

**DATA MODELING FOR BIOCHEMICAL PATHWAY AND  
MICROARRAY GENE EXPRESSION**

**YIMIN LIU**

**A THESIS  
IN  
THE DEPARTMENT  
OF  
COMPUTER SCIENCE**

**PRESENTED IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF MASTER OF  
COMPUTER SCIENCE  
CONCORDIA UNIVERSITY  
MONTREAL, QUEBEC, CANADA**

**AUGUST 2003**

**©YIMIN LIU, 2003**

National Library  
of Canada

Bibliothèque nationale  
du Canada

Acquisitions and  
Bibliographic Services

Acquisitions et  
services bibliographiques

395 Wellington Street  
Ottawa ON K1A 0N4  
Canada

395, rue Wellington  
Ottawa ON K1A 0N4  
Canada

*Your file    Votre référence*

*ISBN: 0-612-83913-3*

*Our file    Notre référence*

*ISBN: 0-612-83913-3*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

**Canada**

# **Abstract**

## **Data Modeling for Biochemical Pathway and Microarray Gene Expression**

Yimin Liu

Biochemical pathways are typically thought as complex networks of chemical compounds and reaction in the living organisms. A great amount of gene expression data that have a bearing on pathway is created at an increasing rate by microarray technology. The large ensemble of information that the gene expression experiments produce contains pattern that are a reflection of pathway dynamic, and therefore can be used to deduce pathway causal structure. Conceptual data modelling involves the development of implementation-independent models that capture and make explicit the principal structural properties of the data. This thesis develops and presents such a data modeling for biochemical pathway and microarray gene expression. The conceptual data models can be transformed in systematic ways for implementation using different platforms, e.g. traditional database management system and visualisation pathway application. This conceptual data model is described by widely used conceptual modelling notation: the class diagrams of UML (unified modelling language).

## Acknowledgements

I would like to express my warmest gratitude to my thesis supervisor, Dr. Gregory Butler, for his patience and invaluable guidance. His profound knowledge in computer science and bioinformatics is highly appreciated. I am grateful for his teaching in bioinformatics, such as Bioinformatics Algorithms (COMP691) and Bioinformatics Databases and Systems (COMP691S). Without them, this thesis would have been impossible to achieve.

I would like to thank all graduate students in Dr. Butler's group especially Yan Meng and Jian Sun for their helpful discussion and friendships. Here, I also express my thanks to the reviewers who give me good comments on the thesis draft.

I dearly thank my parents and my wife for their understanding and encouragement during my graduate study.

# Table of Contents

<b>LIST OF FIGURES .....</b>	<b>VII</b>
<b>LIST OF TABLES .....</b>	<b>VIII</b>
<b>CHAPTER 1. INTRODUCTION .....</b>	<b>1</b>
<b>1.1 MOTIVATION.....</b>	<b>3</b>
1.1.1 The Need for Modeling Biological Data.....	3
1.1.2 The Benefits of Conceptual Data Modeling.....	3
1.1.3 Final Goal for Our Data Models.....	5
<b>1.2 CONTRIBUTION OF THE THESIS.....</b>	<b>6</b>
<b>1.3 ORGANIZATION OF THE THESIS .....</b>	<b>7</b>
<b>CHAPTER 2. BACKGROUND .....</b>	<b>9</b>
<b>2.1 DATA MODELS .....</b>	<b>9</b>
<b>2.2 CONCEPTUAL DATA MODEL .....</b>	<b>9</b>
2.2.1 Type of Model.....	12
2.2.2 Selection of Data Model.....	13
<b>2.3 OVERVIEW OF MICROARRAY AND GENE EXPRESSION.....</b>	<b>14</b>
2.3.1. Gene Expression Analysis Technologies .....	14
2.3.2 Microarray Technology .....	14
2.3.3 Biological Assumptions for Microarray Technology .....	15
2.3.4 GeneChips.....	16
2.3.5 cDNA Microarrays.....	17
<b>2.4 BIOCHEMICAL PATHWAYS AND REGULATION.....</b>	<b>18</b>
2.4.1 Biochemical Pathway Introduction.....	18
2.4.2 Regulation .....	20
2.4.3 Metabolic Regulatory Pathway Case Study .....	21
2.4.4 Gene Expression in Regulating Biochemical Pathways.....	23
<b>CHAPTER 3. THE STATE-OF-THE ART .....</b>	<b>25</b>
<b>3.1 BIOCHEMICAL PATHWAY DATABASES .....</b>	<b>25</b>
3.1.1 Metabolic Pathway Database .....	26
3.1.2 Contents of Metabolic Pathway Databases .....	27
3.1.3 KEGG.....	29
3.1.4 EcoCyc/MetaCyc.....	33
3.1.5 WIT (What Is There?).....	35
<b>3.2 BIOCHEMICAL PATHWAY VISUALIZATION.....</b>	<b>36</b>
3.2.1 Biochemical Pathway Requirement Analysis .....	36
3.2.2 Representation Levels of Pathway .....	37
3.2.3 Construction of Pathway Diagram .....	38
3.2.4 Pathway Visualization Tools.....	39
<b>3.3 MICROARRAY GENE EXPRESSION DATA STANDARD AND MAGE-ML .....</b>	<b>41</b>
3.3.1 Minimum Information about a Microarray Experiment – MIAME.....	42
3.3.2 MAGE-ML.....	50
3.3.3 ArrayExpress, A Public Repository for Gene Expression Data.....	52
<b>CHAPTER 4. CONCEPTUAL DATA MODELS OF METABOLIC PATHWAY .....</b>	<b>55</b>
<b>4.1 DEVELOPING A DATA MODEL FOR BIOCHEMICAL PATHWAYS .....</b>	<b>55</b>
<b>4.2 THE POSSIBLE SCOPES AND CONSTRAINTS IN OUR CONCEPTUAL DATA MODEL .....</b>	<b>55</b>
4.2.1 <i>Compound</i> Object.....	56
4.2.2 <i>Enzyme</i> Object .....	57
4.2.3 <i>Reaction</i> Object .....	57
4.2.4 <i>Gene</i> Object .....	57
4.2.5 <i>GeneExpression</i> Object .....	58
<b>4.3 UML MODEL FOR METABOLIC PATHWAY COMPONENTS .....</b>	<b>58</b>

4.4 UML MODEL FOR BIOCHEMICAL PATHWAY CLASSIFICATION.....	61
CHAPTER 5. DISCUSSION AND CONCLUSION .....	64
5.1 THE CONSTRUCTION OF DATA MODEL IS AN ITERATIVE PROCESS.....	64
5.2 THE UML DATA MODEL DESCRIBES THE DATA IN BIOCHEMICAL PATHWAY .....	65
5.3 THE UML DATA MODEL CAN GENERATE PATHWAY DIAGRAM.....	68
5.4 THE UML DATA MODEL ARE DIFFERENT FROM KEGG AND ECO CYC DATA MODEL .....	70
5.5 MORE STANDARDS FOR NOMENCLATURE ARE NEEDED .....	71
5.6 MIAME AND MAGE ARE USEFUL .....	72
REFERENCES .....	74
APPENDIX A: LAC OPERON MODEL <sup>(64)</sup> .....	81
APPENDIX B MIAME VERSION 1.0 .....	87
APPENDIX C: MAGE-ML AND MAGE-OM .....	99
APPENDIX D: GLOSSARY IN BIOLOGY AND BIOINFORMATICS <sup>(27)</sup> .....	113

## List of Figures

FIGURE 1. ELMASRI R. AND NAVATHE S. PROPOSE THE DESIGN PROCESS <sup>(13)</sup> .....	11
FIGURE 2. AN OVERVIEW OF A METABOLIC PATHWAY: PROLINE BIOSYNTHESIS.....	22
FIGURE 3. ENZYMATIC CATALYSIS AND ITS REGULATION.....	23
FIGURE 4. THE ROLE OF GENE EXPRESSION IN REGULATING METABOLIC PATHWAY.....	24
FIGURE 5. THE DATA MODEL CONCEPT OF KEGG AND ITS RELATION TO DBGET <sup>(65)</sup> .....	32
FIGURE 6. THE TOP OF THE CLASS HIERARCHY FOR ECOCYC DATABASE <sup>(66)</sup> .....	34
FIGURE 7. SIX MIAME COMPONENTS <sup>(54)</sup> .....	47
FIGURE 8. PATHWAY DATA MODEL SCOPE.....	56
FIGURE 9. PATHWAY ELEMENT CONCEPTUAL DATA MODEL .....	59
FIGURE 10. BIOCHEMICAL PATHWAY CLASSIFICATION CONCEPTUAL DATA MODEL.....	63
FIGURE 11. PROLINE BIOSYNTHESIS PATHWAY ARE DISPLAYED WITH UML CONCEPTUAL DATA MODEL.....	66
FIGURE 12. PROLINE BIOSYNTHESIS PATHWAY IS VISUALIZED AS DIRECTED GRAPHS WITH THE UML DATA MODEL.....	69

## List of Tables

TABLE 1. STEREOTYPES TO INDICATE MODEL TYPES (CORE NOTATION) BY SCOTT W. AMBLER.....	13
TABLE 2. METABOLIC PATHWAY DATABASES AND URL .....	28
TABLE 3. COMPARISON OF UML CONCEPTUAL, KEGG, AND ECOCYC DATA MODEL.....	71



# Chapter 1. Introduction

Data modeling techniques are used during the information system design and analysis and are important kinds of techniques. Conceptual data modeling is the process of developing a conceptual data model that is a complete and accurate representation of an organization's data requirements<sup>(1)</sup>. In the thesis, I am going to develop a conceptual data model, which can represent and describe biochemical pathway data and microarray gene expression data. In chapter 3, I will give a detailed discussion about what is biochemical pathway and microarray gene expression(see chapter 3 and appendix D).

Object-oriented Modeling has become the prime methodology for modern software design<sup>(2)</sup>. In the mean time, more and more attention has been paid on object-oriented data modeling. Some approaches to conceptual data modeling advocate the use of abstract formalisms for describing data, mostly based on the notion of class<sup>(3)</sup>.

Object-orientation in software creation appears to be relatively simpler than object-oriented data modeling, since a specific program represents one approach to a solution, and hence one point-of-view. Data are usually shared, and participants can hence approach the modeling from multiple points-of-view<sup>(4)</sup>.

In the thesis, I am trying to use object-oriented conceptual data modeling for biochemical pathway and microarray gene expression. As we already know, using

object-oriented data modeling can give us a better understanding of the requirements, clear designs, and more maintainable systems.

Today, the data gained is quite different from which gained in traditional biology. Biological data are flooding in at unprecedented rate<sup>(5)</sup>. Before, biology is a kind of experimental science. The knowledge biologist gained is through doing experiments. However, the experiment is done manually by biologists, which takes too long time, so as the data gain slowly. Since 1987, the human genome project<sup>(6)</sup> has made biology into a field overwhelmed by data. Biologists face a big challenge in computing. Automatically Sequencing DNA and microarray technologies give a lot of data in such a short time that there are lots of issues about data processing. This is the field of bioinformatics, which involves at least three areas, computer science, mathematics, and statistics. It also includes the integration and mining of the ever-expanding databases of information.

Microarray technologies monitor the combinatorial interaction of a set of molecules, such as DNA fragments and proteins, with a predetermined library of molecule probes. The most advanced of these technologies currently is the use of DNA arrays, also called DNA chips, for simultaneously measuring the level of the mRNA gene products of a living cell<sup>(7)</sup>. In chapter 3, I will give more detailed discussion.

A design activity may involve a combination of a design process and a modeling language. I will use UML as my modeling language for the design activity.

## 1.1 Motivation

There is to be prepared for the flood of gene expression data from microarray experiments<sup>(5)</sup>. In particular, to be ready to extract knowledge about cell control and regulation mechanism that may involve cell death. Hence, we wish to clearly model gene expression metabolic pathway, biosynthesis pathway and regulatory mechanism.

### 1.1.1 The Need for Modeling Biological Data

Biological database development and maintenance are in the scope of bioinformatics. Due to the flood and heterogeneity of biological data, database systems today are facing the task of serving ever increasing amounts of data of ever growing complexity to a user community that is growing nearly as fast as the data, and is getting more and more demanding<sup>(5)</sup>. A sound database design is more important and costs less.

To set up biological database system (for instance, microarray gene expression databases), we must have sound data modeling, as the system design and analysis change cost much less than code change.

### 1.1.2 The Benefits of Conceptual Data Modeling

The two properties of abstraction and transparency of a conceptual data model provide the benefits of accuracy and clarity.

Since data modeling is doing data abstraction. Lots of same instances are abstracted into one abstract data model. Many repeated situations are modeled as a single general situation. For instance, in a biochemical pathway, lots of different metabolites in living organisms may be considered as one reactant A, which reacts with another reactant B, and then the reaction produces the product C and product D under the enzyme catalysis. The only thing different may be that the product C and D may work as new reactants of product D and E. It will happen like this all the time recursively. So, when we do some acts of abstraction, we will get the benefit from this abstraction, as we save the labor to record all the instances.

Also, data modeling is one kind of process that classifies data into different groups by type and groupings, which is kind of transparency property. For instance, in biochemical pathways, there are at least six major different objects if we do data modeling(see section 4.2 and 4.3). They are *Metabolite*, *Cofactor*, *Reaction*, *Enzyme*, *Gene*, and *GeneExpression*. The six different objects, *Metabolite*, *Cofactor*, *Enzyme* and *Gene* may be better considered as one similar group. In the mean time, *Metabolite* and *Cofactor* may go further and fall into one subgroup of those four object group, as they can be thought as a compound. *Gene expression* and *transformation* may be considered into the other group, as *gene expression* and *transformation* object can make chemicals changes.

Now, it is not difficult to see what the benefit is to do data modeling. It makes your data have clarity and accuracy. I use the UML conceptual data model to do data modeling and reach the above two goals, since these two properties belong to conceptual data modeling.

### **1.1.3 Final Goal for Our Data Models**

Genome sequencing and DNA microarray gene expression analysis has become the most widely used source of genome-scale data in the life sciences. Microarray expression studies are producing massive quantities of data. It promises to provide key insights into gene function and interactions within and across metabolic pathways<sup>(8-10)</sup>.

Genome sequence data have standard formats for presentation and widely used tools and databases. However, although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. The data models described in this thesis try to seek a standard from many potential standards to follow. In the meantime, we pursue a clear and agreed view for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of biochemical pathway database or public repositories and enable the development of biochemical pathway data analysis tools at Concordia.

In the life sciences, our Know-It-All database framework<sup>(11)</sup> provide new data models natural to life science, enhanced operations on these data types, and optimized performance.

## **1.2 Contribution of the Thesis**

The principal contribution of this thesis is that it explicitly provides conceptual data model that the data involved in regulatory metabolic pathway and gene expression in our understanding. Not only are all kinds of objects modeled, but also, various relationships among them are tried to dig out. This model is an object-oriented conceptual data model and may be used for pathway visualization.

I summary the extent databases and data models for pathways and gene expression. I briefly review MIAME, MAGE-ML, and ArrayExpress to determine whether they meet our needs in future, as MIAME has become the standard for the field of microarray, which is one of the developments of suitable microarray data standards we can follow. I also discuss some visualization tools including pathway visualization tools.

In short, we design our conceptual data model by object-oriented data modeling. Obtaining a clear understanding of the semantics of a piece of data in life science is a real challenge, as different people will see things in different ways, and use different features. The situation is further complicated since life science is non-axiomatic and the views on the same or similar concepts vary strongly among different communities. However, our

conceptual data models may be really helpful to this situation in developing, making explicit and communicating clear and detailed descriptions of biochemical pathway data that is available or about to be produced.

### **1.3 Organization of the Thesis**

There are five chapters and three appendixes in the thesis.

Chapter 2: I provide a detailed description of the background that includes introducing some basic knowledge about conceptual data models and some knowledge about biochemical pathway and microarray gene expression technology. The information in this chapter is quite helpful in understanding our data models in the thesis.

Chapter 3: I first give an overview of some biochemical pathway databases that are mainly focused on metabolic pathway. I discuss basic requirements about pathway visualization and then review some tools for pathway visualization since visualization tools is an important requirement for the interpretation of pathway data. Finally, I present and analyse MIAME (Minimum Information About Microarray Experiment) and MAGE-ML (MicroArray and Gene Expression Markup Language.) as well as ArrayExpress that is the database system which implemented the MAGE-OM object model. Clearly understanding them will be helpful for the design of our own microarray experiment and data management system in future.

Chapter 4: The UML conceptual data models for metabolic regulatory pathway are developed and presented.

Chapter 5: I conclude this thesis and check the UML data model if it meets the pathway requirements mentioned in Chapter 2. Last, I give some suggestions for our future work.

Appendix A is the lac operon model of *E. coli*.

Appendix B is MIAME (Minimum Information about Microarray Experiment) version 1.0.

Appendix C is MicroArray Gene Expression Markup Language and MicroArray Gene Expression Object Model.

Appendix D is a glossary of bioinformatics terms in the thesis.



## **Chapter 2. Background**

In this chapter, I will briefly introduce some basic knowledge about data models and some background knowledge about biochemical pathway and microarray gene expression technology. The information in this chapter is quite helpful to get a clear understanding our data models in the thesis.

### **2.1 Data Models**

A data model is the collection and identification of concepts for describing data, relationship between data and constrains on the data in an application domain. An accurate and clear representation of a data application domain (here, biology and life science is our application domain) is the key to successfully develop complex bioinformatics applications.

### **2.2 Conceptual Data Model**

In June 2001's DAMA(Data Modeling) Chicago Meeting, Dr. Duncan Dwelle, the president of Applied Information Science says: "the conceptual model is concerned with the real world view and understanding of data."

Dr. William G. Smith says: "The purpose of a Conceptual Data Model is 'to clear up a few things around here.' Whether the scope of the data model is the entire enterprise

or a single system project, the object is to identify and clearly define the entities (persons, places, things, concepts and events) about which the business must keep data, and to identify and clearly define the important associations between those entities.”

A conceptual data model (CDM) can give a notation by which the structural properties of data (the structuring of data and their relationships) from a certain domain (a field of knowledge such as biochemical pathway or gene expression in this thesis) can be described in a precise but implementation-independent manner<sup>(12)</sup>.

The role of the conceptual data model in the design process is to allow precise statements to be made about the data of interest in a manner that can be communicated to others<sup>(13)</sup>. The comprehensibility of a conceptual model is important, as it is used both in discussions with subject experts whose understanding of the relevant data is to be described, and by the developers of software who are to construct applications. A usual remark on conceptual data models is that they are usually much easier to read than to construct.

Conceptual modeling considers the transformation specifications into implementation as a subsequent stage in the design process, to denote that the emphasis is placed on clean concepts rather than on implementation technique. Conceptual data models make explicit the structural properties of data, and as such are useful for capturing, refining and communicating details about the data in a database or a

laboratory<sup>(12)</sup>. Constructing conceptual models is considered as a challenging, often iterative process(see Figure 1).

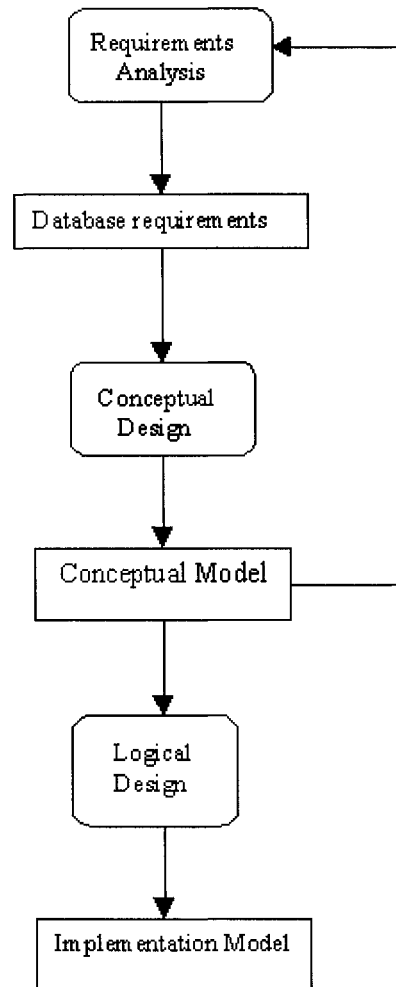


Figure 1. Elmasri R. and Navathe S. propose the design process<sup>(13)</sup>

In this thesis, our data models discussed will be mainly focused on the conceptual data model (Figure 1).

### 2.2.1 Type of Model

There are many different data modeling notations, such as entity-relationship data model<sup>(15)</sup>, object-relational data model<sup>(14)</sup>, and object-oriented data model<sup>(13)</sup>, although the most well-known families are the entity-relationship (ER) data models and the object-oriented models.

According to the profile of UML<sup>(16)</sup> data modeling proposed by Scott W. Ambler<sup>(17)</sup>, there are four model types we can use (see Table 1).

1. Object-oriented data models allow real world data to be represented as objects and allow new classes to be created by extending the description of the parent class. Objects encapsulate the data and provide methods to access or manipulate it.
2. Object-relational data models are improved relation models by adding some features from object data models. Information is represented as in relational models.
3. Logical data models (LDMs) are used to set up either the conceptual design of a database or the detailed data architecture of your application domain. LDMs describe the logical data entities, typically referred to simply as data entities, the data attributes depicting those entities, and the relationships between the entities.

4. Physical data models (PDMs) are used to design the internal schema of a database, depicting the data tables, the data columns of those tables, and the relationships between the tables.
5. Conceptual data models are typically used to explore domain concepts with project stakeholders. Conceptual data models are often created as the precursor to LDMs or as alternatives to LDMs.

Table 1. Stereotypes to Indicate Model Types (Core notation) by Scott W. Ambler

Stereotype	Model Type
<<Class Model>>	Object-oriented or object-relational model
<<Conceptual Data Model>>	Conceptual data model
<<Logical Data Model>>	Logical data model (LDM)
<<Physical Data Model>>	Physical data model (PDM)

### 2.2.2 Selection of Data Model

In this thesis, what I want to pursue is a kind of ‘clean’ design, which is sort of implementation-independent. So, I will select the conceptual data model (see Figure 1) as an approach to develop our biochemical pathway database design. I will use the UML class diagram to represent our conceptual data model. UML is one of standards object-oriented modeling language. The focus is on class diagrams, which are used to model the structural aspects of data within UML. As an object-oriented modeling language, the central notation in UML is the class diagram.

## **2.3 Overview of Microarray and Gene Expression**

### **2.3.1. Gene Expression Analysis Technologies**

Genome-wide expression information is principally generated by three technologies: cDNA microarrays<sup>(18)</sup>, GeneChips<sup>(19)</sup> (also called high-density oligonucleotide arrays) and SAGE<sup>(20)</sup> (serial analysis of gene expression). These technologies are all new and rapidly evolving.

GeneChips and SAGE measure the absolute gene expression levels, which is mRNA level in living organisms. Since the technology itself reasons, cDNA microarray primarily measure the relative level of gene expression, which yield an ‘expression ratio’. From data analysis view, there are major two widely used gene analysis measures in bioinformatics.

In the thesis, I focus on the microarray technology itself, which is used to measure all gene expression in the living organisms and get ‘casual’ information of biochemical pathway<sup>(38)</sup>. Some biology and bioinformatics terms can be found in Appendix D.

### **2.3.2 Microarray Technology**

Microarray technologies monitor the combinatorial interaction of a set of molecules, such as DNA fragments and proteins, with a predetermined library of

molecular probes. The current widely used implications are DNA arrays, also called DNA chips.

The advent of microarray technology will allow the analysis of gene expression of thousands of genes simultaneously, so it creates a comprehensive transcriptional profile of condition studies. Computational biologists use them in order to compare these profiles taken from organisms under control condition and an alternative (e.g., pathogenic) condition, or compare these gene expression profiles between two systems under one or several conditions. For the first time, investigators can relatively quickly measure the expression of a complete genome across a large number of environmental stimuli. This awe-inspiring technological breakthrough has the potential to impact some previously intractable scientific realms and aid in the elucidation of complex models and systems<sup>(21)</sup>.

### **2.3.3 Biological Assumptions for Microarray Technology**

Gene expression information is very important to the understanding of many aspects of cellular and organism function. Regardless of the gene expression technology to be adopted, all of them have the following three general and fundamental biological assumptions<sup>(22)</sup>.

- \* There is a close correspondence between mRNA transcription and its associated protein translation. As mentioned by Brown and Botstein<sup>(21)</sup>, one

would ideally like to measure the final products of every gene, such as proteins, or even better, the biochemical activity of these products, which are directly related to biological functionality. Such quantity would provide a link between chemical DNA bases at microscopic levels with biological aspects that are manifest at macroscopic scales such as phenotype and physiology.

- \* All mRNA transcripts have identical life span. There are several well-known exceptions. For instance, the length of 3' poly-A tail of an mRNA appears to be related to its stability.
- \* All cellular activities and responses are entirely programmed by transcriptional events. There is also a much larger class of biological processes that do not primarily operate at the transcriptional level. These include muscular contraction, nerve excitation, and hormonal release, but the pattern of gene expression would probably not reveal the control process that govern them at the sub-genomic time scale.

#### **2.3.4 GeneChips**

GeneChips<sup>(19)</sup> (Affymetrix GeneChip) technology provides an array of 250,000 probes, each probe containing a set of oligonucleotides of approximately 25 base pairs each representing a region of interest within a gene. The array is exposed to cDNA developed from a cell in which it is hypothesized these genes are expressed. The cDNA then hybridizes (attaches) to complementary sequences on the array. As with other microarray technologies, the strategy is to identify which genes in the cell are expressed,



and to what degree, based on the extent of hybridization observed at each probe. Fluorescent molecules attached to the cDNA create an intensity of light that corresponds to the degree of hybridization.

### **2.3.5 cDNA Microarrays**

A second technology is cDNA microarrays, namely, the robotically spotted cDNA glass slide, where the mechanical deposition of entire cDNA onto an array is done by using carefully designed metal pen nibs controlled by a robotic arm.

The array is exposed to equal amounts of green and red fluorescent dyed samples, corresponding to normal and affected cells, respectively. The color at each site indicates the relative amount of hybridization corresponding to the relative expression in the two cells of the cDNA at that site. That is, yellow indicates expression in both cells; black in neither; red in only the affected cell; green only in the normal cell.

This technology was introduced into common use at Stanford University and first described by Mark Schena<sup>(23)</sup> et al. in 1995. They are also known as cDNA microarray.

In making the array, a robotic spotter mechanically picks up specific cDNA sequences, which are amplified from vectors in bacterial clones using PCR(see Appendix D), from separate physical containers and deposits them in specific locations in the grid on the glass slide to create specific probes. Each cDNA drop should ideally be equal in

quantity. This fabrication approach epitomizes the do-it-yourself tendency in microarray measurement, even though there are several commercial ready-to-use versions available. The frequently home-grown quality of these arrays has led to do the production of highly localized and customized microarrays which pose specific signal amplification during subsequent data analysis stages.<sup>(24)</sup>

## **2.4 Biochemical Pathways and Regulation**

The behavior of an organism depends not only on the nature of the proteins that are expressed, but also on the extent to which they are expressed and the environmental conditions under which this expression occurs. Of course, some housekeeping proteins that build cell architecture must be made all the time. The ability to synthesize materials only as needed, therefore, would make sense for economy and adaptation. In general, the synthesis of particular gene products is controlled by mechanisms that are collectively called “gene regulation.” Regulation of gene expression not only exists, but also makes cellular adaptation, variation, differentiation, and development possible<sup>(26)</sup>.

### **2.4.1 Biochemical Pathway Introduction**

Biochemical pathways are bioprocess structures that researchers use to describe the dependencies and consequences of a system of bio-molecular interactions, thereby providing insight into the significance and purpose of a pathway. These constructs often

begin as a sequence of biochemical steps that either process material or transduce information, and then eventually increase in complexity as more knowledge is obtained.

In living organism, cells function as organized chemical engines carrying out a large number of transformations, called bio-reactions or biochemical reactions, in a coordinated manner. These reactions are catalyzed by enzymes and exhibit great specificity and rates much higher than the rates of non-enzymatic reactions. Enzymes are neither transformed nor consumed, but they facilitate the underlying reactions by their presence. The coordination of the extensive network of biochemical reactions is achieved through regulation of the concentrations and the specific activities of enzymes. Single enzyme catalyzed steps in succession form long chains, called biochemical pathways, achieving the overall transformation of substrates to far the removed products.

From an artificial intelligence researcher, Michael L. Mavrovouniotis' point of view<sup>(25)</sup>, biochemical pathways are often described in symbolic terms, as a succession of transformations of one set of molecules (called reactants) into another set (called products); reactants and products are collectively referred to as metabolites. In the construction of metabolic pathways one uses enzyme-catalyzed bio-reactions as building blocks, to assemble pathways that meet imposed specifications. A class of specifications can be formulated by classifying each available building block, i.e., each metabolite and each bio-reaction, according to the role it can play in the synthesized pathways. For example, a set of specifications may include some metabolites designated as required

final products of the pathways, other metabolites as allowed reactants or by-products, and some bio-reactions as prohibited from participating in the pathways.

### **2.4.2 Regulation**

The regulatory systems of prokaryotes and eukaryotes are somewhat different. For convenience, I will briefly discuss on the basic model in bacteria, as it is one of the best-understood regulatory mechanisms. In bacterial systems, when several enzymes act in sequence in a single metabolic pathway, usually either all or none of these enzymes are produced. This phenomenon, coordinated regulation, results from control of the synthesis of a single polycistronic mRNA molecule that encodes all the gene products. So, I will be focus on transcriptional regulation.

There are several mechanisms for regulation of transcription<sup>(26)</sup>. The particular one used often depends on whether the enzymes being regulated act in degradative or synthetic metabolic pathways. That is, does the action of the enzymes in question break down a substance into a more useful compound (degradative), or is the desired molecule being “built”? In a multi-step degradative system, the availability of the molecule to be degraded frequently determines whether or not the enzymes involved in the pathway will be synthesized. In contrast, in a biosynthetic pathway the final product is often the regulatory molecule. The molecular mechanisms for each of the two regulatory patterns vary widely, but usually fall into one of two major categories: negative or positive regulation.

In negatively regulated systems, a specific protein (called a repressor protein) that inhibits transcription of a specific gene may be present in the cell. In some cases the repressor alone acts to prevent transcription, and a molecule (called an inducer) that is an antagonist of the repressor is needed to allow transcription. In other instances of negative regulation, the repressor on its own does not inhibit transcription—it does so only when combined with a specific signal molecule. In a positively regulated system, a protein called an activator works to increase the frequency of an operon<sup>(26)</sup> (see Appendix A). In the thesis, I have an appendix for The Lac Operon model and its regulation to give more detailed information about transcriptional regulation.

### **2.4.3 Metabolic Regulatory Pathway Case Study**

#### **2.4.3.1 Metabolic Regulation: Proline Biosynthesis in *E. coli***

The example used in Figure 2 is proline biosynthesis in *E. coli*, which involves a chain of generated protein shown. One of the final products of the chain, proline, inhibits the initial reaction (see Figure 2), which has started the whole process. This “feedback inhibition” pattern is highly typical to metabolic pathways and genetic networks, and serves to regulate the process execution rate. This example tries to give you an overview of biochemical pathway. There are two major parts. On the left part is kind of regulation part, for instance, proB gene codes for gamma-glutamyl kinase, which specifically catalyzes the reaction 2.7.2.11. The right part is kind of the whole proline synthesis

pathway. Glutamate is transformed into gamma-glutamyl phosphate and along with ATP is transformed into ADP. And, the next reaction 1.2.1.41 will transform gamma-glutamyl phosphate into 1-pyrroline-carboxylate. Similarly, it happens in same time that NADPH is transformed into NADP. This character is typical of biochemical pathway. One metabolite is transformed into another. Another one then will become the third different metabolite.

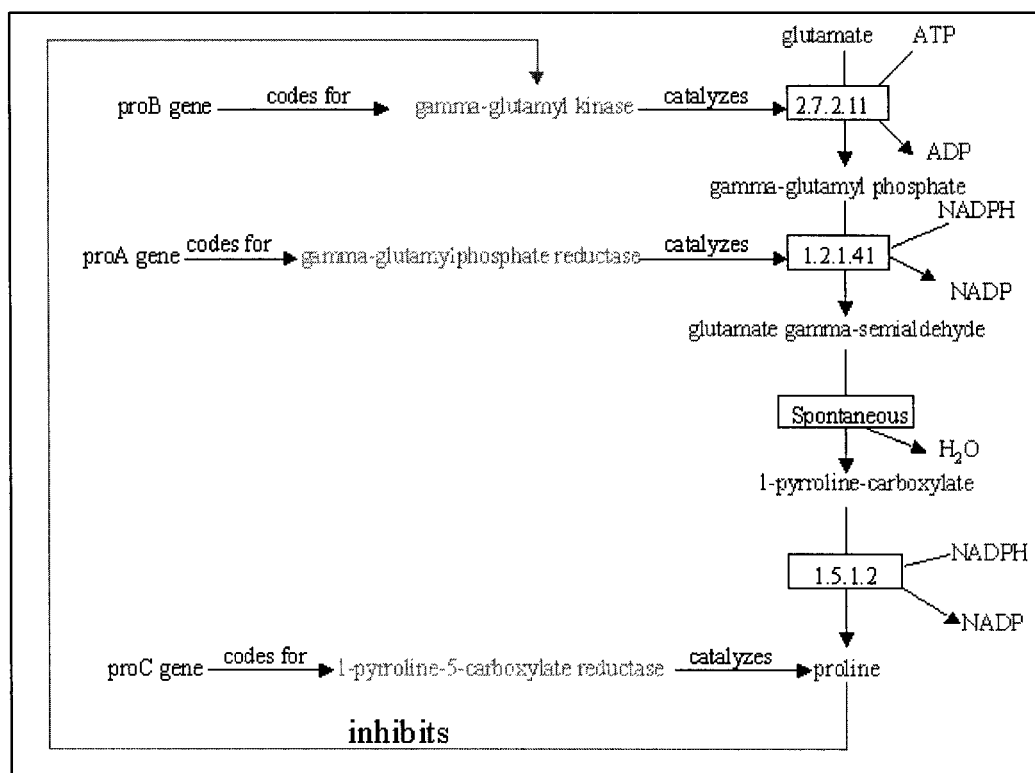


Figure 2. An overview of a metabolic pathway: Proline Biosynthesis

#### 2.4.3.2 Enzyme catalysis and its regulation in metabolic pathway

This example used in Figure 3 is the catalysis of the phosphorylation of glutamate by  $\gamma$ -glutamyl kinase, which is regulated by proline. The final product, proline, in the proline synthesis metabolic pathway inhibits  $\gamma$ -glutamyl kinase activity(the initial reaction). The catalyzed reaction is indicated by EC number (2.7.2.11). The compound names are marked as labels.

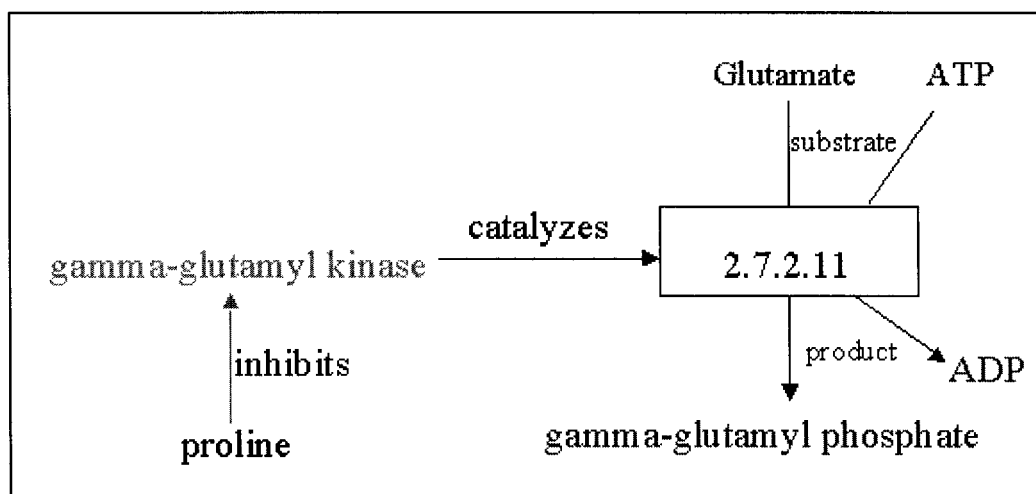


Figure 3. Enzymatic catalysis and its regulation

#### 2.4.4 Gene Expression in Regulating Biochemical Pathways

Figure 4 depicts gene expression and its role in catalyzing certain chemical reactions in metabolic pathways. The proB gene is being expressed into  $\gamma$ -glutamyl kinase protein, which catalyzes a reaction involving glutamate and ATP, which produces  $\gamma$ -glutamyl phosphate and ADP compounds. This example tries to give you closely look at how reaction happen between two compounds. The reaction 2.7.2.11 will be catalyzed by gamma-glutamyl kinase. This kinase is coded by proB gene in Ecoli. ProB gene expression has a very close connection with proline synthesis pathway. So, if proB gene

is lowly expressed, this kinase would have a tiny amount left in the living Ecoli, proline synthesis will be greatly effective so as proline will become lowly produced in Ecoli.

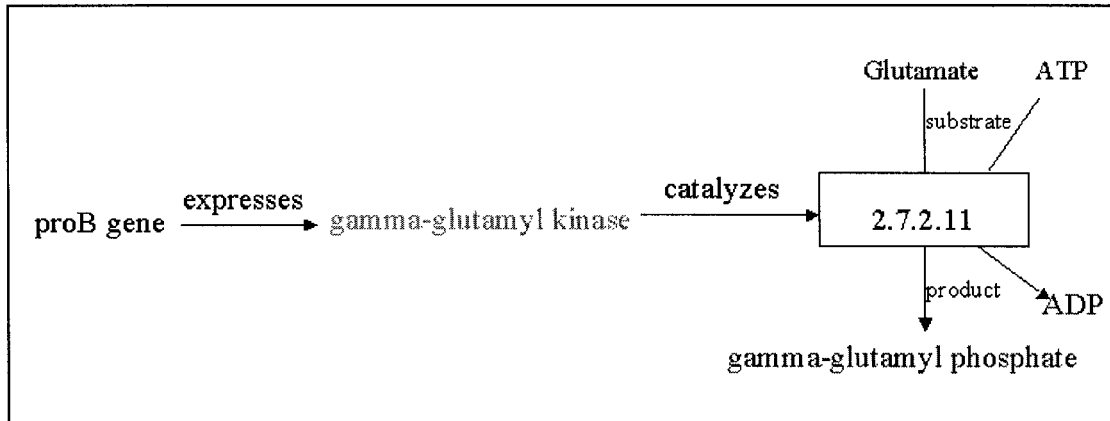


Figure 4. The role of gene expression in regulating metabolic pathway



## **Chapter 3. The State-of-the Art**

The understanding of the interplay of gene and gene products is the new challenge in functional genomics study. From various kinds of interactions (protein-gene, protein-protein), causal, regulated networks of biochemical pathway arise. Such networks are responsible for the development, maintenance, and responsiveness of all living systems. The collection and organization of pathway information is critical and needs to be effectively addressed.

In this chapter, I will briefly overview some biochemical pathway databases and mainly I will focus on metabolic pathway. Also, I will discuss basic requirements about metabolic pathway visualization and then review some tools for pathway visualization since visualization tools is an important requirement for the interpretation of pathway data. Finally, I will briefly introduce and discuss Minimum Information About a Microarray Experiment (MIAME) and Microarray Gene Expression Markup Language (MAGE-ML) as microarray gene expression data is a source of pathway ‘casual’ information, which promise to provide a key insight into gene function and interaction within and across metabolic pathway.

### **3.1 Biochemical Pathway Databases**

Databases developed can be classified into different categories, including genome database, protein databases, enzyme databases, pathway databases, literature databases

and some very specific databases. This classification of databases is based on their biological content. Although the content of the database is mostly restricted to specific biochemical compounds or functions, a lot of overlap occurs<sup>(28)</sup>.

Enzyme databases mainly contain information about enzymes and their properties. On the other hand pathway databases reflect information about reaction and pathway in general, data related to organism-specific information about genes, their related gene products, protein functions, expression data, data about enzymatic activities, kinetic data, etc.

Pathway databases can also be sub-classified into databases containing metabolic pathways, signaling pathways and gene regulatory pathway. In the thesis, I will mainly focus on metabolic pathway database.

### **3.1.1 Metabolic Pathway Database**

Metabolic pathway databases are a new kind of bioinformatics resource with a wide variety of potential uses in academia and in industry. These databases can serve as online resources. It makes biochemical pathway information readily accessible via the Internet. Metabolic pathway databases can also let scientists who study metabolism to pose new questions about metabolic networks<sup>(25)</sup>.

### 3.1.2 Contents of Metabolic Pathway Databases

Collections of enzymes, reactions and biochemical pathways are typically used to describe metabolic pathway databases. Software usually is coupled with these databases to query and visualize metabolic pathway data information. Metabolic databases can describe either the biochemistry of a single organism or many organisms. Some metabolic pathway databases provide a more approximate collection of pathway data information that is not specific to any organism. Some databases are either derived from, and tightly linked to the primary bio-medical literature or derived from secondary sources. Each collection of actual data varies significantly. Enzymes, pathways and chemical substrates, which also include genomic information are described in the most comprehensive databases. They also tightly link the data to the primary literature, providing citations for most information. Other metabolic databases provide only a subset of this information. Computer algorithms produce automatically graphical drawings in these databases. There are also some hand drawings of pathway diagrams. The visualizations may include pathways, reactions, substrate structures and entire metabolic networks. Some metabolic databases, but not all, can contain links to other biological databases, such as to the SWISS-PROT protein sequence database<sup>(29)</sup>.

Generally, metabolic pathway databases use EC numbers(see chapter 2.4.3). It consist of 4-digit number. EC stands for Enzyme Commission. The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) devises the enzyme nomenclature system. For instance, enzymes that catalyze the

reaction  $\text{O}_2 + \text{oxalate} = \text{H}_2\text{O}_2 + \text{CO}_2$  have the EC number 1.2.3.4, which designates an oxidoreductase (class 1) that acts on the aldehyde or oxo groups of donors (subclass 1.2) with an oxygen as acceptor (subsubclass 1.2.3). The number 4 is designated for the fourth reaction in this class<sup>(25)</sup>. A strong advantage of EC numbers is that they provide unique identifiers for enzyme functions, no matter what confusing name is used for that enzyme in different organisms.

Currently, the most commonly used metabolic pathway databases are those listed in Table 2.

Table 2. Metabolic Pathway Databases and URL

Database name	URL
BBID	<a href="http://BBID.GRC.NIA.NIH.GOV">http://BBID.GRC.NIA.NIH.GOV</a>
BIND	<a href="http://WWW.BIND.CA">http://WWW.BIND.CA</a>
BioCarta	<a href="http://WWW.BIOCARTA.COM">http://WWW.BIOCARTA.COM</a>
BioCyc	<a href="http://WWW.BIOCYC.ORG">http://WWW.BIOCYC.ORG</a>
BRITE	<a href="http://WWW.GENOME.AD.JP/BRITE">http://WWW.GENOME.AD.JP/BRITE</a>
CSNDB	<a href="http://GEO.NIHS.GO.JP/CSNDB">http://GEO.NIHS.GO.JP/CSNDB</a>
EcoCyc/MetaCyc	<a href="http://ecocyc.pangeasystems.com/ecocyc/">http://ecocyc.pangeasystems.com/ecocyc/</a>
ExPASy-Biochemical Pathway	<a href="http://expasy.proteome.org.au/cgi-bin/search-biochem-index">http://expasy.proteome.org.au/cgi-bin/search-biochem-index</a>
GeneNet	<a href="http://WWW.MGS.BIONET.NSC.RU/MGS/SYSTEMS/GENENET/">http://WWW.MGS.BIONET.NSC.RU/MGS/SYSTEMS/GENENET/</a>
KEGG	<a href="http://www.genome.ad.jp/kegg/kegg.html">http://www.genome.ad.jp/kegg/kegg.html</a>
Metabolic Database	<a href="http://CGSC.BIOLOGY.YAHLE.EDU/METAB.HEML">http://CGSC.BIOLOGY.YAHLE.EDU/METAB.HEML</a>

Metabolic Pathways of Biochemistry	<a href="http://WWW.GWU.EDU/~MPB/">http://WWW.GWU.EDU/~MPB/</a>
PathDB	<a href="http://www.ncgr.org/software/pathdb/">http://www.ncgr.org/software/pathdb/</a>
UM-BBD	<a href="http://www.labmed.umn.edu/umbbd/index.html">http://www.labmed.umn.edu/umbbd/index.html</a>
SPAD	<a href="http://www.grt.kyushu-u.ac.jp/spad">http://www.grt.kyushu-u.ac.jp/spad</a>
WIT	<a href="http://wit.mcs.anl.gov/wit2/">http://wit.mcs.anl.gov/wit2/</a>

One major advantage of pathway databases over other biological databases is the possibility of providing several types of information in the context of the graphical representation of pathways. For example pathway database are able to represent the high complexity of all of biochemical reactions within a single cell or a complete organism. There are different ways to show a graphic view of a pathway.

### 3.1.3 KEGG

KEGG (Kyoto Encyclopedia of Gene and Genomes) contains all known metabolic pathways and a limited number of regulatory pathways and transport mechanisms<sup>(30)</sup>. The KEGG system consists of three main databases which are tightly connected: LIGAND, with information about compounds, enzymes and reactions stored in flat files<sup>(31)</sup>; PATHWAY, which contains the graphical representations of the pathways and lists of enzymes and reactions within the pathways; and GENES, which contains organism-related genome and gene information and lists of genes within an organism and pathway. Furthermore, KEGG provides many links to other databases that are integrated within the DBGET integrated database retrieval system<sup>(32)</sup>.

Pathways in KEGG are classified according to the chemical structures of their main compounds, e.g. carbohydrates, lipids, amino acids. All specific pathway diagrams and overviews are manually drawn pictures where pathway maps consist of links to specific information about compounds, enzymes and genes. There are all known reactions catalyzed by proteins/enzymes derived from gene products in the pathway maps. Reactions within the pathway maps do not represent side compounds, eg ATP (adenosine triphosphate) or NADH (reduced nicotinamide adenine dinucleotide). Also, there are links to some other related pathways in the pathway maps connected by their contributing compounds. This allows the users to get an overview about connections to other pathways<sup>(28)</sup>. From the user's point view, either graphical diagrams or hierarchical texts represent the KEGG's data.

Users can use EC numbers for enzymes to search the KEGG pathways, by compound numbers for chemical compounds, and by gene accessions for specific genes. The KEGG pathways can also be searched by sequence similarity. This is especially useful for identifying orthologues and reconstructing pathways from the gene catalogue. For instance, by taking the *E. coli* pathways as references, the user can check if a functional unit can be formed from the gene catalogue of a specific organism. Alternatively, the user can search against a KEGG orthologue table that contains a multiple alignment of gene orders in the pathway, as well as in the genome (operon), for predictions of ABC transporters<sup>(33)</sup>, bacterial two-component systems<sup>(34)</sup> and others.

The pathway query result shows all pathways containing a given enzyme or a given compound<sup>(35)</sup> but offer no graphical representation. Organism-specific information is also not available within such queries. Searching for a pathway by selecting a specific organism does not provide information about the enzymes available but only links to the gene information related to that organism. A comparison of pathways of two or more organisms cannot be implemented. Based on binary relations, two algorithms (Dijkstra/Floyd) are used to find the shortest pathway between two compounds. In such a case that the reaction has two or more substrates, the implementation is not able to distinguish between different substrates and their correlating products or compartments and transport mechanisms. The pathway created based on one main substrate pays no attention to side substrates or products, compounds/enzyme locations and cofactors. Results of the pathway creation contain only the compound ID and EC numbers.

From data model concept, KEGG may consist of three interconnected sections: pathways, genes, and molecules, which are also linked to a number of existing databases through DBGET (Fig. 5). In KEGG, binary relations, hierarchies, and pathways represent functional aspects of genes and molecules<sup>(65)</sup>.

Recently, KEGG announced the release of KEGG 23.0, and Kanchisa<sup>(36)</sup> et al. presented a detailed description of the database. The primary objective of KEGG is to computerize the current knowledge of molecular interactions, namely metabolic pathways, regulatory pathways and molecular assemblies. At the same time, KEGG maintains gene catalogues for all the organisms that have been sequenced and links each

gene product to a component on the pathway. It currently represents most of the known metabolic pathways and some of the known regulatory pathways in about 100 graphical diagrams and 60 orthologue group tables. The database is cross-linking with WIT (see below), an interacting metabolic reconstruction in the web, which provides a more detailed picture of the metabolic pathways. In contrast, KEGG attempts to cover a wider range of biochemical pathways at a higher level of abstraction. Matching the enzyme gene in the gene catalogue with enzymes on the reference pathway diagrams generates organism-specific pathways.

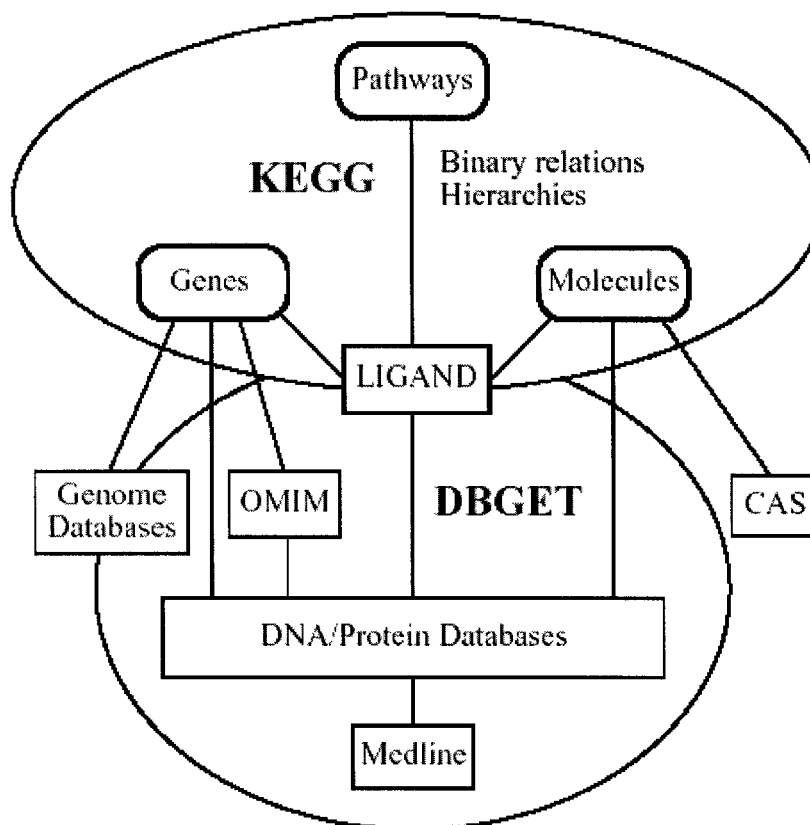


Figure 5. The data model concept of KEGG and its relation to DBGET<sup>(65)</sup>



There are two versions of KEGG. One is the Internet version, the other is the CD (local copy) version. A Web browser, such as Netscape Navigator or Microsoft Internet Explorer can browse both of them.

### 3.1.4 EcoCyc/MetaCyc

EcoCyc is a metabolic pathway database that describes the genome and the biochemistry of *Escherichia coli* based on information from EcoGene database<sup>(37)</sup>, SWISS-PROT and the scientific literature<sup>(38)</sup>. The database consists of all sequences and functional annotations of *E. coli* genes. Pathways of *E. coli* and its reactions and enzymes are annotated with references to the literature. The query interface of EcoCyc provides search options for *E. coli* genes, proteins, reactions, compounds and pathways by names, sub-string, classification hierarchy, EC number or chemical structure. The EcoCyc overview is a bird's-eye view of the complete *E. coli* biochemical pathways<sup>(39)</sup>.

The EcoCyc data are stored within a frame knowledge representation system(FRS) called Ocelot. FRSs use an object-oriented data model<sup>(66)</sup>. They organize information within classes. The EcoCyc schema is based on the class hierarchy shown in Figure 6. Each EcoCyc frame contains *slots* that describe attributes of the biological object that the frame represents, or that encode a relationship between that object and other objects.

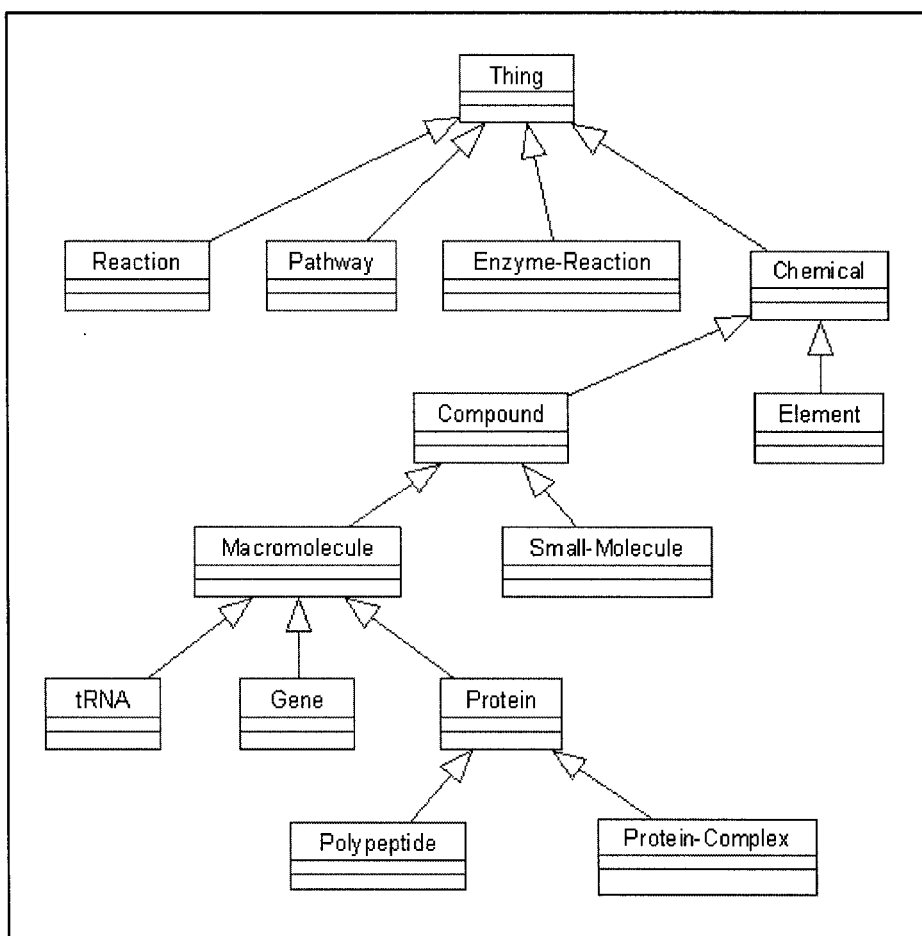


Figure 6. The top of the class hierarchy for EcoCyc Database<sup>(66)</sup>

Pathway diagrams are drawn automatically using graph-drawing algorithms. Also, its size may be adjusted. There are different levels of data information provided by the graphical representation of pathways. The lowest level contains only the major reactions of a given pathway. Information about enzymes and genes, etc. is not available. In the more detailed level, all enzymes with their corresponding EC numbers and corresponding genes are given. Some pathways show chemical structures of the reactants on that level, but this feature is not consistent. Some pathways also include green arrows that indicate regulatory mechanisms (solid lines). In the graphical representation of pathways, the

inhibitors and activators of reactions are represented, but detailed compound information is not accessible. More data about the enzyme are given by connecting to the ExPASy-ENZYME database. It also includes reactions that can be catalyzed by more than one enzyme. A linkage of the enzyme to the related gene and vice versa is implemented<sup>(28)</sup>.

In addition, MetaCyc describes pathways, reactions and enzymes of a variety of organisms, with a microbial focus. But MetaCyc does not contain organism-specific genome or protein information such as genomic maps or sequences. MetaCyc uses the same database schema and visualization software as EcoCyc<sup>(40)</sup>. In contrast to EcoCyc, the MetaCyc query interface only offers searching for pathways, reactions and compounds. Proteins or genes can be selected for querying but no matches will be found by searching. MetaCyc offers no general pathway overview.

### **3.1.5 WIT (What Is There?)**

The WIT system<sup>(41)</sup> connects data about genes and genomes, enzymes, reactions and pathways. The Enzyme and Metabolic Pathways (EMP) database is currently embedded within WIT. It offers a general functional overview as one possible starting point represented as a classification table of pathways. The ‘View Annotation’ window gives names, EC number and functional description of an enzyme. Graphical representations contain comments and information on the compartments where the reaction occurs and links to information about the enzyme. The descriptive page of enzyme is linked to KEGG, EMP and Medline. The link to the EMP database allows the retrieval of literature-related and organism-specific enzyme information. The ‘Diagram

Data' window contains detailed information on the overall balance of the reaction, every compound involved in the reaction and the location or compartment of the cell where the reaction occurs. These data are used for the graphical representation of pathways that are represented in the "Diagram Picture' view. The 'Assertions Table' contains information about existing enzymes of a given pathway in complete sequenced organisms. WIT system has no graphic representation for all pathways overview<sup>(28)</sup>.

## **3.2 Biochemical Pathway Visualization**

In this section, I will discuss basic requirements about metabolic pathway visualization and then review some tools for pathway visualization since visualization tools is an important requirement for the interpretation of pathway data.

### **3.2.1 Biochemical Pathway Requirement Analysis**

The physiological functions of an organism are accomplished through the coordinated regulation of the expression of a large of number of genes. The functional elements of a gene network and pathway may have the following biology conceptual requirements<sup>(42)</sup>:

- A gene ensemble interacting when certain biological functions are performed.
- Protein expressed by these genes. To ensure the performance of an appropriate function, the protein can be modified (phosphorylated or glycosylated), or can form different complexes<sup>(43-44)</sup>

- Biochemical pathways providing gene activation in response to an external stimulus
- A set of positive and negative feedback stabilizing the parameters of the pathway or providing a transition to a new functional state<sup>(45)</sup>.
- External signal, hormone and metabolites that trigger the pathway or correct its operation in response to the changes in physiological parameters.

### 3.2.2 Representation Levels of Pathway

Compartmentalization is a characteristic feature of the processes occurring in the real biological gene network. The components of the gene network are scattered through organs, tissues, cells, and cell compartments. So, in the description of the pathway three hierarchical levels<sup>(42)</sup> are considered in the representation of pathway.

1. Organism level: organs, tissue, and particular types of cell
2. Cell level: compartment locations in single-cell
  - intercellular space
  - cell membrane
  - cytoplasm
  - nucleus
3. Gene level: the regulation of gene transcription

### 3.2.3 Construction of Pathway Diagram

Metabolic pathways can be modeled as directed graphs, which are a collection of interconnected biochemical reactions. The diagram of the pathways is a graph with nodes corresponding to entities and arrows representing relationships between the pathway components. It typically consists of nodes and arrows (links). Main reactants and products (metabolites) are represented as nodes and the reactions as arrow edges of the graph. Usually the enzymes catalyzing the reaction are displayed as edge labels. Side substrates (low-molecular weight compounds) are drawn near the arrow edge, connected to the edge by curved arcs. The arrows represent interactions between the nodes. You may see my conceptual data model generate proline pathway (see section 5.3). In real pathway, those nodes also have hypertext links between the diagram and molecular data. Thus, the problem of visualizing pathways can be formulated as a graph layout problem.

There are two major ways to show a graphical view of a pathway. One is graphical representation of the pathways that are drawn manually, which is visualized in a static and non-dynamic way (like KEGG, WIT see section 3.1). Pathway diagrams are manually drawn and stored as bitmap image files. These diagrams are displayed as interactive image maps with links to additional information on enzymes and to adjacent pathways. The others are produced automatically on demand and are user-dependent (PathDB, EcoCyc see section 3.1).

For a user-dependent interaction, flexible representations will be more and more important because they enable different levels of information to be represented. To avoid information overload, but on the other hand to be able to differentiate between general overviews and detailed representation of reactions or pathways containing structural formulas of chemical compounds, users should be able to zoom into their specific level of interest. A real challenge of pathways representation is to combine a flexible drawing of pathways, with the content behind all the elements within that given pathway without losing clarity.

### **3.2.4 Pathway Visualization Tools**

#### **3.2.4.1 GeneMAPP**

GeneMAPP (Gene MicroArray Pathway Profiler)<sup>(46)</sup> is a new computer software that is helping genomic researchers to make sense of the reams of data—a massive collection of numbers and decimals—that result from using DNA microarrays (see Chapter 2.3). It is one kind of program that displays the gene expression data in the context of known biological pathways. So, scientists can see how their results fit in with real life data examples. The flood of sequences from various genome-sequencing projects has paved the way for large-scale experiments to study gene expression. Just one experiment can yield information from thousands of genes. GeneMAPP organizes the results by biological process, allowing researchers to see coordinated changes in gene expression that would be difficult to see when looking at all the data at once<sup>(47)</sup>.

#### 3.2.4.2 The Pathway Tools

P. Karp<sup>(48)</sup> et al developed the pathway tools, which is a software environment for creating a type of model-organism database (MOD) called a Pathway/Genome Database(PDGB). These tools integrate information about the genes, proteins, metabolic pathway, and genetic network of an organism. The pathway tools in this software can provide two different modalities for interacting with a PGDB. First, it provides a graphical environment that allows users to visualize the contents of a PGDB and to interactively update a PGDB; and then, it provides a sophisticated ontology and database API that allow programs to perform complex queries, symbolic computations, and data mining on the contents of PGDBB.

#### 3.2.4.3 JDIP, a protein network visualization tool

JDIP at first provides the means for a fast, visual evaluation of protein's interaction environment, represented as a static graph. Now, it is developed as a tool providing means of protein-network oriented data retrieval, visualization and analysis<sup>(49)</sup>, which is a stand-alone Java application that provides a generic framework for integration of heterogeneous data from other biological databases. It is a graphic display of protein interaction network such as protein expression level, focused on any given protein contained in DIP<sup>(50)</sup>, the database of interacting proteins. JDIP tool is an applet available within the web interface and an independent, cross-platform java application.



In addition, after installation as a browser ‘helper’ application, JDIP can be fully integrated with DIP web interface by providing two-directional queries between the DIP server and JDIP. XML files using JIN (Java Interaction Network) syntax specified as XML schema <sup>(51)</sup> fulfill the data exchanged between the DIP database and JDIP.

### **3.3 Microarray Gene Expression Data Standard and MAGE-ML**

Microarray technology is a high-throughput functional-genomics method for obtaining gene expression data from thousands of genes simultaneously<sup>(21)</sup>, allowing biologists potentially to study the transcription of an entire set of genes for a species. Data from various kinds of experiments that have a bearing on pathway are being created at an increasing rate. The large ensemble of information they produce contains patterns that are reflection of pathway dynamics, and therefore can be used to deduce pathway causal structures. Microarray gene expression data is a source of pathway ‘causal’ information<sup>(38)</sup>, which promise to provide a key insight into gene function and interactions within and across metabolic pathway. To help illustrate these functional relationships, researchers are applying a wide range of approaches to analyzing microarray. There are lots of computational approaches involved, but there are serious challenges emerging to these approaches. Microarray data is being produced by many independent organizations, is defined and described in a variety of ways, and is being stored and displayed in multiple locations using a variety of technologies. In addition, microarrays do not measure gene expression levels in any objective units. Data communication is one of most significant challenge microarray present.

Microarray data requires data structures that are both multidimensional and varied, and no natural or standard ways to move results between research groups yet exist. This applies to both underlying gene expression data and the descriptive biological annotations that provide context for the gene expression measurements<sup>(52)</sup>. Recently, the microarray gene expression data group<sup>(53)</sup> (MGED) has published a specification describing MIAME, the minimal information for the annotation of a microarray experiment.<sup>(54)</sup>

In this section, I will briefly introduce and discuss MIAME and MAGE-ML as well as ArrayExpress that is the database system which implemented the MAGE-OM object model. Clearly understanding them will be helpful for the design of our own microarray data management system in future.

### **3.3.1 Minimum Information about a Microarray Experiment – MIAME**

#### **3.3.1.1 Introduction**

Although every experiment may be different, MIAME aims to define the core that is common to most microarray experiments. MIAME is not a formal specification, but a set of guidelines<sup>(54)</sup>.

A major objective of MIAME is to guide the development of microarray databases and data management software. A standard microarray data model and exchange format MAGE,<sup>(53)</sup> which is able to capture information specified by MIAME,

has been submitted by European Bioinformatics Institute (EBI) for MGED and Rosetta Biosoftware and recently became an adopted specification of the OMG standards group<sup>(25)</sup>. Many organizations, including Agilent, Affymetrix, and Iobion, have contributed ideas to MAGE. Links to software tools supporting the MIAME information capture and management are available.

MIAME tries to define the minimum information that must be reported, to ensure the interpretability of the experimental results generated using microarrays as well as their potential independent verification. Although MIAME concentrates on the content of the information and should not be confused with a data format, it also tries to provide a conceptual structure for microarray experiment descriptions. It is focused on microarray-based gene expression data, which will facilitate the establishment and usefulness of microarray databases<sup>(54)</sup>.

MIAME is platform-independent but includes essential evidence about how the gene expression level measurements have been obtained (see appendix B). It should be noted that MIAME does not specify the format in which the information should be provided, but only its content.

### 3.3.1.2 Gene expression conceptual model

Collections of gene expression data can be abstractly viewed as a table with rows representing genes, columns representing various samples and each position in the table describing the measurement for a particular gene in a particular sample, which is called a

gene expression matrix in MIAME. The information described for a microarray experiment can be conceptually divided into three logical parts: gene annotation, sample annotation and a gene expression matrix. In addition, not only the final gene expression matrix needs to be recorded, but also a detailed description of how the expression values obtained is necessary. So, if the data verification is to be ensured, the nature of the data recorded may become more complex.

Sample annotations recorded the information about the context of the particular biological sample and the exact conditions under which the samples were taken. Gene annotation should provide a full and detailed description of each gene element on the array.

There are three levels of data relevant to a microarray experiment:

1. The scanned images (raw data)
2. The quantitative outputs from the image analysis procedure (microarray quantitation matrices)
3. The derived measurements (gene expression data matrix)

#### 3.3.1.3 MIAME Experiment design

Brazma et al<sup>(54)</sup> propose the data and annotations from microarray experiment should meet the following requirements:

1. The recorded information about each experiment should be sufficient to interpret the experiment and should be detailed enough to enable comparisons to similar experiments and permit replication of experiments.
2. The information should be structured in a way that enables useful querying as well as automated data analysis and mining.

The first requirement may imply that a detailed annotation of the sample and other experimental conditions should be recorded and its reliability should be given. The second one may imply some necessities for controlled vocabularies or ontologies to represent data as well as the need to limit free-format text only.

#### 3.3.1.4 MIAME components

MIAME defines six main components (see Figure 7) for minimum information about a published microarray-based gene expression experiment. They are as follows:

1. *Experimental design*: the set of hybridization experiments as a whole
2. *Array design*: each array used and each element (spot, feature) on the array
3. *Samples (targets)*: samples used, extract preparation and labeling
4. *Hybridization*: procedures and parameters
5. *Normalization control*: types, values and specifications
6. *Measurements*: images, quantification and specifications

Each of these components contains information provided using controlled vocabularies or free-text format.

#### 3.3.1.4.1 *Experimental design*

The minimal information in this part is as follows:

- Experiment type: normal-versus-diseased comparison, time course, dose response and so on.
- Experimental variable: parameters or conditions tested (time, dose, genetic variation or response to a treatment or compound)
- General quality-related indicators: usage and types of replicates
- Quality-control steps: nonspecific hybridization
- Experimental relationship to: array and sample entities (which samples and arrays were used in each hybridization assay)

#### 3.3.1.4.2 *Array design*

This entity will provide information about a systematic definition of all arrays used in the experiment, including the genes represented and their physical layout on the array. There are two parts in the entity as follows:

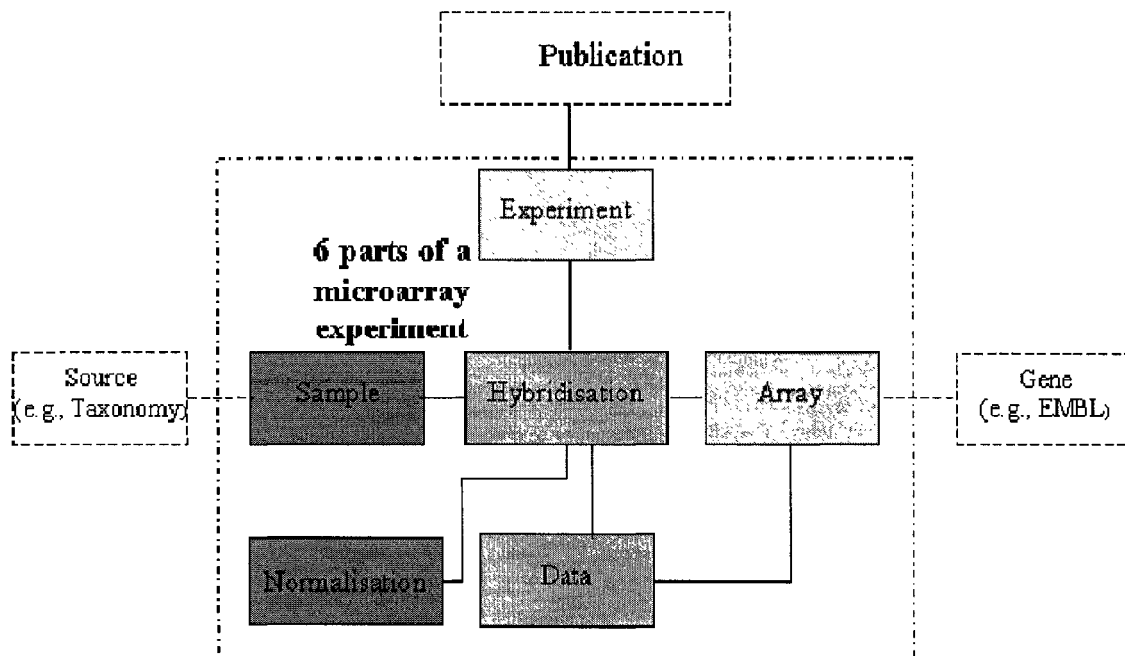


Figure 7. Six MIAME components<sup>(54)</sup>

1. A list of the physical array: unique ID and a simple description
2. Array-type definition:
  - A description of array as a whole: platform type, provider, and surface type
  - A description of each type of element or spot used: synthesized oligonucleotides or PCR products from cDNA clones
  - A description of the specific properties of each element: DNA sequence or quality control indicators

#### 3.3.1.4.3 *Samples*

*Sample* entity represents the biological materials for which the gene expression profile is being established. It describes the source of the original sample (such as organism taxonomy and cell type) and any biological in vivo or in vitro treatments applied, the technical extraction of the nucleic acids and their subsequent labeling.

#### 3.3.1.4.4 *Hybridization*

*Hybridization* entity defines the laboratory conditions under which hybridization was carried out, other than a free-text description of hybridization protocol. The critical hybridization parameters are specified as follows:

- Choice of hybridization solution
- Nature of the blocking agent
- Wash procedure
- Quantity of labeled target used
- Hybridization time, volume, and temperature
- Description of hybridization equipment

#### 3.3.1.4.5 *Normalization controls*



The aim of microarray experiment is to identify relative changes in expression levels, identify differentially expressed genes, and identify classes of genes or samples with similar patterns of expression after analysis of the data from multiple samples. Hybridization intensity derived from image processing must first be normalized to get a comparison. MIAME proposes the standard that include four parts as follows:

1. The normalization strategy: spiking, housekeeping genes, total array, other approaches
2. The normalization and quality control algorithm used
3. The identities and location of the array elements serving as controls as well as their type
4. Hybridization extract preparation

#### 3.3.1.4.6 *Measurements*

It defines the actual experimental results processing from raw to processed data, which consists of three parts as follows:

1. The original scans of the array (images)
2. The microarray quantification matrices based on image analysis
3. The final gene expression matrix

Image data should be given as raw scanner files, accompanied by scanning information that includes relevant scan parameters and laboratory protocols. Storing the primary image files may need a significant quantity of disk space. For each experimental image, a microarray quantification matrix contains the complete image analysis output as directly generated by the image analysis software. This is a 2D matrix for given image files, where array elements (spots or features) constitute one dimension and quantification types (such as mean and median intensity, mean or median background intensity) are the second dimensions.

The gene expression matrix consists of sets of gene expression levels for each sample. The microarray quantification matrices can be considered as spot/image centric and the gene expression matrix is gene/sample centric.

### **3.3.2 MAGE-ML**

MAGE-ML stands for MicroArray and Gene Expression Markup Language. Briefly, MAGE-ML is a language designed to describe and communicate information about microarray-based experiments. MAGE-ML is based on XML and can describe microarray design, microarray manufacturing information, microarray experiment set up and execution information, gene expression data and data analysis results. In fact, MAGE-ML is XML representation of MAGE-Object Model. As MAGE-ML data can be expressed in XML, it is both human-readable and machine-readable with their relationships in a pre-defined DTD. Full MAGE specification can be found at

(<http://cgi.omg.org/cgi-bin/doc?lifesci/01-10-01>). MAGE-ML Document Type Definition (DTD) is at (<http://cgi.omg.org/cgi-bin/doc?lifesci/01-11-02>).

There is a detailed introduction to MAGE-ML and MAGE-OM in the Appendix C of the thesis, in which MAGE-OM is expressed in UML. A few rules were used to translate MAGE-OM into the DTD named MAGE-ML as follows:

1. Each class in the object model is represented as an element with an attribute list matching the attributes of the class
2. For each association of that class, a daughter element having the role's name with 'assn' appended
3. If the association is by reference, 'ref' is appended and if the cardinality of the association is greater than one 'list' is appended

So, MAGE-ML is predictable and the future addition and extensions to MAGE-ML will be compatible. It is a standard format for exchanging data among microarray databases and data analysis tools.

To make it as easy as possible for researchers to be MIAME compliant, Microarray Gene Expression Data Society (MGED) has developed a mark-up language, MAGE-ML, for communicating MIAME-compliant data. MAGE-ML makes it easier to transfer MIAME-compliant data between microarray applications. In addition, EBI

(European Bioinformatics Institute) has developed ArrayExpress<sup>(57)</sup> as a database schema to allow MIAME/MAGE-ML – compliant data to be stored in database repositories.

MAGE-ML is set to become the *de facto* standard for exchanging microarray data between applications. If data not either in MAGE-ML's format or capable of being converted to MAGE-ML, it will prove very difficult to analyze or combine with other forms of data. Some of software packages, such as MaxD<sup>(55)</sup> or GeneX<sup>(56)</sup>, are now available that allow MIAME-compliant data to be stored in database.

### **3.3.3 ArrayExpress, A Public Repository for Gene Expression Data**

ArrayExpress <sup>(57)</sup> is an international public repository for microarray gene expression, whose data are based at the EBI. It aims to store and provide access to well-annotated data from microarray experiments. It is an Oracle implementation of the MAGE-OM object model (see appendix C).

There are four major components in ArrayExpress:

1. The database itself
2. A web-based query interface
3. A data submission and annotation tool called MIAMExpress
4. An online data analysis tool called Expression Profile

MIAMExpress is the EBI's MIAME compliant web-based annotation tool and data submission tool for ArrayExpress. MIAMExpress is a generic array-platform independent tool that allows users to submit experiments, array descriptions and protocols. The tool is based on the MIAME questionnaire and is used for bench biologists who are wishing to submit data. Data is stored in a MySQL database during the submission process and is parsed to MAGE-ML after curation by the ArrayExpress database staff prior to loading into ArrayExpress. MIAMExpress uses MGED ontology terms in order to limit free text within the submissions, this might speed up submissions and is designed to allow automated data mining. The definitions for the terms can be found in the MGED ontology, developed by the MGED ontology-working group and full contextual help is provided within the tool<sup>(57)</sup>.

ArrayExpress has publicly available data sets loaded that can be queried, and is accepting submissions in the MAGE-OM derived MAGE-ML data exchange format and via MIAMExpress. The ArrayExpress staff are establishing MAGE-ML pipelines with major microarray producers and experimenters, including the Sanger Institute, Affymetrix, TIGR and MIMR. Data can be exported from ArrayExpress to Expression Profiler, an integrated set of web-based tools for the analysis and visualization of functional genomics data, loosely with powerful data selection and filtering mechanisms, and numerous clustering and pattern discovery algorithms (hierarchical, K-means).

ArrayExpress supports the microarray community standards MIAME and MAGE-ML. Its data submissions are divided into three parts: Experiment, Protocol and Array.

Each is given an accession number so that an Array or Protocol can be referenced by many Experiments. ArrayExpress can be downloaded freely to analyze publicly available microarray data locally, or to store your own experimental data.

All the data are stored in the central ArrayExpress database, from which they can be accessed using a web-based query tool. The data can be imported directly into Expression Profiler for analysis, or you can export data to analyze them locally using other your own tools.

## **Chapter 4. Conceptual Data Models of Metabolic Pathway**

To understand the molecular logic of cells, we must be able to analyze metabolic processes in qualitative and quantitative terms. Therefore, data modeling is one of the most important methods. In this chapter, I will develop and present a conceptual data model for biochemical pathway in my understanding.

### **4.1 Developing a Data Model for Biochemical Pathways**

Our data model is based on the idea that there exist a relatively small number of fundamentally important classes of biological objects that tend to resist major schematics variation: e.g. genes, proteins, enzymes, pathways, and reactions. Along with their core attributes, these classes form a common denominator for biological databases, with differences among schemas tending to occur in less essential details.

### **4.2 The Possible Scopes and Constraints in our Conceptual Data Model**

There is one assumption for our conceptual data models. As mentioned before, our data model is a conceptual data model (section 2.2). This is one dimension in our database design. Another dimension is what is the system architecture in our database design. Is it the distributed database system? No. So, here I will assume that our database management system is the traditional standard system architecture that is the three-schema ANSI/SPARC (American National Standards Institute/Standards Planning and

Requirements Committee)<sup>(58, 59)</sup>. ANSI/SPARC divided database-centric systems into three models: the internal, conceptual, and external.

Since our data models belong to conceptual data model, the scope of it may be hard to be given at this stage. So, I just give the possible scope and constraints for our pathway element conceptual data model (see Figure 8).

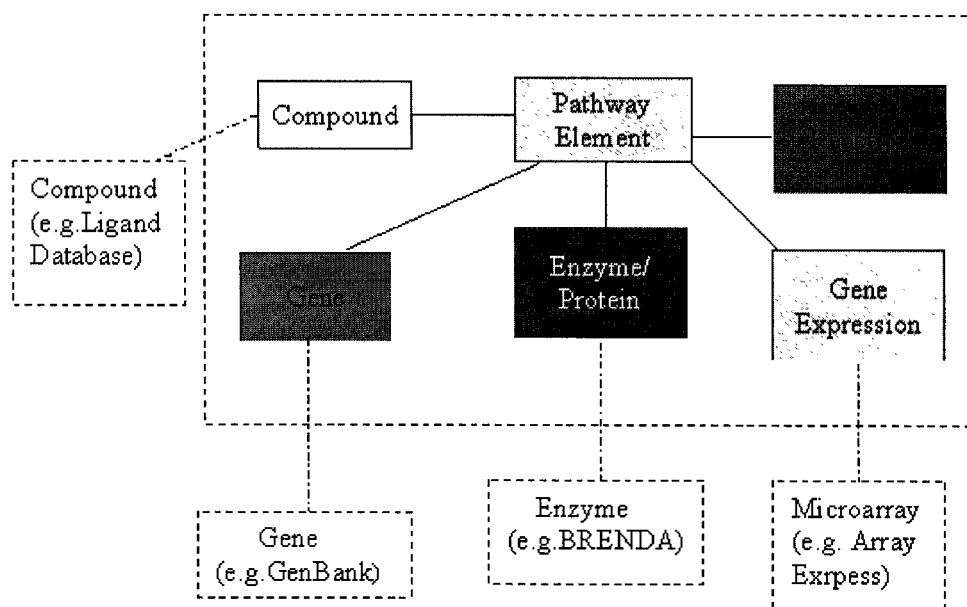


Figure 8. Pathway Data Model Scope

#### 4.2.1 Compound Object

This is one big object where there are two sub-objects under it, *metabolite* and *Cofactor*. These two objects are specifically involved in metabolic pathway. In *compound*



object, it may contain some compound ID outside the databases for user to link to other compound databases(see figure 8). The core attributes are listed in the data model.

#### **4.2.2 Enzyme Object**

It contains enzyme information in the metabolic pathway. Some enzymes that are not involved in metabolic pathway will not be encompassed in this object. Each enzyme can involve more than one chemical reaction. In this object, it may contain some enzyme ID outside our system that hyperlinks to other enzyme databases (figure 8).

#### **4.2.3 Reaction Object**

This object defines chemical reaction that involves in cell metabolic pathway. Typically, each reaction is constrained by EC number. It has a one-to-one relationship with EC number.

#### **4.2.4 Gene Object**

This object defines the gene in cells that has a direct connection in the metabolic pathway. Some genes that involve other cell process but not in metabolic pathway can not be described with this object, for instance, actine gene. This means that gene in *gene* object is specifically defined for metabolic pathway. However, some genes in this object may have other databases gene ID for instance, GenBank<sup>(60)</sup> (see Figure 8).

#### 4.2.5 *GeneExpression* Object

Here, *GeneExpression* object is specifically defined for the data from microarray gene expression experiment. We do not consider other gene expression data from non-microarray gene expression experiment such as SAGE (see 3.3.1). From this object, we can do microarray gene expression data analysis to get some information about metabolic pathways in organisms. As we know, MIAME (see chapter 3.3) has become one accepted microarray gene expression data standard, this object may have a link to specific microarray gene expression databases (see figure 8).

### 4.3 UML Model for Metabolic Pathway Components

The UML model provided in Figure 9 is used to describe metabolic pathway element information. As shown in Figure 9 under the *PathwayElement* class, two main classes of objects are defined. The first class, *BiochemicalEntity*, represents structural units. Here, I try to classify biochemical entity into compound, protein, enzyme, and gene. So, these can be defined as objects like *compound*, *protein*, *gene*, and *enzyme*. These objects have core attributes, which describe their physical characteristics (chemical formula, structural formula, molecular weight, sequence, gene and so on). In pathway visualization, it could be displayed as nodes. The second class, *Interaction*, represents interactions between those biochemical entities. There are two kinds of interaction that I can classify them, which are *transformation* and *regulation*. The *interactions* objects considered here are not simple links between entities as in many databases, but are objects in their own rights.

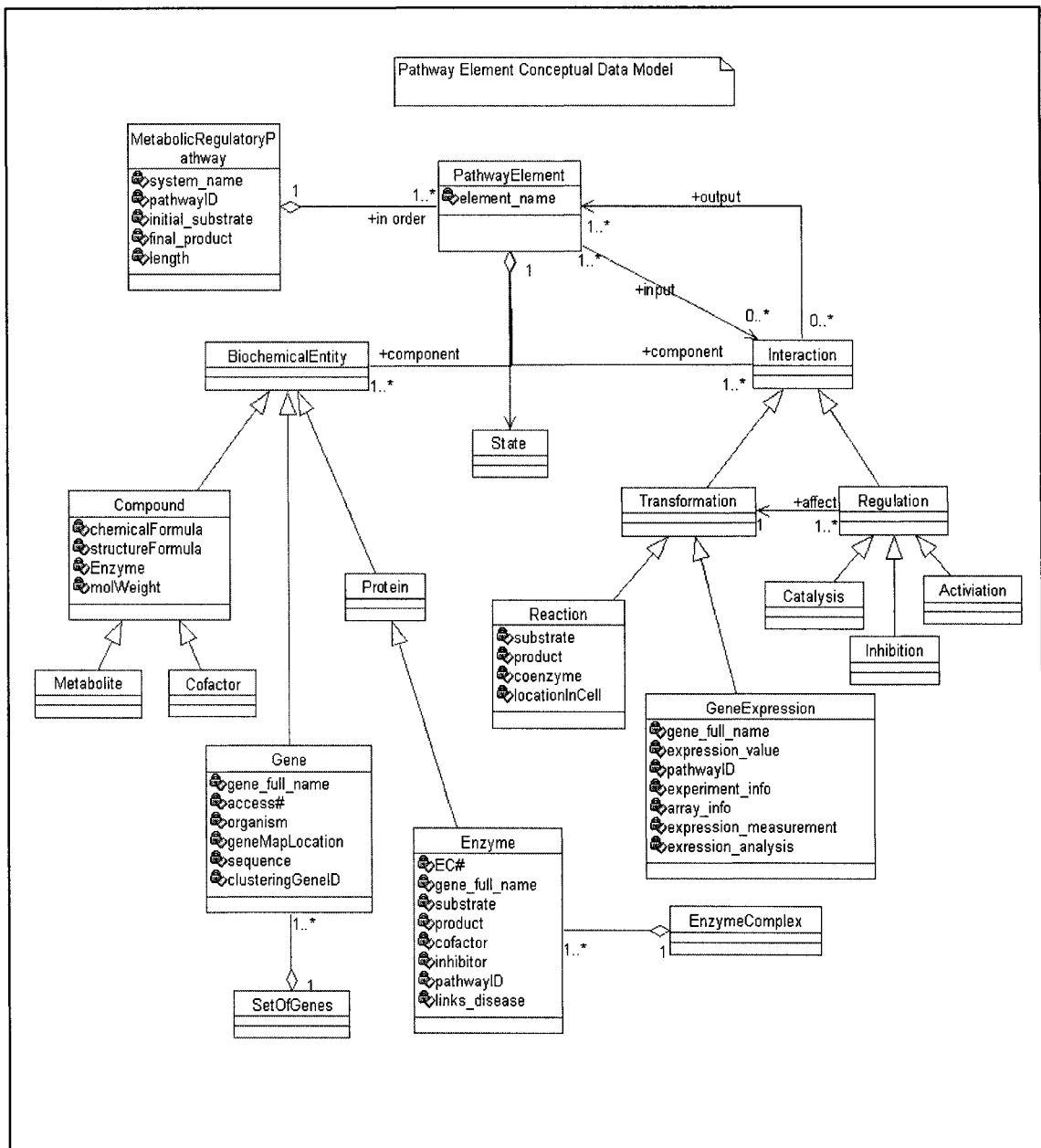


Figure 9. Pathway element conceptual data model

A list of inputs and a list of outputs characterize each interaction object. Here I make some kind of abstraction. That the metabolites in pathway are transformed into others can be considered as a list of inputs that are transformed into a list of outputs

through *transformation* object. In addition, it may have a rich collection of attributes, which describe the properties of the interaction. An interaction can be a *Reaction*, which converts a set of substrates (the input) into a set of products (the output). Another example is *Expression*, which has a gene as input and an enzyme as output.

The *interaction* (in Figure 9) object has the fact that both their inputs and outputs are sets of entities. Also, our pathway database should describe all the intervening steps for metabolic pathway. It can be represented as using *MetabolicRegulatoryPathway*. Each such object refers to the entities and interactions of all its intervening steps. Not only have the *Interaction* objects entities as input/output, but also can have other interactions as output. This is exactly the case for the *Inhibition* and *Activation* in Figure 9. This is the case for the object *Catalysis* that represents the action of *Enzyme* in accelerating a chemical reaction under *Regulatory* object. *Regulatory* object is the effect of an entity (it may be regarded as a catalyst or inhibitor) on certain reaction.

Since gene expression data is a source of pathway “casual” information<sup>(38)</sup>, we may use microarray gene expression data to analyze metabolic pathways in organisms. Our conceptual data model should cover gene expression in metabolic pathway components. Therefore, there is another subclass *GeneExpression* under the *Interaction* class. If in future we would like to know the detailed transcriptional regulation information, we can still extend this object under *Interaction* class. Finally, I define one class called as *state* to represent time or process dependent pathway situation to model a

dynamic pathway. In this data model, some core attributes are listed. This is still a conceptual data model, whose design process is iterative to validate attributes.

#### 4.4 UML Model for Biochemical Pathway Classification

The structure of biochemical pathway classification is important at least for two reasons. First, we can think of the classification as providing definitions of biological terms. Second, the classification of biochemical pathway is important because it influences the ease with which users can query the database system. As user queries often refer to the class hierarchy. Our biochemical pathway classification is such a kind of classification to guide our users to do query.

The UML diagram is given in Figure 10. The model includes primary, secondary and tertiary structure information. The topmost class in Figure 10 is *BiochemicalPathway*, of which all other classes are either directly or indirectly components. All the relationships between classes in Figure 10 are either aggregation or generalization relationships.

A *BiochemicalPathway* object has three subclasses that are *SignalTransduction*, *GeneRegulatory*, *MetabolicRegulatory*, which represent three major kinds of biochemical pathways available at present in molecular biology<sup>(56)</sup>. I will mainly focus on modeling metabolic pathway. In *MetabolicRegulatory* object, there are five properties to define this class, which are as follow:

- \* Sys\_pathway\_name: to define the organism, tissue or cell name

- \* PathwayID: to identify the pathway in the system
- \* Initial\_substrate: to define the first component in the pathway
- \* Final\_product: to define the final component in the pathway
- \* Length: to give the component number in the pathway

Metabolic regulatory pathways typically consist of four kinds of metabolic pathway<sup>(40)</sup>. So, a *MetabolicRegulatory* are composed by four subclasses. They are *Biosynthesis*, *Degradation*, *Energy*, and *OtherIntermediaryMetablism*. They are actually an abstract class, which should have been depicted in the diagram. An abstract class is one for which no direct instance objects are ever created, but which can play a useful organizational role in the diagram.

There are five sub-classes under *Biosynthesis* and four sub-classes under *Degradation* classes. Since we know there are five kinds of biosynthesis, they are classified into amino acid, carbohydrate, cell structure, fatty acid and lipids, and nucleotide. So, we define five subclasses as *Aminoacid*, *Carbohydrate*, *CellStructure*, *FattyAcidLipid*, and *Nucleotides* under a Biosynthesis class. It is quite similar that there are *Aminoacid*, *Carbolism*, *FattyAcid*, and *OtherDegradation* four subclasses under a *Degradation* class.

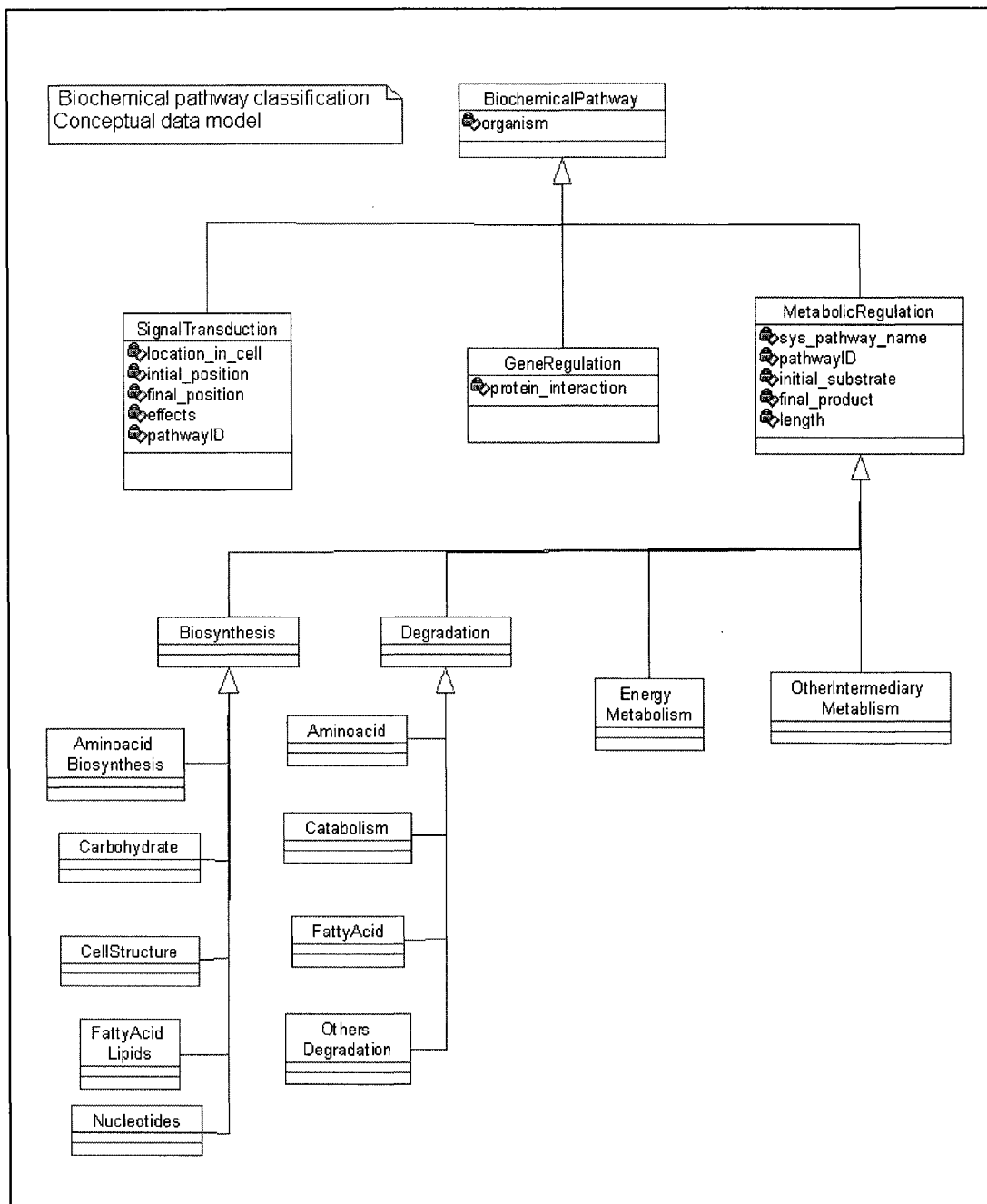


Figure 10. Biochemical pathway classification conceptual data model

## **Chapter 5. Discussion and Conclusion**

The modeling process is just the start of design. Once we have a conceptual data model, the next step is to relate the model back to needs, then move to forward to adding the structures that support both reuse and system function. In this chapter, I will briefly discuss our future work about our data models and give some conclusions about the thesis.

### **5.1 The Construction of Data Model is an Iterative Process**

During a conceptual data model construction, some issues on data model need to be clarified by revisiting the tasks to support or the sources of information to be described. In the design process, the role of conceptual data model is to allow precise statements to be made about the data of interest in a manner that can be communicated to others. The extensibility of a conceptual data model is very important, as it is used both in discussions with experts or scientists whose understanding of the relevant data is to be described, and by the developers of software that makes use of the data. Elmasri and Navathe<sup>(13)</sup> give a more detailed description on the design process that involves a combination of a design process and a modeling language.




The conceptual data models for biochemical pathway are built up in the thesis. I go back to double check if the data model describe the data identified in requirement analysis(see next section). Also, we still need to identify the needs of application and



sources of information that the modeling activity seeks to support are completely met and described in the conceptual data model. In section 5.2, I use the conceptual data model to describe the Figure 2 to see if that UML data model can model Proline Biosynthesis Pathway or not. It is displayed in the Figure 11.

## 5.2 The UML Data Model Describes the Data in Biochemical Pathway

In the Figure 11, it uses the data in Figure 2 to check if the UML conceptual data model can describe the proline synthesis pathway or not. We can see the following objects clearly in the data model.

1. *BiochemicalEntity*:  *Metabolite, Cofactor, Enzyme, and Gene*
2. *Transformation*:  *Reaction and GeneExpression*
3. *Regulation*:  *Catalysis and Inhibition*

As I mentioned in Figure 9, *metabolite* and *cofactor* objects are used as input and output for *interaction* object. When they are modeled as list of input for *reaction* object, they are the substrates of *enzyme* object. They will be the products of the *reaction* when they are modeled as output of *reaction* object. For instance, in Figure 9, Glutamate and ATP are the instances *BiochemicalEntity* object, as a list of input of *Reaction*. EC 2.7.2.11 is an instance of *reaction* object. The reaction is catalyzed by Enzyme, ?-glutamyl kinase which is an instance of *Enzyme* object. The products of that reaction are a list of output of the *Reaction* 2.7.2.11, which are ?-glutamyl phosphate and ADP, respectively, an instance of *Metabolite* and *Cofactor*. Therefore, the UML data model can describe the data in Proline biosynthesis pathway very well.

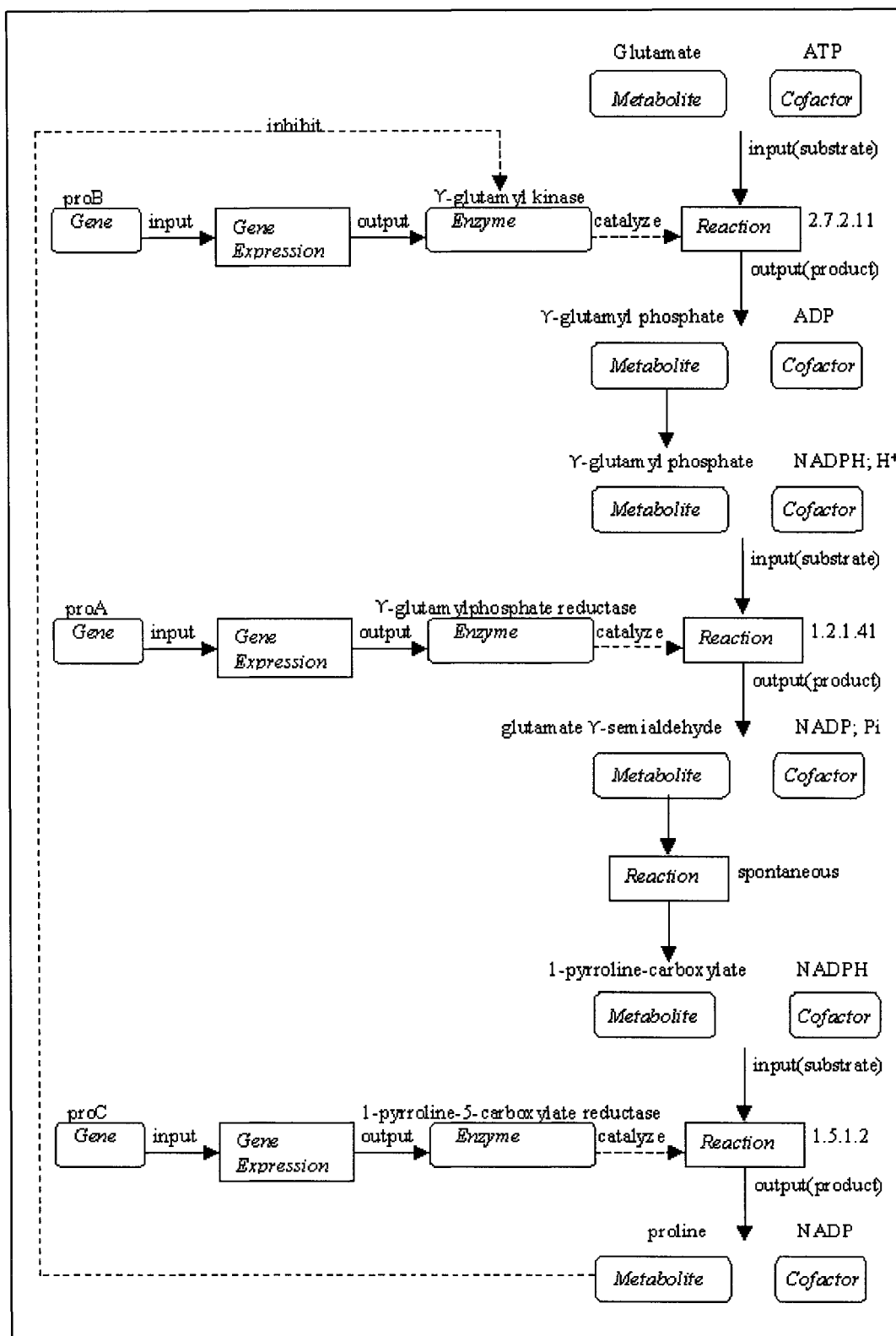


Figure 11. Proline biosynthesis pathway are displayed with UML conceptual data model

On the other hand, the biochemical pathway data models I designed and proposed in Chapter 4 can be used as the basis for an implementation using an object database. In fact, UML models are independent of the implementation platform to be used. In practice, mapping UML models, including class diagrams, onto object-oriented implementation platforms is more straightforward and intuitive than mapping onto non-object-based platforms. However, this is not to say that it cannot apply a relational data storage engine for an object-oriented data model. It also has the similar situation of implementing from ER model to relational database system. Blaha and Premerlani<sup>(61)</sup> provide a comprehensive description of how to map class diagrams onto relational tables. This process is along the similar lines as the implementing process for ER models, but no key in UML models and the tendency for inheritance to be used more widely in object models, often makes the mapping process more involved.

In addition, class diagrams are not targeted at any specific category of application. Mapping of these diagrams onto implementation platforms can be less direct or systematic than in the narrower context within which ER is used, but it is often straightforward to map class diagrams onto object-oriented implementation platforms. As mentioned above, the conceptual data model can be implemented into relational database management system, but the necessary overhead for the assembly/disassembly process for objects and the separation of data and functions should be taken into consideration. This may limit the usage of this approach.

### 5.3 The UML Data Model Can Generate Pathway Diagram

Most of metabolic pathway databases have interactive diagrams that are drawn manually, which is in a static way to visualize the pathway (mentioned in Chapter 3.2.3). Automated construction of diagrams from formalized information appears to be a promising direction, which the visualization process is performed dynamically at runtime based on the information provided by the database. EcoCyc was the first convincing demonstration of the efficiency of automated generation of diagrams for metabolic pathway<sup>(62)</sup>.

Static visualization has many server disadvantages. Whenever the data has been updated, the corresponding images have to be edited manually to reflect the changes. Furthermore, there is no way to specify the amount of detail to be displayed or to hide parts of the pathway. Last, when it comes to visualizing user defined or novel pathways, static visualization is not applicable at all<sup>(63)</sup>.

Automated construction of diagrams is based on object-oriented data model. The UML conceptual data model in the thesis can display biochemical pathway as directed graphs(see Figure 12). It is illustrated that Proline biosynthesis pathway can be visualized as directed graphs by using the conceptual data model very well. The problem of dynamically drawing a pathway is a graph layout problem. The UML conceptual data model in the thesis (Chapter 4.2) can be used to develop such a dynamic visualization pathway application or pathway drawing editor. Given as input a combinatorial

description of graph, a graph layout algorithm should compute geometric positions for the graph elements according to a set of rules.

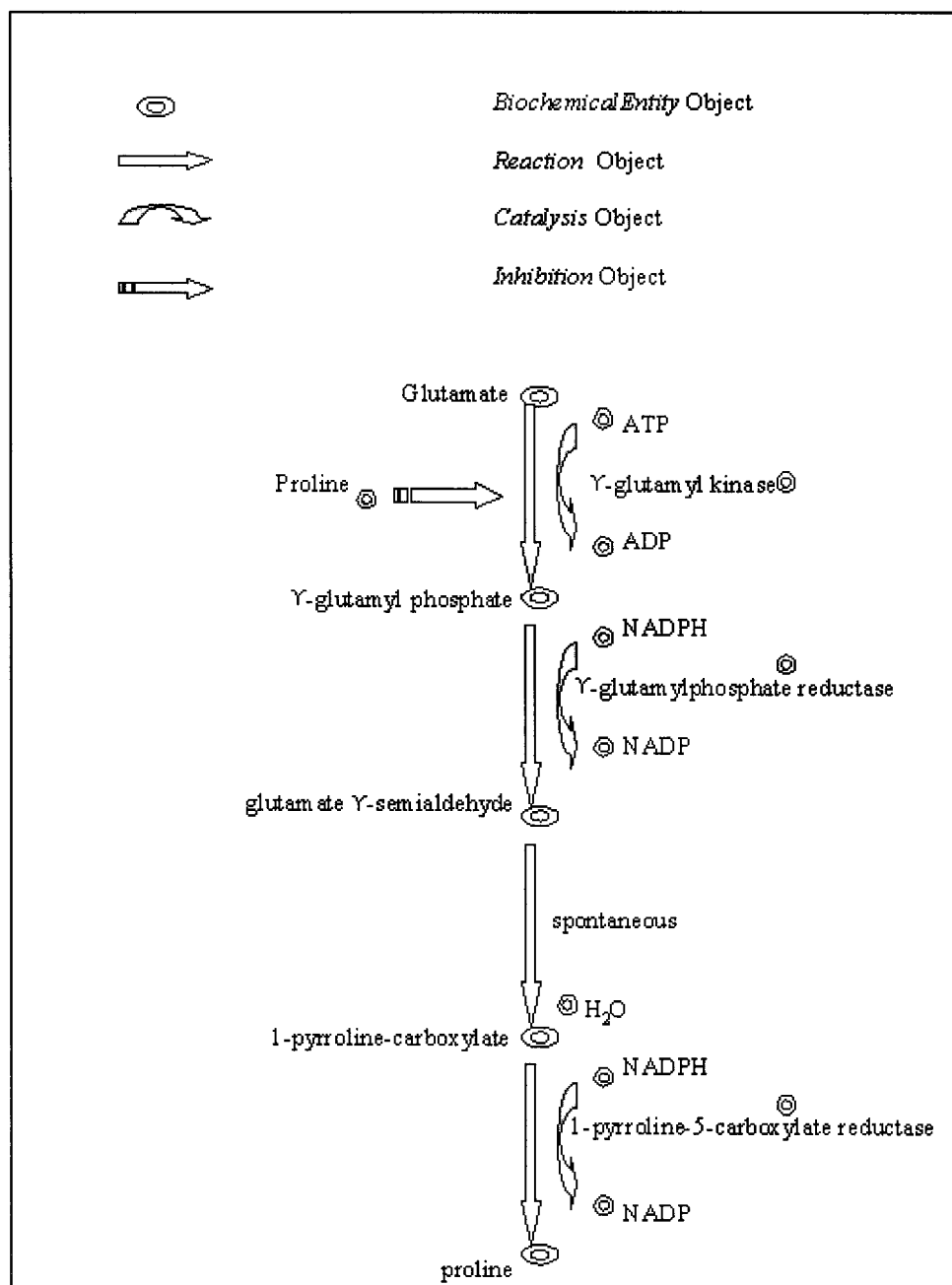


Figure 12. Proline biosynthesis pathway is visualized as directed graphs with the UML data model

Karp et al<sup>(62)</sup> devised an algorithm for drawing metabolic pathways automatically that breaks the graph into cyclic, linear and tree-structured components and then applies different layout methods to each of these individually. The algorithm has been implemented in EcoCyc system<sup>(40)</sup>(see section 3.1.4), which allows biologists to visualize a collection of biochemical information dynamically.

Dynamic visualization in contrast to static visualization provides high flexibility, which is necessary for complex queries and the construction of novel pathways. One of the advantages of this approach is the possibility of automated diagram updated when new data are obtained. This is quite helpful for biologists to analyze the gene experiments since computer does some computation tasks for human. So, we need to pay the price. The problem of dynamically drawing a pathway is a graph layout problem. So, we may use the UML data model and plus the robust algorithm to do dynamic visualization in future.

## **5.4 The UML Data Model are Different from KEGG and EcoCyc Data Model**

If we compare our UML conceptual data model with possible KEGG and EcoCyc data model, it may conclude that the UML data model is quite different others. Here, I assume KEGG has one abstract data model as in Figure 5 and EcoCyc has one abstract data model as in Figure 6. The comparisons of these data models are illustrated in table 3.

Table 3. Comparison of UML Conceptual, KEGG, and EcoCyc Data Model

Data Model	Object-Oriented	Relationship among Data	Knowledge Representation
KEGG	No	binary	No
EcoCyc	Yes	hierarchy and generalization	Yes
UML conceptual	Yes	hierarchy, generalization, aggregation	kind of

It is hard to make a conclusion and say that our UML model is better or not, as our data model is just conceptual data model. What we can tell at this moment is the differences among them in table 3.

## 5.5 More Standards for Nomenclature Are Needed

A comparison of data of pathway database is relatively complicated because of different classification of compounds, genes, proteins, pathways and gene/protein functions. For instance, EcoCyc classified compounds into macromolecules and small molecules, so proteins, genes, and polypeptides under the macromolecules. That's quite different from KEGG and what I am doing in pathway data model construction (see chapter 3). The good thing is the classification of enzymes recommended by the IUBUB (International Union of Biochemistry and Molecular Biology) that is used as a standard. That makes it really easy and simple to define enzyme data structure. One suggestion would be to have more standards for nomenclature. In database development, a standardization of data structures and file formats would make further applications more powerful.

Therefore, having data standards would be ideal, but it is unrealistic to expect them soon. The key technical challenge toward this goal is to develop standardized semantics. Due to the complexity of biological data, its rapidly evolving nature, and problems with synonymy (different names with the same meaning) and polysemy (the same name for different concepts), standards tend to be several steps behind. For this reason, it is concluded that using temporary standards or continuing efforts would be important in merging standards among multiple groups with such similar domains as metabolic pathways and networks.

## **5.6 MIAME and MAGE are Useful**

MIAME requires information on experiment design, sample preparation and labeling, hybridization procedures and parameters, measurement data and specifications, and array design – dozens of bits of information. It facilitates the interpretation and verification of microarray results. MIAME tries to ensure that all the relevant information is captured in a principled way.

The MIAME standards define a benchmark for the minimum standard for the type of information that needs to be recorded during the microarray process. If some of the MIAME data is missing, the experimental data will be regarded as being of insufficient quality to be entered into the public repositories. In addition, most of the main chip-manufacturing and microarray-software vendors are supporting this standard activity. If



data is to be useful and usable in new systems, MIAME compliance needs to be addressed now.

## References

- (1) English L., *Conceptual Data Modeling Student Guide*. Oakbrook Terrace, IL: Platinum Technology, pp.1-10, 1994.
- (2) Barsalou T., Siambela N., Keller A.M., and Wiederholod G. "Updating Relational Databases Through Object-based Views"; *ACM-SIGMOD 91*, Boulder CO, May, 248-257, 1991.
- (3) R. B. Hull and R. King. Semantic database modeling: Survey, applications and research issues. *ACM Computing Surveys*, 19(3): 201-260, September 1987.
- (4) Papazoglou M.P., Spaccapietra S. and Tari Z, *Advances in Object-oriented Data Modeling*, MIT press, pp. 1-20, 2000.
- (5) Reichhardt T. It's sink or swim as a tidal wave of data approaches, *Nature*; 399(6736): 517-520,1999.
- (6) Roberts L, Davenport RJ, Pennisi E, Marshall E. A history of the Human Genome Project, *Science*, Feb 16;291 (5507):1195, 2001.
- (7) Baldi P. and Wesley H. G., *DNA Microarrays and Gene Expression*, Cambridge University press, pp1-10, 2002.
- (8) Brown P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays, *Nature genet.* 21: 33-37, 1999.
- (9) Young R. Biomedical discovery with DNA microarrays. *Cell* 102: 9-16, 2000.
- (10) Lockhart D. & Winzeler E. Genomics, gene expression and DNA arrays. *Nature* 405: 827-836, 2000.

- (11)Butler G. and Denommee P. *Documenting Frameworks to Assist Application Developers, in Building Application Frameworks: Object-oriented Foundations of Framework Design*. John Wiley & Sons Inc. 1999.
- (12)Bornberg-Bauer E. and Paton N.W. Conceptual Data Modeling for Bioinformatics. *Briefings in Bioinformatics*, June 3:166-180 2002.
- (13)Elmasri R. and Navathe S. *Fundamentals of Database System*, 3rd edn, Addison-Wesley, Reading, MA, 2000.
- (14)Fladnders D.J., Weng J., Petel F.X., and Cherry J.M. AtDB, the Arabidopsis thaliana database and graphical-web-display of progress by the Arabidopsis Genome Initiative, *Nucleic Acid Res.*, 26: 80-84, 1998.
- (15)Booch G., *Object-oriented Design with Application*, Benjamin/Cummings, Redwood City, CA, 1991.
- (16)Booch G., Rumbaugh J., and Jacobson I. *The Unified Modeling Language User Guide*, Addison-Wesley, Reading, MA, 1999.
- (17)Ambler Scott, A UML Profile for Data Modeling.  
<http://www.agiledata.org/essays/umlDataModelingProfile.html>
- (18)Schena M., Shalon D, Davis R.W., and Brown P.O. Quantitative monitoring of gene expression pattern with a complementary DNA microarray. *Science*, 270(5235): 467-470, 1995.
- (19)Lockhart D.J. and Dong H. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14: 1675-1680, 1996.
- (20)Velculescu V.E. and Zhang L. et al. Characterization of the yeast transcriptome. *Cell*, 88:243-251, 1997.

- (21) Brown P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays, *Nature genet.* 21: 33-37, 1999.
- (22) Kohane I.S., Kho A.T., and Butte A.J., *Microarrays for an Integrative Genomics*, MIT press, p31-33, 2003.
- (23) Schena M., Shalon D, Davis R.W., and Brown P.O. Quantitative monitoring of gene expression pattern with a complementary DNA microarray. *Science*, 270(5235): 467-470, 1995.
- (24) Velculescu V.E. and Zhang L. et al. Characterization of the yeast transcriptome. *Cell*, 88:243-251, 1997.
- (25) Mavrovouniotis M.L., Identification of Qualitatively Feasible Metabolic Pathways, Chapter 9 of *Artificial Intelligence & Molecular Biology*, e-book, 2001.  
<http://www.biosino.org/mirror/www.aaai.org/Press/Books/Hunter/hunter-contents.html>
- (26) Malacinski G.M., Freifelder D., Regulation of Gene Activity In Prokaryotes, Chapter 11 of *Essentials of Molecular Biology*, Jones and Bartlett Publishers, Inc. 1998.
- (27) Network Science Corporation's terms and definitions in biology and bioinformatics:  
<http://www.netsci.org/Science/Bioinform/definitions.html>
- (28) Wittig U., and Beuckelaer A.D., Analysis and Comparison of Metabolic Pathway Databases, *Briefings in Bioinformatics*, May 2: 126-142, 2001.
- (29) Karp P.D., Metabolic databases. *TRENDS in Biochemical Sciences*, 23:114-116, March 1998.
- (30) Ogata H., S. Goto, K. Sato. KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acid Res.* 27(1): 29-34, 1999.
- (31) Goto S., Nishioka T. and Kanehisa M. LIGAND: Chemical database for enzyme reactions, *Nucleic Acid Res.* 28(1): 380-382, 2000.

- (32)DBGET/LinkDB - Integrated database retrieval system (2000).  
<http://www.genome.ad.jp/dbget/>
- (33)Locher K. P., Lee A. T., Rees, D. C. The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* 296: 1091-1098, 2002.
- (34)Hoch J.A. and Silhavy T.J. *Two-Component Signal Transduction*, ASM Press, 1995.
- (35)Goto S., Nishioka, T., et al. LIGAND: Chemical database for enzyme reactions, *Bioinformatics*, 14(7): 591-599, 1998.
- (36)Kanehisa M., Goto S., Kawashima S., and Nakaya, A. The KEGG databases at Genome Net, *Nucleic Acids Res.* 30: 42-46, 2002.
- (37)EcoGene database (2000). URL: <http://bmb.med.miami.edu/EcoGene/index.html>
- (38)Karp P. D. and Riley M. "EcoCyc: The resource and the lessons learned" in Letovsky, S. Ed., *Bioinformatics Databases and Systems*, Kluwer Academic, Boston, MA pp47-62, 1999.
- (39)Letovsky S. Chapter four of *Bioinformatics: Databases and Systems*, Kluwer Academic Publishers, 1999.
- (40)Karp P. D., Riley M., and Saier M. EcoCyc and MetaCyc databases, *Nucleic Acids Res.* 28(1): 56-59, 2000.
- (41)Overbeek R., Larsen N. Pusch G. D. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acid Res.* 28(1): pp123-125, 2000.
- (42)Kolpkov F.A., Ananko E.A., Kolesov G.B., and Kolchanov N.A., GeneNet: a gene network database and its automated visualization, *Bioinformatics*, 14: (6) 529-537, 1998.

- (43) Anan'ko, E.A., Bazhan, S.I., Belova, O.E. and Kel', A.E. Mechanisms of transcription of the interferon-induced genes: a description in the IIG-TRRD information system. *Mol. Biol. (Mosk)*, 31, 592-605, 1997.
- (44) Harada, H., Takahashi, E.I., Itoh, S., Harada, K., Hori, T.A. and Taniguchi, T. Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system. *Mol. Cell. Biol.*, 14, 1500-1509, 1994.
- (45) Kolchanov, N.A. Transcription regulation in eukaryotic genes: databases and computer analysis. *Mol. Biol. (Mosk)*, 31, 581-583, 1997.
- (46) GenMAPP.org organization's GenMAPP (Gene MicroArray Pathway Profiler), 2002, <http://www.genmapp.org/intro.html>
- (47) Dahlquist K.D., Salomonis N., Vranizan K et al. *Nature Genetics*, 31: 19-20, 2002.
- (48) Karp P. and Paley S. and Romero P. The pathway tools software, *Bioinformatics* 18: 225-232, 2002.
- (49) Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D. DIP: The Database of Interacting Proteins, *Nucleic Acids Research*, 29:239-241, 2001.
- (50) Xenarios I., Salwinski L., Duan X. J., Higney P., Kim S. and Eisenberg D. DIP, the Database of Interesting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Research*, 30(1): 303-305, 2002. <http://dip.doc-mbi.ucla.edu>
- (51) The University of California's DIP database, 1999. <http://dip.doc-mbi.ucla.edu/jin.xsd>

- (52) Spellman P.T., Miller M., Stewart J. et al. Design and implementation of microarray gene expression markup language (MAGE-ML) *Genome Biology*, 3(9): research, 0046.1–0046.9, 2002.
- (53) MGED – Microarray Gene Expression Data Society (<http://www.mged.org>)
- (54) Branzma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., et al. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat Genet*, 29: 365-371, 2001.  
<http://www.mged.org/miame>
- (55) Microarray Bioinformatics Group in the University of Manchester, a data warehouse and visualisation environment for genomic expression data, named maxd, 2002.  
<http://www.bioinf.man.ac.uk/microarray/maxd/index.html>
- (56) NCGR. National Center for Genome Resources, GeneX-Lite : Gene Expression Lite.  
<http://www.ncgr.org/genex/>
- (57) The European Bioinformatics Institute (EBI), MIAMExpress, a public repository for microarray based gene expression data, (2001).  
<http://www.ebi.ac.uk/microarray/MIAMExpress/miamexpress.html>
- (58) American National Standards Institute. ANSI/X3/SPARC Study Group on Data Base Management Systems; *Interim Report. FDT (Bulletin of ACM SIGMOD)* 7:2, 1975.
- (59) Date C.J. *An Introduction to Database System, Second Edition*. Reading, MA: Addison-Wesley, 1977.
- (60) NIH genetic sequence database, GenBank. <http://www.ncbi.nlm.nih.gov/Genbank/>
- (61) Blaha M. and Premerlani W. *Object-oriented Modeling and Design for Database Applications*, Prentice-Hall, Englewood Cliffs NJ, 1998.

- (62)Karp P. and Paley S. Automated drawing of metabolic pathway. In Lim, H. (ed.), *Processing of the Third International Conference on Bioinformatics and Genome Research*, World Scientific Publishing Co. pp225-238, 1995.
- (63)Brandenburg F. J., Gruber B., Himsolt M. and Schreiber F. *Automatische Visualisierung Biochemischer Information*, In Proceedings of The Workshop Molecular Bioinformatics, GI Jahrestagung, pp. 24-38, 1998.
- (64)Dr. Don's Biology course: <http://www.cariboo.bc.ca/schs/biol/FacPgs/Nelson/>
- (65)Kanehisa, M., Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan*, No. 59, pp. 34-38 (1996).
- (66)Karp P. and Riley M. EcoCyc: The Resource And The Lessons Learned in Chapter 4 of *Bioinformatics: Databases and Systems*, Kluwer Academic Publishers, 1999.



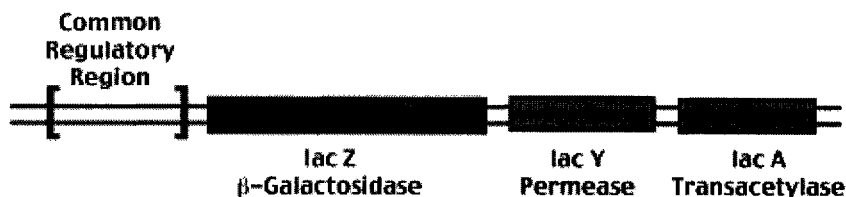
## Appendix A: LAC OPERON MODEL <sup>(64)</sup>

### 1. The Lac Operon

The lac operon of *E. coli* is a genetic unit that encodes the biochemical pathway that allows the bacterium to utilize lactose as a carbon-source. Lactose is a disaccharide made up of a galactose and a glucose linked in a beta-1,4 glycosidic linkage.

The enzyme **beta-galactosidase**, encoded by the gene **lac Z**, cleaves the beta-1,4 glycosidic linkage to release the sugars galactose and glucose. In *E. coli*, the transport of lactose across the cell membrane requires a second gene - **lac Y** - which encodes the **permease**. A third function- **transacetylase** - is encoded by **lac A**. These three proteins constitute the biochemical pathway for lactose utilization.

The term, **operon**, refers to a set of structural regions (usually encoding proteins) which are clustered together and whose expression is under the control of a single regulatory region. The overall structure of the lac operon is illustrated below:



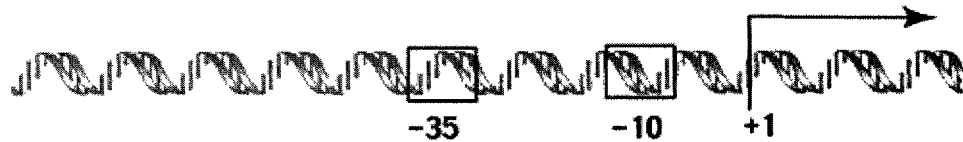
In 'standard' terminology, each protein coding unit is called a **cistron**. Since the entire operon is expressed as a single mRNA from a common regulatory region, the primary transcript encodes three separate proteins. Such a message is called **polycistronic**. What is the rationale for the operon organization of multiple coding regions under the control of a common regulatory region? Two rationales are commonly offered.

The first emphasizes the coordinate regulation of multiple functions required for a single biochemical pathway. Since all three functions are required, it makes energetic sense to express all three from a single mRNA.

The second rationale emphasizes recent evidence for the lateral transmission of heritable information between individuals. This rationale suggests that the three functions are clustered together to facilitate their transfer as a complete unit rather than as individual units. This allows the transfer of a complete functional biochemical pathway rather than those of individual functions that are ineffective in isolation.

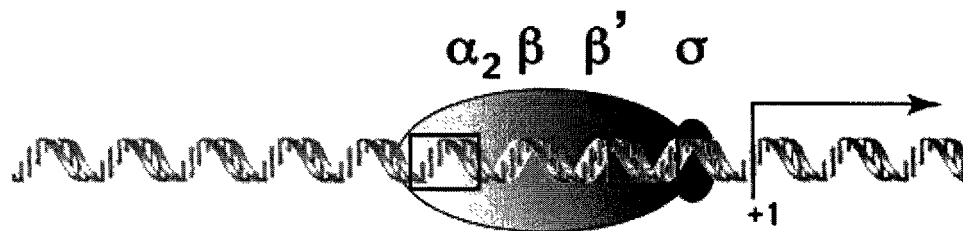
The initiation of transcription by RNA polymerase involves the binding of the polymerase to specific template sequences called the promoter. The structure of the promoter was originally revealed by comparing the DNA sequences' upstream of many different transcription initiation sites. Two blocks of highly conserved sequence were identified centered 10 bp and 35 bp upstream of position +1 - the first nucleotide of the transcript.

These two 'promoter' sequences were subsequently shown to interact directly with RNA polymerase during the initiation phase.



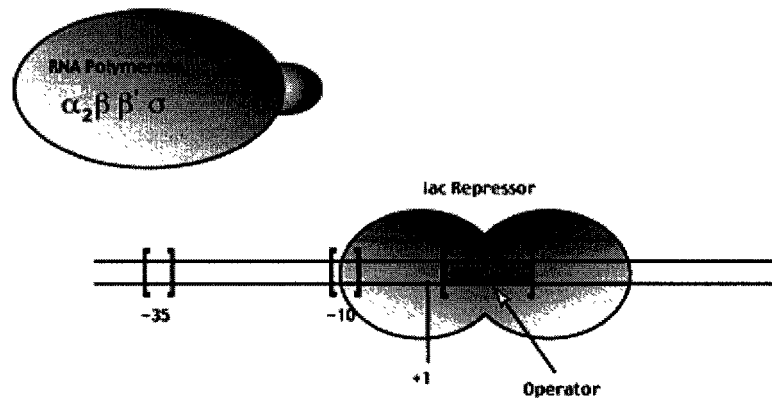
## 2. Basic Features of the Promoter

While the alpha (2) beta - beta-prime complex is the active polymerase during the elongation phase, it is unable to recognize the promoter sequence on its own. To initiate, the polymerase must interact with an initiation factor called sigma. This polymerase complex is then capable of recognizing and binding to the promoter as shown below



The lac repressor, coded for by the lac I gene, is a sequence specific DNA binding protein. The specific sequence recognized by the lac repressor is located just downstream of the lac promoter. Binding of the lac repressor to this recognition sequence (called an

**operator** as it operates on the adjacent promoter) negatively regulates the initiation phase of the transcription process by sterically blocking access of the RNA polymerase to the adjacent promoter.

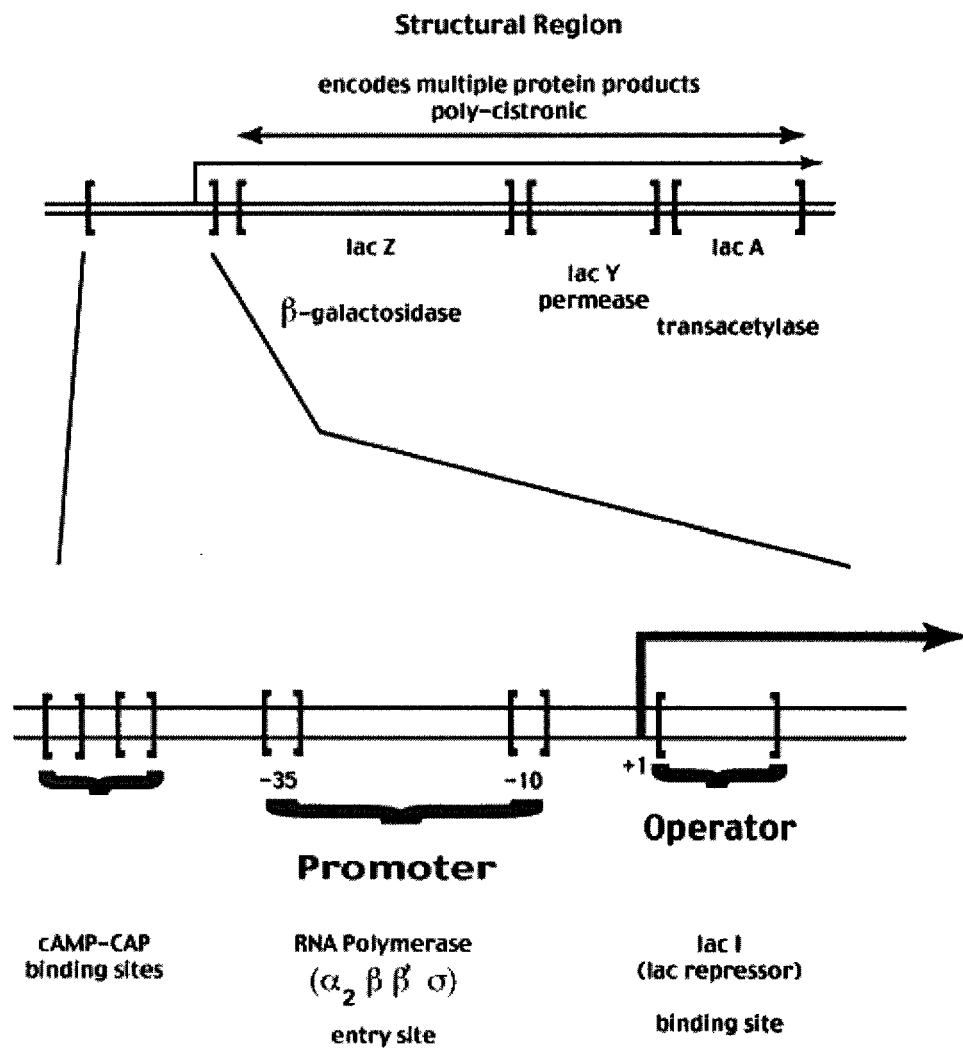


*E. coli* expresses the lac I gene constitutively - there is always lac repressor running around the cell. So how do we turn on transcription of the lac operon when we need it? Activation of the lac operon requires the presence of the substrate of the biochemical pathway it encodes - lactose. A metabolic derivative of galactose - allolactose - is the **inducer** of the operon. Allolactose binds to the lac repressor, causing a conformational shift in the protein that results in the loss of its sequence specific DNA binding ability (allosteric regulation). Thus, in the presence of the inducer, the repressor no longer binds to the operator - thus freeing the promoter that is now accessible to the RNA polymerase + sigma factor and transcription can initiate as discussed previously.

This mechanism for the negative regulation of transcription initiation by steric interference with RNA polymerase - promoter interactions is very common in

prokaryotes. In addition to this steric blockage mode of negative regulation, prokaryotes also have mechanisms that positively regulate the initiation of transcription. For the lac operon, positive regulation involves a second regulatory protein - the cAMP-CAP (Catabolic Activation Protein) (a sequence specific DNA binding protein) and its target binding sites. The diagram below shows how the various regulatory sequence elements are organized in the lac operon regulatory region.

The binding of cAMP-CAP to its target sequence(s) increases the frequency with which RNA polymerase initiates at the adjacent promoter. This enhanced frequency of initiation is due to the affinity of the cAMP-CAP for RNA polymerase itself (via protein-protein interaction). This affinity for the transcription enzyme results in an increase in the local RNA polymerase concentration in the immediate vicinity of the lac promoter. This increase in local concentration increases the frequency with which transcription initiates from this promoter.



## Appendix B MIAME VERSION 1.0

### The MIAME structure

MIAME recommendations include sections that will usually be provided in a free text format, along with information that are recommended to be given by maximum use of controlled vocabularies or external ontologies (such as species taxonomy, cell types, anatomy terms, chemical compound nomenclature). The use of controlled vocabularies is needed to enable database queries and automated data analysis. Since few controlled vocabularies have been fully developed, MIAME encourages the users, if necessary, to provide their own qualifiers and values identifying the source of the terminology. This is achieved through the use of

*(qualifier, value, source)*

triplets, for instance,

*(qualifier: 'cell type', value: 'epithelial', source: 'Gray's anatomy, 38<sup>th</sup> ed.'),*

which are recommended instead or in addition to free text format descriptions wherever possible. This will allow the community to build up a knowledge base of the most useful controlled vocabularies for describing microarray experiments. The MGED group is

developing an ontology for microarray experiment description, and where the ontology is sufficiently mature, the MIAME document recommends its use.

Microarrays are often manufactured independently of particular experiments and their design description can be given separately. Therefore MIAME has two major sections

- (1) array design description;
- (2) gene expression experiment description.

Another potentially reusable part of the experiment description is laboratory protocols, including data processing methods (e.g., normalization). MIAME encourages the user to assign unique identifiers to all reusable parts of experiment description and to reference these when the respective parts are reused (possibly indicating the deviations). A standard for the description of protocols, including the data transformation protocols is being developed by MGED.

## 1. Array design description

The array design specification consists of the description of the common features of the array as the whole and the description of each array design elements (e.g., each spot). Following terminology used in MAGE, we distinguish between three levels of array design elements:



feature - the location on the array,

reporter - the nucleotide sequence present in a particular location on the array,

composite sequence - a set of reporters used collectively to measure an expression of a particular gene, exon, or splice-variant.

The details that should be given of each of them are described below.

#### 1) Array related information

- array design name
- platform type: in situ synthesized, spotted or other
- surface and coating specification
- physical dimensions of array support (e.g. of slide)
- number of features on the array
- availability (e.g., for commercial arrays) or production protocol for custom made arrays

#### 2a) For each reporter type

- the type of the reporter: synthetic oligo-nucleotides, PCR products, plasmids, colonies, other
- single or double stranded

2b) For each reporter

- sequence or PCR primer information:
  - o sequence or a reference sequence (e.g., for oligonucleotides), if known
  - o sequence accession number in DDBJ/EMBL/GenBank, if exists
  - o primer pair information, if relevant
- approximate lengths if exact sequence not known
- clone information, if relevant (clone ID, clone provider, date, availability)
- element generation protocol that includes sufficient information to reproduce the element for custom-made arrays that are not generally available

3a) For each feature type

- dimensions
- attachment (covalent/ionic/other)

3b) For each feature

- which reporter and the location on the array

4) For each composite sequence

- which reporters it contains
- the reference sequence

- gene name and links to appropriate databases (e.g., SWISS-PROT, or organism specific databases), if known and relevant

#### 5) Control elements on the array

- position of the feature (the abstract coordinate on the array)
- control type (spiking, normalization, negative, positive)
- control qualifier (endogenous, exogenous)

For each array that is not generally available (e.g., commercially available), the provided information should be sufficient to reproduce the array and all its design features.

## 2. Experiment description

By experiment we understand a set of one or more hybridizations that are in some way related (e.g., related to the same publication). The minimum information includes a description of the following four parts.

1. Experimental design
2. Samples used, extract preparation and labeling
3. Hybridization procedures and parameters
4. Measurement data and specifications of data processing

MIAME recommends the following details on each of these sections.

(1) Experimental design

This section that is common to all the hybridizations done in the experiment, such as the goal, brief description, experimental factors tested. It includes the following.

1) Authors, laboratory, contact

2) Type of the experiment, for instance,

- normal vs. diseased comparison
  - treated vs. untreated comparison
  - time course
  - dose response
  - effect of gene knock-out
  - effect of gene knock-in (transgenics)
- (multiple types possible)

3) Experimental factors, i.e. parameters or conditions tested, for instance,

- time
- dose
- genetic variation

- response to a treatment or compound

(also, see <http://www.mged.org/ontology>)

4) How many hybridizations in the experiment?

5) If a common reference is used for all the hybridizations?

6) Quality control steps taken:

- if any replicates done (yes/no), what type of replicates, description?
- whether dye swap is used (only for two channel platforms)?
- other (e.g., polyA tails, low complexity regions, unspecific binding)

7) A brief description of the experiment and its goal and a link to a publication if exists

8) Links (URL), citations

## (2) Samples used, extract preparation and labeling

By a sample we understand the biological material (biomaterial), from which the nucleic acids have been extracted for subsequent labeling and hybridization. In this section all steps that precedes the hybridization with the array are described. We can usually distinguish between the source of the sample (bio-source, e.g., organism, cell type or line), its treatment, the extract preparation, and its labeling. MGED is developing an

ontology for sample description (see <http://www.mged.org/ontology>) the use of which is encouraged. Here we list the most essential items that are usually needed.

#### 1) Bio-source properties

- organism (NCBI taxonomy)
- contact details for sample
- descriptors relevant to the particular sample, such as
  - o sex
  - o age
  - o development stage
  - o organism part (tissue)
  - o cell type
  - o animal/plant strain or line
  - o genetic variation (e.g., gene knockout, transgenic variation)
  - o individual genetic characteristics (e.g., disease alleles, polymorphisms)
  - o disease state or normal
  - o is additional clinical information available (link)
  - o the individual (for interrelation of the samples in the experiment)

#### 2) Bio-material manipulations: laboratory protocol, including relevant parameters, e.g.,

- growth conditions
- in vivo treatments (organism or individual treatments)
- in vitro treatments (cell culture conditions)
- treatment type (e.g., small molecule, heat shock, cold shock, food deprivation)
- compound
- separation technique (e.g., none, trimming, microdissection, FACS)

For recommendations for controlled vocabularies that can be used see

<http://www.mged.org/ontology>

### 3) Hybridization extract preparation protocol for each extract prepared from the sample, including

- extraction method
- whether total RNA, mRNA, or genomic DNA is extracted
- amplification (RNA polymerases, PCR)

### 4) Labeling protocol for each labeling prepared from the extract, including

- amount of nucleic acids labeled
- label used (e.g., A-Cy3, G-Cy5, 33P, ....)
- label incorporation method

5) External controls added to hybridization extract(s) (spiking controls)

- element on array expected to hybridize to spiking control
- spike type (e.g., oligonucleotide, plasmid DNA, transcript)
- spike qualifier (e.g., concentration, expected ratio, labelling methods if different than that of the extract)

(3) Hybridization procedures and parameters

Each hybridization description should include

- 1) information about which labeled extract (related to which sample, which extract) and which array (e.g., array design, batch and serial number) has been used in the experiment; and
- 2) the hybridization protocol, normally including
  - the solution (e.g., concentration of solutes)
  - blocking agent
  - wash procedure
  - quantity of labeled target used
  - time, concentration, volume, temperature



- description of the hybridization instruments

#### (4) Measurement data and specifications of data processing

We distinguish between three levels of data processing - raw data (images), image quantitations and gene expression data matrix. Each hybridization has at least one image, each image has a corresponding image quantitation table, where a row represents an array design element and a column to a different quantitation types, such as mean or median pixel intensity. Several quantitation tables can be combined using data processing metrics to obtain the 'final' gene expression measurement table associated with the experiment.

##### 1) Raw data description should include

- for each scan laboratory protocol for scanning, including scanning hardware and software, scan parameters, including laser power, spatial resolution, pixel space, PMT voltage;
- scanned images;

It should be noted that MGED does not have consensus whether the provision of images is a part of MIAME.

##### 2) Image analysis and quantitation

- image analysis software specification and version, availability, and the description or identification of the algorithm and all the parameters used
- for each image the complete image analysis output (of the particular image analysis software)

### 3) Normalized and summarized data - gene expression data matrix

- data processing protocol, including normalization algorithm (for detailed recommendations, see <http://www.mged.org/normalization>)
- gene expression data table(s) derived from the experiment as the whole,
  - o derived measurement value summarizing related elements and replicates as used by the author (this may constitute replicates of the element on the same or different arrays or hybridizations, as well as different elements related to the same entity e.g., gene)
  - o providing a reliability indicator for each datapoint (e.g., standard deviation) is encouraged

This ends the experiment description. The document is based on the earlier version MIAME 1.0 and discussions at MGED 4 meeting. The more detailed information about MIAME array design and description as well as gene expression experiment description can be found at (27).

## **Appendix C: MAGE-ML and MAGE-OM**

### **MicroArray Gene Expression Markup Language and MicroArray Gene Expression Object Model**

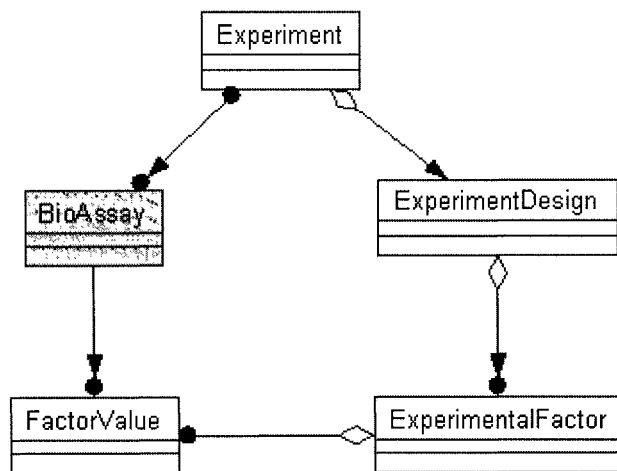
Microarray Gene Expression Markup Language (MAGE-ML) is a language designed to describe and communicate information about microarray based experiments. MAGE-ML is based on XML and can describe microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results.

MAGE-ML has been automatically derived from Microarray Gene Expression Object Model (MAGE-OM), which is developed and described using the Unified Modelling Language (UML) – a standard language for describing object models. Descriptions using UML have an advantage over direct XML document type definitions (DTDs), in many respects. First they use graphical representation depicting the relationships between different entities in a way which is much easier to follow than DTDs. Second, the UML diagrams are primarily meant for humans, while DTDs are meant for computers. Therefore MAGE-OM should be considered as the primary model, and MAGE-ML will be explained by providing simplified fragments of MAGE-OM, rather than XML DTD or XML Schema.

MAGE-OM is a bit too large to be represented on a single diagram in a readable way. In order to structure the model the UML notion of *packages* is used. Related classes are grouped together into packages, and quite often represented on the same diagrams. MAGE-OM will be explained package-by-package; a tool able to export MAGE-ML will probably have separate modules and/or user interface sections for separate packages, e.g., you can enter information about array designs in one UI section and information about steps of your microarray experiment using another UI section. On diagrams classes belonging to the package under discussion are coloured yellow, while classes belonging to other packages, therefore detailed elsewhere, but drawn on the current diagram for the purposes of showing inter-package relationships, are coloured grey.

### ***Experiment***

This package is for describing a microarray experiment as a unit. Note two parallel branches on the diagram. On the right-hand side we have experiment blueprint information – experiment design and one or more experimental factors that are changed in the course of the experiment to explore whether and how gene expression levels change (e.g., time or drug concentration). On the left-hand side there is experiment execution information. An experiment consists of one or more bioassays (experiment steps), and each bioassay can test for gene expression with one or more experimental factor values fixed (e.g., time = 30min, drug concentration = 15ug/ml).

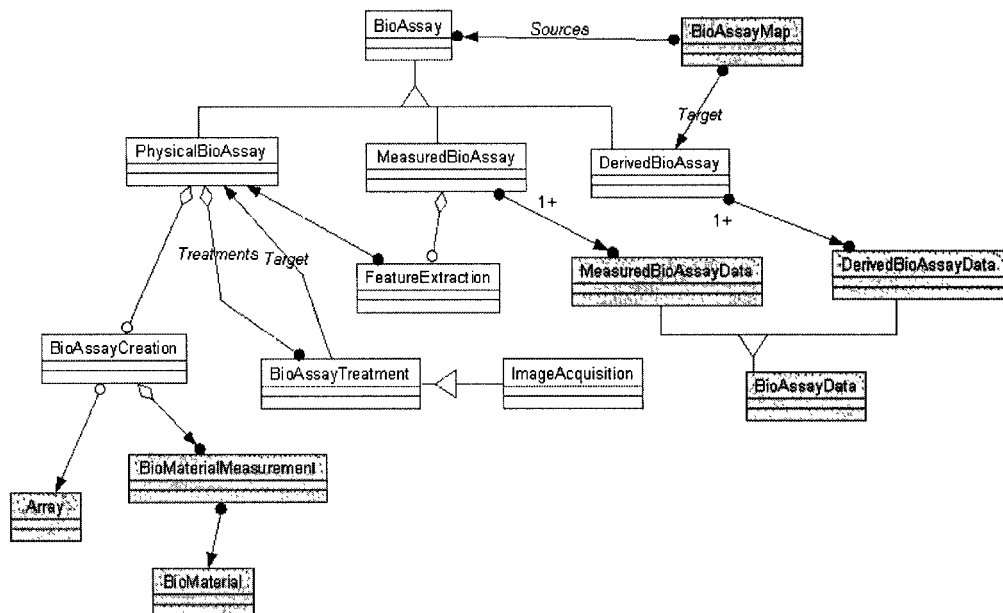


### ***BioAssay***

A bioassay is a single step within a microarray experiment. There are 3 types of bioassays. A physical bioassay corresponds to wet-lab microarray experimental step. A measured bioassay corresponds to a situation after feature extraction has been performed. A derived bioassay corresponds to data processing experimental steps.

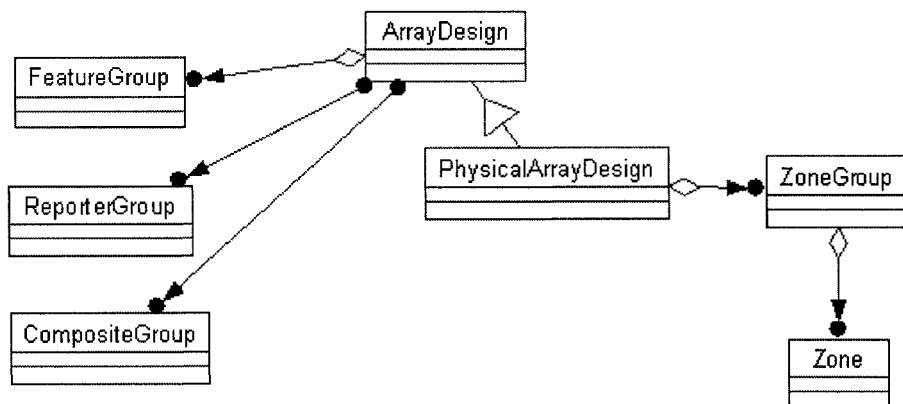
A physical bioassay is created by applying some amount of some biomaterial to a microarray. Bioassay treatment events (e.g., wash, apply blocking agent etc.) transform physical bioassays into new physical bioassays. A particular type of bioassay treatment is image acquisition.

Measured bioassays can have corresponding MeasuredBioAssayData objects (raw data). Derived bioassays are obtained by data transformations, they are linked by BioAssayMap objects.



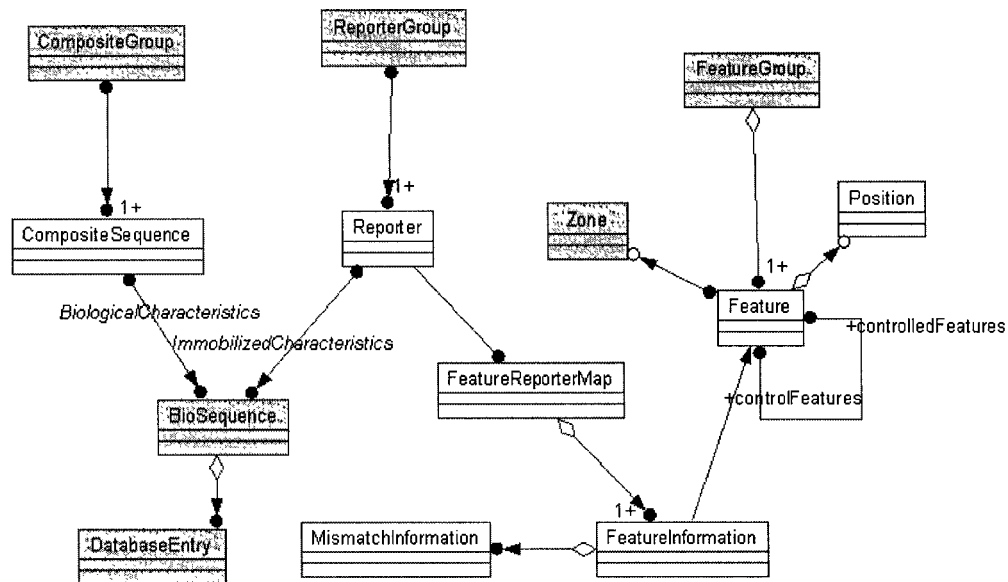
## ArrayDesign

An array design consists of design element groups as well as information about element zone layout. Physical array design has been made as a subclass of array design, to allow “virtual” array designs with element groups but no zone layout information; such “virtual” designs can be used, e.g., to define different reporter-composite sequence mappings for the same physical array design. Zones can be grouped together to form zone groups; zones within the same zone group would have the same spacing between them, whereas zones from different zone groups can have different spacing. Zones/zone groups sometimes are referred to as blocks/metablocks.



### ***DesignElement***

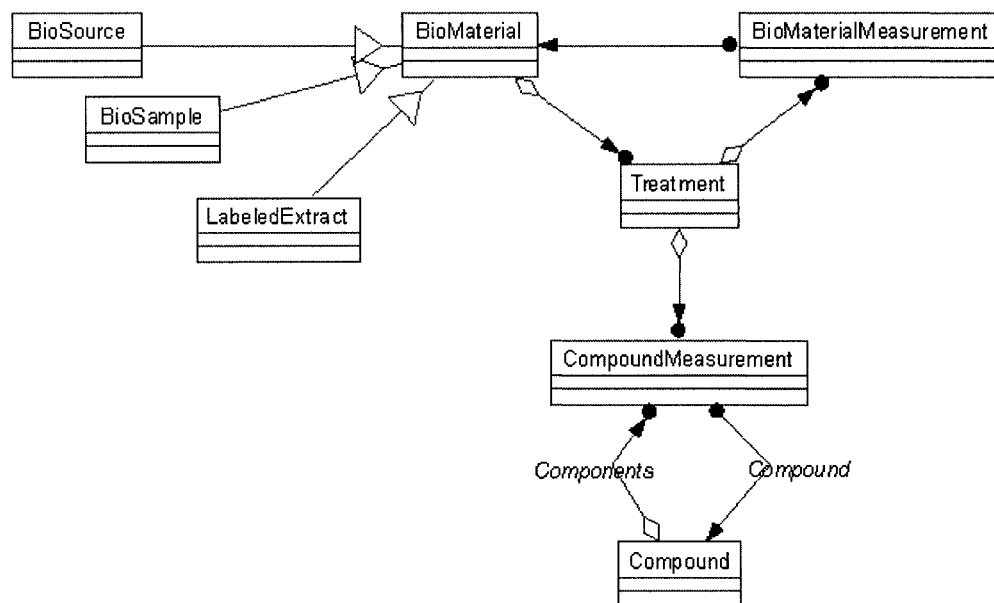
There are three types of design elements. A feature has a position on the array (within some zone), it can be a control feature for other features or controlled by other features. A reporter corresponds to the physical substance synthesized/printed on the array, it can be characterized by one or more biosequence objects which in turn can be characterized by database entries. There can be many features for the same reporter, and features can have one or several mismatches compared to reporter's reference sequence. The third, most abstract kind of design element is a composite sequence. It can have more than one reporter on the same array (e.g., different splice variants) and is characterized by biological characteristics, which are actually again sequences with corresponding database entries. The mapping between reporters and composite sequences is not shown on this diagram, but this is similar to the model of feature-reporter mapping. Also, composite sequences can be aggregated into more abstract composite sequences (also not shown here), e.g., genes of the same functional group etc.



## ***BioMaterial***

BioMaterial is an abstraction of various states of biology-based materials used in various stages of the microarray experiment. Biosource refers to the initial source of material used in hybridization (e.g., cell line or tissue). Biosample is what is extracted from the biosource, and labeled extract is the last state of the biomaterial before hybridization. A biomaterial can be a result of a chain of treatments, each treatment involving one or more biomaterials in some amounts. A special kind of treatment is treatment with some amount of a compound. A simple way to model compounds consisting of other compounds is provided.

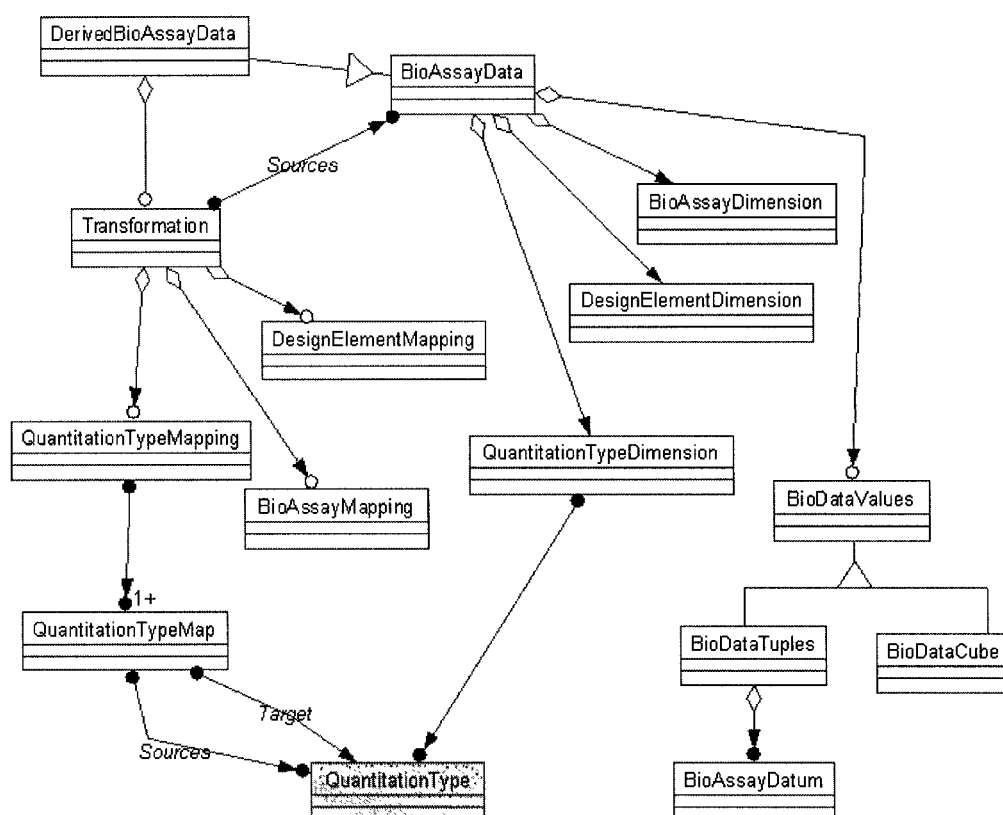




### ***BioAssayData***

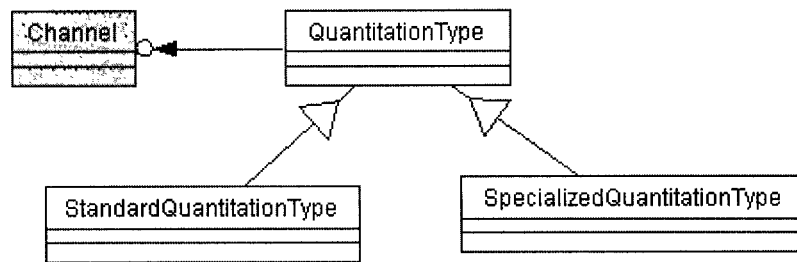
One of the central principles of MAGE is that data objects are regarded as 3-dimensional matrices, where there are bioassays (experimental steps or conditions) along one dimension, design elements (spots) along the other dimension and quantitation types (e.g., signal intensity, background intensity) along the 3<sup>rd</sup> dimension. Bioassay data objects can be represented in one of two ways: as a set of vectors in the form (value, dimension1, dimension2, dimension3) (useful for small amounts of data), or as a 3-D matrix (BioDataCube). Transformations (e.g., filtering, normalization) can be applied to one or more bioassay data objects, resulting in derived data objects. A transformation involves computing values of the resulting 3-D matrix from the values of source matrices, and it also transforms dimensions. On this diagram just the mapping of quantitation types into new quantitation types has been shown; DesignElementMapping and

BioAssayMapping are modeled similarly. A quantitation type mapping transforms a list of quantitation types into another list of quantitation types, and it consists of maps that deal with single target quantitation types.



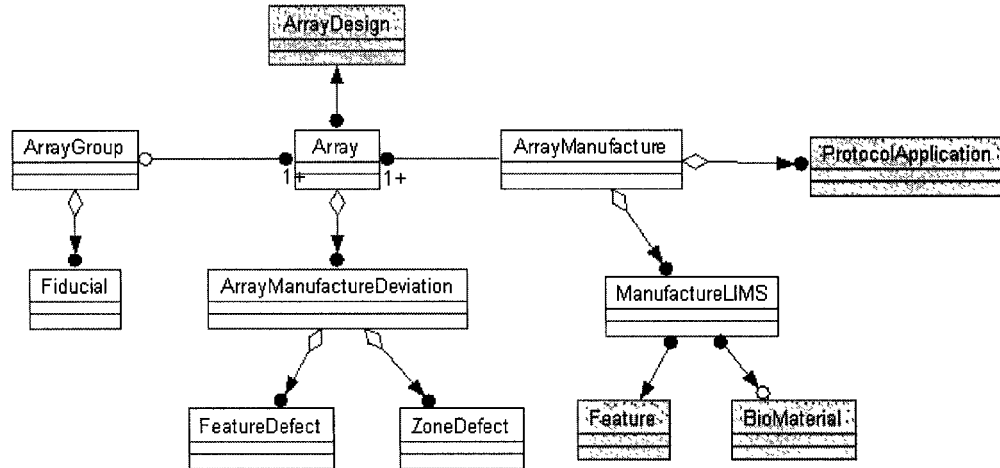
### Quantitation Type

A quantitation type can be either a standard quantitation type (a list of these is provided within MAGE) or a specialized quantitation type that should be described in detail. A quantitation type may reference a channel (e.g., Cy3 green signal intensity).



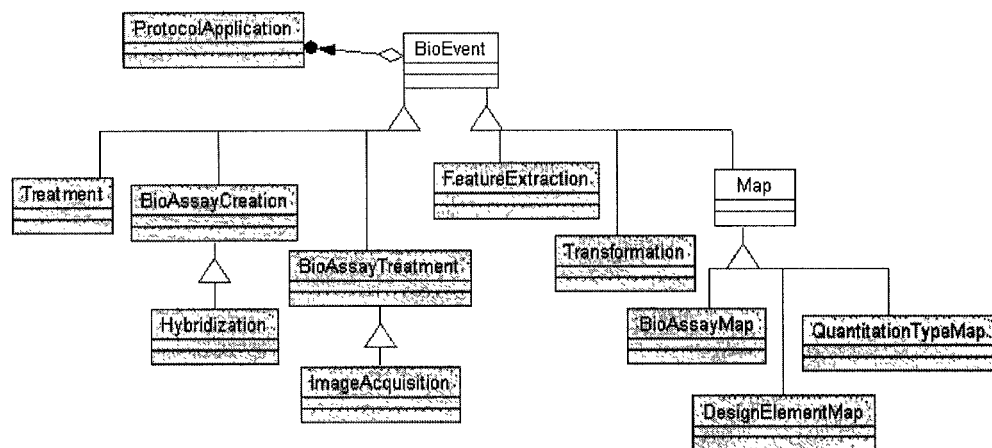
### *Array*

An array is a physical array that corresponds to some array design. There are three types of information that can be captured about individual arrays and array production process. An individual array can have deviations from the design, either zone defects (e.g., a whole zone of spots is shifted) or individual feature defects. An array group can consist of more than one array printed on the same slide, and fiducials (markings on the surface of the slide that can be used to identify arrays' origins) can be printed to facilitate feature detection software accuracy. Array manufacture information also can refer to more than one individual array (sometimes referred to as an array batch), and it can contain protocol information (how the arrays were manufactured) as well as some limited LIMS information (what was printed on the array, on a feature-by-feature basis).



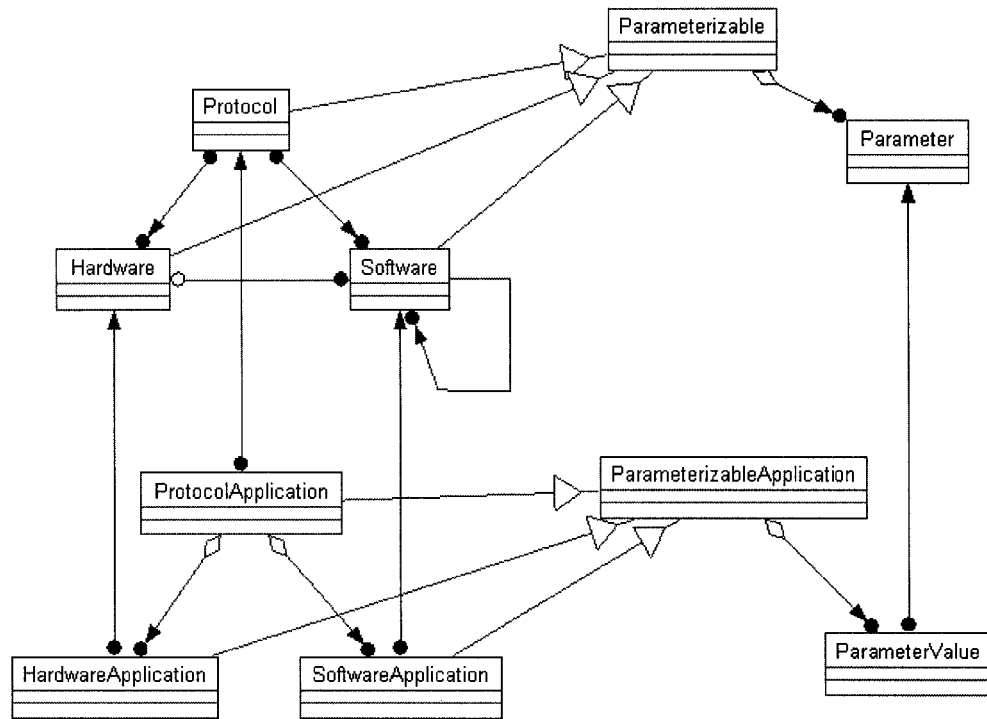
### ***BioEvent***

This diagram is provided to summarize what kinds of events are possible to describe in MAGE. Each event can have a sequence of protocol applications. On the left-hand side there are physical events (biomaterial treatment, bioassay creation as a generalization for hybridization, and bioassay treatment and image acquisition as a special case), while on the right-hand side there are information processing events (feature extraction, data transformation, maps of data dimensions).



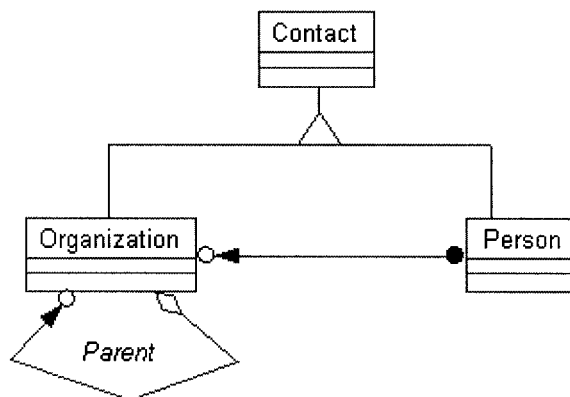
## ***Protocol***

There are two parts for this package. On the upper part there are Protocol, Hardware and Software classes, representing abstract entities. All the “parameterizable” objects can have parameters, a protocol can involve usage of specific hardware and software, specific hardware might be needed to run some software, and software objects can be composed of other software objects (modules). On the lower part there are classes representing application of abstract entities at a given time point, with parameters filled in by some parameter values.



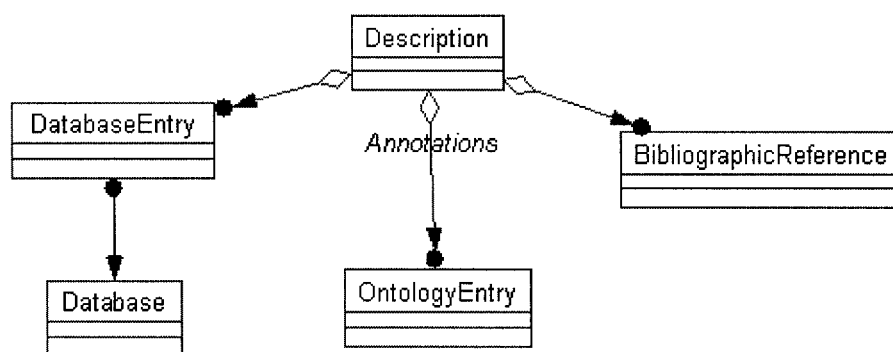
### ***AuditAndSecurity***

A contact can be either an organization or a person. A person can work for an organization, and an organization can consist of other sub-organizations.



### ***Description***

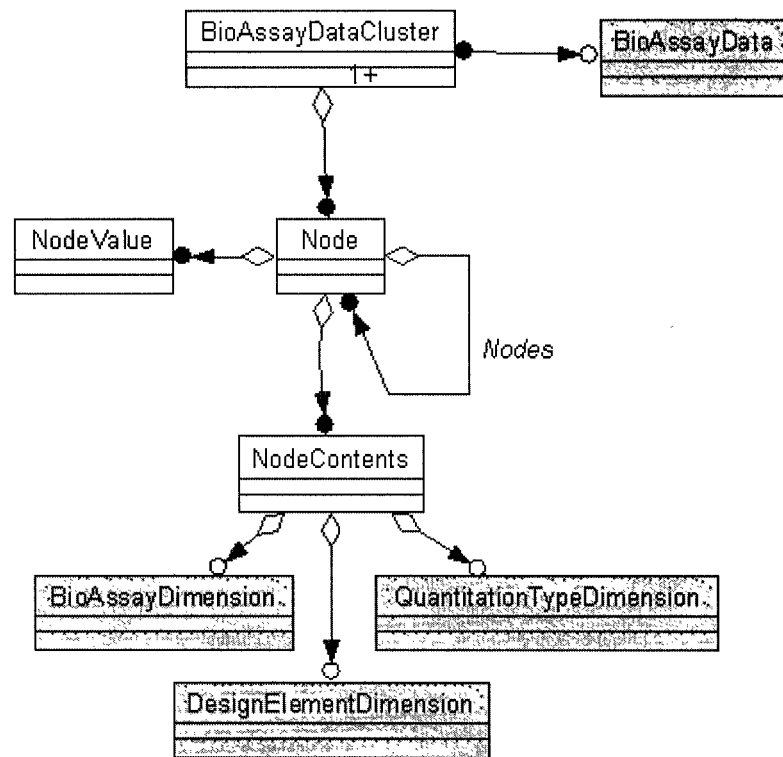
Many MAGE objects can be further characterized by attaching Descriptions to them, if there is some important information that cannot be recorded using provided attributes and relations. A Description can consist of a piece of free text, references to database entries, references to ontology entries (annotations) and one or more bibliographic references. A generic NameValueType class also is provided, objects of which can be attached to every MAGE object if even the Description functionality is not sufficient.



### ***HigherLevelAnalysis***

Experimental data (BioAssayData) can be clustered, obtaining one or more top level clusters. Each cluster consists of nodes, where each node in turn can contain subnodes (in the case of hierarchical clustering). A node can be characterized by its values (e.g., some metric of cluster quality), and a node groups together design elements (e.g., spots, genes) or bioassays (i.e., experimental conditions). BioAssayDimension is

just an ordered list of bioassays, and DesignElementDimension is an ordered list of design elements.





## APPENDIX D: GLOSSARY IN BIOLOGY AND BIOINFORMATICS<sup>(27)</sup>

**Amino acid:** An  $\alpha$ -amino carboxylic acid of the general form  $\text{H}_3\text{N}-\text{CHR}-\text{COO}^-$ . There are 20 common amino acids, defined by the R group on the alpha-carbon ([A listing of common amino acids is available](#)), that are used to build proteins and peptides.

**Base:** One of five molecules that are assembled, along with a ribose and a phosphate, to form nucleotides Adenine (A), guanine (G), cytosine (C), and thymine (T) are found in DNA while RNA is made from adenine (A), guanine (G), cytosine (C), and uracil (U).

**Base pair (BP):** The complementary bases on opposite strands of DNA which are held together by hydrogen bonding. The atomic structure of these bases pre-select the pairing of adenine with thymine and the pairing of guanine with cytosine (or uracil in RNA).

**Cell:** The smallest functional structural unit of living matter. Cells are classed as either procaryotic or eucaryotic.

**cDNA (complementary DNA):** An artificial piece of DNA that is synthesized from an mRNA (messenger RNA) template and is created using reverse transcriptase. The single stranded form of cDNA is frequently used as a probe in the preparation of a physical map of a genome. cDNA is preferred for sequence analysis because the introns found in DNA are removed in translation from DNA ----> mRNA ----> cDNA.

**Chromosome:** A collection of DNA and protein which organizes the human genome. Each human cell contains 23 sets of chromosomes; 22 pairs of autosomes (non sex determining chromosomes) and one pair of sex determining chromosomes. The human

genome within the 23 sets of chromosomes is made of approximately 30,000 to 100,000 genes which are built from over 3 billion base pairs. While eukaryotic chromosomes are complex sets of proteins and DNA, prokaryotic chromosomal DNA is circular with the entire genome on a single chromosome.

**Complementarity:** The sequence-specific or shape-specific recognition that occurs when two or more molecules bind together. DNA forms double stranded helixes because the complementary orientation of the bases in each strand facilitates the formation of the hydrogen bonds that hold the strands together.

**Computational biology:** See bioinformatics

**Deoxyribose:** A five carbon sugar lacking a hydroxyl group on position 2 (beta-d-2-deoxyribose) which is used in the construction of DNA

**DNA (deoxyribonucleic acid):** A double stranded molecule made of a linear assembly of nucleotides (See Figure 3). DNA holds the genetic code for an organism in the arrangement of the bases. The double strand of DNA results from the hydrogen bonds formed between bases when two polynucleotide chains, identical, but running in opposite directions, associate.

**DNA polymerase:** The enzyme that assembles DNA into a double helix by adding complementary bases to a single strand of DNA. Linkages are formed by adding nucleotides at the 5' hydroxyl group to the phosphate group located on the 3' hydroxyl.

**Enzyme:** A protein which catalyzes (or speeds the rate of reaction for) biochemical processes, but which does not alter the nature or direction of the reaction.

**Eukaryote:** An organism whose genomic DNA is organized as multiple chromosomes within a separate organelle -- the cell nucleus.

**Exon:** The region of DNA that encodes proteins. These regions are usually found scattered throughout a given strand of DNA. During transcription of DNA to RNA, the separate exons are joined to form a continuous coding region.

**Gene:** A section of DNA at a specific position on a particular chromosome that specifies the amino acid sequence for a protein.

**GenBank:** The NIH genetic sequence database. An annotated collection of all publicly available DNA sequences which is located at <http://www.ncbi.nlm.nih.gov>. There are approximately 2,162,000,000 bases in 3,044,000 sequence records as of December 1998. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

**Gene expression:** The conversion of the information encoded in a gene to messenger RNA that is in turn converted to protein.

**Genome:** The total genetic material of a given organism.

**Genomics:** The mapping, sequencing, and analysis of an organism's genome.

**Hybrides (or hybride molecular complexes):** The formation of a complimentary complex between a probe molecule and a target molecule. This complex is generally tagged with a radioactive label on the probe molecule so that the complex can be located and isolated for further study. Hybrid molecular complexes of the type DNA-DNA, DNA-RNA, and Protein-Protein are frequently used in genetic analysis. Since hybridization reactions are specific, they can be used to locate one DNA, RNA, or protein molecule within complex mixtures of similar molecules.

**Hybridization:** The formation of a double stranded DNA, RNA, or DNA/RNA from two complementary oligonucleotide strands.

**Hydrogen bond:** A dipole-dipole attraction in which a hydrogen atom bridges two electronegative atoms. One half of the hydrogen bond is a covalent bond and the other is an electrostatic bond.

**Induction:** The switching of cells between pathways under the influence of an adjacent group of cells. It is possible to generate several different cells through a series of inductions between a limited number of cell types.

**Intron:** The portion of a DNA sequence which interrupts the protein coding sequences of the gene. Most introns begin with the nucleotides GT and end with the nucleotides AG.

**Kilobase (kb):** A length of DNA equal to 1,000 nucleotides.

**Microarray:** DNA that has been anchored to a chip as an array of microscopic dots, each one of which represents a gene. Messenger RNA that encodes for known proteins is added and will hybridize with its complementary DNA on the chip. The result will be a fluorescent signal indicating that the specific gene has been activated.

**Motifs:** A pattern of DNA sequence that is similar for genes of similar function. Also a pattern for protein primary structure (sequence motifs) and tertiary structure that is the same across proteins of similar families.

**mRNA (messenger RNA):** RNA that is used as the template for protein synthesis. The first codon in a messenger RNA sequence is almost always AUG

**NCBI:** The National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>), a division of the NIH, is the home of the BLAST and Entrez servers.

**Nucleotide (nt):** A molecule which contains three components: a sugar (deoxyribose in DNA, ribose in RNA), a phosphate group, and a heterocyclic base.

**Oligos (Oligonucleotides):** A chain of nucleotides.

**Operon:** The group of contiguous genes in a bacterial chromosome that are transcribed into an mRNA molecule.

**PCR (polymerase chain reaction; in vitro DNA amplification):** The laboratory technique for duplicating (or replicating) DNA using the bacterium *Thermus aquaticus*, a heat stable bacterium from the hot springs of Yellowstone. As with the polymerase reaction that occurs in cells, there are three stages of a PCR process: separation of the DNA double helix, addition of the primer to the section of the DNA strand which is to be copied, and synthesis of the new DNA. Since PCR is run in a single reaction vessel, the reactor contains all of the components necessary for replication: the target DNA, nucleotides, the primer, and the bacterial DNA polymerase. PCR is initiated by heating the reaction vessel to 90° which causes the DNA chains to separate. The temperature is lowered to 55° to allow the primers to bind to the section of the DNA that they were designed to recognize. Replication is then initiated by heating the vessel to 75°. The process is repeated until the quantity of new DNA desired is obtained. Thirty cycles of PCR can produce over 1 million copies of a target DNA.

**PDB (Protein Data Bank):** An international repository for the results of macromolecular studies using NMR, X-ray crystallography, or homology methods. The results of structural studies of proteins, RNA, DNA, viruses, and polysaccharides are presently available. The term PDB also defines a standard file format for publishing protein and nucleotide structures for use in computer programs.

**Peptide:** A small chain of amino acids (see protein).

**Polymerase:** The process of copying DNA in each chromosome during cell division. In the first step the two DNA chains of the double helix unwind and separate into separate strands. Each strand then serves as a template for the DNA polymerase to make a copy of each strand starting at the 3' end of the chain.

**Polypeptide:** A linear chain of amino acids joined head to tail via a peptide bond between the carboxylic acid group of one amino acid and the amino group of the next amino acid.

**Post translational modification:** Changes that occur to a protein after translation from mRNA. Modifications can include cleavage of a small number of residues, the addition of carbohydrates, phosphorylation of hydroxyl groups, acetylation, etc.

**Primer:** The short sequence of nucleotides (usually eight) which serve to prime the DNA polymerase process during cell division. Primers are produced by the enzyme primase. Primers also can be customized to 'isolate' specific sections of DNA for replication using PCR.

**Probe:** A radiolabeled or fluorescent oligonucleotide used to locate complementary sequences in a hybridization experiment.

**Prokaryote:** An organism whose DNA is not enclosed in a separate organelle.

**Promoter:** The short sequence on nucleotides on DNA that start the transcription of RNA by RNA polymerase.

**Protein:** A linear chain of variable length that is constructed from the 20 basic amino acids (also referred to as a peptide or as a polypeptide). The linear arrangement of the amino acids is known as the protein's primary structure. The local three-dimensional

arrangement (or folding pattern) of the main portion of the chain (the polypeptide backbone) is known as the protein's secondary structure. The overall three-dimensional arrangement of all atoms in a single chain in the protein is termed the protein's tertiary structure. The three dimensional shape, in conjunction with the chemical properties of the amino acids contained in the protein, determines the protein's function.

**Proteome:** The full complement of proteins produced by a particular genome.

**Proteomics:** The study of protein expression, structure, and function, and the interactions of all proteins of a specific organism.

**Reading frame (also open reading frame):** The stretch of triplet sequence of DNA that encodes a protein. The reading frame is designated by the initiation or start codon and is terminated by a stop codon. DNA (through RNA) uses a triplet code to specify the amino acid for a given protein. As can be seen above, a given strand of DNA has three possible starting points (position [or reading frame] one, two, or three). Since both strands of DNA can be translated into RNA and then into protein, a sequence of double helical DNA can specify six different reading frames.

**Recombinant DNA:** Partial strands of DNA from different sources that are joined outside of a cell.

**Recombination:** The exchange of regions of DNA on chromosomes via cross over during meiosis (see crossover).

**Regulatory region:** The segment of DNA that controls whether and to what degree, a gene will be expressed.

**Restriction enzyme:** A protein that recognizes specific sites on nucleotides or proteins and hydrolyzes the nucleotide or protein at these points.

**Ribonucleic acid (RNA):** Nucleotide made from a ribose, a base [adenine (A), guanine (G), cytosine (C), and uracil (U)], and a phosphate group. RNA is generally found in the cell nucleus or cytoplasm.

**Ribose:** A five carbon sugar (b-d-ribose) which is used in the construction of RNA.

**Ribosome:** Cellular components made of ribosomal RNA and proteins that are the site of protein synthesis (translation).

**Sequencing:** Determining the order of nucleotides in a gene or the order of amino acids in a protein.

**Structural genomics:** The prediction of the 3-D structure of proteins encoded by genes using both experimental and computational techniques.

**SWISS-PROT:** An annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library (now the EMBL Outstation - The European Bioinformatics Institute (EBI)). The SWISS-PROT protein sequence data bank consists of sequence entries. Sequence entries are composed of different line-types, each with their own format. For standardization purposes the format of SWISS-PROT follows as closely as possible that of the EMBL Nucleotide Sequence Database.

**TrEMBL:** The supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.

**Transcription:** The process of copying a strand of DNA to yield a complementary strand of RNA

**Transcription Factors:** The class of proteins that bind to DNA and promote or inhibit the initiation of transcription.



**Transfection:** Introduction of a foreign DNA molecule into a eucaryotic cell and subsequent expression of the genes of the new DNA.

**Transfer RNA (tRNA):** Specialized RNA that transfers single amino acids to a growing protein chain. tRNA has a complementary codon to the codon on the mRNA.

**Translation:** The process of sequentially converting the codons on mRNA into amino acids that are then linked to form a protein.