Directed Evolution of Glutathione Transferases Guided by Multivariate Data Analysis

SANELA KURTOVIC
Dissertation presented at Uppsala University to be publicly examined in B7:101a, BMC, Box 576, Uppsala University, SE-75123 Uppsala, Friday, May 23, 2008 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Evolution of enzymes with novel functional properties has gained much attention in recent years. Naturally evolved enzymes are adapted to work in living cells under physiological conditions, circumstances that are not always available for industrial processes calling for novel and better catalysts. Furthermore, altering enzyme function also affords insight into how enzymes work and how natural evolution operates.

Previous investigations have explored catalytic properties in the directed evolution of mutant libraries with high sequence variation. Before this study was initiated, functional analysis of mutant libraries was, to a large extent, restricted to uni- or bivariate methods. Consequently, there was a need to apply multivariate data analysis (MVA) techniques in this context. Directed evolution was approached by DNA shuffling of glutathione transferases (GSTs) in this thesis. GSTs are multifarious enzymes that have detoxication of both exo- and endogenous compounds as their primary function. They catalyze the nucleophilic attack by the tripeptide glutathione on many different electrophilic substrates.

Several multivariate analysis tools, e.g. principal component (PC), hierarchical cluster, and K-means cluster analyses, were applied to large mutant libraries assayed with a battery of GST substrates. By this approach, evolvable units (quasi-species) fit for further evolution were identified. It was clear that different substrates undergoing different kinds of chemical transformation can group together in a multi-dimensional substrate-activity space, thus being responsible for a certain quasi-species cluster. Furthermore, the importance of the chemical environment, or substrate matrix, in enzyme evolution was recognized. Diverging substrate selectivity profiles among homologous enzymes acting on substrates performing the same kind of chemistry were identified by MVA. Important structure-function activity relationships with the prodrug azathioprine were elucidated by segment analysis of a shuffled GST mutant library. Together, these results illustrate important methods applied to molecular enzyme evolution.

Keywords: DNA shuffling, substrate selectivity, mutant library, glutathione transferase, multivariate data analysis, prodrug

Sanela Kurtovic, Department of Biochemistry, Box 576, Uppsala University, SE-75123 Uppsala, Sweden

© Sanela Kurtovic 2008

ISSN 1651-6214
ISBN 978-91-554-7194-1
urn:nbn:se:uu:diva-8718 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8718)
Nothing in life is to be feared. It is only to be understood.

Marie Curie

To my family
Papers included in this thesis

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals:


Reprints of the articles were made with permission from the publishers.
Contents

General background ........................................................................................................ 11

Introduction .................................................................................................................. 11

Enzyme evolution ........................................................................................................ 13
  Natural evolution ......................................................................................................... 13
  Enzyme engineering .................................................................................................. 18

Multivariate data analysis .......................................................................................... 21
  Pretreatment of data .................................................................................................. 22
  Principal components ............................................................................................... 23
  Biplots ....................................................................................................................... 24
  Cluster analysis .......................................................................................................... 25
  Canonical correlation analysis .................................................................................. 26

Glutathione transferases ............................................................................................. 26
  Glutathione ............................................................................................................... 27
  Function of GSTs ....................................................................................................... 27
  Classification and nomenclature .............................................................................. 30
  Structure and mechanism ......................................................................................... 30
  Evolution .................................................................................................................... 32

Present Investigation .................................................................................................. 33

Objectives ..................................................................................................................... 33

Discovery of novel quasi-species in enzyme evolution (Papers I and II) ................................................................. 33
  The functional profile of GSTs acting on electrophilic substrates is not normally distributed ............................................. 34
  What is a quasi-species? ........................................................................................... 36
  Application of the concept of quasi-species to a GST mutant library .......................... 37
  Coping with the problem of expression levels in screening ......................................................................................... 38
  Determination of the number of clusters in a GST mutant library .............................................. 40
  Structure-function activity relationships in a Mu class GST mutant library ................................................................. 43

Elucidation of important structure-function activity relationships of a shuffled Alpha class GST mutant library screened with azathioprine (Paper III) ................................................................................................. 43
  Azathioprine – a widely used immunosuppressive prodrug ........................................ 44
Shuffling of Alpha class GSTs and screening of the library with azathioprine .................................................................44
Structure-function activity relationships illuminated by segment analysis .................................................................................45

Synthesis and characterization of a novel substrate for GSTs that mimics azathioprine (Paper IV) ..................................................48

Synthesis and characterization of NPTI ..................................................48
NPTI – a good model substrate for azathioprine ......................................49

Catalytic capacities and selectivity profiles in a GST mutant library screened with haloalkanes (Papers V and VI) .........................50

Haloalkanes – examples of both detoxication and bioactivation reactions with GSTs .................................................................51
Development of a colorimetric endpoint assay useful in screening large mutant libraries ............................................................51
Introduction of new descriptors useful in enzyme engineering .............52
Screening of a mutant library with haloalkane substrates ..................54

Importance of the substrate matrix in recognition of novel quasi-species (Paper VII) .................................................................57
PC analysis of screening data reveals two quasi-species ........................57
Measurements with additional substrates expose yet another quasi-species ..................................................................................60

Birth of a novel glutathione transferase (Paper VIII) ............................61

Summary ..............................................................................................66

Summary in Swedish ..................................................................................69

Bakgrund ..............................................................................................69
Inledning ............................................................................................69
Enzymer – uppbyggnad .......................................................................69
Enzymevolution ..................................................................................69
Glutationtransferaser .........................................................................70
Multivariata dataanalyser ....................................................................70

Min forskning ........................................................................................71
Varför denna avhandling? ...................................................................71
Utnyttjande av multivariat dataanalys inom enzymevolution ..........71
Proläkemedlet azatioprin – GST struktur-funktions samband ............71
Födelse av ett nytt GST ........................................................................72

Acknowledgments ..................................................................................73

References ............................................................................................75
Abbreviations

PC  Principal component
UV  Unit variance scaling
UL  Unit length scaling
GST  Glutathione transferase
CDNB 1-chloro-2,4-dinitrobenzene
pNPA para-nitrophenylacetate
AD  Δ^5^-androstene-3,17-dione
Nonenal trans-2-nonenal
NPTI 1-methyl-4-nitro-5-(4-nitrophenylthio)-1H-imidazole
MCB Monochlorobimane
NPG 3-(4-nitrophenyl)-glycidol
tPBO trans-4-phenyl-3-buten-2-one
EPNP epoxy-3-(4-nitrophenoxo)-propane
NCH 1-nitro-1-cyclohexene
PEITC phenethyl isothiocyanate
NPB 4-nitrophrenethyl bromide
tSBO trans-stilbene oxide
Cyano-DMNG 2-cyano-1,3-dimethyl-1-nitrosoguanidine
NBD-Cl 4-chloro-7-nitro-1,2,3-benzoxadiazole
Diiodoethane 1,2-diiodoethane
Diiodobutane 1,4-diiodobutane
Diiodohexane 1,6-diiodohexane
Diiodooctane 1,8-diiodooctane
Iodohexane 1-iodohexane
3-IPB 3-iodopropylbenzene
CuOOH Cumene hydroperoxide
EA Ethacrynic acid
human GST M1-1 hM1-1
human GST M2-2 hM2-2
human GST A1-1 hA1-1
human GST A2-2 hA2-2
human GST A3-3 hA3-3
bovine GST A1-1 bA1-1
rat GST A2-2 rA2-2
rat GST A3-3 rA3-3
General background

Introduction

Proteins are large biomolecules functioning as building stones of life on Earth as we know it. They are involved in a variety of different tasks in living cells, ranging from binding of small molecules and triggering cellular responses (receptors) to being part of a defense system against toxic molecular species. Every cell performs a large number of chemical reactions every minute that life is dependent on. However, most of these reactions proceed too slowly in aqueous solution, so catalysts are needed. Nature has designed biological catalysts that are proteins called enzymes.

Enzymes are fascinating catalysts that are able to perform impressive rate enhancements in living cells. Enhancements of $1.4 \times 10^{17}$ compared to the corresponding reaction in an aqueous solution have been reported for an enzyme called orotidine 5’-phosphate decarboxylase (Radzicka and Wolfenden, 1995). The question of how enzymes are capable of achieving such high catalytic rates has interested scientists for many decades. In fact, even ancient philosophers like Aristotle were interested in the process of making cheese, which results from milk curdling via action of enzymes. One of the first systematic experiments on enzymes was performed by the famous French scientist Réaumur (1683-1757) (Copeland, 2000), who investigated how predatory birds digested meat. In 1946, Linus Pauling proposed that the catalysis of enzymes is achieved by lowering the activation energy of the reaction by binding the transition state better than the reactants of the reaction (Pauling, 1946).

However, not only the question concerning how enzymes work is interesting to address, but also how it is possible to evolve enzymes with novel properties. Today, enzymes are used in food and beverage manufacturing, stereospecific chemical synthesis and laundry detergents, but also as targets for many drugs (Powell et al., 2001; Yuan et al., 2005). One of the most exciting applications of enzymes at the moment is research concerning gene therapy, whereby recombinant DNA is introduced into the cell to effect expression of recombinant protein (Budak-Alpdogan et al., 2005). The idea is to use certain enzymes for treatment of particular diseases such as cancer. Novel and better biological catalysts can find their applications within these widespread areas. Directed evolution of novel catalysts with altered catalytic
functions and navigation in a multidimensional substrate-activity space is the main focus of this thesis.

How are enzymes with altered properties evolved in the laboratory? One method involves redesign of already existing structures, which is a powerful approach to evolve new catalysts (Anantharaman et al., 2003; Pettersson et al., 2002). If the structure is known, it is possible to pinpoint important residues that are directly involved in catalysis. By altering these residues, new functionality of the enzyme can be evolved. However, this kind of rational design of enzymes based on structural data has limitations, since even highly similar protein sequences can sometimes display totally different functions (White, 2006). Therefore, the more empirical approach of directed evolution has gained much attention as an alternative method of creating novel enzyme catalysts (Svendsen, 2004). There are many different ways of creating large mutant libraries with high sequence variation, one of them being DNA shuffling (Stemmer, 1994a). Many directed evolution experiments that have been conducted have resulted in changing of different enzyme properties. Examples include evolution of higher catalytic efficiency and improved thermal stability (Ness et al., 1999), as well as changes in the substrate selectivity profile of an enzyme (Hansson et al., 1999a). Furthermore, different approaches concerning analysis of structural data have been introduced (Orengo and Thornton, 2005). However, methods for analyzing the functional profiles of large mutant libraries created by directed evolution have not gained comparable attention. The optimal goal of creating enzymes with tailor-made function is dependent on being able to explore distributions of members of a mutant library in functional substrate activity space. In contrast to the structural space of proteins where a one-dimensional entity in primary structure becomes a three-dimensional unit in the folded structure, functional properties can be defined by many dimensions, depending on the parameters investigated. Because of this complexity, functional analysis becomes a multidimensional problem which requires multivariate data analysis as a tool. Before the present study was initiated, very little thorough analysis of large mutant libraries was done using multivariate data analysis methods. An early study addressed the question of how many different classes of GSTs there is by monitoring the functional properties of the enzyme variants in terms of specific catalytic activities with nine substrates and IC_{50} values for 11 inhibitors (Mannervik et al., 1985). By this approach three different classes were identified. Joern and coworkers used probe hybridization in a microarray format to analyze chimeric DNA libraries created by DNA shuffling (Joern et al., 2002). Raillard et al. revealed novel enzyme activities and functional plasticity upon recombining highly homologous triazine hydrolases when analyzing the substrate specificities of the resulting enzymes (Raillard et al., 2001). However, a thorough multivariate analysis of relationships between mutants and substrates that might guide the subse-
quent evolution of functional properties was lacking. Therefore, several questions were raised that make a foundation for this thesis:

1. Is it possible to identify groupings in a library of mutants generated by DNA shuffling where activities have been monitored with a battery of electrophilic substrates?
2. How will substrates themselves group? Based on the reaction mechanism or something else?
3. Is it possible to find evolving units in molecular evolution by multivariate analysis?
4. Can we identify any confidence intervals in a multidimensional space that will define these evolvable units?
5. Is the substrate matrix important when grouping enzymes into different evolvable units? Are all substrates equally important or can some be excluded from the analysis?

However, during the work with this thesis, several interesting studies using multivariate methods have been reported. Enzyme activity fingerprints with substrate cocktails were identified by hierarchical clustering methods (Goddard and Reymond, 2004; Grognux and Reymond, 2004). Also, Taly and coworkers have recently used hierarchical clustering as well as principal component analysis to identify sequence-activity relationships within a three-dimensional functional space of a mutant library (Taly et al., 2007). Furthermore, a synthetic library of cytochrome P450s made by SCHEMA recombination was tested for activity on 11 different substrates and subjected to K-means clustering, which allowed the enzymes to be classified into five distinct groups (Landwehr et al., 2007).

In summary, central to the articles presented in this thesis is the application of many different multivariate techniques to issues concerning directed as well as rational evolution of GSTs. By this approach, insights into enzyme evolution have been gained. For example, novel evolvable units (quasi-species) in molecular evolution as well as diverging substrate selectivity profiles among shuffled GST variants have been recognized.

**Enzyme evolution**

**Natural evolution**

Before going into details about laboratory-evolved enzymes, it is important to consider how natural enzymes have evolved. Nature has had millions of years to fine tune the remarkable specificity and capacity of enzymes. Many enzymes are also promiscuous, meaning that they display another function than that originally expected, sometimes even several functions. It is also clear from research in our group and also by others that certain enzymes
have specific selectivity profiles that define the working repertoire of the enzyme. Put together, all of these characteristics are quite fascinating and raise many questions about how natural enzyme evolution operates.

In 1859, Charles Darwin published *On the Origin of Species* where he described how new species arose and discussed the origin of life itself. Darwin proposed that natural selection is the basic mechanism responsible for the origin of new species (Darwin, 1859). According to Darwin, the processes of evolution, occur gradually maintaining organism fitness intact all the time. A natural consequence of this idea is that already existing genes have been “tinkered with” or modified in order to generate novel structures and functions in molecular evolution (Jacob, 1977).

**Gene duplication**

Gene duplication is thought to be the primary source of material for novel gene functions (Lynch and Conery, 2000; Muller, 1936; Ohno, 1970; Sidow, 1996). Gene duplication events can involve duplication of genes, chromosomal segments or even entire genomes. It is, however, not clear how often these events occur. It is also not clear how evolution navigates from the initial state of a duplicated gene to a state where both copies are maintained by natural selection. Several models based on both theoretical and experimental results have been proposed. The existing theory on gene duplication illustrates three possible outcomes for a duplicated gene: (i) nonfunctionalization, where a copy is silenced by nonadvantageous mutations; (ii) neofunctionalization, where one copy shows a novel function and is preserved by natural selection, whereas the other copy keeps its original function; or (iii) subfunctionalization, where both copies are partially compromised by deleterious mutations (Force *et al.*, 1999; Lynch and Force, 2000; Ohno, 1970; Sidow, 1996; Walsh, 1985). Almost all theoretical models predict that nonfunctionalization is the most common fate of one copy of a duplicated gene (Lynch and Conery, 2000). The early theory on origins of new genes has been developed in 1970 by Ohno who stated that duplication will create a redundant gene copy being free from selection for its original function. Eventually, the newly formed gene copy will accumulate earlier forbidden mutations, thereby hopefully acquiring a novel function (Ohno, 1970). Later Kimura and Ohta deduced five important principles concerning molecular evolution, “gene duplication must always precede the emergence of a gene having a new function” being one of them. This model is known as mutation during nonfunctionality (MDN) (Kimura and Ota, 1974). However, the MDN theory of duplicate genes possesses some problems. The newly duplicated gene will be exposed to loss by drift and common inactivating mutations, *e.g.* deletions, frameshifts, nonsense mutations, implying that this extra copy of the gene must persist long enough in a functionally intact state to be able to obtain a new function by advantageous mutations. A recent article
Bergthorsson et al., 2007) deals with this problem and introduces a new proposal for how novel gene functions arise. The authors state that new genes arise under continuous positive selection rather than a gene being duplicated before a new function has been acquired and maintained by selection for its original function (the MDN model). The new proposal involves innovation, amplification, and divergence (IAD) and suggests that new genetic functions arise from a minor side function of the parental gene that eventually becomes valuable before the parent allele is duplicated. This idea has its origin in the experimental evidence that many genes (especially enzymes) have minor secondary activities that are addressed in the next section.

**Enzyme promiscuity**

In recent years enzymes that possess latent function different from the original one have gained attention. These enzymes are sometimes called promiscuous or moonlighting. In an early review, Jensen proposed that starting points of evolutionary scenarios are broad substrate specificities, or promiscuous functions (Jensen, 1976). Thus, very old enzymes show broader substrate specificities compared to modern enzymes that have evolved through gene duplication events and divergence (discussed below). This hypothesis provided primordial enzymes with the advantage of being few molecules performing the large amount of functions that were needed in order to maintain ancestral organisms.
A protein with low promiscuous function having a certain original function (black circle) will gradually evolve. At the end of the evolutionary process, during many generations of mutation and selection, a new function is adopted at the expense of the original one (gray circle). The mechanism of this process may vary. The conversion of one specialist to a specialist with novel function may trade off linearly or follow either a concave or convex route. Directed evolution in the laboratory suggests that the convex route (weak negative trade-off) is the more likely one. By this approach large increases in the promiscuous function imply only small changes in the original function. The intermediates created (generalists, white circle) can proceed prior to gene duplication events. On the contrary, the concave route is dependent on gene duplication taking place before acquisition of new function because acquisition of even very small levels of new function will imply large loss in the original one.

There are several examples of enzymes that have been found to possess more than one function\(^1\), e.g., chymotrypsin, that catalyzes hydrolysis of many types of different compounds, bovine carbonic anhydrase II that, in addition to its carbon esterase activity, also displays phosphotriesterase activity (O’Brien and Herschlag, 1999). What are the underlying mechanisms of new function acquisition? Tawfik and coworkers (Khersonsky et al., 2006) propose a theory of an evolutionary pathway that originates from a specialist and goes via a generalist in turn to a new specialist (Figure 1) that several directed evolution experiments support (Matsumura and Ellington, 2001; O’Loughlin et al., 2006; Rothman and Kirsch, 2003). Their conclusion is that “tinkering” (Jacob, 1977) seems to be a very important process that nature exploits.

**Enzyme families and superfamilies**

Processes of enzyme promiscuity and gene duplication have led to the creation of enzyme families (a group of homologous proteins sharing the same

---

\(^1\) When discussing promiscuity, the definition adopted here is applied to enzymes being able to catalyze multiple chemical transformations that are not regarded as same types of reactions.
function) and superfamilies (proteins with less than 40% sequence identity but with the same overall fold displaying more divergent function than enzyme families) as evidenced by structures of >40,000 proteins (January 2008) in the protein data bank as well as sequences of hundreds of thousands. The number of unique protein folds is estimated to be no more than a few thousand which is much less than the number of proteins encoded by the human genome (Gerlt and Babbitt, 2001). In order to be able to supply the many functions represented throughout the entire protein universe, evolution must adopt a mechanism of convergence (independent evolution of structures not related to a common ancestor that have the same function) or divergence (evolution from a common ancestor) (Gerlt and Babbitt, 2001). In the SCOP and CATH databases, between 1200 and 1400 domain families are classified wherein around 70% of the domain sequences in the genomes can be assigned to these families (Orengo and Thornton, 2005). Of particular interest for prediction of protein functions and protein engineering is the understanding of the existence of mechanistically diverse superfamilies. These superfamilies are composed of homologous members that catalyze different overall reactions but share a common aspect of catalysis. There are several known examples of mechanistically diverse superfamilies, of which the enolase superfamily is one of the best-characterized. This family is known to catalyze at least 14 different reactions involving a common partial reaction where an enolate anion intermediate is formed (Glasner et al., 2006). Other examples include crotonase, amidohydrolase, vicinal oxygen chelate and nudix superfamilies. The explanation for functional diversification in enzyme superfamilies can be that new protein functions actually evolve through promiscuous intermediates discussed above (Khersonsky et al., 2006).

However, it is important to note that, in eukaryotes, alternative splicing (Kopelman et al., 2005; Talavera et al., 2007) and posttranslational modifications can have an impact on protein function (Bork and Orengo, 2004).

The RNA world
Many issues regarding enzyme evolution have been addressed so far in this thesis, but one important question still remains: how did it all start? One theory that has been adopted is that of an ancient “RNA world”. In this theory, it is proposed that ribozymes, rather than protein enzymes, were responsible for the catalysis of the reactions needed for the maintenance as well as the proliferation of life (Gilbert, 1986). The validity of the hypothesis relies on the proposed existence of two types of ribozymes, i.e., an RNA replicase that can function both as a repository of genetic material and as an RNA polymerase for replication of its own sequence. Some support for the theory can be found today in self-splicing introns as well as the fact that modern ribosomal RNA is the catalytic core of the ribosome (Hager et al., 1996). Szostak and coworkers have studied the hypothesis of an ancient RNA world.
by *in vitro* selection of large libraries of nucleic acid sequences for rare functionalities (Ellington and Szostak, 1990; Lorsch and Szostak, 1996). They even discuss the synthesis of life in the laboratory by designing a primitive protocell that is composed of an RNA replicase and a vesicle (a simple lipid membrane). The possibility then emerges of studying how a simple protocell could develop into a living cell, thus providing important clues about early evolution of life (Szostak *et al*., 2001).

Enzyme engineering

The need for new and better catalysts is gaining momentum in many different fields of applications. In organic synthesis, novel catalysts are wanted that are able to perform enantioselective catalysis. However, enzymes that can withstand high temperatures, organic solvents and other different physical and chemical conditions that often prevail organic synthesis need to be engineered (Powell *et al*., 2001). Also in medicine, new and better enzymes that are responsible for metabolizing certain anticancer prodrugs or have protective properties for the fast proliferating cells are needed in a growing field of gene therapy (Budak-Alpdogan *et al*., 2005; Encell *et al*., 1999; Sorrentino, 2002; Wadhwa *et al*., 2002).

There are three main ways to design enzymes with new properties: rational design, "irrational" design and *de novo* design. Among these techniques the *de novo* enzyme design is the most challenging, since it introduces not only a problem of catalysis but also a problem of protein folding into the experiment. One of the most impressive examples of *de novo* design is a study performed by Seelig and Szostak using a messenger RNA display that yielded new enzymatic activities (Seelig and Szostak, 2007). Limitations of different strategies in enzyme engineering are constantly being overcome and new ways of evolving enzymes with interesting properties are developing.

**Rational design**

In rational design, an enzyme with known three dimensional structure can be subjected to site-directed mutagenesis\(^2\), simultaneous mutations by cassette mutagenesis (Reidhaar-Olson and Sauer, 1988) and exchange of whole catalytic domains. By this method, important catalytic residues in the active site can be elucidated. In order to make a rational design experiment successful, knowledge about the structure and mechanism of the enzyme studied is often required. However, the structure is not always available and the catalytic mechanism is not always elucidated, which severely hinder this approach.

Rational design experiments have successfully been conducted where substrate specificity of human glutathione transferase (GST) A1-1 was re-

---

\(^2\) Introduction of one or several mutations at a specific site in a studied gene
designed to mimic the substrate specificity of human GST A4-4 having high catalytic activity with alkenals (Nilsson et al., 2000). Also human GST A2-2 was subjected to rational active-site directed mutagenesis in order to incorporate high steroid isomerase activity not present in the wild-type enzyme (Pettersson et al., 2002). Norrgård et al have altered a previously recognized hypervariable (Ivarsson et al., 2003) residue in the active site of human GST M2-2 to all other 19 amino acids which gave altered substrate selectivity profiles with diverse electrophilic substrates (Norrgård et al., 2006). Completely novel enzyme active sites have been introduced into already existing scaffolds using rational computational design (Bolon and Mayo, 2001; Röthlisberger et al., 2008).

**Directed evolution**

In order to circumvent problems encountered with rational design approaches, directed evolution was introduced as an alternative method for protein engineering in the early 1990s (Brannigan and Wilkinson, 2002). In contrast to rational design, directed evolution experiments do not require prior knowledge about catalytic mechanism or structural information. The method of directed evolution is dependent on screening or selection of a large number of mutants created by random mutagenesis and/or gene recombination (Powell et al., 2001; Yuan et al., 2005). By this approach, many properties such as activity, stability, selectivity, specificity and affinity have been successfully engineered (Johannes and Zhao, 2006). As it mimics natural evolution (Babushok et al., 2007), the method of directed evolution is also a powerful tool for the study of structure-function relationships in proteins (Kurtovic et al., 2008b; Meyer et al., 2006; Yano et al., 1998). Directed evolution is not restricted to engineering novel or improved protein properties, but can also be employed to evolve operons, pathways, viruses, and even whole organisms (Schmidt-Dannert, 2001).

**DNA Shuffling**

In 1994, a new technique for *in vitro* evolution, DNA shuffling, was introduced (Stemmer, 1994a; Stemmer, 1994b). By this method, new combinatorial DNA variants can be created from several templates (Crameri et al., 1998). Using this approach Stemmer was able to increase β-lactamase resistance to the antibiotic cefotaxime 32 000-fold, illustrating the power of the combinatorial method (Stemmer, 1994b).
The first step in DNA shuffling is random DNA fragmentation by digestion with DNase I. A primerless PCR reaction follows. In this way, small pieces of DNA will anneal to each other in a combinatorial manner, giving rise to new variants that are composed of the starting material. The next step in a DNA shuffling experiment involves primers in order to select only the DNA having the same length as the starting template. Additionally, spurious point mutations are introduced in the PCR reaction to contribute to the diversity. The starting material can be one cDNA, as exemplified by β-lactamase (Stemmer, 1994b), as well as cDNA encoding several homologous proteins (Crameri et al., 1998; Hansson et al., 1999a; Kurtovic et al., 2008b). However, DNA shuffling is not restricted to proteins only but can be applied to a complete plasmid (Stemmer, 1994a) or an entire operon (Crameri et al., 1997). In DNA shuffling, crossovers occur at random but homologous positions (Moore et al., 2001), which is a limitation of the method. However, methods are being developed for shuffling of cDNAs encoding proteins displaying low homology (see next section). In humans, exons occupy only 1% and introns 24% of the genome, implying that most of the natural crossovers occur between exons and not within. The natural process of new recombinations of exons is called exon shuffling (de Souza et al., 1996) and can be mimicked in vitro (Kolkman and Stemmer, 2001). DNA shuffling together with structural analysis has found its use in the prediction of the emergence of antibiotic resistance (Orencia et al., 2001). Whole genome shuffling has also been demonstrated for Streptomyces fradiae (Zhang et al., 2002). Altered substrate selectivity profiles of mutant libraries have been recognized utilizing DNA shuffling (Emrén et al., 2006; Kurtovic et al., 2008c; Raillard et al., 2001; Taly et al., 2007). Catalytic activity (Ness et al., 1999), thermal

Figure 2. Scheme of DNA shuffling. For details see the text below.
stability (Minagawa et al., 2007), acid tolerance\(^3\) (Patnaik et al., 2002) and enantioselectivity (Rui et al., 2005) have been improved compared to wild-type proteins by DNA shuffling.

**Recombining proteins with low homology**

In order to expand the sequence space that can be explored by directed evolution, several techniques for recombining proteins with low homology have been developed. Arnold and coworkers have developed a method where cDNAs encoding low-homology proteins can be shuffled – Sequence-independent site-directed chimeragenesis (SCHEMA) (Hiraga and Arnold, 2003). Lutz and coworkers have developed a methodology named SCRATCHY that makes combinatorial engineering independent of sequence identity possible (Lutz et al., 2001). This technique is a combination between ITCHY (incremental truncation for the creation of hybrid enzymes) (Ostermeier et al., 1999) and DNA shuffling. The advantage of SCRATCHY compared to many other techniques employed when recombining proteins with low homology (such as ITCHY) is that it is able to generate several crossovers, whereas the others only generate one.

**Combination of rational design and directed evolution**

In directed evolution, the number of sequence variations of a protein that can be generated may be too large to be screened in the laboratory. This is one of the limitations of the directed evolution approach. Lately, a combination between rational or computational design and directed evolution has found increasing interest in order to circumvent the limitations of the directed evolution approach (Zhao, 2007). By utilizing the combination of directed evolution and rational design, the activity of β–lactamase has been introduced into the αβ/βα metallohydrolase scaffold of glyoxalase II (Park et al., 2006). In this way a protein with novel function has been evolved, a task that is still a challenge for directed evolution studies (Brannigan and Wilkinson, 2002).

**Multivariate data analysis**

Researchers in all disciplines have a key feature in common: collection of data sets that need to be evaluated and understood in order to draw some conclusion. In many cases, the data sets are large and contain many variables. Many data sets are thus multivariate in nature and need special multivariate statistics for analysis. In contrast to simple statistics where each variable is investigated separately, multivariate statistics allow many variables to be investigated simultaneously. Simultaneous analysis of all the variables included in the study is important, since they might not be separate but rather

\(^3\) A whole species, *Lactobacillus*, has been improved for acid tolerance by genome shuffling
dependent on each other. The multivariate approach will thus make it possible to fully understand the relationships among variables. Multivariate techniques have found their use in widespread areas such as drug discovery, statistical process control, experimental design and in the design of enzymes with novel functions.

In this thesis, multivariate techniques constitute a very important “tool kit” for the study of enzyme evolution. Therefore, the background of different methods used will be presented in this section.

Pretreatment of data

Before any multivariate analysis is performed it is important to investigate the data in a univariate fashion with the help of e.g., scatter plots for each variable, histograms, and normal probability plots. An effort should be made to get “acquainted” with the data set of interest. By examining each variable at a time, drifts in experimental data can be exposed that might disturb the analysis later on. Whether a particular data set is normally distributed can also play an important role in detecting outliers and assigning the data some level of statistical confidence. In addition, observations may differ in their responses to different variables so that some display very high absolute values, whereas others are very low. If this is the case, some variables will dominate the analysis which might not be desirable. In this case mathematical operations of scaling can be applied. However, if one has prior knowledge about a higher importance of certain variables, a scaling procedure is not recommended.

There are several techniques that can be used when there is a need for scaling the data, where both unit variance (UV) and unit length (UL) scaling are very common and have both used extensively in this thesis. In UV scaling long variables are shrunken and short variables are stretched by dividing each response with the standard deviation of all the observations, \( \frac{x_{np}}{s_p} \) where \( x_{np} \) is the value of the \( p \)th variable for the \( n \)th observation and \( s_p \) represents the standard deviation of variable \( p \). In addition, it is also common to use mean-centering in combination with UV scaling where simply the average value for each variable is calculated and subtracted from the data. In UL scaling the same idea of making long variables shorter and short variables longer is adopted. However, different mathematical operations are performed where a unit circle is used (Kurtovic et al., 2008c) (Figure 3).
Figure 3. Schematic explanation of unit-length scaling. Variable 1 displays high catalytic activity on both observation 1 and 2 whereas Variable 2 displays very low catalytic efficiency with observation 1 as well as observation 2. Variable 1 is thereby suppressed down to the unit circle whereas variable two is stretched out. The new values for each observation can now be obtained (open and filled circles). (Figure from Paper VI)

Principal components

Principal component (PC) analysis was originally described by Pearson (Pearson, 1901) and Hotelling (Hotelling, 1933). Algebraically, a PC analysis involves explaining the variance-covariance structure of a set of variables by using linear combinations of the variables. Therefore, the geometrical interpretation is that the linear combinations will represent a new coordinate system that is obtained by rotating the original system (Figure 4). The new axes obtained in this way will represent the directions with maximum variability. The calculation of PCs is not dependent on the variables analyzed being normally distributed. The mathematical derivation of PCs is shown below.

PCs depend only on the covariance matrix $\Sigma$ of $X_1, X_2, \ldots, X_p$ random variables

Let the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\Sigma$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Consider the linear combinations

$$Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\vdots \quad \vdots$$
$$Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

Then, we obtain
The first PC will thus be the linear combination with maximum variance, that is, it maximizes \( \text{Var}(Y_i) = \mathbf{a}_i' \Sigma \mathbf{a}_i \).

Principal components can be visualized in two different graphs denoted score plot and loading plot. Two PCs together will define a two-dimensional plane in a \( p \)-dimensional variable space. A score plot is created by projecting all the observations onto this lower-dimension plane (Figure 4). Therefore, a score plot gives information about the new observations that summarizes the behavior of the original ones. A loading plot, on the other hand, reveals how the PC model plane is inserted into the variable space. The loadings can be used in order to interpret the meaning of the scores (Eriksson et al., 2001; Everitt and Dunn, 2001; Johnson and Wichern, 2007; Krzanowski, 2000).

How many PCs should be used? There is no definitive answer to this question. There are several techniques that can be adopted where a scree plot (Cattell, 1965) is one of the most commonly used. It is based on the examination of \( \lambda_i \) against \( i \), where the number of components selected is the point at which the remaining eigenvalues are relatively small and of about equal size.

Biplots
So far, plots where either the observations or variables are plotted in a single plot have been considered. However, there is a possibility to plot both vari-
ables and observations simultaneously in the same plot, denoted a biplot (Gabriel, 1971). This can be performed by representing a matrix by two sets of vectors, a vector for each row and another set of vectors for each column. The vectors are chosen so that the elements of the data matrix are the inner products of the vectors. If a matrix has a rank higher than two it cannot be represented exactly by a biplot. However, if a matrix of higher rank can be approximated by a matrix of rank 2, the biplot of this new matrix can be investigated. In this case, a biplot will give approximate information about the original matrix itself. In order to reduce any rectangular matrix to a space of lower dimensions, an algorithm of singular value decomposition (Golub and Reinsch, 1970) can be used.

Cluster analysis
Cluster analysis relates to grouping a collection of objects to “clusters” or groups where those individuals belonging to the same group are more similar to each other than are the individuals from different groups. There are two different ways to assign observations to different clusters, the first one being a top down approach where a number of clusters has been chosen prior to analysis, and the other a bottom up approach where there is no requirement for selection of a number of clusters. Both of these approaches have been used in the thesis and examples of different techniques belonging to each approach are discussed below (Eriksson et al., 2001; Everitt and Dunn, 2001; Johnson and Wichern, 2007; Krzanowski, 2000).

K-means – an iterative descent clustering method
A K-means cluster analysis starts with a guess about the number of cluster centers. The algorithm then alternates two steps until convergence is reached: in Euclidean distance the closest cluster center for each data point is defined whereafter each center is replaced by the average of all data points included that are closest to it. The algorithm is intended for situations in which all variables are quantitative (Hastie et al., 2001; Johnson and Wichern, 2007).

Usually when a K-means cluster analysis is performed the problem addressed is concerned with the extent to which the observations fall into natural distinct groupings. The number of clusters in the data are unknown. However, a good starting point is a PC analysis where it is easy to see trends among observations and variables, thereby providing a clue about how many clusters the data set is composed of.

Hierarchical clustering methods
In contrast to K-means clustering algorithms, hierarchical clustering does not require a prior decision regarding the number of clusters. However, these methods require a specification of the measure of the dissimilarity between
groups of measures. These algorithms make hierarchical representations where clusters of lower levels are merged to create the next cluster level. At the lowest level of hierarchy, each cluster will be composed of only one individual, whereas at the highest level there will only be one cluster that is composed of all the individuals in the analysis. There are two different kinds of strategies adopted in order to create a hierarchical clustering tree, agglomerative (bottom-up) and divisive (top-down), where agglomerative strategies are being most frequently used. The agglomerative approach starts at the lowest hierarchy level and at each consecutive level merges the individuals from the lower. The chosen pair for merging is the one having the smallest intergroup dissimilarity, or the highest similarity. With this technique it is up to the user to decide at which level there is a situation where individuals within each of its groups are significantly more similar to each other than to individuals belonging to clusters at the lower level. One obvious measure to consider in this regard is the level of standard deviation among the observations beneath which only experimental error will be displayed. A graphical display of hierarchical clustering is called a dendrogram (Hastie et al., 2001; Johnson and Wichern, 2007).

**Canonical correlation analysis**

If the number of clusters in a data set is known, canonical variate analysis can be performed in order to highlight differences between groups. Normal probability assumption of variables included is not needed for this analysis. However, if the variables are normally distributed, a 95% confidence region can be plotted around each cluster that will show how well the clusters are separated. In contrast to PCs, canonical variates are not obtained by a rotation where the variates are orthogonal to each other. The transformation of canonical variates is obtained by non-orthogonal rotation that best represents the Mahalanobis distance between the predefined groups (Everitt and Dunn, 2001; Krzanowski, 2000).

**Glutathione transferases**

In this thesis, directed and rational evolution has been applied to an enzyme family of glutathione transferases (GSTs). GSTs are multifunctional enzymes where the major role is believed to be protection of cells against oxidative damage. These enzymes are widely distributed in nature and can be found in animals, plants and bacteria (Hayes and Pulford, 1995). They were discovered in 1961 in rat liver extracts where they were found to catalyze conjugation of the tripeptide glutathione with arylhalides (Booth et al., 1961; Combes and Stakelum, 1961). GSTs are well suited for protein engineering experiments because they display activity with many different electrophilic
compounds. The activities are moderate in some cases, which makes them highly evolvable. Members of the GST enzyme family also display different substrate selectivity profiles, thus making them even more interesting in a directed evolution context.

**Glutathione**

The tripeptide glutathione, γ-L-glutamyl-L-cysteinyl-glycine (*Figure 5*), is present at the millimolar level in almost all aerobic cells, making it the most abundant low-molecular-mass thiol in mammals (Anderson, 1998; Reed, 1990). Glutathione is involved in many different functions in the cell, the most central ones including defense against oxidizing and electrophilic chemical species and a role as a reducing cofactor in the biosynthesis of deoxyribonucleotides. In addition, glutathione is also involved in redox regulation of molecular processes and acts as a carrier of certain signal molecules such as nitric oxide (Josephy and Mannervik, 2006).

*Figure 5.* Structure (left) and crystal structure (right) of glutathione (γ-L-glutamyl-L-cysteinyl-glycine) (PDB-code 1xw5).

The chemical structure of glutathione exhibits two negative charges and one positive at physiological pH, making it highly soluble in aqueous solution. One of the most important properties of the tripeptide is its ability to act as a good nucleophile in reactions with electrophilic substances. This is due to its functional sulfhydryl (-SH) group. However, the pKa of the sulfhydryl group is quite high, 9.2, implying that only a very small portion will be in the thiolate anion form at neutral pH (Josephy and Mannervik, 2006). GSTs are able to lower this pKa thereby enhancing the catalytic rate of glutathione conjugation with electrophiles, as will be discussed later.

**Function of GSTs**

GSTs are a part of the phase II detoxication system having a role of protecting cells against both endo- as well as exogenous compounds (Armstrong, 1997; Hayes and Pulford, 1995). As a part of the phase II system GSTs have been found to catalyze conjugation of glutathione with different electrophilic compounds that have been activated by phase I enzymes (*e.g.*, cytochromes
P450). This will make the conjugates more water soluble and facilitate excretion from the body constituting a part of phase III xenobiotic metabolism (Ishikawa, 1992).

GSTs are able to catalyze conjugation of a broad range of electrophilic compounds with glutathione. The conjugation is conducted by addition or substitution mechanisms. Some of the biologically relevant substrates include *ortho*-quinones (Baez *et al*., 1997), $\alpha,\beta$-unsaturated carbonyl compounds (Habig *et al*., 1974), as well as isothiocyanates (Kolm *et al*., 1995). GSTs also catalyze the glutathione-dependent double-bond isomerization of $\Delta^5$-3-ketosteroids, such as $\Delta^5$-androstene-3,17-dione, important in the pathway leading from cholesterol to steroid hormones (Johansson and Mannervik, 2001; Raffalli-Mathieu and Mannervik, 2005). GSTs are also able to catalyze epoxide ring openings displaying regio- and enantioselectivities (Ivarsson and Mannervik, 2005). Substrates used in this thesis include both substitution and addition reactions as well as compounds that have biological relevance. Reaction mechanisms of some of those substrates are presented in *Figure 6*. }
Figure 6. Examples of glutathione conjugations with different electrophilic substrates. (A) 1-chloro-2,4-dinitrobenzene is a common GST substrate and represents an aromatic substitution reaction with glutathione. (B) para-nitrophenylacetate is also conjugated to glutathione by a substitution reaction. (C) trans-2-nonenal is conjugated to glutathione by a Michael addition reaction and is particularly active with hGST A4-4 (Hubatsch et al., 1998). (D) The double-bond isomerization of Δ^5- androstene-3,17-dione to Δ^4-androstene-3,17-dione, a precursor of steroid hormone testosterone (Payne and Hales, 2004; Raffalli-Mathieu and Mannervik, 2005). (E) Reaction of immunosuppressive prodrug azathioprine with glutathione catalyzed by GSTs leads to the formation of the active compound 6-mercaptopurine and simultaneous formation of 1-methyl-4-nitroimidazole conjugate of glutathione (Eklund et al., 2006). (F) 1-methyl-4-nitro-5-(4-nitrophenylthio)-1H-imidazole, a model substrate for azathioprine (Kurtovic et al., 2008a).
Conjugation of glutathione with electrophilic substrates does not always result in detoxication, but can also involve bioactivation reactions, as is the case with certain dihaloethanes (Guengerich, 2005).

GSTs display other functions than detoxication, e.g., intracellular binding of heme, bilirubin and other ligands (Ketley et al., 1975; Litwack et al., 1971) as well as involvement in regulation of Jun N-terminal kinase (Adler et al., 1999). Proteins that display a similar fold as GSTs are involved in fundamentally different functions than that of catalysis. Examples of this secondary function involve cephalopod lens crystallins (Tomarev et al., 1995; Tomarev et al., 1991) and eukaryotic elongation factor 1γ (Koonin, 1994).

Classification and nomenclature

GSTs can be divided into three different categories: the soluble or cytosolic GSTs, the membrane-bound microsomal GSTs, and the bacterial fosfomycin resistance protein (Hayes and Pulford, 1995). This thesis is based on the soluble GSTs.

Based on their sequence similarity the soluble GSTs are divided into seven different classes designated Alpha, Mu, Pi, Theta, Zeta, Omega, and Sigma and abbreviated in Roman capitals (Mannervik et al., 2005). Within each GST class the protein members are named by Arabic numerals, e.g., GSTA1-1 (a protein composed of two copies of subunit 1 belonging to the Alpha class). If the enzymes are more than 50% sequence identical, they are usually regarded as belonging to the same class. In order to specify which species the enzyme belongs to, a lower case Roman letter is used, e.g., hGSTA1-1.

Structure and mechanism

Soluble GSTs are enzymes that are composed of two, approximately 25 kDa, subunits (Figure 7). Each subunit of GST embraces one active site. Both homo- and heterodimers are found in nature. GSTs are regarded to be inactive as monomers, as has been evidenced by GST P1 monomer that did not display any measurable activity (Abdalla et al., 2002). However, recently a catalytically active monomer from class Mu GST from rat has been constructed by introducing mutations in the electrostatic region of the subunit interface that governs monomerization. In addition, the monomers were stabilized by adding potassium bromide to the buffer solutions (Hearne and Colman, 2006). The paper reported that the monomer could be almost as active as the dimer protein.
Figure 7. Dimer (left) and monomer (right) of human GST M2-2 in complex with glutathione (PDB-code 1xw5).

The three-dimensional structures of a large number of GSTs have been determined by X-ray crystallography (Figure 7). In some cases the structure has been determined with ligands and in other cases the apo form has been investigated. Crystal structures exist for many human enzymes including GST A1-1 (Sinning et al., 1993), GST M2-2 (Raghunathan et al., 1994), GST P1-1 (Reinemer et al., 1992), GST T1-1 (Tars et al., 2006), GST O1-1 (Board et al., 2000), and GST Z1-1 (Polekhina et al., 2001). In spite the low sequence similarity between different classes of GSTs (less than 30% being strictly conserved), the tertiary and quaternary structures of these enzymes are astonishingly similar.

However, significant differences occur between the active site regions of GSTs belonging to different classes. Each subunit of a GST is built up of two domains: the N-terminal domain that is composed of an α/β structure and harbors interactions needed for glutathione to bind, and a C-terminal all-α-helical domain that contributes interactions needed for the second electrophilic substrate to bind (Figure 7) (Armstrong, 1997; Board et al., 2000; Sinning et al., 1993). The glutathione binding site that contributes multiple polar bonds (G-site) is heavily conserved among GSTs, whereas the hydrophobic binding site of the electrophilic substrates (H-site) varies considerably among the enzymes (Josephy and Mannervik, 2006). It is this variability in the H-site among different enzymes that is responsible for the diverging substrate selectivity profiles in the GST enzyme family.

The catalytic mechanism of GSTs involves the enzyme providing binding sites for glutathione and the electrophilic substrates as well as positioning them in the right orientation to favor product formation and stabilize the transition state. However, one of the most important features of GSTs in catalysis is their capability to lower the pKa of the sulfhydryl group of glutathione in the active site, thereby increasing the reactivity of glutathione. As mentioned above, the pKa of the sulfhydryl group of glutathione is around 9.2 in solution, but in the enzyme active site it is lowered to 6.1-7.5 depending on the enzyme. In this way, the sulfhydryl group of glutathione becomes
a strong nucleophile. What is responsible for this lowering of pKa? The nucleophilic active-site residue (Tyr or Ser, depending on the enzyme) that is situated close to the sulfur of glutathione in the active site is not directly involved in ionization of the sulfhydryl group. This has been evidenced by site-directed mutagenesis of Tyr in Alpha class GSTs (Pettersson and Mannervik, 2001). However, this active-site nucleophilic residue promotes catalysis by orienting the bonding orbitals of the sulfur of glutathione. The correct orientation of sulfur for the nucleophilic attack is achieved by hydrogen bonding from Tyr or Ser (Josephy and Mannervik, 2006).

Evolution

All soluble GSTs share a common $\alpha/\beta$ fold that is similar to a thioredoxin molecule. It is believed that most soluble GSTs have arisen through addition of an all-$\alpha$-helical part to the already existing thioredoxin fold (Armstrong, 1997). The “ancestral GST” is not known, but it has been proposed by Pemble and Taylor that the Theta class is the precursor of all other GSTs (Pemble and Taylor, 1992). GSTs are also believed to have arisen by both divergent and convergent evolutionary mechanisms.

Many studies of directed as well as rational evolution in the laboratory have been performed on GSTs. Directed evolution experiments on enzymes belonging to several GST classes have led to understanding of structure-function activity relationships among enzymes monitored with several substrates. Among the Theta class GSTs a mutant has been constructed by DNA shuffling that had 1600% increased activity with 1-menaphthyl sulfate and a 60% lower activity with 4-nitrophenethyl bromide (Broo et al., 2002). In Mu class GSTs, DNA shuffling of hM1-1 and hM2-2 led to elucidation of a common structural basis for the enzymatic mechanism for conjugation of the physiologically relevant substrate aminochrome and denitrosation of 2-cyano-1,3-dimethyl-1-nitrosoguanidine (cyano-DMNG) (Hansson et al., 1999a; Hansson and Mannervik, 2000). Important structure-function activity relationships have been investigated by site-directed mutagenesis in Mu class GSTs, where hM2-2 residues were introduced into the framework of hM1-1 (Hansson et al., 1999b). Alpha class hA2-2 has been subjected to rational design restricted to substrate-binding residues, whereby high steroid double-bond isomerase activity, otherwise characteristic of hA3-3, was successfully obtained in hA2-2 (Pettersson et al., 2002). In another study hA4-4 typical H-site residues have been incorporated into hA1-1, resulting in high activity with alkenals that is typical of the hA4-4 enzyme (Nilsson et al., 2000).
Present Investigation

Objectives
A common aim of the papers presented in this thesis is to study enzyme evolution in the laboratory. Questions of how evolution is operating have been addressed by means of multivariate data analysis. By this approach functional traits in both rational and directed evolution have been studied. Another important task was to study structure-function activity relationships among chosen mutants from shuffled libraries in order to find key structural features for a certain function. Directed evolution studies have been performed by means of DNA shuffling of several members of the GST enzyme family. In order to be able to screen a large number of mutants generated by this approach, questions concerning suitable assays have been addressed, thereby leading to development of novel screening methods.

Utilizing different tools in both molecular biology and mathematics we have thus gained insight into how novel quasi-species fit for further evolution evolve, how we can cluster different mutants with regard to their functional capabilities, how altered catalytic capacities and substrate selectivity profiles evolve, and also what structural elements are important in the function of Alpha class GSTs with the clinically used prodrug azathioprine.

Discovery of novel quasi-species in enzyme evolution (Papers I and II)
Today, large emphasis is placed on structural analysis of mutant libraries generated in various ways (Bloom et al., 2005; Meyer et al., 2006; Valetti and Gilardi, 2004). Much less effort is devoted to their functional properites. However, tailor-made novel functions in enzyme evolution are very difficult to achieve on the basis of structural data alone. It is thus desirable to study functional relationships of large mutant libraries in molecular evolution, even though functional relationships are much harder to elucidate since they are usually expressed in many dimensions. In this study we sought to investigate the functional space of a DNA-shuffled Mu class GST library in order
to find out whether it is possible to recognize distinct quasi-species among the library mutants. Another goal of this study was to use multivariate tools in order to detect groupings in the data set that can be coupled to structural data.

A library of cDNA encoding hGST M1-1 (hM1-1) and hGST M2-2 (hM2-2) had been constructed by DNA shuffling (Hansson et al., 1999a). A sample of 384 library mutants was analyzed in bacterial lysates with nine electrophilic substrates, whereupon an extensive multivariate data analysis was performed on the data set.

Mu class GSTs are clustered on chromosome 1 (DeJong et al., 1988) and display high sequence similarities. A typical feature of the three-dimensional structure of Mu class GSTs is the so-called Mu-loop that partly covers the active site (Armstrong, 1997) (Figure 7). The two human enzymes studied here display high sequence identity (84% at the protein level), but despite that show diverging substrate selectivity profiles (Hansson et al., 1999a) as well as tissue distributions (Rowe et al., 1997). hM1-1 displays high activity with certain epoxide substrates and is abundantly expressed in the human liver. hM2-2, on the other hand, has high activity with the substrate aminochrome, a reactive metabolite of dopamine, and is highly expressed in heart, skeletal muscle, and brain (Rowe et al., 1997).

The functional profile of GSTs acting on electrophilic substrates is not normally distributed

The substrates chosen for screening of the library undergo different kinds of chemical transformations. The substrates 1-chloro-2,4-dinitrobenzene (CDNB) (Figure 6A), 4-nitrophenethyl bromide (NPB), para-nitrophenylacetate (pNPA) (Figure 6B) and monochlorobimane (MCB) are conjugated with glutathione by substitution reactions, whereas 3-(4-nitrophenyl)-glycidol (NPG), trans-4-phenyl-3-buten-2-one (tPBO), epoxy-3-(4-nitrophenoxo)-propane (EPNP), 1-nitro-1-cyclohexene (NCH) and phenethyl isothiocyanate (PEITC) are conjugated through addition reactions. CDNB is a standard GST substrate, but the specific activity differs among the different members of the enzyme family, where hGSTs M1-1 and M2-2 display high catalytic activity. The reaction of glutathione with the electrophile CDNB is a nucleophilic aromatic substitution reaction. NPB undergoes an aralkyl substitution reaction with glutathione. The glutathione-dependent release of p-nitrophenol from pNPA represents a transacylation reaction. The fourth substrate chosen to monitor substitution reactions is MCB, a fluorogenic substrate that has activity with both hM1-1 and hM2-2. The opening of the oxirane ring of epoxides NPG and EPNP could occur via nucleophilic addition on either one of the oxirane carbons. tPBO undergoes an addition reaction on the β-carbon of the enone. NCH undergoes addition of glu-
tathione to the double bond of the cyclohexene ring. PEITC is an isothiocyanate that forms a dithiocarbamate when glutathione is added.

The activity values with different substrates were plotted as histograms in order to reveal whether the data follow a Gaussian distribution.

![Histograms showing distributions of activities with nine alternative substrates determined in bacterial lysates of clones chosen from the M1/M2 library.](image)

*Figure 8.* Histograms showing distributions of activities with nine alternative substrates determined in bacterial lysates of clones chosen from the M1/M2 library. The graphs show untransformed data and are mean values of duplicate measurements for every clone and substrate analyzed; the clones encompass 384 randomly picked mutants from the M1/M2 library and the parental enzymes hM1-1 and hM2-2. The bars represent bins of clones with similar activities, the x-axis showing activity and the y-axis the number of clones with the corresponding activity.

*Figure 8* shows histograms of the activities determined in bacterial lysates of individual mutants of the GST M1/M2 library with each of the electrophilic substrates used in the screening. The distributions of the activities with CDNB, MCB, NPG, tPBO, EPNP, NCH, and PEITC are all skewed to the right and clearly do not display a normal probability distribution according to normal probability plots for each substrate (data not shown). The pNPA and NPB data do not deviate strongly from normally distributed values. In fact, NPB, which is a good substrate for hGST T1-1 (Broo *et al.*, 2002), did not show significant activity with any of the GST M1/M2 mutants. The approximately normal distribution of the NPB data is therefore equivalent to sampling values of the nonenzymatic background reaction, which could be expected to have a Gaussian scattering, in distinction from the envelope of the different activities of emerging enzymes with novel properties.
In order to illustrate the functional diversity of the GST M1/M2 library, eleven mutants were plotted from an arbitrarily chosen region of the data set to exemplify the variable activities of individual clones with different substrates (Figure 9). For example, variant 74 has high activities with CDNB, MCB and NCH, but quite low activity with all other substrates used in the screening. On the other hand, variant 71 has high activity with PEITC, but relatively low activity with all other substrates.

Since NPB did not show any significant activity with any of the library mutants or with the parental enzymes, it was regarded as merely showing experimental error background and excluded from subsequent analysis.

What is a quasi-species?

When choosing offspring for a new generation, it is important not to choose only the mutants that have gained the highest activities with desired substrates, since their genetic properties may be too narrow and lead to dead-ends in evolution (Ness et al., 1999). Eigen et al. have proposed by theoretical calculations that the target of selection in evolution, i.e., “the fittest” is not a single type but rather a distribution of neutral types including their related less adapted mutants (Biebricher and Eigen, 2005). Therefore, en ensemble of mutants is more desirable to choose than single individuals as parents for next generation. These evolving units in molecular evolution can be regarded as “quasi-species” (Eigen et al., 1988). The concept of quasi-species has been mostly applied to RNA viruses (Biebricher and Eigen, 2005; Eigen, 1996).
Application of the concept of quasi-species to a GST mutant library

**PC analysis reveals functionally diverging quasi-species**

In order to investigate the functional space of eight dimensions (eight electrophilic substrates used in screening of the library mutants), a PC analysis was adopted (Figure 10). This analysis revealed the existence of three distinct distributions, or quasi-species, among the 384 GST mutants: hM1-1-like, hM2-2-like, and a novel distribution of mutants (Paper I).

---

**Figure 10.** PCA of 384 mutant and parental enzymes assayed with eight electrophilic substrates. The distribution of enzyme variants is represented in A, C, and E whereas the contribution of different substrates to the PCs are shown in B, D, and F. Enzyme variants are color-coded according to their distribution in C with arbitrary boundaries. Black-colored mutants represent low or insignificant activities, whereas red (hM2-2-like) and blue (hM1-1-like) display high overall activities. Green-colored mutants represent novel quasi-species displaying high activities with pNPA and EPNP. (A-D from Paper I)
A scree plot (Paper I) revealed that 77% of the variance is expressed by the first two principal components. In Figure 10A the projection of the substrate-activity data onto the PC2/PC1 plane is shown. All substrates were found to contribute to a similar extent to PC1 (Figure 10B). It could also be seen that PC2 is responsible for separating the parentals hM1-1 and hM2-2 in their catalytic activities, where NCH, CDNB and MCB are signature substrates for hM2-2 and tPBO, NPG and PEITC are signature substrates for hM1-1, as evidenced from the loading plot. Furthermore, it was clear from the analysis that mutants similar to hM1-1 (colored blue in Figure 10) and hM2-2 (colored red in Figure 10) segregate in the PC2/PC1 plane. However, when the PC3/PC2 plane was examined (Figure 10C) a third distribution of mutants (colored green in Figure 10), orthogonal to the parental distributions, emerged. The signature substrates for this novel distribution of mutants are pNPA and EPNP, as judged by the loading plot (Figure 10D). A novel quasi-species was thus recognized displaying characteristics deviating from those of the parental enzymes. The novel distribution of mutants has a changed substrate selectivity profile compared to the parental enzymes so that some activities typifying parentals have dwindled whereas activities with pNPA and EPNP are prominent. When PCs of higher order were examined (PC3/PC5 plane, Figure 10 E and F) it seemed that the novel quasi-species branched off in two directions with EPNP and pNPA as signature substrates for each direction. However, the number of mutants representing the novel quasi-species is too small to permit a clear decision on this point.

Purified mutants display diverging activity fingerprints compared to parental GSTs

So far, the analysis has been performed in bacterial lysates where factors such as expression level can disturb the analysis. In order to find out whether the novel quasi-species is diverging from the parental distributions, two mutants belonging to the new distribution have been purified and, together with parental enzymes, subjected to specific activity measurements with the eight electrophilic substrates used in the screening. The mutants chosen were 342 and 383 (Figure 10). The analysis of purified mutants showed that they indeed displayed activity fingerprints distinct from those of the parental GSTs, thus further supporting the emergence of novel quasi-species in evolution of GSTs (Paper I).

Coping with the problem of expression levels in screening

Mutants having high expression levels may bias the identification of individuals with enhanced or altered catalytic properties. A consequence of this behavior can be that an enzyme produced in low or very high amounts is inappropriately assigned to a certain group. One way of coping with this problem is direct monitoring of the expression of a certain protein, which,
however, can be difficult to perform in practice. One aim of the following study was to find a mathematical transformation capable of taking account of these differences. When the diverging expressions levels have been removed, the possibility to explore groupings in the data set emerges, which is another objective of this study (Paper II).

![Figure 11. Plots of kinetic parameter values in lysate measurement. Correlation between untransformed (slope = 2.4 ± 0.27, R² = 0.929, P = 0.0001) (A) and transformed (slope = 0.94 ± 0.13, R² = 0.899, P = 0.0003) (B) values between parental enzyme M1-1 and mutant 72. (C) Correlation between untransformed activity values between variant 72 versus variant 71 (slope = 1.10 ± 0.016, R² = 0.999, P < 0.0001). (D) Correlation between transformed activity values between variant 72 versus variant 71 (slope = 1.02 ± 0.018, R² = 0.998, P < 0.0001). The parental enzyme hGST M1-1 is abbreviated to M1.]

Sequence analysis of several mutants from the library (discussed in more detail below), revealed that the parental enzyme hM1-1 has the same amino acid sequence as mutants 72 and 119 (Paper II). By plotting the functional parameters of the two enzyme variants against each other and performing a regression analysis, different expression levels for the two enzymes could be elucidated (Figure 11). In the case of M1-1 and 72 it was clear that the lysate containing variant 72 contained 2.4 ± 0.27 (R² = 0.929, P = 0.0001) times more active protein than M1-1 (Figure 11A). In order to compensate for this behavior, each row in the data matrix was transformed to unit length. By this transformation, clones producing functionally identical GSTs are treated as though they were present at the same concentration in the lysates. Furthermore, the eight columns describing the activities with different substrates were all individually normalized to a mean value of zero and unit variance.
By this approach the substrates having high absolute activity values were not allowed to contribute more than those substrates displaying low absolute values in the analysis. When the data were treated with this transformation procedure, the slope of the regression line became close to 1 (slope = 0.94 ± 0.13, R² = 0.899, P = 0.0003), as expected if both enzymes were expressed at the same levels (Figure 11B). A situation can exist where there is an equal amount of enzyme with the same functional activity in two individual clones. In this case both the untransformed and transformed slopes should approach a value of 1. This has been shown in the GST Mu class library for individuals 71 and 72 (Figure 11C and D).

Determination of the number of clusters in a GST mutant library
The primary analysis performed on the data set involved scatter plots of correlations between activities of different substrates that clearly indicated subgrouping of the data (Paper II). In order to statistically investigate these groupings, several different cluster analyses were performed on the data set. In order to avoid the contribution of different expression levels among the library mutants, the data set was transformed according to the procedure described above.

Multivariate data mining reveals four clusters in lysate measurements
Dendrogram clustering as well as K-means cluster analysis were applied to the transformed data set (Paper II). Dendrogram analysis suggested at least four different clusters, the number that subsequently was chosen for a K-means cluster analysis (Figure 12).
Figure 12. K-means cluster analysis of activity values from lysate measurements divided into four clusters (four nodes colored black). The 386 rows of activity values representing parental enzymes as well as library mutants were scaled to unit length prior to normalizing each substrate column to unit variance and zero mean. The clustering was based on Euclidian distances and the members of the corresponding clusters are color-coded according to the following: blue squares, hM1-1-like mutants; red triangles, hM2-2-like mutants; yellow diamonds, mutants with novel properties and brown circles, mutants with impaired or zero activities. The two-dimensional projection of the clusters in eight-dimensional substrate-activity space is shown with EPNP versus CDNB (left) and MCB versus pNPA (right).

The K-means cluster analysis (Figure 12) showed that two clusters involved the parental GSTs M1-1 (blue) and M2-2 (red). Additionally, two clusters with diverging properties were identified (yellow and brown). The brown cluster was found to be composed of mutants having low absolute values of catalytic activities overall. On the other hand, the yellow cluster found intermediate between the parental clusters encompassed mutants with an altered substrate selectivity profile compared to the parental clusters. A PC analysis of the data set that had been transformed by means of both rows and columns, displayed that pNPA and EPNP are signature substrates of the yellow cluster. This is in accordance with the initial PC analysis, where the transformation had been performed only with regard to columns (Paper I).

In addition, a canonical variate analysis was performed that underscored the distinction between the clusters (Paper II). The separation was obvious between the hM1-1-like and the hM2-2-like clusters, but somewhat less clear between the “null” individuals (brown cluster) and the novel distribution (yellow). A clear border between the brown and yellow cluster cannot be expected since the functional properties are stochastic variables subjected to variance.

Characterization of purified variants belonging to three different clusters

Three additional substrates, trans-stilbene oxide (tSBO), cyano-DMNG, and 4-chloro-7-nitro-1,2,3-benzoxadiazole (NBD-Cl) were used for specific ac-
tivity measurements together with the eight substrates used in screening (Papers I and II). Mutants belonging to hM1-1-like (blue), hM2-2-like (red) and the yellow cluster were chosen for purification. Variants 72 and 295 belong to the hM1-1-like cluster (blue), variant 74 belongs to the hM2-2-like cluster (red) whereas variant 383 belongs to the novel distribution (yellow). Mutant 342, however, was assigned to a novel quasi-species cluster in the initial PC analysis (Paper I) whereas it belonged to hM2-2-like quasi-species in cluster analysis with both transformations. The previous clustering is more likely, since mutant 342 is more similar to 383 in functional behavior, as is evidenced by specific activity measurements (Figure 13) (Paper II). The different cluster assignments depending on the transformation procedure could be explained by the fact that mutant 342 lies somewhere between the yellow and red clusters in the eight dimensional substrate activity space, thus making both groupings possible.

![Figure 13](image.png)

*Figure 13.* Specific activity of purified GST variants (72, 295, 74, 342, and 383) together with the parental enzymes, hGST M1-1 (M1) and hGST M2-2 (M2), measured with tSBO, white; cyano-DMNG, gray; and NBD-Cl, black. The specific activities were normalized to unit length for each substrate in order to facilitate comparison.

The data with purified enzymes made it clear that mutants 72 and 295 showed functional properties similar to those of hM1-1, whereas mutant 74 had a similar substrate selectivity profile as hM2-2 (Figure 13). Variants 342 and 383 did not follow a substrate activity profile similar to hM1-1 or to hM2-2, but had acquired novel functional properties (Paper II).
Structure-function activity relationships in a Mu class GST mutant library

DNA sequences were determined for 34 mutants representing all four clusters discussed above (Paper II). The main conclusions were that most of the variants in the novel quasi-species were structurally similar to the hM1-1-like GSTs, with a difference in the active site where a Ser was replaced by a Thr at position 210. This residue has previously been identified as hyper-variable and under positive selection in evolution (Ivarsson et al., 2003; Norrgård et al., 2006). Paper II thus demonstrated the importance of this residue from a totally different approach, suggesting that alterations at this position might give rise to Mu class GSTs with novel functional properties.

Enzymes as vectors in multivariate functional space

In a multidimensional substrate-activity space, each individual enzyme can be represented as a point and be thought of as a vector (Paper II). For two enzymes having the same functional profile in lysate measurements but diverging in the expressivity the vectors will have the same orientation but different lengths. On the other hand, for two enzymes having distinct functional properties the vectors will have different orientations as well. However, all experimental data are subject to some experimental error, implying that the two enzymes having the same function but different expression levels will fall along the same line within a conical confidence contour.

Elucidation of important structure-function activity relationships of a shuffled Alpha class GST mutant library screened with azathioprine (Paper III)

It has for a long time been believed that activation of the prodrug azathioprine is non-enzymatic in nature. However, recent findings show that 6-mercaptopurine, the active form of azathioprine, is released via a displacement effect by glutathione catalyzed by GSTs (Figure 6E) (Eklund et al., 2006). Among 14 human GSTs tested, hGST A1-1 (hA1-1), hGST A2-2 (hA2-2) and hM1-1 were found to have the highest catalytic activity with azathioprine. In the present study, six Alpha class GSTs were recombined by DNA shuffling in order to elucidate important structure-function relationships in GST activity with azathioprine. This study demonstrates the power of DNA shuffling in identifying important structural elements that are responsible for a particular function. This approach can be applied to predict
mutations in drug targets that can give rise to drug resistance even before they occur in a clinical situation.

Azathioprine – a widely used immunosuppressive prodrug

Azathioprine is widely used as an immunosuppressive agent in inflammatory bowel disease, as well as following kidney transplantation (Calne et al., 1962; Elion, 1989). The active compound, 6-mercaptopurine, interferes with DNA synthesis, which is believed to be the main mechanism of action of thiopurine drugs (Lennard, 1992). However, azathioprine treatment shows some severe side-effects that involve bone marrow suppression, liver toxicity, pancreatic toxicity, neuropathy, and arthralgia, among others (Stocco et al., 2007). Bone-marrow toxicity has been shown to be related to patients that are homozygous for thiopurine S-methyltransferase (TPMT) mutations, whereas other side-effects cannot be correlated to a mutated TPMT genotype (Lennard, 2002; Schwab et al., 2002; Stocco et al., 2005). GSTs are believed to be involved in cytotoxic side-effects of azathioprine treatment by glutathione depletion, leading to mitochondrial injury and depletion of ATP leading to cell death by necrosis as a consequence (Lee and Farrell, 2001). Stocco et al. have studied the relationship between the genetic polymorphism of TPMT, as well as human GSTs GSTM1, GSTP1, and GSTT1 and the occurrence of adverse side-effects in patients having inflammatory bowel disease that were treated with azathioprine (Stocco et al., 2007). The conclusion of their study is that patients with a wild type GSTM1 genotype had increased probability of developing severe side-effects when treated with azathioprine.

Shuffling of Alpha class GSTs and screening of the library with azathioprine

Human GSTs A2-2 and A3-3 as well as active site mutants of hA1-1 (Widersten and Mannervik, 1995) together with bovine GST A1-1 (bA1-1), as well as rat GSTs A2-2 (rA2-2) and A3-3 (rA3-3) were subjected to DNA shuffling. A total of 1570 library mutants were screened with azathioprine in bacterial lysates. 62 mutants displaying low or high activities were chosen for further study. Important structure-function activity relationships of hA2-2 (the enzyme having highest specific activity with azathioprine) could be elucidated by this approach.

Of the chosen chimeras, 45 had a unique DNA sequence, and analysis of these mutants displayed that the library had a satisfactory structural diversity. The mean value of recombination sites of the sequenced chimeras was 5.9 ± 2.6 at the DNA level among the mutants that were composed of more than one parental primary structure. The mean number of parental structures
for each variant was estimated to be 4.1. In addition, 0.6 non-synonymous point mutations were detected among the sequenced library mutants.

The activities from screening in 1570 bacterial lysates could be represented by the sum of two Gaussian distributions, where the lowest rates are normally distributed around a value of $7.0 \pm 1.0 \times 10^{-3} \Delta A_{320}/\text{min}$.

Structure-function activity relationships illuminated by segment analysis

The sequenced chimeras could be divided into 23 different segments, excluding a few single sites of variation, at the protein level (Figure 4 in Paper III). The segments are separated by conserved regions and vary in size between two and 11 amino acid residues. Segments 2, 8, and 20-23 contain active site H-site residues, which makes them important in the subsequent analysis (Figure 14). The H-site is partitioned into three different regions denoted A, B, and C (Figure 14 and Figure 15).
Figure 14. Specific activities in μmol min⁻¹ mg⁻¹ (left) and amino acid sequence alignment (right) of 10 purified chimeras. The alignment here represents only the active site regions A, B, and C of 10 purified chimeras, but is based on the whole sequence comparison of all parentals together with 45 chimeras isolated from the Alpha class library (Paper III). Based on this alignment, the sequences have been divided into 23 different segments color-coded according to the following: conserved regions (gray), hGST A2-2 (hA2) sequence all segments as well as other enzymes having the same primary structure in a particular segment as hA2 (dark blue), segments of hGST A1-1 (hA1) sequence that are differing from hA2 as well as all other enzymes with the same primary structure as hA1 (light blue), hGST A3-3 (hA3) specific sequence as well as all other enzymes with this particular sequence in a segment (green), bGST A1-1 (bA1) specific sequence as well as all other enzymes with bA1 primary structure in a segment (yellow), rGST A2-2 (rA2) specific sequence and all subsequent chimeras with the same protein sequence (pink), rGST A3-3 (rA3) particular sequence as well as all other enzymes with this sequence (orange). Spurious point mutations are marked “m” in the figure. H-site residues are indicated by “H” above the alignment. Different segments are marked below the sequence alignment. Parental enzymes as well as different chimeras are marked on the left with parental enzymes above and the different chimeras below. The length of each segment is stated above the alignment.

Of the 62 sequenced chimeras, 10 mutants were chosen for purification and specific activity measurements (Figure 14). These mutants were chosen so that both high activity and low activity with azathioprine would be represented. Of the purified mutants, 1362 and 1347 displayed the highest specific activity with azathioprine. However, none of the purified mutants had a higher specific activity than hA2-2.
Upon analyzing all of the sequences for the purified enzymes together with specific activities we found that segments 20-22, together with segment 2 are the most important contributors to azathioprine activity (Figure 14). This conclusion was based on the assumption that segments displaying large variation in the primary structure are not important for the azathioprine activity, whereas those segments that are similar in structure are indeed important for the activity. Segments 2 and 20-22 are all composed of sequences originating from hA2-2 among the six most active enzymes (hA2-2, hA1-1, 1362, 1347, 608, and 1108), one exception being hA1-1. Among the enzymes having very low activity with azathioprine, none was found to be composed of hA2-2 sequence elements in segments 20-22. However, even some enzymes displaying low specific activities were found to possess hA2-2 derived structure in segment 2. This implied that hA2-2 structural elements in segment 2 alone is not enough to confer azathioprine activity. Rather, a combination of an enzyme having hA2-2 typical structure in segments 2 and 20-22 is crucial for the activity with azathioprine.

The H-site residues found to be important for the azathioprine activity are situated in regions A and C (Figure 14 and Figure 15). hA2-2 contains Ser in position 10 (Phe in hGST A1-1), and Ile in position 12 (Ala in hA1-1) constituting region A. Region C is composed of Met in position 208, Leu in position 213 and Ser in position 216 (Ala in hA1-1) (Figure 15). Region B of the active site is composed of segments belonging to many different parental enzymes thereby not showing any obvious influence on the azathioprine activity.
Synthesis and characterization of a novel substrate for GSTs that mimics azathioprine (Paper IV)

In Paper III a prodrug, azathioprine, was used to study structure-function activity relationships with GSTs to reveal crucial segments for high azathioprine activity. This structural knowledge may serve in engineering enzymes with high catalytic turnover for activation of the prodrug. Evolving GST variants with higher activity towards azathioprine is desirable, since they may find novel medical applications such as antibody-directed enzyme prodrug therapy (ADEPT) and gene-directed enzyme prodrug therapy (GDEPT) for cancer (Denny, 2004). In addition, GSTs could also serve a role in gene transfer as selectable markers (Budak-Alpdogan et al., 2005). However, screening large mutant libraries with azathioprine is cumbersome due to the low activity of the substrate with GSTs. In order to facilitate the screening procedure a novel substrate, 1-methyl-4-nitro-5-(4-nitrophenylthio)-1H-imidazole (NPTI) with higher catalytic rate was synthesized (Paper IV). NPTI mimics azathioprine by sharing the same substituted imidazole moiety whereas the leaving groups differ. 4-Nitrothiophenol and 6-mercaptopurine are released from the reaction of NPTI and azathioprine, respectively, with glutathione (Figure 6 E and F). Furthermore, there is evidence that the imidazole moiety of azathioprine rather than 6-mercaptopurine is involved in the mechanism of immunosuppression (Crawford et al., 1996). Thus, NPTI may also find its use in elucidating the exact mechanism of the biotransformation of azathioprine.

NPTI was proven to be a good model substrate for azathioprine as evidenced by kinetic characterization of several GSTs. The novel substrate had a higher catalytic efficiency, making it suitable in screening large amounts of mutants for improved azathioprine activity.

Synthesis and characterization of NPTI

NPTI was successfully synthesized as verified by NMR spectroscopy as well as matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF) analysis (Paper IV). After measuring specific activity with several classes of GSTs (Alpha, Mu and Theta) it became clear that Alpha class GSTs displayed the highest specific activities with the novel substrate, whereas Theta class GSTs displayed the lowest, and Mu class GSTs were in between. The same ranking order could be observed with azathioprine, implying that NPTI is a good model substrate.
Table 1. $k_{cat}/K_m$ values of several GSTs with azathioprine and NPTI (Table from Paper IV).

<table>
<thead>
<tr>
<th>GST</th>
<th>Azathioprine (s$^{-1}$mM$^{-1}$)</th>
<th>NPTI (s$^{-1}$mM$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hA1-1</td>
<td>0.48 ± 0.04*</td>
<td>198 ± 13</td>
</tr>
<tr>
<td>hA2-2</td>
<td>1.17 ± 0.07*</td>
<td>137 ± 4.6</td>
</tr>
<tr>
<td>hA3-3</td>
<td>0.14 ± 0.03</td>
<td>62.0 ± 3.4</td>
</tr>
<tr>
<td>hA4-4</td>
<td>0.28 ± 0.04</td>
<td>1.0 ± 0.1</td>
</tr>
<tr>
<td>bA1-1</td>
<td>0.03 ± 0.02</td>
<td>22.9 ± 4.5</td>
</tr>
<tr>
<td>rA2-2</td>
<td>0.03 ± 0.00</td>
<td>0.32 ± 0.03</td>
</tr>
<tr>
<td>rA3-3</td>
<td>0.20 ± 0.05</td>
<td>13.0 ± 1.5</td>
</tr>
<tr>
<td>GIMFhelix</td>
<td>0.06 ± 0.02</td>
<td>4.8 ± 1.5</td>
</tr>
</tbody>
</table>

*aFrom (Eklund et al., 2006)*

For a more incisive study of NPTI activity, catalytic efficiencies, $k_{cat}/K_m$, were determined with both NPTI and azathioprine for several Alpha class GSTs (Table 1). Alpha class GSTs were chosen because they displayed the highest specific activities among the GSTs tested. Overall catalytic efficiency was higher with NPTI than with azathioprine, *e.g.*, 760-fold higher with bA1-1, 400-fold higher with hA1-1 and hA3-3.

NPTI – a good model substrate for azathioprine

In order to investigate whether NPTI is a good surrogate substrate for azathioprine, the $k_{cat}/K_m$ data obtained with Alpha class GSTs were illustrated in a bar plot (*Figure 16*). The much higher activity with NPTI was compensated for by scaling the two series of catalytic efficiencies to unit length so that all values could be plotted simultaneously. The results showed that enzymes having high activity with azathioprine also displayed high activity with NPTI (*e.g.*, hA1-1, hA2-2), although the absolute value was somewhat higher with azathioprine than NPTI for hA2-2, but somewhat lower with azathioprine than NPTI with A1-1. Similarly, the enzymes having low azathioprine activity also displayed overall low activity with NPTI (*Figure 16* and (Table 1). NPTI is thus a good model substrate for azathioprine in a screening situation where it is important to choose variants having higher activity for further studies.
Figure 16. Three-dimensional bar plot of $k_{cat}/K_m$ values for several Alpha class enzymes together with azathioprine (light gray) and NPTI (black) as substrates. The catalytic efficiencies were scaled to unit length prior to analysis.

Catalytic capacities and selectivity profiles in a GST mutant library screened with haloalkanes (Papers V and VI)

It is well established that many enzymes demonstrate promiscuous functionalities, or broad substrate specificities, on their way towards a novel function, both in natural and laboratory evolution (Khersonsky et al., 2006). Usually, these secondary catalytic activities are smaller than already existing activities. Thereby, it is becoming increasingly important to detect diverging substrate selectivity fingerprints in enzyme evolution that may allow for a higher catalytic efficiency or even a novel function to evolve. The importance of altered substrate selectivities has been recognized in several studies dealing with directed evolution of proteins (Goddard and Reymond, 2004; Grognux and Reymond, 2004; Hansson et al., 1999a; Varadarajan et al., 2005).

In the present study we addressed evolution of novel substrate selectivity profiles by applying a directed evolution approach to Alpha class GSTs and screen the mutant library with several iodoalkanes (Paper VI). The substrates used share similar structure and undergo essentially the same kind of chemical reaction with glutathione. The results showed that it is possible to recombine GSTs belonging to the same class and thereby sharing similar structures to obtain novel catalytic capacities as well as altered substrate selectivity profiles with iodoalkanes.
Haloalkanes – examples of both detoxication and bioactivation reactions with GSTs

Figure 17. Reaction of glutathione with alkyl halides (Nucl, nucleophile). Figure adopted from (Guengerich, 2005).

Haloalkanes constitute an important family of compounds from an environmental point of view. They are used as pharmaceuticals, herbicides, fungicides, insecticides, flame retardants, intermediates in organic synthesis, etc. (Fetzner and Lingens, 1994; van Pee and Unversucht, 2003). However, many of these compounds have the potential to induce toxicity as well as tumor formation, which has been shown in model studies (Anders, 2004). The reaction of haloalkanes with GSTs results in both detoxication and bioactivation reactions, depending on the substrate (Guengerich, 2005). The substitution reaction of 1,2-diiodoethane with glutathione catalyzed by GSTs is an example where the product is more reactive than the original substrate (Figure 17). If the reaction is left to run for a sufficient period of time, a cyclic episulfonium ion will be formed with certain dihalogens because of the instability of the GS-haloalkane complex. This episulfonium ion might react with nucleophiles such as DNA, which can lead to damage. However, this secondary substitution is very slow compared to the enzyme-catalyzed displacement of the halogen ion. The conclusion was therefore that GST activities based on initial rates only display monosubstitutions of iodine even with the diiodoalkane substrates (Paper VI). This conclusion was further supported by high performance liquid chromatography (HPLC) experiments that did not show more than one conjugation product within the short initial time frame of the reactions measured (Paper VI).

Development of a colorimetric endpoint assay useful in screening large mutant libraries

The reaction of GSTs with haloalkanes involves initial substitution of halide ion (Figure 17) and subsequent halide ion release. If the halide ion is iodide it can be measured spectrophotometrically at 226 nm (Shokeer et al., 2005). However, when screening large amounts of mutants in bacterial lysates the measurements performed at very low wavelength will be subjected to considerable background interference. Therefore, a fast and simple assay suit-
able for screening large number of mutants in bacterial lysates was in demand.

A colorimetric endpoint assay that can be measured spectrophotometrically at 610 nm or simply scored with the human eye was developed (Paper V). The assay is dependent on the classical reaction of iodine with starch giving rise to blue color development. The iodide ion liberated in the reaction of iodoalkanes with glutathione, catalyzed by GSTs, was oxidized to iodine with acidified hydrogen peroxide (Equation (1)). Thereafter, the iodine reacted with amylose in starch, giving rise to the blue color.

\[
\text{GSH + C}_2\text{H}_5\text{I} \xrightarrow{\text{GST}} \text{GS} - \text{C}_2\text{H}_5 + \text{I}^- + \text{H}^+ \\
2\text{I}^- + \text{H}_2\text{O}_2 + 2\text{H}^+ \rightarrow \text{I}_2 + 2\text{H}_2\text{O}
\]

The assay was tested successfully in bacterial lysates of several wild type GSTs belonging to different classes as well as in the screening of a mutant library.

**Introduction of new descriptors useful in enzyme engineering**

The issues of enzyme promiscuity and broad substrate selectivity profiles in enzyme evolution are important concepts both in nature as well as in laboratory designed enzymes. Therefore it is important to monitor a certain enzymatic activity in relation to a pool of other activities that a particular enzyme possesses. However, it is also important to compare different enzymes that exist together in a population, acting on a defined pool of substrates when evolving enzymes with novel properties serving a defined function in a certain environment. Throughout different studies in this thesis it has been shown that multivariate analysis can serve an important tool in understanding the complex relationships among substrates and enzymes in a multi-substrate activity space. By this approach the guide for evolution of novel functions can be illuminated. In order to facilitate analysis of large mutant libraries tested with many substrates, several descriptors were introduced in Paper VI. The descriptors are presented in *Figure 18* and are described further below.
Figure 18. Schematic figure of descriptors and multivariate scaling of enzymes active with alternative substrates. Activities of three enzymes, E1, E2, and E3 are measured with three substrates, S1 (black), S2 (gray), and S3 (white). For each figure (A) – (C) a table shows activity values in arbitrary units for a particular enzyme with corresponding substrate. For further explanation, see the text below. (Figure from Paper VI)

(A) “Specific catalytic capacity” is the sum of several enzymatic activities for a given enzyme measured separately. This measure reflects the magnitude of the catalytic competence of a certain enzyme in relation to all substrates considered.

(B) “Substrate selectivity” describes the catalytic discrimination among the substrates to which an enzyme is being exposed. This measure is a vector that is composed of fractions of activities with the different substrates compared to a background of the total sum of all activities.

(C) “Unit-scaled substrate selectivity” is a vector that can be used to compare different enzymes to each other. Here the comparison is not only dependent on the substrates present but also on the ensemble of enzyme variants compared to each other.
Screening of a mutant library with haloalkane substrates

A GST mutant library (Paper III) constructed by shuffling six different Alpha class GSTs was subjected to screening with several iodoalkanes, including both diiodoalkanes and monoiodoalkanes (Paper VI). The substrates were: 1,2-diiodoethane (diiodoethane), 1,4-diiodobutane (diiodobutane), 1,6-diiodohexane (diiodohexane), 1,8-diiodooctane (diiodooctane), 1-iodohexane (iodohexane), and 3-iodopropylbenzene (3-IPB). The screening method utilized was the endpoint assay based on the reaction of iodine with starch (Paper V).

Functional characterization of bacterial lysates

PC analysis was performed on the data obtained from screening the library in bacterial lysates. It clearly demonstrated that the library mutants display diverging substrate-selectivity profiles with the iodoalkanes used in screening (diiodoethane was excluded from the screening because of the high non-enzymatic background) compared to the parental enzymes (Paper VI). Furthermore, several chimeras displayed higher catalytic activity compared to the best parental enzyme with a certain substrate.

Functional and structural characterization of purified mutants

In lysate measurements, factors such as expressivity, solubility, and stability might influence the outcome of the analysis. Therefore, eight mutants displaying diverging substrate selectivity profiles as well as higher catalytic competence with some substrates compared to parental enzymes were selected for purification and DNA sequencing (Paper VI).

![Figure 19. Multivariate biplots of specific activities of purified enzyme variants and parental enzymes. Analysis made on raw data without any prior scaling (A) and scaled data to unit length making each substrate contribute equally to the analysis (B). Enzyme vectors are represented by solid arrows and substrate vectors by dashed arrows. (Figures from Paper VI)](image-url)
Diverging substrate selectivity profiles could be confirmed for all chosen mutants. However, higher catalytic competence could only be seen for mutants 244 and 419, indicating that different expression levels in lysate measurements were responsible for higher activities among the rest of the mutants. This has been confirmed when screening the whole library with CDNB, also showing much higher activities in lysate measurements but not in purified form. Nevertheless, novel substrate selectivity profiles could be detected by using biplot (Figure 19) and dendrogram analyses (Figure 20). Biplot analysis of untransformed specific activities revealed that, among parentals, hA2-2 had the highest specific catalytic capacity, whereas rA2-2 had the lowest catalytic capacity (Paper VI). The human enzyme, hA2-2, also displayed the most narrow substrate selectivity profile with the highest discrimination towards diiodoethane. rA2-2, on the other hand, had the broadest substrate selectivity profile among the parental enzymes. The remainder of the parental enzymes were placed somewhere between hA2-2 and rA2-2 as regards both their catalytic capacities and substrate selectivities (Figure 19A). This grouping has a counterpart in an hierarchical clustering of the purified variants presented as a dendrogram, where it is clear that hA2-2 makes a cluster of its own, and has the most narrow substrate selectivity, whereas rA2-2 is part of a large cluster and displays the broadest substrate acceptance (Figure 20). In order to make the iodoalkanes contribute equally to the analysis, the data were scaled to unit length with regard to the different substrate activities (Figure 18C). By this approach, a unit-scaled substrate selectivity profile was illustrated. It revealed that mutant 419 distinguished itself from the other variants by having high activity with both 3-IPB and iodohexane.

Figure 20. Hierarchical clustering analysis of specific activities of purified enzyme variants and parental enzymes. Data have not been scaled prior to analysis.
Among the library mutants, chimera 371 had both the highest specific catalytic capacity and the highest selectivity for the most active substrate diiodoethane, as evidenced by biplot- (*Figure 19A*) and dendrogram analyses (*Figure 20*). On the other hand, the lowest specific catalytic capacity was shown by mutant 330, which, together with mutants 146, 197, and 244 had the least narrow substrate selectivity.

*Figure 21.* Three-dimensional bar plots displaying specific activities of purified rGST A2-2 and four chimeras found to have identical H-site residues with this parental enzyme (146, 197, 244, and 330). The specific activities have been scaled to unit length prior to analysis so that each substrate is allowed to contribute equally to the analysis. (A) Specific activities with iodoalkanes and (B) specific activities with additional eight electrophilic substrates, CDNB, EPNP, AD, PEITC, pNPA, nonenal, butynitrite, and cumene hydroperoxide (CuOOH).

In their structural properties, mutants 146, 197, 244, and 330 displayed identical H-site residues to those of rA2-2, which was in accordance with the functional profile of these chimeras (*Figure 19* and *Figure 20*). However, a closer examination of these mutants revealed divergent catalytic properties compared to rA2-2 (*Figure 21*). A three-dimensional plot with normalized specific activities showed that, *e.g.*, both higher and lower catalytic activities compared to rA2-2 with iodohexane were represented among these four mutants (*Figure 21A*). Furthermore, this behavior among chimeras 146, 197, 244, and 330 was confirmed by making an equivalent analysis with additional substrates, indicating that the difference in functionality compared to rA2-2 indeed is real (*Figure 21*). These findings demonstrate the importance of unit-scaled substrate specificity in directed evolution of enzymatic function.
Importance of the substrate matrix in recognition of novel quasi-species (Paper VII)

Throughout this thesis, the concept of enzyme promiscuity has been discussed. Many enzymes acting on xenobiotics and foodstuffs are active with many different substrates and can thereby be considered promiscuous (Khersonsky et al., 2006). However, enzymes that are able to accept many different substrates do it at the expense of low catalytic turnover (Jakoby and Ziegler, 1990). On the other hand, certain enzymes that need to be highly selective in their function are able to catalyze only one reaction. It is essential that enzymes are able to discriminate among different substrates that they might be exposed to in their environment (e.g., cell). However, many naturally evolved enzymes that display highly specific catalytic functionalities might be active with unnatural substrates that they have not been tested for. Their substrate pool, or substrate matrix in the cell, does not contain these additional substrates and no drive for suppressing activities with these compounds has appeared during evolution. For a very specific enzyme to evolve it is thus equally important to be able to suppress certain functions, and not only to evolve a high catalytic rate with a given substrate. As a result, the chemical environment that an enzyme is part of will influence the evolutionary trajectories of that particular enzyme. In Paper VII we introduced the concept of a “substrate matrix”, which in our definition is the array of alternative substrates considered that an enzyme (or several enzymes) are exposed to. In this study, we addressed the question of how alternative substrate matrices may influence the evolutionary trajectories of evolving enzymes.

In previous papers (I, and II), the concept of molecular quasi-species defining evolvable units in molecular evolution was introduced. This concept is applied again in Paper VII.

The results obtained by multivariate data analysis demonstrated the importance of the substrate matrix in molecular enzyme evolution.

PC analysis of screening data reveals two quasi-species

Previously (Paper III) an Alpha class GST library was created by DNA shuffling. The 413 library mutants chosen randomly including the six parental enzymes were subsequently screened with six different iodoalkane substrates (Paper VI). Screening results obtained with four of the iodides were included in this study: diiodobutane, diiodohexane, 3-IPB, and iodohexane. The library mutants were also screened with an additional five substrates: CDNB, PEITC, pNPA, cumene hydroperoxide (CuOOH), and butynitrite. The substrates used undergo different kinds of chemical transformations. Iodoalkanes are examples of alkyl substitution reactions, CDNB is a standard GST substrate that reacts with glutathione in a nucleophilic aromatic
substitution (Figure 6A), pNPA represents a transacylation reaction (Figure 6B), PEITC is conjugated to glutathione via an addition reaction, CuOOH is example of a reduction, and butylnitrite displays a transnitrosylation reaction with glutathione.

Figure 22. Principal component (PC) analysis of activities of the Alpha class GST mutant library. The 407 mutants together with parental lysates assayed with nine alternative substrates are presented in two orthogonal plane projections. The data have been scaled to unit variance and mean centered prior to analysis thus making each substrate to contribute equally to the analysis. (A) and (C) show PC score plots displaying enzyme variants where parental enzymes (bold) and purified mutants are marked with mutant number. (B) and (D) Corresponding loading plots showing how the nine alternative substrates contribute to the three different PCs. The parental enzymes are abbreviated as follows: hA1-1 (hA1), hA2-2 (hA2), hA3-3 (hA3), bA1-1 (bA1), rA2-2 (rA2), and rA3-3 (rA3). (Figure from Paper VII)

The results of the PC analysis of screening data revealed two quasi-species including parental enzymes hA1-1 and hA2-2 in one and rA2-2 in the other. The two directions are evident in the score plots in both PC2/PC1 and the PC3/PC2 planes (Figure 22 A and C). Library mutants from both quasi-species were chosen for subsequent purification and specific activity measurements as indicated in Figure 22. The signature substrates for hA1/hA2 quasi-species are butylnitrite, CuOOH, PEITC, and pNPA. Corresponding substrates for rA2-2 quasi-species are diiodobutane, diiodohexane, iodohexane, 3-IPB, and CDNB (Figure 22 B and D). In this case there is a relation
between the distribution of the library variants and the reaction mechanism of the substrates used for this particular data set.

**Figure 23.** Confidence intervals (95%) of three enzymes belonging to one of the quasi-species, the rA2 quasi-species. The data have been scaled to unit variance and mean centered prior to analysis thus making each substrate contribute equally to the analysis. Confidence intervals have been calculated on individuals belonging only to this quasi-species. The boundaries between hA1/hA2 quasi-species and rA2 quasi-species are based on the PC2/PC1 score plot. (A) Column score plot displaying parental enzymes bA1 and rA2 together with mutant 244 in both PC1 and PC2. Corresponding confidence intervals are shown. (B) Scores (dots) and confidence intervals in the PC2/PC1 plane for parental enzymes bA1 and rA2 together with mutant 244. The conical shape of the confidence contour for the quasi-species is marked. The enzymes are abbreviated as follows: bA1-1 (bA1), and rA2-2 (rA2). (Figure from Paper VII)

Another question addressed in the study considered whether it is possible to define confidence intervals for a certain quasi-species? In Paper II we envisioned enzymes as vectors in substrate-activity space, where two enzymes having the same function but different expression levels will fall along the same line within a conical confidence contour. Accordingly, it should be feasible to find confidence intervals for an entire quasi-species in a multidimensional substrate-activity space. For this purpose, the rA2-2 quasi-species was chosen and subjected to PC analysis alone in nine-dimensional substrate-activity space. The calculated confidence contours (95%) for each member of the quasi-species were plotted for PC1 and PC2. The individual confidence contours for three enzymes are shown in Figure 23A. Further-
more, the score values for each enzyme were plotted in a two-dimensional PC2/PC1 plane together with the confidence intervals (Figure 23B), thereby outlining the confidence contour for this particular quasi-species. The conical shape is explained by the fact that the confidence contours are expected to be lower for the low-activity enzymes and vice versa.

Measurements with additional substrates expose yet another quasi-species

The nine purified variants belonging to the two quasi-species recognized in PC analysis of screening data were, together with parental enzymes, subjected to specific-activity measurements. These measurements were conducted with the substrates used in screening as well as with four additional substrates, \( \Delta^5 \)-androstene-3,17-dione (AD) (Figure 6D), trans-2-nonenal (nonenal) (Figure 6C), diiodoethane and EPNP, that were not used in screening of the library. Multivariate biplot analysis of the specific activities showed that the two quasi-species detected in lysate screening, hA1-1/hA2-2 and rA2-2 quasi-species, could be recognized also in measurements with the purified proteins (Figure 24). The distributions of specific activities followed the same behavior with signature substrates as in the screening of lysates, thus giving further support for the analysis of the lysate activities (Figure 22 and Figure 24). However, one additional quasi-species can be recognized in specific activity measurements that could not be seen in lysate data. The novel hA3-3 quasi-species had AD as a signature substrate. AD was not used as substrate when screening the library, and therefore this quasi-species remained undetected.
Figure 24. Multivariate biplot showing how specific activities of parental enzymes and purified library mutants are governed by 13 alternative substrates. The substrates used in measurements with purified variants contain the nine substrates used in screening as well as additional four substrates (AD, EPNP, DIE, and Non). The substrates are abbreviated as follows: Δ^1-androstene-3,17-dione (AD); 1-chloro-2,4-dinitrobenzene (CDNB); butynitrite (BuONO); Cumene hydroperoxide (CuOOH); trans-2-nonenal (Non); para-nitrophenylacetate (pNPA); phenethyl isothiocyanate (PEITC); epoxy-3-(4-nitrophenoxy)-propane (EPNP); 3-iodopropylbenzene (3-IPB); 1,2-diiodoethane (DIE); 1,4-diiodobutane (DIB); 1,6-diiodohexane (DIH); 1-iodohexane (IH). In the 3D biplot enzyme and substrate vectors are represented by solid and dashed arrows respectively. The data were not subjected to any scaling procedure prior to the analysis. A third quasi-species composed of hA3 having the additional substrate AD as signature substrate can be recognized that is not seen in lysate screening thus showing the importance of substrate matrix in identification of different quasi-species. The enzymes are abbreviated as follows: hA1-1 (hA1), hA2-2 (hA2), bA1-1 (bA1), and rA2-2 (rA2). (Figure from Paper VII)

Birth of a novel glutathione transferase (Paper VIII)

One of the most central tasks in enzyme evolution is to evolve catalysts with tailor-made functions. In this respect, an important question needs to be addressed: when has an enzyme gained a novel function? For an enzyme to alter its functional profile, a requirement of higher catalytic efficiency with a certain substrate alone is not enough. The aspect of an enzyme suppressing activities with other substrates is also essential, leading to a more restricted substrate selectivity profile. It has been proposed that enzymes showing both high catalytic capacity as well as high substrate selectivity have evolved
from descendants with lower catalytic efficiency and broader substrate selectivities, i.e., more promiscuous generalists (Figure 1) (Jensen, 1976; Khersonsky et al., 2006). The study of a rationally designed hA1-1 to mimic first sphere interactions of hA4-4 gave evidence for a biocatalyst capable of accepting many different substrates evolving into an enzyme with high catalytic capacity as well as highly restricted substrate selectivity (Paper VIII).

A mutant called GIMFhelix had previously been constructed (Nilsson et al., 2000) from hA1-1 (showing high catalytic efficiency with CDNB but low with alkenals) to mimic first sphere interactions of hA4-4 (displaying high catalytic efficiency with alkenals but low with CDNB). In this way hA1-1 was redesigned to change the preferred catalyzed reaction from nucleophilic substitution (CDNB) to Michael addition (alkenals) (Figure 6). The site-directed mutagenesis included introduction of hA4-4 specific residues in three important regions of the H-site (Figure 25). In Paper VIII, the functional behavior of GIMFhelix was further investigated, together with naturally evolved GSTs hA1-1 and hA4-4, by measuring catalytic efficiency $k_{cat}/K_m$ with nine substrates and subjecting the data to a multivariate analysis. The questions addressed involve whether GIMFhelix is a new enzyme distinct from hA1-1 and if it is an isoenzyme of hA4-4.

A question of hA1-1 and hA4-4 functionalities has been addressed in two recent publications. In one of the articles the authors have calculated a quantitative index for substrate promiscuity that is dependent on the set of substrates used in the calculation (Nath and Atkins, 2008). This index has been calculated for some members of protease; CYP 450; as well as GST enzyme families. The choice for these calculations when it comes to GSTs was human GSTs A1-1 and A4-4. Since the indices calculated are independent of an enzyme’s overall activity, the authors were able to see whether there is a general correlation between promiscuity and catalytic efficiency. The authors asked the question: “whether functional promiscuity can be achieved simultaneously with catalytic efficiency”. If there is a correlation between substrate activity and catalytic efficiency, then a strong negative correlation between promiscuity and the catalytic efficiency for the most favored substrate could be expected. The conclusions were that this correlation was very weak for the two of the three enzyme classes studied (proteases and CYP 450s). However, the correlation was very good for the GST enzymes A1-1 and A4-4. In another study, Atkins and coworkers addressed the question of correlation between molecular dynamics and enzyme promiscuity (Hou et al., 2007). In this regard, they studied hA1-1 and hA4-4 and concluded that there were global differences in protein dynamics between the two enzymes. Their results also suggested that it might be important for an enzyme to have an extensive conformational plasticity in order to achieve high levels of functional promiscuity.
In paper VIII, GIMFhelix was characterized with a battery of different substrates representing different chemical mechanisms. Three of the chosen substrates fall into the definition of being endogenous cellular compounds that may have governed natural evolution of GSTs. These compounds are trans-2-nonenal (nonenal), cumene hydroperoxide (CuOOH) and $\Delta^5$-androstene-3,17-dione (AD). A 330-fold higher catalytic efficiency was observed for GIMFhelix compared to hA1-1 with nonenal. However, a significant decrease in catalytic efficiency with CuOOH (32-fold) and AD (600-fold) was also observed. GIMFhelix has thus altered its substrate selectivity.
profile compared to hA1-1 to become more narrow towards one substrate, nonenal, illustrated in Figure 26B. However, GIMFhelix has also gained higher catalytic capacity compared to hA1-1 (Figure 26A). In comparison with hA4-4, GIMFhelix has a lower catalytic capacity, but also a broader substrate selectivity (Figure 26).

Figure 26. Catalytic capacity (A) and substrate selectivity (B) for hA1-1, GIMFhelix, and hA4-4. The catalytic capacity has been monitored with nine substrates: nonenal, ethacrynic acid (EA), PEITC, pNPA, diiodobutane, CuOOH, NPTI, butynitrite, and AD.

A multivariate data biplot showed that the GIMFhelix vector is almost perpendicular to that of hA1-1, thereby displaying novel catalytic properties (Figure 27A). Vectors of GIMFhelix and hA4-4 almost coincided in this analysis. However, many enzymes possess secondary dormant activities with alternative substrates, GIMFhelix being no exception. In order for these residual lower activities not to be hidden by larger catalytic efficiencies, the original data matrix was subjected to transformation of all substrate activities to unit length (Figure 27B). After this scaling, the vectors of hA4-4 and GIMFhelix clearly diverged, whereas the functional properties of GIMFhelix still were different from those of hA1-1. Therefore, GIMFhelix can be regarded as an isoenzyme of hA4-4 with slightly different functional properties, but a novel enzyme compared to hA1-1, where the functional profile has changed drastically (Figure 26 and Figure 27). Furthermore, GIMFhelix is an example of divergent evolution from hA1-1, but convergent evolution compared to hA4-4.
Figure 27. Multivariate biplots showing catalytic efficiencies ($k_{cat}/K_m$) of hA1-1, hA4-4 and GIMFhelix together with nine substrates projected to two dimensions. Enzyme vectors are represented by solid arrows and substrate vectors by dashed arrows. Untransformed catalytic efficiencies (A) and transformed catalytic efficiencies to unit length (B). (Figure from Paper VIII)

In evolving enzymes for tailor-made functions, nature has adopted a trade-off between broad substrate specificity and high catalytic capacity of an enzyme. This has been demonstrated also in laboratory enzyme evolution in Paper VIII. We proposed that the mechanism of enzymes with high catalytic capacity having restricted substrate selectivity is dependent on the demand of transition-state stabilization in enzyme catalysis.
Common to several papers included in this thesis is the application of multivariate data analyses to directed as well as rational evolution of enzymes. Prior to the start of this PhD project, not many studies in enzyme evolution had been conducted using multivariate analysis. Instead, the functional data were explored in a more uni- and bivariate fashion. However, as demands in protein engineering are gaining more and more in complexity, refined multivariate methods tailor-made for functional analysis of large mutant libraries are needed. Many well-known multivariate techniques are thus adopted and further refined to be suitable in an enzyme engineering context in this thesis. Furthermore, new descriptors are introduced that facilitate the multivariate analyses. Many aspects of enzyme evolution are addressed in this thesis, where the conclusions are based on extensive multivariate analyses. These involve identification of novel quasi-species, elucidation of diverging substrate selectivities in a glutathione transferase (GST) mutant library, recognition of a novel enzyme constructed by rational design, as well as coping with the problem of different expression levels when screening large mutant libraries in bacterial lysates.

In Paper I and Paper II, novel quasi-species fit for further evolution were identified by extensive multivariate analysis of functional data. Quasi-species are evolving units in molecular evolution that are characterized by being a distribution of mutants with characteristic properties rather than a single individual. Two Mu class GSTs, hM1-1 and hM2-2 were subjected to DNA shuffling, whereupon library mutants in bacterial lysates were screened with eight alternative substrates belonging to addition and substitution reaction mechanisms. Three quasi-species could be recognized based on the multivariate analysis of functional data: hM1-1 quasi-species, hM2-2 quasi-species, and a novel quasi-species having altered functional properties compared to both parental enzymes. These novel quasi-species have the potential of functional progression in molecular enzyme evolution. The problem of different expression levels in bacterial lysates that can influence the functional analysis has been addressed by mathematical scaling operations prior to the analysis. The three identified quasi-species in bacterial lysates were confirmed to have the same properties among purified representatives from each quasi-species, thereby further supporting the correct identification of the novel quasi-species. Use of multivariate data analysis in identification of molecular quasi-species is important for identification of indi-
viduals fit for further evolution both in natural molecular evolution as well as in protein engineering.

Important structural determinants for azathioprine activity with GSTs were found by DNA shuffling of six Alpha class members belonging to human (A1-1, A2-2 and A3-3), bovine (A1-1) and rat (A2-2 and A3-3) origins in **Paper III**. Azathioprine is a prodrug that is widely used as an immunosuppressive agent. The activation of azathioprine to its active form 6-mercaptopurine, is achieved by reaction with the tripeptide glutathione catalyzed by GSTs. Previous studies in our group recognized hA2-2 as the GST variant displaying the highest catalytic activity of all GSTs tested. In this study, the primary structure of Alpha class library mutants was divided into 23 exchangeable segments in order to find important structural determinants for the azathioprine activity. The primary structure was thereafter compared to the azathioprine activity of purified variants leading to recognition of important segments 2, 20, 21, and 22 for the activation of the prodrug. These segments are part of the substrate-binding H-site, where segments 21 and 22 are situated in the C-terminal helix that is characteristic of Alpha class GSTs and also very important in their catalytic function. Therefore, this study demonstrates the power of DNA shuffling in identifying segments of primary structure that are important for a particular catalytic function. Novel GST variants with higher catalytic activity towards azathioprine might find medical applications, such as antibody-directed enzyme prodrug therapy (ADEPT) and gene-directed enzyme prodrug therapy (GDEPT). In addition, they might also find use as selectable markers in gene transfer. However, evolving novel GSTs that have higher activity with azathioprine than the parental enzymes is a time-consuming process, due to azathioprine being a slow substrate in comparison with the most effective GST substrates. Therefore, a novel chromogenic substrate, 1-methyl-4-nitro-5-(4-nitrophenylthio)-1H-imidazole (NPTI), that is faster than azathioprine, was synthesized and characterized in **Paper IV**. The substrate is a structural analog of azathioprine and was shown to mimic activation of the prodrug with GSH catalyzed by Alpha class GSTs, which makes it a valuable surrogate in screening large mutant libraries. Furthermore, there is evidence that the imidazole moiety of azathioprine rather than 6-mercaptopurine is involved in the immunosuppressant effect of the prodrug, a mechanism that might be elucidated with NPTI.

In **Paper V**, an endpoint assay for screening large mutant libraries of GSTs with iodoalkanes was developed. The assay is based on the oxidation of iodide that is formed in the enzymatic reaction. The iodine formed gives rise to blue color in the presence of starch that can be measured at 610 nm or scored with the human eye. This assay was used in screening an Alpha class GST mutant library with six iodoalkanes (**Paper VI**). Both lysate activities and data obtained with purified mutants from the library as well as parental enzymes were subjected to a refined multivariate data analysis that has been
fine-tuned to show how closely related enzymes act on similar substrates. In order to facilitate the analysis three important descriptors were introduced: ‘specific catalytic capacity’, ‘substrate selectivity’, and ‘unit-scaled substrate selectivity’. The chimeras altered both their catalytic capacities as well as their substrate selectivity profiles toward alkyl iodides used in screening and characterization of purified enzymes. The results demonstrate that it is possible to evolve GSTs belonging to the same class to obtain altered catalytic capacities as well as diverging substrate selectivity profiles with electrophilic compounds that are similar in structure and undergo essentially the same chemical reaction. These results are relevant in the choice of parental enzymes for the next generation of mutants in protein engineering. Furthermore, multivariate data analysis facilitates rational design of directed evolution in the laboratory, providing a more efficient tool to functional fitness than the random trial-and-error approach that is believed to drive natural enzyme evolution.

The importance of the substrate matrix in identification of quasi-species was highlighted in Paper VII. An Alpha class GST mutant library constructed by DNA shuffling was screened with iodoalkanes (Paper VI) and additional five electrophilic substrates in this study. The results analyzed by PC analysis of screening data suggested that the data could be divided into two different quasi-species. Nine mutants belonging to both quasi-species were chosen for purification and subjected to specific activity measurements with substrates used in screening as well as four additional substrates. The two quasi-species recognized in lysate data could be confirmed even in specific activity measurements. However, a third quasi-species could be recognized in specific activity analysis that was not seen in lysate data. The signature substrate for this quasi-species was $\Delta^2$-androstene-3,17-dione (AD), that had not been used in screening the lysates. These results demonstrate thus the importance of the chemical environment, or substrate matrix, in recognizing the evolutionary trajectories of evolving enzymes.

A mutant called GIMFhelix has previously been constructed in our laboratory by exchanging the substrate-binding residues in hA1-1 to mimic first-sphere interactions in hA4-4. In Paper VIII, this study was extended to comprise catalytic efficiencies, $k_{cat}/K_m$, for nine alternative substrates, together with hA1-1, hA4-4 and GIMFhelix, in order to monitor the functional behavior of the mutant. Multivariate analysis of the catalytic efficiencies provided evidence that GIMFhelix is a novel enzyme compared to hA1-1. The conversion of hA1-1 into GIMFhelix, represents divergent evolution, whereas GIMFhelix, being an isoenzyme of hA4-4, exemplifies convergent evolution. Major characteristics of the novel enzyme were the narrowed substrate selectivity profile and enhanced catalytic efficiency compared to its parental hA1-1 towards trans-2-nonenal. These results demonstrated how a promiscuous generalist enzyme with broad substrate acceptance could be tailor-made for high substrate selectivity with enhanced catalytic efficiency.
Summary in Swedish

Bakgrund

Inledning


Enzymer – uppbyggnad


Enzymevolution

Naturliga enzymer är resultatet av miljoner år naturlig evolution av genduplicering och mutation. En mutation är en förändring av en viss aminosyra på ett givet ställe hos ett enzym. Mutationer kan också innefatta hela segement av flera aminosyror på rad. Inom laboratorieevolution av nya enzymer går det att utnyttja båda dessa mutationsmekanismer. Ibland resulterar dessa förändringar i struktur också i förändringar av enzymets funktion. Med hjälp av en metod som heter DNA shuffling går det att klippa i DNA:t så att många små segment genereras som senare ihopklistras slumpmässigt till en
ny struktur vilket kan ge upphov till tusentals nya enzymer. Alla dessa metoder har använts i denna avhandling varav DNA shuffling är den mest representerade.

Glutationtransferaser

Multivariata dataanalyser
Min forskning

Varför denna avhandling?

Innan denna studie initierades var multivariata statistiska metoder något som inte användes flitigt inom enzymevolutionen. Det fanns vissa studier gjorda, men inga som tillräckligt djupt hade undersökt möjligheterna med denna analysmetod. Nyckelfrågorna jag ställde mig var:

  Kan vi hitta kluster mellan olika enzymvarianter genererade med DNA shuffling?
  Kan multivariat dataanalys användas för att guida evolvering av nya enzymvarianter med nya egenskaper?
  Hur kommer olika GST substrat som tillhör olika typer av reaktioner att grupperas?
  Hur vet vi när ett enzym har fått nya egenskaper jämfört med parentala enzymer och kan kallas för ett ”nytt enzym”?

Svar på dessa frågor kan leda till högre förståelse av hur enzymevolutionen går till samt också hur det går att utveckla skräddarsydda enzymer som kan användas i många industriella och medicinska tillämpningar.

Utnyttjande av multivariat dataanalys inom enzymevolution

I Paper I, II, VI och VII undersöks ett stort antal mutanter från GST bibliotek genererade med DNA shuffling och screenade med många olika substrat. För att kunna screena med en viss typ av substanser som går under namnet haloalkaner, utvecklades en ny screeningsmetod i Paper V. Dessa data undersöks med många olika multivariata metoder syftandes till att ge svar på en del av ovanstående frågor. Vad vi kom fram till var att evolverbara enheter inom enzymevolution (s. k. quasi-species) kunde identifieras med hjälp av principal komponent analys. Dessa quasi-species kunde vidare delas upp i tre kluster där två bestod av egenskaper som liknar ursprungsenzymerna vid DNA shufflingen, medan det tredje klustret bestod av mutanter med helt nya funktionella egenskaper. Vad gäller substraten och hur dessa grupperar sig i förhållande till mutanterna, står det klart att det inte nödvändigtvis behöver vara samma kemiska mekanism som ligger bakom grupperingarna av substraten (även om det är sant i vissa fall). Dessa studier visar alltså att det går att använda sig av multivariat dataanalys vid identifiering av nya evolverbara enheter i den molekylära evolutionen.

Proläkemedlet azatioprin – GST struktur-funktions samband

I Paper III studerades struktur-funktionssamband mellan det immunsuppressiva proläkemedlet azatioprin och ett Alfa klass GST-bibliotek skapat med DNA shuffling. Ett proläkemedel är ett läkemedel som i sig inte är ak-

Födelse av ett nytt GST

Acknowledgments

Först och främst vill jag rikta ett stort tack till min handledare, Bengt Mannervik. Tack för att du gav mig möjligheten att forska i din grupp, friheten att i denna forskning få ledsagas av min nyfikenhet och för alla berikande diskussioner. Dessutom vill jag tacka dig för den entusiasm och glädje för forskning som du oavbrutet delar med mig av. Ett stort tack också till Anne-Charlotte Mannervik för alla trevliga stunder på Gräsö och i San Francisco.

Tack till min biträdande handledare, Ulf Hellman, för ett fint samarbete, granskningen av avhandlingen och för att du alltid ställer upp.

Jag vill även rikta mina varmaste tankar till Per Lundahl, min första handledare, som tyvärr fick lämna oss alldeles för tidigt.

Jag vill också tacka David Eaker för språkgranskning av min avhandling.

Tack till...

...alla mina medförfattare för ett lyckat samarbete:

...alla de som jag har undervisat med för roliga, givande och lärorika stunder samt ett gott samarbete:
Gun, Helena D, Ann, Malena, Göran, Micke, Gunnar, och Erik. Ett stort tack också till Lillian för all ovärderlig hjälp på kurslabbet.

...alla från min förra grupp för att ni tog emot mig så väl i Uppsala och lärde mig så mycket om membranforskning:
Lisa, Christine, Caroline, Anna, Shusheng och Ingo
...alla före-detta och nuvarande medarbetare vid f. d. Institutionen för naturvetenskaplig biokemi för att ni gjort arbetsplatsen så trevlig med alla onsdagsfikan och trevliga sammankomster även på fritiden: Ylva, Birgitta E, Usama, Natalia, Wei, Emilia, Birgitta T, Ann-Christine, Lisa T, Maryam, Nisse, Maria, Thomas, Mathis, Helena N, Angela, Johan, Angelica, Sofia, Sandra, Nasim och Marcus.

...Inger för all hjälp med administrativa frågor och...

...Per-Axel för hjälpen med apparaterna på labbet.

...alla på aminosyralabbet: Anna, Bo, Ingrid och Josef, för all hjälp med torris och frystorkning av olika substanser.

...alla mina exjobbare: Andreas, Cissi, Sofia, Ronnie, och Olle

...alla mina studenter – det har varit jätteroligt och lärorikt att få undervisa er!

Ett stort tack till fina vänner som jag har fått under min doktorandtid: Malena och Françoise, för allt ert stöd genom alla åren, för att ni är så bra rumsamrater och underbara vänner. Omar and Leticia thank you for your kindness and friendship! Dan för alla tågresor som blev mycket trevligare i ditt sällskap. Ann och Diana för att ni har så lätt till skratt, det smittar av sig! Malin för den trevliga middagen hemma hos dig och för hjälpen med lösenord till studentdatorerna. Lisa för alla filmkvällar och härliga fikan hos dig.

Dessutom vill jag tacka mina vänner utanför forskningen för all vänskap, trevliga stunder tillsammans och allt stöd: Sara, Per, Saida och Maarit

Sist, men inte minst, ett speciellt tack till min familj för allt ert stöd, engagemang och kärlek: Amira, Rolf, Erik, Hamed, Fatima, Zlatan, Mersa, Denis, Demir, Damir, Eivor och Anki. Ni betyder alla väldigt mycket för mig! Tack också till Eriks familj, Ingrid, Börje, Johan och Göta för uppmuntran och visat intresse för min forskning.
References


Broo, K., Larsson, A.-K., Jemth, P. and Mannervik, B. (2002) An ensemble of theta class glutathione transferases with novel catalytic properties generated by sto-


Hubatsch, I., Ridderström, M. and Mannervik, B. (1998) Human glutathione transferase A4-4: an alpha class enzyme with high catalytic efficiency in the conjuga-


Raghunathan, S., Chandross, R.J., Kretsigner, R.H., Allison, T.J., Penington, C.J. and Rule, G.S. (1994) Crystal structure of human class mu glutathione trans-


Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 431

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-8718