Thesis for the degree of Doctor of Philosophy

Towards Classification and Functional Description of Enzymes

A case study of feruloyl esterases

D.B.R.K. GUPTA UDATHA

Industrial Biotechnology
Department of Chemical and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2013

Towards Classification and Functional Description of Enzymes

A case study of feruloyl esterases

D.B.R.K. GUPTA UDATHA

ISBN 978-91-7385-770-3

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3451
ISSN 0346-718X

Industrial Biotechnology
Department of Chemical and Biological Engineering
Chalmers University of Technology
SE-41296 Gothenburg
Sweden

Telephone: +46 (0) 31-772 1000

Cover illustration:
Scheme to functionally describe an enzyme through substrate interactions and chemical features, by D.B.R.K. Gupta Udatha

Printed by Chalmers Reproservice
Gothenburg, Sweden 2013

To my family and friends…

“Touch a scientist and you touch a child”

(Ray Bradbury)

Preface

This PhD dissertation serves as a partial fulfillment of the requirements to obtain the PhD degree at the Department of Chemical and Biological Engineering, Chalmers University of Technology, Sweden. The PhD project was carried out between 2009 and 2012, under the supervision of Professor Lisbeth Olsson. Associate Professor Gianni Panagiotou acted as co-supervisor in 2009-2011. The major part of the research work during my PhD was focused on integration of chemical informatics, biophysical characterization and protein engineering for modeling the substrate specificity of feruloyl esterases. As a guest PhD student at the Technical University of Denmark for a period of 6 months, I initiated collaboration with Associate Professor Irene Kouskoumvekaki (CBS, Computational Chemical Biology group) for the integration of bioinformatics and cheminformatics tools towards functional enzyme classification schemes.

The knowledge and expertise gained through integration of *in silico*- and experimental- biology was also applied to other collaborative projects (which are outside the border of my PhD project) focused on metabolic engineering and functional genomics.

My PhD project was mainly funded by the Swedish Research Council (Vetenskapsrådet) however personal scholarships partially supported my course work and conference trips.

Dasaradhi Bala Rama Krishna Gupta Udatha

January 2013

Towards Classification and Functional Description of Enzymes

A case study of feruloyl esterases

D.B.R.K. GUPTA UDATHA
Industrial Biotechnology, Department of Chemical and Biological Engineering, Chalmers University of Technology

Abstract

The prediction of enzyme functionality from sequence or structure data remains a challenging task that can be best addressed by studying the structure-function relationships determined from previously available information. This thesis work was focused on developing a reliable classification and functional description for the feruloyl esterase (FAE) enzyme family, whose members' possess both structural and catalytic promiscuity. To establish functional subgrouping of feruloyl esterases a combination of computational and experimental resources was used. The major challenge for FAEs, which often share little sequence similarity to each other and show varied substrate specificity catalyzing the conserved reaction involving an ester bond, is to represent the function in a computationally accessible format. For the analysis of FAEs with overlapping and unique specificity to individual substrates there is a need to capture the chemical function in terms of overall substrate specificity. To meet this requirements, the classification of FAEs was performed by incorporating the information of sequence properties, common-feature based pharmacophore models and the knowledge of active-residue constellations of the FAE binding pockets. Using machine learning techniques an automated descriptor-based classification system for FAEs was proposed that resulted into 12 FAE families. Based on catalytic residue constellations these families were sub-grouped into 32 functionally distinct sub-families. The biological relevance of the descriptor based classification system was validated with experimental data obtained from biochemical and biophysical characterization of FAEs. Challenges in the selection of the appropriate docking algorithm and scoring function combination for the prediction of substrate specificity of FAEs were addressed using molecular docking approaches. The evaluation of 88 docking algorithm-scoring function combinations from leading commercial docking programs for substrate specificity predictions revealed large differences in their performances that could be attributed to the differences in properties of the target proteins. Using the combination of *in silico* approaches and enzymology, structure-function relationships of FAEs were probed, especially in case of an exceptional Multiple Nucleophilic Elbowed Esterase (MNEE) from *Sorangium cellulosum* with four functionally distinct and catalytically promiscuous active-sites. Finally, this thesis demonstrates the application of structure-function relationship studies to obtain insights on the promiscuity of enzymes in their evolutionary path and to explain their structure-activity changes in immobilization based biosynthetic reactions.

Keywords: feruloyl esterases, functional classification, enzyme promiscuity, molecular docking, descriptors, pharmacophore, catalytic triad, structure-function relationship, protein evolution, enzyme immobilization

List of Publications

This thesis is based on the work described in the following publications, referred to as Paper I-V in the text:

PAPER I: D.B.R.K. Gupta Udatha, Irene Kouskoumvekaki, Lisbeth Olsson, Gianni Panagiotou (2011)
The interplay of descriptor-based computational analysis with pharmacophore modeling reveals a new classification scheme for feruloyl esterases.
Biotechnology Advances 29(1):94-110.

PAPER II: D.B.R.K. Gupta Udatha, Nobuyoshi Sugaya, Lisbeth Olsson, Gianni Panagiotou (2012)
How well do the substrates KISS the enzyme? Molecular docking program selection for feruloyl esterases.
Scientific Reports 2:323.

PAPER III: D.B.R.K. Gupta Udatha, Valeria Mapelli, Gianni Panagiotou, Lisbeth Olsson (2012)
Common and distant structural characteristics of feruloyl esterase families from *Aspergillus oryzae*.
PLoS ONE 7(6): e39473.

PAPER IV: D.B.R.K. Gupta Udatha, Karina Marie Madsen, Gianni Panagiotou, Lisbeth Olsson
Multiple nucleophilic elbows leading to multiple active sites in a single module esterase from *Sorangium cellulosum*.
Submitted for Publication.

PAPER V: Christian Thörn$^{=}$, D.B.R.K. Gupta Udatha$^{=}$, Hao Zhou, Paul Christakopoulos, Evangelos Topakas, Lisbeth Olsson
Understanding the pH dependent immobilization efficacy of Feruloyl esterase-C on mesoporous silica and its structure-activity changes.
Submitted for Publication.
=These authors contributed equally to this work

Throughout the course of this doctoral research contributions have been made to additional publications that are not included in this thesis. These publications are listed below:

PAPER VI: Otero JM, Papadakis MA, D.B.R.K. Gupta Udatha, Nielsen J, Panagiotou G (2010)
Yeast biological networks unfold the interplay of antioxidants, genome and phenotype, and reveal a novel regulator of the oxidative stress response.
PLoS ONE 5(10):e13606.

PAPER VII: Karina M. Madsen[=], D.B.R.K. Gupta Udatha[=], Saori Semba[=], Jose M. Otero, Peter Koetter, Jens Nielsen, Yutaka Ebizuka, Tetsuo Kushiro, Gianni Panagiotou (2011)
Linking genotype and phenotype of *Saccharomyces cerevisiae* strains reveals metabolic engineering targets and leads to triterpene hyper-producers.
PLoS ONE. 6(3): e14763.
=These authors contributed equally to this work

PAPER VIII: D.B.R.K. Gupta Udatha, Karina M. Madsen, Lisbeth Olsson, Gianni Panagiotou (2012)
Synthesis of adenosine ferulate using a multiple nucleophilic elbowed enzyme from *Sorangium cellulosum*. *Manuscript in preparation.*

PAPER IX: Evangelos Topakas[=], Melanie Z. Khodaie[=], D.B.R.K. Gupta Udatha[=], Kasper Jensen, Margarita Salazar, Mikael R. Anderssen, Lisbeth Olsson, and Gianni Panagiotou (2012)
Mapping the plant cell wall degradation capacity of *Aspergillus oryzae* using gene expression signatures.
Manuscript in preparation.
=These authors contributed equally to this work

Book Chapter

PAPER X: D.B.R.K. Gupta Udatha, Simon Rasmussen, Thomas Sicheritz-Pontén, Gianni Panagiotou.
Targeted metabolic engineering guided by computational analysis of Single Nucleotide Polymorphisms (SNPs). *Systems Metabolic Engineering: Methods and Protocols, Methods in Molecular Biology,* vol. 985, Springer (2013).

Contribution summary

A summary of my contributions to each of the above listed publications is given below:

PAPER I: Designed and performed research; analyzed data; wrote the manuscript.

PAPER II: Designed and performed research; analyzed data; wrote the manuscript.

PAPER III: Designed and performed research; analyzed data; wrote the manuscript.

PAPER IV: Designed and performed research; analyzed data; wrote the manuscript.

PAPER V: Designed and performed *in silico* research work; analyzed data; co-authored the manuscript.

PAPER VI: Designed and performed *in silico* research work; analyzed data; co-authored the manuscript.

PAPER VII: Designed and performed *in silico* research work; analyzed data; co-authored the manuscript.

PAPER VIII: Designed and performed research; analyzed data; co-authored the manuscript.

PAPER IX: Designed and performed *in silico* research work; analyzed data; co-authored the manuscript.

PAPER X: Co-authored the book chapter.

TABLE OF CONTENTS

INTRODUCTION

The most difficult challenge in catalysis has been solved by living organisms through harnessing the specificity and reactivity of enzymes to build and degrade diverse molecules. Enzymes, the cell's own catalysts, are challenging to understand in detail knowing that the efficiency of enzyme catalyzed reactions can reach ~10^9 M^{-1} sec^{-1} (i.e. k_{cat}/K_M) with the substrate in solution (Wolfenden & Snider, 2001). This catalytic ability of an enzyme is located in its binding pocket or cavity, called active site (Koshland, 1958) that varies among different enzymes in size, shape and the constellation of catalytically active group of amino acid residues. These structural differences in enzymes are the roots for variations in the reactions they catalyze.

The presence of diverse substrates or nutrients in the habitat drives the evolution of species by imparting selection for new functions on enzymes to metabolize or recognize nutrients or toxic compounds in the environment, hence new enzyme activities arise in species that adapt to changing environments (Hegeman & Rosenberg, 1970). In fact, the plasticity of enzymes to attain new functions in the path of evolution has allowed living organisms to flourish in diverse environments (Zalatan & Herschlag, 2009). Even though a hypothesis that has been proposed back in 1976 (Jensen, 1976) indicate that enzymes can catalyze secondary reactions in addition to the one they are evolved to catalyze, still several biochemistry books define these macromolecules as being highly specific. Along with high substrate and reaction selectivity many enzymes are known today to be able to process several substrates, a property called enzyme promiscuity (Hult & Berglund, 2007). Developments in enzymology, from an early focus on the catalytic mechanisms of individual enzymes to recent efforts to understand enzyme action in the context of dynamic and functional biological systems consisting of many interacting molecules, are continuously filling the gaps in our knowledge on the Darwinian assumption of 'one enzyme-one function' evolution under which every protein has evolved to perform a unique function that ultimately benefits the host organism (Reymond et al., 2009; Simon & Cravatt, 2010).

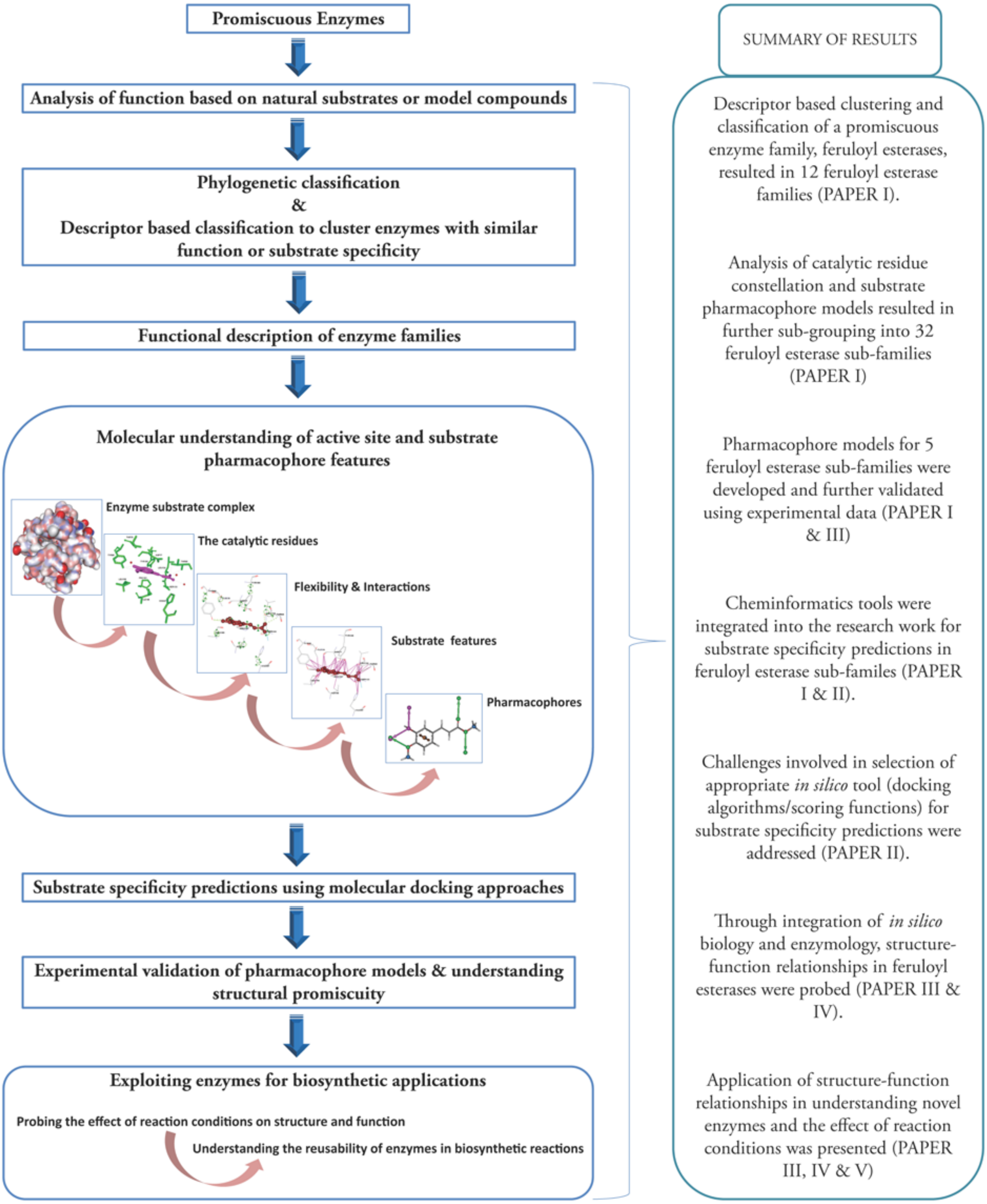
Enzymes are generally classified either based on function or sequence/structure similarity. Therefore, there are a few questions to be solved: does the functional promiscuity shown by enzymes correlates to sequence or structural promiscuity? How can enzymes with substrate promiscuity be classified to develop a toolbox for biocatalytic applications? Do promiscuous enzymes possess more than one active site or an active site with flexible catalytic residue constellations?

The work described in this thesis demonstrates ways to deal with and understand enzymes with functional promiscuity[1] (Carbonell & Faulon, 2010). I present how the combination of *in silico* approaches and protein biochemistry can be used to classify and explore enzyme families with functional promiscuity. Feruloyl esterases (FAEs) are taken as a case study in this thesis, as they are featured by broad substrate specificity, a property that has been exploited in biosynthetic applications (presented in **CHAPTER I**). The framework presented in **Figure 1** was followed to study the sequence-structure-function relationships in the feruloyl esterase group of enzymes and can be applied to understand any promiscuous enzyme family. The information provided on the molecular signatures for functional sub-classification of enzymes might also be of value for enzyme engineers in designing novel biocatalysts.

The specific aims of this work are:

1) To develop novel classification schemes to group enzymes into clusters or families that reflect their substrate specificity and to develop an automated classification system for enzyme families (**PAPER I**).
2) To develop substrate pharmacophores for the classified enzyme families and to further experimentally validate the developed pharmacophore features (**PAPER I** & **III**).

[1] Enzymes often possess the capability of functional promiscuity, i.e. to catalyze more than one reaction (catalytic promiscuity) or to show broad substrate specificity (substrate promiscuity).


Promiscuous Enzymes
Analysis of function based on natural substrates or model compounds
Phylogenetic classification
&
Descriptor based classification to cluster enzymes with similar function or substrate specificity
Functional description of enzyme families
Molecular understanding of active site and substrate pharmacophore features
Enzyme substrate complex
The catalytic residues
Flexibility & Interactions
Substrate features
Pharmacophores
Substrate specificity predictions using molecular docking approaches
Experimental validation of pharmacophore models & understanding structural promiscuity
Exploiting enzymes for biosynthetic applications
Probing the effect of reaction conditions on structure and function
Understanding the reusability of enzymes in biosynthetic reactions
SUMMARY OF RESULTS

Descriptor based clustering and classification of a promiscuous enzyme family, feruloyl esterases, resulted in 12 feruloyl esterase families (PAPER I).

Analysis of catalytic residue constellation and substrate pharmacophore models resulted in further sub-grouping into 32 feruloyl esterase sub-families (PAPER I)

Pharmacophore models for 5 feruloyl esterase sub-families were developed and further validated using experimental data (PAPER I & III)

Cheminformatics tools were integrated into the research work for substrate specificity predictions in feruloyl esterase sub-familes (PAPER I & II).

Challenges involved in selection of appropriate *in silico* tool (docking algorithms/scoring functions) for substrate specificity predictions were addressed (PAPER II).

Through integration of *in silico* biology and enzymology, structure-function relationships in feruloyl esterases were probed (PAPER III & IV).

Application of structure-function relationships in understanding novel enzymes and the effect of reaction conditions was presented (PAPER III, IV & V)

Figure 1. A combined approach of *in silico* biology and enzymology towards classification, structural and functional analysis of enzymes with catalytic promiscuity. The steps involve the classification of enzymes into functional groups based on primary amino acid sequences followed by the molecular understanding of enzyme and substrate structures for substrate specificity predictions to be used in biosynthetic applications. Brief summary of results is shown at the respective steps of the scheme.

3) To predict substrate specificity of enzymes using *in silico* approaches and selection of efficient molecular docking programs to explore overlapping substrate promiscuity in enzymes (**PAPER II**).
4) To probe the relation between structural and functional promiscuity in enzymes (**PAPER III** & **IV**).
5) To probe the structure-function changes during enzyme immobilization process for biosynthetic applications (**PAPER V**).

I have divided the first part of my thesis into four chapters, which are briefly introduced below. The second part of my thesis contains the articles (**PAPER I-V**), which have been published or submitted for publication. The articles cover the different steps illustrated in **Figure 1**.

CHAPTER 1 gives an overview and describes limitations of the enzyme classification system introduced by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. In the same chapter there is an introduction to the Carbohydrate Active Enzymes and feruloyl esterases are presented, which have been the focus of my PhD thesis work.

CHAPTER 2 introduces the novel classification approaches for promiscuous enzyme families and has been written as a background for the work published in **PAPER I. PAPER I** reviews the literature regarding the hydrolytic and synthetic specificities of FAEs generated via a variety of enzymatic assays. In order to assay FAE activity, researchers have used different model substrates and the information available from recent works on hydrolytic specificity of FAEs have challenged the previously proposed (empirical) classification system that was based on the specificity for only four substrates (Crepin et al., 2004). I apply an array of computational tools and succeeded to develop a new classification scheme for FAEs, which can be selectively used in biocatalytic transformations. In addition, I demonstrate that amino acid sequence information can be used to develop models that are able to suggest the underlying structural characteristics that determine substrate specificities.

Virtual screening of compound libraries will suggest opportunities to use members of specific enzyme groups on completely novel substrates. There is a large number of docking programs available for virtual screening of compounds; while new programs are released by the molecular docking companies every year, many existing programs are upgraded with new technology (i.e. docking algorithms and scoring functions). A number of docking program evaluations in recent years has indicated that different docking algorithms and scoring functions showed performances that are target specific. This is an important issue that has not been well addressed so far in the literature. In this work, Different commercial docking programs-scoring function sets are evaluated for the selection of a best program-scoring function that can reproduce the experimental substrate specificity of FAE families. A novel framework and assessment measure for the evaluation and selection of molecular docking programs for a specific protein of interest is proposed in **PAPER II**. In the work published as **PAPER III**, I investigate the structure-function relationships of three recombinant enzymes by collecting experimental data determining the substrate specificity, enzyme activity and biophysical characterization using circular dichroism (CD) spectroscopy as a function of pH. By applying chemoinformatics tools, I succeeded to develop pharmacophore models for the respective FAE families and I further experimentally validated the pharmacophores proposed in **PAPER I**. In **CHAPTER 3**, I introduce the application of bio- and chemo-informatics tools to elucidate the substrate recognition mechanisms for catalytically promiscuous enzyme systems like FAEs. The sub-chapters of this part of the thesis focus on **PAPER II** and **PAPER III**.

Much has been described in literature about FAE enzymes; however, I strongly believe that **PAPER I**, **PAPER II** and **PAPER III** represent a significant contribution to the field in terms of actually *demonstrating* the structure-function complexity in this enzyme family.

In **CHAPTER 4**, I present a novel enzyme with multiple active sites and discuss about its enzyme evolutionary trajectories. The subchapters deal with **PAPER IV** and

PAPER V. In **PAPER IV**, I present the integration of *in silico* biology and enzymology to elucidate the interplay of multiple binding pockets of this special enzyme and its catalytic promiscuity. **PAPER V** deals with the structural features involved in the successful reuse of enzymes through enzyme immobilization and further the relationship of enzymatic activity to material properties to aid in the development of improved biocatalysts.

On the whole, this thesis comprises three themes. The first theme is the classification of enzymes into functional groups, where the sequence and the structural properties that reflect their function are explored. The second theme is the study of enzyme structure-function relationships with the integration of *in silico* approaches and protein biochemistry, in which a case study showed that multi-functional enzymes emerge in the path of enzyme evolution. The third theme is about the application of the tools to study the enzyme structure-function relationships to design biosynthetic processes.

CHAPTER I

Enzyme Classification

Grouping enzymes in different classes based on the reaction/type of reaction they catalyze is a possible way to gain an understanding of the bonds they create or break. Ambiguities in the terms used for enzymes according to their function might cause a great deal of confusion. Generally, the suffix 'ase' has been added to the name of the substrate the enzyme acts on (e.g. Urease) or the name of the enzyme gives some indication of the reaction it catalyzes (e.g. glucose oxidase). Furthermore, names with no indication of the reaction catalyzed or the substrate involved still persist (e.g. rhodanase, barnase). There are about 20 different enzymes in the human liver that have been coined the same name 'alcohol dehydrogenase' but show specificity for primary aliphatic alcohols with different chain length. The need of universally accepted grouping for enzymes has given birth to the classification system by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) that gives each enzyme a four-digit Enzyme Commission (EC) number denoting the reaction type. The terms of EC, established by the IUBMB in 1956 were '*To consider the classification and nomenclature of enzymes and coenzymes, their units of activity and standard methods of assay, together with the symbols used in the description of enzyme kinetics*' (IUBMB, 1965). In 1958, the EC reported the task of grouping enzymes to a satisfactorily accepted level into six enzyme classes, namely Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases (**Table 1**). Later, in 1964 the enzyme classification system has been published as a book (Dixon & Webb, 1964). To fit the enzymes under the EC classification scheme, several *subclasses* were made under each *class* of enzyme. With the increased research reports on substrate specificity of enzymes several sub-subclasses have been created under each sub-class. Enzyme classification is constantly developing and one current issue is that the recommendations for enzyme classification and nomenclature are inappropriate for several enzyme groups (e.g. carbohydrate-active enzymes), especially in case of enzymes with multiple substrate specificity and for isoenzymes.

The IUBMB enzyme classification system gives us a starting point of information, which is a tribute to the perseverance of recommendations set by the Enzyme Commission more than five decades ago. The enzyme classification system is being constantly updated with new enzymes or corrections to existing entries and the details of recommendations for enzyme classification are provided at the World Wide Web: http://www.chem.qmul.ac.uk/iubmb/enzyme/ <23 October. 2012>

Efforts to understand the sequence-structure-function relationships in enzymes and their classification have given rise to online enzyme databases that use several bioinformatics approaches. For example, BRENDA (Schomburg et al., 2002) provides information on experimental results; whereas relational databases, like KEGG ENZYME (Kanehisa, 1997), depend on the combination of *in silico* approaches providing additional annotations from sequence data links. Sequence comparison is the most common method of assigning functions to novel proteins, however, it has been shown that more than 60% of global sequence identity is required to functionally annotate novel proteins with 90% accuracy (Tian & Skolnick, 2003). Nevertheless, there are several counterexamples that render the sequence identity thresholds inappropriate (Babbitt, 2003).

Experimental structural biology efforts provide information of 3-dimensional (3D) structure for proteins with insights into the functional relationships that could not be found by primary structure analysis. A 3D structure or an amino acid sequence alone is not enough to assign an EC number to a protein due to the fact that the EC classification system has been developed before the era of sequencing and crystallography. The EC nomenclature for enzymes is based on their substrate specificity and the type of reaction they catalyze. For example, the EC number for feruloyl esterase is 3.1.1.73, where the first digit indicates that it belongs to the enzyme *class* hydrolases,

Table 1. The list of enzyme classes and their sub-classes according to the IUBMB Enzyme nomenclature.

Class	Subclass	Name
EC 1		**Oxidoreductases**
	EC 1.1	Acting on the CH-OH group of donors
	EC 1.2	Acting on the aldehyde or oxo group of donors
	EC 1.3	Acting on the CH-CH group of donors
	EC 1.4	Acting on the CH-NH_2 group of donors
	EC 1.5	Acting on the CH-NH group of donors
	EC 1.6	Acting on NADH or NADPH
	EC 1.7	Acting on other nitrogenous compounds as donors
	EC 1.8	Acting on a sulfur group of donors
	EC 1.9	Acting on a heme group of donors
	EC 1.10	Acting on diphenols and related substances as donors
	EC 1.11	Acting on a peroxide as acceptor
	EC 1.12	Acting on hydrogen as donor
	EC 1.13	Acting on single donors with incorporation of molecular oxygen (oxygenases)
	EC 1.14	Acting on paired donors, with incorporation or reduction of molecular oxygen
	EC 1.15	Acting on superoxide radicals as acceptor
	EC 1.16	Oxidising metal ions
	EC 1.17	Acting on CH or CH_2 groups
	EC 1.18	Acting on iron-sulfur proteins as donors
	EC 1.19	Acting on reduced flavodoxin as donor
	EC 1.20	Acting on phosphorus or arsenic in donors
	EC 1.21	Acting on X-H and Y-H to form an X-Y bond
	EC 1.22	Acting on halogen in donors
	EC 1.97	Other oxidoreductases
EC 2		**Transferases**
	EC 2.1	Transferring one-carbon groups
	EC 2.2	Transferring aldehyde or ketonic groups
	EC 2.3	Acyltransferases
	EC 2.4	Glycosyltransferases
	EC 2.5	Transferring alkyl or aryl groups, other than methyl groups
	EC 2.6	Transferring nitrogenous groups
	EC 2.7	Transferring phosphorus-containing groups
	EC 2.8	Transferring sulfur-containing groups
	EC 2.9	Transferring selenium-containing groups
	EC 2.10	Transferring molybdenum- or tungsten-containing groups
EC 3		**Hydrolases**
	EC 3.1	Acting on ester bonds
	EC 3.2	Glycosylases
	EC 3.3	Acting on ether bonds
	EC 3.4	Acting on peptide bonds (peptidases)
	EC 3.5	Acting on carbon-nitrogen bonds, other than peptide bonds
	EC 3.6	Acting on acid anhydrides
	EC 3.7	Acting on carbon-carbon bonds
	EC 3.8	Acting on halide bonds
	EC 3.9	Acting on phosphorus-nitrogen bonds
	EC 3.10	Acting on sulfur-nitrogen bonds
	EC 3.11	Acting on carbon-phosphorus bonds
	EC 3.12	Acting on sulfur-sulfur bonds
	EC 3.13	Acting on carbon-sulfur bonds
EC 4		**Lyases**
	EC 4.1	Carbon-carbon lyases
	EC 4.2	Carbon-oxygen lyases
	EC 4.3	Carbon-nitrogen lyases
	EC 4.4	Carbon-sulfur lyases
	EC 4.5	Carbon-halide lyases
	EC 4.6	Phosphorus-oxygen lyases
	EC 4.99	Other lyases
EC 5		**Isomerases**
	EC 5.1	Racemases and epimerases
	EC 5.2	*cis-trans*-Isomerases
	EC 5.3	Intramolecular isomerases
	EC 5.4	Intramolecular transferases (mutases)
	EC 5.5	Intramolecular lyases
	EC 5.99	Other isomerases
EC 6		**Ligases**
	EC 6.1	Forming carbon—oxygen bonds
	EC 6.2	Forming carbon—sulfur bonds
	EC 6.3	Forming carbon—nitrogen bonds
	EC 6.4	Forming carbon—carbon bonds
	EC 6.5	Forming phosphoric ester bonds
	EC 6.6	Forming nitrogen—metal bonds

the second digit denotes that it belong to the *subclass* of hydrolases that act on ester bonds, the third digit is for the *sub-subclass* of hydrolases that act on carboxylic esters, and the fourth digit defines its substrate specificity and indicates that feruloyl esterase catalyzes the hydrolysis of the feruloyl group from an esterified sugar or feruloyl-polysaccharide. Even though the EC number is useful to avoid ambiguities, it is not appropriate for the enzymes with structural and functional divergence (Udatha et al., 2012a). The EC system is based on qualitative description of the transformation catalyzed by the enzymes, and is too broad to consider the structure-function correlations. Therefore, a classification system that is solely based on substrates considers neither the evolutionary events nor the structural divergence of enzymes (Babbitt, 2003).

Recent studies probing the enzyme structure-function relationships have shown two major points: (i) a common ancestor often generates superfamilies of enzymes catalyzing a diversity of reactions through divergent evolution; (ii) the convergent evolution generates unrelated enzymes that catalyze the same type of reaction (Gerlt & Babbitt, 2001; Glasner et al., 2006; Omelchenko et al., 2010b). Several recently published case studies continue to provide evidences that all members of a superfamily possess at least one common mechanistic aspect linked to conserved features of their substrate binding pockets (Chiang et al., 2008; Linsky & Fast, 2010; Nowotny, 2009). For example, the carbohydrate esterases employ the conserved feature of Ser-His-Asp catalytic triad to catalyze the reactions, but the members within the enzyme family possess varied substrate specificity. For such enzyme groups that need sub-grouping beyond the EC four digit system of classification, it is worthwhile to adopt a combined system of sequence-structure-function, in which individual enzymes are assigned unique identifiers that reflects their substrate specificity. Such a database that follows the semi-automatic modular assignment for classification of enzymes is CAZy (Carbohydrate-Active EnZymes) database, where a module can be defined as a structural and functional unit (Cantarel et al., 2009).

1.1 Carbohydrate Active Enzymes

Enzymes that act on the structurally diverse, complex carbohydrates and glycoconjugate substrates are collectively designated as Carbohydrate-Active enZymes or CAZymes (Cantarel et al., 2009). As the diversity of carbohydrates exceeds the number of known protein folds, CAZymes have evolved from a small pool of ancestors by acquiring novel structural features and thus novel substrate specificities (Henrissat, 1991; Henrissat & Bairoch, 1993; Laine, 1994). The information on CAZymes is available at CAZy database (www.cazy.org). CAZymes are classified into four enzyme classes (**Table 2**) and class has been sub-grouped into multiple families. In addition to the four enzyme classes, CAZy database also contains the carbohydrate-binding module (CBM) family divided into 64 sub-families. A carbohydrate-binding module (CBM) is defined as a contiguous amino acid sequence within a carbohydrate-active enzyme with a discrete fold having carbohydrate-binding activity. So, a CBM can be an integral part of the enzymes present in the four CAZy classes.

Table 2. Number of enzyme families, classified and unclassified modules in the four CAZy Enzyme classes as per 10th December 2012.

CAZy Enzyme Class	Reaction type	Number of Families	Modules in present families	Non-Classified modules
Glycoside Hydrolases (GHs)	Hydrolysis and/or rearrangement of glycosidic bonds	131	133637	1542
GlycosylTransferases (GTs)	Formation of glycosidic bonds	94	101926	2289
Polysaccharide Lyases (PLs)	Non-hydrolytic cleavage of glycosidic bonds	22	3451	172
Carbohydrate Esterases (CEs)	Hydrolysis of carbohydrate esters	16	13891	1212

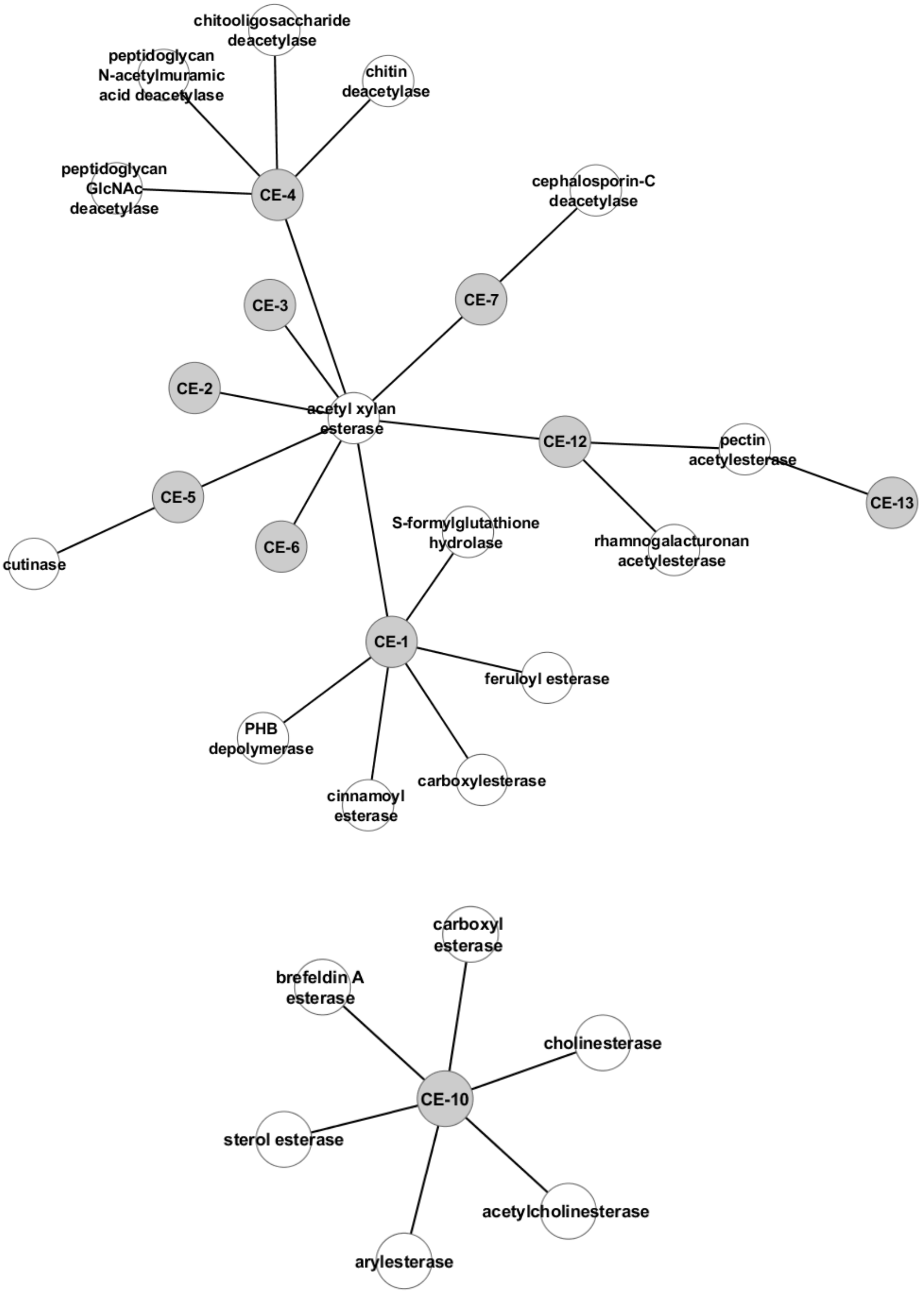


Figure 2. Overlapping and multiple enzyme activities among carbohydrate esterase families (CE-1, CE-2, CE-3….) according to the CAZy database; as presented in Supplementary File S1 of **PAPER IV**. Several carbohydrate esterase families contain proteins with different substrate specificity, but have been grouped together based on the sequence similarity of the conserved modules.

The feature used to classify the enzymes in CAZy is protein sequence similarity to experimentally characterized enzymes, which serves as a seed for the family that is gradually extended with sequences that share statistically significant (>85%) sequence similarity. As more enzymes are catalogued through the genome projects, the number and diversity of the sequences grow at a rapid pace, which further poses a challenge to CAZy for structure-function mapping. As shown in **Table 2**, the carbohydrate esterase class has comparatively low number of families and modules, but a large number of non-classified modules. The fact that members of some carbohydrate esterase families in the CAZy database are able to hydrolyze the substrates specific for other carbohydrate esterase families raises questions on the accuracy of the automated classification of carbohydrate esterases. The conserved modules of carbohydrate esterase families also possess overlapping substrate specificities as shown in **Figure 2**. With the increase in the gap between the automatically annotated and biochemically characterized sequences, the number of non-classified sequences and complexity of overlapping substrate specificity among the CAZy families also multiplies. Sub-classification of CAZy families based on the functional motifs or structural properties and constellation of the active sites may provide a possibility for a better functional classification system.

1.2 Feruloyl esterases

Feruloyl esterases (FAEs) fall under the sub-subclass E.C. 3.1.1 of hydrolases that catalyze the hydrolysis of carboxylic ester linkages in plant cell wall materials, releasing ferulic acid (FA) and other hydroxycinnamic acids (**Figure 3**).

IUBMB comments on FAEs: catalyze the hydrolysis of the 4-hydroxy-3-methoxycinnamoyl (feruloyl) group from an esterified sugar, which is usually arabinose in natural substrates. They are sometimes called hemicellulase accessory enzymes, since they "help" xylanases and pectinases to break down plant cell wall hemicellulose (http://www.chem.qmul.ac.uk/iubmb/enzyme/EC3/1/1/73.html <23 October. 2012>)

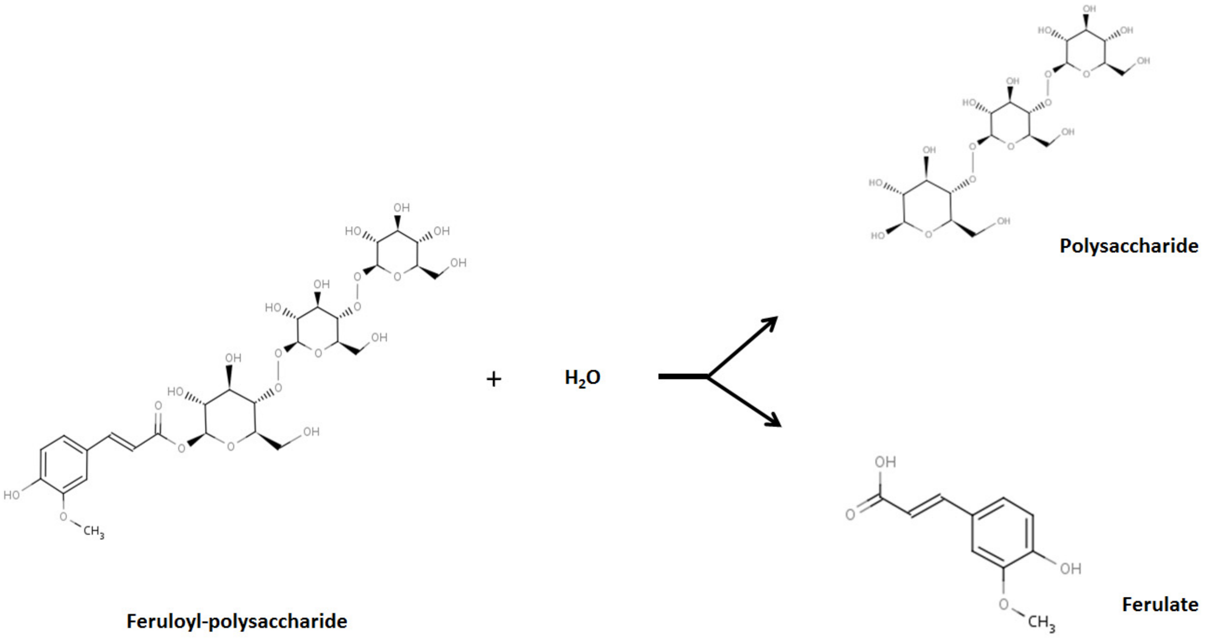


Figure 3. Hydrolysis of carboxylic ester linkage between a sugar polysaccharide and phenolic moiety by feruloyl esterase.

According to the CAZy database, FAEs falls under the family CE-I of Carbohydrate esterases. Since an ester = acid + alcohol, two classes of substrates for carbohydrate esterases exist: those in which the sugar plays the role of the "acid", such as pectin methyl esters and those in which the sugar behaves as the alcohol, such as in acetylated xylan. A number of possible reaction mechanisms may be involved: the most common is a Ser-His-Asp catalytic triad catalyzed deacetylation analogous to the action of classical lipase and serine proteases (Ekici et al., 2008a).

As described in **PAPER I**, Feruloyl esterases (FAEs) have gained importance in biofuel, medicine and food industries due to their capability of hydrolyzing carbohydrate esters in wood polymers and synthesizing high added-value molecules through esterification and transesterification reactions (Benoit et al., 2008; Koseki et al., 2009; Wong, 2006a). An abstracted version of the feruloyl esterase applications is shown in **Figure 4**.

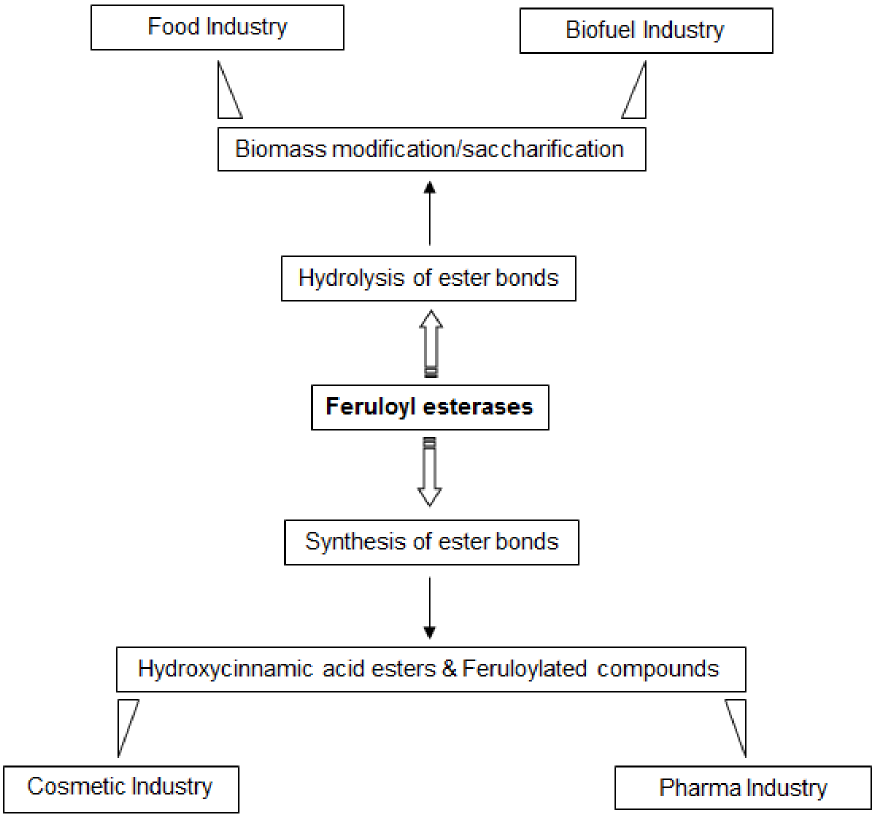

Food Industry
Biofuel Industry
Biomass modification/saccharification
Hydrolysis of ester bonds
Feruloyl esterases
Synthesis of ester bonds
Hydroxycinnamic acid esters & Feruloylated compounds
Cosmetic Industry
Pharma Industry

Figure 4. Applications of Feruloyl esterases. The hydrolytic and synthetic capacity of feruloyl esterases have been explored in various industries (Faulds, 2010; Fazary & Ju, 2007; Wong, 2006b).

Ferulic acid, one of the most abundant hydroxycinnamic acids liberated from the action of FAEs on agricultural by-products, has gained importance in food industry as it can be further transformed into vanillin, a flavouring food additive (Lesage-Meessen et al, 1996). Other types of hydroxycinnamic acids liberated from the action of FAEs have importance in cosmetic and pharmaceutical industries due to their antioxidant properties (Kikuzaki et al, 2002). During the last decade, FAEs have gained increased attention in the area of biocatalytic transformations for the synthesis of hydroxycinnamic acid esters with medicinal and nutritional applications. Feruloylation of D-arabinose by a FAE and its potential application as anti-mycobacterial agent has been demonstrated (Vafiadi et al, 2007b). Furthermore, the potential of a FAE as a synthetic tool of various phenolic esters and their inhibitory effect on LDL (Low-Density-Lipoproteins) oxidation has been investigated *in vitro* towards the prevention of atherosclerosis (Vafiadi et al, 2008).

Researchers have been generating data to determine the hydrolytic specificity of FAEs using several synthetic or model substrates such as substituted methyl cinnamate compounds. Different FAEs were able to catalyze the hydrolysis of model substrates with different specificities. For example, the three FAEs that fall under three different FAE sub-families of descriptor based classification system (**PAPER I**) possess both overlapping as well as unique specificity to the individual substrates shown in **Figure 5.**

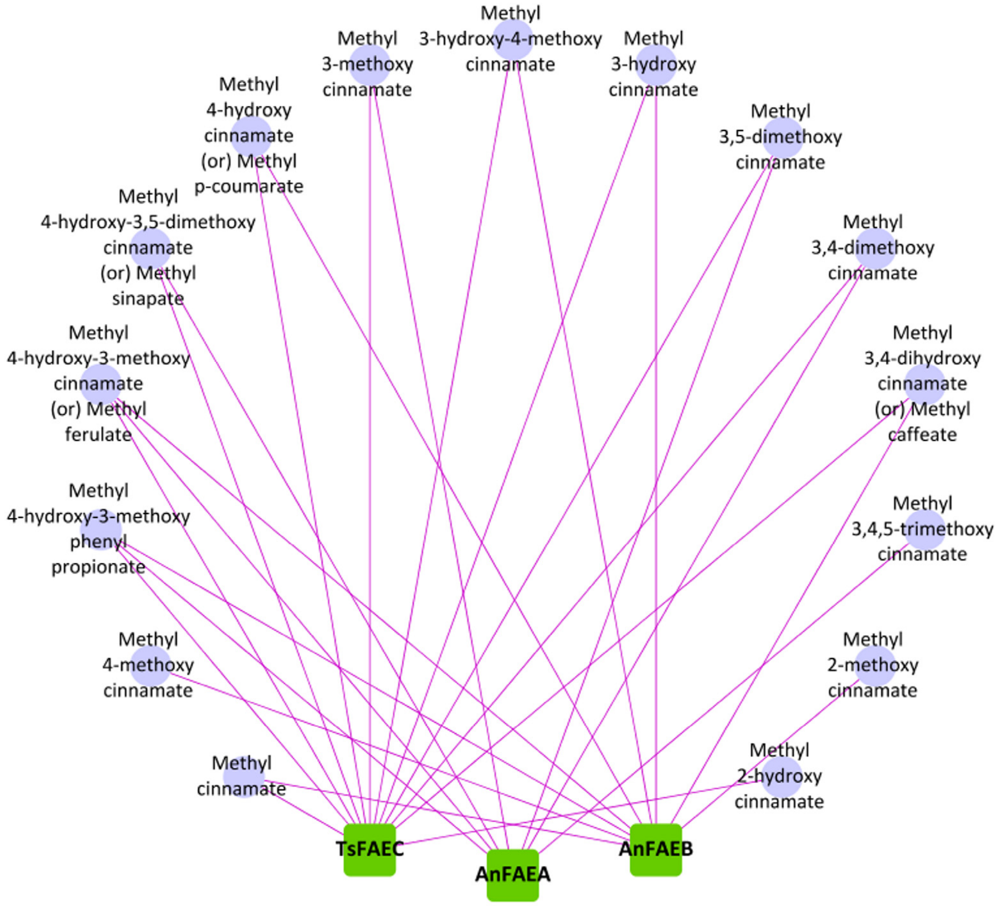

Methyl 3-hydroxy-4-methoxy cinnamate
Methyl 3-methoxy cinnamate
Methyl 3-hydroxy cinnamate
Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate
Methyl 3,5-dimethoxy cinnamate
Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate
Methyl 3,4-dimethoxy cinnamate
Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate
Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate
Methyl 4-hydroxy-3-methoxy phenyl propionate
Methyl 3,4,5-trimethoxy cinnamate
Methyl 4-methoxy cinnamate
Methyl 2-methoxy cinnamate
Methyl cinnamate
Methyl 2-hydroxy cinnamate
TsFAEC
AnFAEA
AnFAEB

Figure 5. Overlapping substrate specificities among the three different FAEs viz.,Feruloyl esterase type-C from *Talaromyces stipitatus* (TsFAEC), Feruloyl esterase type-A from *Aspergillus niger* (AnFAEA) and Feruloyl esterase type-B from *A. niger* (AnFAEB) for the 15 methyl cinnamate esters; as presented in Figure 1 of **PAPER II**. The substrate specificity information of the three enzymes was extracted from the experimental binding affinity data that has been published previously (Topakas et al., 2005; Vafiadi et al., 2006).

The varied and overlapping substrate specificity profile of FAEs for the small compounds like substituted methyl cinnamates is a major challenge to understand the small structural differences of the FAE binding pockets and to propose a classification system that reflects their function. Enzyme reactions are enabled by the structural elements in the enzymes that catalyze them, so sub-grouping them based on the bound ligands seems a more functionally defined approach. Since the techniques for comparing the ligand similarities and algorithms for computing protein sequence descriptors are already mature, I envisage that these methods will help in generating sequence-structure-function links and further sub-classification of enzyme families based on specificity as described in the following chapters.

CHAPTER 2

Approaches for Classification of Enzymes beyond the EC system

With the rapid increase of sequenced genomes in the post-genomic era, functional annotation of proteins has become both a necessity and a challenge. The first step in the functional annotation of proteins is to detect the homologous relationship between proteins through pairwise sequence similarity using sequence alignment algorithms. The second step is to infer the functional similarity from the homology (Altschul & Gish, 1996; Altschul et al., 1997; Larkin et al., 2007; Pearson, 2000). Classification approaches designed based on sequence similarity rely on the assumption that similarities between the sequences of two proteins imply similarities between the structures and thus also the function of these proteins. Each protein sequence is assumed to fold into a unique three dimensional structure. However, when all proteins are considered, this one-to-one correspondence is no longer valid (Koehl & Levitt, 2002). The size of the protein structure space is much smaller than the size of the protein sequence space: it is commonly assumed that there are 1000 different protein folds, covering 10,000 different protein sequence families (Govindarajan et al., 1999; Orengo et al., 1994). Surprisingly, the average sequence identity between pairs of proteins with similar structures has been found to be in range of 8–10% (Rost, 1997) and thus most of the evolutionary related proteins or homologous proteins must have different function, which makes the functional annotation based on sequence similarity a challenging task (Brenner, 1999; Brenner et al., 1997; Chothia, 1992; Devos & Valencia, 2001).

The substrate specificity of a reaction for an enzyme is represented by the last digit of the EC number, while the first three digits describe the type of the reaction. It has been shown that all the four digits of the EC number start to diverge quickly when the sequence identity is below 70% (Rost, 2002), which raised questions on the functional schemes based on sequence similarities. Hence, annotation errors could be easily spread among the enzyme classification system if the functional annotation is not done carefully. This creates an urgent requirement to choose alternative methods to sub-group enzymes that reflects their function or substrate specificity.

Researchers have tried to use enzyme structural information in the process of predicting enzyme specificity, where they infer a specificity-conferring code of the active site (Challis et al., 2000; Stachelhaus et al., 1999). Identification of the binding pocket residues and mapping of the binding pocket from the protein primary structure are not straight forward procedures and are non-practical for hundreds of sequences generated through genome sequencing projects. Later on, a significant improved prediction performance has been achieved for substrate specificity predictions using machine learning approaches (Rausch et al., 2005). For a dataset of functionally known protein sequences belonging to different enzyme groups, group-specific features can be extracted to build models using machine learning algorithms to predict the function of an unknown protein sequence or to assign a group label to it (Juncker et al., 2009; Ong et al., 2007).

Three prominent approaches have been widely experimented for classification of enzymes based on the protein feature space. The first approach involves assigning a class to an enzyme based on sequence similarity between enzymes belonging to the same functional class (Shah & Hunter, 1997) and the second approach is based on protein structure comparison (Wang et al., 2003b). The inefficiency of the first two approaches gave birth to a third approach which involves representing enzymes using sequence and structure driven features that do not use sequence similarity as a classifier (Han et al., 2004; Syed & Yona, 2009). As shown in **Table 3**, the classification accuracy is superior for schemes based on protein sequence driven features when compared to approaches based on sequence or structural similarity.

Table 3. List of enzyme classification attempts based on sequence similarity, structural similarity and protein descriptors.

Method	Feature Used	Classification accuracy/Result	References
BLAST, FASTA	Sequence information	40% of enzyme classes predicted correctly	(Shah & Hunter, 1997)
BLAST	Sequence information	Found putative analogy of 40.5% for all EC classes	(Audit et al., 2007)
Bayesian	Structural information	45% of enzyme classes predicted correctly	(Borro et al., 2006)
Support Vector Machine	Structural properties	60% accuracy in functional annotation of enzymes	(Dobson & Doig, 2005)
Structure template matching	Structural information	87% accuracy in functional annotation of enzymes	(Kristensen et al., 2008)
Nearest neighbor algorithm	Sequence Descriptor: Amino acid composition	95% accuracy to the level of enzyme class	(Nasibov & Kandemir-Cavas, 2009)
Nearest neighbor algorithm	Domain composition and pseudo amino acid composition	98% accuracy to the level of enzyme class	(Cai et al., 2005)
Self-organizing maps	Reaction descriptors	Accuracies up to 92, 80 and 70% for class, subclass and sub-subclass levels, respectively	
Support Vector Machine	Amino Acid Composition and Conjoint triad feature	81% to 98% accuracy in predicting the first three EC digits	(Wang et al., 2011)

2.1 Sequence based classification of enzymes

Sequence homology between a group of proteins or against a protein family database are done by sequence homology tools like BLAST (Altschul & Gish, 1996), PSI-BLAST (Altschul et al., 1997), FASTA (Pearson, 2000; Pearson, 1990) and HMMER (Finn et al., 2011; McClure et al., 1996). There are a few studies suggesting that sequence homology tools are able to determine the EC number for the query sequence, but the coverage is achieved only till the second digit of the EC number (Audit et al., 2007; Shah & Hunter, 1997).

Shah and Hunter (1997) showed that ~60% of enzyme classes of the EC system could not be discriminated by sequence similarity at any threshold, and their work strongly suggests that functional assignment of enzymes should attempt to delimit functionally significant sub-regions, or domains, before matching to EC classes. Furthermore, Audit and his co-workers (2007) found that most classification errors occur between closely related EC classes. An attempt to automate the identification of analogous[2] and homologous enzymes based on sequence similarity has been made by Otto et al., (2008), which resulted in the identification of 986 EC classes with a putative analogy of 40.5% for all EC classes. In addition, enzymes without detectable sequence similarity to each other have been found for 105 EC numbers (Galperin et al., 1998a). Even though the percentage of sequence identity is helpful in detecting remote homology of proteins, there is no clear indication on the functional relationship among them. Furthermore, due to the lack of established sequence identity score thresholds (such as *E*-Value of BLAST and FASTA), classification of an enzyme class into sub-classes require human intervention (Hannenhalli & Russell, 2000).

2.2 Structure based classification of enzymes

Classification schemes based on structural similarity are assumed to be more tolerant to errors due to the belief that structural information is more conserved in evolution (Almonacid & Babbitt, 2011). Surprisingly, classification of enzymes based on protein structure alignments achieved an accuracy of 45% (Borro et al., 2006), not far to the accuracy obtained through sequence based classification schemes. Interestingly, the use of structural properties (secondary structure content, amino acid propensities, surface properties and ligands) instead of structural alignments pushed the level of accuracy for structure based enzyme classification to 60%, which shed light on the simple structural attributes in protein function prediction (Dobson & Doig, 2005).

To deal with functionally analogous enzymes, a new idea of incorporating evolutionarily important amino acids in the structure based enzyme classification scheme has been employed by Kristensen et al., (2008). With the addition of the information on evolutionarily important amino acids in the template 3D structure, an accuracy of 87% has been achieved in the prediction of enzyme function.

[2] Functionally analogous enzymes are those that catalyze similar reactions on similar substrates but do not share common ancestry

Even though the structure based enzyme classification schemes are better than sequence similarity schemes, the coverage of structure-based schemes is low due to the fact that the 3D structure data for all enzyme classes are not available in Protein Data Bank (PDB). This drives the necessity for alternative approaches that can utilize the vast amount of available primary sequence data.

2.3 Descriptor based classification of enzymes

Classification of enzymes based on the third approach captures the biochemical characteristics of an enzyme from its amino acid sequence and the enzyme properties or descriptors are represented in the form of vectors (Ong et al., 2007). Sequence-derived descriptor features can effectively represent and distinguish proteins with different functional and interaction profiles, irrespective of sequence similarity (Han et al., 2004). Every enzyme sequence can be represented by its respective descriptor vectors from encoded representations of twenty amino acid residues (Cai et al., 2004). Several types of descriptor sets can be extracted from protein sequences which serve to represent and distinguish proteins of different structural and functional profiles by exploring features in amino acid composition, physicochemical properties, correlations, di-peptide distributions etc. An attempt to measure the efficiency of just one sequence descriptor, the amino acid composition, on enzyme classification showed an accuracy of 95%, but was limited to the level of the Enzyme Class (Nasibov & Kandemir-Cavas, 2009). Accommodating additional sequence descriptors, such as pseudo-amino acid composition in the classification scheme, further increased the accuracy of annotation among the 6 enzyme family classes to 98% (Cai et al., 2005). A different approach in the descriptor based enzyme classification has been implemented by the use of enzyme reaction descriptors that resulted in the accuracies up to 92, 80 and 70% for enzyme class, subclass and sub-subclass levels, respectively (Latino et al., 2008). The combination of amino acid composition and amino acid neighbour relationship descriptors proved promising, with an accuracy of 81% to 98% in predicting the first three EC digits of the Enzyme Commission's classification scheme (Wang et al., 2011).

Apart from enzyme classification, protein sequence based descriptors have been successfully exploited in the machine learning prediction of protein functional classes, protein-protein interactions, subcellular locations of proteins and secondary structure predictions. Furthermore, these descriptors sets and their combinations have shown varied degree of accuracy in the functional sub-grouping of protein families (Ong et al., 2007). Machine learning approaches help us to gain knowledge from complex patterns in data. One of the latest applications of machine learning is the successful use of physicochemical properties and sequence derived descriptors for the classification of proteins, for example, G-protein coupled receptors (Karchin et al., 2002) and nuclear receptors (Bhasin & Raghava, 2004).

Machine learning approaches for protein classification involve clustering of instances (in this case, the instances are individual proteins) followed by classification of instances. The goal of clustering is to group data based on common traits, whereas classification deals with the assignment of an unknown instance to a specific class among a predefined number of classes. Clustering is an unsupervised technique that reveals how instances are naturally grouped in the descriptor space. In the clustering process, the classes are unknown and are identified by the cluster analysis of the data.

In simple terms, the overall idea of clustering is to group similar elements together. A problem with most of the clustering methods is that the input data are forced into clusters even though in reality they do not share any similarities. The solution to this problem is to carefully inspect the variance of instances within the cluster and the variance between the clusters. Clusters with low variance within the group and high variance between the groups can be considered ideal. A simple illustration of the 'Variance' concept is shown in **Figure 6**.

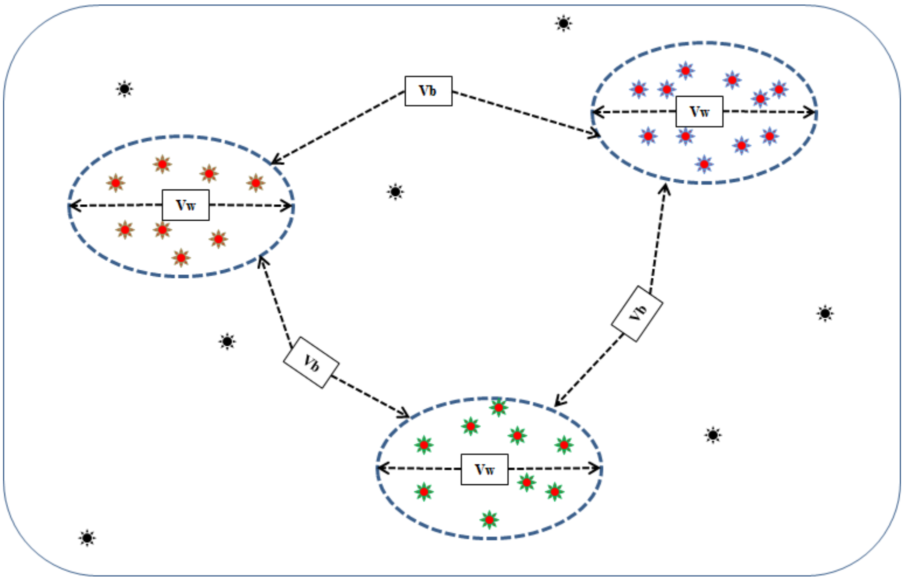


Figure 6. A schematic drawing of ideal clusters with quality variance. Instances (red spots) with similar properties (respectively coloured surrounding the instance spots) grouped together will have low variance between them. Variance within the cluster (Vw); Variance between clusters (Vb).

The initial part of my PhD research work has been focused on functional classification of feruloyl esterases aiming to represent each enzyme group reflecting their substrate specificity. As presented in PAPER I, the classification system of the putative and known FAEs involved unsupervised clustering of sequences based on a large number of amino acid sequence properties or descriptors. Later, support vector machine learning algorithm was trained to predict the class of new FAEs.

With the intention to select the best descriptor set that clusters FAEs with low variance within clusters and high variance between clusters, I evaluated the effectiveness of different descriptor sets listed in Table 4, as well as combinations of the ones showing the highest performance. The evaluation of the formed clusters was based on inspection of the within and between clusters variance. As mentioned above, clusters with low within variance and high variance between them, is what characterizes a good clustering output.

Table 4. A summary of analysis on efficacy of different sequence derived descriptors. Based on the variance scores within and between clusters, descriptor set combination DS14 was chosen as the best set; as presented in **PAPER I**.

Set code	Descriptor sets*	Descriptor Components	Number of clusters	Variance score	
				Within Cluster	Between Clusters
DS1	Amino acid composition	20	12	0.001	6.62
DS2	Dipeptide composition	400	12	0.001	0.07
DS3	Normalized Moreau-Broto autocorrelation descriptors	240	17	8.92	0.005
DS4	Moran autocorrelation descriptors	240	23	1.86	0.001
DS5	Geary autocorrelation descriptors	240	13	2.91	0.001
DS6	Composition, transition, distribution	147	13	2755	1024
DS7	Sequence order coupling numbers (Schneider-Wrede physicochemical distance matrix)	30	13	392027	14.4
DS8	Sequence order coupling numbers (Grantham chemical distance matrix)	30	13	16499	8.52
DS9	Quasi sequence order descriptors (Schneider-Wrede physicochemical distance matrix)	50	12	0.001	0.001
DS10	Quasi sequence order descriptors (Grantham chemical distance matrix)	50	13	0.001	0.001
DS11	Pseudo amino acid composition	50	10	0.001	0.001
DS12	Physicochemical composition	11	12	16.14	147.57
DS13	Amino acid composition and dipeptide composition	420	12	0.001	1.42
DS14	Amino acid composition and physicochemical composition	31	13	14.15	157.44
DS15	Dipeptide composition and physicochemical composition	413	11	15.44	20.54
DS16	Amino acid composition, dipeptide composition and physicochemical composition	433	12	13.13	20.53

*Descriptors of amino acid composition (DS1), dipeptide composition (DS2) and physicochemical composition (DS12) showed satisfactory variance scores within and between clusters. On the other hand, the rest of the descriptor sets (DS3, DS4, DS5, DS6, DS7, DS8, DS9, DS10 and DS11) showed poor performance with low quality variance scores. It should be noted that the combined use of descriptor sets containing similar information adds redundancy without improving the performance of the model. For example, the three autocorrelation descriptor sets (Set codes: DS3, DS4, DS5) utilize the same physicochemical properties and differ only in the correlation algorithm that they are based upon. The combination of three autocorrelation descriptors generates noise in clustering and does not add information in relation to the individual descriptor sets. Even though descriptor sets DS1, DS2 and DS12 showed satisfactory variance scores within cluster and between clusters, the distribution of experimentally characterized FAEs among the clusters was better in DS14, which also showed good variance scores. Combination of well-performing descriptors sets improved further the clustering of data, as evident from DS14.

After clustering, automated classification of FAEs was performed using support vector machines. Support vector machines (SVM) are supervised learning methods that learn by example to assign labels to objects (Noble, 2006) and perform the classification by constructing an *N*-dimensional hyperplane that optimally separates the data with different labels. The 10-fold cross validation of the SVM model using different ratios of training and test sets resulted in accuracies ranging from 96% to 100% (**Table 5**), which further shows efficiency of the FAE clustering based on selected protein sequence descriptor sets.

Table 5. Performance of SVM model in the cross-validations for classification of FAEs.

Ratio (Training set: Blind Test set)	Correctly Classified Instances in blind set	Incorrectly Classified Instances in blind set	% Accuracy
09:01	37	0	100%
08:02	72	1	98.63%
07:03	108	1	99.08%
06:04	142	4	97.26%
01:01	176	6	96.70%

A bird's eye view on the percent identities of sequences within respective FAE clusters given in **Table 6** shows the reason for the failure in the attempts that have been made for functional classification of FAEs based on protein sequence similarity (Benoit et al., 2008; Crepin et al., 2004).

The functional sub-grouping of the resulted FAE clusters requires the structural analysis of protein and ligand structures and their interactions. So, the next step of the FAE classification scheme involved the sub-grouping of FAEs and prediction of their substrate specificities using common feature based pharmacophore models and molecular docking methods which are discussed in **CHAPTER 3**.

Table 6. Protein sequence identities within the FAEs clustered based on sequence derived descriptors. The use of protein sequence derived descriptors to group functionally similar FAEs overcomes the challenge of low sequence identity among them.

FAE Clusters	Mean percent identity within the cluster
Cluster 1	26.30%
Cluster 2	One sequence in this cluster
Cluster 3	35.03%
Cluster 4	29.82%
Cluster 5	19.61%
Cluster 6	32.32%
Cluster 7	26.66%
Cluster 8	20.10%
Cluster 9	25.20%
Cluster 10	14.49%
Cluster 11	14.19%
Cluster 12	29.37%
Cluster 13	28.92%

CHAPTER 3
Substrate Specificity Predictions

Experimental screening to identify the substrate specificity profile of an enzyme often suffers from limitation with respect to the possible number of compounds that can be used in high-throughput assays, which are time consuming and costly. The understanding of key interactions between an enzyme and a substrate can ease the task of substrate selection. Within this context, *in silico* approaches like pharmacophore modeling[3] (Stoll et al., 2002) and molecular docking are proven to be successful to understand the biological target structure and supramolecular interactions (Guner et al., 2004; Kurogi & Guner, 2001; Langer & Krovat, 2003; Stoll et al., 2002). The way to determine a pharmacophore can be based on either the protein (protein structure-based pharmacophores) or on the compounds interacting with the binding pocket of the protein (ligand based pharmacophores).

Keeping in view the macromolecular structure of a protein and the number of rotatable/flexible bonds in its binding pocket, a major challenge in the design of protein structure-based pharmacophores is the reduction of the high number of features to those features that are related to the biological activity. Ligand based pharmacophores can be modeled when the activity data of protein on a certain number of ligands are available and the key elements involved in their modeling might be a group of atoms or pharmacophoric features like H-bond acceptors, H-bond donors, hydrophobic groups, ionizable groups, aromatic rings and can also involve geometrical constraints (Wolber et al., 2008).

Pharmacophore models can be used as a tool to identify novel compounds or substrates that have high probability of interacting with the protein target and thus are

[3] A pharmacophore model can be defined as the ensemble of steric and electrostatic features of different compounds which are necessary to ensure optimal supramolecular interactions with a specific biological target structure.

biologically active based on the fulfilment of the pharmacophore feature requirements. In other words, pharmacophores can be considered as *in silico* filters in the search of novel substrates or ligands. Even though the relative performance of ligand versus structure based pharmacophore models in virtual screening can be target dependent, recent studies have revealed that ligand-based methods outperform protein structure based methods (Evers et al., 2005; Guner et al., 2004; Guner, 2011; Kitchen et al., 2004). As shown in **Figure 7**, building a pharmacophore involves the analysis of the training set compounds itself to identify the pharmacophore features and further alignment of known active compounds to determine the best overlay of corresponding features.

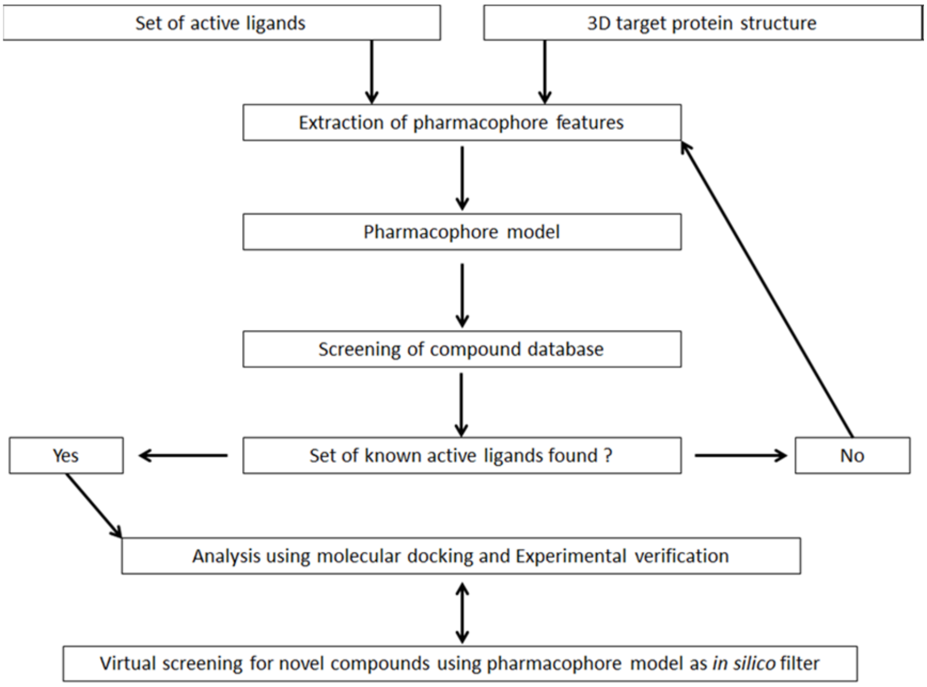

Set of active ligands
3D target protein structure
Extraction of pharmacophore features
Pharmacophore model
Screening of compound database
Yes
Set of known active ligands found ?
No
Analysis using molecular docking and Experimental verification
Virtual screening for novel compounds using pharmacophore model as *in silico* filter

Figure 7. Workflow of virtual screening to identify novel compounds using pharmacophore model. The pharmacophore model built on the basis of the chemical signatures of known active ligands can be used as 3D filter in the screening of compound libraries to identify the compounds that match the chemical features of the pharmacophore.

Several players on the market like Accelrys Inc (USA), BioSolveIT GmbH (Germany), Chemical Computing Group (Canada), Tripos (USA), Molecular Networks GmbH (Germany), etc., offer software solutions for pharmacophore modeling and pharmacophore based database search algorithms. The software solution, Catalyst®, available from Acceryls Inc, is by far the most used by researchers due to the flexibility it offers during pharmacophore modeling with integrated database search options. Comparison of pharmacophore solutions has shown that Catalyst and HipHop algorithm from Accelrys Inc outperform the other software packages (Sutter et al., 2011). Catalyst checks the surface accessibility of molecules available for receptor interaction and further defines the position of pharmacophore features rather than by inter-feature distances in the training set compounds. The HipHop algorithm evaluates members of a training set based on the type of chemical features they contain, along with the ability to adopt a conformation that allows those features to be superimposed on a particular configuration (Barnum et al., 1996).

Fuelled by the availability of the algorithms for extracting the enzyme-substrate interactions and mapping of substrate features, common feature ligand-based pharmacophore models were modeled for the five feruloyl esterase sub-groups (of the descriptor based classification system) based on the experimental substrate specificity data using the software solutions offered by Accelrys Inc (**PAPER I** & **III**). Both active and inactive substrates were given as input for pharmacophore model development with a constraint that the active substrates of respective enzyme must map completely or partially to the pharmacophore; while the features from inactive substrates (on which the respective enzyme has no observed activity) must be considered as "NOT" features. This option resulted into broader and more diverse pharmacophores as shown in **Figure 8**. The pharmacophore models can be further used in a virtual screening workflow to identify novel substrates for the use FAEs in biocatalytic applications.

In the case of FAE sub-groups with no available experimental enzyme activity data, molecular docking approach was chosen to predict the substrate specificity. Molecular docking programs are used to position potential substrates within a three-dimensional structure of the enzyme. Careful choice of model ligand and protein structures, as well as the selection of appropriate docking program, is important for reliable substrate specificity predictions.

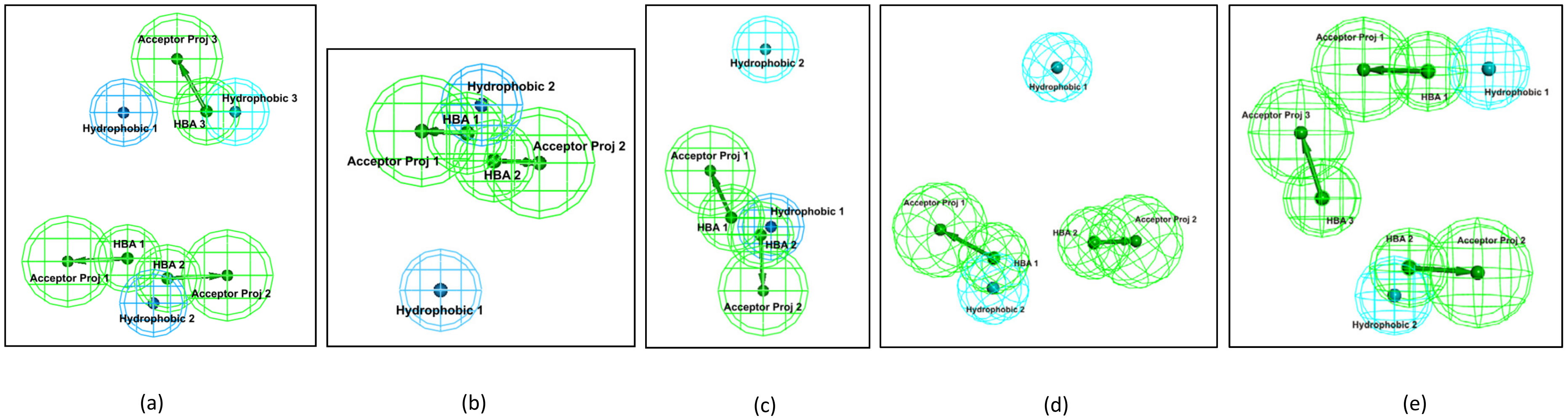


Figure 8. Common feature-based pharmacophore models for the substrates of five feruloyl esterases viz., AnFAEA, AnFAEB, TsFAEC, AoAO8, AoAO10, represent members of the sub-families FEF 12A, FEF 4A, FEF 4B, FEF 7A and FEF 12B respectively, and can, thus, be used for the prediction of their substrate binding profiles. (a) The best performing HipHop pharmacophore model for the substrates of AnFAEA – *Aspergillus niger* feruloyl esterase A. (b) The best performing HipHop pharmacophore model for the substrates of AnFAEB - *Aspergillus niger* feruloyl esterase B. (c) The best performing HipHop pharmacophore model for the substrates of TsFAEC - *Talaromyces stipitatus* feruloyl esterase C. (d) The best performing HipHop pharmacophore model for the substrates of AoAO8 - *Aspergillus oryzae* feruloyl esterase 8. (e) The best performing HipHop pharmacophore model for the substrates of AoAO10 - *Aspergillus oryzae* feruloyl esterase 8. Blue: hydrophobic features; Green: H-bond acceptor features and their projections from the molecule to hydrogen-bond donors and acceptors or charged groups in the binding site; Figure adapted from **PAPER I** and **PAPER III**.

3.1 Challenges in selecting molecular docking programs in prediction of substrate specificities

Even the experienced researchers in drug discovery and molecular docking agree upon the difficulties in choosing the best docking program. In the constantly shifting landscape of new molecular docking programs, several publications comparing their performance have been published, yet there is a question on how many of those evaluation studies can be considered free of bias or run using the 'black box' protocols provided by the software companies. (Bissantz et al., 2000; Bursulaya et al., 2003; Chen et al., 2006; Cross et al., 2009a; Cummings et al., 2005; Ferrara et al., 2004; Kellenberger et al., 2008; Kellenberger et al., 2004; Kontoyianni et al., 2004; Kontoyianni et al., 2005; McGaughey et al., 2007; Onodera et al., 2007; Perola et al., 2004; Schulz-Gasch & Stahl, 2003; Stahl & Rarey, 2001; Wang et al., 2003a; Warren et al., 2006; Xing et al., 2004; Yang et al., 2005). The intention it is not to criticize the authors of these studies, but to make clear the unintentional flaws in selecting the molecular docking programs as there are no universally accepted set of standards in designing the evaluation studies.

Recent findings published by Cross et al., (2009) have put an end to the trend of evaluating molecular docking programs using mix of protein structures from all the families in standard datasets like Directory of Useful Decoys (Huang et al., 2006; Irwin, 2008; von Korff et al., 2009). Their studies have indicated that the differences in performance of the molecular docking programs could be attributed to the composition of the training sets used while developing particular docking programs with different intended goals. The molecular docking community has now realized that the evaluation of docking programs should be done against your protein or protein family of interest, not using a mix of structures that belong to different protein families (Hevener et al., 2009; Udatha et al., 2012b; Zeragraf et al., 2007).

A molecular docking evaluation study should carefully consider the following points:

i) Representative target protein structures

ii) Known active, inactive ligand molecules and their structures

iii) Binding pocket information of the target protein

iv) Flexible protocols to optimize the options in the algorithms

v) Performance measures to evaluate the programs

The quality of the protein structure needs to be assessed even for X-ray crystal structures, as the areas that might not be well-resolved may result in either multiple models, or data being absent altogether. The severity of missing data ranges from occasional missing atoms to entire sections of the structure being absent. In many cases the missing data need to be modeled and fixed before subsequent computational analyses can proceed. Few software packages like Accelrys Discovery Studio (Accelrys Inc, USA) and Schrödinger suite (Schrödinger LLC, USA) offer protocols for pre-processing of protein structures. The pre-processing resolves the missing hydrogen atoms, incorrect bond order assignments, charge states or orientations of various groups and generate the protein structures to a state in which they are properly prepared for molecular docking. The differences in the ligand-receptor interactions, as shown in **Figure 9**, potentially affect the molecular docking calculations and therefore pre-processing of structures should be considered as a critical step before starting the docking process.

Misrepresentation of docking studies using a single conformation of the ligand structures has to be avoided by incorporating the step to generate multiple forms of a ligand like tautomers, ring conformers, ligands with different protonation states etc (Bursulaya et al., 2003; Cross et al., 2009a; Hevener et al., 2009; Wang et al., 2003a). As explained in **PAPER II**, it should be noted that different scoring functions in the docking programs may perform better on a certain protein target than on another, even if both belong to the same protein family. Furthermore, the differences in performance of the molecular docking programs could be attributed to the composition of the training sets used while developing particular docking programs that have different intended goals (Cross et al., 2009b).

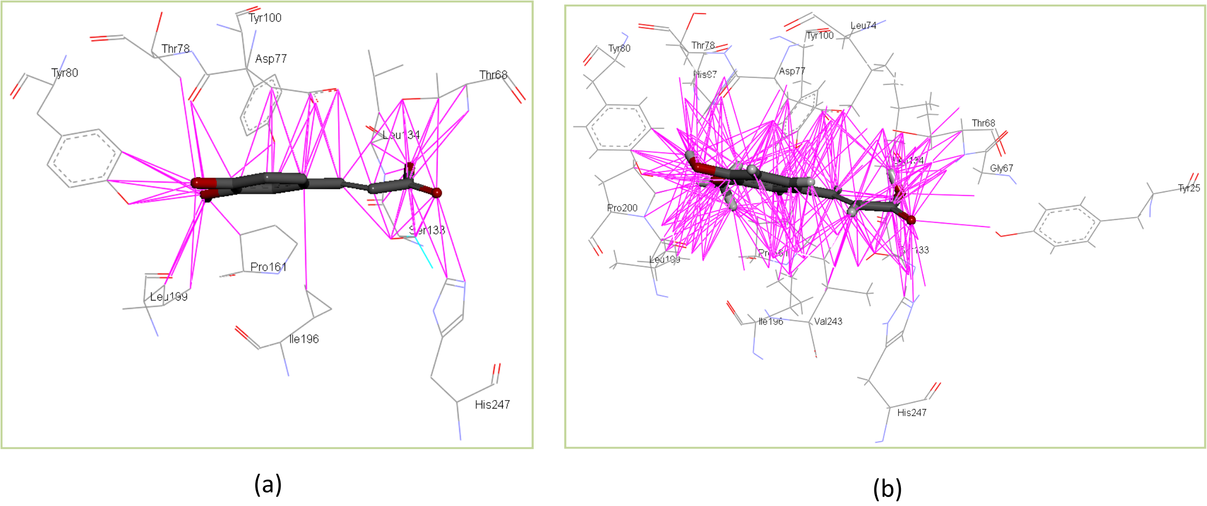


Figure 9. Comparison of ligand-receptor interactions in (a) unprocessed and (b) processed feruloyl esterase crystal structure (PDB ID: 1UWC), as shown in supplementary information of **PAPER II**. The polar and non-polar contacts between the ligand (stick model) and the amino acid residues of the protein binding pocket were depicted as magenta lines.

It is often ignored or forgotten by docking program evaluators that the docking process consists of two steps: i) an algorithm that is used to place representations of ligands in the protein structure, which is referred to as '*docking*', ii) estimation of binding enthalpies of the docked ligands by evaluating their complementarity to the target that finally leads to the prediction of the binding free energy or affinity, which is referred to as '*scoring*'. The two major technical challenges for a docking program consist of the correct prediction of the ligand binding mode (called as 'pose prediction accuracy') and the reliable rank-ordering of ligands that reflects the experimental binding affinity. It is unlikely to calculate a meaningful score for rank-ordering of ligands by the scoring function, if they are not properly docked into the target protein by the docking algorithm. Thus, the accuracy in the first step is prerequisite for the reliably processing the second step by the docking program.

Most of the commercial software packages simply termed as 'docking programs' contains both the docking algorithms and scoring functions to carry out the two steps mentioned above in the docking process. The major difference in docking algorithms is the degree to which the respective algorithm implements the flexibility of ligand and receptor. A large variety of scoring schemes also exists to rank-order ligand poses. Ligand scoring is a method to rapidly estimate the binding affinity of a ligand, based on a candidate ligand pose geometry docked into a target receptor structure. Scoring methods typically use empirical functions developed by fitting various functional forms (described in the supplementary information of PAPER II), which characterize various aspects of the receptor-ligand interactions against binding affinity data. So, different combinations of docking algorithms and scoring functions should be evaluated: a procedure that we generally do not observe in the published evaluation studies. The software evaluation studies described in PAPER II address the problem of selecting an appropriate docking and scoring function combination among 88 docking algorithm-scoring function sets.

The evaluation of the docking programs should be based on reliable performance measures. Measures like root mean square deviation (RMSD), enrichment factor, area under the received operating characteristic curve (ROC), exist for determining the pose prediction accuracy and identification of active ligands by the docking programs, while each measure has certain disadvantages. For example, it is assumed that the higher the RMSD of the docked pose, the most likely it is classified as inactive ligand or incorrect pose. As shown in PAPER II, having a low RMSD between the docked and the crystallographic pose does not necessarily mean that the ligand can actually form similar interactions or similar binding modes and that a high RMSD value does not indicate the opposite situation. Further, the enrichment factors are highly sensitive to the ratio of active and inactive ligands sets, which makes it difficult to compare the evaluation studies using different ligand sets. The ROC approach does not say anything about whether the docked poses make any interactions that are biologically meaningful. For evaluating molecular docking programs, the combination of RMSD and Key Interaction Score System (KISS) was proposed as described in PAPER II, which paves the way towards providing a biological meaning to the docking program evaluation studies.

3.2 Validation of substrate specificity predictions

The pharmacophore models for FAE sub-families and the combined RMSD & KISS for molecular docking predictions were validated using experimental substrate activity data. First, the reliability of the generated pharmacophore models (shown in **Figure 8**) was validated for the presence of the chemical features necessary to interact with the amino acid residues in the binding pocket of the respective enzyme that represents the corresponding FAE sub-family. The pharmacophore models generated for the FAE sub-families were ranked based on how well the known active substrates mapped on the proposed pharmacophores, as well as on the rarity or infrequency of the pharmacophore model. A pharmacophore model that is less likely to map to an inactive substrate will be given a higher rank. As a validation, each pharmacophore model was mapped against 25 compounds, which comprised of 15 training substrates on which the pharmacophore models were built and additional 10 substrates that were not involved in the pharmacophore modeling. For example, the heat map shown in **Figure 10** indicates how well the active substrates map to the respective pharmacophore models generated for AnFAEA that belongs to FAE sub-family 12A of the descriptor based classification system (**PAPER I**).

In **PAPER I**, it was assumed that the pharmacophore model developed for the substrate specificity of one enzyme of a particular FAE sub-family may represent the substrate specificity of all the members in that sub-family. To validate this assumption, a predicted feruloyl esterase (A.O.2) from *A. oryzae* that belongs to FAE sub-family 4A of the descriptor based classification system was recombinantly expressed in *Pichia pastoris* strain SMD1168H and tested for its experimental substrate specificity (**PAPER III**). The experimental substrate specificity profile for the A.O.2 showed ~95% match with the substrate specificity profile of AnFAEB, based on which the pharmacophore model for FAE sub-family 4A was developed (**Table 7**). This also shows how the right combination of protein sequence descriptors and molecular signatures can be successfully used in the functional classification of enzymes.

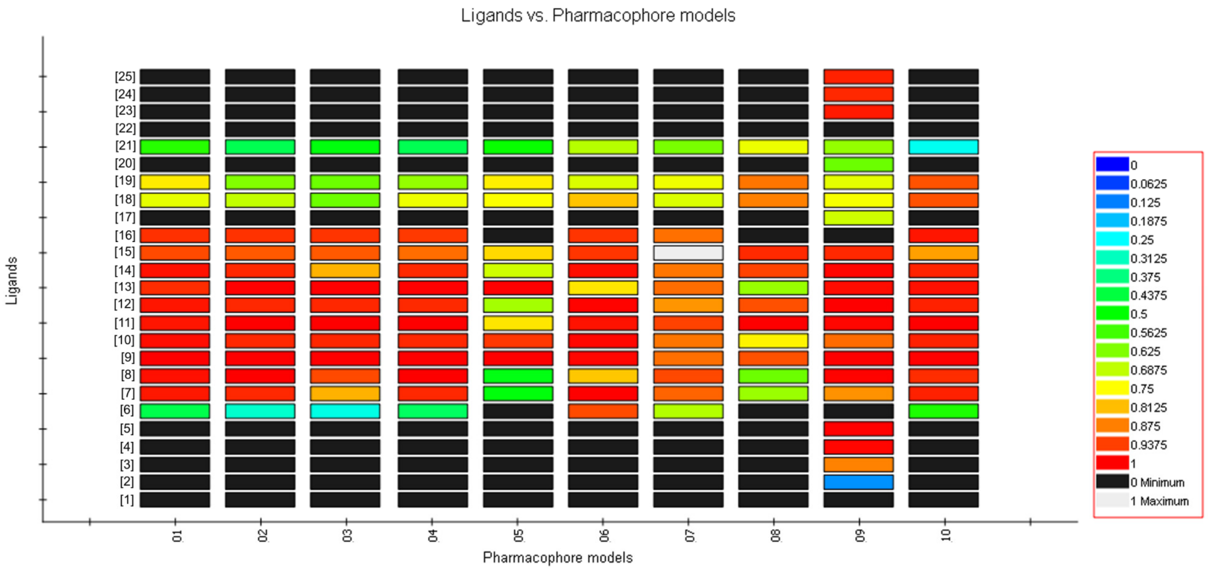


Figure 10. Ligand pharmacophore mapping for AnFAEA belonging FAE sub-family 12A; as shown in **PAPER I**. The heat map values show how well compounds map to pharmacophore models; higher values indicate better mapping of compounds to pharmacophore model. The colour legend corresponds to the alignment score and is in the range between 0 and 1.0 with high values above 0.9 (red) indicating a good match. Substrates on which AnFAEA can act are highlighted in bold. The pharmacophore model 06 maps well against all of the known AnFAEA substrates and therefore selected as the best pharmacophore model that describes the enzyme's substrate selectivity profile. The pharmacophoric features of this model are mapped to the features of the AnFAEA active substrates with an average alignment score of 0.98.

Substrates shown on Y-axis are [1] Methyl cinnamate; [2] Methyl 2-hydroxy cinnamate; [3] Methyl 3-hydroxy cinnamate; [4] Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate; [5] Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate; [6] Methyl 2-methoxy cinnamate; **[7] Methyl 3-methoxy cinnamate;** [8] Methyl 4-methoxy cinnamate; **[9] Methyl 3,4-dimethoxy cinnamate; [10] Methyl 3,5-dimethoxy cinnamate; [11] Methyl 3,4,5-trimethoxy cinnamate; [12] Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate;** [13] Methyl 3-hydroxy-4-methoxy cinnamate; **[14] Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate; [15] Methyl 4-hydroxy-3-methoxy phenyl propionate;** [16] Methyl 3,4-dichloro phenyl propionate; [17] Methyl 4-hydroxy phenyl acetate; [18] Methyl 4-hydroxy-3-methoxy phenyl acetate; [19] Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate; [20] Methyl 4-hydroxy benzoate; [21] Methyl 4-hydroxy-3-methoxy benzoate; [22] Methyl 5-phenylpenta-2,4-dienoate; [23] Methyl L-tyrosine; [24] Methyl 3,4-methylene dioxy phenyl propionate; [25] Methyl 3,4-methylene dioxy cinnamate. AnFAEA can hydrolyze the substrates highlighted in bold above.

Table 7. Comparison of substrate specificity profile of AnFAEB and A.O.2 that belong to FAE sub-family 4A. The values given are K_m expressed as mM.

Substrates	FAE sub-family 4A	
	AnFAEB[a]	A.O.2[b]
MFA	1.32	1.39
MCA	0.22	2.36
MPC	0.014	1.51
MSA	ND	ND
M2C	ND	1.73
M3C	0.55	2.55
MC	0.79	3.14
MTM	ND	ND
M2M	0.72	0.73
M3M	ND	ND
M4M	0.31	0.55
M34DC	ND	ND
M35DC	ND	ND
M34MC	0.85	1.47
M43PP	3.17	8.64

[a]Values taken from Topakas *et al* (2005).
[b]Values taken from **PAPER III**.

Evaluation of molecular docking programs to predict the substrate specificity of FAEs showed that the performance of each program varies for the three FAEs considered. The three FAEs studied in PAPER II are members of different FAE families and present high diversity in their binding sites (as shown in Figure 6 of PAPER II). It has been proposed that the differences in performance of the molecular docking programs could be attributed to the composition of the training sets used while developing particular docking programs that have different intended goals (Cross et al., 2009b). Several factors like binding pocket environment, volume of the binding pocket and number of rotatable bonds that deal with the flexibility of the binding pocket play significant role on the performance of the docking algorithms/scoring functions.

Can we use molecular docking programs to predict the substrate specificity of FAEs? The answer is 'yes', but only if the information of key enzyme-substrate interactions are included in docking studies as described in PAPER II. The reverse validation for the combination of the key interaction system with docking score was

Table 8. Comparison of docking score and the combination of key interaction information with docking score in the rank-ordering of active/inactive substrates. (a) Rank-ordering of the substrates based on docking score from Glide SP program. (b) Rank-ordering of the substrates based on the combination of Key Interaction System and docking score.

(a)

Rank	Compound	Docking score (kcal/mol)	K_m (mM)
1	**Methyl 4-hydroxy-3,5-dimethoxy cinnamate (Methyl sinapate)**	**-6.15**	**0.45**
2	**Methyl 4-hydroxy-3-methoxy cinnamate (Methyl ferulate)**	**-6.11**	**0.72**
3	**Methyl 3,4-dimethoxy cinnamate**	**-6.05**	**1.36**
4	Methyl 3-hydroxy cinnamate	-5.94	Inactive
5	**Methyl 3,5-dimethoxy cinnamate**	**-5.89**	**0.92**
6	**Methyl 3,4,5-trimethoxy cinnamate**	**-5.79**	**1.63**
7	**Methyl 3-methoxy cinnamate**	**-5.74**	**1.99**
8	Methyl 3-hydroxy-4-methoxy cinnamate	-5.63	Inactive
9	Methyl 4-methoxy cinnamate	-5.50	Inactive
10	Methyl 2-hydroxy cinnamate	-5.39	Inactive
11	Methyl 4-hydroxy cinnamate (Methyl p-coumarate)	-5.33	Inactive
12	Methyl 3,4-dihydroxy cinnamate (Methyl caffeate)	-5.33	Inactive
13	**Methyl 4-hydroxy-3-methoxy phenyl propionate**	**-5.29**	**2.08**
14	Methyl 2-methoxy cinnamate	-4.73	Inactive
15	Methyl cinnamate	-4.68	Inactive

(b)

Rank		[a] HBI with Thr 68	[a] HBI with Leu 134	Glide SP score (kcal/mol)	K_m (mM)
1	**Methyl 4-hydroxy-3,5-dimethoxy cinnamate (Methyl sinapate)**	**Yes**	**Yes**	**-6.15**	**0.45**
2	**Methyl 4-hydroxy-3-methoxy cinnamate (Methyl ferulate)**	**Yes**	**Yes**	**-6.11**	**0.72**
3	**Methyl 3,4-dimethoxy cinnamate**	**Yes**	**Yes**	**-6.05**	**1.36**
4	**Methyl 3,5-dimethoxy cinnamate**	**Yes**	**Yes**	**-5.89**	**0.92**
5	**Methyl 3,4,5-trimethoxy cinnamate**	**Yes**	**Yes**	**-5.79**	**1.63**
6	**Methyl 3-methoxy cinnamate**	**Yes**	**Yes**	**-5.74**	**1.99**
7	**Methyl 4-hydroxy-3-methoxy phenyl propionate**	**Yes**	**Yes**	**-5.29**	**2.08**
8	Methyl 3-hydroxy-4-methoxy cinnamate	No	No	-5.63	Inactive
9	Methyl 2-hydroxy cinnamate	No	No	-5.39	Inactive
10	Methyl 3,4-dihydroxy cinnamate (Methyl caffeate)	No	No	-5.33	Inactive
11	Methyl 4-methoxy cinnamate	No	No	-5.50	Inactive
12	Methyl 3-hydroxy cinnamate	No	No	-5.94	Inactive
13	Methyl 2-methoxy cinnamate	No	Yes	-4.73	Inactive
14	Methyl 4-hydroxy cinnamate (Methyl p-coumarate)	No	No	-5.33	Inactive
15	Methyl cinnamate	No	No	-4.68	Inactive

[a] HBI = Hydrogen Bond Interaction

performed using experimental substrate specificity data. For example, when the top scoring docked poses obtained for AnFAEA by the Glide SP algorithm were analyzed, it was observed that all the active substrates were able to form hydrogen bond interactions with Thr 68 and Leu 134 amino acid residues of the binding pocket; whereas the inactive substrates were not able to interact in the same way. As shown in **Table 8b**, it was observed that if the enzyme-substrate interaction information is applied as a constraint for docking, the accuracy in the identification of the actives from inactives, and thus the prediction of substrate specificity, improves. A limitation to this approach is the requirement of enzyme-substrate interaction information, which is not straight forward for all proteins due to the lack of 3D structures and experimental substrate specificity data.

A reliable and universally applicable docking program is still far from reach in the near future and the work reported in **PAPER II** indicated that developing docking algorithms/scoring functions towards specific target classes may provide reliable substrate-specificity predictions using *in silico* approaches.

CHAPTER 4

Enzyme structure-function relationships

An important goal of exploring sequence-structure-function relation studies of enzymes in this thesis work is to predict without laboratory experimentation, the substrate specificity of a given sequence or structure. The strategy begins with seeking an understanding of how enzymes with identical protein fold can show different substrate specificity. In the beginning of the 20th century, studies on enzyme structure-function relationships have proposed that enzymes maintain the ability to bind a particular substrate although the structural regions involved in the catalytic process change throughout enzyme evolution (Bryson & Vogel, 1965; Horowitz, 1945). However, several recent case studies have shown that conserved structure of enzymes during the evolution of new functions is for maintaining the ability to perform a catalytic step in the chemical reaction but not to bind a specific substrate (Almonacid & Babbitt, 2011; Babbitt & Gerlt, 1997; Gerlt & Babbitt, 2009; Gerlt & Raushel, 2003; Todd et al., 2001). This chemistry-constrained enzyme evolution has resulted into several enzyme families consisting of homologous enzymes that can act on a wide variety of substrates, while maintaining a key mechanistic step of the catalytic process guided by the conserved structural features in the active site (Gerlt & Babbitt, 2001; Udatha et al., 2012a).

The feruloyl esterase family studied in this research work shows how the structural insights can lead to functional sub-grouping of enzymes (PAPER I & III). Fifteen methyl cinnamate substrates were assayed for FAEs belonging to different sub-families. Understanding the structure-function relationships requires mapping of protein structural features. This will help in identifying the less conserved structural features as well as in identifying the structural elements that are conserved for shared catalytic capabilities at the superfamily level. The preserved structural features in the path of evolution may be the basis for promiscuous activity shown by enzymes (Galperin & Koonin, 1999; Galperin et al., 1998b; James & Tawfik, 2003; Khersonsky & Tawfik, 2010; Nobeli et al., 2009; Omelchenko et al., 2010a). For example, the three-dimensional structure alignments of thirteen feruloyl esterases from *A. oryzae* that are distributed among four different FAE sub-families, revealed the structural elements that are common within FAE families. The cross-structure statistics for structural alignment of the 13 FAEs are shown in **Table 9**. The proteins within a certain FAE family show significant structural similarity especially in the secondary structural elements surrounding the binding pocket. The FAEs A.O.1, A.O.2, A.O.3 and A.O.11 which were predicted (based on the descriptor based classification system) as members of the FAE family 4, showed structural similarity with an overall RMSD of 3.3 Å over 187 structurally aligned residues. The FAEs A.O.4 and A.O.5, which were predicted as members of the FAE family 6, showed structural similarity with an overall RMSD of 2.7 Å over 134 structurally aligned residues. The FAEs A.O.6, A.O.7, A.O.8 and A.O.9, which were predicted as members of the FAE family 7, showed structural similarity with an overall RMSD of 3.1 Å over 192 structurally aligned residues. In addition, the FAEs A.O.10 and A.O.13, predicted as FEF 12 members, showed structural homology with an overall RMSD of 2.9 Å over 262 structurally aligned residues. Taking into consideration the number of amino acid residues of each protein, approximately 50% of the residues were structurally aligned with its family member proteins, but it should not be overlooked that the sequence homology is still quite low even between the members of the same FAE family. The structural alignment of those 13 FAEs showed that an average of only 37 residues was structurally aligned. So, it is evident that despite of low sequence similarity, a certain extent of structural homology is preserved within each FAE family to catalyze the ester hydrolysis, as it is well known that enzymes with the same fold catalyze the same reaction even in absence of significant sequence similarity (Omelchenko et al., 2010a). Analysis of the modeled structures of the 13 FAEs showed that, with a limited set of structural scaffold variations, FAEs evolved into different families with varied substrate specificities guided by topological variations of the binding pockets.

Table 9. Cross-structure statistics for protein structure alignments of 13 FAEs from *A. oryzae* (A.O.1, A.O.2, A.O.3… A.O.13) belonging to four different FAE sub-families; as presented in **PAPER III**.

Cross-structure statistics: RMSD[a]

FEF	Structure	A.O.1	A.O.2	A.O.3	A.O.11
FEF 4	A.O.1		1.862	3.755	3.893
	A.O.2	1.862		3.548	3.694
	A.O.3	3.755	3.548		2.707
	A.O.11	3.893	3.694	2.707	
		A.O.4	A.O.5		
FEF 6	A.O.4		3.380		
	A.O.5	3.380			
		A.O.6	A.O.7	A.O.8	A.O.9
FEF 7	A.O.6		3.077	3.034	3.369
	A.O.7	3.077		3.388	3.722
	A.O.8	3.034	3.388		1.887
	A.O.9	3.369	3.722	1.887	
		A.O.10	A.O.13		
FEF 12	A.O.10		3.583		
	A.O.13	3.583			

Cross-structure statistics: Sequence Identity[b]

FEF	Structure	A.O.1	A.O.2	A.O.3	A.O.11
FEF 4	A.O.1		0.396	0.123	0.123
	A.O.2	0.396		0.160	0.144
	A.O.3	0.123	0.160		0.299
	A.O.11	0.123	0.144	0.299	
		A.O.4	A.O.5		
FEF 6	A.O.4		0.142		
	A.O.5	0.142			
		A.O.6	A.O.7	A.O.8	A.O.9
FEF 7	A.O.6		0.229	0.193	0.188
	A.O.7	0.229		0.177	0.234
	A.O.8	0.193	0.177		0.323
	A.O.9	0.188	0.234	0.323	
		A.O.10	A.O.13		
FEF 12	A.O.10		0.191		
	A.O.13	0.191			

[a]RMSD stands for the Root Mean Square Deviation, calculated between Cα-atoms of matched residues at best 3D superposition of the query and target structures. RMSD is presented in angstroms. In simple words, RMSD gives you an idea how separated, at best 3D superposition, a "typical" pair of matched Ca-atoms is.

[b]Sequence identity is a quality characteristic of Cα-alignment. It is calculated from structure (3D), rather than sequence alignment. Therefore, two almost identical sequences may be estimated at low sequence identity if they fold into slightly different structures.

Furthermore, the binding pockets or active sites of FAEs seem to have evolved from a common ancestor with the classic constellation of the SER-HIS-ASP catalytic triad (McAuley et al., 2004). In the reactions catalyzed by FAEs, serine acts as nucleophile, histidine as the general acid-base, and the aspartic acid helps to orient the histidine residue and further neutralize the charge that forms on histidine during the catalytic process (Ekici et al., 2008b). However, the experimental data (**PAPER III**) on substrate specificity indicates that the presence of a common domain with the classic constellation of the SER-HIS-ASP catalytic triad among different FAEs does not imply that they have the same catalytic function or can act on the same substrates.

Identification of catalytic residues and their constellation in FAE protein structures plays an important role in complementing the experimental characterization of the enzyme. Identification of the active site serine residue is relatively easy and can be done by analyzing the presence of the conserved nucleophilic elbow pattern 'G-X-S-X-G' (where G is glycine; X denotes 'any' amino acid; S is for Serine) in the amino acid sequences of FAEs (Cygler et al., 1993; Dodson & Wlodawer, 1998; Ghosh et al., 2001; McAuley et al., 2004). As described in **PAPER III**, *in silico* approaches can be used in predicting the active site residues accurately using amino acid titration curves and further can provide insights into the binding pocket microenvironments of FAEs. As shown in Figure 5 of **PAPER III**, identification of active ASP and HIS residues in all FAEs was based on analyzing the residues possessing different predicted titration function from the amino acid titration curves obtained through molecular simulations. Inspecting the structural framework of amino acids surrounding the catalytic triad residues provided a basis for understanding the factors responsible for different titration curves of the active site residues among different FAEs (**PAPER III**). Furthermore, it was observed that the position of each amino acid residue in the FAE binding pockets, in addition to their collective motions, determines their capability to accommodate different substrates, which dictates their substrate specificity (amino acid residue constellations in the binding pockets of three FAEs with different substrate specificity are shown in the Supplementary Figure S3 of **PAPER III**). The substrate selectivity of an enzyme is dependent on the constellation of amino acid residues forming the active site and can be changed by the mutations that occur during the evolutionary trajectory. Even though the details remain unclear, at some point in the evolution of FAEs there should be a handful of peptides with esterase activity that diversified into enzymes with varied substrate specificity through acquisition of new variations in their binding pockets. Therefore, a protein with certain substrate specificity can evolve into any other protein with different substrate specificity through a series of functional intermediates.

4.1 Enzyme classification *vs* Enzyme Evolution: a case study

Why does a protein or enzyme need to acquire novel function or substrate specificity? The answer for this is often associated with the process of adaptive evolution, which in simple terms reflects the adaptation of an organism towards a phenotype that best fits its current environment. One such example is the adaptation of bacteria to novel environmental conditions or carbon sources (Elena & Lenski, 2003). At the molecular level, this adaptation is driven by protein evolution to acquire novel functions guided by the adaptive amino acid substitutions in their sequences (Bull & Otto, 2005). Novel substrate specifities can be acquired in a matter of few years, as the ability to degrade synthetic chemicals appeared in microorganisms during the 20th century (Wackett, 2004).

Another question that pops-up is how a protein can suddenly gain a novel function without sacrificing the old one. A novel function may not be mutationally adjacent to the ancestral function and only beneficial mutations result into additive substrate specificity (Lynch et al., 2001). Furthermore, it has been observed that several enzymes are promiscuous and can act on different substrates with varying efficiencies (Khersonsky et al., 2006; Khersonsky & Tawfik, 2010; Yang & Metcalf, 2004).

In the case of FAEs, the promiscuous nature has led to the creation of 12 FAE families based on the differences in the protein properties acquired through evolution. As described before, the active serine residue in FAEs was identified based on the presence of the nucleophilic elbow pattern. I came across a total of 70 putative FAEs (out of 365 FAE sequences used for the descriptor based classification system described in **PAPER I**) with more than one nucleophilic elbow, which potentially indicate gene recombination events in the evolution of the FAE proteins. This observation led to a major outstanding question whether the members of FAE sub-families are still functional intermediates in the evolutionary path. But, it is clear that evolution readily derives novel functions from existing proteins.

One particular amino acid sequence of the descriptor based classification that drew my attention is a putative esterase from *Sorangium cellulosum* Soce56 that possesses five nucleophilic elbows, which I termed as 'Multiple Nucleophilic Elbowed Esterase' (MNEE) in **PAPER IV**. Biochemical characterization to probe the function of each binding pocket with a nucleophilic elbow in this enzyme revealed its ability to act on substrates of six different esterase activities (**Figure 11**). Does this particular enzyme of

S. cellulosum Soce56 evolved (or still evolving?) to become a generalist[4] instead of a specialist[5]? Can this evolutionary process show any positive effect for the organism? I can hypothesise that, being a soil-dwelling bacterium, to have promiscuous enzymes involved in plant polysaccharide degradation with broad substrate specificity can be an advantage to *S. cellulosum*. Such promiscuous enzymes help the microorganisms in adaptation to novel habitats with a myriad of substrates and altered environmental conditions. It has been shown that enzyme groups that act on plant biomass, that constitute a structural diverse set of substrates, can generally hydrolyze several alternative substrates and therefore possess the promiscuous behaviour of multiple substrate specificity (Cantarel et al., 2009; Turcot-Dubois et al., 2007). The results presented in **PAPER IV** indicate that broad substrate specificity acquired by MNEE comes at the price of low reaction turnover number for its original feruloyl esterase activity; whereas the nature of the reaction catalyzed is unchanged (MNEE was predicted as putative feruloyl esterase as part of descriptor based classification system). From the specific activity data (presented in Table 2 and Table 3 of **PAPER IV**), it can be speculated that MNEE subsequently sacrificed its efficiency of FAE activity with the emerging new additional binding pockets in its protein scaffold with non-FAE activities. Even though enzymes cannot be freely mutated for acquiring novel substrate specificities without interruption of their original or starting substrate specificity, it has been proposed that evolutionarily adapting enzymes have promiscuous activities (Khersonsky et al., 2006). Generally, enzyme evolution focuses on the acquisition of novel activities and during this process suppression of the original activity is important to be selective in their action. Being highly selective is an advantage for the enzymes involved in metabolic pathways, but for the enzymes that are involved in plant polysaccharide degradation, being promiscuous with broad substrate specificity can be an advantage.

[4] Enzymes that can promiscouously catalyze reactions on a variety of substrates or display multifunctionality or with different active sites are termed as ‘generalists’.

[5] Enzymes that can specifically catalyze one reaction or display activity on a unique substrate are termed as ‘specialists’.

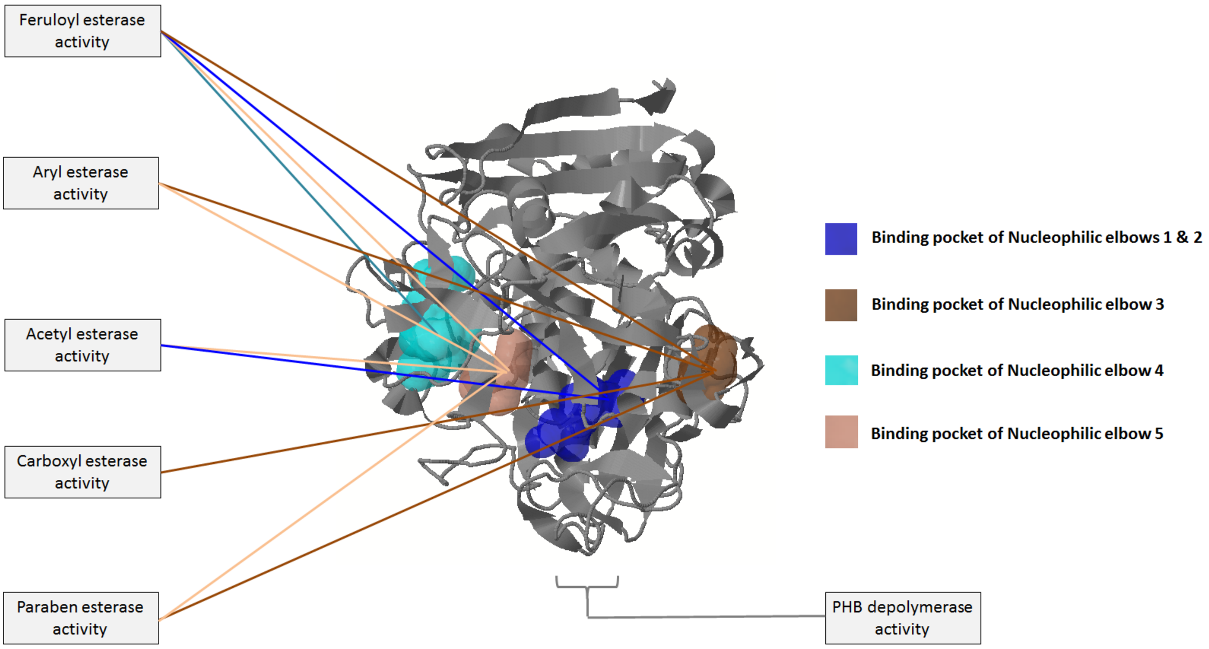

Feruloyl esterase activity
Aryl esterase activity
Acetyl esterase activity
Carboxyl esterase activity
Paraben esterase activity
PHB depolymerase activity
Binding pocket of Nucleophilic elbows 1 & 2
Binding pocket of Nucleophilic elbow 3
Binding pocket of Nucleophilic elbow 4
Binding pocket of Nucleophilic elbow 5

Figure 11. Structure of Multiple Nucleophilic Elbowed Esterase (MNEE) and its binding pockets. MNEE showed hydrolytic activity on the substrates of six different enzymes and mutational analysis (**PAPER IV**) revealed that each binding pocket possess both unique and overlapping substrate specificities.

Does the protein evolution have bias between metabolic enzymes which are generally assumed as 'specialists' and secreted biomass degrading enzymes which are generally assumed as 'generalists' (Copley, 2012; Nam et al., 2012)? How common is the promiscuous behaviour in metabolic enzymes? A very recent study published few months ago showed that an estimated 37% of enzymes in *Escherichia coli* are generalists and exhibit substrate promiscuity (Nam et al., 2012). Why a fraction of generalist metabolic enzymes are maintained in the evolutionary path? It might be the flux of metabolites that renders selective pressure on the organism to carry out the different catalytic processes while maintaining the low levels of total enzyme concentration. The same assumption can be applied to the activities observed in the MNEE of *S. cellulosum* Soce56 (Table 2 & 3 of PAPER IV). It is evident from the enzyme activity data, that MNEE has low feruloyl esterase activity. This might be due to the presence of low amount of feruloyl groups compared to other ester bonds in the plant biomass (Caffall & Mohnen, 2009; Faulds, 2010; Heredia et al., 1995) present in the habitat of *S. cellulosum* Soce56. This might have provided selective pressure to retain low amount of the feruloyl esterase activity in MNEE. Furthermore, the classification of enzymes with multiple active sites arisen from the selective pressure remains challenging and it would be counterproductive to attempt a classification based on function or structure (**Figure 12**).

Functional evolution can be inferred from the changes in protein structural dynamics (Lai et al., 2012). When the function is conserved, the structural dynamics relevant to enzyme function is also expected to be evolutionarily conserved. Mutations in the protein sequence in the process of evolution can have an effect on the catalytic activity through small changes in local structure of the active site. Few mutations in the evolutionary path of enzymes may affect the local structure, that does not change the catalytic activities, but may change the catalytic parameters of the enzyme and thus creating merely an enzyme variant with different substrate affinity (Kurtovic et al., 2008). It is also worthwhile to mention here that, when FAEs from different sub-families were studied in **PAPER III**, functional promiscuity in FAEs was found to be linked with conformational diversity of the active site for accommodating different substrates. Can a protein like MNEE with multiple active-sites be generated by mutations in the protein evolution? Such proteins must be the result of events like gene duplications, gene transfers and rearrangements of DNA sequences encoding different enzymes, resulting in redesigning of entire structure to form proteins with promiscuous activity or proteins with multiple active sites (Gerlt & Babbitt, 2009; Innan & Kondrashov, 2010; Voigt et al., 2001). Recombination between different genes or gene copies allows further exploration of combination of mutations leading to proteins like MNEE, in which each binding pocket is featured by promiscuous activity (**PAPER IV**).

In addition, it is apparent from the FAEs shown in Figure 1 of **PAPER III** that proteins can undergo significant changes in sequence in the evolutionary path and can still inherit the structural folds that are responsible for maintaining the same or similar substrate specificity. The evolutionary relationships in sequence-structure-function indeed exist between FAEs that were classified into sub-families by the combination of descriptor based classification system and catalytic triad constellations (PAPER I). It is unlikely that the FAE sub-families appeared independently, but they most probably evolved from a smaller set of generalist and less diverse ancestral proteins (PAPER III).

The points discussed above regarding the evolutionary space of protein sequence-structure-function is complex and in many ways defy classification systems based on only sequence or structural similarity (see CHAPTER 2). As shown in Figure 12, difficulties in defining the function or substrate specificity of an enzyme occur at all levels of classification hierarchy, due to the promiscuous nature of proteins in the evolutionary path. The research work described in this thesis suggests that strategies using structure-function relationships may offer a more reliable classification and a robust approach for function annotation for the sequences within an enzyme family.

Structure-function relationship studies are not only useful in understanding the substrate specificity or function of the enzymes and further their classification based on it, but are also useful in understanding how the functional efficiency of the enzymes changes according to reaction conditions. Two of such reaction condition cases were studied as part of the research work of this thesis: the first is the understanding of the effect of pH on the activity of enzyme (PAPER III) and the second is the understanding the pH dependent immobilization efficacy on mesoporous silica for biocatalytic synthesis (PAPER V).

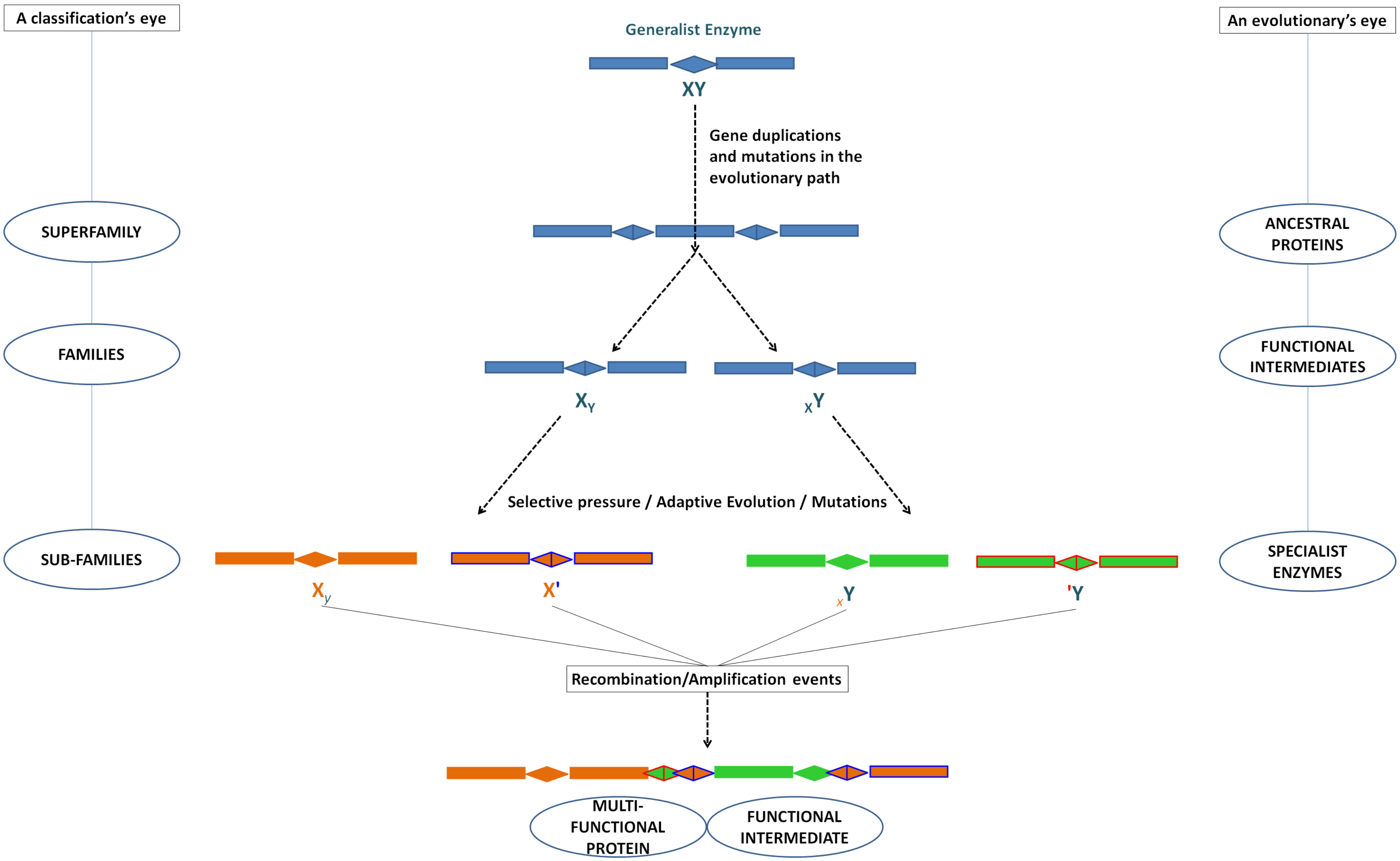

A classification's eye
SUPERFAMILY
FAMILIES
SUB-FAMILIES
Generalist Enzyme
XY
Gene duplications and mutations in the evolutionary path
X_Y
$_XY$
Selective pressure / Adaptive Evolution / Mutations
X_y
X'
$_xY$
'Y
Recombination/Amplification events
MULTI-FUNCTIONAL PROTEIN
FUNCTIONAL INTERMEDIATE
An evolutionary's eye
ANCESTRAL PROTEINS
FUNCTIONAL INTERMEDIATES
SPECIALIST ENZYMES

Figure 12. The emergence of specialist enzymes from the ancestral generalist enzymes in the view of classification and evolution (Bergthorsson et al., 2007; Freilich et al., 2005; Hughes, 1994; Park et al., 2006). Classification is feasible or functionally meaningful till the evolutionary path of specialist enzymes with defined substrate specificity. A challenge is posed for the classification system, when the evolutionary path gives rise to proteins with multiple active sites through duplication and divergence of genes. Novel enzymes can be further evolved from the specialized enzymes by entering again the phase of functional intermediates.

The scheme depicted is for a hypothetical generalist enzyme that possesses two different activities. The starting point in the generation of specialist enzymes is a generalist enzyme (**XY**), where duplication of genes encoding it leads to division of its ancestral functions and generate enzymes ($\mathbf{X_Y}$, $\mathbf{{}_XY}$) which are catalytically promiscuous with varied affinity towards substrates. Further mutations guided by adaptive evolution may give rise to specialist enzymes. Even though the active sites of few specialist enzymes ($\mathbf{X}_y$, ${}_x\mathbf{Y}$) are very specific to a set of substrates, molecules that bear resemblance to their natural substrates can bind with lower affinity. When such molecules bind in correct orientations the reactive functional groups of the enzymes active site catalyze chemical reactions, which gives the promiscuous property to specialist enzymes. Further mutations from this point may also give rise to more specific enzymes (**X'**, **'Y**) evolved to catalyze a reaction with more specificity and catalytic efficiency. Still, multi-functional enzymes can be emerged from specialist enzymes by gene duplication and recombination events. These multi-functional enzymes further may repeat the entire cycle described above giving rise to novel and multiple promiscuous enzymes.

CONCLUSIONS

The work presented in this thesis integrates bioinformatics, cheminformatics and protein biochemistry tools to explore the sequence-structure-function relationships of enzymes, especially for the structurally and catalytically promiscuous enzyme group, FAEs. Based on the results, observations presented and the methodology developed, I strongly believe that advancement was made in classification and functional description of a promiscuous enzyme family.

Establishment of a classification scheme reflecting the function/substrate specificity of FAEs

The information on an enzyme's functional specificity is necessarily contained in its protein sequence, but the classification schemes using sequence information alone are not successful in sub-grouping enzymes according to their functional specificity. Even the classification schemes based on structural information alone do not perform well in functional sub-grouping of enzymes. Building a predictive model for functional classification was performed by incorporation of information on enzyme properties through their protein sequence descriptors, along with the substrate pharmacophore features. This approach resulted in a reliable classification of FAEs (**PAPER I**). Even though FAEs possess common characteristics, such as the classic constellation of the Ser-His-Asp triad, variations in amino acid sequences forming surface loops and additional domains allow them to accommodate diverse substrates. Using the properties of the whole protein sequence, a new classification system was proposed for FAEs resulting into 12 distinct families, while by careful inspection of the catalytic residues constellation in the sequences of each FAE family they were divided into 32 sub-families reflecting substrate specificity. I should emphasize the fact that the classification system proposed in **PAPER I**, does not contradict, but rather significantly expands, the current knowledge in the area and allows a systematic sub-grouping of FAEs.

Reliable prediction of substrate specificity in FAEs

As the FAE enzymes show substrate promiscuity, it is important that a classification system can capture the great variety in substrate specificity these enzymes show. A molecular docking approach was used for the prediction of unique and overlapping substrate specificities in FAE families. The problem generally arises in the first step of molecular docking process i.e., choosing the right docking program among several commercial and academic softwares. Evaluation of 88 docking algorithm-scoring function sets was performed (**PAPER II**) with the aim to identify the docking program that can predict the substrate-activity maps of the members of the various FAE families. The ultimate challenge for a docking program is to correctly predict the overlapping and unique substrate specificity profiles of the FAE families, which will position it superior among the others. Comparison of molecular docking programs for pose prediction and enrichment showed that there is significant variability on the performance of docking programs based on the specific target protein. Studies on evaluation of docking programs are problematic by the fact that docked ligand poses are penalized and considered incorrect from 2Å to an infinitely poor RMSD. Such a crude RMSD cut-off cannot rescue correct ligand poses with high RMSD. Even though the traditional approach of evaluating the docking programs using the RMSD is commonly used, the main drawback is not taking into account the interactions between the ligand and the receptor. An assessment measure called Key Interaction Score System (KISS) was proposed to overcome the drawbacks mentioned above. The KISS has the ability to identify the beneficial docking poses irrespective of the RMSD value. RMSD is strictly a measure of fit based on the proportion of atoms aligned with the crystallographic pose, whereas the KISS also considers docked poses with badly aligned atoms if they were able to form the same hydrogen bond interactions observed in the crystallographic pose. The KISS thus reduces the problem of flexibility arising from the large number of poses or conformers. Though KISS may not solve all the issues with the current docking algorithms and scoring functions, combining with RMSD will avoid discarding realistic poses. The approach of combining RMSD and KISS was able to predict the experimental substrate specificity of FAEs, when the best performing docking program was used.

Understanding the structure-function relationships in FAEs

PAPER III and **PAPER IV** deal with understanding the structure-function relationships of FAEs from *Aspergillus niger* and a promiscuous enzyme featured by multiple nucleophilic elbows previously predicted as a putative FAE, respectively. In this part of the work I have used the theoretical framework established in the first papers to gain insights of the selected enzymes. Furthermore, by combining the *in silico* work with the experimental investigations, the capabilities of the theoretical framework were confirmed. **PAPER V** was a starting point towards probing how structure-function relationships for biocatalytic reaction are influenced by an enzyme immobilization process.

In **PAPER III**, through the structural analysis of 13 FAEs from *A. oryzae*, high similarity in the secondary structure elements (SSEs) was observed between the members that belong to the same FAE family. However, there was no consensus on the structural features that contribute to the substrate specificity between different FAE family members (**PAPER III**). The modeled FAE structures suggested that, with a limited set of structural scaffolds, FAEs evolved into different families and further analysis of binding pockets indicated the topological variations of FAEs that led to a wide spectrum of substrate specificities. The active site residues of FAEs were identified using amino acid titration curves obtained through molecular dynamics simulations. Together with 3D mapping of the enzyme binding pockets the microenvironment of amino acid residues that dictates the enzyme activity were revealed.

In **PAPER IV**, probing the function of each nucleophilic elbow of *S. cellulosum* MNEE revealed that each nucleophilic elbow forms a local active site with one or more enzyme activities. To the best of my knowledge, this is the first study in the literature that showed the presence of four binding pockets in a single protein domain and further proves the interplay of multiple nucleophilic elbows and catalytic promiscuity of esterases. All the binding pockets of MNEE showed ester hydrolysis capability with different substrate specificities. My analysis also showed that broad substrate specificity acquired by MNEE comes at the price of low reaction turnover number for its assumed original feruloyl esterase activity; whereas the nature of the reaction catalyzed is unchanged. The work presented in **PAPER IV** indicates that substrate selectivity of an enzyme is dependent on the constellation of amino acid residues forming the active site and can be changed by the mutations that occur during the evolutionary trajectory. Protein domains are the evolutionary units of the enzyme structure; furthermore, their combinations gives rise to multi-domain enzymes (Vogel et al., 2004). In such a case, each domain can have an independent function or contribute to the common function of the enzyme (Teichmann et al., 1998). However, MNEE is a small protein comprised of a single domain and still possesses four distinct binding pockets. Investigation of MNEE protein 3D structure using Domain Reconstruction Algorithm (Gelly et al., 2006) showed that MNEE is comprised of ten small protein units. Further hierarchical splitting of protein units in MNEE 3D structure indicated the possibility of MNEE being an intermediate enzyme resulted from recombination of protein coding DNA sequences in the evolutionary process. Using the framework presented in **PAPER IV**, identification of multiple nucleophilic elbows that form distinct binding pockets in enzymes can help to identify new catalytic sites. Furthermore, it represents a starting point to understand the multi-dimensional nature of enzyme evolution. The strategy used by nature to evolve unique enzyme activities can be transferred as principles to be used in enzyme engineering.

Understanding structure-function relationships can also help in optimizing the reaction conditions of enzyme based biocatalytic applications. For satisfactory stability and easy recovery of enzyme based biosynthetic reactions, it is often necessary to immobilize the enzymes to a solid support material. The work presented in **PAPER V** deals with combining experimental results with *in silico* modeling in order to analyse the environment of the enzyme binding pocket and protein surface factors involved in immobilization process. In PAPER V, the modeled FAE structure enabled us to inspect the structural changes at different pH conditions and further understand its immobilization profile. Through molecular simulations, the pH dependent immobilization and activity profile of FAE was found to depend on the charged surface interactions and binding pocket microenvironment, respectively.

PERSPECTIVES

Knowledge about the enzyme function is of utmost importance for taking full advantage of an enzyme's full capacity in biocatalytic applications. For a promiscuous enzyme group like FAEs, some of the challenges would be to identify the enzyme that has the best potential to perform a selected reaction.

The framework (**Figure 1**) for functional grouping and description of FAEs that were put up in the thesis included implementation of novel classification approaches (**PAPER I**), computational prediction of substrate selectivity (**PAPER II**), experimental validation of the computational predictions (**PAPER II** & **III**), unravelling the structure-function relationships of a putative FAE possessing multiple active sites (**PAPER IV**), and understanding the molecular effects of reaction conditions (**PAPER III** & **V**). I believe that the framework of integrating *in silico* biology and enzymology developed during my thesis work can be applied towards functional classification and understanding of the sequence-structure-function relationships within any promiscuous enzyme family.

Several methods that are based on sequence and structural similarities for classification of proteins suffer from limitations in annotating promiscuous enzymes. As presented in **PAPER I**, machine learning methods work well only for certain combinations of protein sequence properties or descriptors. It seems that there is no preferred combination of descriptor sets that could be utilized for sub-grouping of enzymes that reflects their substrate specificity, as the clustering performance does not differ significantly for few descriptor combinations (**Table 4**). The selection of the best descriptor set was based on available experimental substrate specificity data. Even though the work presented in **PAPER I** shows that the choice of an optimal descriptor set and machine learning algorithm are critical for the classification, more accurate functional sub-grouping required integration of structural features involved in catalytic function, e.g. constellation of catalytic triad in FAEs and pharmacophore features of substrates.

However, we still have to elucidate the relationship between the sequence properties that guided clustering and the structural properties that guided the functional sub-grouping of FAEs. Such sequence to structure relationship information can be used for reliable classification and engineering the selected enzymes for a required substrate specifity. The use of common feature based pharmacophores to represent the substrate specificity of FAE sub-families in **PAPER I** indicated that the time has come for the utilization of established cheminformatics tools in enzymology. I envisage that the future development of algorithms for automated identification and extraction of features responsible for functional classification may provide opportunities for ensemble approaches in the classification and functional description of any poorly understood protein or enzyme family. The pharmacophore models developed for FAE sub-families could be applied for virtual screening of compound databases for the identification of potential substrates through molecular docking approach. Pharmacophores can also be used as 3D filters for post-processing the docked ligand poses to remove the false positives in the molecular docking process.

In **PAPER II**, several molecular docking programs were evaluated for substrate specificity predictions of FAEs. A docking program that can predict the unique and overlapping substrate specificity profile of the FAE families, will position it superior among the others and more suitable for enzymes with promiscuous properties. The combination of RMSD and KISS proved to be more meaningful in measuring the docking accuracy for selection of an appropriate docking program. Generally, docking programs include both a docking algorithm for the analysis of different ligand confirmations and a scoring function that should ideally be able to rank the ligands according to the experimental binding affinity. The scoring functions evaluated in **PAPER II** still remain weak predictors of binding affinity and are not able to rank-order the substrates of FAEs according to experimental data. Assigning the lowest energy score to the correct binding pose proved to be a major challenging task for the scoring functions, which is the major reason for the inability to rank-order the compounds. The binding affinity of a ligand also depends upon the collective interactions with binding pocket residues of the receptor, which makes the rank-ordering task more challenging for scoring functions. Unfortunately, the K_m values (the measure of affinity) of the FAEs used in the evaluation study are quite close among different substrates: this poses a major challenge for the scoring schemes to rank-order the substrates. Development of target protein dependent scoring functions may help in reliable rank-ordering of the substrates that can be preferentially catalyzed by enzymes.

The research work presented in **PAPER III** was focused on experimental validation of computational predictions and analysis of structure-function relationships in FAEs. Future structural studies with different cognate ligand-receptor complexes using X-ray crystallography/NMR complemented with analysis of cognate ligand-mutated receptor complexes will further extend our understanding of characteristic fingerprints that guide the varied substrate specificities among the members of different FAE families. Nevertheless, using the experimental data, the predicted 3D structures can be verified and the advancements can be made in the algorithms by knowing the regions that were modeled or predicted incorrectly. Protein structure modeling often involves human interventions for the selection of template protein structures and subsequent loop modeling process. Consequently, development of fully-automated algorithms for reliable protein structure predictions will remove the human bias for template structure or protein loops in the structure refinement process. Predicting high quality protein structures using *in silico* approaches has also the advantage of providing an ensemble of engineered enzyme structures for desired biosynthetic reactions in short time.

Experimental analysis of substrate specificity of each binding pocket of MNEE described in **PAPER IV** indicates that the possible enzyme active sites in proteins have not yet been fully explored. Advances can be made with high-throughput screening methods of such multiple nucleophilic elbowed enzymes (or any other enzymes with multiple conserved catalytic signatures) that would have a major impact on the development of new biocatalysts. Such enzymes with multiple substrate specificity have potential industrial application for the development of 'in-pot enzyme processes' (Kim et al., 2011; Yu et al., 2006). In general, in-pot enzyme processes are characterized by the mixture of enzymes that catalyze several reactions in a single pot. In-pot enzyme processes eliminate the need of purification steps of intermediate products and reduce the downstream processing and operating costs.

ACKNOWLEDGEMENTS

Pursuing my doctoral studies in Sweden has given me the greatest opportunity I could imagine till now. First of all, I would like to thank the education system and research settings at the Chalmers University of Technology that have made me evolving much faster from a researcher into a young independent scientist. It would not have been possible to write this doctoral thesis without the help and support of the people around me, to only some of whom it is possible to give particular mention here.

I would like to express my heartfelt gratitude to my supervisor, Professor Lisbeth Olsson. I could not have asked for better role model, who has been inspirational and supportive in the up and downs of the PhD journey. I hope that in my next steps of the career I can accommodate the research values and patience learned from you. With my approach of involving in several research projects at the same time, I would have been lost without your advice on time management and priority setting. I admire your ability to tackle unexpected problems both at scientific and administrative levels, hats off to you, Lisbeth.

I would like to give my highest appreciation to my co-supervisor, Associate Professor Gianni Panagiotou. I can read, write and speak three languages. But, I found it difficult to find and explain in words the guidance, mentoring, support, and friendship that you have extended to me. You gave me unlimited freedom of thinking in the projects we have been collaborating, both in and out of the framework of my PhD thesis, and trusted in my potential to execute them since the time we met. I have often witnessed your unwavering support to your students in all situations that makes you truly unique.

I am very grateful to Associate Professor Irene Kouskoumvekaki for her supervision and support during the first critical step of my PhD period. Thank you for the valuable discussions and your encouragement during my stay as guest PhD student in your research group at the Technical University of Denmark, where the idea of integrating bio-cheminformatics in enzymology was sowed in my mind.

I should not forget to acknowledge all the members of Industrial Biotechnology group for providing an intellectually inspiring and healthy working environment. Thanks to the efforts of Lisbeth to ensure a constructive working environment for the group. I would like to thank Valeria Mapelli for providing guidance in molecular biology steps of my thesis work. During the last 4 years whenever I contacted Hampus Sunner for practical issues related to my computer or programming, he always helped me. Hampus, without your timely help, this work would have taken much longer time. Special thanks to Erica Dahlin for taking care of practicalities involved in my stay Chalmers. I would like to thank Jenny Nilsson, who helped not only to me, but to the Industrial Biotechnology group in general, for smooth functioning of both old and new research labs. I would like to extend my thanks to all the co-authors and collaborators of my manuscripts.

My deepest gratitude belongs to my mother Durga Devi and to my grandfather Upendra Rao. Unfortunately, they are not here to share the PhD celebration with me, but I am sure that they keep an eye on me from heaven. Special thanks to my brother GSRK Gupta, who has always supported and motivated me in life so far. Lastly, and most importantly, I would like to thank Yassell Lopez Rodriguez, without whom life would be bleak. I will never forget your ever-present support. I thank your dance lessons that make my mind to refresh every day. Your music and beautiful singing, and the special batch of friends (Kamilla Wöldike Breum, Laurry Ramirez, Lilian Lillkung, Marianne Steen) that I got from you made me laugh every day to forget the stress that often arise from research projects. This thesis work is dedicated to you….

Dasaradhi Bala Rama Krishna Gupta Udatha
January 2013

REFERENCES

Almonacid, D.E., Babbitt, P.C. 2011. Toward mechanistic classification of enzyme functions. *Curr Opin Chem Biol,* **15**(3), 435-42.

Altschul, S.F., Gish, W. 1996. Local alignment statistics. *Methods Enzymol,* **266**, 460-80.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* **25**(17), 3389-402.

Audit, B., Levy, E.D., Gilks, W.R., Goldovsky, L., Ouzounis, C.A. 2007. CORRIE: enzyme sequence annotation with confidence estimates. *BMC Bioinformatics,* **8 Suppl 4**, S3.

Babbitt, P.C. 2003. Definitions of enzyme function for the structural genomics era. *Curr Opin Chem Biol,* **7**(2), 230-237.

Babbitt, P.C., Gerlt, J.A. 1997. Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. *J Biol Chem,* **272**(49), 30591-4.

Barnum, D., Greene, J., Smellie, A., Sprague, P. 1996. Identification of common functional configurations among molecules. *J Chem Inf Comput Sci,* **36**(3), 563-71.

Benoit, I., Danchin, E.G., Bleichrodt, R.J., de Vries, R.P. 2008. Biotechnological applications and potential of fungal feruloyl esterases based on prevalence, classification and biochemical diversity. *Biotechnol Lett,* **30**(3), 387-96.

Bergthorsson, U., Andersson, D.I., Roth, J.R. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences of the United States of America,* **104**(43), 17004-17009.

Bhasin, M., Raghava, G.P. 2004. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem,* **279**(22), 23262-6.

Bissantz, C., Folkers, G., Rognan, D. 2000. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem,* **43**(25), 4759-67.

Borro, L.C., Oliveira, S.R.M., Yamagishi, M.E.B., Mancini, A.L., Jardine, J.G., Mazoni, I., dos Santos, E.H., Higa, R.H., Kuser, P.R., Neshich, G. 2006. Predicting enzyme class from protein structure using Bayesian classification. *Genetics and Molecular Research,* **5**(1), 193-202.

Brenner, S.E. 1999. Errors in genome annotation. *Trends Genet*, **15**(4), 132-3.

Brenner, S.E., Chothia, C., Hubbard, T.J. 1997. Population statistics of protein structures: lessons from structural classifications. *Curr Opin Struct Biol*, **7**(3), 369-76.

Bryson, V., Vogel, H.J. 1965. Evolving Genes and Proteins. *Science*, **147**(3653), 68-71.

Bull, J.J., Otto, S.P. 2005. The first steps in adaptive evolution. *Nat Genet*, **37**(4), 342-3.

Bursulaya, B.D., Totrov, M., Abagyan, R., Brooks, C.L. 2003. Comparative study of several algorithms for flexible ligand docking. *Journal of Computer-Aided Molecular Design*, **17**(11), 755-763.

Caffall, K.H., Mohnen, D. 2009. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydr Res*, **344**(14), 1879-1900.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, Y.Z. 2004. Enzyme family classification by support vector machines. *Proteins*, **55**(1), 66-76.

Cai, Y.D., Zhou, G.P., Chou, K.C. 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. *J Theor Biol*, **234**(1), 145-149.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*, **37**(Database issue), D233-8.

Carbonell, P., Faulon, J.L. 2010. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics*, **26**(16), 2012-9.

Challis, G.L., Ravel, J., Townsend, C.A. 2000. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol*, **7**(3), 211-24.

Chen, H., Lyne, P.D., Giordanetto, F., Lovell, T., Li, J. 2006. On evaluating molecular-docking methods for pose prediction and enrichment factors. *Journal of Chemical Information and Modeling*, **46**(1), 401-15.

Chiang, R.A., Sali, A., Babbitt, P.C. 2008. Evolutionarily Conserved Substrate Substructures for Automated Annotation of Enzyme Superfamilies. *PLoS Comput Biol*, **4**(8).

Chothia, C. 1992. Proteins. One thousand families for the molecular biologist. *Nature*, **357**(6379), 543-4.

Copley, S.D. 2012. Toward a systems biology perspective on enzyme evolution. *J Biol Chem*, **287**(1), 3-10.

Crepin, V.F., Faulds, C.B., Connerton, I.F. 2004. Functional classification of the microbial feruloyl esterases. *Appl Microbiol Biotechnol,* **63**(6), 647-652.

Cross, J.B., Thompson, D.C., Rai, B.K., Baber, J.C., Fan, K.Y., Hu, Y.B., Humblet, C. 2009a. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J Chem Inf Model,* **49**(6), 1455-1474.

Cross, J.B., Thompson, D.C., Rai, B.K., Baber, J.C., Fan, K.Y., Hu, Y.B., Humblet, C. 2009b. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *Journal of Chemical Information and Modeling,* **49**(6), 1455-1474.

Cummings, M.D., DesJarlais, R.L., Gibbs, A.C., Mohan, V., Jaeger, E.P. 2005. Comparison of automated docking programs as virtual screening tools. *J Med Chem,* **48**(4), 962-76.

Cygler, M., Schrag, J.D., Sussman, J.L., Harel, M., Silman, I., Gentry, M.K., Doctor, B.P. 1993. Relationship between Sequence Conservation and 3-Dimensional Structure in a Large Family of Esterases, Lipases, and Related Proteins. *Protein Science,* **2**(3), 366-382.

Devos, D., Valencia, A. 2001. Intrinsic errors in genome annotation. *Trends Genet,* **17**(8), 429-31.

Dixon, M., Webb, E.C. 1964. Enzymes. *Longmans, Green & Co., London, and Academic Press, New York,* **2nd Edition**.

Dobson, P.D., Doig, A.J. 2005. Predicting enzyme class from protein structure without alignments. *J Mol Biol,* **345**(1), 187-199.

Dodson, G., Wlodawer, A. 1998. Catalytic triads and their relatives. *Trends Biochem Sci,* **23**(9), 347-352.

Ekici, O.D., Paetzel, M., Dalbey, R.E. 2008a. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. *Protein science : a publication of the Protein Society,* **17**(12), 2023-37.

Ekici, O.D., Paetzel, M., Dalbey, R.E. 2008b. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. *Prot Sci,* **17**(12), 2023-37.

Elena, S.F., Lenski, R.E. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet,* **4**(6), 457-69.

Evers, A., Hessler, G., Matter, H., Klabunde, T. 2005. Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J Med Chem,* **48**(17), 5448-65.

Faulds, C.B. 2010. What can feruloyl esterases do for us? *Phytochemistry Reviews*, **9**(1), 121-132.

Fazary, A.E., Ju, Y.H. 2007. Feruloyl esterases as biotechnological tools: current and future perspectives. *Acta Biochim Biophys Sin (Shanghai)*, **39**(11), 811-28.

Ferrara, P., Gohlke, H., Price, D.J., Klebe, G., Brooks, C.L., 3rd. 2004. Assessing scoring functions for protein-ligand interactions. *J Med Chem*, **47**(12), 3032-47.

Finn, R.D., Clements, J., Eddy, S.R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, **39**(Web Server issue), W29-37.

Freilich, S., Spriggs, R.V., George, R.A., Al-Lazikani, B., Swindells, M., Thornton, J.M. 2005. The complement of enzymatic sets in different species. *J Mol Biol*, **349**(4), 745-763.

Galperin, M.Y., Koonin, E.V. 1999. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica*, **106**(1-2), 159-70.

Galperin, M.Y., Walker, D.R., Koonin, E.V. 1998a. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*, **8**(8), 779-90.

Galperin, M.Y., Walker, D.R., Koonin, E.V. 1998b. Analogous enzymes: independent inventions in enzyme evolution. *Genome research*, **8**(8), 779-90.

Gelly, J.C., de Brevern, A.G., Hazout, S. 2006. 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments. *Bioinformatics*, **22**(2), 129-33.

Gerlt, J.A., Babbitt, P.C. 2001. Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem*, **70**, 209-246.

Gerlt, J.A., Babbitt, P.C. 2009. Enzyme (re)design: lessons from natural evolution and computation. *Curr Opin Chem Biol*, **13**(1), 10-8.

Gerlt, J.A., Raushel, F.M. 2003. Evolution of function in (beta/alpha)8-barrel enzymes. *Curr Opin Chem Biol*, **7**(2), 252-64.

Ghosh, D., Sawicki, M., Lala, P., Erman, M., Pangborn, W., Eyzaguirre, J., Gutierrez, R., Jornvall, H., Thiel, D.J. 2001. Multiple conformations of catalytic serine and histidine in acetylxylan esterase at 0.90 A. *J Biol Chem*, **276**(14), 11159-66.

Glasner, M.E., Gerlt, J.A., Babbitt, P.C. 2006. Evolution of enzyme superfamilies. *Curr Opin Chem Biol*, **10**(5), 492-497.

Govindarajan, S., Recabarren, R., Goldstein, R.A. 1999. Estimating the total number of protein folds. *Proteins*, **35**(4), 408-14.

Guner, O., Clement, O., Kurogi, Y. 2004. Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. *Curr Med Chem,* **11**(22), 2991-3005.

Guner, O.F. 2011. Pharmacophore modeling in drug design: recent advances. *Curr Comput Aided Drug Des,* **7**(3), 158.

Han, L.Y., Cai, C.Z., Ji, Z.L., Cao, Z.W., Cui, J., Chen, Y.Z. 2004. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res,* **32**(21), 6437-44.

Hannenhalli, S.S., Russell, R.B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol,* **303**(1), 61-76.

Hegeman, G.D., Rosenberg, S.L. 1970. The evolution of bacterial enzyme systems. *Annu Rev Microbiol,* **24**, 429-62.

Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal,* **280 (Pt 2)**, 309-16.

Henrissat, B., Bairoch, A. 1993. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal,* **293 (Pt 3)**, 781-8.

Heredia, A., Jimenez, A., Guillen, R. 1995. Composition of Plant-Cell Walls. *Zeitschrift Fur Lebensmittel-Untersuchung Und-Forschung,* **200**(1), 24-31.

Hevener, K.E., Zhao, W., Ball, D.M., Babaoglu, K., Qi, J., White, S.W., Lee, R.E. 2009. Validation of molecular docking programs for virtual screening against dihydropteroate synthase. *Journal of Chemical Information and Modeling,* **49**(2), 444-60.

Horowitz, N.H. 1945. On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A,* **31**(6), 153-7.

Huang, N., Shoichet, B.K., Irwin, J.J. 2006. Benchmarking sets for molecular docking. *J Med Chem,* **49**(23), 6789-801.

Hughes, A.L. 1994. The Evolution of Functionally Novel Proteins after Gene Duplication. *Proceedings of the Royal Society B-Biological Sciences,* **256**(1346), 119-124.

Hult, K., Berglund, P. 2007. Enzyme promiscuity: mechanism and applications. *Trends Biotechnol,* **25**(5), 231-8.

Innan, H., Kondrashov, F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet,* **11**(2), 97-108.

Irwin, J.J. 2008. Community benchmarks for virtual screening. *J Comput Aided Mol Des*, **22**(3-4), 193-9.

IUBMB. 1965. Enzyme nomenclature. Report on the recommendations (1964) of the International Union of Biochemistry on Nomenclature and Classification of Enzymes. *Science*, **150**(3697), 719-21.

James, L.C., Tawfik, D.S. 2003. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences*, **28**(7), 361-8.

Jensen, R.A. 1976. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, **30**, 409-25.

Juncker, A.S., Jensen, L.J., Pierleoni, A., Bernsel, A., Tress, M.L., Bork, P., von Heijne, G., Valencia, A., Ouzounis, C.A., Casadio, R., Brunak, S. 2009. Sequence-based feature prediction and annotation of proteins. *Genome Biol*, **10**(2), 206.

Kanehisa, M. 1997. A database for post-genome analysis. *Trends Genet*, **13**(9), 375-6.

Karchin, R., Karplus, K., Haussler, D. 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**(1), 147-59.

Kellenberger, E., Foata, N., Rognan, D. 2008. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *Journal of Chemical Information and Modeling*, **48**(5), 1014-25.

Kellenberger, E., Rodrigo, J., Muller, P., Rognan, D. 2004. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, **57**(2), 225-42.

Khersonsky, O., Roodveldt, C., Tawfik, D.S. 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol*, **10**(5), 498-508.

Khersonsky, O., Tawfik, D.S. 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual review of biochemistry*, **79**, 471-505.

Kim, J.H., Wang, R., Lee, W.H., Park, C.S., Lee, S., Yoo, S.H. 2011. One-Pot Synthesis of Cycloamyloses from Sucrose by Dual Enzyme Treatment: Combined Reaction of Amylosucrase and 4-alpha-Glucanotransferase. *Journal of Agricultural and Food Chemistry*, **59**(9), 5044-5051.

Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J. 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, **3**(11), 935-49.

Koehl, P., Levitt, M. 2002. Sequence variations within protein families are linearly related to structural variations. *J Mol Biol*, **323**(3), 551-62.

Kontoyianni, M., McClellan, L.M., Sokol, G.S. 2004. Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem*, **47**(3), 558-65.

Kontoyianni, M., Sokol, G.S., McClellan, L.M. 2005. Evaluation of library ranking efficacy in virtual screening. *J Comput Chem*, **26**(1), 11-22.

Koseki, T., Fushinobu, S., Ardiansyah, Shirakawa, H., Komai, M. 2009. Occurrence, properties, and applications of feruloyl esterases. *Applied microbiology and biotechnology*, **84**(5), 803-10.

Koshland, D.E. 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **44**(2), 98-104.

Kristensen, D.M., Ward, R.M., Lisewski, A.M., Erdin, S., Chen, B.Y., Fofanov, V.Y., Kimmel, M., Kavraki, L.E., Lichtarge, O. 2008. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, **9**.

Kurogi, Y., Guner, O.F. 2001. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr Med Chem*, **8**(9), 1035-55.

Kurtovic, S., Shokeer, A., Mannervik, B. 2008. Emergence of novel enzyme quasi-species depends on the substrate matrix. *J Mol Biol*, **382**(1), 136-153.

Lai, J., Jin, J., Kubelka, J., Liberles, D.A. 2012. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. *J Mol Biol*, **422**(3), 442-59.

Laine, R.A. 1994. A Calculation of All Possible Oligosaccharide Isomers Both Branched and Linear Yields 1.05x10(12) Structures for a Reducing Hexasaccharide - the Isomer-Barrier to Development of Single-Method Saccharide Sequencing or Synthesis Systems. *Glycobiology*, **4**(6), 759-767.

Langer, T., Krovat, E.M. 2003. Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. *Curr Opin Drug Discov Devel*, **6**(3), 370-6.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. 2007. Clustal W and clustal X version 2.0. *Bioinformatics*, **23**(21), 2947-2948.

Latino, D.A.R.S., Zhang, Q.Y., Aires-De-Sousa, J. 2008. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics*, **24**(19), 2236-2244.

Linsky, T., Fast, W. 2010. Mechanistic similarity and diversity among the guanidine-modifying members of the pentein superfamily. *Biochimica Et Biophysica Acta-Proteins and Proteomics*, **1804**(10), 1943-1953.

Lynch, M., O'Hely, M., Walsh, B., Force, A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics*, **159**(4), 1789-804.

McAuley, K.E., Svendsen, A., Patkar, S.A., Wilson, K.S. 2004. Structure of a feruloyl esterase from Aspergillus niger. *Acta Crystallographica Section D-Biological Crystallography*, **60**, 878-887.

McClure, M.A., Smith, C., Elton, P. 1996. Parameterization studies for the SAM and HMMER methods of hidden Markov model generation. *Proc Int Conf Intell Syst Mol Biol*, **4**, 155-64.

McGaughey, G.B., Sheridan, R.P., Bayly, C.I., Culberson, J.C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.F., Cornell, W.D. 2007. Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling*, **47**(4), 1504-19.

Nam, H., Lewis, N.E., Lerman, J.A., Lee, D.H., Chang, R.L., Kim, D., Palsson, B.O. 2012. Network context and selection in the evolution to enzyme specificity. *Science*, **337**(6098), 1101-4.

Nasibov, E., Kandemir-Cavas, C. 2009. Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction. *Comput Biol Chem*, **33**(6), 461-464.

Nobeli, I., Favia, A.D., Thornton, J.M. 2009. Protein promiscuity and its implications for biotechnology. *Nature biotechnology*, **27**(2), 157-67.

Nowotny, M. 2009. Retroviral integrase superfamily: the structural perspective. *Embo Reports*, **10**(2), 144-151.

Omelchenko, M.V., Galperin, M.Y., Wolf, Y.I., Koonin, E.V. 2010a. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biology direct*, **5**, 31.

Omelchenko, M.V., Galperin, M.Y., Wolf, Y.I., Koonin, E.V. 2010b. Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. *Biology Direct*, **5**.

Ong, S.A., Lin, H.H., Chen, Y.Z., Li, Z.R., Cao, Z. 2007. Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics*, **8**, 300.

Onodera, K., Satou, K., Hirota, H. 2007. Evaluations of molecular docking programs for virtual screening. *Journal of Chemical Information and Modeling*, **47**(4), 1609-18.

Orengo, C.A., Jones, D.T., Thornton, J.M. 1994. Protein superfamilies and domain superfolds. *Nature*, **372**(6507), 631-4.

Park, H.S., Nam, S.H., Lee, J.K., Yoon, C.N., Mannervik, B., Benkovic, S.J., Kim, H.S. 2006. Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, **311**(5760), 535-538.

Pearson, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol*, **132**, 185-219.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, **183**, 63-98.

Perola, E., Walters, W.P., Charifson, P.S. 2004. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, **56**(2), 235-49.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., Huson, D.H. 2005. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res*, **33**(18), 5799-808.

Reymond, J.L., Fluxa, V.S., Maillard, N. 2009. Enzyme assays. *Chemical Communications*(1), 34-46.

Rost, B. 2002. Enzyme function less conserved than anticipated. *J Mol Biol*, **318**(2), 595-608.

Rost, B. 1997. Protein structures sustain evolutionary drift. *Fold Des*, **2**(3), S19-24.

Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., Schomburg, D. 2002. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem Sci*, **27**(1), 54-6.

Schulz-Gasch, T., Stahl, M. 2003. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J Mol Model*, **9**(1), 47-57.

Shah, I., Hunter, L. 1997. Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol*, **5**, 276-83.

Simon, G.M., Cravatt, B.F. 2010. Activity-based Proteomics of Enzyme Superfamilies: Serine Hydrolases as a Case Study. *Journal of Biological Chemistry*, **285**(15), 11051-11055.

Stachelhaus, T., Mootz, H.D., Marahiel, M.A. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol*, **6**(8), 493-505.

Stahl, M., Rarey, M. 2001. Detailed analysis of scoring functions for virtual screening. *J Med Chem*, **44**(7), 1035-42.

Stoll, F., Liesener, S., Hohlfeld, T., Schror, K., Fuchs, P.L., Holtje, H.D. 2002. Pharmacophore definition and three-dimensional quantitative structure-activity relationship study on structurally diverse prostacyclin receptor agonists. *Molecular Pharmacology*, **62**(5), 1103-11.

Sutter, J., Li, J.B., Maynard, A.J., Goupil, A., Luu, T., Nadassy, K. 2011. New Features that Improve the Pharmacophore Tools from Accelrys. *Curr Comput Aided Drug Des*, **7**(3), 173-180.

Syed, U., Yona, G. 2009. Enzyme function prediction with interpretable models. *Methods Mol Biol*, **541**, 373-420.

Teichmann, S.A., Park, J., Chothia, C. 1998. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci U S A*, **95**(25), 14658-63.

Tian, W.D., Skolnick, J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*, **333**(4), 863-882.

Todd, A.E., Orengo, C.A., Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, **307**(4), 1113-43.

Topakas, E., Christakopoulos, P., Faulds, C.B. 2005. Comparison of mesophilic and thermophilic feruloyl esterases: Characterization of their substrate specificity for methyl phenylalkanoates. *Journal of Biotechnology*, **115**(4), 355-366.

Turcot-Dubois, A.L., Le Moullac-Vaidye, B., Despiau, S., Roubinet, F., Bovin, N., Le Pendu, J., Blancher, A. 2007. Long-term evolution of the CAZY glycosyltransferase 6 (ABO) gene family from fishes to mammals--a birth-and-death evolution model. *Glycobiology*, **17**(5), 516-28.

Udatha, D.B., Mapelli, V., Panagiotou, G., Olsson, L. 2012a. Common and Distant Structural Characteristics of Feruloyl Esterase Families from *Aspergillus oryzae*. *PLoS One*, **7**(6), e39473.

Udatha, D.B., Sugaya, N., Olsson, L., Panagiotou, G. 2012b. How well do the substrates KISS the enzyme? Molecular docking program selection for feruloyl esterases. *Sci Rep*, **2**, 323.

Wackett, L.P. 2004. Evolution of enzymes for the metabolism of new chemical inputs into the environment. *J Biol Chem*, **279**(40), 41259-62.

Vafiadi, C., Topakas, E., Christakopoulos, P., Faulds, C.B. 2006. The feruloyl esterase system of Talaromyces stipitatus: Determining the hydrolytic and synthetic specificity of TsFaeC. *Journal of Biotechnology*, **125**(2), 210-221.

Wang, R., Lu, Y., Wang, S. 2003a. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*, **46**(12), 2287-303.

Wang, X.Y., Schroeder, D., Dobbs, D., Honavar, V. 2003b. Automated data-driven discovery of motif-based protein function classifiers. *Information Sciences*, **155**(1-2), 1-18.

Wang, Y.C., Wang, Y., Yang, Z.X., Deng, N.Y. 2011. Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context. *Bmc Systems Biology*, **5**.

Warren, G.L., Andrews, C.W., Capelli, A.M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., Tedesco, G., Wall, I.D., Woolven, J.M., Peishoff, C.E., Head, M.S. 2006. A critical assessment of docking programs and scoring functions. *J Med Chem*, **49**(20), 5912-31.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., Teichmann, S.A. 2004. Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol*, **336**(3), 809-23.

Voigt, C.A., Kauffman, S., Wang, Z.G. 2001. Rational evolutionary design: The theory of in vitro protein evolution. *Advances in Protein Chemistry, Vol 55*, **55**, 79-160.

Wolber, G., Seidel, T., Bendix, F., Langer, T. 2008. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today*, **13**(1-2), 23-9.

Wolfenden, R., Snider, M.J. 2001. The depth of chemical time and the power of enzymes as catalysts. *Acc Chem Res*, **34**(12), 938-45.

von Korff, M., Freyss, J., Sander, T. 2009. Comparison of ligand- and structure-based virtual screening on the DUD data set. *Journal of Chemical Information and Modeling*, **49**(2), 209-31.

Wong, D.W. 2006a. Feruloyl esterase: a key enzyme in biomass degradation. *Applied biochemistry and biotechnology*, **133**(2), 87-112.

Wong, D.W. 2006b. Feruloyl esterase: a key enzyme in biomass degradation. *Appl Biochem Biotechnol*, **133**(2), 87-112.

Xing, L., Hodgkin, E., Liu, Q., Sedlock, D. 2004. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput Aided Mol Des*, **18**(5), 333-44.

Yang, J.M., Chen, Y.F., Shen, T.W., Kristal, B.S., Hsu, D.F. 2005. Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling*, **45**(4), 1134-46.

Yang, K., Metcalf, W.W. 2004. A new activity for an old enzyme: Escherichia coli bacterial alkaline phosphatase is a phosphite-dependent hydrogenase. *Proc Natl Acad Sci U S A*, **101**(21), 7919-24.

Yu, H., Chokhawala, H.A., Huang, S., Chen, X. 2006. One-pot three-enzyme chemoenzymatic approach to the synthesis of sialosides containing natural and non-natural functionalities. *Nat Protoc*, **1**(5), 2485-92.

Zalatan, J.G., Herschlag, D. 2009. The far reaches of enzymology. *Nature Chemical Biology*, **5**(8), 516-20.

Zeragraf, M., Steuber, H., Koch, C., La Motta, C., Sartini, S., Sotriffer, C.A., Klebe, G. 2007. How reliable are current docking approaches for structure-based drug design? Lessons from aldose reductase. *Angewandte Chemie-International Edition*, **46**(19), 3575-3578.