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

DECEMBER 2002
© RONGHUA SHU, 2002
This is to certify that the thesis prepared

By: Ronghua Shu

Entitled: Design of An Enzyme Activity Mapping Database

And submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulation of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_______________________________ Chair

Examiner

Examiner

Supervisor

Approved ____________________________

Chair of Department of Graduate Program Director

_________ 20 ______

Dr Nabil Esmail, Dean
Faculty of Engineering And Computer Science
Design of An Enzyme Activity Mapping Database

Abstract

Ronghua Shu

This thesis presents the design of a relational database supporting enzyme catalytic activity high throughput assay – Enzyme Activity Mapping Database (EAMDB). The database is composed of five modules according to the information to store: enzyme module, chemical module, activity mapping module, experimental module, and data module. Each module has one or more tables to cover relevant information. There are 12 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 functional information to serve scientists for their various research interests.
Acknowledgement

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 enzyme assay introduction and discussion. Special thanks go to Dr. Justin Powlowski for his detailed correction and discussion of the requirement document.

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

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

List of Tables 7  
List of Figures 8  

Chapter 1. Introduction 9  
1.1 Enzyme Catalytic assay project 9  
1.2 Public enzymatic reaction databases 9  
1.3 Why a new enzyme catalytic database of our own 12  
1.4 Contribution of the thesis 13  
1.5 Layout of the thesis 14  

Chapter 2. Background 15  
2.1 Brief review of Bioinformatics 15  
  2.1.1 Scientific problems for bioinformatics 15  
  2.1.2 Bioinformatics databases 19  
  2.1.3 Applications for Bioinformatics 20  
2.2 Enzyme catalysts and enzyme kinetics 22  
  2.2.1 Enzyme catalysts 22  
  2.2.2 Enzyme kinetics 26  
  2.2.3 High throughput enzyme assay 29  
2.3 Database for enzyme assay 35  
  2.3.1 Database availability 35  
  2.3.2 Database design 35  

Chapter 3 Enzyme Catalytic Activity Mapping Database 36  
3.1 Enzyme catalytic assay 36  
  3.1.1 Enzyme Preparation 36  
  3.1.2 Chemical preparation 38  
  3.1.3 Assay reaction operation and data collection 39  
  3.1.4 Data processing 39  
3.2 Use case analysis 41  
  3.2.1 Database administration 42  
  3.2.2 Internal scientists 43
Chapter 3

3.2.2.1 Information collection 43
3.2.2.2 Information application 44
3.2.3 External scientists 47
3.2.4 Bioinformatists 47
3.3 Main modules of the database 48
3.3.1 Enzyme activity mapping module 48
   3.3.1.1 Reaction table 48
   3.3.1.2 KineticParameter table 50
3.3.2 Enzyme module 51
3.3.3 Chemical module 54
3.3.4 Experimental module 56
3.3.5 Data module 58
3.4 Primary keys and foreign key constraints 59

Chapter 4 Concluding Remarks 60
4.1 Conclusion 60
4.2 Contribution of the thesis 63
4.3 Suggestions for future work 63

References 65
Appendix 69
### List of Tables

<table>
<thead>
<tr>
<th>Table 3.1</th>
<th>List of fields in reaction table</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.2</td>
<td>List of fields in KineticParameter table</td>
<td>51</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>List fields in Enzyme table</td>
<td>52</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>List of fields in EnzymeStability table</td>
<td>53</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>List of fields in EnzymeSolution table</td>
<td>54</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>List of fields in EnzymePurification table</td>
<td>54</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>List of fields in Chemical table</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.8</td>
<td>List of fields in ChemicalSolution table</td>
<td>56</td>
</tr>
<tr>
<td>Table 3.9</td>
<td>List of fields in Experiment table</td>
<td>57</td>
</tr>
<tr>
<td>Table 3.10</td>
<td>List of fields in Protocol table</td>
<td>57</td>
</tr>
<tr>
<td>Table 3.11</td>
<td>List of fields in AssayData table</td>
<td>58</td>
</tr>
<tr>
<td>Table 3.12</td>
<td>List of fields in AssayResult table</td>
<td>59</td>
</tr>
</tbody>
</table>
List of Figures

**Figure 2.1** Activation Energy Comparison Between Catalyzed and Un-catalyzed Reaction 23
**Figure 2.2** Catalytic scheme of enzyme 23
**Figure 2.3** TPP acting as a coenzyme 25
**Figure 2.4** Transfer Reaction Solutions to 96-well Micro-Plate 31
**Figure 2.5** Profile of Product Concentration vs. Time 32
**Figure 2.6** Reaction Rate vs. Substrate Concentration 33
**Figure 2.7** Scheme of effect of pH on Initial Reaction Rate 34
**Figure 3.1** Enzyme catalytic assay workflow diagram 37
**Figure 3.2** Michaelis-Menton plot 41
**Figure 3.3** Lineweaver-Burk plot 41
**Figure 3.4** Use cases of EAMDB 42
**Figure 3.5** ER Model of EAMDB 49
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 interests in their application in various chemical reactions. 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 projects, a group of enzymes will be assayed against substrates of interests with the hope of finding enzyme catalysts suitable for application in 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 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-BBD⁵ are the main examples. The enzymes covered in LIGAND and BRENDA (BRaunschweiger ENzyme DAtabase)⁶ are
mainly in the category of metabolism central to functioning of cell life. UM-BBD\textsuperscript{5} is a database specifically focusing biodegradation pathway information.

The Ligand Chemical Database for Enzyme Reactions (\textbf{LIGAND})\textsuperscript{3} is designed to provide the linkage between chemical and biological aspects of life in the light of enzymatic reactions. \textbf{LIGAND} is part of the \textbf{KEGG}\textsuperscript{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 (\textbf{IUBMB}, 1992) and International Union of Pure and Applied Chemistry (\textbf{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 \textbf{KEGG/PATHWAY} database and in the ENZYME section, as are other compounds found in literature. The REACTION section is a collection of reactions, mostly enzymatic reactions, involving the compounds covered in the COMPOUND section.

\textbf{UM-BBD}\textsuperscript{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 \textbf{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) strings and with Molfiles. The two-dimensional chemical structures 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 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 performed organism specifically.

Information on some enzymes and their associated human diseases has been included into the **BRENDA** database.

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.

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 values. 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 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 presented a database design to store all the necessary information about enzyme catalytic assay research 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 experimental oriented. Biochemists can record all detailed experimental information, experimental data, and conclusion. 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 bioinformatics system covering the functional information of enzymes. The system can provide valuable information to chemists, biochemists, and bioinformatists.
1.5 Layout of the thesis

There are four chapters and one appendix in the thesis discussing the justification, backgrounds, database design and conclusions respectively, namely: Chapter 1 presents the justification of a new database. Chapter 2 presents the background knowledge of bioinformatics, and enzyme catalytic assay. 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 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 and nucleic acids – the information carrier. 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, constrain programming, formal language theory, neural nets, machine learning, motif 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 biomedicine and computer technology offers substantial benefits to all scientists involved in research of biology, biochemistry, and chemistry etc.

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 object 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 are believed to 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 protein or on the nucleic acid level. The results desired are
relationships between different proteins that allow for making conclusions about protein function, cellular localization and the like.

As 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 molecule 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 analysis and prediction of protein structures. Though many successes have been accomplished, a long journey is still right ahead of bioinformatics to solve the general problem. What can be achieved is to model (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 out 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 pockets.

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 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 molecules 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 much 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 and regulatory network is another major study subject 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 collection EMBL\(^9\) and GENEBank\(^10\), 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 contained 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 the enzyme will be active under.

2.1.3 Applications for Bioinformatics

The exponentially growing biological data initiated the development of bioinformatics. To solve the scientific challenges elaborated above, bioinformatics poses three aims since its beginning. First, 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. 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 and PSI-BLAST, 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 aiming at uncovering common principles that apply across many systems and highlight novel features while with traditional biological studies can only compare a few related systems.

Bioinformatics has some practical applications in 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 helps 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 not successful. 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 mostly 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. Enzymes are proteins that enable
thousands of essential chemical reactions in the body. Function assignment to enzymes is
an indispensable part of 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 the biochemistry. It is also necessary for the simulation and construction of
whole metabolic paths 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 $E_a$
is needed for a reaction to take place. The lower the $E_a$ the easier 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. The most striking characteristics of enzymes
are their catalytic power and specificity. They are proteins produced or derived from
some 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 a 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 and several orders of magnitude higher than chemically catalyzed reactions.  
(b) Enzyme-catalyzed reactions occur under mild reaction conditions (e.g. temperature, pressure, aqueous solutions and pH)  
(c) Enzymes have great 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 group at 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, 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 catalyze specific reactions.

Figure 2.3 TPP acting as a coenzyme

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 substrate. Inhibitors are one of major research subjects for bio-scientists. They are harmful for the application of enzymes as catalysts; but on the hand they may be effective drugs to cure some diseases.\textsuperscript{18}

2.2.2 Enzyme kinetics

Enzyme kinetic is to study the rate of an enzyme catalyzed reaction under various conditions. These conditions include pH, temperature, and substrate concentrations etc. In enzyme kinetic study, one of the most important and mostly used chemical reaction mechanisms is Michaelis-Menten mechanism.\textsuperscript{16} Michaelis-Menten mechanism is a general explanation of the velocity 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., when the substrate concentration is very high. These relationships provide the basis for all 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 chemical reaction equation, Michaelis-Menten mechanism for enzyme kinetics is expressed as:

\[
E + S \xrightleftharpoons[k_4]{k_1} ES \xrightarrow[k_2]{k_1} E + \text{products}
\]

Eq. 2.1
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 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:

\[
\frac{d[\text{product}]}{dt} = k_2[ES]
\]  

Eq. 2.2

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

\[
\frac{d[ES]}{dt} = k_1[E][S] - k_{-1}[ES] - k_2[ES] \approx 0
\]  

Eq. 2.3

Solve for [ES],

\[
[ES] = \frac{k_2[E][S]}{k_{-1} + k_2}
\]  

Eq. 2.4

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

\[
\text{Rate} = \frac{d[\text{product}]}{dt} = k_2 \frac{k_1[E][S]}{k_{-1} + k_2}
\]  

Eq. 2.5

The Michaelis-Menten constant \( K_M \) is defined by following equation:

\[
\frac{1}{K_M} = \frac{k_1}{k_{-1} + k_2}
\]  

Eq. 2.6

so that the rate becomes,

\[
\text{Rate} = \frac{k_2}{K_M} [E][S]
\]  

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, \([E]_0\), but

\[
[E]_0 = [E] + [ES] = [E] + \frac{[E][S]}{K_M}
\]

\(= [E] \left( 1 + \frac{[S]}{K_M} \right) \)  

Eq. 2.8

from which we obtain,

\[
[E] = \frac{[E]_0}{1 + \frac{[S]}{K_M}} \]

Eq. 2.9

The rate becomes, then,

\[
\text{Rate} = \frac{k_2}{K_M} [S] \frac{[E]_0}{1 + \frac{[S]}{K_M}}
\]

\(= k_2 \frac{[E]_0 [S]}{K_M + [S]} \)  

Eq. 2.10

Define the reaction velocity as \(v = \text{Rate}\). So,

\[
v = k_2 \frac{[E]_0 [S]}{K_M + [S]} \]

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. Define the maximum velocity, \(v_{\text{max}}\), as,
\[ v_{\text{max}} = \lim_{[S] \to \infty} \frac{k_2 [E]_0 [S]}{K_M + [S]} = k_2 [E]_0 \quad \text{Eq. 2.12} \]

then

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

Note that the kinetics of the reaction are characterized by two parameters, \( v_{\text{max}} \) 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.

**Eq. 2.13** is normally written in the following format to deal with experimental data,

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

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

\[ K_M = \text{slope} \times v_{\text{max}} = \frac{\text{slope}}{\text{intercept}} \quad \text{Eq. 2.15} \]

The assumptions for Michalis-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.\(^{19}\)

### 2.2.3 High throughput enzyme assay

High throughput enzyme assay is the process to screen 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 work. 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 common 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 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 some of 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 the 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 micro-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 of 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.\textsuperscript{16} To measure these two parameters, the \( V_{\text{max}} \) and \( K_M \) are determined for each enzyme-substrate reaction by applying Michaelis-Menten equation discussed above. The relationship between \( k_{\text{cat}} \), \( V_{\text{max}} \), and \( K_M \) can be expressed by the following equations:

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

\[
\text{Catalytic Efficiency} = \frac{k_{\text{cat}}}{K_M}
\]
In most cases, an enzyme converts one chemical (the substrate), into another (the product). 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 doesn’t 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 which 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](image-url)
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 $V_{\text{max}}$.

![Reaction Rate vs. Substrate Concentration](image)

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

$V_{\text{max}}$ 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 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 enzyme kinetic study. 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 under suitable temperature but may show no activity beyond this temperature range.

![Figure 2.7 Scheme of effect of pH on Initial Reaction Rate](image)

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 enzyme. 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 is accompanied by the appearance of many publicly available databases covering various biological information. However, there is no database to support enzyme assay experiments that is publicly available or accessible. This may due to the nature of the information. Experimental detail information sometimes is considered sensitive especially when intellectual right is involved. While recording experimental information is gaining its popularity, each companies 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, relational data model is much better developed although there are limitations.

The design of a database normally consists of four major activities: 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 is one of the important processes used to understand the biochemical pathways, to give information about the enzyme catalytic properties, and to identify potential pharmacophores and inhibitors. The essential piece of these experiments is the measurement of the kinetic parameters ($k_{\text{cat}}$, $K_m$, $K_i$). These enzyme kinetic parameters can be evaluated 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, and solution preparation. There are lots of enzyme purification protocols categorized in column chromatography, solvent extraction, and re-crystallization etc. The main goal of the purification is to remove the contaminants that may deteriorate the required properties and stabilities of the enzyme or may be incompatible with further experimental procedures. This step is sometimes 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 that the purified enzyme is much less active or simply not active at all. To determine the effect of the purification on the enzyme activity, assay experiments need to be performed. The result of the assay is then serving 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.)

Enzyme stability study is another prerequisite for the appropriate application of enzymes as catalysts. The enzyme activity and stability is not only under the effect of impurities but very often is sensitive to environmental conditions, such as temperature,
pH. While the effect of experimental condition 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 various conditions on the stability of the enzyme. An enzyme can be very stable in a certain range of pH but decompose seriously at other pH of the solution. Some enzymes can be stored at a temperature above freezing point while others need to survive under a temperature as low as –20 °C. Solvent is another effect on 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 usually rare and expensive. Stability study will provide the information about how the enzyme should be stored and under what kind of conditions the enzyme should be used. The results of the study should also be applied to the design of enzyme catalytic assay, as the experiments should avoid the conditions that are not practical to the application of certain enzymes.

In enzyme catalytic assay, 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 study. The stability study should conclude which kind of solvent, what range of pH should be avoided if any applicable.

3.1.2 Chemical preparation

Chemicals to the interest of enzyme catalytic assay are usually commercially available in appropriate forms and purification is not part of the concern for the assay. The final goal of a specific assay might aim at the application of the enzyme to the chemicals from a non-commercial source. In this case purification prior to the assay is
more inappropriate. However, the solvent and pH of the 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 chemical solution and to initiate the catalytic reaction. Usually an instrument with parameters pre-set 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 that 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 used to interpret the concentration change of the starting material or product.

### 3.1.4 Data processing

There are three categories of data for enzyme catalytic assay: data of enzyme stability study, data of optimal condition study, and data of enzyme kinetic 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 optimal condition study can also be analyzed by direct comparison. A graph of the initial reaction rate against the change of the condition in concern such as pH and temperature is normally applied 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 with \(K_m\) and \(V_{max}\) has been expressed in the forms of Michaelis-Menton equation and Lineweaver-Burke equation as discussed in Chapter 2. Lineweaver-Burke equation was obtained by the rearrangement of Michaelis-Menton equation.

\[
\text{The Michaelis-Menton Equation}
\]
\[\nu = \frac{V_{max} [S]}{K_m + [S]}\]

\[
\text{The Lineweaver-Burke Equation}
\]
\[\frac{1}{\nu} = \frac{K_m}{V_{max}} \frac{1}{[S]} + \frac{1}{V_{max}}\]

The values \(K_m\) and \(V_{max}\) are normally derived by using Michaelis-Menton plot (Figure 3.2) or Lineweaver-Burke plot (Figure 3.3).
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 ways is far superior.
There are four groups of potential users of this database system: database administrator, internal scientists, external scientists, and bioinformaticians.

![Diagram of EAMDB Use Cases]

**Figure 3.4 Use cases of EAMDB**

### 3.2.1 Database administration

Database administrator has full control of the database. Its 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.
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 the 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 \textit{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 scientists working on chemical is the properties related to chemical reactions, \textit{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 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 occurred during the data processing, they are needed for result verification.

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

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, the storage such information for EAMDB is not optional for a academic research lab due to the issue of intellectual 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. 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 substitute and enzyme.

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

Information about the enzyme stability in various chemical environments, including such factors as pH, concentrations, buffer constitutes, presence of activators and/or inhibitors, and organic solvents. 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, it is also needed that a scientist 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 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 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 as part of internal scientist activities and can be served by the information excluding the experimental part.

### 3.2.4 Bioinformaticians

EAMDB is of interest to bioinformatician. Enzyme catalytic properties are determined by the enzyme structure. Two enzymes of very similar structures 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 their various needs discussed above, the database should be composed of following modules: enzyme activity mapping module, enzyme module, chemical module, experiment module, and data 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: Reaction table, and KineticParameter table.

3.3.1.1 Reaction 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 information is indicated in the field of References. The chemical information of the main substrates and products is stored in
the chemical module. The fields and the corresponding descriptions are listed in Table 3.1.

Figure 3.5 ER Model of EAMDB
Table 3.1. List of fields in reaction table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReactionId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>SubstrateId</td>
<td>varchar (100)</td>
<td>Main substrate in the reaction, referenced to Chemical table</td>
</tr>
<tr>
<td>ProductId</td>
<td>varchar (100)</td>
<td>Major product in the reaction, referenced to Chemical table</td>
</tr>
<tr>
<td>ReactionEquation</td>
<td>varchar (200)</td>
<td>Full reaction equation</td>
</tr>
<tr>
<td>ReactionType</td>
<td>varchar (30)</td>
<td>Classification of the reaction</td>
</tr>
<tr>
<td>References</td>
<td>varchar (100)</td>
<td>Literature sources of the reaction</td>
</tr>
</tbody>
</table>

3.3.1.2 KineticParameter table

The information in KineticParameter table covers the enzyme identity information, reaction information, and major catalytic activity parameters (i.e. $K_m$, $k_{cat}$), 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 also shows negative results. The information may be summarized from the results of several experiments. In that case the related experiments will be identified. The information can also be extracted from a public literature with which the source of the information will also be indicated.

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 out this enzyme can be applied as a catalyst for which substrate.
Table 3.2. List of fields in KineticParameter table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KineticId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>ReactionId</td>
<td>varchar (20)</td>
<td>Foreign key from Reaction table in this module</td>
</tr>
<tr>
<td>EnzymeId</td>
<td>varchar (20)</td>
<td>Foreign key from Enzyme table in enzyme module</td>
</tr>
<tr>
<td>pHRange</td>
<td>varchar (20)</td>
<td>pH range within which the specified catalytic activity is obtained</td>
</tr>
<tr>
<td>pHOptimum</td>
<td>varchar (8)</td>
<td>pH at which the best catalytic activity is obtained</td>
</tr>
<tr>
<td>TempRange</td>
<td>varchar (20)</td>
<td>Temperature range within which the specified catalytic activity is obtained</td>
</tr>
<tr>
<td>TempOptimum</td>
<td>varchar (20)</td>
<td>The temperature at which the best catalytic activity is obtained</td>
</tr>
<tr>
<td>Cofactor</td>
<td>varchar (20)</td>
<td>Co-enzyme or cofactor used in the reaction</td>
</tr>
<tr>
<td>Activator</td>
<td>varchar (20)</td>
<td>The activator used in the reaction</td>
</tr>
<tr>
<td>Km</td>
<td>varchar (10)</td>
<td>Km value</td>
</tr>
<tr>
<td>kcat</td>
<td>varchar (10)</td>
<td>Turnover Number of the enzyme in the reaction</td>
</tr>
<tr>
<td>SpecificActivity</td>
<td>varchar (20)</td>
<td>Specific activity is a unit to express the amount of enzyme</td>
</tr>
<tr>
<td>AssayResultsId</td>
<td>varchar (50)</td>
<td>List of AssayResultsIds. AssayResultsId is the primary key for the AssayResults table in the data module.</td>
</tr>
<tr>
<td>Reference</td>
<td>varchar (200)</td>
<td>Information source other than our own experiments</td>
</tr>
</tbody>
</table>

3.3.2 Enzyme module

As some of the enzyme information, such as sequence and structure, has been dealt in the biological information system, this module is mainly for the information related to chemical assay, namely enzyme purification, enzyme solution, and enzyme stability. There are four tables in this module: enzyme table, EnzymeStability table, EnzymeSolution table, and EnzymePurification table.
Enzyme table is to store the information about the enzyme identity information. Although IUPAC and IUBMB have recommended a nomenclature for enzymes, due to historical reasons and scientist personal preferences, the same enzyme may be referred to with different names. Enzyme table try to cover all the possible names used for the enzyme. The class of the enzyme is also indicated. The fields and their description are listed in Table 3.3.

**Table 3.3. List fields in Enzyme table**

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

EnzymeStability table stores the information about stability of the enzyme under various conditions, such as pH, temperature, and solvent etc. Enzyme catalytic information is important for optimizing conditions for the enzyme catalytic reactions. The fields and their descriptions are listed in Table 3.4.
EnzymeSolution table stores the information about the enzyme solution used in experiments. The information covers the concentration of the enzyme, solvent, and pH etc. The fields and their descriptions are listed in Table 3.5.

Pure enzymes are required for the kinetic study of enzyme-catalyzed reactions. Each enzyme needs specific purification strategy. 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 EnzymePurification table and their description are listed in Table 3.6.

Table 3.4. List of fields in EnzymeStability table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>StabilityId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>EnzymeId</td>
<td>varchar (20)</td>
<td>Foreign key from Enzyme table</td>
</tr>
<tr>
<td>pHStability</td>
<td>varchar (20)</td>
<td>Suitable pH for catalytic reaction</td>
</tr>
<tr>
<td>TempStability</td>
<td>varchar (20)</td>
<td>Suitable temperature for catalytic reaction</td>
</tr>
<tr>
<td>GeneralStability</td>
<td>varchar (300)</td>
<td>This field summarizes general information on stability, e.g., increased stability of immobilized enzymes, stabilization by SH-reagents, detergents etc.</td>
</tr>
<tr>
<td>SolventStability</td>
<td>varchar (150)</td>
<td>List of types of suitable solvents for storage and reaction, e.g. water-miscibility, hydrophobicity, polarity</td>
</tr>
<tr>
<td>OxidationStability</td>
<td>varchar (100)</td>
<td>Stability in the presence of oxidizing agents, e.g. O₂, H₂O₂</td>
</tr>
<tr>
<td>StorageStability</td>
<td>varchar (100)</td>
<td>Suitable pH and temperature for storage</td>
</tr>
<tr>
<td>ExperimentId</td>
<td>varchar (20)</td>
<td>The experiments for determine the stability properties</td>
</tr>
</tbody>
</table>
**Table 3.5.** List of fields in EnzymeSolution Table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnzymeSolutionId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>EnzymeId</td>
<td>varchar (20)</td>
<td>Foreign key from the Enzyme table in this module</td>
</tr>
<tr>
<td>Solvent</td>
<td>varchar (20)</td>
<td>Solvent used in the solution</td>
</tr>
<tr>
<td>PH</td>
<td>varchar (5)</td>
<td>PH value of the solution</td>
</tr>
<tr>
<td>Buffer</td>
<td>varchar (20)</td>
<td>Buffer solution used for the solution</td>
</tr>
<tr>
<td>Stabilizer</td>
<td>varchar (20)</td>
<td>Stabilizer if used to make the solution</td>
</tr>
<tr>
<td>Concentration</td>
<td>varchar (10)</td>
<td>Concentration of the enzyme</td>
</tr>
</tbody>
</table>

**Table 3.6.** List of fields in EnzymePurification table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnzymePurificationId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>EnzymeId</td>
<td>varchar (20)</td>
<td>Foreign key from the Enzyme table</td>
</tr>
<tr>
<td>Method</td>
<td>varchar (200)</td>
<td>Method used for the purification</td>
</tr>
<tr>
<td>Procedure</td>
<td>Text</td>
<td>Detail procedure</td>
</tr>
<tr>
<td>SpecificActivity</td>
<td>varchar (20)</td>
<td>Enzyme activity after purification</td>
</tr>
<tr>
<td>Yield</td>
<td>varchar (10)</td>
<td>Yield of the purification</td>
</tr>
<tr>
<td>Reference</td>
<td>varchar (200)</td>
<td>Reference list</td>
</tr>
</tbody>
</table>

### 3.3.3 Chemical module

This module is for the information of substrates and products. A wide range of information is needed to characterize the properties of a compound. As some of the properties are only to the interests of 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 reactions. There are two tables in this module: chemical table and solution table.

Chemical table holds the information of chemical structure, namely identity of the substrate, molecule formula, functional groups, stereo structure, physical properties, spectrum information, and 3D picture or drawing if available. The fields are listed in Table 3.7. 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 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>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemicalId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>ChemicalName</td>
<td>varchar (50)</td>
<td>Compound name used in the system</td>
</tr>
<tr>
<td>CaName</td>
<td>varchar (50)</td>
<td>Name registered with CA</td>
</tr>
<tr>
<td>IupacName</td>
<td>varchar (50)</td>
<td>Name according to IUPAC nomenclature</td>
</tr>
<tr>
<td>Synonym</td>
<td>varchar (50)</td>
<td>Other name used for the compound</td>
</tr>
<tr>
<td>CaNumber</td>
<td>varchar (20)</td>
<td>CA registry number of the compound</td>
</tr>
<tr>
<td>StructuralFormula</td>
<td>varchar (100)</td>
<td>Structural formula of the compound, e.g. CH₃CH₂ for ethane</td>
</tr>
<tr>
<td>Class</td>
<td>varchar (50)</td>
<td>Type of the compound, e.g. acid, ester etc.</td>
</tr>
<tr>
<td>MolecularWeight</td>
<td>varchar (10)</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>StereoCenter</td>
<td>varchar (50)</td>
<td>Location and configuration of the stereo center</td>
</tr>
<tr>
<td>FunctionalGroup</td>
<td>varchar (50)</td>
<td>Major functional groups</td>
</tr>
<tr>
<td>BoilingPoint</td>
<td>varchar (10)</td>
<td>Boiling point of the compound</td>
</tr>
<tr>
<td>MeltingPoint</td>
<td>varchar (10)</td>
<td>Melting point of the compound</td>
</tr>
<tr>
<td>Nmr</td>
<td>varchar (100)</td>
<td>Characteristic NMR peak chemical shift and split pattern</td>
</tr>
<tr>
<td>Ir</td>
<td>varchar (100)</td>
<td>Type and location of the characteristic IR peak</td>
</tr>
<tr>
<td>Reference</td>
<td>varchar (200)</td>
<td>List of references</td>
</tr>
</tbody>
</table>
In enzyme assay experiments, chemicals are used in the form of solutions. ChemicalSolution table stores the information about the chemical solutions used in the experiments, namely compound identity, solvent, and concentration (Table 3.8).

**Table 3.8. List of fields in ChemicalSolution table**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemicalSolutionId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>ChemicalId</td>
<td>varchar (20)</td>
<td>Foreign key from the compound table</td>
</tr>
<tr>
<td>Solvent</td>
<td>varchar (20)</td>
<td>Solvent used for the solution</td>
</tr>
<tr>
<td>pH</td>
<td>varchar (5)</td>
<td>pH value in the solution</td>
</tr>
<tr>
<td>Buffer</td>
<td>varchar (20)</td>
<td>Buffer solution used to make the solution</td>
</tr>
<tr>
<td>Concentration</td>
<td>varchar (10)</td>
<td>Concentration of the substrate</td>
</tr>
</tbody>
</table>

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 the great insight of 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.

**3.3.4 Experimental module**

Experimental module stores the information about experiments. There are two tables in this module: experiment table, and protocol table.

The experiment 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 graph or picture. They are stored as files in specified directories. The fields and their descriptions are listed in Table 3.9.

### Table 3.9. List of fields in Experiment table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExperimentId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>ExperimentTitle</td>
<td>varchar (100)</td>
<td>Brief description of the goal of the experiment</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
<td>Experiment date</td>
</tr>
<tr>
<td>Researcher</td>
<td>varchar (50)</td>
<td>Researcher’s name</td>
</tr>
<tr>
<td>ProtocolId</td>
<td>varchar (20)</td>
<td>Foreign key in the Protocol table</td>
</tr>
<tr>
<td>EnzymeSolutionId</td>
<td>varchar (20)</td>
<td>Foreign key in the EnzymeSolution table</td>
</tr>
<tr>
<td>ChemicalSolutionId</td>
<td>varchar (20)</td>
<td>Foreign key in the ChemicalSolution table</td>
</tr>
<tr>
<td>LayoutFileName</td>
<td>varchar (20)</td>
<td>The file name of the layout file</td>
</tr>
<tr>
<td>LayoutFilePath</td>
<td>varchar (20)</td>
<td>The file path of the file</td>
</tr>
<tr>
<td>DataFileId</td>
<td>varchar (20)</td>
<td>Foreign key in the DataFile table</td>
</tr>
<tr>
<td>RelatedExperimentId</td>
<td>varchar (30)</td>
<td>List of ExperimentId of the related experiments.</td>
</tr>
</tbody>
</table>

### Table 3.10. List of fields in Protocol table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProtocolId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>ProtocolName</td>
<td>varchar (50)</td>
<td>Name of the protocol</td>
</tr>
<tr>
<td>TargetEnzyme</td>
<td>varchar (50)</td>
<td>Target enzyme name</td>
</tr>
<tr>
<td>TargetChemical</td>
<td>varchar (50)</td>
<td>Target chemical name</td>
</tr>
<tr>
<td>Cofactor</td>
<td>varchar (50)</td>
<td>Cofactor used in the assay</td>
</tr>
<tr>
<td>Coenzyme</td>
<td>varchar (50)</td>
<td>Coenzyme used in the assay</td>
</tr>
<tr>
<td>Activator</td>
<td>varchar (50)</td>
<td>Activator used in the assay</td>
</tr>
<tr>
<td>CharacterizationMethod</td>
<td>varchar (50)</td>
<td>Measurement method for collecting the kinetic data, e.g. photo spectroscopy, titration, etc.</td>
</tr>
<tr>
<td>ProtocolDescription</td>
<td>Text</td>
<td>i.e. detailed procedure</td>
</tr>
<tr>
<td>References</td>
<td>varchar (200)</td>
<td>Original source of the protocol</td>
</tr>
</tbody>
</table>

Protocol table stores the information of all the technique protocols used for enzyme assay. 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 3.10.

3.3.5 Data module

The data module stores the information of the data obtained from the enzyme catalytic assay experiments. There are two tables in this module: AssayData table, and AssayResults.

The AssayData will store the information about the original data obtained from 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 3.11.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssayDataId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>ExperimentId</td>
<td>varchar (20)</td>
<td>Foreign key from experiment table in the experiment module</td>
</tr>
<tr>
<td>DataFileName</td>
<td>varchar (30)</td>
<td>File name of the data file</td>
</tr>
<tr>
<td>DataFilePath</td>
<td>varchar (50)</td>
<td>Directory path of the data file</td>
</tr>
</tbody>
</table>

AssayResults table is used to store the kinetic study results from the assay experiments. The information includes $K_m$, $V_{max}$, $K_{cat}$, 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 AssayResults table. The fields and their descriptions are listed in Table 3.12.

Table 3.12. List of fields in AssayResult table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssayResultId</td>
<td>varchar (20)</td>
<td>Primary key</td>
</tr>
<tr>
<td>AssayDataId</td>
<td>varchar (20)</td>
<td>Foreign key from AssayData table</td>
</tr>
<tr>
<td>GraphFileName</td>
<td>varchar (30)</td>
<td>File name of the graph to derive the results</td>
</tr>
<tr>
<td>GraphFilePath</td>
<td>varchar (50)</td>
<td>Directory path of the graph file</td>
</tr>
<tr>
<td>Km</td>
<td>varchar (10)</td>
<td>Km value</td>
</tr>
<tr>
<td>Kcat</td>
<td>varchar (10)</td>
<td>Kcat value</td>
</tr>
<tr>
<td>Vmax</td>
<td>varchar (10)</td>
<td>Vmax value</td>
</tr>
</tbody>
</table>

3.4 Primary keys and foreign key constraints

The database module is composed of five modules as discussed above covering. Each module is composed several tables. The relational database discussed 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: high selectivity (stereo-selectivity, region-selectivity, and chemo-selectivity), mild reaction conditions, high catalytic efficiency, and environmental friendliness. Their applications are growing rapidly. On the other hand, the lack of the 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 the direct reflect 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 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 pursue of environmental friendliness and economical catalyst for the pulp and paper industry. EAMDB is a relational database with 12 tables. According to the information to store, it consists of five parts: enzyme module, chemical module, experiment module, catalytic mapping module, and data module.
Enzyme module is to store the information about enzyme information 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.

Chemical 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 of chemical solutions.

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

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.

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 reaction table and KineticParameter table for the corresponding information.
EAMDB is designed for the usage of 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 environmental degradation etc. The information can also be applied to bioinformatics system development since the enzymes catalytic activity information can be interpreted as direct reflect of enzyme structures.

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 conclusions. 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 quickly and efficiently without compromising the integrity of the data in storage.
4.2 Contribution to knowledge

The thesis presents a database design to support the enzyme catalytic assay experiments. Unlike the other databases such as BRENDA, LIGAND, or UM-BBD, EAMDB presented in this thesis is experimental oriented. Scientists can document detail information of assay experiments. The information stored will cover chemical information, chemical reaction, 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 serve the needs of biologists with the functional information of enzymes. It is expected that the database will be an important part of the bioinformatics system with unique characteristics.

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 presentations of biochemists. The final product is expected to provide fully support to enzyme assay experiments with high accuracy, efficiency, and convenience. To fulfill this achievement, 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 and 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 application and user interface to enforce its functionality requirements, e.g. access privilege control.

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.
References


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

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