	2 Hereditary Diseases
2 Hepatitis, Toxic	2 Hyperglycemia
2 Hyperinsulinemia	2 Hyperthyroidism
2 Hyperthyroidism	2 Hypothyroidism
2 Hypochondroplasia	2 Infections of musculoskeletal system
2 Immunologic Deficiency Syndromes	2 Inflammatory Bowel Diseases
2 Infections of musculoskeletal system	2 Ischemic stroke NOS
2 Inflammatory Bowel Diseases	2 Kidney Failure, Chronic
2 Ischemic stroke NOS	2 Lens Diseases
2 Kidney Failure, Chronic	2 Liver Failure, Fulminant
2 Labor, Premature	2 Motor Neuron Disease
2 Lens Diseases	2 Nervous System Diseases
2 Liver Cirrhosis, Alcoholic	2 Peripheral Vascular Diseases
2 Liver Diseases, Alcoholic	2 Proliferative diabetic retinopathy
2 Liver Failure, Fulminant	2 Prostatic Hypertrophy, Benign
2 Mastocytosis	2 Psoriasis
2 Mastocytosis, Systemic	2 Salivary Gland Diseases
2 Motor Neuron Disease	2 Thrombocytopenia
2 Nervous System Diseases	2 Thyroid Diseases
2 Osteopetrosis	2 Viral hepatitis
2 Peripheral Vascular Diseases	3 Acquired Immunodeficiency Syndrome
2 Piebaldism	3 Adenovirus Infections
2 Proliferative diabetic retinopathy	3 Adhesions
2 Prostatic Hypertrophy, Benign	3 Cerebrovascular accident
2 Psoriasis	3 Chronic liver disease
2 Salivary Gland Diseases	3 Fetal Alcohol Syndrome
2 Streptococcal lymphadenitis of swine	3 Glomerulonephritis
2 Thrombocytopenia	3 Gonorrhea
2 X-linked agammaglobulinemia	3 Heart Failure, Congestive
2 alcohol flush reaction	3 Hepatitis
3 Acquired Immunodeficiency Syndrome	3 Huntington Disease
3 Acute pancreatitis	3 Kidney Diseases
3 Arthritis	3 Kidney Failure
3 Cerebrovascular accident	3 Muscular Dystrophy, Duchenne
3 Chronic liver disease	3 Parkinson Disease
3 Duodenal Ulcer	3 Prostatic Diseases
3 Glomerulonephritis	3 Retinal Diseases
3 Gonorrhea	3 Septicemia
3 Heart Failure, Congestive	4 Acute myocardial infarction
3 Hepatitis	4 Consumption-archaic term for TB
3 Hepatitis, Chronic	4 Diabetes Mellitus, Non-Insulin-Dependent
3 Kidney Diseases	4 HIV Infections
3 Kidney Failure	4 Hypertension induced by pregnancy

3 Muscular Dystrophy, Duchenne 3 Parkinson Disease 3 Retinal Diseases 3 Septicemia 3 Thyroid Diseases 3 Viral hepatitis 4 Acute myocardial infarction 4 Adenovirus Infections 4 Multiple Sclerosis 4 Prostatic Diseases 5 Consumption-archaic term for TB 6 Asthma 6 Cirrhosis 10 Bacterial Infections 10 Cerebral Vasospasm 10 Common Variable Immunodeficiency 10 Diabetic complication, NOS 10 Endothelial dysfunction 10 Glomerulosclerosis, Diabetic 10 Hypos 10 Keratoconjunctivitis Sicca 10 Mixed Connective Tissue Disease 10 Myelofibrosis 10 Myopathy 10 Myotonic Dystrophy 10 Pneumonia 10 Severe Combined Immunodeficiency 10 Subarachnoid Hemorrhage 11 Acromegaly 11 Amyotrophic Lateral Sclerosis 11 Anemia, Sickle Cell 11 Cardiomyopathy, Congestive 11 Disseminated Intravascular Coagulation 11 Epilepsy, Temporal Lobe 11 Mole NOS 11 Muscular Dystrophies 11 Myocardial Ischemia 11 Neurodegenerative Diseases 11 Systemic vasculitis 11 Uterine Diseases 12 Diabetes Mellitus, Insulin-Dependent 12 Erythema gyratum repens 12 Essential hypertension, NOS 12 HELLP Syndrome 12 Heart failure, NOS 12 Hereditary Diseases 12 Hypothyroidism 13 Huntington Disease 13 Hyperglycemia 14 Adhesions 14 Hypertension induced by pregnancy 14 Pre-Eclampsia 14 Virus Diseases 15 Diabetes Mellitus, Non-Insulin-Dependent 15 Fetal Alcohol Syndrome 15 HIV Infections 15 Lupus Erythematosus, Systemic 16 Lung diseases 18 Colonic Diseases 19 Rheumatoid Arthritis	4 Multiple Sclerosis 4 Pre-Eclampsia 4 Virus Diseases 5 Asthma 5 Cirrhosis 5 Colonic Diseases 5 Lung diseases 5 Lupus Erythematosus, Systemic 8 Rheumatoid Arthritis

vita

Ming Chen (陈铭) was born in Yueqing, Zhejiang province, China, on November 13, 1972. After graduating from Yueqing High School, he enrolled at Zhejiang University of Technology, where he received the degree of Bachelor of Science in Biochemical Engineering in 1995, and of Master in Biochemical Engineering in 1998. After one and half years working in Zhejiang University of Technology, he began his Ph.D. studying at the Otto-von-Guericke University Magdeburg in February 2000. Then he entered the Graduiertenkolleg “Bioinformatik” of the University of Bielefeld in July 2001.

E-mail address: mchen@techfak.uni-bielefeld.de
ming.chen@web.de