# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

# DESIGN OF AN ENZYME ACTIVITY MAPPING DATABASE

RONGHUA SHU

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
MONTRÉAL, QUÉBEC, CANADA

JANUARY 2003

# ABSTRACT

## Design of An Enzyme Activity Mapping Database

**Ronghua Shu**

This thesis presents the design of a relational database supporting high-throughput assays to determine enzyme catalytic activity – The Enzyme Activity Mapping Database (**EAMDB**). The database is composed of six modules: enzyme module, chemical compound module, enzyme activity mapping module, experiment module, data module, and reference module. Each module has one or more tables to cover relevant information. There are 17 tables in the database. Primary key and foreign key constraints have been introduced to enforce data integrity. The database can be integrated into a bioinformatics database system to provide enzyme function information to serve scientists for their various research interests.

# Acknowledgements

I would like to express my warmest gratitude to my supervisor, Dr. Gregory Butler, for his patience and invaluable guidance. His profound knowledge in computer science and bioinformatics is highly appreciated.

Dr. Justin Powlowski and Dr. Paul Joyce are gratefully acknowledged for their introduction to enzyme assays and for their discussions. Special thanks go to Dr. Justin Powlowski for his detailed correction and discussion of the requirements document.

I would like to thank all fellow students in Dr. Butler's group for their helpful discussions and friendship.

I dearly thank my parents and all my family members for their understanding and support during my long school years.

I dedicate this thesis to my beloved father who passed away during the preparation of this thesis.

# Contents

# List of Figures

# List of Tables

# Chapter 1. Introduction

## 1.1 Enzyme catalytic assay project

Enzymes are biological catalysts responsible for supporting almost all of the chemical reactions that maintain homeostasis of organisms. The versatility of enzyme catalysts has attracted interest in their application. Enzyme catalysts exhibit high selectivity (stereo-selectivity, regio-selectivity, and chemo-selectivity), high catalytic efficiency, work under mild reaction conditions, and are environmentally friendly.[1]

As part of the fungal genomics project, a group of enzymes will be assayed against substrates of interest with the hope of finding enzyme catalysts suitable for application in the pulp and paper industry, as environmental concerns have become a big hurdle to its development.[2] The Enzyme Activity Mapping Database (**EAMDB**) under design is to support the enzyme assay experiments of the project. The design of **EAMDB** is based on the requirements of the enzyme catalytic activity assay part of this project. **EAMDB** is also expected to be applicable to general enzyme assay experiments used to search for enzyme catalysts for chemical reactions. The information stored can be used in catalysis research and organic synthesis research.

## 1.2 Public enzymatic reaction databases

There are a few databases available on the web providing information about enzymatic reactions. **LIGAND**,[3] **BRENDA**,[4] and **UM-BBD**[5] are the main examples. The enzymes covered in **LIGAND** and **BRENDA** (**BR**aunschweiger **EN**zyme **DA**tabase)[4] are mainly in the category of metabolism central to functioning of cell life. **UM-BBD**[5] is a database specifically focusing on biodegradation pathway information.

The Ligand Chemical Database for Enzyme Reactions (**LIGAND**)[5] is designed to provide the linkage between chemical and biological aspects of life from the perspective of enzymatic reactions. **LIGAND** is part of the **KEGG**[6] metabolic pathway database. It is a composite database consisting of three sections: the ENZYME section, the COMPOUND section and the REACTION section. The ENZYME section is based on the enzyme nomenclature of the International Union of Biochemistry and Molecular Biology (**IUBMB**, 1992) and International Union of Pure and Applied Chemistry (**IUPAC**). The COMPOUND section is a collection of metabolic compounds, including substrates, products, inhibitors, cofactors and effectors, and other chemical compounds that play important functional roles in living cells. These compounds are also found in the KEGG/PATHWAY database and in the ENZYME section, as are other compounds found in the literature. The REACTION section is a collection of reactions, mostly enzymatic reactions, involving the compounds covered in the COMPOUND section.

**UM-BBD**[5] (The University of Minnesota Biocatalysis/Biodegradation Database) is an online compilation of microbial catabolic enzymes, reactions, and pathways for primarily synthetic organic chemical compounds. This information is directly applicable toward enhancing the understanding of biocatalysis leading to specialty chemical manufacture and the biodegradation of environmental pollutants. Unlike the other pathway databases, where the focus is primarily on intermediary metabolism, the **UM-BBD** is a key resource for biodegradation pathway information and is recently evolving to include the prediction of specialized catabolic routes for new compounds.

**BRENDA** is a collection of enzyme functional data available to the scientific community. It is a comprehensive relational database of functional and molecular

information about enzymes, based on primary literature. All data and information are manually extracted and evaluated from the primary literature by scientists. A major part of **BRENDA** is the information about ligands. These ligands function as natural or *in vitro* substrates/products, inhibitors, activating compounds, cofactors, bound metal, etc. **BRENDA** stores about approximately 320,000 enzyme-ligand relationships with more than 33,000 different chemical compounds functioning as 'ligand'. The ligands are stored as compound names, SMILES (Simplified Molecular Input Line Entry System)[7] strings and Molfiles. The two-dimensional chemical structure of these compounds can be displayed as images. It is claimed to be an important tool for biochemical and medical research covering information on properties of all classified enzymes, kinetics, substrates/products, inhibitors, cofactors, activators, structure and stability.

The data and information in **BRENDA** are stored in 52 tables containing approximately 460,000 entries directly from the primary literature in a relational database system to enable different search features. Enzymes can be searched by their EC numbers, their names or synonyms, or by the organisms from which the enzyme was isolated. All other information fields can be searched individually or by combination searches, which can be organism-specific.

Information on some enzymes and their associated human diseases has been included in the **BRENDA** database.[4] Based on data from **BRENDA**, the calculation and simulation of metabolic pathways can be performed by using the information of substrate/product chains and the corresponding kinetic data of the preceding and following enzymes in **KEGG** metabolism pathway.

### 1.3 Why a new enzyme catalytic database of our own

Compared with the databases discussed above and other web accessible bioinformatics databases, our proposed database is more experimental information oriented, *i.e.* the detailed experimental information will be recorded along with the information about the enzymes, chemical compounds, and reactions. Detailed experimental information is rarely made available when the results are of high market value. People normally would patent the procedures to protect their intellectual property rights. One has to keep careful track of experimental information so that it may be used in a patent application. The same practice would also be applied to the results obtained by the scientists at Concordia. If the enzyme assays reveal enzymes with new and useful properties, they could potentially be applied to the pulp and paper industry, and the results would be patented. The information stored in the database will also be a useful resource for the research of scientists in related scientific fields.

Another major difference between **BRENDA** or **UM-BBD** and **EAMDB** is that **BRENDA** and other databases are heavily dependent on literature and other public information sources while **EAMDB** would mainly contain the information about the research results of our scientists regarding the catalytic activities of enzymes. The data in **EAMDB** would include both positive and negative results. Information about experiments yielding positive results can be used to guide further research or even industrial production. Information about experiments with negative conclusions can help scientists to avoid some trials doomed to failure.

**EAMDB** will also store some information extracted from public information sources to serve our own research needs as well as the public interests. Excepting

confidential experimental details, the **EAMDB** system will also be made accessible to the public. Adequate information will be provided to interested users about the catalytic activity mapping between enzymes and chemical compounds. Of course for internal users, the corresponding details about how the conclusions have been reached will be accessible.

### 1.4 Contribution of the thesis

Based on the requirements of enzyme catalytic assay research, the thesis presents a database design to store all the necessary information about enzyme catalytic assay experimental details and conclusions. The information stored will cover chemical information, chemical reaction, enzyme information, and enzyme catalytic activity information. The major difference between this database and other public enzymatic reaction database resources is that our system is experiment oriented. Biochemists can record all detailed experimental information, experimental data, and conclusions. Scientists can not only get the information about the catalytic activity of enzymes but also get the information about the related experiments. The database system can be integrated into a bioinformatics system covering the functional information of enzymes. The system can provide valuable information to chemists, biochemists, and bioinformaticians.

### 1.5 Organization of the thesis

There are four chapters and one appendix in the thesis discussing the justification, backgrounds, database design and conclusions respectively: **Chapter 1** presents the justification of a new database. **Chapter 2** presents the background knowledge of

bioinformatics, and enzyme catalytic assays. **Chapter 3** presents the detailed design of the database. **Chapter 4** presents the concluding remarks of the thesis. Molecular biology databases discussed in the thesis are listed in the **Appendix**.

# Chapter 2. Background

## 2.1 Brief review of Bioinformatics

Bioinformatics is the application of information technology to store, organize and analyze the vast amount of biological data that is available in the form of sequences and structures of proteins – the building blocks of organisms. It is a multidisciplinary field, which encompasses molecular biology, biochemistry and genetics on the one hand, and computer science on the other. Bioinformatics uses methods from various areas of computer science, such as algorithms, combinatorial optimization, integer linear programming, constraint programming, formal language theory, neural nets, machine learning, pattern recognition, inductive logic programming, database systems, knowledge discovery and data mining. The exponential growth in biological data, generated from national and international genome projects, offers a remarkable opportunity for the application of modern computer science. The fusion of life science and computer technology offers substantial benefits to all scientists involved.

### 2.1.1 Scientific problems for bioinformatics

The scientific problems for bioinformatics are sequencing support, analysis of nucleic acid and protein sequences, analysis and prediction of molecular structure (**DNA**, **RNA**, protein, lipids, and carbohydrates), molecular interactions (protein-ligand, protein-protein, protein-**DNA** etc.), and metabolic and regulatory networks.

The aim of sequencing support of bioinformatics is to interpret experimental data that are generated by sequencing efforts. There are three challenges in sequencing support: base calling, physical mapping, and fragment assembly. Base calling is the interpretation of the signals output by sequencers in terms of nucleic acid sequence. Physical mapping is to provide a rough map of relevant loci along the genome. Fragment assembly is the process of piecing together short segments of sequenced **DNA** to form a contiguous sequence of the genome or chromosome. Sequencing support plays a special role, as it is a scientific problem of bioinformatics and an application scenario of bioinformatics. Sequencing support provides the raw genomic sequence, which is the base for further bioinformatics work.

Analysis of nucleic acid sequences is concerned with annotating the raw genomic sequence with information that can be derived directly from the sequence. The problems involved are gene finding and the analysis of non-coding regions. Gene finding is the identification of those stretches of genomic **DNA** that code for protein. This is the entry to understanding the proteome. Analysis of non-coding regions currently concentrates on the upstream of sequences that encode proteins. These regions contain patterns that govern the regulation of the expression of the genes, *i.e.* their translation to proteins. Gene finding and analysis of non-coding regions are considered basic research problems and scientific grand challenges.

Analysis of protein sequences annotates protein sequences. The most voluminous source of information for the analysis is the comparison with other homologous sequences, either on the level of protein or the level of nucleic acid. The results desired

are relationships between different proteins that allow scientists to make conclusions about protein function, cellular localization and the like.

As the structure of a molecule is the key to its function, modeling molecular structures is another very important part of bioinformatics. The molecules studied are **DNA**, **RNA**, proteins, lipids, and carbohydrates.

The well-known double helix structure of **DNA** is very well preserved throughout nature. The small differences in the fine structure of **DNA** can be modeled in the computer by using methods like energy minimization and molecular dynamics.

Compared with **DNA**, **RNA** is structurally more flexible. As a matter of fact, modeling three-dimensional structures of **RNA** basically remains a challenge for bioinformatics.

Proteins display a wide variety of structures. Great efforts have been made in the analysis and prediction of protein structures. Though many successes have been accomplished, a long journey is still ahead of bioinformatics to solve the general problem. What can be achieved is to model a (target) protein given a structurally resolved protein that acts as a template. If sufficient similarity exists between the target protein and template protein, a successful result can be expected; otherwise an accurate full-atom model of the protein cannot be generated. However, often one can still find many significant aspects of the structure of the protein, such as the overall architecture, the coordinates of the protein backbone, or even more accurate models of relevant active sites.

The analysis and prediction of molecular structures of lipids and carbohydrates is another subject of bioinformatics. Lipids form membranes inside and around the cell.

Carbohydrates form complex tree-like molecules that become attached to the surface of proteins and cellular membranes. Their three-dimensional molecular structures are not unique, but the molecular assemblies are highly flexible. Thus analyzing the molecular structure involves the inspection of a process in time. So far the analysis and prediction of the structures of lipids and carbohydrates has revealed relatively few results compared with protein structures.

Study of molecular interactions, such as protein-ligand, protein-protein, protein-DNA, DNA-ligand, is one of the major subjects of bioinformatics. As these interactions are essential for living organisms, the study is of great importance to bio-scientists.

In protein-ligand interactions, one molecular partner is a protein and the other is a small, often flexible, organic molecule. The issue is a basic research subject of bioinformatics and, at the same time, is of prime importance for the research of applications where the docking between a protein molecule and small molecules is a prerequisite, such as drug design and enzyme catalysis. The study of protein-ligand docking has two aspects: determine the correct geometry of the molecular complex, and provide an accurate estimate of the differential free energy of binding. Whereas much progress has been made on the first aspect, the second aspect remains a tremendous challenge.

Protein-protein docking is different from protein-ligand docking in several aspects. In protein-ligand docking, the binding mode is mostly determined by strong enthalpic forces between the protein and the ligand. In addition, the contribution of desolvation (replacing the water molecules inside the pocket by the ligand) are essential. The notion of molecular surface is not as relevant, especially since the ligand can be

highly flexible and does not have a unique surface. In contrast, genomic complementary of both proteins is a dominating issue in protein-protein docking, where both partners meet over a much larger contact surface area. Issues of desolvation can be essential here. Induced fit, *i.e.* subtle structural change on the protein surface to accommodate binding, is important in both problems.

Other molecular interactions, such as protein-**DNA** and reactions involving RNA or lipids, are also important subjects of bioinformatics.

Metabolic pathways and regulatory networks are another major subjects of bioinformatics. It focuses on biological interaction networks. It integrates the information about genes and proteins generated by genome sequencing, functional genomics, and proteomics experiments with the computerized reference knowledge on molecular interaction networks, *i.e.* it has two aspects: database aspect and algorithmic aspect. Database aspect collects the voluminous data and makes them generally accessible. Algorithmic aspect performs simulations on the resulting networks. Both aspects are in a preliminary stage. The most development has taken place in metabolic databases.

### 2.1.2 Bioinformatics databases

With the development of bioinformatics, many databases, along with applications and other molecular biology resources, have given rise to the need for bioinformatics solutions. Currently there are about 335 molecular databases of value to the biological community.[8] Some of the databases are freely available, such as the DNA sequence collections **EMBL**[9] and **GENEBANK,**[10] while others are only freely available to the

academic community, such as the protein sequence database **SWISS-PROT**;[11] and others are only available on subscription. Academics and pharmaceutical companies also have their own proprietary data which must be integrated into a system so that relationships with publicly available data can be found.

Bioinformatics databases can be divided into sequence databases (*e.g.* **EMBL**, **NCBL**, **DDBJ**, and **GENBANK**), sequence related databases (*e.g.* **PDB**, **DSSP**, and **HSSP**), genome databases (*e.g.* Genome Databank), pathway and chemical compounds, and others.

Enzymes that are involved in a large number of reactions are catalogued in databases such as **ENZYME**, **BRENDA**, **LIGAND** etc. Each enzyme with known enzyme function are catalogued and named by a nomenclature committee. Also included in these databases is information on the reaction and specificity of the enzyme and the various conditions under which the enzyme will be active.

### 2.1.3 Applications for Bioinformatics

The exponentially growing biological data initiated the development of bioinformatics. To solve the scientific challenges elaborated above, bioinformatics has posed three aims since its beginning. First, it organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced, *e.g.* the Protein Data Bank.[12, 13] The second aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare with previously characterized sequences. These programs, such as **FASTA**[14] and **PSI-BLAST**,[15] must consider what comprises a biologically significant match. The

third aim of bioinformatics is to use these tools to analyze the data and interpret the results in a biologically meaningful manner. With the help of bioinformatics, it is possible to globally analyze all the available data to uncover common principles that apply across many systems and highlight novel features. With traditional biological studies, only a few related systems may be compared.

Bioinformatics has some practical applications in the biological sciences including finding homologues, and large-scale censuses. Finding homologues not only enables systematic organization of data, but also helps characterization of proteins. Large-scale censuses condense all the information related to genomes, structures and expression datasets. Through large-scale censuses, broad generalizations help identify interesting subject areas for further detailed analysis, and place new observations in a proper context.

Another important application of bioinformatics is aiding rational drug design. The identification of a disease-related protein provides great guidance to the search for an effective drug. Basically there two approaches to development a new drug. One is to create a drug from scratch. This approach is based on the knowledge of the structure of the binding site of the protein. Due to the fact that often the developed molecules were hard to synthesize and it was hard to optimize the drug lead, this approach is problematic. The second approach is to screen through a large database of known molecules and check their binding affinity to the target protein. The advantage of this approach is that most of the properties of the compounds in the database such as bio-accessibility and toxicity have been studied.

Enzyme structural and functional information in bioinformatics system can be applied to other research areas with environmental degradation and chemical synthesis (with drug design at its core) as the prominent ones.[17] Enzymes are proteins that enable thousands of essential chemical reactions in the cells. Function assignment to enzymes is an indispensable part of a bioinformatics system. It is becoming more and more essential with the ongoing development and progress on structural and functional genomics. A complete information system on enzymes (including structural and functional information) is an important tool in the field of bioinformatics to understand biological functions and biochemistry. It is also necessary for the simulation and construction of whole metabolic pathways and networks. Due to the environmental friendliness and special versatile catalytic properties of enzymes, their application in chemical and biochemical fields has attracted much interest.

## 2.2 Enzyme catalysts and enzyme kinetics

### 2.2.1 Enzyme catalysts

A catalyst is a substance that increases the rate of a chemical reaction by reducing the activation energy, but which is left unchanged by the reaction. Activation energy **Ea** is needed for a reaction to take place. The lower the **Ea** the easier it is for the reaction to occur. **Figure 2.1** shows a simple comparison of the activation energies of a reaction in the presence of a catalyst and in the absence of a catalyst. Much lower activation energy is expected in the presence of a catalyst thus a much higher reaction rate. Catalysts are important to many industrial processes. Without catalysts, some reactions cannot take place spontaneously or may be too slow to be industrially useful.

Enzymes are biological catalysts.[16] The most striking characteristics of enzymes are their catalytic power and specificity. They are proteins produced or derived from living organisms. Enzymes are very specific in nature. Each enzyme can act to catalyze only very select chemical reactions and only with very select substances. All enzyme-catalyzed reactions have at least three steps (**Figure 2.2**):

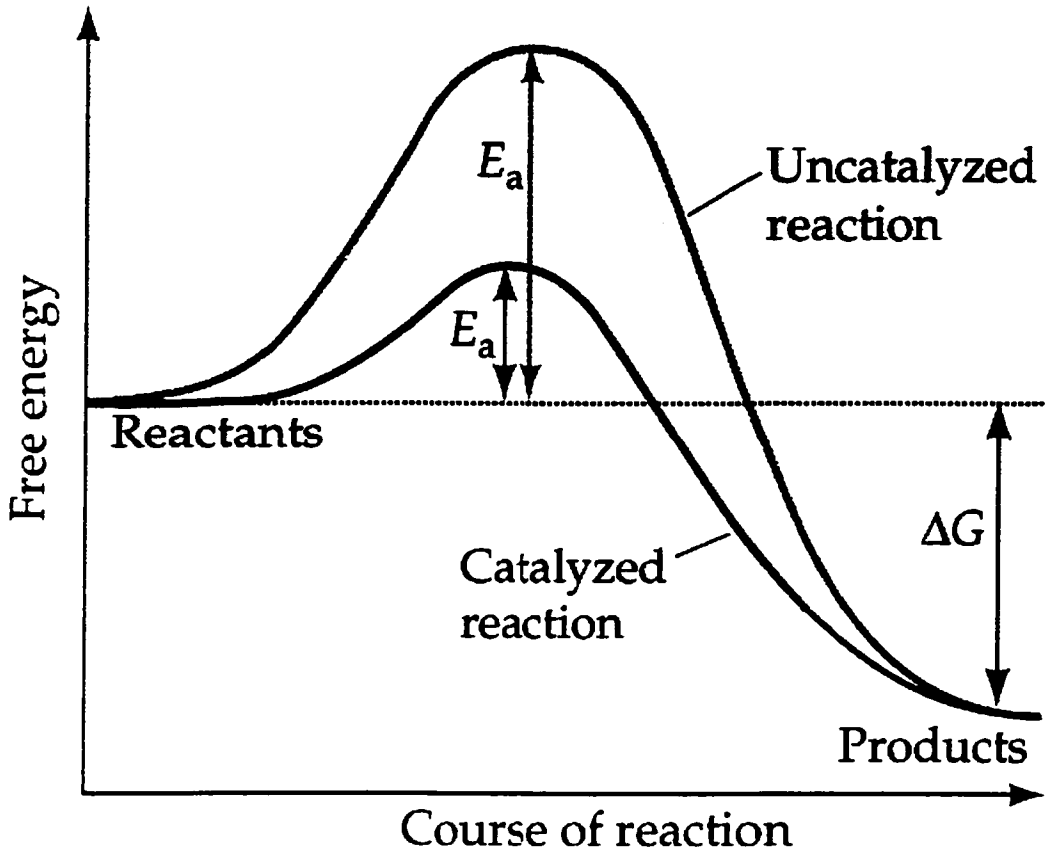


**Figure 2.1 Activation Energy Comparison Between Catalyzed and Uncatalyzed Reaction.**

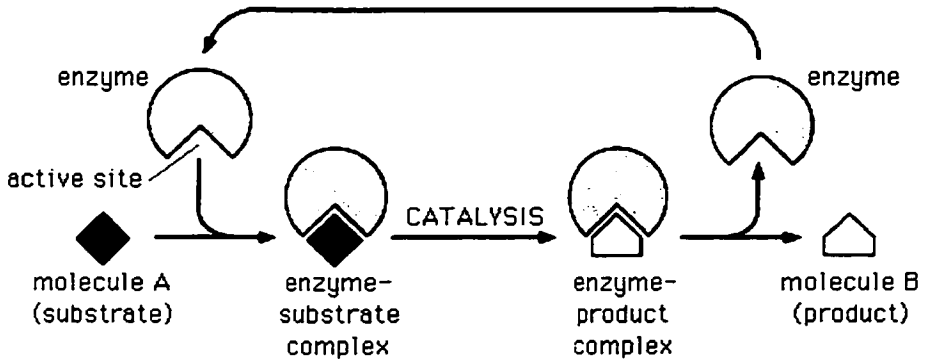


**Figure 2.2 Catalytic scheme of enzyme**

(a) each enzyme has an active site to which one or two substrate molecules bind, forming an enzyme-substrate complex (step 1);

(b) a catalyzed reaction occurs at the active site, forming an enzyme-product complex (step 2)

(c) the product is then released (step 3), allowing the enzyme to bind additional substrate molecules.

Enzymes are thus unchanged after participating in reactions (and therefore are able to catalyze the reaction over and over again).

Compared with other chemical catalysts, enzyme catalysts have their own unique properties:

(a) Enzyme-catalyzed reactions have higher reaction rates: $10^6 - 10^{12}$ higher than uncatalyzed reactions and several orders of magnitude higher than chemically catalyzed reactions.

(b) Enzyme-catalyzed reactions occur under mild reaction conditions (e.g. temperature, pressure, pH, and in aqueous solutions)

(c) Enzymes have greater reaction specificities than chemical catalysts. The binding of the enzyme and its substrate(s) is highly selective.

Enzyme active sites are usually formed by a surface indentation or cleft that is complementary in shape to the substrate. This maximizes the number of non-covalent interactions that can occur between the enzyme and substrate molecule(s). Thus, the chemical characteristics of the specific amino acids that comprise the binding site are a major determinant of enzyme specificity (e.g. geometric specificity). Enzymes are also stereo-specific (e.g. they have much higher reaction rates with one configuration versus the other) and region-specific (e.g. they have higher reaction rates with functional groups at a specific position). The structure of most enzyme binding sites is preformed (lock and key fit), although some binding sites assume their final structure following substrate binding (an induced fit). Many enzymes that carry out certain types of reactions (such as oxidation/reduction and group-transfer reactions) require the association and assistance of co-factors,[16] which are obtained in the diet and which include: metal ions, such as $Cu^{2+}$, $Fe^{3+}$ and $Zn^{2+}$; coenzymes (small organic molecules). Coenzymes, such as thiamine pyrophosphate (**TPP**, **Figure 2.3** in gray), are small organic molecules that bind to an enzyme's surface and help catalyze specific reactions.

Various substances can reduce enzyme catalytic activity. These substances are inhibitors of enzymes. They slow down the rate of enzyme-catalyzed reactions, generally by interacting specifically with the enzyme's active site in such a way as to reduce access to the active site by the substrate. Inhibition is a major research subject for bio-scientists. They are harmful for the application of enzymes as catalysts, but on the other hand they may be effective drugs to cure some diseases. [18]

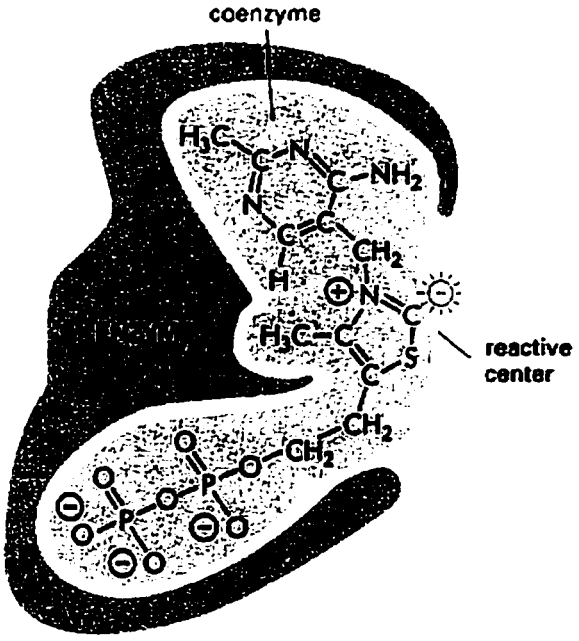


**Figure 2.3 TPP acting as a coenzyme**

### 2.2.2 Enzyme kinetics

Enzyme kinetics studies the rate of an enzyme catalyzed reaction under various conditions. These conditions may include variations in pH, temperature, and substrate concentrations. In enzyme kinetics, one of the most important and widely used chemical reaction models is the Michaelis-Menten model.[16] This model is a general explanation of the kinetics and gross mechanism of enzyme-catalyzed reactions. First stated in 1913, the hypothesis assumes that a complex is formed between an enzyme and its substrate, which complex then decomposes to yield free enzyme and the reaction product: the latter rate determines the overall rate of substrate-product conversion. The velocity of such a reaction is greatest when all the sites at which catalytic activity can take place on the enzyme molecules (active sites) are filled with substrate; i.e., enzyme is saturated with substrate. These relationships provide the basis for many kinetic studies of enzymes and also have been applied to investigations of the effects of carriers upon the transport of

substances through cell membranes. In terms of the chemical reaction equation, the Michaelis-Menten model for enzyme kinetics is expressed as:

$$\mathbf{E} + \mathbf{S} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathbf{ES} \xrightarrow{k_2} \mathbf{E} + \mathbf{products} \qquad \text{Eq. 2.1}$$

where E is the enzyme, S is the "substrate" (the molecule on which the enzyme does its work), and ES is an enzyme-substrate complex. (It is thus presumed that the substrate must somehow bind to the enzyme before the enzyme can do its work.)

The reaction rate is defined as the rate of formation of the product. The kinetic equation implied by this mechanism can be expressed as **Eq.2.2**:

$$\frac{d[\mathbf{product}]}{dt} = k_2[\mathbf{ES}] \qquad \text{Eq. 2.2}$$

The enzyme-substrate complex, ES, is a transient species. By applying the steady state approximation,[19] following equation can be set up for the change rate of ES,

$$\frac{d[\mathbf{ES}]}{dt} = k_1[\mathbf{E}][\mathbf{S}] - k_{-1}[\mathbf{ES}] - k_2[\mathbf{ES}] \approx 0 \qquad \text{Eq. 2.3}$$

Solve for [ES],

$$[\mathbf{ES}] = \frac{k_1[\mathbf{E}][\mathbf{S}]}{k_{-1} + k_2} \qquad \text{Eq. 2.4}$$

and substitute it into the **Eq. 2.2** for the rate,

$$\mathbf{Rate} = \frac{d[\mathbf{product}]}{dt} = k_2 \frac{k_1[\mathbf{E}][\mathbf{S}]}{k_{-1} + k_2} \qquad \text{Eq. 2.5}$$

The Michaelis-Menten constant, $K_M$, is defined by following equation:

$$\frac{1}{K_M} = \frac{k_1}{k_{-1} + k_2} \qquad \text{Eq. 2.6}$$

so that the rate becomes,

$$\text{Rate} = \frac{k_2}{K_M}[E][S] \qquad \text{Eq. 2.7}$$

In the above equation, [E] is the concentration of free (uncomplexed) enzyme and this is usually not known. What is known is the total enzyme concentration, $[E]_o$, but

$$[E]_o = [E] + [ES] = [E] + \frac{[E][S]}{K_M}$$

$$= [E]\left(1 + \frac{[S]}{K_M}\right) \qquad \text{Eq. 2.8}$$

from which we obtain,

$$[E] = \frac{[E]_o}{\left(1 + \frac{[S]}{K_M}\right)} \qquad \text{Eq. 2.9}$$

The rate becomes, then,

$$\text{Rate} = \frac{k_2}{K_M}[S]\frac{[E]_o}{\left(1 + \frac{[S]}{K_M}\right)}$$

$$= k_2 \frac{[E]_o[S]}{K_M + [S]} \qquad \text{Eq. 2.10}$$

Define the reaction velocity as $v$ = Rate. So,

$$v = k_2 \frac{[E]_o[S]}{K_M + [S]} \qquad \text{Eq. 2.11}$$

Note that the reaction velocity, $v$, is zero when [S] is zero and that the reaction velocity increases as we increase [S]. The reaction velocity reaches a maximum when [S] becomes very large relative to $K_M$. Defining the maximum velocity, $v_{max}$, as,

$$v_{max} = \lim_{[S]\to\infty} k_2 \frac{[E]_o[S]}{K_M + [S]} = k_2[E]_o \qquad \text{Eq. 2.12}$$

then

$$v = \frac{v_{max}[S]}{K_M + [S]} \qquad \text{Eq. 2.13}$$

Note that the kinetics of the reaction are characterized by two parameters, $v_{max}$ and $K_M$. These are the parameters that are pursued by biochemists and are usually given in the literature in studies of the kinetics of biochemical reactions.

The lineweaver-Burk equation, **Eq. 2.13**, is often used to deal with experimental data,

$$\frac{1}{v} = \frac{K_M}{v_{max}[S]} + \frac{1}{v_{max}} \qquad \text{Eq. 2.14}$$

In an experiment one measures $v$ as a function of [S]. The plot of $1/v$ against $1/[S]$ (the Lineweaver-Burk plot) should give a straight line with slope, $K_M/v_{max}$ and intercept $1/v_{max}$. This gives us both parameters,

$$K_M = \text{slope} \times v_{max} = \frac{\text{slope}}{\text{intercept}} \qquad \text{Eq. 2.15}$$

From the definition of $K_M$, in the cases of when $k_{-1} >> k_2$ we can obtain

$K_M = k_{-1}/k_1$ = ES dissociation constant

In other words, $K_M$, is a measure of the affinity of the enzyme for substrate. A low $K_M$ indicates tight binding of the substrate by the enzyme, whereas a high $K_M$ indicates weak binding. Note however, that the $K_M$ is only a measure of the strength of substrate binding in the special case where $k_{-1} >> k_2$. In such cases it is a useful measure of the relative strengths of the binding of an enzyme for different substrates. This might be very useful information if one is trying to design an inhibitor for example.

Michaelis-Menten constants have been determined for many of the commonly used enzymes. The size of $K_M$ tells us several things about a particular enzyme.

(1) A small Km indicates that the enzyme requires a small amount of substrate to become saturated. Hence the maximum velocity is reached at relatively low substrate concentrations.

(2) A large Km indicates the need for high substrate concentration to achieve maximum reaction velocity.

The substrate with the lowest $K_M$ upon which the enzyme acts as a catalyst is frequently assumed to be enzyme's natural inhibitor, though this is not true for all enzymes.

The assumptions for the Michaelis-Menten mechanism are not always true for enzyme-catalyzed reactions. The case discussed above is a very simplified one. The actual reactions might be more complicated.[9]

### 2.2.3 High throughput enzyme assay

High throughput enzyme assay is the process of screening large numbers of compounds for binding activity or catalytic activity against target enzymes (or vice versa) rapidly and in parallel. The study includes the determination of optimal reaction conditions.

To determine suitable conditions for the reaction of an enzyme and a substrate is by no means trivial. One must not only screen the available enzyme libraries, but also search through various conditions to find out the optimal ones, such as concentrations (substrate and enzyme), temperature, pH, and buffer. It is likely that several hundreds of experiments might be involved to determine the suitable reaction conditions for a certain enzyme and substrate, with several hundreds or even thousands of samples analyzed each day. Some experimental approaches and analytical methods such as **HPLC** (High Performance Liquid Chromatography) and **NMR** (Nuclear Magnetic Resonance) are impractical, too expensive, or too slow. Fortunately, high throughput assays based on simple chromogenic or fluorogenic tests may be used for many enzyme assays.

Among the widely-used high throughput techniques for the important enzyme assay process, microtiter plate-based assays are among the most applied ones. The main principle behind the technique is to study the kinetics of the enzyme catalyzed reaction by measuring the change of a chromogenic or fluorogenic property in the reaction process. The chromogenic or fluorogenic property can be either substrate-based or product-based. From the change in chromogenic or fluorogenic property over time, the rate of change in concentration of the substrate and/or product can be determined. In a typical procedure, reaction samples are added to the wells of a 96-well microtitre-plate (**Figure 2.4**). The reaction samples are prepared according to the specific experimental purpose. e.g., if the experiment is designed to determine the effect of the substrate concentration on the reaction velocity, the reaction samples should be prepared with different concentrations of the substrate while keeping the other reaction conditions the same. The microtiter-plate with samples is then placed in a sample chamber of a plate reader. The samples are then

automatically mixed and scanned periodically for changes in absorbance or fluorescence. The readings are recorded and exported in various formats, e.g. Excel file, for data processing.

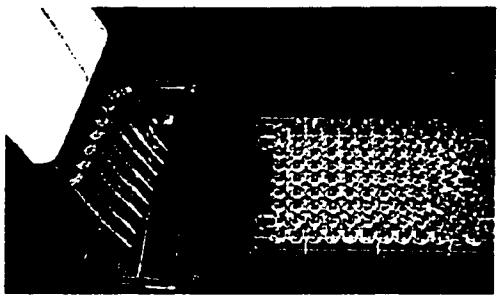

**Figure 2.4. Transfer Reaction Solutions to 96-well Micro-Plate**

Catalytic efficiency and turnover number are the two useful kinetic parameters to characterize the catalytic activity of enzymes.[16] To measure these two parameters, the $V_{max}$ and $K_M$ are determined for each enzyme-sustrate reaction by applying the Michaelies-Menten equation discussed above. The relationship between $k_{cat}$, $V_{max}$, and $K_M$ can be expressed by the following equations:

$$k_{cat} = Turnover\ Number = \frac{V_{max}}{[E]}$$

$$Catalytic\ Efficiency = \frac{k_{cat}}{K_M}$$

In most cases, an enzyme converts one chemical, the substrate, into one or more. A graph of product vs. time follows three phases as shown in the following graph (or simply a graph of signal intensity vs. time) (**Figure 2.5**). At the very early time points,

the rate of product accumulation increases over time. This transition phase usually lasts less than a second (in **Figure 2.5**, the first phase is greatly exaggerated and represents what happens in the pre-steady state). For an extended period of time, the product concentration increases linearly with time. At later times, the substrate is depleted, so the curve starts to level off (In some cases, the accumulation of product may lead to inhibition, and enzyme instability in the assay may also cause the rate to decrease). Eventually the concentration of product reaches a plateau and does not change with time. The second phase is the part in which we are most interested. Due to the linearity of the graph during this stage, the slope of the curve is easy to determine. The slope represents the initial reaction rate (Vo). By varying the experimental conditions, we can determine the effects of various assay conditions on the reaction rate.

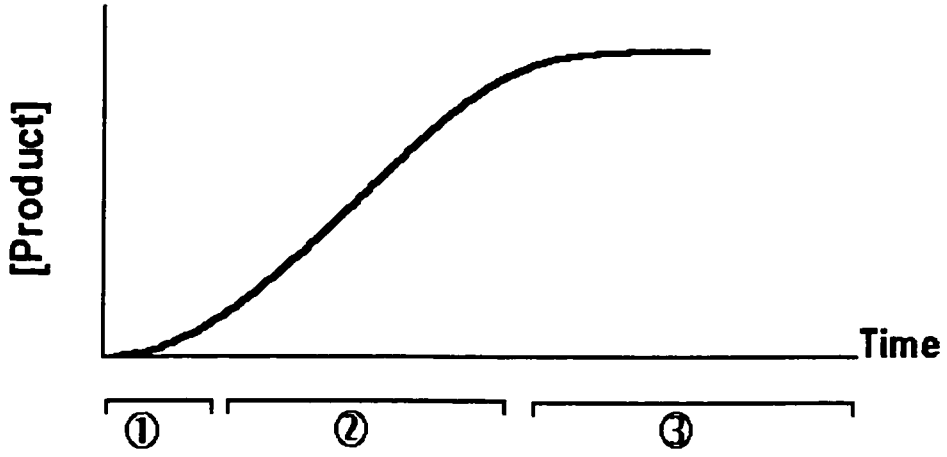


**Figure 2.5 Profile of Product Concentration vs. Time**

**Figure 2.6** shows the effect of substrate concentration on the initial reaction rate. At first the initial reaction rate increases as the substrate concentration increases according to the Michaelis-Menten equation. Gradually the slope of the curve decreases until nearly 0 where the initial reaction rate approaches $V_{max}$.

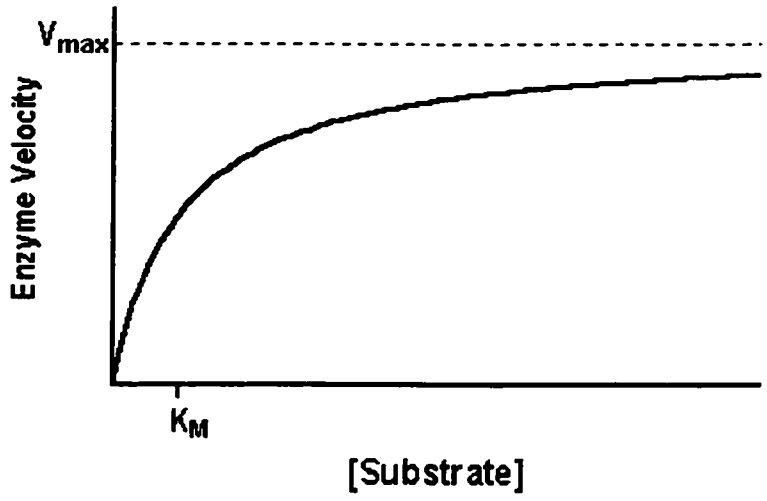


**Figure 2.6 Reaction Rate vs. Substrate Concentration**

Vmax and $K_M$ give the information about whether an enzyme is active for a reaction and how active the enzyme is. This is determined by the properties of the enzyme and substrate. However, assay conditions such as pH, and temperature can also play important roles. The catalytic activity of the enzyme may be dramatically changed at different pHs or temperatures. Therefore the study of the effect of pH and temperature is also an integral part of a complete enzyme kinetics. **Figure 2.7** shows an example of the effect of pH on the initial rate of an enzyme-catalyzed reaction. As shown in **Figure 2.7**, the rates of enzyme-catalyzed reactions vary with pH and often pass through a maximum as the pH is varied. The pH at which the rate is a maximum is called the pH optimum. Similarly, temperature also has significant effects on the catalytic activity of an enzyme. An enzyme might be active for a reaction at suitable temperatures but may show no activity beyond this temperature range.

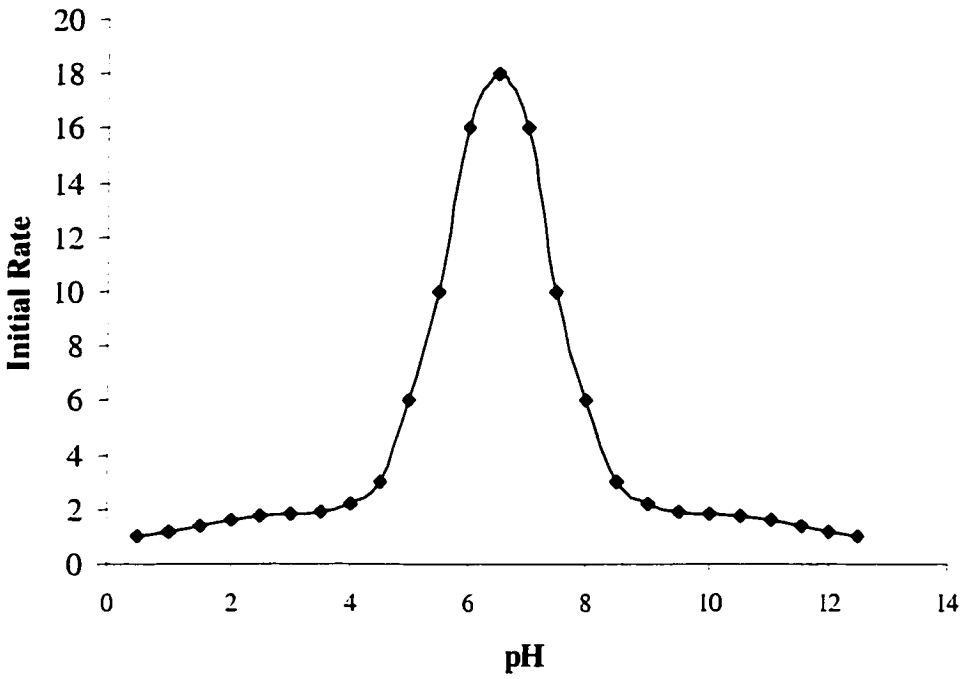


**Figure 2.7 Scheme of effect of pH on Initial Reaction Rate**

Factors such as pH, temperature, or even substrate itself may sometimes have great effect on the stability of an enzyme. An enzyme may be very active for a reaction of a substrate at certain pH, but it is also possible for it to be inactivated rapidly at the same time. Similarly temperature can exert similar effects on an enzyme: make it more active but also less stable. Temperature extremes, particularly high temperature used in industrial processes, often rapidly inactivate enzymes. Therefore characterization of enzyme stability is an important part of the enzyme catalytic study.

## 2.3 Database for enzyme assay

### 2.3.1 Database availability

Progress in bioinformatics has been accompanied by the appearance of many publicly available databases covering various biological information.[8] However, there is

no database to support enzyme assay experiments that is publicly available or accessible. This may be due to the nature of the information. Experimental detail information sometimes is considered sensitive especially when intellectual property rights are involved. While recording experimental information is gaining in popularity, each company's academic unit would develop their own information system, including databases, according to its needs and scientific orientation.

### 2.3.2 Database design

Relational, object-relational, and object-oriented data models are the three major data architectures that are current contenders for the attention of database designers. Among these three data architectures, the relational data model is much better developed although there are limitations.[21]

The design of a database normally consists of four major activities:[21] gathering requirements, modeling requirements with use cases, testing the system, and building data models. Gathering requirements is to find what the end users need. Modeling requirements with use cases is to analyze the requirements rigorously. Testing the system is to verify the requirements. Building data models is to transform the user requirements into data models.

# Chapter 3 Enzyme Catalytic Activity Mapping Database

## 3.1 Enzyme catalytic assay

Enzyme assays are important processes used to understand biochemical pathways, to give information about the enzyme catalytic properties, and to identify potential pharmacophores and inhibitors. The essential goal of these experiments is the measurement of the kinetic parameters ($k_{cat}$, $K_m$, $K_i$). These enzyme kinetic parameters can be evaluated using assay data obtained chromogenically, fluorogenically, electronically, or calorimetrically depending on the detectable property changes.

Enzyme catalytic assay is comprised of four major steps (**Figure 3.1**): (a) Enzyme preparation; (b) Chemical preparation; (c) Assay reaction operation and data collection; (d) Data processing.

### 3.1.1 Enzyme Preparation

Enzyme preparation includes enzyme purification, stability study, inhibition study, and solution preparation.

There are lots of enzyme purification protocols including column chromatography, solvent extraction, and re-crystallization etc. The main goal of the purification is to remove the contaminants that may deteriorate the stability of the enzyme or may be incompatible with further experimental procedures. This step is usually accompanied by catalytic assay experiments. The impurities present in an enzyme sample can be co-enzymes, co-factors, activators, inhibitor, or just an impurity that has nothing to do with catalytic activity of the enzyme. Therefore the purification sometimes enhances the catalytic activity of the enzyme, but sometimes may show no effect on the

enzyme catalytic activity. The worst case is that the purification achieves the very opposite, so that the purified enzyme is much less active or simply not active at all. To determine the effect of the purification procedure on the enzyme activity, assay experiments need to be performed. The result of the assay then serves as a guide for further purification.

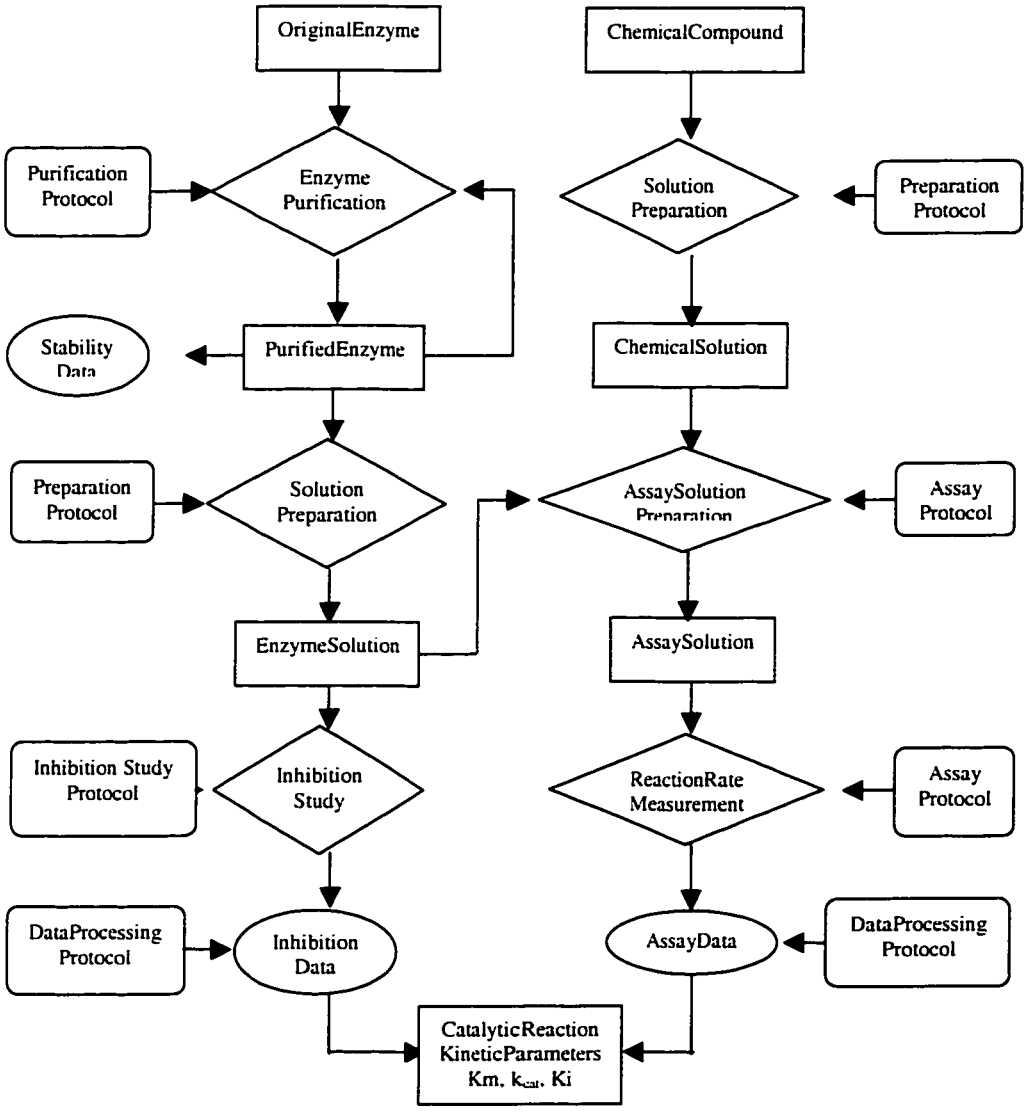


**Figure 3.1 Enzyme catalytic assay workflow diagram**
(Rectangles represent physical things, diamonds represent events, ovals represent data, and rounded rectangles represent methods.)

An enzyme stability study is another prerequisite for the appropriate application of enzymes as catalysts. The enzyme activity and stability are not only affected by impurities but very often are sensitive to environmental conditions, such as temperature, and pH. While the effect of experimental conditions on the catalytic activity is an important topic of the enzyme catalytic assay and will be discussed in the following section, the stability study is to investigate the effect of various conditions on the stability of the enzyme. An enzyme can be very stable in a certain range of pH but decomposes seriously at other pH values. Some enzymes can be stored at temperatures above freezing point while others are more stable at lower temperatures. Solvent is another variable that affects the stability of enzymes. Some solvents may help to stabilize the enzyme while some others may initiate or accelerate the decomposition of the enzyme. A careful stability study is required for the best use of enzymes as they are often rare and expensive. The stability studies will provide information about how the enzyme should be stored and under what kind of conditions the enzyme is most effective. The results of the study should also be applied to the design of the enzyme catalytic assay, as the experiments should avoid the conditions that are not practical to the application of certain enzymes.

Inhibition study is another important part of enzyme catalytic study. The presence of inhibitors can dramatically decrease the catalytic activity of an otherwise very active enzyme. To choose a suitable enzyme as a catalyst for a specific application, it is necessary to know whether the substrates, products, and solvents are possible inhibitors of the enzyme. A very effective enzyme may turn out to be ineffective when it is applied to a practical solution due to the presence of inhibitors. For scientists involved in drug

design, the inhibition information may provide a starting point. As some effective drugs are good inhibitors of the target enzymes.

In enzyme catalytic assays, enzymes are normally used in the form of solutions. Enzyme solutions should be prepared according to the requirements of the catalytic assay experiments and the results of stability studies. The stability study should conclude which kind of solvent and what range of pH are suitable.

### 3.1.2 Chemical preparation

Chemicals used in catalytic assays are usually commercially available in appropriate forms and purification is not necessary for the assay. However, the final goal of a specific assay might involve application of the enzyme to chemicals from a non-commercial source. In this case purification prior to the assay is inappropriate. However, the solvent and pH of the substrate solution should be suitable for the corresponding enzymes.

### 3.1.3 Assay reaction operation and data collection

Assay reaction operation is to mix the enzyme solution and substrate solution and to initiate the catalytic reaction. Usually an instrument with pre-set parameters based on some preliminary experiments performs the measurement of concentration change and recording of the data automatically. If the concentration change of the reacting compounds or the product can be directly determined, no other compound needs to be added for the sake of measurement of the concentration change. In cases where both starting compound and product are not suitable for direct measurement, some other

compounds need to be introduced to determine the concentration indirectly. The compound introduced should have no significant effect on the catalytic reaction itself but may react with the product or starting compound to make it detectable or simply act as an indicator to signal a certain property change of the solution which can then be use to interpret the concentration change of the substrate or product.

### 3.1.4 Data processing

There are four categories of data for enzyme catalytic assay: data of enzyme stability study, data of optimal condition study, data of enzyme kinetics study, and data of enzyme inhibition study.

The data from the enzyme stability study is usually analyzed by direct comparison. They can be listed in tables or displayed graphically. No special analysis methods need to be applied.

The data collected from the optimal condition study can also be analyzed by direct comparison. A graph of the initial reaction rate against the change of the condition of concern such as pH and temperature is normally constructed, which is intuitive and straightforward.

The data collected from the kinetic study experiments are then processed to derive the kinetic parameters of the enzyme catalytic reaction, *i.e.* $K_m$, $V_{max}$, and $k_{cat}$. Substrate concentration is one of the most fundamental factors affecting the enzyme activity. Its relation to $K_m$ and $V_{max}$ has been expressed in the forms of the Michaelis-Menten equation and Lineweaver-Burke equation as discussed in **Chapter 2**. The Lineweaver-Burk equation was obtained by the rearrangement of the Michealis-Menten equation.

The values Km and Vmax are normally derived by using a Michaelis-Menton plot (**Figure 3.2**) or a Lineweaver-Burk plot (**Figure 3.3**).

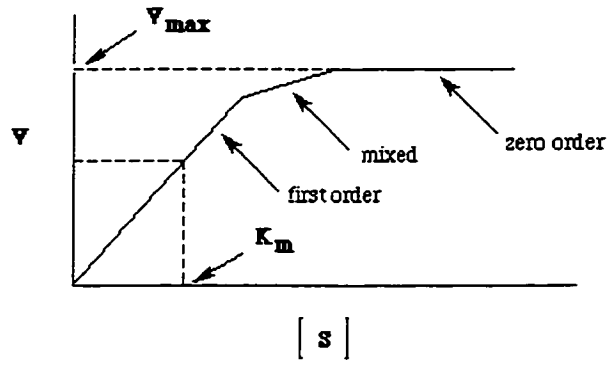


**Figure 3.2 Michaelis-Menten plot**

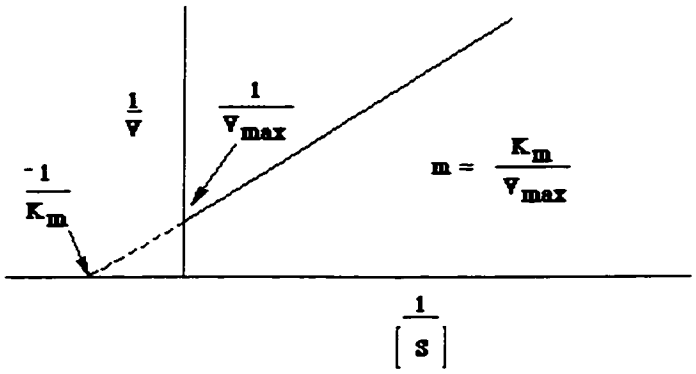


**Figure 3.3 Lineweaver-Burk plot**

Enzyme inhibition study is to obtain the inhibition constant $K_i$. By definition, $K_i$ is the inhibitor concentration when the Km is half of the Km value in the absence of the inhibitor. The process of inhibition data from inhibition study is performed in two steps: first to determine the $K_M$ and $V_{max}$ values of the enzyme catalyzed reaction in the presence and absence of the inhibitor respectively, then to derive the inhibition constant $K_i$. The first step is the same as discussed above for $K_M$ and $V_{max}$.

The traditional data processing of the enzyme kinetics study as discussed above introduced linear regression. The data transformation (reciprocals) distorts the experimental error, so the double-reciprocal plot does not obey the assumption of linear regression. Some modern software have been developed using nonlinear regression to fit data to Michaelis_menten equation to obtain the most accurate kinetic parameters for enzyme kinetic study.

## 3.2 Use case analysis

A database to support enzyme assays should cover all the information that scientists need to make use of the results directly, or for reference. The information should satisfy various needs of scientists. Although these needs can also be met through traditional ways such as using hard copies of experimental records, and various reference books, a centralized database that can be accessed electronically is far superior.

There are four groups of potential users of this database system: database administrator, internal scientists, external scientists, and bioinformaticians.

### 3.2.1 Database administrator

A database administrator has full control of the database. The DBA's activity includes database maintenance, and data manipulation based on the requirements of the users of the database system, *i.e.* scientists and bioinformaticians. The administrator can add new parts to the database, modify or delete existing parts in the database. The administrator's activity is dependent on requirements of the management. For example, the administrator may regroup the stored information to optimize the process of data

retrieval while keeping the information unchanged. In general the administrator's activity has effects on the database structure and how the information is retrieved but little direct effects on what data should be stored in the database.

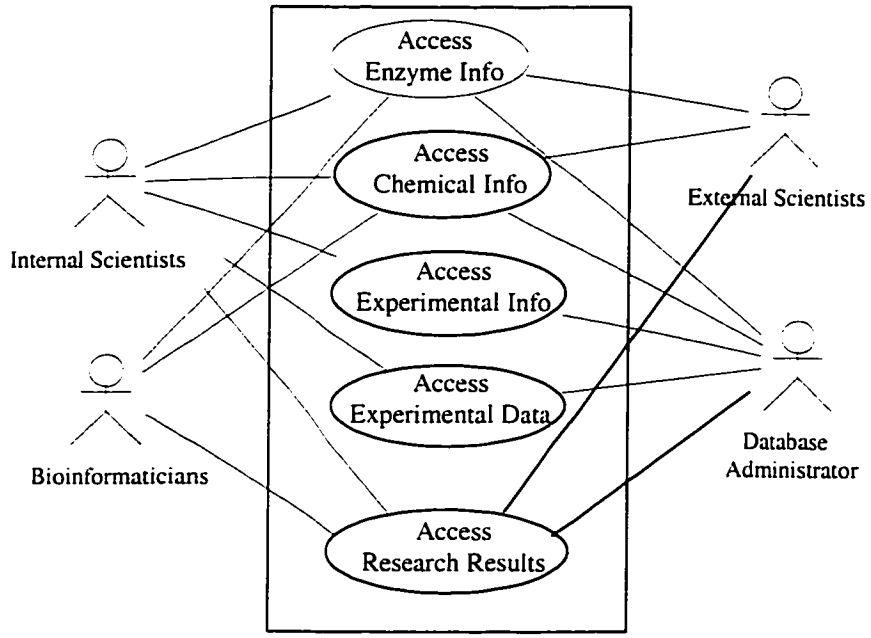


**Figure 3.4 Use Cases of EAMDB**
(Internal scientists input and use data, see elaboration in discussion)

### 3.2.2 Internal scientists

Internal scientists are the users of the database and the data sources as well. The database is designed to record the research results of the internal scientists. Their research work is the main data sources. The internal scientists have full access to the database. Their activities with the database include information collection and information application.

#### 3.2.2.1 Information collection

The data for a scientist performing enzyme catalytic assays to record in the database are the information about the substrates, catalysts (*i.e.* enzymes), reactions, experiments, experimental data, and conclusion.

The information about the substrates covers the chemical structure, physical properties, and solutions used in the experiments.

As enzyme structure information is a major concern on the biology side and is covered by corresponding database, the information for a scientist working on chemical assay is the properties related to chemical reactions, *i.e.* stability under various conditions, and information about the enzyme solutions used in the experiments.

Under different reaction conditions, same substrates can react differently yielding different products. The information about the reaction is important for chemical assays as different enzymes may lead to different products. It is especially useful for synthetic chemists due to the special catalytic properties of enzymes.

Information about experiments is another major data component for internal scientists to record. This information is needed when experiments need to be verified or to be repeated.

Experimental data are collected from the assay experiments. They are the basis for reaching the final conclusion. In case of any errors introduced during the data processing, the experimental data is needed for verification.

The conclusion is an essential part of the database and is directly derived from the experimental data. Basically it is the mapping between the substrates and the enzyme catalytic activity.

#### 3.2.2.2 Information application

Another major activity of internal scientists with the database system is to make use of the information in the database system, *i.e.* to retrieve the information from the database to guide their research activities.

(a) Choose an enzyme as a catalyst for a given substrate or application.

A scientist may need to find out which enzyme is active for the reaction of a given substrate, and how active the enzyme is, what kind of conditions should be applied for achieving the optimal results. Furthermore, the information should be available about what kind of chemical reaction is expected and what the final products would be. To serve this research activity is one of the major reasons to justify the necessity of such a database. Although similar information is available from public databases like **BRENDA**,[4] the storage of such information for **EAMDB** is not optional for a academic research lab due to the issue of intellectual property rights.

(b) Choose a substrate for a given enzyme.

Information about which enzymes are catalytically active for specific substrates under defined conditions, and the expected reactions and products is retrieved. This need is complementary to (a) and is also a major part for the justification of such a database.

(c) Find out the optimal reaction conditions for a given enzyme and substrate.

A user must be able to find information about the optimal reaction conditions and the possible reaction and products for a given pair of substrate and enzyme.

(d) Check the enzyme stability in various chemical environments.

Information about the enzyme stability in various chemical environments, including such factors as pH, concentrations, buffer constituents, presence of activators and/or inhibitors, and organic solvents is retrieved. These are important factors to consider when making decisions on what reaction conditions to employ for a reaction. An enzyme can be active to catalyze the reaction of a substrate under a wide range of conditions. However, the stability of the enzyme may prevent the application of the enzyme under certain conditions.

(e) Check the structural information of a substrate.

Structural information for substrates that are subject to the catalytic activity of an enzyme must be available. Vice versa, a scientist may also need to check that a specific enzyme is active toward what kind of substrates, *i.e.* the structural characteristics of these substrates. The structural information on these substrates can provide useful information for further development of the enzyme assay. With this information, for a new substrate to assay, the research can focus on a certain group of enzymes in the available library instead of a wild screening through the whole library.

(f) Check the original data and details of an experiment.

It should be possible for a user to check the details of an experiment. It is possible that an error or mistake occurred during the experiment set up, data collection, or data processing, which could affect the final conclusion. This conclusion may conflict with later research results. Recording the details of the experiment is

necessary for scientists to trace down any errors or mistakes introduced during the experimental stage.

(g) Check the technical protocol of the experiment.

A user should be able to check the technical protocol employed in the assay of enzymes and substrates. This information on the one hand can be used to evaluate the protocol for the assay of a specific enzyme and provide a basis for possible future improvement. On the other hand, it may also provide some useful guidance for the assay of similar enzymes and substrates.

(h) Obtain a detailed experimental procedure for a given enzyme and substrate.

This is one of the main reasons for the creation of such a database. Once an enzyme is found to effectively and efficiently catalyze the reaction of a substrate and the reaction conditions have been optimized, all the detailed information should be recorded in the database. A workable standard procedure must be available for routine assay of the substrate and enzyme.

(i) Check available protocols for the assay of an enzyme

It should be possible for a user to get the information about the protocols that can be used for a known enzyme.

### 3.2.3 External scientists

There are two major differences between external and internal scientists regarding the activities with the **EAMDB** system.

External scientists do not input any data into **EAMDB**. Although **EAMDB** will certainly contain the research achievements of the external scientists, the information will be collected and uploaded into **EAMDB** by the internal scientists.

External scientists have no privilege to access the experimental details as the experimental details are sometimes patent-related. Therefore the activities of external scientists are the same in part as the internal scientist activities and can be served by the database information by excluding access to the experimental part.

### 3.2.4 Bioinformaticians

**EAMDB** is of interest to bioinformaticians. Enzyme catalytic properties are determined by the enzyme structure. Two enzymes of very similar structure should have similar catalytic properties and different catalytic properties imply different structure. Bioinformaticians can make use of the enzyme catalytic assay results to verify the results of enzyme sequence analysis, analysis and prediction of 2D and 3D structures, and hence to modify the corresponding analysis system. The research conclusion section of the database can provide valuable information for bioinformaticians.

## 3.3 Main modules of the database

To store all the information of enzyme assay experiments and serve the scientists and bioinformaticians for the various needs discussed above, the database should be composed of the following modules: enzyme activity mapping module, enzyme module, chemical compound module, experiment module, data module, and reference module.

Each module is composed of one or more tables to hold the related information as shown in **Figure 3.5**.

### 3.3.1 Enzyme activity mapping module

The enzyme activity module is to store the final results of the research work. It maps the enzyme activity to the corresponding substrates and reactions, *i.e.* providing detailed information about the catalytic activity of enzymes towards specific substrates and reactions. It contains all conclusions about enzyme catalytic activities obtained from the enzyme assay experiments. It is the central part of the **EAMDB** system. There are two tables in this module: the Reaction table and the KineticParameter table.

The Reaction table is used to store the reactions catalyzed by enzymes. The information includes the substrates, products, and reaction classification. The reactions stored may be a novel reaction discovered by internal scientists or a well-known reaction but catalyzed by the enzymes under study. They can also be of interest to our scientists and remain as a subject of future chemical assay. The Note field in the table is to indicate whether the reaction is a novel or a known one. The Reference field is to give literature source if the reaction is reported by literature. The chemical information of the main substrates and products is stored in the chemical compound module. The fields in the Reaction table and the corresponding descriptions are listed in **Table 3.1.**

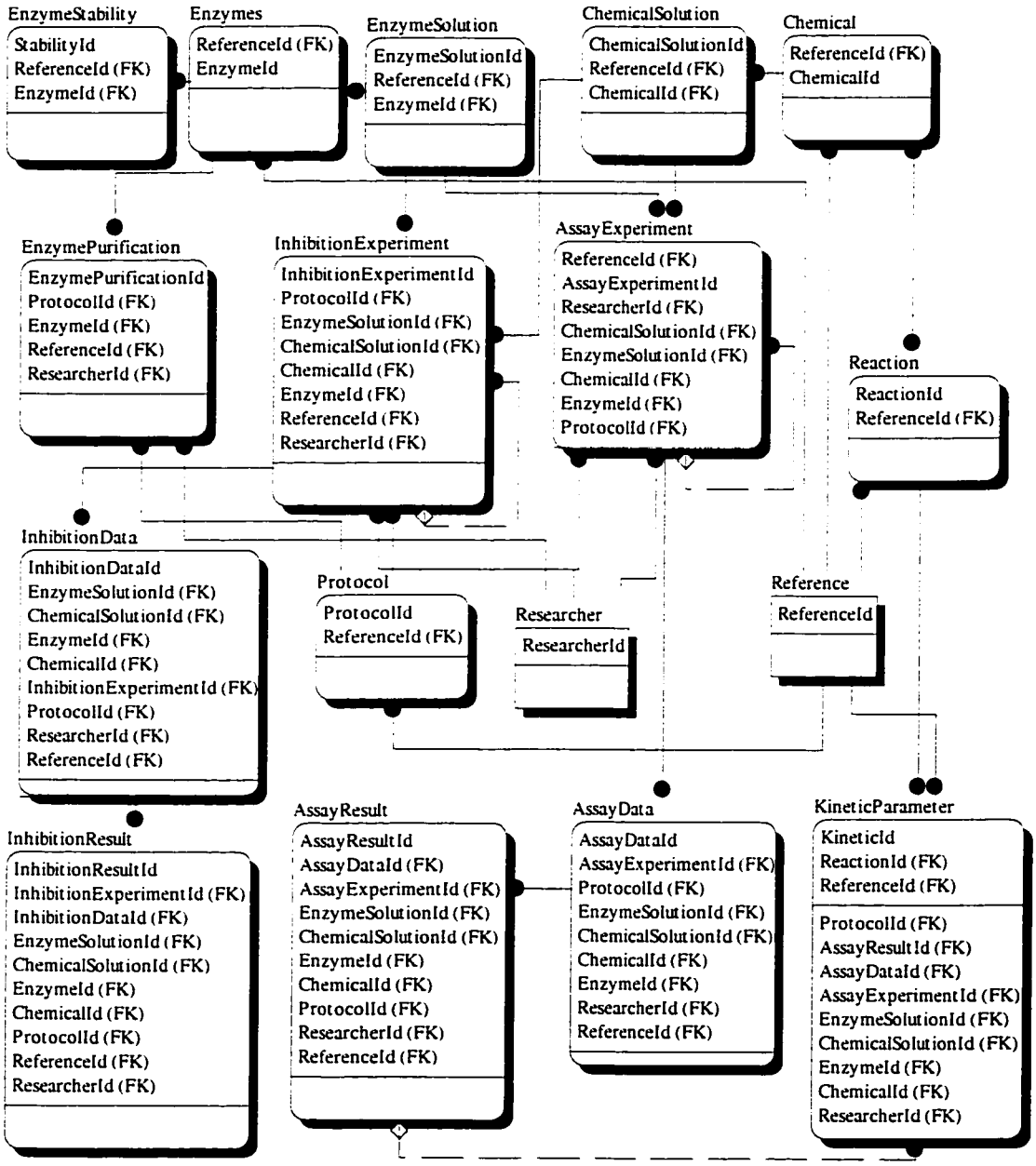


**Figure 3.5 ER Model of EAMDB**

(For clarity purpose, all the non-key attributes are omitted.
For full list of all attributes, see tables follow)

**Table 3.1**. List of Fields in the Reaction Table

| Field Name | Data type | Description |
|---|---|---|
| ReactionId | varchar (20) | Primary key |
| SubstrateId | varchar (20) | Main substrate in the reaction, referenced to Chemical table |
| ProductId | varchar (20) | Major product in the reaction, referenced to Chemical table |
| ReactionEquation | varchar (200) | Full reaction equation |
| ReactionType | varchar (30) | Classification of the reaction |
| Note | varchar (200) | Indicate whether the reaction is new or well-known |
| ReferenceID | varchar (20) | Literature sources of the reaction, referenced to Reference table |

The information in the KineticParameter table covers the enzyme identity information, reaction information, and major catalytic activity parameters (*i.e.* Km, $k_{cat}$), optimal pH, buffer solution, optimal temperature, optimal concentrations, reaction, and related experiments. The fields in this table and their descriptions are listed in **Table 3.2**.

The information stored not only shows positive assay results, but also shows negative results. The information may be summarized from the results of several experiments. In that case the experimental results of the related experiments will be identified in the field of AssayResultId. The information can also be extracted from a public literature with which the source of the information will be specified in the field of ReferncedId.

By using this module, scientists are able to identify which enzyme is active for which substrate under what kind of optimal conditions and how active the enzyme is towards the reaction. According to the information retrieved from this module, with a

given substrate, a scientist will be able to select a suitable enzyme to catalyze the reaction of the substrate; with a given enzyme, a scientist will be able to find to which substrate the enzyme can be applied as a catalyst.

**Table 3.2.** List of Fields in the KineticParameter Table

| **Field Name** | **Data type** | **Description** |
|---|---|---|
| KineticId | varchar (20) | Primary key |
| ReactionId | varchar (20) | Foreign key from Reaction table in this module |
| EnzymeId | varchar (20) | Foreign key from Enzyme table in enzyme module. |
| PhHi | varchar (20) | The high bound of pH that the specified catalytic activity is obtained |
| PhLo | varchar (20) | The low bound of pH that the specified catalytic activity is obtained |
| PhOptimum | varchar (8) | pH at which the best catalytic activity is obtained |
| TempHi | varchar (20) | Temperature high bound that the specified catalytic activity is obtained |
| TempLo | varchar (20) | Temperature low bound that the specified catalytic activity is obtained |
| TempOptimum | varchar (20) | The temperature at which the best catalytic activity is obtained |
| Cofactor | varchar (20) | Co-enzyme or cofactor used in the reaction |
| Activator | varchar (20) | The activator used in the reaction |
| KmOp | varchar (20) | Km value under optimal conditions |
| KmRm | varchar (20) | $K_m$ value at room temperature |
| KcatOp | varchar (20) | Turnover Number of the enzyme in the reaction under optimal conditions |
| KcatRm | varchar (20) | $k_{cat}$ value at room temperature |
| SpecificActivity | varchar (20) | Specific activity is a unit to express the amount of enzyme |
| AssayResultId | varchar (50) | List of AssayResultsIds. AssayResultsId is the primary key for the AssayResults table in the data module. |
| ReferenceId | varchar (20) | Foreign key from reference table |

### 3.3.2 Enzyme module

As some of the enzyme information, such as sequence and structure, has been stored in other databases within the bioinformatics platform, this module is mainly for the information related to enzyme catalytic assays, namely enzyme solution, and enzyme stability. There are three tables in this module: the Enzyme table, the EnzymeStability table, and the EnzymeSolution table.

The Enzyme table is to store the information about the enzyme identity. Although IUPAC and IUBMB have recommended a nomenclature for enzymes, [20] due to historical reasons and scientist personal preferences, the same enzyme may be referred to with different names. The Enzyme table covers all the possible names used for the enzyme. The class of the enzyme is also indicated. The fields in the Enzyme table and their descriptions are listed in **Table 3.3**.

**Table 3.3**. List Fields in the Enzyme Table

| **Field Name** | **Data type** | **Description** |
|---|---|---|
| EnzymeId | varchar (20) | Primary key. |
| EnzymeName | varchar (100) | The name used by our lab. It may come from references. |
| EcNumber | varchar (20) | EC number of the enzyme |
| RecommendedName | varchar (100) | Recommended name of the enzyme |
| Synonym | varchar (100) | Other names used for the enzyme. |
| CasNumber | varchar (20) | CAS registry number |
| Class | varchar (50) | Class of the enzyme |
| Comment | varchar (200) | Description of storage temperature, specific activity, and physical form |
| UnitDefinition | varchar (200) | How the activity unit is defined. |
| Source | varchar (150) | The origin of the enzyme |
| ReferenceId | varchar (20) | Foreign key from reference table |

The EnzymeStability table stores the information about stability of the enzyme under various conditions, such as pH, temperature, and solvent. Enzyme catalytic information is important for optimizing conditions for the enzyme catalytic reactions. The fields and their descriptions are listed in **Table 3.4.**

**Table 3.4.** List of Fields in the EnzymeStability Table

| **Field Name** | **Data type** | **Description** |
|---|---|---|
| StabilityId | varchar (20) | Primary key |
| EnzymeId | varchar (20) | Foreign key from Enzyme table |
| PhStability | varchar (20) | Suitable pH for catalytic reaction |
| TempStability | varchar (20) | Suitable temperature for catalytic reaction |
| GeneralStability | varchar (300) | This field summarizes general information on stability, e.g., increased stability of immobilized enzymes, stabilization by SH-reagents, detergents etc. |
| SolventStability | varchar (150) | List of types of suitable solvents for storage and reaction, e.g. water-miscibility, hydrophobicity, polarity |
| OxidationStability | varchar (100) | Stability in the presence of oxidizing agents. e.g. $O_2$, $H_2O_2$ |
| StorageStability | varchar (100) | Suitable pH and temperature for storage |

The EnzymeSolution table stores the information about the enzyme solution used in experiments. The information covers the concentration of the enzyme, solvent, and pH. The fields and their descriptions are listed in **Table 3.5.**

**Table 3.5**. List of Fields in the EnzymeSolution Table

| EnzymeSolutionId | varchar (20) | Primary key |
|---|---|---|
| EnzymeId | varchar (20) | Foreign key from the Enzyme table in this module |
| Solvent | varchar (20) | Solvent used in the solution |
| Ph | varchar (5) | pH value of the solution |
| Buffer | varchar (20) | Buffer solution used for the solution |
| Stabilizer | varchar (20) | Stabilizer if used to make the solution |
| Concentration | double | Concentration of the enzyme |
| ConUnit | varchar(5) | Concentration unit |

### 3.3.3 Chemical compound module

This module is for the information about substrates and products. A wide range of information is needed to characterize the properties of a compound. As some of the properties are only of interest to the study of the compound itself but not its reactions, they will not be covered in **EAMDB**. The information in this module will focus on the properties related to enzyme-catalyzed reactions. There are two tables in this module: the Chemical table and the ChemicalSolution table.

The Chemical table holds the information about chemical structure, namely identity of the substrate, molecular formula, functional groups, stereo structure, physical properties, spectrum information, and 3D picture or drawing, if available. The fields are listed in **Table 3.6**. Since NMR and IR are the routine techniques to characterize a chemical compound, the chemical shifts and coupling patterns of characteristic NMR

signals, and the positions and intensities of the characteristic IR peaks are stored in the table. If required the NMR and IR spectra can be stored in specified directories with the file names and directory path stored in the database.

**Table 3.6**. List of Fields in the Chemical Table

| **Field Name** | **Data type** | **Description** |
|---|---|---|
| ChemicalId | varchar (20) | Primary key |
| ChemicalName | varchar (50) | Compound name used in the system |
| CaName | varchar (50) | Name registered with CA |
| IupacName | varchar (50) | Name according to IUPAC nomenclature |
| Synonym | varchar (50) | Other name used for the compound |
| CaNumber | varchar (20) | CA registry number of the compound |
| StructuralFormula | varchar (100) | Structural formula of the compound, e.g. $CH_3CH_3$ for ethane |
| Class | varchar (50) | Type of the compound, e.g. acid, ester etc. |
| MolecularWeight | varchar (10) | Molecular weight |
| StereoCenter | varchar (50) | Location and configuration of the stereo center |
| FunctionalGroup | varchar (50) | Major functional groups |
| BoilingPoint | varchar (10) | Boiling point of the compound |
| MeltingPoint | varchar (10) | Melting point of the compound |
| Nmr | varchar (100) | Characteristic NMR peak chemical shift and split pattern |
| Ir | varchar (100) | Type and location of the characteristic IR peak |
| ReferenceId | varchar (200) | Foreign key from the Reference table |

In enzyme assay experiments, chemicals are used in the form of solutions. The ChemicalSolution table stores the information about the chemical solutions used in the experiments, namely compound identity, solvent, and concentration (**Table 3.7**).

From the chemical module scientists can retrieve the detailed information about the substrates. The chemical information along with enzyme sequence and structural information (from the enzyme sequence and structure section) can provide scientists with great insight into the mapping between structure and catalytic activity. This may enable

scientists to predict the catalytic activity of an enzyme towards the reaction of an specific substrate.

**Table 3.7**. List of Fields in the ChemicalSolution Table

| Field Name | Data type | Description |
|---|---|---|
| ChemicalSolutionId | varchar (20) | Primary key |
| ChemicalId | varchar (20) | Foreign key from the compound table |
| Solvent | varchar (20) | Solvent used for the solution |
| pH | varchar (5) | pH value in the solution |
| Buffer | varchar (20) | Buffer solution used to make the solution |
| Concentration | varchar (10) | Concentration of the substrate |

### 3.3.4 Experiment module

The Experiment module stores the information about experiments. There are four tables in this module: AssayExperiment table, Protocol table, EnzymePurification table, InhibitionExperiment table, and Researcher table.

The AssayExperiment table stores the detailed information about the experiments including experimental conditions, technique protocol employed, enzyme, substrate, experimental layout, and experimental data. The information about the experimental layouts stored in the table is file names and directory path since the experimental layouts are usually described in the forms of a graph or picture. They are stored as files in specified directories. The fields and their descriptions are listed in **Table 3.8**.

The Protocol table stores the information about all the technique protocols used for enzyme assay, enzyme purification, and enzyme inhibition study. Some of the protocols are used in the current chemical assay project; some may only be useful for future projects. The information in this table includes target enzyme, target substrate,

characterization method, experimental conditions, references etc. The tables and corresponding fields are listed in **Table 3.9**.

**Table 3.8**. List of Fields in the AssayExperiment Table

| Field Name | Data type | Description |
|---|---|---|
| AssayExperimentId | varchar (20) | Primary key |
| ExperimentTitle | varchar (100) | Brief description of the goal of the experiment |
| Date | Date | Experiment date |
| ResearcherId | varchar (20) | Foreign key from People table |
| ProtocolId | varchar (20) | Foreign key in the Protocol table |
| EnzymeSolutionId | varchar (20) | Foreign key in the EnzymeSolution table |
| ChemicalSolutionId | varchar (20) | Foreign key in the ChemicalSolution table |
| LayoutFileName | varchar (20) | The file name of the layout file |
| LayoutFilePath | varchar (20) | The file path of the file |
| DataFileId | varchar (20) | Foreign key in the DataFile table |
| RelatedExperimentId | varchar (30) | List of ExperimentId of the related experiments. |

**Table 3.9**. List of Fields in the Protcol Table

| Field Name | Data type | Description |
|---|---|---|
| ProtocolId | varchar (20) | Primary key |
| ProtocolName | varchar (50) | Name of the protocol |
| TargetEnzyme | varchar (50) | Target enzyme name |
| TargetChemical | varchar (50) | Target chemical name |
| Cofactor | varchar (50) | Cofactor used in the assay |
| Coenzyme | varchar (50) | Coenzyme used in the assay |
| Activator | varchar (50) | Activator used in the assay |
| CharacterizationMethod | varchar (50) | Measurement method for collecting the kinetic data, e.g. photo spectroscopy, titration, etc. |
| ProtocolDescription | Text | Detailed procedure of the protocol |
| ReferencesId | varchar (20) | Original source of the protocol |

Pure enzymes are required for the kinetic study of enzyme-catalyzed reactions. Each enzyme needs specific purification strategy. The EnzymePurification table is used to store the information about the possible purification methods, procedures, and the final specific activity of the enzyme. The fields in the EnzymePurification table and their description are listed in **Table 3.10**.

**Table 3.10**. List of Fields in the EnzymePurification Table

| EnzymePurificationId | varchar (20) | Primary key |
|---|---|---|
| EnzymeId | varchar (20) | Foreign key from the Enzyme table |
| Method | varchar (200) | Method used for the purification |
| Procedure | Text | Detail procedure |
| SpecificActivity | varchar (20) | Enzyme activity after purification |
| Yield | varchar (10) | Yield of the purification |
| ProtocolId | varchar (20) | Foreign key from the Protocol table |

The InhibitionExperiment table stores the detailed information about the inhibition experiments including experimental conditions, technique protocol employed, enzyme, chemical (substrate, and inhibitor), experimental layout, and experimental data. The information about the experimental layouts stored in the table is file names and directory path since the experimental layouts are usually described in the forms of graph or picture. They are stored as files in specified directories. The fields and their descriptions are listed in **Table 3.11**.

The Researcher table is to store the information about the researchers. The information is mainly for research purposes rather than personnel management. Therefore

no personal information, such as phone number and home address, will be covered. The fields and their descriptions are listed in **Table 3.12**.

**Table 3.11**. List of Fields in the InhibitionExperiment Table

| Field Name | Data type | Description |
|---|---|---|
| InhibitionExperimentId | varchar (20) | Primary key |
| ExperimentTitle | varchar (100) | Brief description of the goal of the experiment |
| Date | Date | Experiment date |
| ResearcherId | varchar (20) | Foreign key from Researcher table |
| ProtocolId | varchar (20) | Foreign key in the Protocol table |
| EnzymeSolutionId | varchar (20) | Foreign key in the EnzymeSolution table |
| SubstrateSolutionId | varchar (20) | Referenced to ChemicalSolutionId in ChemicalSolution table |
| InhibitorSolutionId | varchar (20) | Referenced to ChemicalSolutionId in ChemicalSolution table |
| LayoutFileName | varchar (20) | The file name of the layout file |
| LayoutFilePath | varchar (20) | The file path of the file |
| RelatedExperimentId | varchar (30) | List of ExperimentId of the related experiments. |

**Table 3.12**. List of Fields in the Researcher Table

| Field Name | Data type | Description |
|---|---|---|
| ResearcherId | varchar (20) | Primary key |
| ResearcherName | varchar (50) | Name of the protocol |
| ResearchGroup | varchar (50) | Target enzyme name |

### 3.3.5 Data module

The data module stores the information of the data obtained from the enzyme catalytic assay experiments and inhibition experiments. There are four tables in this

module: AssayData table, the AssayResult table, the InhibitonData table, and the InhibitionResult table.

The AssayData table will store the information about the original data obtained from assay experiments. It is conceivable that storing the original data file in the database directly is not an elegant way as the file size and format might be different due to the fact that different assay protocols might be applied. The original data files will be stored in specified directories on the server. The file names and path will be stored in the AssayData table. The application software will handle the access to the data files according to the information retrieved from the database. The fields and their descriptions in the AssayData table are listed in **Table 3.13**.

**Table 3.13**. List of Fields in the AssayData Table

| Field Name | Data type | Description |
|---|---|---|
| AssayDataId | vachar (20) | Primary key |
| ExperimentId | vachar (20) | Foreign key from experiment table in the experiment module |
| DataFileName | vachar (30) | File name of the data file |
| DataFilePath | vachar (50) | Directory path of the data file |

The AssayResult table is used to store the kinetic study results from the assay experiments. The information includes values Km, Vmax, $k_{cat}$, the graphs derived from the original experimental data, and information about the original data file. Graphs derived from original experimental data will also be stored as files in specified directories and the corresponding file names and the directory paths will be stored in the AssayResult table. The fields and their descriptions are listed in **Table 3.14**.

The InhibitionData will store the information about the original data obtained from inhibition experiments. As the original experimental data may be in different size

and format, the original data files will be stored in specified directories on the server while the file names and paths will be stored in the InhibitionData table. The fields in the table and their descriptions are listed in the **Table 3.15**.

**Table 3.14**. List of Fields in the AssayResult Table

| Field Name | Data type | Description |
|---|---|---|
| AssayResultId | vachar (20) | Primary key |
| AssayDataId | vachar (20) | Foreign key from AssayData table |
| GraphFileName | vachar (30) | File name of the graph to derive the results |
| GraphFilePath | vachar (50) | Directory path of the graph file |
| Km | vachar (10) | Km value |
| $k_{cat}$ | vachar (10) | Kcat value |
| Vmax | vachar (10) | Vmax value |

**Table 3.15**. List of Fields in the InhibitionData Table

| Field Name | Data type | Description |
|---|---|---|
| AssayDataId | vachar (20) | Primary key |
| ExperimentId | vachar (20) | Foreign key from experiment table in the experiment module |
| DataFileName | vachar (30) | File name of the data file |
| DataFilePath | vachar (50) | Directory path of the data file |

InhibitionResult table is used to store the results of inhibition experiments. The information includes $K_i$, graphs derived from original experimental data, and information about the original data file. Graphs derived from original experimental data will also be stored as files in specified directories and the corresponding file names and the directory paths will be stored in the InhibitionResult table. The fields in the table and descriptions are listed in **Table 3.16**.

**Table 3.16**. List of Fields in the InhibitionResult Table

| Field Name | Data type | Description |
|---|---|---|
| InhibitionResultId | vachar (20) | Primary key |
| InhibitionDataId | vachar (20) | Foreign key from AssayData table |
| GraphFileName | vachar (30) | File name of the graph to derive the results |
| GraphFilePath | vachar (50) | Directory path of the graph file |
| Ki | vachar (10) | Value of inhibition constant Ki |

### 3.3.6 Reference module

There is one table in this module: Reference table. It stores the literature sources cited in all other modules. The fields and their descriptions in the Reference table are listed in **Table 3.17**.

**Table 3.17**. List of Fields in the Reference Table

| Field Name | Data type | Description |
|---|---|---|
| ReferenceId | vachar (20) | Primary key |
| Title | vachar (200) | Title of the literature |
| Source | vachar (200) | Journal or book title |
| Author | varchar(200) | List of author(s) |
| PublishDate | varchar (50) | Publish date (year, month) |
| Volume | varchar (10) | Journal volume |
| Issue | varchar (10) | Journal issue |
| Page | varchar (10) | Page of the literature |

## 3.4 Primary keys and foreign key constraints

The database is composed of five modules. Each module is composed of several tables. The relational data model prescribed above is well normalized. Redundant storage of information has been avoided by introducing primary keys and foreign key constraints. Each table stores specific aspect information of the chemical assay experiments. Primary

keys are defined for every table. Any information stored in other tables is referenced with foreign keys. For example, all enzyme information is stored in the three tables in the Enzyme module. In the KineticParameter table, the information on an enzyme is given by a foreign key from the Enzyme table. The detailed information can be retrieved easily from the Enzyme table.

# Chapter 4 Concluding Remarks

## 4.1 Conclusion

Enzymes are proteins. They are nature's catalysts with special properties that are not often found in other catalysts. The advantages of enzymes as catalysts are their high selectivity (stereo-selectivity, region-selectivity, and chemo-selectivity), the mild reaction conditions, high catalytic efficiency, and environmental friendliness. Their applications are growing rapidly. On the other hand, the lack of structural information of the enzymes makes it hard to predict the activity and selectivity of an enzyme towards certain substrates. With the development of modern molecular biology and bioinformatics, more and more protein sequences have been determined. The better understanding of the enzyme structure will provide the insight of the special catalytic properties of enzymes and hence guide the catalytic assay to high efficiency. Enzyme structure determines the catalytic activity and enzyme catalytic activity is a direct reflection of the enzyme structure. Enzyme catalytic results should be integrated into the bioinformatics system and be used as hard evidence to cross-check the enzyme structure obtained from biological sequencing and bioinformatics analysis. The design of the Enzyme Activity Mapping Database (**EAMDB**) presented in this thesis is part of the effort to meet the challenge.

**EAMDB** has been designed to accommodate the enzyme catalytic assay results in pursuit of environmentally friendly and economical catalysts for the pulp and paper industry. **EAMDB** is a relational database with 17 tables organized into six parts: the Enzyme module, the Chemical Compound module, the Experiment module, the Catalytic Mapping module, the Data module, and the Reference module.

The Enzyme module is to store the information about enzymes including enzyme names, physical properties, purification method, and corresponding solutions. This module can serve as a connection point to integrate **EAMDB** into the bioinformatics database by sharing the enzyme identity information.

The Chemical Compound module stores the information about the chemicals including the catalytic substrates and products. Essential information to characterize chemicals is included such as structure formula, functional groups, stereo-center, IR, NMR, and melting point. It also includes the information about chemical solutions.

The Experiment module is used to store the detailed information about individual experiments. Two parts form this module: the Protocol, and the Experiment. The Protocol part covers the information about assay protocols used in the catalytic activity assay experiments, enzyme purification experiments, and inhibition experiments. The Experiment part records the details of each experiment.

The Data module is to store the information about the data obtained from assay experiments and conclusions derived from the experimental data. The experimental data exported by the measurement instrument will be stored in its original format while the corresponding file name and path will be stored in the database. The graphs leading to the conclusion will be stored as files with the file names and directory path stored along with conclusion.

The Enzyme Activity Mapping module is the central part of the **EAMDB** system. It stores information of enzyme-catalyzed reactions, corresponding kinetic parameters, and optimal reaction conditions. The information is the conclusions of the enzyme

catalytic assay. This module has the Reaction table and the KineticParameter table for the corresponding information.

The Reference module records the literature sources related to the information stored in other modules.

**EAMDB** is designed for usage by internal scientists, external scientists, and bioinformaticians. The information stored in **EAMDB** can be applied in various scientific research areas: drug design, organic synthesis, enzyme structure analysis, and biodegradation etc. The information can also be applied to bioinformatics system development since the enzyme's catalytic activity directly reflection of enzyme structure.

Internal scientists can use **EAMDB** to record the experimental details, experimental data, and experimental conclusions. They can retrieve the information whenever needed for research reference and guidance of their further research design.

External scientists can retrieve the information about the enzyme catalytic activity information for their research reference.

Bioinformaticians can use the information to verify their structure analysis. Enzymes with similar structures should have similar catalytic activity. Different catalytic activity implies different structures. A structure conclusion from a bioinformatics system should be in conformity with catalytic activity assay results if available. An inconformity may imply a flawed system. Bioinformaticians can also use the information to predict enzyme structures.

The design of **EAMDB** is well normalized. All data are arranged into logical groupings. Each group describes a small part of the whole. Duplicate data stored in the database have been minimized. The data have been organized in the way such that any

modification can be achieved by making the change in only one place. The design of the database is also aimed at making the access and manipulation quick and efficient without compromising the integrity of the data in storage.

## 4.2 Contribution to knowledge

The thesis presents a database design to support enzyme catalytic assay experiments. Unlike the other databases such as **BRENDA**, **LIGAND**, or **UM-BBD**, **EAMDB** presented in this thesis is experiment-oriented. Scientists can document detailed information of assay experiments. The information stored will cover chemical compound information, chemical reaction information, and enzyme information. The data in this database comprise important information about the functional aspects of enzymes. The database design presented encompasses cheminformatics and bioinformatics. The system can provide organic chemists with valuable information for drug design and organic compounds synthesis. The system can also provide biologists with the functional information about enzymes. It is expected that the database will be an important part of the bioinformatics system.

## 4.3 Suggestions for future work

**EAMDB** is designed to meet the requirements of enzyme catalytic assay requirements. The requirements are specified according to the needs of biochemists. The final product is expected to full support enzyme assay experiments with high accuracy, efficiency, and convenience. Further work is suggested as follows:

1. Verify and validate all the requirements with potential end users, *e.g.* biochemists, and chemists etc. Any software development requires the specification, design, validation, and evolution process to ensure the final product is the right one that is needed. Sometimes the process needs to be repeated. To make the current **EAMDB** design cover the requirements of scientists accurately and completely, further verification and validation should be performed.
2. Implement the **EAMDB** design into MySQL, PostgreSQL, Oracle or any other relational databases that enforce the primary key and foreign key constraints.
3. Design and implement the corresponding application program and user interface. A good database design needs the support of good applications and user interfaces to enforce its functionality requirements, *e.g.* access privilege control, data entry validation, and enforcement of data integrity.
4. Integrate **EAMDB** into the bioinformatics system. The information of enzyme catalytic activity is part of the functional information about the enzymes. To complete the bioinformatics system, **EAMDB** should be considered as an essential part of it.

# References


1. John Westley. Enzyme Catalysts. John Westley, U.S.A. (1969).
2. Samadni, G. Pulp bleaching – the race for safer methods. *Chem. Eng. (Int Ed)* **98** (1991), 37-43.
3. (a) Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30** (2002), 402-404. (b) Goto, S., Nishioka, T. and Kanehisa, M. LIGAND: chemical database for enzyme reactions. *Bioinformatics* **30** (1998), 402-404.
4. (a) Schomburg, I., Chang, A. and Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **30** (2002), 47-49. (b) Schomburg, I., Hofmann, O., Baensch, C., Chang, A. and Schomburg, D. Enzyme data and metabolic information: BRENDA, a source for research in biology, biochemistry, and medicine. *Gene Funct. Dis.* (**3-4**) (2000): 109-118.
5. (a) Burgard, A. P. and Maranas, C. D. Review of the Biocatalysis/Biodegradation Database (UM-BBD). *Metabolic Engineering* **4** (2002), 111-113. (b) Ellis, L. B. M., Hershberger, C. D., and Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation database: Specialized Metabolism for functional genomics. *Nucleic Acids Res.* **27** (1999), 373-376. (c) Ellis, L. B. M., Hershberger, C. D., and Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation database: Specialized Metabolism for functional genomics. *Nucleic Acids Res.* **28** (2000), 377-379. (d) Ellis, L. B. M., Hershberger, C. D., and Wackett, L. P. The University of Minnesota

Biocatalysis/Biodegradation database: Specialized Metabolism for functional genomics. *Nucleic Acids Res.* **29** (2001), 340-343.

6. Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K. and Kanehisa, M. *Organizing and Computing Metabolic Pathway data in Terms of Binary Relations.* In Altman, R. B., Dunker, A. K., Hunter, L. and Klein, T. E. (eds), Pacific Symposium on Biocomputing '97. World Scientific, Maui, pp. 175-186, (1997).
7. Weininger, D. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **28** (1988), 31-36.
8. Andreas D. B. The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res.* **30**(1) (2002), 1-12
9. Guenter Stoesser, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Nicole Redaschi, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara and Robert Vaughan. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **30**(1) (2002), 21-26
10. Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp and David L. Wheeler. GenBank. *Nucleic Acids Res.* **30**(1) (2002), 17-20
11. Bairoch, A. and Apweiler,R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28** (2000), 45–48.

12. Bernstein F. C., Koetzle T. F., Williams G. J., Mey E. F., Jr., Brice M. D., Rodgers J. R., et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**(2) (1977), 319-24.

13. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., et al. The Protein Data Bank. *Nucleic Acid Res.* **28**(1) (2000) 235-42

14. Pearson W. R., Lipman D. J., Improved tools for biological sequence comparison. Proc. *Natl Acad Sci U S A* **85**(8) (1988): 2444-2448.

15. Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17) (1997), 3389-3402.

16. Berg, J. M.; Tymoczko, J. L.; Stryer, L. Biochemistry 5$^{th}$ Edition. pp. 189-222,W. H. Freeman and Company, New York, (2002)

17. Lengauer, T. (Ed.). Bioinformatics – From Genomes to Drug. Wiley-VCH Verlag GmbH, Weinheim (2002).

18. Cheeseman, J. D.; Corbett, A. D.; Shu, R.; Croteau, J.; L. Gleason, J. L.; and Kazlauskas, R. J. Amplification of Screening Sensitivity Through Selective Destruction. Theory and Screening of a Library of Carbonic Anhydrase Inhibitors. *J. Am. Chem. Soc.* **124**(2002), 5692-5701 and references therein.

19. Connors, K. A. Chemical Kinetics. pp. 100-105, VCH Publishers, Inc. U. S. A. (1990)

20. *Enzyme Nomenclature 1992* [Academic Press, San Diego, California, ISBN 0-12-227164-5 (hardback), 0-12-227165-3 (paperback)] with Supplement 1 (1993), Supplement 2 (1994), Supplement 3 (1995), Supplement 4 (1997) and

Supplement 5 (in *Eur. J. Biochem.* 1994, **223**, 1-5; *Eur. J. Biochem.* 1995, **232**, 1-6; *Eur. J. Biochem.* 1996, **237**, 1-5; *Eur. J. Biochem.* 1997, **250**; 1-6, and *Eur. J. Biochem.* 1999, **264**, 610-650; respectively)

21. Muller R. J. Database Design for Smarties Using UML for Data Modelling. pp. 29-35, Morgan Kaufmann Publishers, Inc. U.S.A. (1999)

# Appendix

**BRENDA** http://www.brenda.uni-koeln.de

Extensive functional data on enzymes

**DDBJ (DNA Data Bank of Japan)** http://www.ddbj.nig.ac.jp

All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration

**EMBL** http://www.ebi.ac.uk/embl.html

All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration

**ENZYME** http://www.expasy.ch/enzyme/

Enzyme nomenclature

**GENE-BANK** http://www.ncbi.nlm.nih.gov/

All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration

**HSSP** **http://www.sander.ebi.ac.uk/hssp/**

Structural families and alignments; structurally-conserved regions and domain architecture

**KEGG (Kyoto Encyclopedia of Genes and Genome)**

http://www.genome.ad.jp/kegg/Metabolic and regulatory pathways

**LIGAND** http://www.genome.ad.jp/ligand/

Chemical compounds and reactions in biological pathways

**SWISS-PROT/TrEMBL** **http://www.expasy.ch/sprot**

Curated protein sequences

**UM-BBD** http://umbbd.ahc.umn.edu/

Microbial biocatalytic reactions and biodegradation pathway