Enzyme Engineering of Bovine Trypsin

A thesis submitted to University College London
for the degree of
Doctor of Engineering

Janahan Paramesvaran

Department of Biochemical Engineering
University College London
2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I, Janahan Paramesvaran, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:_____

In loving memory of E. S. Perinbanayagam

(1936 – 2008)

Acknowledgements

I would like to acknowledge the enormous contribution of my supervisor, Dr. Paul Dalby, in helping to develop this project over the last four years.

For inviting me to complete an internship at Eli Lilly, Indianapolis, I wish to thank my industrial supervisor, Dr. Andrew Russell. For making the experience highly productive and enjoyable, I wish to thank Marc Ebtinger, José Hanquier, Jo Ann Henry, Charles V. Wilson II, William Sales, Matthew Cecil, Andrew Cockshott, Barbara Shropshire, Karla Baase, Mark Zweifel, Amanda Parsons, Parviz Shamlou and G. S. Sittampalam.

For funding this project, I wish to acknowledge the EPSRC and Eli Lilly Inc, and for conference grants I wish to acknowledge the Royal Academy of Engineering and the UCL Graduate School.

For their ideas and encouragement, I wish to thank all of my colleagues at UCL. In particular, Sorwar Choudhury, Waqar Hussain, Farjad Ahmed, Julio Martínez-Torres, John Joseph and Pranavan Thillai have been of tremendous help. For their friendship outside college life, I wish to thank Zeeshan Ali, Faisal Siddiqui, Sujay Balasingham, Piranavan Sathiyanathan, Rajaram Ramakrishnan and Gajan Naguleswaran.

Finally, I wish to thank my parents Paramesvaran and Gnanavally, my brothers Muhunthan and Sudarshan, and my uncle Robert Perinbanayagam for their support.

Abstract

Bovine trypsin is a biocatalyst widely used to cleave recombinant proteins during the downstream processing of therapeutic proteins, and is used particularly for insulin bioprocessing. Evolution has produced a wealth of natural biocatalysts over billions of years, which are generally not optimised for specific industrial applications. Bovine trypsin has a relatively broad specificity towards cleavage at the C-terminal end of arginine or lysine residues. Consequently it has a tendency to cleave alternative sites in the insulin process leading to loss of yield and more complex downstream processing. This project describes efforts to alter the primary specificity of bovine trypsin. Trypsin variants were generated using two traditional random mutagenesis methods tailored to improve the chance of producing a useful mutant. These were focussed error prone PCR (fepPCR) and multiple-site saturation mutagenesis (MSSM).

In order to select residues useful for MSSM, a study of the correlation between (1) mutations enhancing specificity or activity and (2) sequence entropy and distance of mutations from the active site was carried out based on past examples of directed and rational evolution. This analysis along with biochemical information for trypsin aided the selection of two specificity "hotspots" for random mutagenesis, each comprising four residues. These hotspots were regions in the trypsin gene close to or directly involved in substrate binding.

Depending on the mutagenesis method used, the size of the mutant libraries differed considerably. For example, fepPCR of a 522 bp region of the trypsin gene required approximately 3,000 mutants to encompass all possibilities whereas the library size for MSSM was 160,000 for each of the selected four-residue regions. Two alternative library screening approaches, with different throughput capabilities, were tested to isolate mutants of interest. Automated colony screening was considered suitable for the smaller fepPCR library and consisted of the following steps: (1)

transformation of a plasmid library into *E. coli* BL21-Gold(DE3) cells; (2) fermentation of individual colonies in 384 square-well microplates; (3) lysis of the cultures; and (4) spectrophotometric activity measurement on a variety of substrates. The best mutant had a 2.54-fold improvement in arginine specificity. For the larger MSSM libraries, a nutritional selection method was developed using *E. coli* arg-auxotrophic strains.

An alternative approach to generating trypsin variants was also explored based on the known ability of bovine trypsin to autolyse into "pseudo-trypsins". Since these pseudo-trypsins are variants of the native form of the enzyme, it was anticipated that they would have specificities different to that of the native enzyme. Efforts were made to separate the variants via novel chromatographic techniques and to characterise them with respect to molecular weight and specificity. Finally, the activity profile of bovine trypsin was comprehensively carried out on a range of novel substrates, and a comparison made between commercially available bovine trypsin and Eli Lilly's recombinant trypsin. Similar reaction profiles were returned by both enzymes on all substrates with the previously unreported finding that there was a preference for cleavage at the C-terminal end of two positively charged basic residues (*i.e.* KR or RR rather than GR).

Table of contents

Abbreviations

AEBSF	4-[2-aminoethyl] benzenesulfonyl fluoride hydrochloride
AMC	7-amino-4-methyl coumarin
ATCC®	American Type Culture Collection
BHI	biosynthetic human insulin
βNA	beta-naphthylamide
Bz	benzoyl blocker group
C-terminus	carboxy terminal end of a polypeptide chain
CCM	combinatorial cassette mutagenesis
CpB	carboxypeptidase B
CEC	cation exchange chromatography
CLERY	combinatorial libraries enhanced by recombination in yeast
DE	directed evolution
DMF	dimethylformamide
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide A/C/G/T triphosphate
DQ	design qualification
E. coli	*Escherichia coli*
EC	Enzyme Commission
EDTA	ethylenediaminetetraacetic acid
ELISA	enzyme-linked immunosorbent assay
EMEA	European Medicines Evaluation Agency
epPCR	error-prone PCR
EPSRC	Engineering and Physical Sciences Research Council
FDA	Food and Drug Administration
fepPCR	focussed error-prone PCR
GMP	good manufacturing practice
HPLC	high performance liquid chromatography
IQ	installation qualification
IVC	*in vitro* compartmentalization
ITCHY	incremental truncation for the creation of hybrid enzymes
kcat	catalytic constant of an enzyme-catalysed reaction
K_M	Michaelis constant
KPB-BHI	Lys Pro chain B - biosynthetic human insulin
KPB-HPI	Lys Pro chain B - human pro-insulin
LDS	lithium dodecyl sulphate
MEGAWHOP	megaprimer PCR of whole plasmid
MES	2-(N-morpholino) ethane sulfonic acid
MHRA	Medicines and Healthcare Regulatory Agency
MSSM	multiple-site saturation mutagenesis
N-terminus	amino terminal end of a polypeptide chain
OQ	operational qualification
pET(T)	pET26b(+) vector with bovine trypsin gene insert
pET(T3)	triple mutant of pET(T) containing K60N, R117N and K145N

pNa	para-nitroanilide
PCR	polymerase chain reaction
PDB	RSCB Protein Data Bank
PQ	performance qualification
RACHITT	random chimeragenesis on transient templates
rN	recombinant form of an enzyme, N
RPR	random-priming recombination
SCOPE	structure-based combinatorial protein engineering
SCRATCHY	ITCHY with DNA shuffling
SDM	site-directed mutagenesis
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate - polyacrylamide gel electrophoresis
SFF	Sepharose™ Fast Flow
SHIPREC	sequence homology-independent protein recombination
SM	saturation mutagenesis
SOP	standard operating procedure
StEP	staggered extension process
TAE	tris-acetate EDTA
TAME	N_{α}-*p*-toluenesulphonyl-L-arginine methyl ester
Taq	*Thermus aquaticus*
UCL	University College London
Vmax	maximum velocity of an enzyme-catalysed reaction
Z	benzyloxycarbonyl

Units

Å	angstrom
Abs_λ	absorbance at λ nanometers
AU	absorbance unit
bp	base pair (of DNA)
c	concentration
°C	degrees celsius
Da	dalton
ε	molar absorptivity
g	gram
l	path length
L	litre
min	minute
M	mole
Ω	ohm
OD_λ	optical density at λ nanometers
ODU	optical density unit
pfu	plaque forming unit
pH	$-\log_{10}[H^+]$
rpm	revolutions per minute
RT	room temperature (maintained at 25 °C)
s	second
U	unit of enzyme activity*
V	volt
v/v	volume/volume
w/v	weight/volume

Unit prefix	Long form	Multiplication factor
M	mega	10^6
k	kilo	10^3
m	milli	10^{-3}
μ	micro	10^{-6}
n	nano	10^{-9}
p	pico	10^{-12}

*One U is defined as the amount of enzyme that catalyses the conversion of 1 μmol of substrate per minute.

List of figures

List of tables

1 Introduction

1.1 Proteases

1.1.1 Protease families

Proteases or peptidases belong to a class of enzyme that catalyses the cleavage of peptide bonds in proteins. Based on this function they are classed as hydrolases, and within this group as peptide hydrolases (EC 3.4). The major classifications of enzymes are summarised in **Table 1-1** based on the reactions they catalyse. The proteases constitute a large family divided into endopeptidases and exopeptidases, according to the point at which they break the peptide chain. Endopeptidases act preferentially on the inner regions of peptide chains while exopeptidases act only near the ends of peptide chains at the N or C terminus. Endopeptidases may be further subdivided, according to the reactive groups at the active site involved in catalysis, into serine, cysteine, threonine, glutamic acid and aspartic acid endopeptidases as well as metalloendopeptidases (Barrett *et al.*, 2003; Rawlings and Barrett, 1993).

1.1.2 Proteases as biocatalysts

Proteases are found in a wide variety of sources including plants, animals and microorganisms playing a crucial role in many physiological processes. They find extensive applications in a number of industries including the processing of food and dairy produce as well as in the detergent and leather industries (Rao *et al.*, 1998). This project is of relevance to the

biopharmaceutical sector, however, which includes all medical drugs produced using biotechnology. Biopharmaceuticals consist of proteins (hormones and antibodies) and nucleic acids used for therapeutic or diagnostic purposes, and derived by means other than direct from a native biological source. While proteases themselves have limited therapeutic potential, they serve as highly useful tools for specific biocatalytic reactions, often making up essential steps of industrial bioprocesses (see Section 1.1.3.2 for uses of bovine trypsin and Section 1.2.1 for uses of proteases in fusion processes). Exoproteases may be used for the cleavage of process intermediates at the N or C terminus of a peptide chain (e.g. carboxypeptidases A, B and dipeptidylaminopeptidase).

1.1.3 Bovine trypsin

1.1.3.1 Classification

Bovine trypsin (EC 3.4.21.4) is a serine endoprotease based on the classification system described in Section 1.1.1. The serine residue at position 195 (chymotrypsinogen numbering system) in the active site forms a covalent bond with the substrate and gives the enzyme its primary classification. Along with His57 and Asp102, the three residues constitute the catalytic triad responsible for the catalytic mechanism of the enzyme (Section 1.1.3.4). Proteases may be grouped according to other properties such as their physiological function as shown in **Table 1-2**.

EC number	Classification	Reaction catalysed
1.	Oxidoreductases	Oxidation-reduction
2.	Transferases	Transfer of functional groups
3.	Hydrolases	Hydrolysis reactions
4.	Lyases	Addition of a group across a double bond
5.	Isomerases	Intra-molecular rearrangements
6.	Ligases	Bond formation using energy derived from the hydrolysis of ATP

Table 1-1. Six major classes of enzymatic reactions. Trypsin catalyses a peptide hydrolysis reaction and therefore falls in the hydrolase category.

Physiological function	Protease
Digestive system	Enteropeptidase, Trypsin, Chymotrypsin, Elastase
Complement system	Factor B, Factor D, Factor I
Coagulation	Thrombin, Factor VIIa, Factor IXa, Factor Xa, Factor XIa Factor XIIa, Kallikrein
Immune system	Chymase, Granzyme, Tryptase, Proteinase 3
Endocrine system	Proprotein convertases (1, 2)
Reproductive system	Acrosin
Snake venom	Ancrod, Batroxobin
Bacteria (various)	Subtilisin, Streptokinase, Pronase

Table 1-2. Serine endoproteases grouped according to their physiological function. Complement system refers to the biochemical cascade that clears pathogens from an organism and is part of the larger immune system as such.

1.1.3.2 *Uses of bovine trypsin*

The natural function of bovine trypsin, along with other mammalian trypsins, is to aid in the digestion of proteins consumed by an organism. This has been exploited for use in various baby foods where trypsin is added to pre-digest certain foods with a high protein content; casein proteins in breast milk may also be digested by trypsin. Furthermore, the enzyme (along with chymotrypsin) has proven useful to patients recovering from burns, hastening the recovery period significantly compared to patients not given the same therapy (RaviKumar *et al.*, 2001; Latha *et al.*, 1997).

In the biotechnology industry, the enzyme has found some important alternative applications. In animal cell culture, trypsin is used to re-suspend cells that would otherwise adhere to cell culture vessels and dishes (Freshney, 2005), making the harvesting process easier. Eli Lilly use the enzyme in their insulin production process where it has a vital role in the biotransformation of human pro-insulin to the active form, biosynthetic human insulin (Kemmler *et al.*, 1971) (see Section 1.2.3).

1.1.3.3 *Structure and specificity*

In its mature form, bovine trypsin has a molecular weight of 23,800 Da (Cunningham, 1954), and is highly specific towards the carboxy-terminal side of arginine and lysine residues (Olsen *et al.*, 2004; Keil-Dlouha *et al.*, 1971b). The negatively charged Asp189 residue at the base of the binding pocket is reported to provide the basis for attracting the positively charged

basic residues arginine and lysine (Graf *et al.*, 1988). The Schechter-Berger notation (Schechter and Berger, 1967) is commonly used to show the association between subsites across the surface of an enzyme and peptide residues comprising a substrate (**Figure 1-1**).

Produced in the pancreas, the inactive zymogen (precursor), trypsinogen, is secreted into the duodenum of the small intestine. There it is activated into mature trypsin (223 residues) by another serine protease, enteropeptidase (formerly known as enterokinase). The recognition sequence of enteropeptidase is DDDDK (Yamashina, 1956), which occurs once in the entire trypsin sequence close to the N-terminus: H_3N^+-**VDDDDK**IVGGYT. Once cleaved, a newly synthesized active trypsin molecule is yielded.

The crystal structure of bovine trypsin has been solved to 1.8 Å resolution (Bode and Schwager, 1975) and is available at the RSCB Protein Data Bank (structure file: 1BTW). The various secondary structural elements of the mature enzyme are shown in **Figure 1-2**. There is a relatively low representation of α-helices with only 10 residues (4%) making up 2 helices while β-strands make up considerably more with 70 residues (30%) comprising 13 strands.

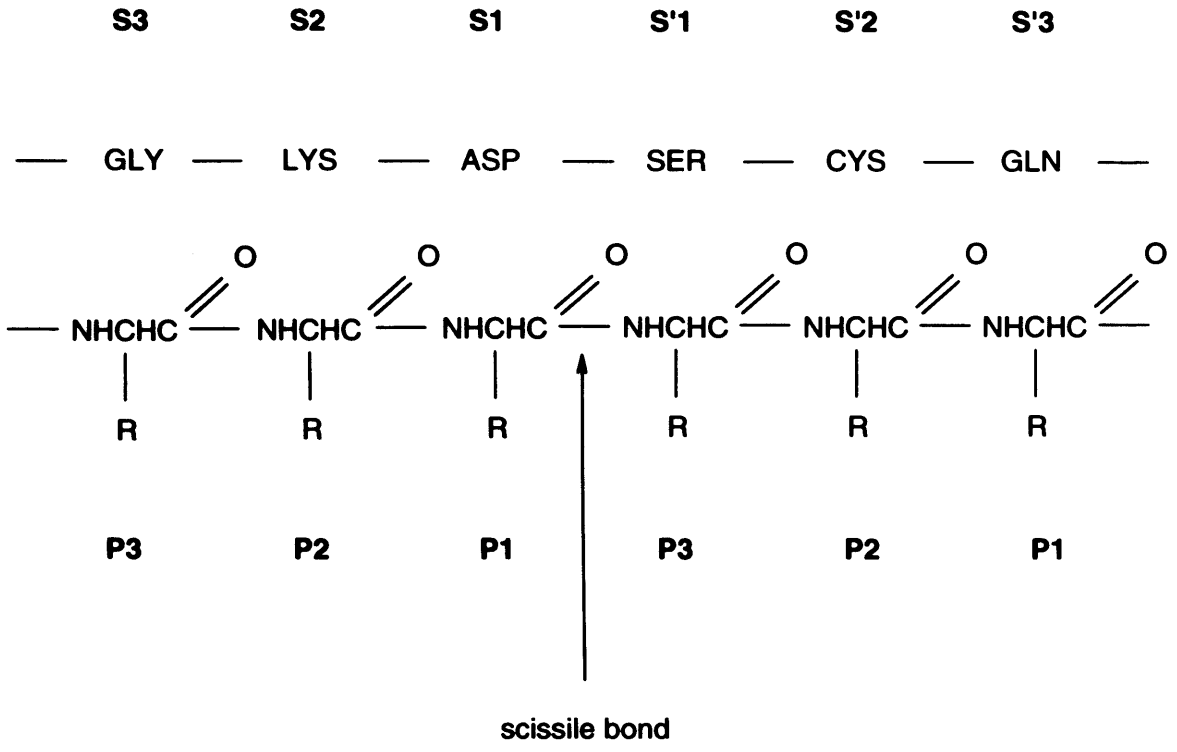

Figure 1-1. Schechter-Berger notation for binding sites in bovine trypsin. Enzyme subsites are denoted by S; peptide sites are denoted by P. The residue at the primary binding site S1 is Asp189 located at the base of the binding pocket. The residue at P1 must be either an arginine or lysine by definition of the trypsin specificity. During peptide hydrolysis at the C-terminal end of an arginine or lysine residue, the scissile bond is cleaved. Figure adapted from Fersht, 2002; Copeland, 2000.

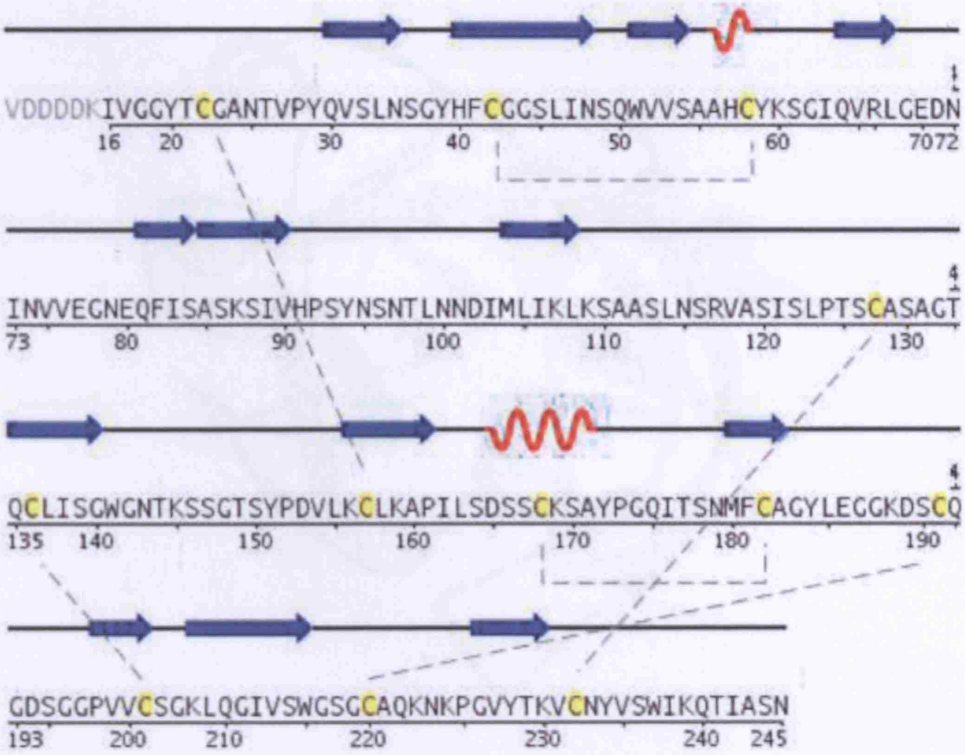

Figure 1-2. Primary and secondary structural elements of bovine trypsin. The chymotrypsinogen numbering system has been used. The enzyme consists of 223 amino acids not including the pro-sequence VDDDDK, which is coloured grey and represents the region that is removed by the action of enteropeptidase. Thirteen β-strands are coloured blue and consist of 70 residues in total (30%); two α-helices are coloured red and consist of 10 residues in total (4%). Cysteine residues are highlighted in yellow with their associated disulphide bridges shown by dashed lines. Figure adapted from Berman *et al.*, 2000.

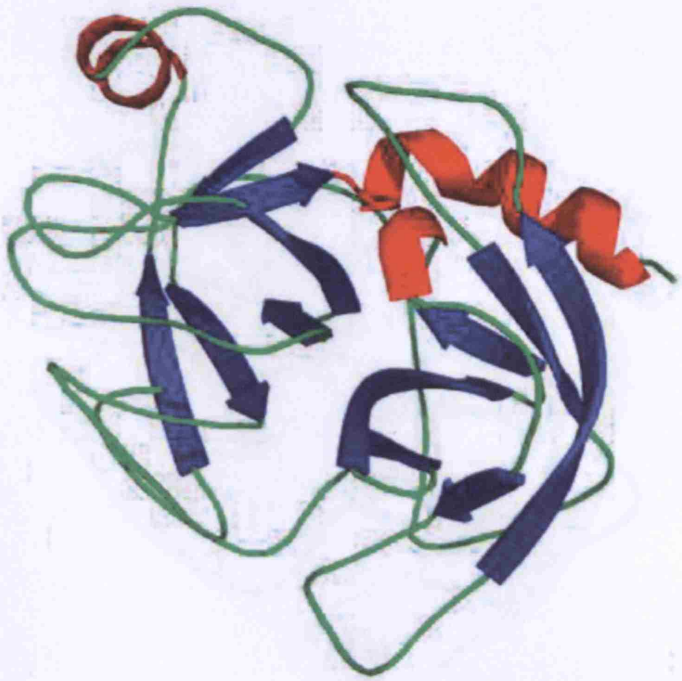

Figure 1-3. Cartoon representation of bovine trypsin. The backbone of the molecule is shown with α-helices coloured red and β-strands coloured blue. Connecting loops are coloured green. The directions of strands are shown from N to C terminus. Two separate β-barrels may be seen on the left and right-hand sides of the molecule respectively. The catalytic triad lies at the interface to these two barrels. Created from the PDB structure file 1BTW.

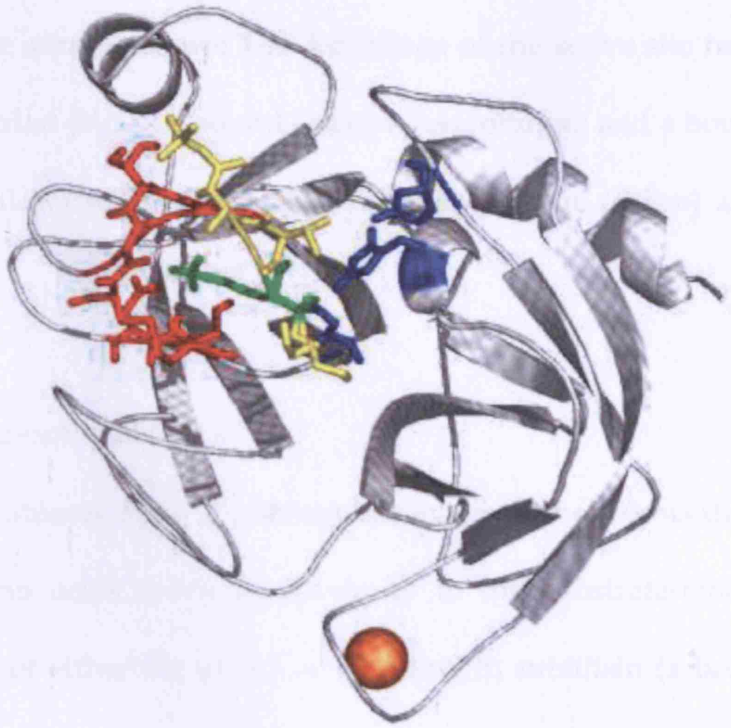

Figure 1-4. Three-dimensional structure of bovine trypsin showing active site and catalytic residues. Active site residues Asp189, Ser190, Cys191, Gly216, Ser217, Gly219 and Cys220 are coloured red; catalytic triad residues His57, Asp102 and Ser195 are coloured blue. The bound inhibitor molecule, T-Butoxy-ala-val-boro-lys 1,3-propanediol monoester, is coloured yellow with the exception of the constituent lysine residue, which is coloured green. The bound calcium ion is coloured orange. Also shown in the background are the α- helices and β-strands coloured in grey. The interaction between the active site and the lysine residue of the inhibitor can be seen clearly. Also of importance is the proximity of the catalytic triad, which provides a mechanism for catalysis. Created from the PDB structure file 1BTW.

The three-dimensional arrangement of the helices and strands of bovine trypsin can be seen in **Figure 1-3**. Locations of the active site residues (red), the catalytic triad (blue), a bound calcium ion (orange) and a bound inhibitor molecule (yellow) with its constituent lysine residue (green) are shown in **Figure 1-4**.

1.1.3.4 Mechanism of catalysis

The serine proteases share a common mechanism of catalysis that requires a triad of amino acids to be in proximity to the substrate-binding pocket. Replacement of either the serine or histidine in subtilisin (a bacterial serine protease) resulted in a reduced rate of reaction of up to 10^6 fold (Carter and Wells, 1988) indicating the immense catalytic efficiency of adopting such a system. It is often referred to as a conserved charge relay system.

The reaction mechanism for peptide hydrolysis is outlined in **Figure 1-5**. A serine residue (hence the family name) acts as the primary nucleophile for attack of the peptide bond. The nucleophilicity of this group is enhanced by interacting with a histidine side chain, which in turn interacts with an aspartate side chain. It was possible to freeze the enzyme during catalysis, in each transition state, by complexing with various small peptide inhibitors (Bode and Schwager, 1975), thereby providing insights into the mechanism.

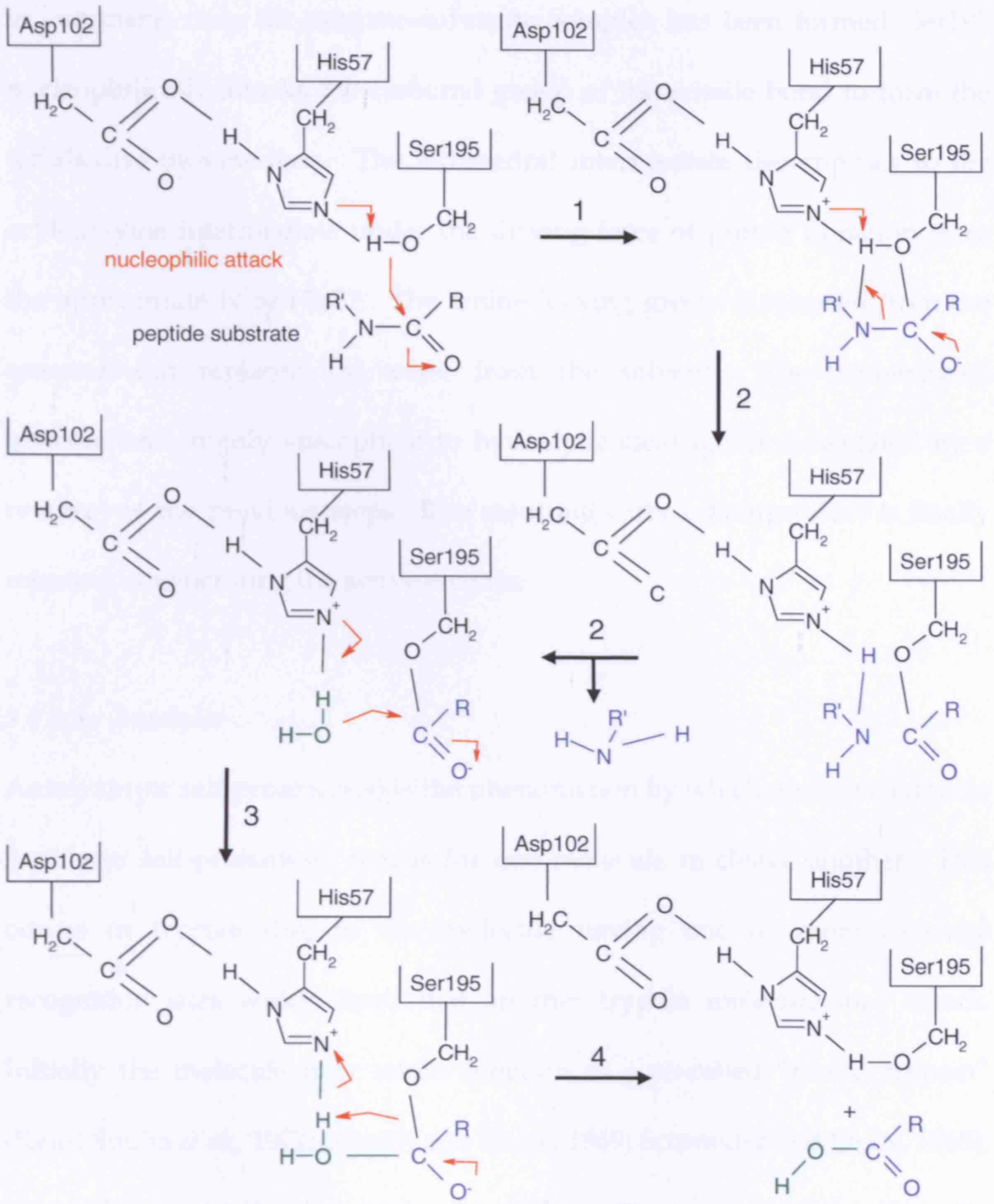

Figure 1-5. Catalytic mechanism of serine endoproteases. The reaction involves (1) nucleophilic attack of Ser195 on the carbonyl C atom of the scissile peptide bond to form the tetrahedral intermediate; (2) decomposition of the tetrahedral intermediate to the acyl-enzyme intermediate through catalysis by the Asp-polarised His, followed by the loss of an amine product and addition of a water molecule; (3) reversal of step 2; and (4) reversal of step 1 to yield a carboxyl product and the active enzyme. Adapted from Voet *et al.*, 2004.

In summary, once the enzyme-substrate complex has been formed, Ser195 nucleophilically attacks the carbonyl group of the scissile bond to form the tetrahedral intermediate. The tetrahedral intermediate decomposes to the acyl-enzyme intermediate under the driving force of proton donation from the appropriate N of His57. The amine leaving group is released from the enzyme and replaced by water from the solvent. The acyl-enzyme intermediate, highly susceptible to hydrolytic cleavage, is deacylated by a reversal of the previous steps. The resulting carboxylate product is finally released, regenerating the active enzyme.

1.1.3.5 Autolysis

Autolysis (or self-proteolysis) is the phenomenon by which an active enzyme is able to self-proteolyse, that is for one molecule to cleave another. This occurs in trypsin due to the molecule having one or more cleavage recognition sites within itself that another trypsin molecule may attack. Initially the molecule may retain function as a so-called "pseudotrypsin" (Keil-Dlouha *et al.*, 1971a; Smith and Shaw, 1969; Schroeder and Shaw, 1968), although eventually, the molecule will inactivate once the active site structure or catalytic mechanism has been compromised beyond a threshold. Efforts to reduce the effect of autolysis by site-directed mutagenesis have met with success in the case of rat trypsin (Varallyay *et al.*, 1998). Three sites were identified as those predominantly responsible for allowing the enzyme to remain susceptible to proteolytic cleavage. Each of these was replaced by

an asparagine residue, which conferred resistance to autolysis as measured by carrying out an activity assay at regular intervals.

1.1.3.6 Expression systems

Bovine trypsin has traditionally been obtained by purifying directly from pancreatic extracts. However, safety concerns have recently emerged over the use of products derived from animal sources in the pharmaceutical industry due to the rise of diseases such as bovine spongiform encephalopathy, for example. Using a recombinant industrial process for expression of the enzyme is the most obvious alternative, and currently Eli Lilly produces the inactive zymogen, trypsinogen, via an *E. coli* expression system. Inclusion bodies produced are refolded, purified and activated to yield mature trypsin. Another alternative, more recently developed, is to produce the zymogen in the yeast strain *Pichia pastoris* (Hanquier *et al.*, 2003), which has the added benefits of high titres (Cregg *et al.*, 1993) and ease of scaling-up (Wegner, 1983).

The procedures described are useful for obtaining industrial quantities of the enzyme, however, they are not practical for screening applications in the laboratory due to the number of steps involved. Expression of the active enzyme in microscale quantities would minimise the number of processing steps required. This objective was successfully realised for both the bovine trypsin gene (Hibbert, 2003) or the rat trypsin gene (Vasquez *et al.*, 1989) by removing the pro-peptide sequence

(VDDDDK) from trypsinogen in the expression vector. In doing so, any perceived problems of toxicity of the enzyme to its *E. coli* host did not prove to be significant enough to prevent cell growth and expression. Problems associated with autolysis, however, were not fully characterised with respect to their effects on assay reproducibility and screening efforts. Expressing the active enzyme also allows for a nutritional selection approach to isolating mutants of interest (see Section 1.3.2.3) in addition to simplifying traditional plate-based screening methods by removing the need for a solubilisation step (Evnin and Craik, 1988).

A recombinant plasmid, given the name pET(T), was constructed by ligating the bovine trypsin gene into the pET26b(+) vector (EMD Biosciences Inc.) (**Figure 1-6**). The vector possessed several useful traits, namely (1) a kanamycin resistance gene; (2) a T7 promoter region to act as a transcription start point; (3) a *lac I* gene, which expresses a *lac* repressor for the T7 RNA polymerase gene present in a host cell; (4) a *pelB* signal sequence for directing newly synthesized proteins to the periplasm of a cell (Martoglio and Dobberstein, 1998); and lastly (5) a His•Tag® to facilitate detection or purification of the protein post-expression. The plasmid described was donated for use in this project by Edward Hibbert, Dept. of Biochemical Engineering, UCL.

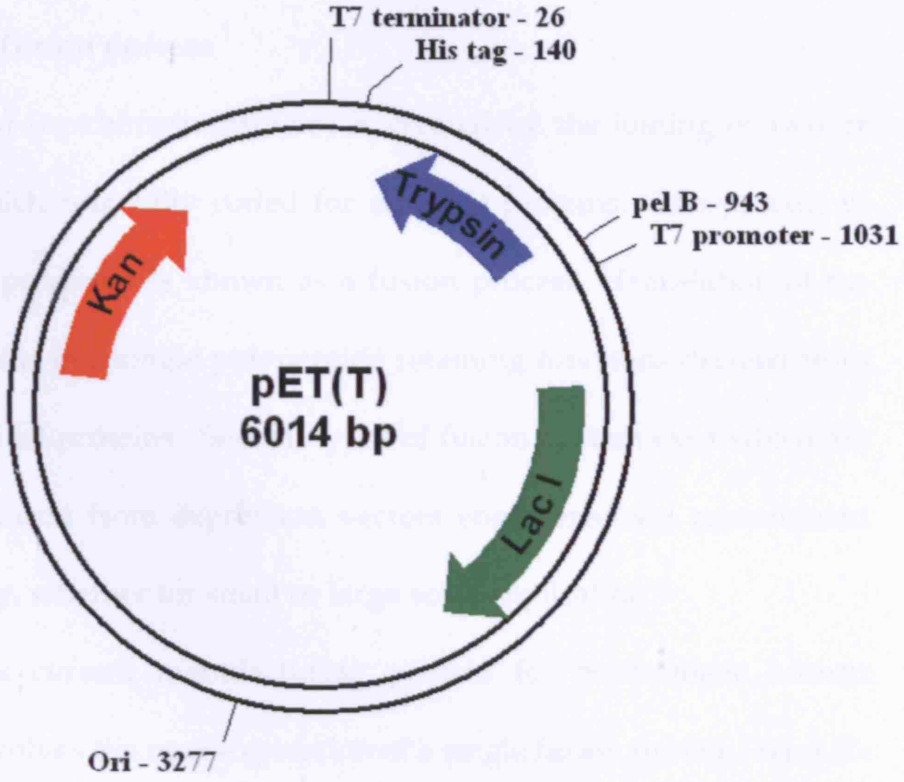

Figure 1-6. Recombinant pET(T) plasmid. The plasmid consists of the pET26b(+) vector with the bovine trypsin gene inserted at the multiple cloning site between restriction sites *Nco* I and *Bam* HI. Directly upstream of the gene is a *pelB* leader sequence. Directly after the multiple cloning site is the His•Tag® .

1.2 Fusion processes

1.2.1 Types of fusion process

A fusion protein (or chimeric protein) is created by the joining of two or more genes, which originally coded for separate proteins. The process in which they are produced is known as a fusion process. Translation of the fusion gene results in a single polypeptide retaining functions derived from each of the original proteins. Several types of fusion protein exist which are essentially produced from expression vectors engineered via recombinant DNA technology, whether for small or large-scale application.

Eli Lilly's current manufacturing process for biosynthetic human insulin (BHI) involves the over-expression of a single fusion protein, Trp-LE'-Met-proinsulin, in *E. coli* (Section 1.2.3). The tag used has a two-fold application. Firstly, the Trp-LE' portion promotes the accumulation of expressed proteins as inclusion bodies (Miozzari and Yanofsky, 1978). Secondly, the single methionine residue at the beginning of the target polypeptide sequence, specifically acts as a recognition site for cyanogen bromide (CNBr) cleavage (Schroeder *et al.*, 1969).

Two different fusion protein concepts are also used in the expression of bovine trypsin via the pET(T) vector. Firstly, the leader (or signal) sequence *pelB* is used for the translocation of newly synthesized proteins to the periplasm. This aids in the expression and folding of non-secreted proteins and prevents the formation of intracellular inclusion bodies. Secondly, the polyhistidine tag (His•Tag®) aids in the purification of a

protein (Hengen, 1995). The tag is an amino acid motif consisting of at least six histidine residues located at the N or C terminus of the target protein. The protein may be purified by metal affinity chromatography whereby the histidine tag binds to nickel ions in an affinity media while contaminant proteins remain unbound and may be separated from the protein of interest. The tagged protein may also be detected using immuno-analytical methods such as Western blotting or ELISA. Glutathione S-transferase (GST) and maltose binding protein (MBP) are also commonly used as tags that increase the solubility of target proteins and aid their purification by affinity methods.

1.2.2 Alternatives to fusion processes

A number of alternatives to the current *E. coli*-based fusion process for the production of human insulin have previously been described (Harrison *et al.*, 2003). These are (1) extraction from human pancreas; (2) chemical synthesis via individual amino acids; (3) conversion of porcine insulin or "semisynthesis" (Morihara *et al.*, 1979); (4) separate expression and combining of the two mature insulin chains (Johnson, 1983); and (5) a yeast expression system to produce single-chain insulin precursor (Barfoed, 1987).

Extraction from the human pancreas is not practiced due to the limited availability of raw material and also because human material may contain harmful pathogens and would require expensive screening. This caveat also applied to the development of an expression system for bovine

trypsin (Section 1.1.3.6). Total chemical synthesis, while technically feasible, is not viable economically due to the very low yield. Semisynthesis transforms porcine insulin, differing by only one amino acid, into an exact replica of human insulin by substituting the amino acid threonine for alanine at position 30. This technology has been developed and implemented by Novo Nordisk A/S. However, this option is also expensive since it requires the collection and processing of large amounts of porcine pancreases.

The two-chain method was the first successful technique for the production of BHI based on recombinant DNA technology. It was developed by Genentech, Inc. and scaled-up by Eli Lilly. Each of the two constituent insulin chains is expressed as a β-galactosidase fusion protein in *E. coli* forming inclusion bodies. The two chains are recovered from the inclusion bodies, purified, and combined to yield mature human insulin. The drawback to this method is the need for two parallel fermentations and purification processes.

Novo Nordisk A/S have also developed a technology based on yeast cells that secrete insulin as a single-chain insulin precursor. It is then converted to human insulin by means of a transpeptidation reaction in organic solvent in the presence of trypsin and a threonine ester followed by de-esterification. An advantage of this technology is the ability to recycle the cells using a continuous bioreactor-cell separator loop. Secretion of the target protein also simplifies isolation and purification of the product.

1.2.3 Insulin production process

The industrial bioprocess for the production of BHI, currently used by Eli Lilly, may be divided into four sections: (1) fermentation; (2) primary recovery; (3) reactions; and (4) final purification. An outline of the major processes is given in **Figure 1-7**.

Initially, the expression vector for the single fusion protein, Trp-LE'-Met-proinsulin is transformed into *E. coli* cells. The cells are used to inoculate a large-scale fermenter (> 50 m^3) with the fermentation lasting 18 hours at 37 °C. The protein accumulates intracellularly as insoluble aggregates (inclusion bodies). Centrifugation follows in which the broth is concentrated approximately four-fold and most of the extracellular impurities are removed. A step is then required to release the product from the cells. A high pressure homogeniser is used to mechanically rupture the cells and release the inclusion bodies. The inclusion bodies are recovered by centrifugation and then solubilised by the addition of urea and 2-mercaptoethanol.

The fusion tag is released from Trp-LE'-Met-proinsulin by CNBr cleavage yielding proinsulin. The proinsulin is not folded correctly, however, since the CNBr has a tendency to break existing disulphide bridges. The subsequent sulfitolysis step, in a solution of guanidine, is necessary to completely unfold the proinsulin and break the disulphide bridges. SO_3 moieties are added to cysteine S groups. A refolding step then removes the SO_3 moieties allowing for

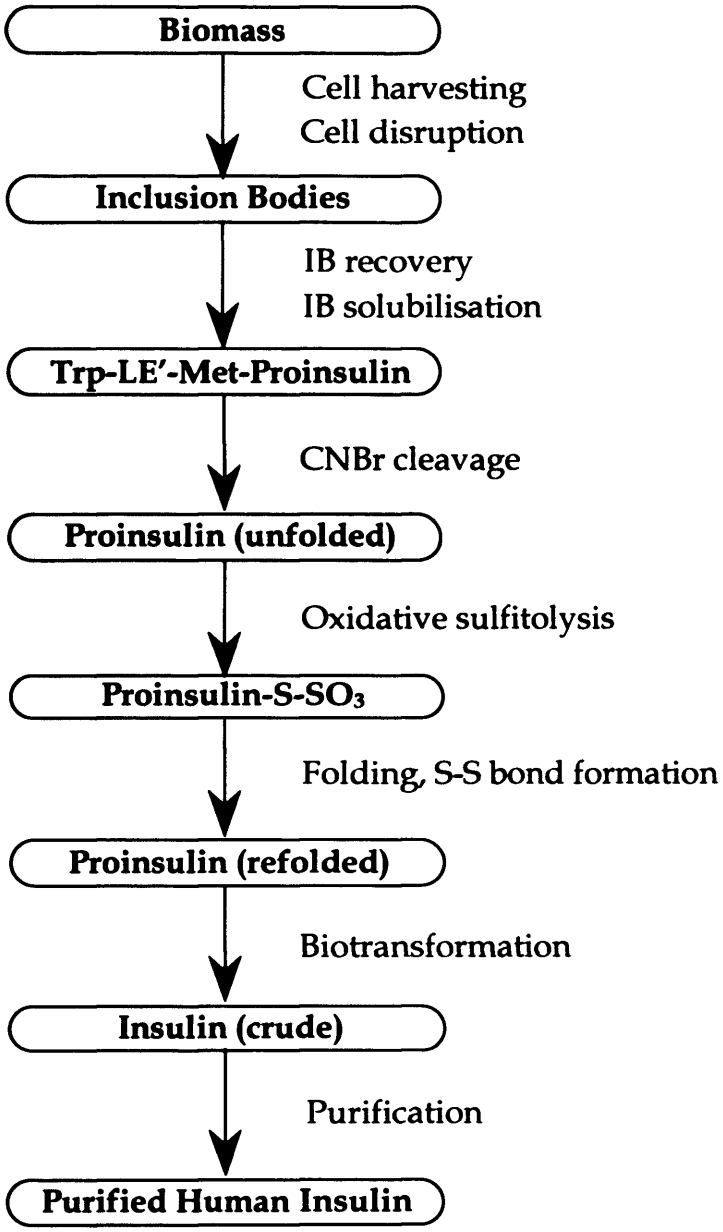

Figure 1-7. Industrial bioprocess for biosynthetic human insulin production.

disulphide bridge formation and correct folding of proinsulin to its mature form. The refolding agent used is mercaptoethanol.

Since mature BHI consists of only two chains A and B, the C chain present in proinsulin must be removed. This is achieved via a biotransformation reaction (see Section 1.2.4) in which trypsin and carboxypeptidase B (CpB) are used, respectively, to cleave the C chain off and remove any C-terminal arginine residues left on chain A. The reaction is quenched after 4 hours. The mixture finally passes through a variety of purification processes designed to separate BHI from contaminants on the basis of its molecular charge (ion exchange chromatography), hydrophobicity (reversed phase chromatography) and size (gel filtration chromatography).

1.2.4 Opportunities for optimisation

The industrial process for the production of BHI may be optimised on several fronts. A variety of both novel and traditional engineering principles may be applied to bioprocess design and optimisation for increased yields and purities. These include the use of operating windows (Salte *et al.*, 2006), computer modelling (Farid *et al.*, 2007), scale-down techniques (Boychyn *et al.*, 2004) and design of experiments (Islam *et al.*, 2007). As valuable and efficient as these methodologies have proved, they cannot address biochemical issues associated with the molecular level of a process.

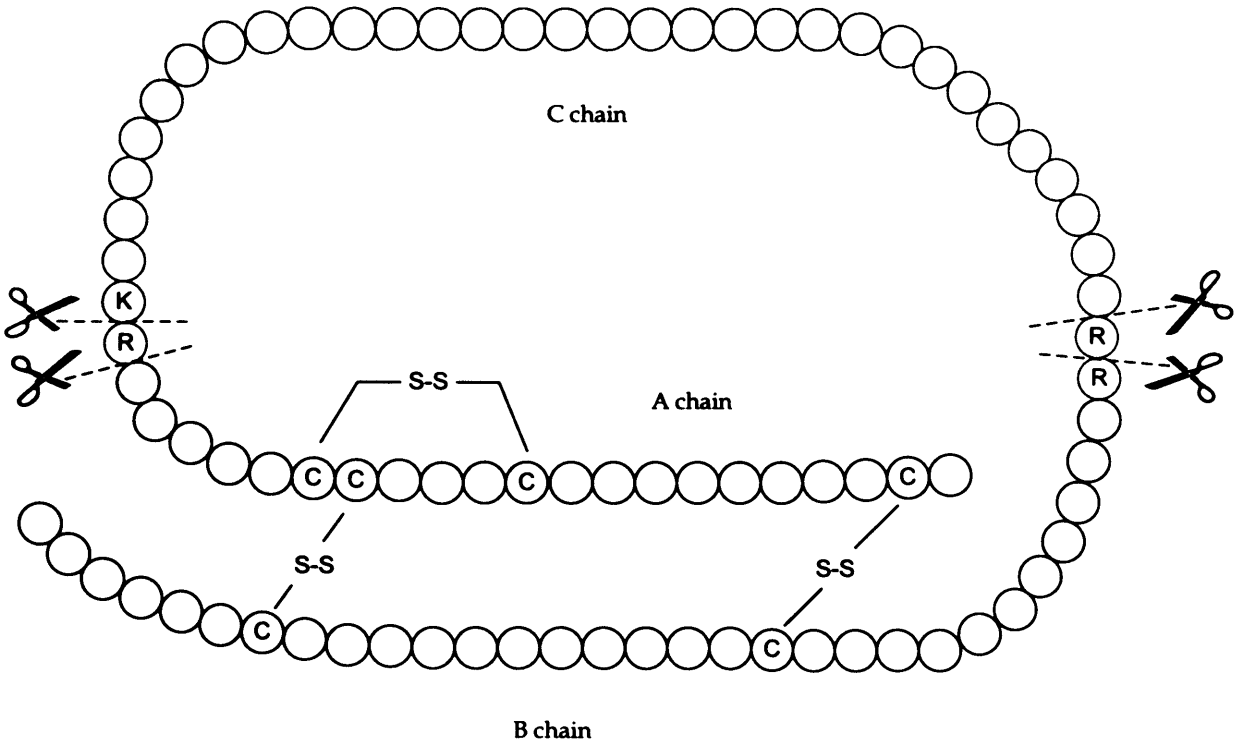

Figure 1-8. Prominent cleavage sites in human pro-insulin. Once the C chain (coloured grey) dissociates, A and B chains (coloured black) fold in the appropriate conformation to yield active insulin. Possible tryptic cleavage sites are indicated by the scissors. Carboxypeptidase B is required to cleave off C-terminal arginine residues left on the B chain. The C chain must be cleaved at the correct place *i.e.* at the C terminal side of the arginine residue to yield active insulin. Cleavage after the lysine residue (grey scissors) leads to the formation of a contaminant molecule. Disulphide bridges (S-S) between the relevant cysteine residues are shown. Note that the arginine or lysine sites unaffected by cleavage, possibly due to the lack of their accessibility to trypsin, are not shown.

In particular, this applies to the biotransformation step for the conversion of the precursor of insulin to active BHI (see **Figure 1-8**).

By definition of trypsin's specificity, cleavage is possible at the C-terminal side of either the arginine or lysine residue (indicated by scissors). Lysine-specific cleavage is not desired as it leaves an N-terminal arginine residue on the A chain. This results in an insulin-like side-product that must be removed in subsequent purification steps. Eli Lilly observe a 10% yield of this contaminant. It is proposed that, by means of enzyme engineering, the specificity of trypsin may be modified to decrease its specificity for lysine residues whilst at the same time retaining arginine specificity. If successful, the engineered trypsin may be phased into the industrial process, leading to an increased product yield in addition to all the associated benefits. The ramifications of such a change on both process economics and validation are discussed in Chapter 8.

1.3 Enzyme engineering

1.3.1 An overview

Used in this context, the term enzyme engineering is used synonymously with protein engineering, that is the application of scientific principles for the development of valuable or useful proteins. Since the focus of this project is on the enzyme trypsin, the term enzyme engineering is used to reflect this. The need for enzyme engineering arises since enzymes in general are not optimised for use in industrial processes. They are naturally

evolved entities that play a part, however critical, in the functioning of their host organism. Since this process has been going on for the last 3 - 4 billion years (by most estimates), the variety of enzymes available to exploit in the present day and age is vast. Protein engineering strategies therefore lean heavily towards methods that fine-tune or tweak a particular characteristic of interest.

This project is concerned primarily with activity and substrate specificity although there is a whole host of other characteristics that may be enhanced including thermostability, solvent stability, catalytic promiscuity, enantioselectivity and resistance to toxic agents examples of which have been reported and reviewed in literature (Morley and Kazlauskas, 2005; Eijsink *et al.*, 2005; Dalby, 2003; Crameri *et al.*, 1997). Historically, two contrasting strategies have been used in protein engineering, "rational design" and "directed evolution", with many variations on these recently reported. A description of the various methods is given in Section 1.3.2 together with the benefits and drawbacks they confer.

There are a number of prerequisites for any enzyme engineering strategy. Firstly, there must be an expression system for the active form of the enzyme; this provides a means for testing enzyme "variants" that may have enhanced functionality. Typically a microbial host is used for expression such as *Escherichia coli* or *Pichia pastoris*. The next requirement is that a screening or selection method (Section 1.3.2.4) must be developed for identifying and isolating mutants of interest. Finally, in order to produce a

"library" of variants, a suitable mutagenesis method must be chosen for creating residue substitutions whether random or targeted.

1.3.2 Strategies

1.3.2.1 Rational design

In this strategy, engineers use available knowledge of an enzyme's structure and function to make desired changes. On the surface this would appear to be an ideal scenario since it is generally inexpensive and easy to implement by using site-directed mutagenesis techniques. However, one of the major drawbacks is that detailed structural information of an enzyme is often unavailable, and even when it is available, it can be difficult to predict mutations that are likely to be beneficial. With so many factors at play influencing structure, this method can usually only be applied to enzymes that have been well characterised. Nevertheless, there have been many recent examples of rational design put to use successfully not only for enzyme stability (Eijsink *et al.*, 2004)[1], but also for the alteration of substrate specificity (Cheon *et al.*, 2004; Suenaga *et al.*, 2002; Whittle and Shanklin, 2001), which arguably requires the most structural information.

[1] This paper gives several examples of enzymes rationally engineered for stability improvements.

1.3.2.2 Directed evolution

1.3.2.2.1 Background

There are two approaches by which a mutant library may be created. Non-recombinative methods are those in which point mutations arise out of misincorporations in the DNA replication process. Conditions are deliberately made favourable for mutagenesis. Recombinative methods involve the breaking and rejoining of DNA in new combinations. The genetic information from one DNA molecule may be lost to another molecule, or the order of genetic information may be changed. Both approaches mimic natural evolution (mutation followed by selection) except that the property to be improved has to be defined before-hand, and also the process takes place in a more practical period of time.

1.3.2.2.2 Non-recombinative methods

The most commonly used method in this category is mutagenic or "error-prone" PCR (epPCR) (**Figure 1-9**). A standard PCR reaction is normally carried out to amplify a given section of DNA with high fidelity (Mullis *et al.*, 1986). The intrinsic $3 \rightarrow 5'$ proof-reading (exonuclease) activity of the polymerase ensures that amplification proceeds to a high degree of accuracy. However, wrong nucleotides are occasionally incorporated during standard PCR, leading to mutations with a frequency of 1.1×10^{-4} per nucleotide in the case of *Thermus aquaticus* (*Taq*) DNA polymerase (Tindall and Kunkel, 1988). The low error-rate during PCR can be increased through several routes

(Cadwell and Joyce, 1995) and used as a mutagenesis method (Zhou *et al.*, 1991): (1) by increasing the concentration of Mg^{2+} ions; (2) by adding Mn^{2+} ions; (3) by increasing the concentrations of dCTP and dTTP; and (4) by increasing the amount of *Taq* polymerase.

A less commonly used option for non-recombinative mutagenesis is by employing a bacterial mutator strain. Wild-type *E. coli* strains naturally exhibit a spontaneous mutation frequency; some strains exhibit a much higher mutation rate if their DNA-repair pathways are compromised. Stratagene Ltd. has created a commercial strain of *E. coli* named XL1-Red that contains mutations in three independent DNA repair pathways (*mutD*, *mutS*, and *mutT*) (Greener *et al.*, 1996). As a result, the strain exhibits a spontaneous mutation frequency of 0.5 mutations per 1000 nucleotides of DNA. To generate a mutant library, a gene is cloned into a suitable plasmid, transformed into *E. coli* XL1-Red, and propagated. There are a number of drawbacks to this method, however: (1) mutations cannot be directed at the target sequence, which may result in unwanted deleterious effects on the plasmid; (2) the number of generations becomes impractically high if multiple mutations are required; and (3) due to the rapid mutation rate, the strain does not remain stable for long periods and therefore cannot be propagated beyond a certain number of generations.

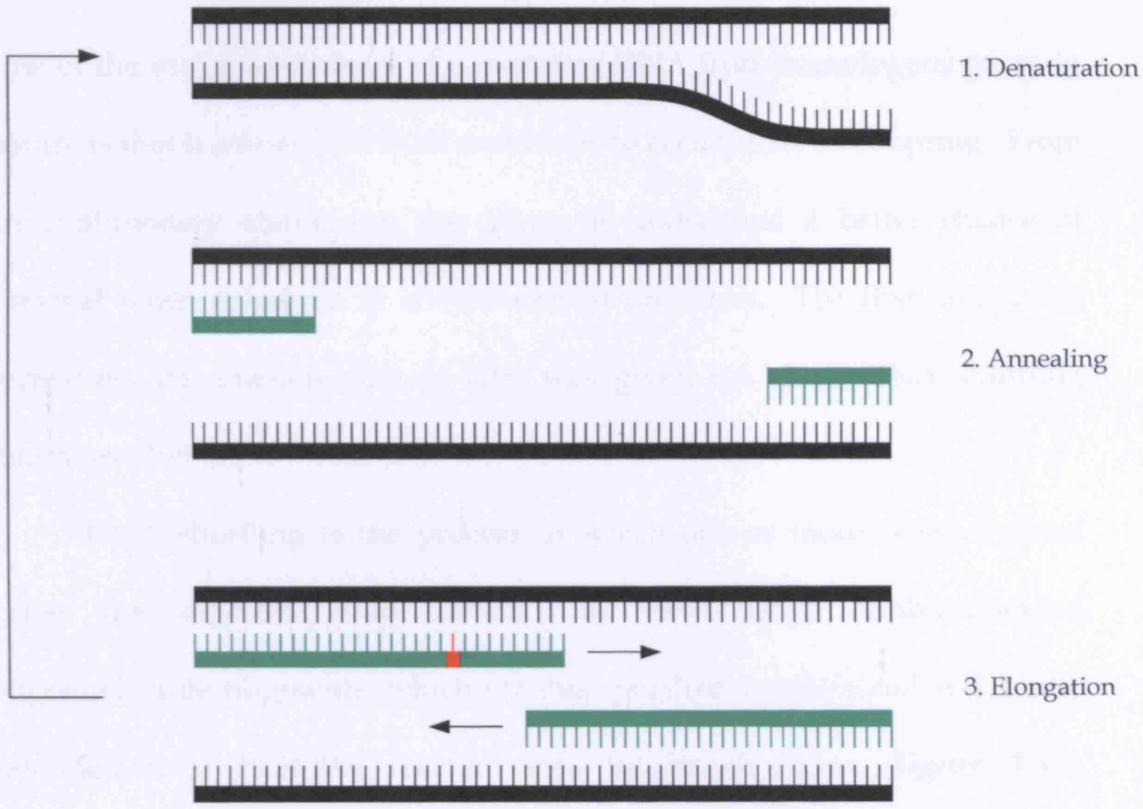

Figure 1-9. Mechanism of an error-prone PCR reaction. (1) Denaturation (94 °C): double-stranded DNA separates to form single-stranded DNA. (2) Annealing (55 °C): a reduction in temperature allows stable bonds to form between the oligonucleotide primers (coloured green) and complementary bases of the template DNA (coloured black). (3) Elongation (72 °C): *Taq* polymerase extends the primers according to the sequence of the template DNA. Occasionally, a mismatched base (coloured red) may incorporate into the growing strand under error-prone conditions. The entire cycle is then repeated with the newly synthesized strands as templates. In this way, the number of copies of the target sequence increases exponentially with mutations at random points.

1.3.2.2.3 Recombinative methods

One of the major advantages of combining DNA from homologous genes in nature is that it allows beneficial mutations to accumulate in offspring. From an evolutionary standpoint this gives an individual a better chance at survival when subjected to environmental pressures. The first account of recreating this phenomenon *in vitro* was given the name DNA shuffling (Stemmer, 1994).

DNA shuffling is the process in which one or more closely-related genes are digested with DNase I to yield small double-stranded oligonucleotide fragments, which are then purified, reassembled and finally extended in a PCR-like reaction into full-length genes (**Figure 1-10**). Recombination occurs by template-switching: fragments from one gene prime at complementary locations on another. The combination of this technique with epPCR or bacterial mutator strains to introduce point mutations allows the simultaneous permutation of both single mutations and large blocks of DNA sequences. Mutants with improved properties can be reshuffled against each other or against the wild-type gene. Eventually, this process may eliminate neutral or deleterious mutations from the gene pool. When homologous genes from different species are used to perform DNA shuffling (family shuffling), the rate of functional enzyme improvement is significantly accelerated since a larger sequence space may be sampled (Crameri *et al.*, 1998). Many variations of the standard DNA recombination protocol have been developed. StEP is a simple and efficient

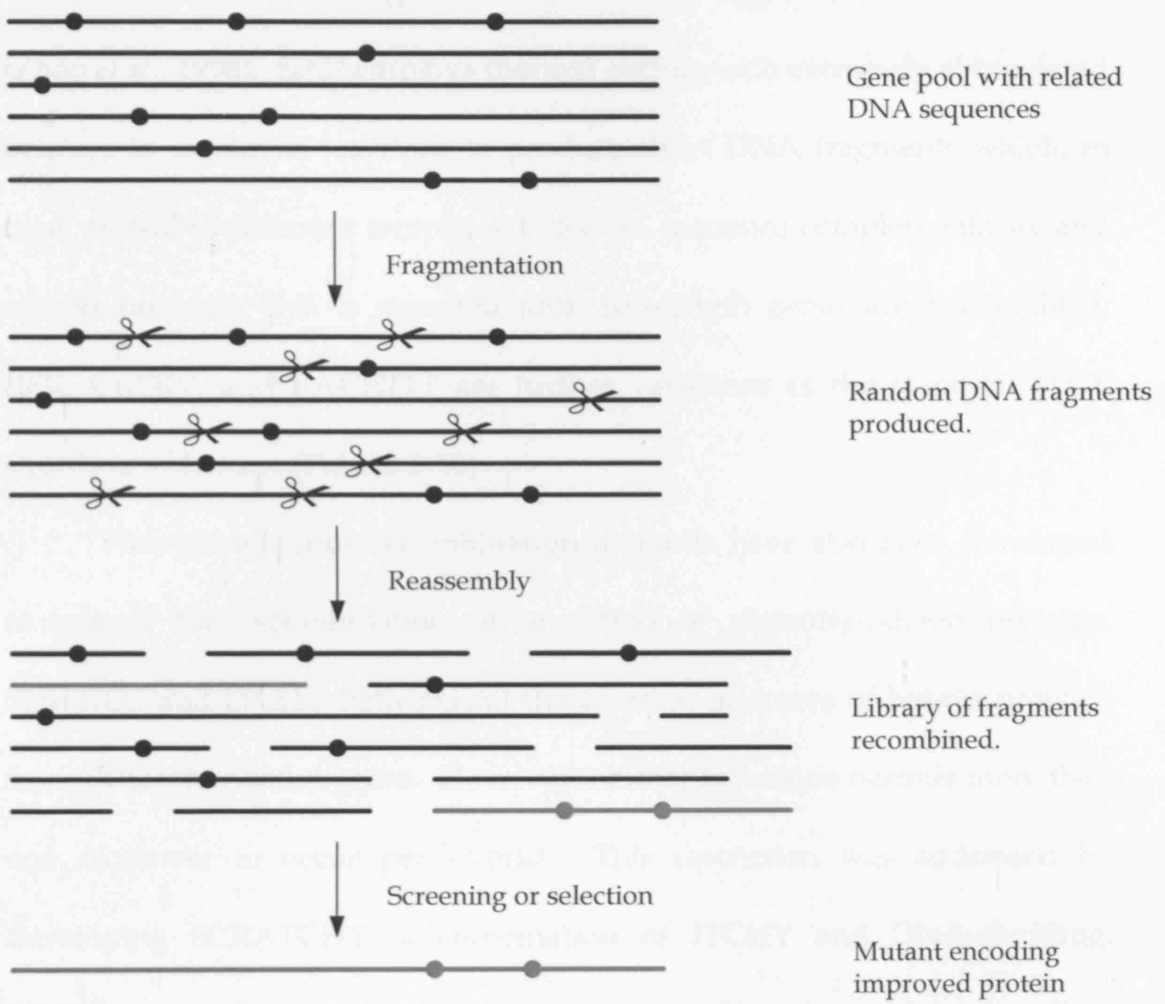

Figure 1-10. DNA shuffling method. Related genes with different beneficial mutations (black dots) are randomly fragmented with DNase I. Fragments are reassembled combining beneficial mutations and isolated (shown in grey). Adapted from Reetz and Jaeger, 1999.

variation of DNA-shuffling that is based on a staggered, PCR-like reaction (Zhao *et al.*, 1998). StEP employs thermal cycling with extremely abbreviated primers in extension reactions to produce short DNA fragments which, in turn, anneal to different templates based on sequence complementarity and extend further. This is repeated until full-length genes are reassembled. RPR, CLERY, and RACHITT are further variations of the standard DNA shuffling technique (**Figure 1-10**).

Non-homologous recombination methods have also been developed to permit the recombination of unrelated or distantly-related proteins. SHIPREC and ITCHY both permit the creation of arrays of hybrid proteins from distantly-related genes. However, neither technique permits more than one crossover to occur per hybrid. This restriction was addressed by developing SCRATCHY, a combination of ITCHY and DNA-shuffling. SCOPE is a different technique to those already mentioned as it utilises protein structural information to guide the generation of multiple-crossover hybrids from non-homologous genes. A summary of the recombinative mutagenesis methods mentioned here is given in **Table 1-3**.

1.3.2.3 Semi-rational methods

Many applications of protein engineering have not relied exclusively on either rational design or directed evolution methods. This section focuses on those methods that can be described as "semi-rational" in that some structural knowledge has been made use of to propose mutation "hotspots"

Method	Traits	Reference
Homologous recombination		
DNA shuffling	Genes with different beneficial mutations are fragmented and reassembled.	Stemmer, 1994
StEP (staggered extension process)	Cycles of short primer extension, re-annealing and extension to full length.	Zhao *et al.*, 1998
RPR (random-priming recombination)	Cycles of primer extension with random-sequence primers, re-annealing and extension to full length.	Shao *et al.*, 1998
CLERY (combinatorial libraries enhanced by recombination in yeast)	Family shuffling *in vitro* followed by *in vivo* homologous recombination in yeast.	Abecassis *et al.*, 2000
RACHITT (random chimeragenesis on transient templates)	Shuffling by annealing randomly-cleaved fragments to full-length template. PCR-free.	Coco *et al.*, 2004
Non-homologous recombination		
SHIPREC (sequence homology-independent protein recombination)	Two non-homologous parent genes are fused end-to-end and randomly fragmented. Single-crossover hybrids are generated by isolating and circularising parent-sized fragments.	Sieber et al., 2001
ITCHY (incremental truncation for the creation of hybrid enzymes)	Nucleotide triphosphate analogues are used to create incremental truncation libraries. Single-crossover hybrids of unrelated or distantly related proteins are generated.	Lutz *et al.*, 2001a
SCRATCHY (ITCHY with DNA shuffling)	Similar to ITCHY but capable of generating multiple-crossover hybrids.	Lutz *et al.*, 2001b
SCOPE (structure-based combinatorial protein engineering)	In a series of PCRs, "hybrid oligonucleotides" act as surrogate introns to direct the assembly of coding segments from two non-homologous genes. Multiple crossovers occur in each hybrid.	O'Maille *et al.*, 2002

Table 1-3. Recombinative mutagenesis methods for application in directed evolution.

in the gene for protein engineering. Mutations are not site-directed in the traditional sense, nor are residues pre-selected to mutate to, thereby adding a random element to the strategy.

In cases whereby structural information of an enzyme is lacking or there is no access to computational programs, the most commonly used semi-rational methods are saturation mutagenesis (SM) (Wells *et al.*, 1985) or combinatorial cassette mutagenesis (CCM) (Reidhaar-Olson and Sauer, 1988). In SM, a single residue may be randomised to any of 19 other possibilities by ligating mutagenic inserts into the host vector. CCM is a simple advance in this technique increasing the region of mutation to two or three residues. Since ligation reactions are notoriously difficult to optimise, PCR-like reactions have taken over as the method of choice for SM and CCM. Working on the same principle as standard PCR, amplification proceeds by the annealing of mutagenic primers containing NNN or NNS codon(s) at the region to be mutated, followed by extension as normal. For both SM and CCM, knowledge of 3-dimensional structure is usually a prerequisite so that mutation hotspots can be proposed such as substrate-binding residues, catalytic residues, cofactor- binding sites and stabilising loops. Many successful examples of applying SM method have been reported (Hibbert *et al.*, 2008; Iyidogan and Lutz, 2008; Mullegger *et al.*, 2005).

Computational protein engineering methods use a variety of predictive tools to propose mutation sites important for enhancing a chosen characteristic. A software package is typically used to predict the effects of

several mutations before narrowing down the residues most likely to yield an improvement if mutated. The mutants are usually expressed and screened experimentally. Large libraries encompassing 2×10^5 mutants, for example, may still need to be screened (Hayes *et al.*, 2002), however, instances of *in silico* pre-screening have been reported in which the number of mutants to be tested experimentally can be narrowed down if their predicted structures are incompatible with protein folds (Ashworth *et al.*, 2006). *De novo* design of an enzyme involves fusing domains or motifs from distinct proteins, with the aid of molecular modelling programs, to yield novel enzymes (Jiang *et al.*, 2008; Ikawa *et al.*, 2004). A summary of the contrasting directed evolution approaches with different degrees of rationalisation is given in **Table 1-4**.

1.3.2.4 Screening and selection

Following the creation of a mutant library, a screening or selection step is used to isolate the best variant(s) for the next cycle. Both approaches typically require a robust system for microbial expression of the mutant genes (see Section 1.1.3.6) in order to identify and isolate mutants of interest. "Screening" is the process of identifying a member or members of a library with outstanding performance in a desired characteristic. Each variant may be assayed for enhanced substrate specificity, for example, under a specific set of conditions. Over the course of an individual research project lasting 3 - 4 years, the realistic

Method	Traits	Reference
In vivo		
Mutator strain: *E. coli* XL1-Red	Mutations in three independent DNA repair pathways cause high levels of spontaneous mutation in a transformed plasmid. Random mutagenesis.	Greener *et al.*, 1996
In vitro		
Error-prone PCR (epPCR)	Mutagenic reaction conditions induced during PCR resulting in random point mutations. Random mutagenesis; semi-rational when mutations are focussed on a region (fepPCR).	Zhou *et al.*, 1991
Saturation mutagenesis	A single codon in a gene is randomised by using mutagenic primers for a single position (NNN). Semi-rational mutagenesis.	Wells *et al.*, 1985
Combinatorial cassette mutagenesis	Two or three consecutive residues randomised by saturation mutagenesis. Semi-rational mutagenesis.	Reidhaar-Olson and Sauer, 1988
Recombinative methods	see Table 1-3. All random mutagenesis.	
In silico		
Computational protein design (semi-rational mutagenesis)	Following an *in silico* screen, an enzyme was redesigned with enhanced specificity.	Ashworth *et al.*, 2006
	Mutations predicted via computer programs in combination with experimental screening.	Dwyer *et al.*, 2004
	Active site redesigned and mutants screened experimentally.	Hayes *et al.*, 2002
	Two domains from distinct but related enzymes fused. *In vivo* protein-folding screen used.	Chevalier *et al.*, 2002
De novo design	Using four different catalytic motifs, several mutants were designed and screened. Semi-rational mutagenesis.	Jiang *et al.*, 2008
	Molecular modelling to create an artificial enzyme. Rational design.	Ikawa *et al.*, 2004

Table 1-4. Directed evolution approaches utilising random, rational and semi-rational techniques.

maximum number of variants that may be screened at high-throughput is in the order of 10^4 - 10^6.

If "selection" is applied instead of screening, then only the desired member of a library appears. It is an efficient method in that it only allows those microorganisms to grow which express a gene encoding a mutant enzyme necessary for survival. Selection processes allow the experimenter to examine libraries up to the order of 10^6 - 10^{12} members. Although this suggests an increased chance of finding the desired enzyme there are some potential drawbacks. Microbial cells are extremely versatile and have the ability to circumvent restrictions imposed by a selection pressure (Reetz and Jaeger, 1999). When available, however, they can deliver dramatic results. In one study, aldolase mutants were transformed into pyruvate kinase deficient cells thereby lacking the ability to generate their own pyruvate for survival. However, a non-natural substrate of the aldolases present in the medium rescued cell growth by releasing pyruvate upon its cleavage. Only those mutants with novel substrate specificity demonstrated this behaviour (Griffiths *et al.*, 2004).

When the desired phenotype cannot be easily linked to cell survival or growth, a screen is still the only viable option. Screens are more versatile than selections with a variety of commercial and synthesized substrates, and different spectrophotometric markers (e.g. fluorometric, colourimetric and luminescent) available for assay development. Reactions that take place in

microwell plates - allowing for spectrophotometric detection of the marker - are still the method of choice in the development of screens.

Phage-display (Sidhu *et al.*, 2000) is an *in vitro* selection method that tests protein-protein and protein-DNA interactions by integrating many genes from a DNA library into the genome of a suitable phage. By immobilising a relevant DNA or protein target(s) to the surface of a microwell, a phage that displays an enzyme that binds to one of those targets on its surface will remain while others are removed by washing. Those that remain can be eluted, used to produce more phage, and finally produce a phage mixture that is enriched with the phage (and hence gene) of interest. Phage eluted in the final step can be used to infect a suitable bacterial host, from which the relevant DNA sequence can be excised and sequenced to identify the enzyme of interest.

An alternative selection method exists called "*in vitro* compartmentalisation" (IVC) (Tawfik and Griffiths, 1998) that dispenses with the need for a microbial expression system and selects mutants of interest. Cells are usually employed to keep together the genes, the RNAs and proteins that they encode, and the products of their activities, thus linking genotype to phenotype. This was reproduced in the test tube by transcribing and translating single genes in the aqueous compartments of water-in-oil emulsions. These compartments, with volumes close to those of bacteria, were recruited to select genes encoding enzymes. An enzyme with a desired catalytic activity converted a substrate, attached to the gene that

encoded it, to product. In other compartments, substrates attached to genes that did not encode catalysts remained unmodified. Subsequently, genes of interest were selectively enriched by virtue of their linkage to the product.

Using IVC, it is possible to select very large libraries of between 10^8 - 10^{11} genes (Miller *et al.*, 2006). Applications of this method have been reported in literature: A variant of *Hae*III methyltransferase was evolved with up to 670-fold improvement in catalytic efficiency for a novel target sequence (AGCC) and 9-fold improvement for the original recognition site (GGCC) (Cohen *et al.*, 2004). Active *Fok*I restriction nuclease variants were also selected by IVC from a large library of mutants with low residual activity (Doi *et al.*, 2004). Additionally, DNA polymerase variants with increased thermal stability were evolved (Ghadessy *et al.*, 2001).

1.4 Project aims

Bovine trypsin is an industrially useful protease with a specific role to play in the processing of human insulin. Given that the enzyme does not act with a high degree of specificity, however, this leads to a loss of yield during the biotransformation step, with a knock-on effect on the rest of the process. The goal of this project was to enhance the specificity of bovine trypsin by means of various enzyme engineering strategies. The wider aims at each stage of the project may be divided and summarised as follows:

Chapter 3 - Random mutagenesis of trypsin

- To test the viability of the pET(T) plasmid for expression of active bovine trypsin.
- To generate a library of bovine trypsin variants by "focussed" epPCR and to evaluate this method as a means of introducing random mutations into gene sections of varying length.
- To evaluate MEGAWHOP (megaprimer PCR of whole plasmid) as a means of re-constructing a plasmid vector around gene sections of varying length.
- To screen the library of variants, at high-throughput, for improvements in specificity towards either arginine or lysine residues.

Chapter 4 - Targets for enzyme engineering

- To establish relationships between (1) residue location, (2) sequence entropy and (3) activation free energy changes, and their influence on enhancing activity and specificity, by analysing past examples of engineered enzymes.
- To identify target sites for engineering the specificity of trypsin and its resistance to self-proteolysis.

Chapter 5 - Targeted mutagenesis of trypsin

- To direct mutations at sites important to the self-proteolysis of bovine trypsin in order to stabilise the molecule in solution.

- To mutate, by MSSM, sites pertaining to the specificity of bovine trypsin and evaluate this method as a tool for the generation of large libraries (> 10^6).
- To develop electrocompetent cells from an auxotrophic strain of *E. coli* for use as a selection tool.
- To select mutants with novel exoprotease activity, directed at N-terminal arginine or lysine residues, by means of a nutritional selection.

Chapter 6 - Characterising the self-proteolysis of trypsin

- To generate bovine trypsin variants by self-proteolytic cleavage and separate these using various chromatographic methods.
- To characterise the variants with respect to enzyme kinetics, molecular weight and on the ability to transform human pro-insulin, and compare the results with commercial bovine trypsin and Eli Lilly's recombinant trypsin.

2 Materials and methods

2.1 General notes

The procedures described in this section are standard methods. Protocols developed for particular applications are described in Chapters 3 - 6. All materials were ordered from Sigma-Aldrich and in the UK unless otherwise stated.

2.2 Laboratory equipment

Water was purified to between 5 - 15 MΩ.cm resistivity by de-ionisation using an Elix 5 water purifier (Millipore, UK). Items to be sterilised were autoclaved to 15 psi for 20 minutes (Thermo Electron, UK). pH adjustments were carried out using 1 M solutions of either NaOH or HCl and verified using a pH electrode and meter (Thermo Scientific, UK). Room temperature was maintained at between 24 - 25 °C in all laboratories. Exceptions to the above are stated in the relevant sections.

2.3 Preparation of media, buffers and reagents

2.3.1 Luria Bertani (LB) medium

LB medium was made by dissolving the following in pure water: 10 $g.L^{-1}$ tryptone, 10 $g.L^{-1}$ sodium chloride and 5 $g.L^{-1}$ yeast extract. Concentrated sodium hydroxide solution was used to adjust the pH to 7.0 when necessary. The medium was sterilised by autoclaving.

2.3.2 M9 minimal medium

M9 medium was made by dissolving the following in pure water: 6 $g.L^{-1}$ sodium hydrogen phosphate, 3 $g.L^{-1}$ potassium dihydrogen phosphate, 1 $g.L^{-1}$ ammonium chloride, 0.5 $g.L^{-1}$ sodium chloride, and 15 $mg.L^{-1}$ calcium chloride. Concentrated sodium hydroxide solution was added to adjust the pH to 7.0 when necessary. The medium was sterilised by autoclaving. Once cooled to room temperature the following supplements were added: 5 mL of 40% (v/v) glycerol solution, 1 mL of 1 M magnesium sulphate solution and thiamine at a final concentration of 0.1 $mg.L^{-1}$. Where arginine was supplemented, this was added to a final concentration of 500 µM and is denoted by arg^+.

2.3.3 Agar plates

LB and M9 agar media were prepared as above but with supplementation of 15 $g.L^{-1}$ agar before autoclaving. Once cooled to between 40 and 50 °C, 20 mL of media was poured into each standard-sized petri dish or 250 mL into each bio-assay dish.

2.3.4 Antibiotics

Where used kanamycin and streptomycin were dissolved in pure water, filter sterilised and added to a final concentration of 40 $mg.L^{-1}$ either to LB or M9 media or to agar medium before it had set. Stocks were stored at -20 °C.

Supplementation of kanamycin is denoted by kan^+ and streptomycin by $strep^+$.

2.4 Standard procedures

2.4.1 Overnight cultures

A single colony was picked from an agar plate into 5 mL of medium in a 50 mL Falcon tube. The tube was incubated for 16 hours at 37 °C with 220 rpm agitation.

2.4.2 Shake flask cultures

1 mL of an overnight culture was added to 99 mL of medium in a sterile 1000 mL shake flask. Shake flasks were incubated for 16 hours at 37 °C with 220 rpm agitation.

2.4.3 Glycerol stocks

A 20% (v/v) glycerol stock was prepared by adding filter-sterilised or autoclaved 40% (v/v) glycerol to an overnight culture in a one to one volume ratio. Aliquots were stored at -80 °C.

2.4.4 Purification of plasmid DNA

Plasmid DNA was purified from 5 mL of an overnight culture using the QIAprep Spin Miniprep Kit (QIAGEN Ltd.). The following steps were carried out in accordance with the manufacturer's protocol: (1) alkaline lysis

of cells; (2) lysate clearing; (3) adsorption of DNA to the QIAprep membrane; (4) washing of DNA; and (5) elution of pure plasmid DNA. The concentration of purified DNA in the product was determined by measuring its absorbance at 260 nm (A_{260}). One A_{260} unit corresponds to 50 $\mu g.mL^{-1}$ dsDNA at neutral pH for a path length of 1 cm. Plasmid DNA samples were stored at -20 °C in 1.5 mL centrifuge tubes.

2.4.5 Transformation by heat-shock

Chemically competent *E. coli* cell strains BL21-Gold(DE3) and XL1-Blue strains were used for all heat-shock transformations (both supplied by Stratagene Ltd.). The relevant competent cells were thawed on ice and 40 µL were transferred to a chilled 15 mL Falcon tube. The plasmid DNA sample was thawed on ice; 1 µL (50 - 150 ng) was added to the cells and mixed gently. The DNA/cells mixture was incubated on ice for 30 minutes. Heat-shock was performed by holding the Falcon tube in a 42 °C water bath for 40 - 45 seconds (three-quarters length submerged). The transformed cells were then incubated on ice for a further 2 minutes. 950 µL of preheated (37 °C) growth medium was added to the tube which was then incubated at 37 °C for 1 hour with 220 rpm agitation. Between 100 and 200 µL of the mixture was transferred to an agar plate containing the relevant medium and antibiotic, spread out using a disposable L-shaped spreader (VWR International) and incubated at 37 °C for 16 - 18 hours.

2.4.6 Transformation by electroporation

All electroporations were carried out using a MicroPulser® electroporator (Bio-Rad Laboratories). Electrocompetent cells were prepared either in the laboratory (Section 5.2.4) or ordered from Invitrogen (*E. coli* TOP10). The electroporation chamber and cuvette with gap width 0.1 cm (Bio-Rad Laboratories) were incubated on ice for 20 minutes prior to electroporation. The relevant competent cells were thawed on ice and 50 µL were transferred to a chilled 1.5 mL centrifuge tube. The plasmid DNA sample was thawed on ice; 1 µL (50 - 150 ng) was added to the cells, mixed gently and incubated on ice for 2 minutes. The DNA/cells mixture was transferred to a pre-chilled electroporation cuvette, shaken to the bottom and placed in the electroporation chamber. An electrical pulse (1.8kV for bacteria) was passed through the sample. 900 µL of LB medium was immediately added to the electroporation cuvette. The mixture was transferred to a 15 mL Falcon tube and incubated for 1 hour at 37 °C (3 - 4 hours for auxotrophic cell strains). 50 µL of the mixture was transferred to an agar plate containing the relevant medium and antibiotic, spread out using a disposable L-shaped spreader and incubated at 37 °C for 16 - 18 hours.

2.4.7 Measurement of absorbance and optical density

Absorbance and optical density measurements were performed in a UV2 spectrophotometer (Unicam Ltd.). An ultra-micro quartz cuvette was used when measuring absorbance in the ultraviolet region (200-400nm

wavelength) whilst 1 mL or 3 mL cuvettes were used for all other measurements. Each sample was diluted sufficiently to register an absorbance of ≤ 1 AU (or an optical density of ≤ 1 ODU). Absorbance of protein was measured at 280 nm and protein concentration was calculated using the Beer-Lambert Law: Abs = ε c l, where ε = molar absorption coefficient (14.3 $L.M^{-1}cm^{-1}$ for trypsin), c = concentration ($g.L^{-1}$) and l = path length (cm). Absorbance of DNA was measured at 260 nm (OD of 0.1 for double-stranded DNA corresponds to a concentration of 5 $ng.\mu L^{-1}$).

2.4.8 Agarose gel electrophoresis

A GNA-100 system (Amersham Biosciences Ltd.) was used for agarose gel electrophoresis of DNA. The amount of agarose used in a gel was either 0.6% (w/v) for loading 0.8–10.0 kbp fragments (plasmid DNA) or 1.5% (w/v) for loading 800–2000 bp fragments (PCR products).

The appropriate amount of agarose was dissolved in 50 mL of 1 × TAE buffer (40 mM Tris-acetate and 1 mM EDTA in pure water) by heating in a microwave. 0.5 $mg.L^{-1}$ ethidium bromide was added to the gel to promote the visualisation of DNA by UV light. A comb was inserted at one end of the gel well to allow the formation of miniature sample wells. Once cooled to between 40 - 50 °C, the gel was poured into the appropriate well for setting. After the gel had set the comb was removed and submerged in 1 × TAE buffer in the gel tank.

Samples were prepared for loading by adding 6 × loading buffer. The wells of sample lanes were loaded with 3 - 5 µL of sample. The wells of marker lanes were loaded with 1.5 µL of Novagen 0.5 - 12.0 kbp Perfect DNA Markers (EMD Biosciences Inc.) or Novagen 50–2000 bp PCR Markers (EMD Biosciences Inc.). Electrophoresis was performed at 65 V for 45 minutes. The gel was visualised and photographed using a Gel Doc 2000 system (Bio-Rad Laboratories).

2.4.9 DNA sequencing

Cycle sequencing (Sambrook *et al.*, 1989) was performed by the DNA sequencing service of the Wolfson Institute for Biomedical Research (UCL). All DNA samples were given to the service at the requested concentration of 50 ng.µL^{-1} (specific for a 6 kbp dsDNA fragment). Primers were designed to bind no less than 20 bp from the intended site of sequence identification to provide a buffer for initial sequencing inaccuracies. All primers were ordered from Operon Biotechnologies and diluted to the requested concentration of 4 pmoles.µL^{-1} for the sequencing service.

3 Random mutagenesis of trypsin

3.1 Introduction

Two strategies for improving an enzyme's properties have been described in Chapter 1: directed evolution (Section 1.3.2.2) and rational design (Section 1.3.2.1). Rational protein design is suitable only in cases where information is available concerning structure-function relationships of particular residues. This requirement appears to have been satisfied given that the structure of bovine trypsin has been elucidated (Bode and Schwager, 1975) and studies on its specificity carried out (Perona and Craik, 1995; Craik *et al.*, 1985a). Attempts to rationally engineer the specificity of trypsin, however, have met with limited success in the past (Evnin *et al.*, 1990; Craik *et al.*, 1985b) through which substrate specificity was enhanced at the expense of overall activity. More recent examples of the rational design of bovine trypsin have not been found.

Directed evolution strategies that utilise a random mutagenesis method for library construction offer a more subtle approach to enzyme engineering that is far less likely to disrupt the catalytic activity of an enzyme given the abundance of sequence space on offer. Furthermore, random mutagenesis requires far less deliberation of the location and type of mutation in order to achieve a goal. Recent examples of directed evolution using random mutagenesis methods to enhance specificity are plentiful utilising both epPCR (Mueller-Cajar *et al.*, 2007; Jennewein *et al.*, 2006; Fujii *et*

al., 2005; Otten *et al.*, 2002) and DNA shuffling (Williams *et al.*, 2003; Broo *et al.*, 2002).

In theory, random mutagenesis by, for example, epPCR can allow for an average of 5.7 out of 19 possible amino acid substitutions at a particular site due to the conservatism of the genetic code (assuming a mutation rate of one nucleotide per codon). The sequence range targeted is far greater with random rather than targeted mutagenesis methods resulting in a greater number of mutations further away from the active site. Morley and Kazlauskas (2005) have reported this to be detrimental to the chances of producing specificity-enhancing mutants, which would seem to contradict the aims of the project. However, random mutagenesis of an entire gene has continued to produce beneficial distant mutations (Jennewein *et al.*, 2006; Fujii *et al.*, 2005).

Developments in epPCR techniques have also dictated that mutagenesis products no longer need to be ligated into a vector but may be inserted via a novel method, MEGAWHOP, instead (Miyazaki and Takenouchi, 2002). In addition, fragments of variable length may be inserted, meaning that different sections of the trypsin gene may be targeted. Based on this principle, focussed epPCR (fepPCR) of particular gene sections was successfully demonstrated for the construction of mutant transketolase libraries (Miller, 2004). It was decided that this approach would constitute the first attempt at enhancing the substrate specificity of bovine trypsin.

In addition to a chosen mutagenesis method, the following are prerequisites for any enzyme engineering strategy: (1) a gene encoding the enzyme of interest; (2) a suitable expression system; and (3) a sensitive high-throughput screen (or selection method) to identify the best performing mutant(s). A wild-type plasmid was constructed and a crude high-throughput screen developed in a previous research project at the Department of Biochemical Engineering, UCL (Hibbert, 2003). The wild-type plasmid was created by ligation of the bovine trypsin gene (originally supplied by Eli Lilly Inc, Fegersheim) into a pET26b(+) host vector and re-named pET(T) (**Figure 1-6**). This was donated for use as a starting point of this project.

The pET26b(+) vector was chosen on the basis that it benefited from several advantages over other vectors. Firstly, kanamycin resistance was considered to be more robust than ampicillin resistance, which usually requires 2 - 3 fold higher concentrations for the selection of transformants on agar medium. Secondly, the *pelB* leader sequence immediately upstream of the multiple cloning site acts to localise recently translated protein to the periplasm avoiding the potential for inclusion body formation within the cell. Finally, the vector has tight control over protein expression levels which was under the influence of the inducible T7 promoter gene.

Traditional microplate-based screening methods were deemed suitable for fepPCR due to the workable library size. The high-throughput screen consisted of the following four steps: (1) transformation of pET(T) into

E. coli BL21-Gold(DE3) competent cells; (2) fermentation of individual colonies in 96-well microplates; (3) lysis of overnight cultures; and (4) spectrophotometric activity assays using a plate reader. A modified version of this screen was finally used. This chapter describes the results of the first attempt at improving trypsin specificity via: (1) fepPCR of six variable trypsin regions; (2) construction of mutant trypsin libraries; and (3) screening at high-throughput.

3.2 Materials and methods

3.2.1 Induced or inhibited 5 mL cultures

Individual colonies of BL21-Gold(DE3) hosting either pET(T) or pET26b(+) were picked to inoculate 5 mL of LB kan^+ medium in 50 mL Falcon tubes. Cultures were incubated at 37 °C with shaking at 200 rpm. Induction was after 4 hours using 0.4 mM final concentration of IPTG. To show the measured activity was trypsin, the serine protease inhibitor, AEBSF (4-[2-aminoethyl] benzenesulfonyl fluoride hydrochloride), was added at the induction time at a concentration of 7 µL per 5 mL culture. Cultures were removed after 18 hours of incubation.

3.2.2 Verification of trypsin activity in microwells

Cultures from the previous section were lysed by the 1:1 addition of BugBuster (Novagen). The activity of lysate containing pET(T) or the pET26b(+) control was measured on the substrate L-BANA (Bz-arg-pNa).

10 µL of cell lysate was added to 0.4 mM L-BANA in 50 mM Tris-HCl, pH 8.0, 20 mM $CaCl_2$, 1% DMF (dimethyl formadide) at a final assay volume of 200 µL per well. Absorbance increases were monitored at 405 nm using a Fluostar Optima (BMG Labtech Ltd, UK) microplate reader over a 24-hour linear range.

3.2.3 Library creation by fepPCR

3.2.3.1 Primer design for fepPCR

Six regions of variable length were targeted as shown in **Figure 3-1**. The software program AnnHyb 4.930 (Friard, 2005) was used to ensure the physical properties of the primers were conducive to PCR. In particular, the following rules of thumb were observed: (1) lengths of primers were kept below 40 bp; (2) GC content of the primers were kept to between 40 - 60% ; and (3) where possible GC content was maximised in the last five bases of the 3' end of primers without going above three of each (also known as a GC clamp). A further rule of thumb conveyed that the melting temperature (T_m) of a primer should be kept below 65 °C.

3.2.3.2 fepPCR amplification of bovine trypsin gene

Six fepPCR reactions were set up. All reaction components were ordered from Stratagene Ltd except dNTP mix (Roche Diagnostics Ltd, UK). A 50 µL fepPCR reaction was made up in a sterile 0.5 mL centrifuge tube and incubated on ice. The reagents were added as shown in **Table 3-1**. A

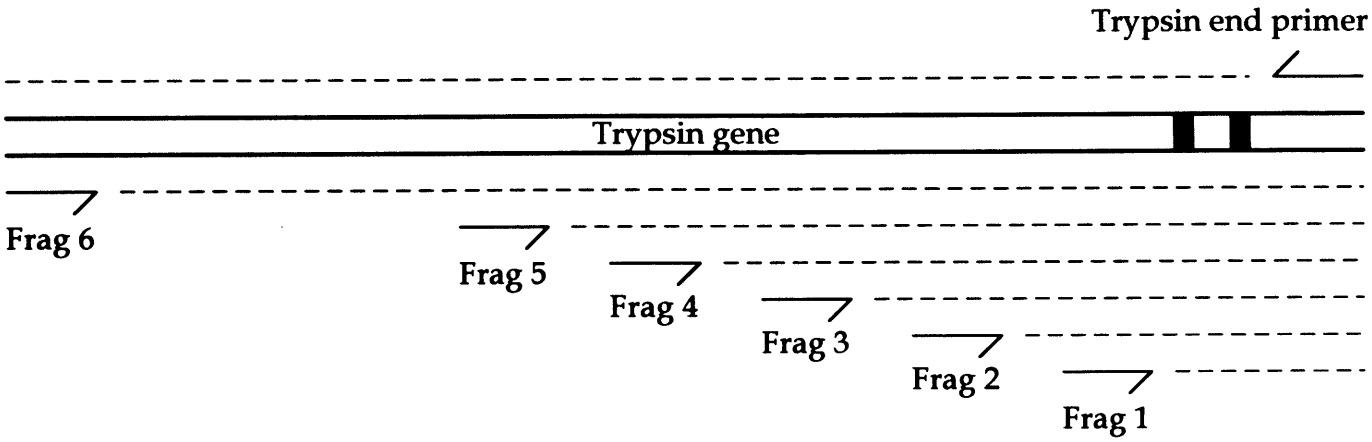

Figure 3-1. Regions targeted for fepPCR. The regions targeted were: Frag 1 - 119 bp; Frag 2 - 222 bp; Frag 3 - 315 bp; Frag 4 - 433 bp; Frag 5 - 522 bp; and Frag 6 - 963 bp. All six fragments included the purported active site residues (indicated by the shaded regions).

Reagent	Volume (µL)	Final amount/ concentration
Sterile pure water	40	-
10 × *Taq* reaction buffer	5	-
dNTP mix	1	not given
Trypsin end primer	0.5	250 ng
Trypsin frag N primer	0.5	250 ng
pET(T) plasmid template	1	150 ng
Taq DNA polymerase	1	2.5 U
$MnCl_2$	1	0.12 - 0.70 mM
Total	50	

Table 3-1. The components of an fepPCR reaction. Trypsin end primer is the end point of all fragments to be amplified. Trypsin frag N primer denotes a variable primer with different starting points to vary the length of fragment to be amplified. Concentrations of $MnCl_2$ in the reaction are shown in **Table 3-2**.

Fragment name	Fragment size (bp)	$[Mn^{2+}]$ (mM)	$[Mn^{2+}] \times 1.2$
F1	119	0.84	0.70*
F2	222	0.45	0.54
F3	315	0.32	0.38
F4	433	0.23	0.28
F5	522	0.19	0.23
F6	963	0.10	0.12

Table 3-2. Standard and amended concentrations of Mn^{2+} ions used in fepPCR reactions. Standard Mn^{2+}concentrations were multiplied by a factor of 1.2. *The maximum practical Mn^{2+}concentration used in a fepPCR reaction was capped at 0.70.

Phase	Cycles	Step(s)	Temperature (°C)	Duration (min)
1	1	Initial denaturation	95	0.5
2	30	Denaturation	95	0.5
		Annealing	55	0.5
		Extension	72	1:5
3	1	Final extension	72	10.0
4	1	Final hold	4	indefinitely

Table 3-3. Program of temperature cycling for a fepPCR reaction.

Reagent	Volume (µL)	Final amount/ concentration
Sterile pure water	40	-
10 × *Pfu Turbo®* reaction buffer	5	-
dNTP mix	1	not given
fepPCR products (Frag 1 - 5)	1	~ 150 ng
pET(T) plasmid template	1	150 ng
Pfu Turbo® DNA polymerase	1	2.5 U
DMSO	2	4%
Total	50	

Table 3-4. The components of a MEGAWHOP reaction. The fepPCR products from single fepPCR reactions (Frag 1, 2, 3, 4 and 5) were used as forward and reverse primers for the amplification of the pET(T) plasmid.

Phase	Cycles	Step(s)	Temperature (°C)	Duration (min)
1	1	Initial denaturation	95	0.5
2	24	Denaturation	95	1.0
		Annealing	65	3.0
		Extension	70	21.0
3	1	Final extension	70	10.0
4	1	Final hold	4	indefinitely

Table 3-5. Program of temperature cycling for a MEGAWHOP reaction.

breakdown of final concentrations of Mn^{2+} is given in **Table 3-2**. The Eppendorf was transferred to a TechGene thermal cycler (Techne Ltd, UK) and submitted to a program of temperature cycling (**Table 3-3**). An agarose gel electrophoresis (Section 2.4.8) was carried out to confirm that the reaction had generated detectable reaction products.

3.2.3.3 *Whole plasmid PCR with fepPCR products as primers*

A PCR reaction was set up in the same way as described in Section 3.2.3.2 based on the MEGAWHOP (megaprimer PCR of whole plasmid) principle (Miyazaki and Takenouchi, 2002). The components of the reaction are shown in **Table 3-4**. The program of temperature cycling is shown in **Table 3-5**. The MEGAWHOP reaction products were digested with 10 U *Dpn* I for 2 hours at 37 °C. An agarose gel electrophoresis (Section 2.4.8) was carried out to confirm that the reaction had generated detectable reaction products.

3.2.3.4 *Verification of fepPCR mutation rate*

The MEGAWHOP reaction products from frag1 and frag5 primers were electroporated into TOP10 electrocompetent cells (Section 2.4.6). Twenty 5 mL overnight cultures (Section 2.4.1) were carried out in LB kan^+ media. Plasmid DNA was purified from these cultures (Section 2.4.4) and sequenced (Section 2.4.9). The DNA sequences returned were analysed in BioEdit for mutations.

3.2.3.5 Amplification of library DNA

MEGAWHOP reaction products were electroporated into TOP10 electrocompetent cells (Section 2.4.6). 2 mL of LB medium was expelled onto the surface of an agar plate with colonies. The colonies were lightly scraped off using a disposable L-shaped spreader and the mixture transferred to a 50 mL Falcon tube. The same procedure was followed for ten plates with colonies and the mixture pooled in the same Falcon tube. Plasmid DNA from this mixture was purified (Section 2.4.4).

3.2.3.6 Generation of library plates

Amplified DNA from the previous section was transformed into BL21-Gold(DE3) cells by heat-shock (Section 2.4.5). The following procedure was adapted from a method previously developed in the Department of Biochemical Engineering, UCL (Miller, 2004). Each well of eight 384 square-well microplates was filled with 60 µL of LB kan$^+$ medium. A QPix2 robot (Genetix Ltd, UK) was programmed to inoculate all but two wells of the microplates with colonies of BL21-Gold(DE3) hosting library DNA. Wells A1 and A2 were reserved for manual inoculation with BL21-Gold(DE3) cells hosting wild-type pET(T) DNA. Each plate was sealed by taking an identical empty plate, inverting it, placing it flush over the top of an inoculated microplate and taping around the edges. The sealed plates were incubated for 18 hours at 1400 rpm on a Variomag Teleshake unit (Camlab Ltd, UK) in a 37 °C incubator. These plates were designated the master library plates

and glycerol solution added to a final concentration of 20% (Section 2.4.3). Reaction plates were generated by replication of the master plates using a program on the QPix2 robot for inoculating microplates containing the appropriate medium and growing for 18 hours. Replicated (reaction) plates were stored at -80 °C.

3.2.4 High-throughput screen of fepPCR library (primary)

A reaction plate containing approximately 30 µL of culture[1] was removed from the -80 °C freezer and defrosted fully at room temperature. The freeze-thaw action constituted the lysing of cells (Miller, 2004). Activity assays were carried out in the wells of a 384 square-well microplate (Fisher Scientific) on the fluorescent substrates Bz-arg-AMC and Ac-lys-AMC (both supplied by Bachem, Switzerland). 30 µL of a substrate solution (2x concentration) was added to each well containing ~ 30 µL of cell lysate. The final concentrations of assay constituents were as follows: 0.2 mM substrate (separately dissolved in 5% DMF), 50 mM Tris-HCl, 20 mM $CaCl_2$ and 1% DMF resulting in a final pH of 8.0. A program was set up on the Fluostar Optima plate reader to measure fluorescence at 340 nm excitation and 450 nm emission every hour for 4 hours. Prior to each reading the reader was set to shake orbitally for 10 seconds.

[1] Evaporation effects during fermentations resulted in a loss of approximately 30 µL per well.

3.2.5 Secondary and tertiary screens

Colonies of interest were picked from master library plates and grown up in 96 deep-square-well microplates (Fisher Scientific) as described in Section 3.2.3.6, but with a culture volume of 600 µL and 900 rpm shaking. Both screens were carried out in 96 round-well microplates (Fisher Scientific). The procedure followed was as described in Section 3.2.2 with some notable differences: A lysate volume of 20 µL was employed (final assay volume remained 200 µL); the substrates used were the same as in the primary screen (*i.e.* Bz-arg-AMC and Ac-lys-AMC) with the same fluorescence measurements and; the lysis method used was freeze-thaw.

3.3 Results and discussion

3.3.1 Verification of trypsin activity in microwells

To confirm the production of soluble bovine trypsin, activity towards L-BANA was measured over 24 hours using lysates of both IPTG induced and uninduced pET(T) and pET26b(+) in BL21-Gold(DE3), in the presence and absence of the specific irreversible serine protease inhibitor AEBSF (Section 3.2.2). It is clear from **Figure 3-2** that the lysate from pET(T) demonstrates cleavage of the L-BANA substrate that is at least 6.5-fold greater than the baseline present with the pET26b(+) control. (Furthermore, AEBSF incubated during cell growth simultaneously with IPTG induction inhibits the L-BANA cleavage by approximately 50% in this experiment, indicating that the L-BANA cleavage is caused by an arginine-specific protease, *i.e.* the

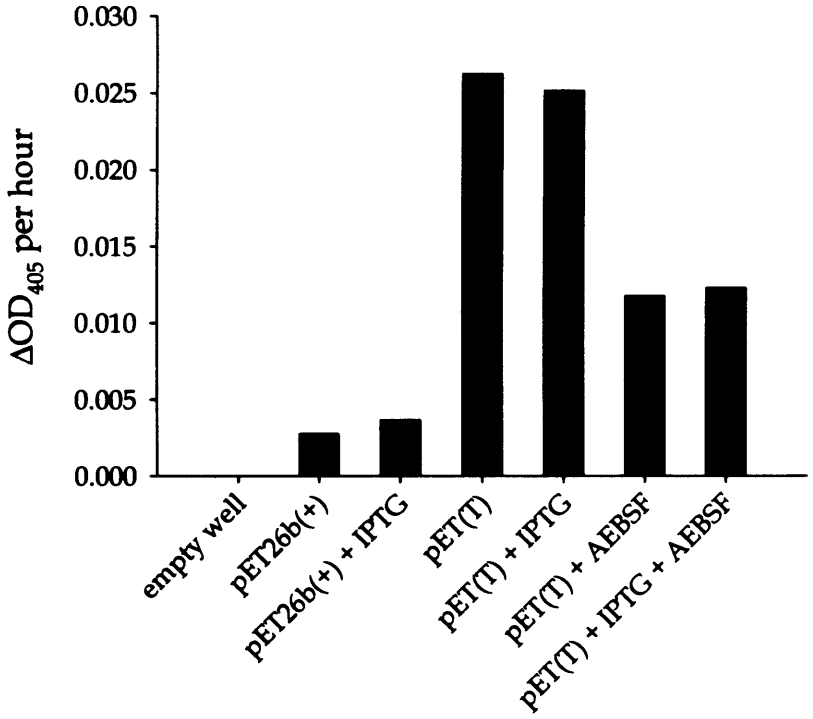

Figure 3-2. The effect of IPTG induction and AEBSF inhibitor on the bovine trypsin activity of BL21-Gold(DE3) cultures containing either pET(T) or the pET26b(+) control. Activity is plotted as ΔOD_{405} per hour or the change in absorbance units (AU) per hour measured over a 24-hour linear range.

bovine trypsin. Incomplete inhibition is probably due to the presence of insufficient AEBSF inhibitor towards the end of the fermentation. One striking feature of this data is the lack of the dependence of activity upon IPTG induction. The trypsin is being expressed under the T7 promoter which is known to be leaky (Studier *et al.*, 1990). This may account for the activity obtained in the absence of IPTG induction.

Assay constituents appear to have been favourable in testing for trypsin activity in microwells. The substrate used was a standard colourimetric reagent, Bz-arg-pNa, consisting of an N-terminal benzoyl blocker group, an arginine residue and a para-nitroanilide reporter group absorbing light at a wavelength of 405 nm. DMF was used to solubilise the substrate before it was diluted, in water, to the appropriate concentration. Based on previous trypsin assays (Evnin and Craik, 1988), a pH of 8.0 was used, which was maintained using Tris-HCl buffer. Finally, the addition of calcium chloride was for the provision of Ca^{2+} ions, previously shown to have had a stabilising effect on the enzyme by slowing the rate of autolysis (Sipos and Merkel, 1970).

Alternatives to this expression system of active trypsin have been considered in previous studies. The expression of trypsinogen in *E. coli* leads to the formation of inclusion bodies that can be refolded with up to 21% yield from urea in a controlled reactor (Buswell *et al.*, 2002). However, to convert this process into a microplate-based high-throughput assay by diluting urea denatured inclusion bodies in cell lysates was not considered to

be efficient or reproducible. The successful expression of rat anionic trypsin in the periplasm, using the OmpT leader sequence (Vasquez *et al.*, 1989; Evnin and Craik, 1988) suggested an alternative approach for the production of soluble and active bovine trypsin in the favoured redox environment of the periplasmic space. The bovine trypsin gene lacking its pro-peptide sequence was therefore isolated by PCR and ligated behind the *pelB* leader sequence of pET26b(+) to create the plasmid pET(T) (Hibbert, 2003). This expression system was expected to direct the mature trypsin gene to the periplasm by fusion to the *pelB* signal sequence, where subsequent signal cleavage was expected to release the mature trypsin to refold in the periplasmic space. Given that the activity of trypsin in microwells has been verified, this system was deemed suitable for high-throughput screening on a variety of substrates available commercially.

3.3.2 Library creation by fepPCR

3.3.2.1 fepPCR amplification of bovine trypsin gene

Five sections of the trypsin gene with varying length were targeted for random mutagenesis by fepPCR (Section 3.3.2.1). The likelihood of a successful reaction occurring was increased by the range of fragment sizes attempted. Any fragments successfully amplified were to be used in the subsequent MEGAWHOP reaction (Section 3.3.2.2). Agarose gel electrophoresis (**Figure 3-3**) confirmed that amplification reactions had worked to varying degrees with all fragments F1 to F5. The lane for F6

shows a band below the 50 bp mark indicating that primers did not get used up properly in the course of fepPCR. In contrast lanes F1 to F5 show a diminished band below the 50 bp mark and clearly visible bands at the expected amplification length. In the case of F6 the amplification length appears to have been unworkable at the reaction conditions used. Changing the conditions to facilitate primer binding was not likely to have addressed the problem since primer binding was not an issue in F1 to F5. However, an increase in time for the elongation stage may have made a difference in allowing strands to be amplified to their full length.

Concentrations of Mn^{2+} higher than 0.70 mM have been shown to inhibit fepPCR (Miller, 2004) and so this was taken as an upper-limit for any fepPCR reaction. This rule only needed to be applied to F1 and even at the capped concentration, the intensity of the band was particularly depleted relative to the other fragments F2 to F5. It is also noticeable that product bands become brighter going from F2 to F5 which is mainly due to the fact that the products themselves are larger resulting in greater ethidium bromide binding but may also be due to the decreasing inhibitory effect of Mn^{2+}.

3.3.2.2 *Whole plasmid PCR with fepPCR products as primers*

The traditional method for inserting epPCR products into a host vector is by ligation using restriction enzymes and ligase. The reaction is notoriously troublesome requiring optimisation of the ratio and concentrations of DNA

insert and vector. Even then erroneous products may result in the form of plasmids with missing or multiple inserts. Often several attempts are required for the reaction to produce fragments detectable by agarose gel electrophoresis. An alternative method for inserting a fragment was developed by Miyazaki and Takenouchi in 2002, which they titled megaprimer PCR of whole plasmid (MEGAWHOP). The procedure (see Section 3.2.3.3) is a modification of the QuikChange® protocol using "megaprimers" hundreds of bases long instead of short primers (20 - 40 bp long) designed for SDM. A wild-type recombinant plasmid is required to act as a template for elongation.

Fragments that were successfully amplified by fepPCR (F1, F2, F3, F4 and F5) were used as megaprimers in separate MEGAWHOP reactions. **Figure 3-4** is an agarose gel electrophoresis picture showing the product bands. Bands are visible in all lanes with differing degrees of intensity. Lanes F2 and F3 have not resolved completely where semi-circular bands may be seen. It was deemed that the reaction products were sufficient for library transformation purposes pending verification of the mutation rate.

Cycle conditions for MEGAWHOP were varied before a set of conditions that worked was eventually found (**Table 3-5**). The first changes from standard PCR were to substantially increase the annealing time from 0.5 minute to 3 minutes and the elongation time from 1 - 1.5 min per kbp to 3.5 min per kbp (21 minutes for the pET(T) plasmid). Overly sufficient times were allotted to these stages in the first instance to ensure that time would

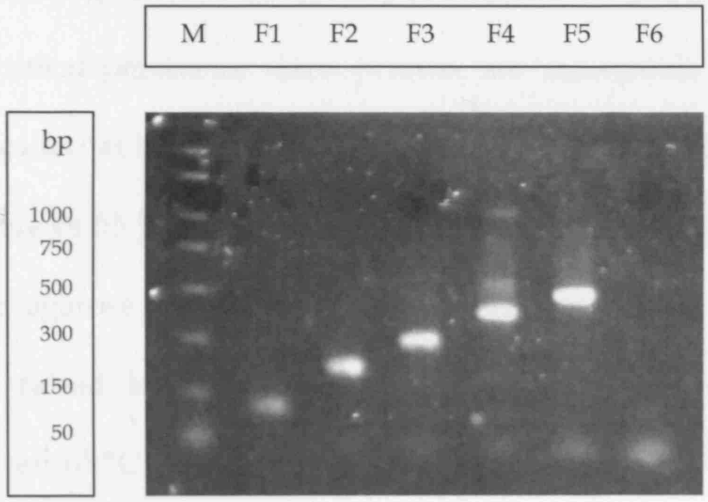

Figure 3-3. Agarose gel electrophoresis picture of fepPCR reactions. F1 = Frag 1 (119 bp), F2 = Frag 2 (222 bp), F3 = Frag 3 (315 bp), F4 = Frag 4 (433 bp), Frag 5 (522 bp), F6 = Frag 6 (963 bp).

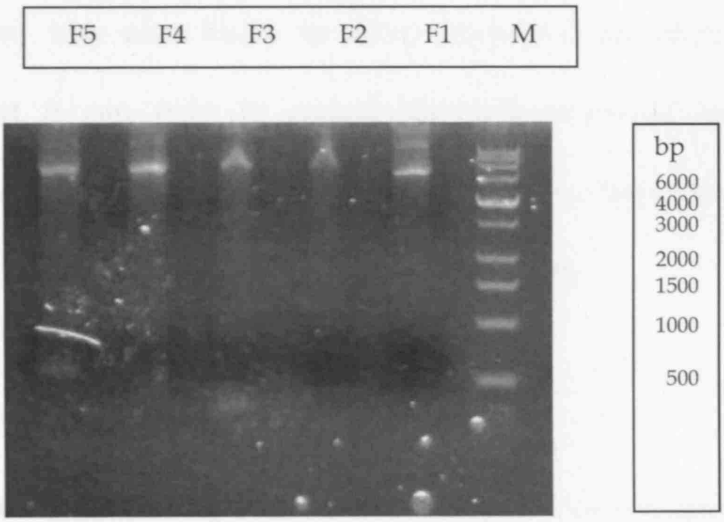

Figure 3-4. Agarose gel electrophoresis picture of MEGAWHOP reactions. Bands of varying intensity may be seen at the 6000 bp mark in all lanes.

not be a limiting factor to the success of a reaction. Annealing temperature is a more critical parameter since primers are susceptible to binding at the wrong location at low enough temperatures. Using the standard annealing temperature of 55 °C of the MEGAWHOP did not yield any product bands following agarose gel electrophoresis. The annealing temperature was therefore raised in 1 - 2 °C increments until a successful reaction was achieved at 65 °C annealing (**Figure 3-4**). For the correct binding of megaprimers it was evident that a highly selective annealing temperature was required which was just 5 °C below the elongation temperature.

As in the QuikChange® SDM reactions the addition of DMSO (4% final concentration) was also likely to have increased the likelihood of success given that it can help to reduce the occurrence of secondary structures. Structures of this nature were more likely to have emerged in megaprimers than in small primers due to the extra length.

3.3.2.3 *Verification of fepPCR mutation rate*

FepPCR differs from standard epPCR in that a gene section is targeted for mutagenesis. The difference between epPCR and standard PCR lies in the accuracy of base insertion in a growing strand of DNA. EpPCR relies on the act of random misincorporation somewhere along a growing strand while standard PCR aims for full integrity of the amplified strand. In order to achieve "error-prone" PCR, the fidelity of DNA polymerase was lowered by the addition of Mn^{2+} to the reactions (Beckman *et al.*, 1985). The standard

concentration used to achieve a mutation rate of one per gene fragment was 0.1 mM Mn^{2+}per kilobase for *Taq* DNA polymerase. Given that a more inclusive library would result from over-mutating rather than under-mutating (which may result in some wild-type contaminants) this value was multiplied by a factor of 1.2. The amended concentrations of Mn^{2+} are shown in **Table 3-2**.

Different regions of the trypsin gene were targeted, all of which included the active site residues. This was to improve the chances of creating a successful library after both fepPCR and MEGAWHOP reactions were carried out. The largest fragment that was successful in the fepPCR reaction was F5 (522 bp). Since the MEGAWHOP reaction was also successful (Section 3.3.2.2), this fragment was chosen for the main random mutagenesis library as it encompassed 522 out of a possible 669 bp of the trypsin gene (the greatest area of coverage out of all fragments that was successfully amplified).

Twenty DNA sequences of F5 amplified by fepPCR were checked for mutations compared to the wild-type bovine trypsin sequence. Once mutations were identified (see **Figure 3-5**) the raw sequencing chromatograms were checked to ensure that purported mutations were correctly interpreted. Misinterpretations could have been the result of a "double signal" at a single position in the sequence or an additional base added to or deleted from a stretch of A, G, T or C bases. In such cases the

```
5' AACATCAACGTCGTGGAGGGCAGTGAGC AGTTCATCTCCGCATCCAAGTCCATCGT
3' AACATCAACGTCGTGGAGGGCAGTGAGC NGTTCATNNCCNCNNCCAAGTCCATCGT

GCACCCGTCCTACAACTCCAACACTCTGAACAATGACATCATGCTGATCAAGCTCAAGT
GCACCCGTCCTACAACTCCAACACTCTGAACAATGACATCATGCTGATCAAGCTCANGT

CCCGCCGCATCCCTGAACTCCCGCGTGGCCTCCATCTCTCTGCCGACCTCCTGTGCCTC
CCCGNCGCATCCCTGAACTCCCGCGTGGCCTCCNTCTCTCTGNCGACCNCCNGTGCCTC

CGCCGGCACGCAGTGCCTCATCTCTGGCTGGGGCAACACTAAGAGCTCTGGCACCTCCT
CGCCGGCACGCAGTGCCTCATNTCTGGCTGGGGCAACNCNANGAGCTCTGGCACCTCCT

ACCCAGACGTGCTGAAGTGCCTGAAGGCTCCTATCCTGAGCGATTCCTCCTGTAAGTCC
ACNCAGACGTGCTGANGTGCCTGAAGGCTCCTATCCTGAGCGATTCCTCCTGTANGTCC

GCCTACCCTGGCCAGATTACCAGCAACATGTTCTGTGCCGGCTACCTGGAGGGCGGCAA
GCNTACCCNGGCCAGATTACCAGCAACANGTTCTGTGCCGGCTACCTGNNGGGCGGCAA

GCATTCCTGTCAGGGTGATTCTGGTGGCCCTGTGGTCTGCTCCGGCAAGCTCCAAGGCA
GGATTCCTGTCAGGGTGATTCTGGTGGCCCTGTGGTCTGCTCNGGCAAGCTCCAAGGCA

TCGTCTCCTGGGGTTCCGGCTGTGC 3'
TCGTCTCCTGGGGTTCCGGCTGTGC 5'
```

Figure 3-5. Mutation sites from fepPCR of Frag 5 compared with the wild-type sequence. The underlined region was the forward primer binding site during fepPCR which was not susceptible to mutation. The region inbetween the forward and reverse primers was amplified by epPCR. Note that the reverse primer was used as the sequencing primer and is not shown in the figure due to customary errors in sequencing the first 50 bases of a DNA strand (approximately). The bases shown are those after the point at which sequencing became reliable (*i.e.* when there were no missing or additional bases compared to the wild-type). DNA was sequenced from twenty discrete colonies. Every mutation has been mapped above and denoted by "N".

Sequence name	Mutations
Frag 5 A	A29T
Frag 5 B	T295C
Frag 5 C	G40C
Frag 5 D	A341T
Frag 5 E	T300C
Frag 5 F	A29T
Frag 5 G	A287T
Frag 5 H	
Frag 5 J	
Frag 5 K	G340A
Frag 5 L	T213C, T320C
Frag 5 M	T37C, T43A, A148G, A248G
Frag 5 N	A42G, C157G, T163A, T166A, C195T
Frag 5 O	
Frag 5 P	C120T, A211G
Frag 5 Q	C36T, A215T
Frag 5 R	
Frag 5 S	
Frag 5 T	A113T, T232C, C392A
Frag 5 U	C235A

Table 3-6. Base substitutions from sequences 5A to 5U. The integrity of each mutation was verified on the sequence chromatogram. Any mutations which may have been wrongly interpreted by the sequencing service were restored to the original wild-type base. A false interpretation could have been the result of a "double signal" at a single position in the sequence or an additional base added to or deleted from a stretch of A, G, T or C bases. A total of 27 mutations were considered to be genuine after removal of the dubious mutations. Five sequences returned no mutations. This gave an average mutation rate of 1.35 per fragment.

Mutation type	Frequency (%)
A → G	14.8
G → A	3.7
T → C	22.2
C → T	11.1
transition	51.8
A → C	0
T → A	11.1
A → T	22.2
T → G	0
G → C	3.7
G → T	0
C → A	7.4
C → G	3.7
transversion	48.2

Table 3-7. Mutation frequencies. Over the 27 documented mutations, there is a small bias in transition mutations (A ↔ G, T ↔ C) over transversion mutations (A/G ↔ T/C) despite the fact that there are twice as many transversion mutations possible than transition mutations.

original wild-type base was restored and the mutation not counted. A total of 27 mutations remained (see **Table 3-6**) following the removal of dubious mutations. This calculates to a rate of 1.35 mutations per gene fragment. Transitions were marginally more common than transversions (see **Table 3-7**), although in theory the bias would be predicted to be greater considering there are twice as many transversion possibilities than transition possibilities (eight compared to four).

3.3.3 High-throughput screen of fepPCR library

3.3.3.1 Primary screen: High-throughput activity assays in microwells

Eight library plates were generated (Section 3.2.3.6) and screened (Section 3.2.4) for activity separately on Bz-arg-AMC and Ac-lys-AMC. The library size of an fepPCR library with an average mutation rate of one per fragment is given by the following formula: fragment length (number of residues subject to mutation) × residue accessibility. The fragment length was 522 bp which equates to $\frac{522}{3}$ = 174 residues. Residue accessibility is the average number of residues (out of a possible 19) that may be accessed by mutating a single bp within any given codon, and calculates to a value of 5.7 (Svendsen, 2003). This gives a library size of 174 × 5.7 = 991.8. However, in reality this value was multiplied by 3 to increase the probability to 95%, that each mutant in the library would possess a mutation at a unique site. This resulted in a projected library size of 991.8 × 3 = 2975. In reality, the library was contained in eight 384 square-well plates, which equated to 3056

mutants taking into consideration that two wells were reserved for wild-type BL21-Gold(DE3):pET(T) inoculations per plate.

Library reaction plates were screened for arginine activity by monitoring the release of the fluorescent AMC marker group catalysed by the action of trypsin. Lysine activity was screened similarly on replica reaction plates. To eliminate the risk of background noise interfering with initial fluorescence measurements, the first true readings were taken as those at a time point of 2 hours with final readings taken at 4 hours. Final minus initial readings were calculated for each well to give a fluorescence increase over time. Values for arginine were divided by values for lysine to give an activity ratio indicating the specificity of trypsin mutants. These data were arranged in ascending order for each plate (A - H) and plotted (**Figure 3-6**). In some cases, wild-type cultures and mutants did not grow in their allotted wells overnight in which case data was excluded from the plots. The data was plotted separately for each plate since the plate reader was subject to day-to-day variations. These variations were the result of inevitable switching on and off of the plate reader from day to day, which led to variable baselines each time. Combining data from all library plates was not possible due to these variations.

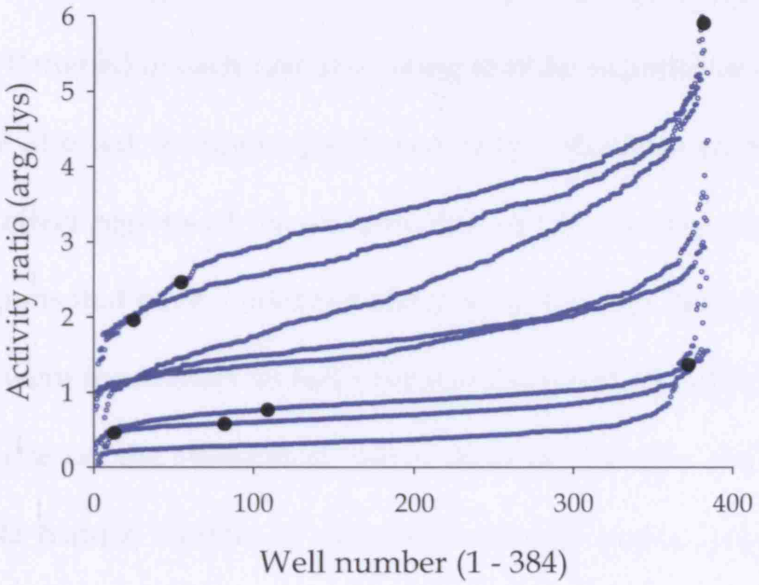

Figure 3-6. Plot of arginine versus lysine activity for each mutant in the fepPCR library (eight library plates treated separately). Increase in intensity of the fluorescent AMC marker, released by tryptic digestion, was measured over 4 hours. Arginine and lysine activity was taken as the fluorescence reading at 4 hours minus the reading at 2 hours. Arginine activity was divided by lysine activity to give an activity ratio, which was then sorted in ascending order (treating each plate separately) to identify those mutants of interest. The outcome for all eight library plates is displayed on the plot (single blue dot for each mutant). Mutants with high ratios were potentially arginine-specific variants while those with low ratios were potentially lysine-specific. Black dots indicate the activity ratio of wild-type pET(T) cultures.

All lines in **Figure 3-6** followed an expected sigmoidal pattern. The sigmoid is flattened in each case indicating that the majority of mutants were not greatly affected by single point mutations. Random mutagenesis of a gene may affect regions of the enzyme that impact greatly on specificity as well as regions that have a minimal effect on specificity. Mutations affecting specificity were most likely to have occurred to residues within 10 - 15 Å of the active site on the strength of recent findings (Morley and Kazlauskas, 2005). Data from a number of directed evolution papers was analysed to correlate mutation distance from the active site with improvements to enzyme properties (including substrate specificity).

The structure file PDB ID: 1BTW is of an inhibitor bound to a bovine trypsin molecule. The number of residues within 10 Å of the lysine side chain of the inhibitor was found to be 61 out of 223 and so the sequence space searched in the fepPCR library was mostly outside of this proposed hotspot. It may have, therefore, been expected that the majority of mutants would not have had a significant change in activity, which was the case. This supports the recent findings on distance correlations. It should be noted, however, that specificity-enhancing mutations were also found at locations distant from the active site (e.g. > 20 Å) in the Morley study. This justified the random mutagenesis strategy used at this stage, which was intended to represent an initial approach to enzyme engineering that was not information-intensive.

Wild-type trypsins (black dots), surprisingly, are scattered across the plots exhibiting a full range of activity ratios. Wild-type assays were expected to indicate the natural activity ratio of trypsin under the experimental conditions used. The reason that they did not appear to do this may have been due to the manner in which wells were inoculated with wild-type pET(T) cells. All wells containing library DNA were inoculated via automated colony picking whilst those containing wild-type plasmid were inoculated manually. With hindsight, manual inoculation using a sterile cocktail stick was open to variations in the amount of viable cells that were present at the start of the culture. It follows that the time to reach maximum OD was also variable. Given the sensitivity of trypsin to autolysis (Section 1.1.3.3), the growth rate of the culture was likely to have affected the final concentration of mature trypsin produced. Such variations may have accounted for the wild-type data appearing to have variable activity ratios. Automated colony picking may also have been affected by this phenomenon. This was countered, to a degree, by setting certain parameters beyond which the robot would not pick colonies. Colony picking parameters that were controlled included the roundness and diameter of colonies as well as the proximity to other colonies. The software on the QPix2 robot allowed for the input of arbitrary values for these parameters rather than absolute units.

It was possible that wild-type trypsins may have been contaminated by wells containing library DNA. However, a study on the danger of contamination from external sources or neighbouring wells found that there

was no significant risk (Miller, 2004). A similar method had been employed here for the generation of mutant libraries in 384 square-well plates.

Possible scenarios that may have caused false positives[2] to crop up have been discussed, which suggested the need for a secondary screen. Data obtained at this stage, however, was not intended to identify mutants of interest conclusively, but to identify candidates with a potentially enhanced specificity from the library at large. From each plate, mutants with a high arginine to lysine activity ratio (or high arginine specificity) were selected for a secondary screen as were those with a low ratio (or high lysine specificity).

3.3.3.2 *Secondary screen*

A repeat of the screening procedure was carried out in triplicate (Section 3.2.5) with the best performing mutants from the primary screen. Some changes were made to the primary screening procedure in an attempt to produce more reliable data and reduce the occurrence of false positives. Firstly, mutants were grown up in 96 deep-square-well plates. This was to reduce the effects of variable growth patterns that were more likely to have occurred in 384 square-well plates. In particular, evaporation effects could be minimised using culture volumes of 600 rather than 60 µL where approximately half of the culture volume was lost after an 18 hour incubation. Also, inoculating a larger volume was more likely to result in more uniform cell growth rates than inoculating a 60 µL culture. Finally,

[2] 'false positives' are mutants with falsely exaggerated activities singling them out in a library.

performing the screen in triplicate (from three unique cultures) was more likely to average out any well to well variations in enzyme concentration.

The screening data was processed in the same manner as the initial screen and an average activity ratio was calculated for each mutant. The results are displayed in **Figure 3-7**. The activity ratio for the wild-type pET(T) control is displayed as a black bar. Mutants selected from the primary screen with purported high arginine and lysine specificity are coloured blue and red respectively. Error bars on the plot show a minimal deviation from the mean activity ratio in all cases indicating that the assay data may be considered accurate.

In carrying out a secondary screen, it was expected that the chosen mutants would fall into one of two categories, either high arginine specificity or low arginine specificity with the wild-type somewhere in between. The pattern observed, however, is a continuum of activities without any clear separation. Nevertheless, mutants at the lower end of the scale were generally those selected for high lysine specificity while those selected for high arginine activity were at the higher end, as should be the case. There were some exceptions with two arginine-specific mutants exhibiting a relatively low activity ratio. These mutants were most likely to be false positives from the primary screen. Given that the screen failed to decisively separate the mutants on the basis of specificity, a tertiary screen was necessary for the most promising candidates identified from the secondary screen.

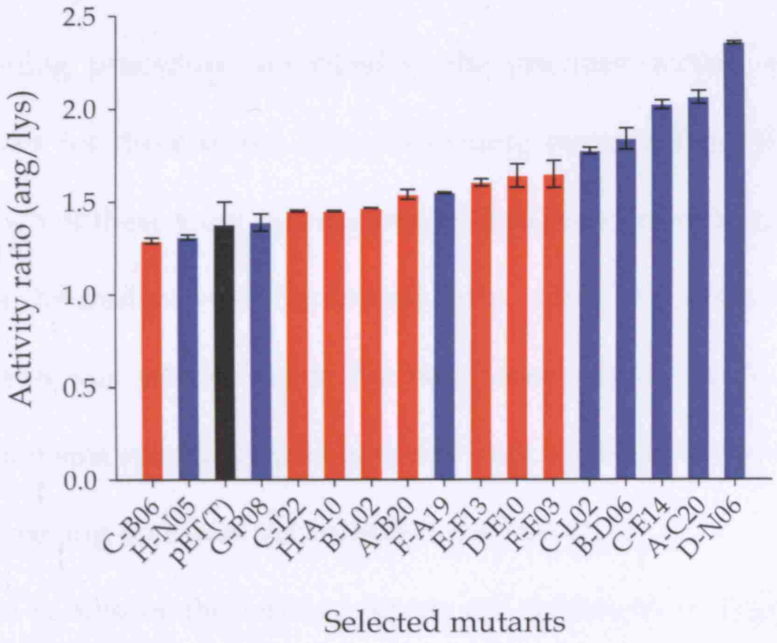

Figure 3-7. Secondary screen for activity ratio of mutants selected from initial screen. The activity ratios are shown for wild-type pET(T) (black bar), mutants selected for arginine specificity (blue bars) and mutants selected for lysine specificity (red bars). Reference points for mutants in the fepPCR library are displayed on the x-axis e.g. C-B06 denotes plate C, well position B06. Assays were carried out in triplicate. Activity ratios were calculated as described in **Figure 3-6**. Error bars correspond to ± one standard deviation about the mean.

3.3.3.3 Tertiary screen

The screening procedure described in the previous section was repeated fifteen times for three of the best performing mutants from the secondary screen. Two of these were the mutants with highest activity ratios while the other was the mutant with the lowest ratio. Only one mutant with a low activity ratio was selected since the next lowest mutant with a purported increase in lysine specificity (red bar) was four positions above and therefore further screening on it was not considered at this stage.

The results of the tertiary screen are displayed in **Figure 3-8**. The activity ratio for the wild-type pET(T) control is displayed as a black bar. Mutants selected from the secondary screen with purportedly high arginine and lysine specificity are coloured blue and red respectively. The most striking feature of the plot is that all three mutants now had high activity ratios (> 2.3 : 1) where previously the mutant selected for a low activity ratio had a value of only 1.29 : 1. This suggested that the secondary screen was also vulnerable to errors as reliable means of separating mutants based on their specificity. The crudeness of the primary screen was expected, to a large extent, bearing in mind that assays were performed only once on each mutant. The secondary screen, however, was expected to root out any false positives. For the mutant with a purportedly high lysine specificity, the tertiary screen has failed to corroborate the results of the secondary screen and instead confers an activity ratio consistent with high arginine specificity. Tertiary screening data for the two mutants with purportedly high arginine

	Mutation(s)	Activity (AU.min^{-1}ug^{-1}ml)		Ratio of arg/lys
		arg	lys	
pET(T)	-	3.93×10^{-3}	2.56×10^{-3}	1.5 : 1
C-B06	N79S, N115S	2.44×10^{-3}	1.06×10^{-3}	2.3 : 1
A-C20	N79S	5.21×10^{-3}	2.03×10^{-3}	2.6 : 1
D-N06	S146G, M180I	2.35×10^{-3}	9.93×10^{-4}	2.4 : 1

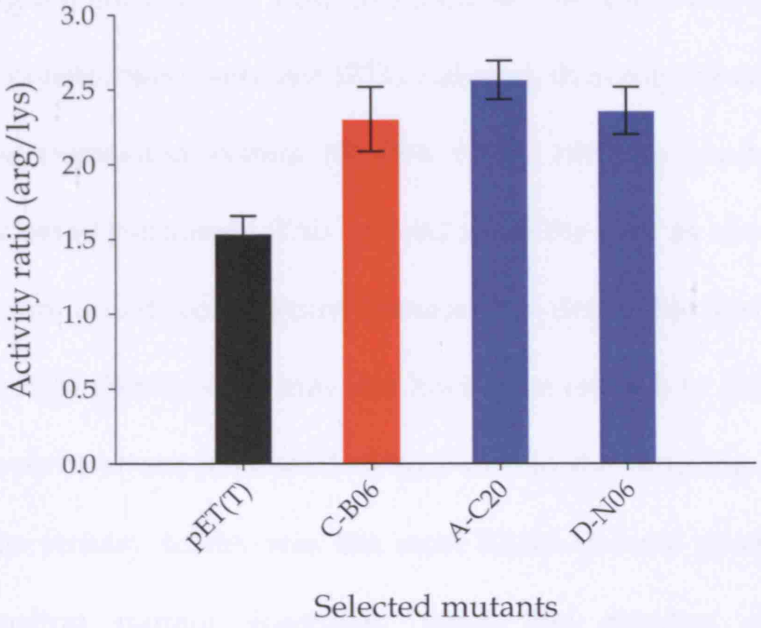

Figure 3-8. Tertiary screen for activity ratio of mutants selected from secondary screen. The activity ratios are shown for wild-type pET(T) (black bars), mutants selected for arginine specificity (blue bars) and mutants selected for lysine specificity (red bars). Assays were carried out in triplicate and activity ratios were calculated as described in **Figure 3-6**. Error bars correspond to ± one standard deviation about the mean.

specificity, however, were consistent with the secondary screen. The wild-type activity ratio of 1.54 also concurred reasonably with the previous value of 1.38. The cause of the anomalous data might have been due to differences in enzyme concentration from well to well. Although this was always a possibility, it may have been aggravated further by the effects of autolysis inactivating the enzyme. In a bid to minimise this, microwell fermentations for library construction were not IPTG induced, thus relying on the leaky T7 polymerase expression system (Studier *et al.*, 1990) to produce sufficient trypsin for assay purposes. This proved to be the case as shown in **Figure 3-2** where an uninduced culture produced a detectable level of trypsin. Nevertheless, such measures may not have been enough to prevent variable enzyme concentrations from well to well due to the differing growth rates. Overall, the tertiary screen was the most likely to have produced reliable data regarding mutant specificity given the number of repetitions encompassed. Error bars shown on the plot do not show a large deviation from the mean activity ratios indicating that the tertiary screen has produced statistically significant values.

All three mutants were sequenced (Section 2.4.9) to identify the mutations responsible for causing the specificity modifications. The mutated residues have been highlighted in **Figure 3-9**. Mutant A-C20, with high arginine specificity, contained a mutation at N79S. Using the PDB structure file, 1BTW, distances of mutations from the lysine side chain of the inhibitor (instance residue) were obtained. The asparagine residue was 24.4 Å away

from the instance residue, but was within close proximity (8.1 Å) to the Ca^{2+} ion binding loop near the surface of the molecule. This suggested there may have been a subtle change to the overall structure of the enzyme increasing the binding pocket's affinity for arginine. Another explanation is that the Ca^{2+} interaction with trypsin might have been enhanced resulting in greater overall stability. This might be due to the molecule following an alternative pathway to self-proteolysis in which the molecule retains its ability to catalyse arginine-specific proteolysis reactions in a substrate for a longer period. The increased arginine specificity of the mutant, therefore, might have been a side-effect of the molecule following an alternative autolysis pathway. If this was the case then the effect on substrate specificity was not due to an altered interaction between the active site and the substrate. This is an interesting theory in itself that would require an investigation into the various states trypsin adopts during the autolysis pathway.

Mutant D-N06, also with high arginine specificity contained two mutations, S146G and M180I. Both were relatively close to the active site at distances of 9.3 and 10.7 Å respectively. This indicated that both mutations may have had a more direct effect on the mechanics of substrate binding. The attraction or ease of interaction, specifically between the binding pocket and the arginine residue of a substrate, may have been enhanced at the expense of a lysine residue. Mutant C-B06 was chosen from primary and secondary screens for its high lysine specificity, however, the tertiary screen relayed that the mutant had a better specificity towards arginine after all.

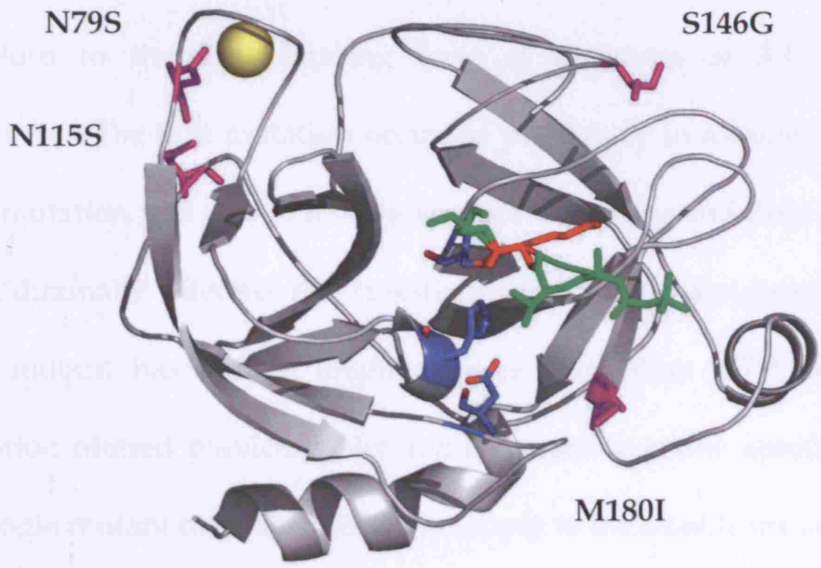

Figure 3-9. Locations of trypsin mutations in relation to the substrate-binding site. Mutations N79S, N115S, S146G and M180I are coloured purple. A bound active-site inhibitor molecule (T-Butoxy-ala-val-boro-lys 1,3-propanediol monoester) is shown in green with the constituent lysine residue in red. The residues of the catalytic triad are coloured blue and the bound Ca^{2+} ion is coloured yellow.

Two mutations, N79S and N115S, were found in the sequence both of which were close to the Ca^{2+} binding loop at distances of 8.1 and 13.0 Å respectively. The first mutation occurred previously in mutant A-C20. The second mutation was also to a surface-exposed residue and does not seem to have additionally affected the specificity of the enzyme greatly since the double mutant has only a slightly lower ratio than N79S alone. The explanation offered previously for the increased arginine specificity in the N79S single mutant may, therefore, also apply to the double mutant.

The changes to mutant specificity were of greater importance in the context that overall activity on either of the arginine or lysine substrates alone was retained. Mutants with similar specificity enhancements have been isolated before (Evnin *et al.*, 1990), although at the expense of overall activity in all cases. Those mutants with retained activity closest to that of wild type were 66% as active on arginine and 74% as active on lysine. From the tertiary screen, mutants C-B06 and D-N06 both had slightly lower activities on arginine (62% and 60% respectively) and lysine (41% and 39%) compared to the wild-type. On the other hand, mutant A-C20 had an improved activity on arginine (133%) and a reduced activity on lysine (79%) compared to the wild-type. This mutant was especially important given that specificity was enhanced without a reduction in both arginine and lysine activity.

The anomalous screening data presented may lead to the conclusion that the screening process had been flawed. However, it is important to note

that random single mutations can result in major enhancements to enzyme activity and specificity and, in many directed evolution examples, improve by a factor of between 2 - 100 fold (Dalby, 2003). Once screened, a trypsin mutant with a specificity enhanced in this range would exhibit a change in specificity large enough to make the error, present during library construction and screening, appear insignificant. Since no such mutant was found, this indicated that limitations of using a random mutagenesis strategy were responsible for the lack of success in finding more significant specificity-enhancing mutations. The limited library size as a result of there being only 5.7 from a possible 19 residue substitutions at each site on average, in addition to the fact that even these were not guaranteed, may have been the cause of the shortcomings experienced. Rather than persevere with this strategy, a more targeted approach was pursued.

4 Targets for enzyme engineering

4.1 Introduction

Bovine trypsin is an endoprotease naturally specific towards cleavage at the carboxy-terminal end of lysine and arginine residues. Due to this relatively broad specificity, unwanted cleavages usually result in a mixture of protein by-products when processing fusion protein biopharmaceuticals and gives rise to further costly purification steps. Furthermore, the presence of several internal arginine and lysine residues within a trypsin molecule leads to the phenomenon of self-proteolysis. One trypsin molecule may proteolyse another at any of the recognition sites which eventually inactivates the molecule (Keil-Dlouha *et al.*, 1971a; Smith and Shaw, 1969; Maroux *et al.*, 1967).

One option available is to optimise the process conditions and minimise both secondary cleavage of the target protein (see **Figure 1-8**) and the self-proteolysis of trypsin. However, as this does not necessarily reach a satisfactory optimum, the alternative is to engineer the enzyme for enhanced specificity towards the desired target site. Directed evolution provides a generally suitable route for improving enzyme activity and specificity without prior knowledge of enzyme structure-function relationships (Stemmer, 1994; Chen and Arnold, 1993). However, there are now a considerable and growing number of complete genome sequences and crystal structures available for many enzymes including bovine trypsin.

Using this information may help to identify, with a better probability, those amino acids responsible for influencing enzyme activity and specificity.

In Chapter 3, fepPCR was used to obtain mutants with up to 2.56 fold improved substrate specificity. In this chapter, the aim was to investigate alternative strategies for obtaining greater enhancements while not generating libraries too large for selection or screening. Analyses of variants obtained from fully random mutagenesis experiments with improved activity, substrate specificity or enantioselectivity, have shown that the majority of mutations are found in regions that contribute to substrate binding, catalysis or the conformation and dynamics of the active site environment (Morley and Kazlauskas, 2005; Dalby, 2003). Consequently, useful variants can be obtained from smaller, more defined enzyme libraries, in which random mutations are focussed to regions of the enzyme more likely to result in beneficial mutations (Kast and Hilvert, 1997), such as the active site (Shinkai *et al.*, 2001), catalytic residues (Mullegger *et al.*, 2005), or where hot-spots have been identified by error-prone PCR (Miyazaki and Arnold, 1999). The targeting of saturation mutagenesis, either experimentally or *in silico*, to active-site residues chosen solely on a structural basis for their proximity to substrate or product has led to the successful improvement of enzyme activities (Hayes *et al.*, 2002), altered substrate specificity (Hibbert *et al.*, 2007; Ashworth *et al.*, 2006; Santoro and Schultz, 2002; Wang *et al.*, 2001; Ting *et al.*, 2001), and altered catalytic reactions (Dwyer *et al.*, 2004; Peimbert and Segovia, 2003; Goud *et al.*, 2001). There are

limited examples reported of improvements made to protease activity or specificity without a major structural overhaul such as in the engineering of trypsin to chymotrypsin-like activity in which nine residues were mutated (Hedstrom *et al.*, 1992).

The overall benefit of targeted mutagenesis has been systematically and directly compared to fully random mutagenesis in a number of studies, including the improved enantioselectivity of the *P. fluorescens* esterase (Park *et al.*, 2005), the improved activity of dihydrofolate reductase (DHFR) (Schmitzer *et al.*, 2004), and the activity of *E. coli* β-galactosidase (Parikh and Matsumura, 2005). It has also been shown recently for the directed evolution of epidermal growth factor (EGF), EGF receptor, interleukin-2 and subtilisin, that mutagenesis improving natural activity is biased towards residue positions that are phylogenetically variant in nature, and that 40 - 80% of the amino-acids obtained at these positions were present in at least one natural ortholog (Cochran *et al.*, 2006). By contrast, loss of function was observed for mutations at highly conserved residues. The directed evolution of transketolase by targeted saturation of both highly conserved and phylogenetically variant active site residues revealed a greater tolerance to mutation at the naturally variant residues (Hibbert *et al.*, 2007). Improved activities towards naturally analogous hydroxylated substrates were obtained from mutations at the most phylogenetically variant residues. Furthermore, three highly conserved residues that normally interact with phosphorylated substrates became tolerant to mutation when assaying

activity towards non-phosphorylated substrates, and yielded mutants with improved activity. In a separate study, a similar mutagenesis strategy was used to produce libraries to screen for improved activity on a non-natural aliphatic aldehyde substrate (Hibbert *et al.*, 2008). Improved mutants were found to have mutations at conserved residues that, again, normally interact with natural substrates. This chapter investigates the relationship between enzyme sequence entropy and measured gains in activity towards both natural and non-natural substrates. The relationship between proximity to the active site (and other regions of association) and activity gains is also assessed. Previously reported directed and rational evolution experiments, and available sequence homologues of the evolved enzymes, were used as a basis for this study. Targets for engineering bovine trypsin specificity were sought based on the findings.

In a parallel study, target sites for engineering the resistance of trypsin to self-proteolysis were investigated. The benefit of directing mutations towards a select few sites has been demonstrated in the case of rat anionic trypsin resulting in a reduced susceptibility to self-proteolysis (Varallyay *et al.*, 1998). Efforts were made to determine homologous sites in bovine trypsin which were likely to be susceptible to self-proteolytic cleavage as well as other sites deemed to be most susceptible based on structural analysis. Rational mutations were introduced to the trypsin gene with the goal of increasing its resistance to self-proteolysis and improving the overall stability of the enzyme (Chapter 5).

4.2 Materials and methods

4.2.1 Selection of enzymes to study

A set of criteria was established for selecting literature examples of directed and rational evolution to use in this study. Examples of directed evolution using a random mutagenesis strategy were separated from those using a rational targeted approach such as saturation or site-directed mutagenesis. In light of the project aims to modify the substrate specificity of trypsin, the first requirement for any chosen example was that the improved function had to be the activity of the enzyme on a natural substrate (natural activity) or the activity of the enzyme on at least one non-natural substrate (specificity). Enhancements to enantioselectivity were included as well as enhancements to general substrate specificity. Mutants of interest had to have a single amino acid substitution in order to show that this mutation alone was responsible for the improvement.

Kinetic data relating to enzyme activity was required both for the wild-type enzyme and for each mutant in order to compare the magnitude of improvement between different enzymes. Typically $kcat/K_M$, $Vmax/K_M$ or *Vrel* values were used although in some cases the paper only reported the specificity or enantioselectivity as a percentage relative to wild-type which was used instead.

For each enzyme a PDB structure file was required to define the active site residues, locate mutations and to measure the distances of mutations

from the active site. The freely available molecular graphics software, PyMOL (Delano, 2002) was used for this purpose.

4.2.2 Change in activation free energy caused by a mutation

The degree of enhancement in enzyme specificity had to be standardised in a form that could allow for a direct comparison between enzymes. The activation free energy change caused by a mutation was used for this purpose as defined by Morley and Kazlauskas, 2005:

$$\Delta\Delta\Delta G = - RT\ln(S_{mutant}/S_{wild\text{-}type}) \quad [4.1]$$

where R (gas constant) = 8.314 $J.K^{-1}M^{-1}$, T (temperature) = 298 K and S = substrate selectivity = $(kcat/K_M)_{non\text{-}natural\ substrate}/(kcat/K_M)_{natural\ substrate}$. For enzymes with improved natural activity the energy change simplifies to:

$$\Delta\Delta G = - RT\ln[(kcat/K_M)_{mutant}/(kcat/K_M)_{wild\text{-}type}] \quad [4.2]$$

Where $kcat/K_M$ is not given in a reference, other useful data was used as long as it allowed for a comparison between activities on different substrates and mutants (*i.e.* $Vmax/K_M$ or $Vrel$).

4.2.3 Construction of sequence alignments

The online basic local alignment search tool, BLAST[1] (Altschul *et al.*, 1990), was used to obtain sequences homologous to the experimentally evolved enzyme. Those with less than 30% similarity were discarded for alignments and entropy calculations as were sequences with 100% similarity. Sequences were aligned using DbClustal, a multiple sequence alignment tool available to users of BLAST, and saved in FASTA file format for manipulation with BioEdit software (Hall, 1999). Positions in alignments that consisted mostly of amino acid omissions were deleted.

4.2.4 Sequence entropy calculations

Entropy values, $H(x)$, for each position within a sequence alignment were calculated using the entropy function in BioEdit. The equation is given as:

$$H(x) = - \Sigma f(\mathrm{b},x)\log(\mathrm{base2})f(\mathrm{b},x) \qquad [4.3]$$

where $H(x)$ is the uncertainty or entropy at position x, b represents a residue (out of the allowed choices for the sequence in question), and $f(\mathrm{b},x)$ is the frequency at which residue b is found at position x. The information content of a position x, then, is defined as a decrease in uncertainty or entropy at that position. The maximum entropy for twenty-one possible amino acids

[1] http://www.ebi.ac.uk/blastall/ [Accessed 01st May 2007]

(including the stop codon) is 3.04. Zero represents a fully conserved residue. The minimum number of homologues used for an entropy calculation was set at 10 whilst the maximum number available in a search was 250.

In some cases, the vast majority of homologues returned from BLAST had high percentage similarities. For example, greater than 90% of the 250 homologues of TEM β-lactamase all had a percentage similarity of 98% or higher. As a result the mean entropy of the enzyme was 0.36 which was clearly biased by the availability of only very similar homologues. Enzymes displaying such bias were discarded from this study. The mean entropies of enzymes included in the study ranged from 0.68 - 1.48.

4.2.5 Considerations for entropy calculations

The entropy of a given residue in a protein sequence is a measure of the degree of randomness at that position relative to a group of homologues (equation 4.1). The absolute maximum value for a given position is 3.04 which represents the highest possible degree of phylogenetic variation. The minimum computed value is zero indicating that the position is fully conserved across all homologues available. Entropy must therefore be influenced by the overall similarity between the reference sequences. A calculation based on a large number of close homologues would give a low entropy whilst a low number of close homologues would give a high entropy. Since the online BLAST search tool returns every available sequence homologue, it was necessary to define constraints on the

homologues that were used in entropy calculations to account for some of the bias.

A study of the "twilight zone" of 20 - 35% sequence identity was carried out (Rost, 1999) to establish the relationship between sequence identity and homology. After checking more than a million sequence alignments, it was found that 90% of pairs were truly homologous only if their sequence identity was greater than 30%. Having a sequence identity of less than 25% resulted in only 10% being homologous. Based on this evidence, sequences below 30% sequence identity were considered to be too distantly related for a robust entropy calculation and therefore discarded.

Homologues with a percentage similarity of exactly 100% were also discarded since, by definition, these sequences would be identical to the input sequence therefore biasing the entropy calculation. Any other duplicate sequences from the same species were removed to leave only one. Putative and fragmented sequences were also removed. The minimum number of homologues returned by BLAST was set at ten for a valid entropy calculation. Having calculated entropies for each enzyme a further constraint was defined; the mean entropy for an enzyme's entire sequence needed to be above 0.6. Where mean entropies fell below this value, it was taken as an indication that the set of reference sequences were too closely related for a balanced calculation (see Section 4.2.4).

4.2.6 Definition of active site residues and distance methods

Three methods were used by which to define the active site residues of an enzyme. For distance method 1, structure files were searched for a key atom from a catalytic residue of the enzyme as used in a previous study (Morley and Kazlauskas, 2005). This was taken as the reference point for each enzyme. A python script run in PyScripter (http://mmm-experts.com) was written and used to generate a file containing the distances between the specified reference point and all residues present in an enzyme (Paul Dalby, Dept. of Biochemical Engineering, UCL). Those residues with a distance of less than 10 Å from the reference point were grouped together and labeled as active site residues.

For distance method 2, a bound substrate together with the chosen key catalytic atom were taken as the collective reference point from which to calculate all distances as the nearest approach to any atom in the group. In some cases there was no substrate included in the structure file and so a reactive substrate was modelled in at the active site. For distance method 3, the catalytic atom, the whole substrate and any cofactors present were taken as the reference point from which to calculate all distances.

4.2.7 Sites important for self-proteolysis

A literature survey was carried out to identify the key residues likely to be responsible for the rapid self-proteolysis of bovine trypsin. In addition, a structural analysis of bovine trypsin was carried in PyMOL to identify

additional residues, not mentioned in the literature, that were likely to be causing self-proteolysis.

4.3 Results and Discussion

4.3.1 Selection of enzymes to study

The enzymes short-listed for this study are shown in **Table 4-1** and **Table 4-2**. They have been enhanced by various directed and rational evolution methods and satisfy the criteria set out in Section 4.2.1 detailing the information required for carrying out this study. The following factors were analysed by a combination of frequency distributions and various scatter plots: (1) sequence entropy; (2) distance methods 1, 2 and 3; and (3) activation free energy change caused by a mutation.

4.3.2 Frequency distributions of distances

4.3.2.1 Directed evolution study

Histograms were plotted showing distance frequency distributions for each of the following: (1) activity-enhancing mutations; (2) specificity-enhancing mutations; (3) whole sequence residues; and (4) active-site-only residues. "Activity-enhancing" mutations denote mutations that have improved activity towards a natural substrate (*i.e.* change in *kcat*). "Specificity-enhancing" mutations denote mutations that have increased specificity towards either the natural substrate or a non-natural substrate (*i.e.* change in

Enzyme	Origin	Random mutagenesis method	Reference
Directed Evolution (DE)			
Glutaryl acylase	*Pseudomonas SY77*	epPCR; spiked oligonucleotide	Otten *et al.*, 2002 Sio *et al.*, 2002
Lipase	*Pseudomonas aeruginosa*	epPCR	Fujii *et al.*, 2005
Cyclodextrin glycosyltransferase	*Bacillus circulans*	epPCR	Leemhuis *et al.*, 2003
D-amino acid oxidase	*Rhodotorula gracilis*	epPCR	Sacchi *et al.*, 2004
Galactose oxidase	*Fusarium graminearum*	epPCR	Wilkinson *et al.*, 2004; Delagrave *et al.*, 2001
Lipase	*Bacillus subtilis*	epPCR	Funke *et al.*, 2005
deoxy-D-ribose 5-phosphate aldolase	*Escherichia coli*	epPCR	Jennewein *et al.*, 2006
Tagatose-1,6-bisphosphatase	*Escherichia coli*	DNA shuffling	Williams *et al.*, 2003
Glutathione-S-transferase	*Rattus norvegicus*	DNA shuffling	Broo *et al.*, 2003
Ribulose 1,5-bisphosphate carboxylase/oxygenase	*Rhodospirullum rubrum*	epPCR	Mueller-Cajar *et al.*, 2007

Table 4-1. Enzymes enhanced by directed evolution used in this study.

Enzyme	Origin	Reference
Rational Evolution (RE)		
Transketolase	*Escherichia coli*	Hibbert *et al.*, 2007 Hibbert *et al.*, 2008
TEM-1 beta-lactamase	*Escherichia coli*	Gaytan *et al.*, 2002 Wang *et al.*, 2002b
Beta-galactosidase	*Escherichia coli*	Parikh and Matsumura, 2005
Penicillin acylase	*Escherichia coli*	Gabor and Janssen, 2004
Aminoacyl-tRNA synthetase	*Methanococcus jannaschii*	Turner *et al.*, 2006 Wang *et al.*, 2003 Wang *et al.*, 2002a
O6-alkylguanine-DNA alkyltransferase	*Homo sapiens*	Juillerat *et al.*, 2003
DNA polymerase I	*Thermus aquaticus*	Xia *et al.*, 2002
DD-transpeptidase	*Streptococcus pneumoniae*	Peimbert and Segovia, 2003
Lipase	*Pseudomonas aeruginosa*	Reetz *et al.*, 2005
Phospholipase C	*Bacillus cereus*	Antikainen *et al.*, 2002
Isocitrate dehydrogenase	*Escherichia coli*	Doyle *et al.*, 2000
Subtilisin	*Bacillus licheniformis*	Strausberg *et al.*, 2005

Table 4-2. Enzymes enhanced by rational evolution used in this study. Mutagenesis methods used were variations of site-directed and saturation mutagenesis.

kcat/K_M). "Whole sequence" denotes every residue in an aligned sequence. "Active-site-only" residues denote residues within 10 Å of a defined reference point according to three different methods (Section 4.2.6). The use of three different reference points led to the calculation of three different distance measurements. Frequencies were converted from sum totals to percentage occurrences by dividing each raw frequency by the total number of occurrences and multiplying the figure by 100. Processing sets of data from DE examples as described above allowed for a direct comparison between enzymes.

The first thing to note in the DE histograms is that whether methods 1, 2 or 3 were applied, the overall pattern does not seem to differ greatly. Going from methods 1 to 2 to 3, however, there is an increase of frequencies in the smaller groupings between 0 and 15 Å. On the histograms this can be seen as a shifting to zero of all frequency groupings. Whether distance method 1, 2 or 3 was applied, every activity-enhancing mutation and every specificity-enhancing mutation was at a distance of less than 35 Å away from the reference point (

Figure 4-1 and **Figure 4-2**). Further to this, there is a general trend of increased frequencies between 8 and 12 Å. A greater representation of residues is shown in these groupings generally resulting in a peak.

In contrast, histograms for all residues (**Figure 4-3**) show a more even spread with only roughly 75% of residues at a distance of less than 35 Å away from the reference point (for all distance methods). This indicated that

having mutations less than 35 Å away from any of the reference points was absolutely essential to improving activity or specificity in the examples studied. This supported findings in a previous study on the importance of proximity (Morley and Kazlauskas, 2005) where no enhanced mutants were found beyond 35 Å from a key reactive atom.

Distance from the reference point has been divided into two scenarios for the case of specificity-enhancing mutations. Those frequencies above the x-axis are cases where the ΔΔΔG value has been positive *i.e.* the mutant had better specificity towards the *non-natural* substrate compared to the wild-type. Frequencies below the x-axis (negative ΔΔΔG values) indicate that the mutant had better specificity towards the *natural* substrate than the wild-type had. Note that for mutants with negative ΔΔΔG values, ΔΔG calculations have also been carried out and included as examples of improvements in activity where possible.

In the case of specificity-enhancing mutations (**Figure 4-2**), the distance method applied had a marked effect on the pattern of improvements for both natural and non-natural specificity. For distance method 1, natural and non-natural improvements had their greatest frequencies between 8 and 10 Å and between 14 and 16 Å, which were relatively close. Distance method 2 did not have a clear distinction in this regard while method 3 had a tighter distribution overall. Improvements towards non-natural and natural substrates were both highly represented in the groupings from 4 to 8 Å.

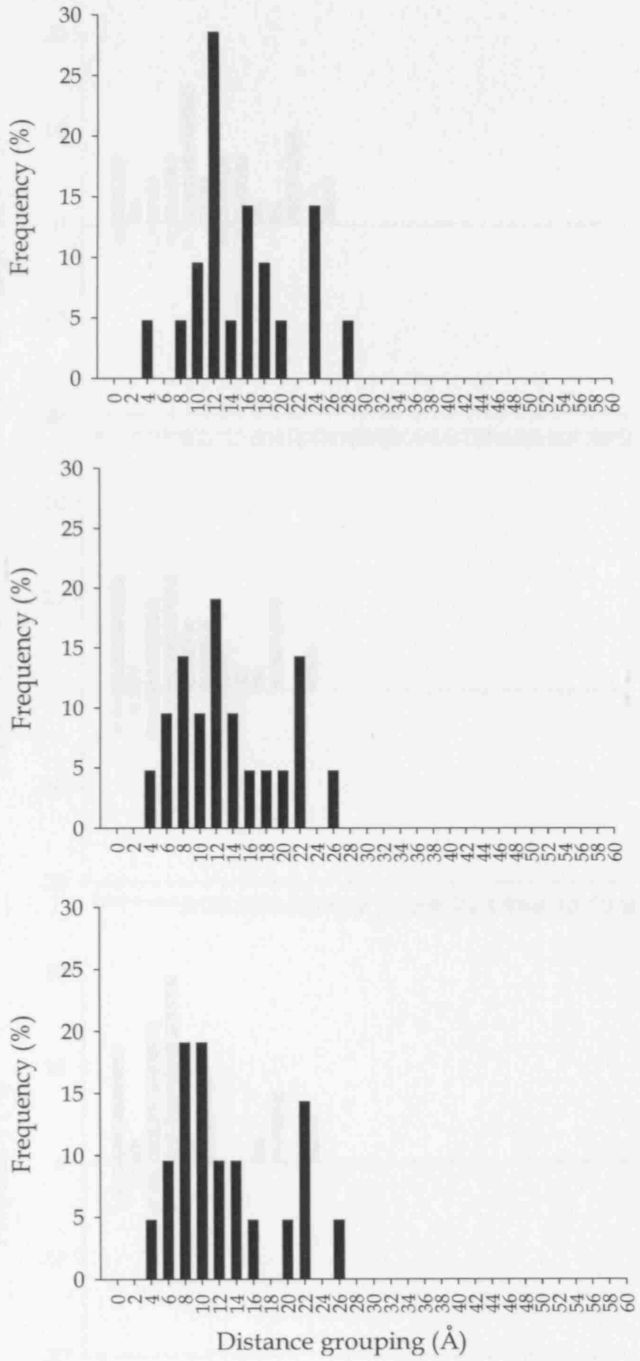

Figure 4-1. Histograms of distances for activity-enhancing mutations (DE). Top = distance method 1, middle = distance method 2, bottom = distance method 3. Each distance grouping spans 2 Å. Total number of mutations = 21.

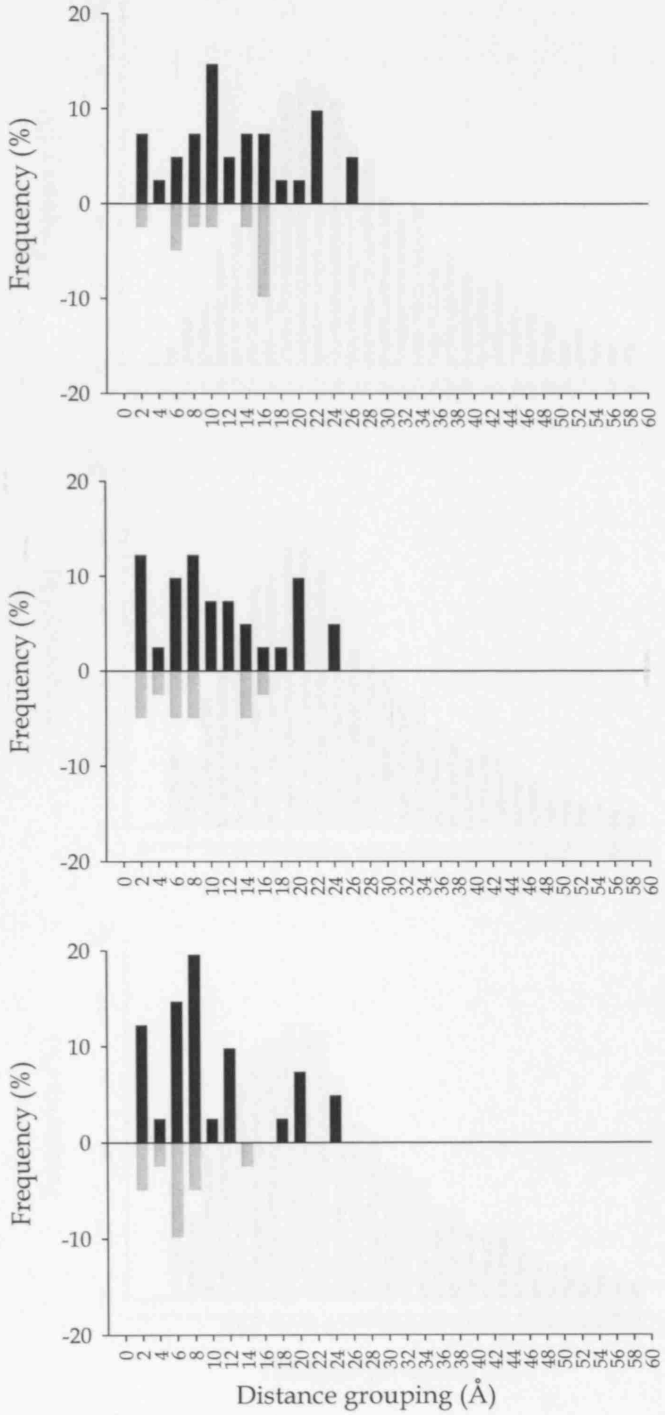

Figure 4-2. Histograms of distances for specificity-enhancing mutations (DE). Top = method 1, middle = method 2, bottom = method 3. Positive frequencies (black bars) indicate enhancements towards non-natural substrates while "negative" frequencies (grey bars) indicate enhancements towards natural substrates. Each distance grouping spans 2 Å. Total number of mutations = 41.

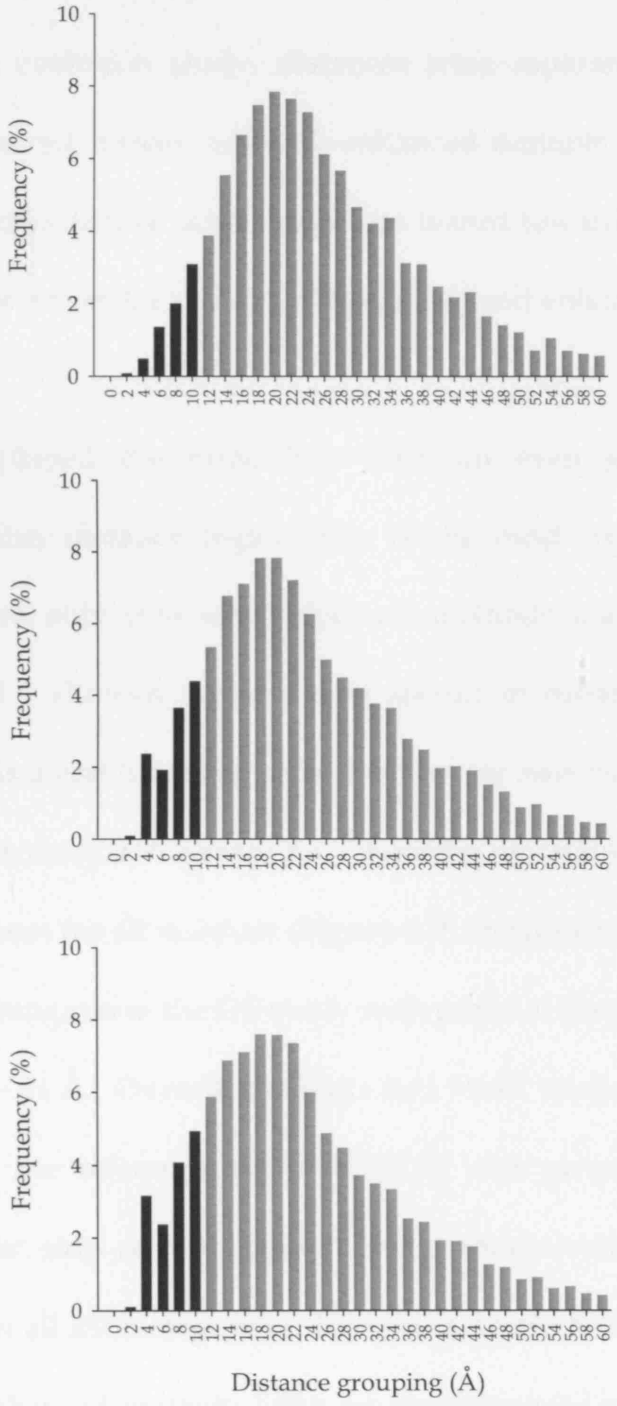

Figure 4-3. Histograms of distances for all residues (DE). Top = distance method 1, middle = distance method 2, bottom = distance method 3. Black bars = active site residues. Each distance grouping spans 2 Å. Total number of residue distances = 5114.

4.3.2.2 Rational evolution study

For the rational evolution study, distances were separated into those of target sites, enhanced mutant sites, all enhanced mutants and all residues. The distance distribution of target sites was biased towards 2 - 4 Å (**Figure 4-6**). Distances of enhanced mutants (**Figure 4-4**) and enhanced mutant sites (

Figure 4-5) displayed the same bias with an even greater emphasis indicating that this distance region was of the most significance. These observations were only true when distance methods 2 and 3 were used; distance method 1 showed a more even spread of distances in all cases. Distance methods 2 and 3 appear to be the best for narrowing the sequence search space for activity and specificity-enhancing mutations.

The distances for all residues (**Figure 4-7**) showed a similar pattern to the equivalent histogram in the DE study with peaks at distances of 15 - 19 Å compared to 14 - 21 Å. Overall, the plots had "bell" shapes with the peaks shifted towards the reference point (< 20 Å) and generally diminishing frequencies either side of the peaks. No residues were found below a distance of 1 Å in all RE histograms. There was a very low frequency in the 1 - 2 Å bin for enhanced mutants, although the frequency for target sites and for all residues in the same bin was also very low indicating that the few residues in this range should not be overlooked in future enzyme engineering strategies.

The most revealing plot is that of the cumulative frequency against cumulative distance for the different distance methods (**Figure 4-8**). It can be seen that the distance method used had a marked effect on the cumulative frequency profiles. Distance method 3 included a larger number of residues at shorter distances than distance method 1 for both enhanced mutants (Δ) and all residues (○). This was expected since the reference point from which to calculate distances using method 3 (catalytic atom, substrate and cofactors) was much larger than for method 1 (catalytic atom). The effect is more pronounced on enhanced mutants, however, in which case the plot for method 3 resembled an exponential rise as opposed to a more linear increase which was seen for method 1.

For a given threshold of success, say 84% of enhanced mutants, it can be seen that the distance cut-off required to achieve this was 14 Å using method 1, and the frequency of all residues within this distance was 23%. Using method 3 for the same threshold of success, it can be seen that the distance cut-off required was 6 Å, and yet only 7% of all residues fall within this distance. This means that selecting more regions of catalytic association effectively reduces the total target residues required for successful enhancement. In other words, the total number of inclusive residues required for a given level of success reduces faster than the total number of residues within the cut-off range increases when using distance method 3 over 1. The discrepancy is 7% of all residues using method 3 compared to 23% of all residues using method 1, which represents at least a 3-fold

reduction in target residues. If a target of 7% of all residues is sought using method 1, the success threshold decreases to only 53%. This highlights the importance of considering substrate and cofactor binding sites when correlating improvements in activity and specificity with distances. In previous studies, only a key catalytic atom had been chosen as a reference point (Morley and Kazlauskas, 2005) neglecting the influence of substrates and cofactors. Distance method 2 (catalytic atom and substrate) gave a similar result to 3 although it did not narrow down the inclusive residues as much, and so the data was not shown.

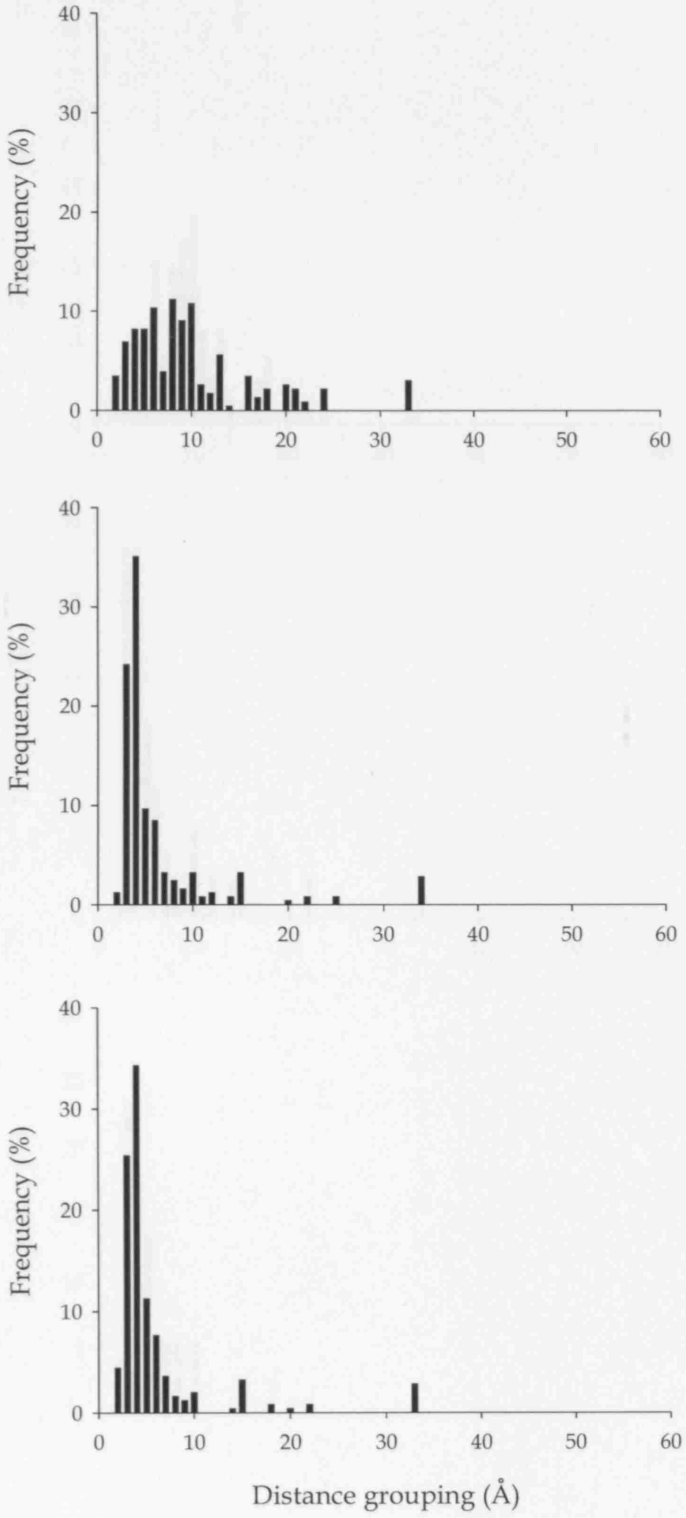

Figure 4-4. Histograms of distances for enhanced mutants (RE). Top = distance method 1, middle = distance method 2, bottom = distance method 3. Each distance grouping spans 1 Å. Total number of mutations = 168.

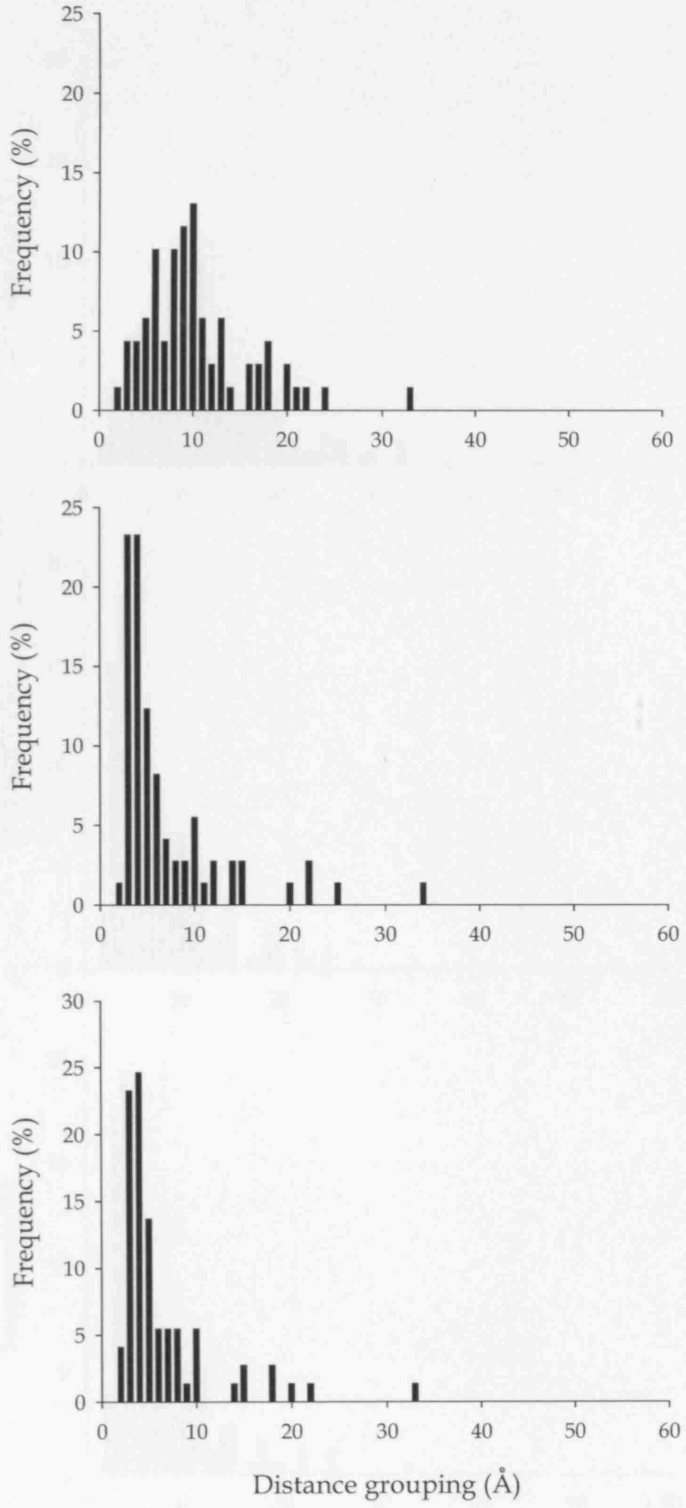

Figure 4-5. Histograms of distances for enhanced mutant sites (RE). Top = distance method 1, middle = distance method 2, bottom = distance method 3. Each distance grouping spans 1 Å. Total number of enhanced mutant sites = 64.

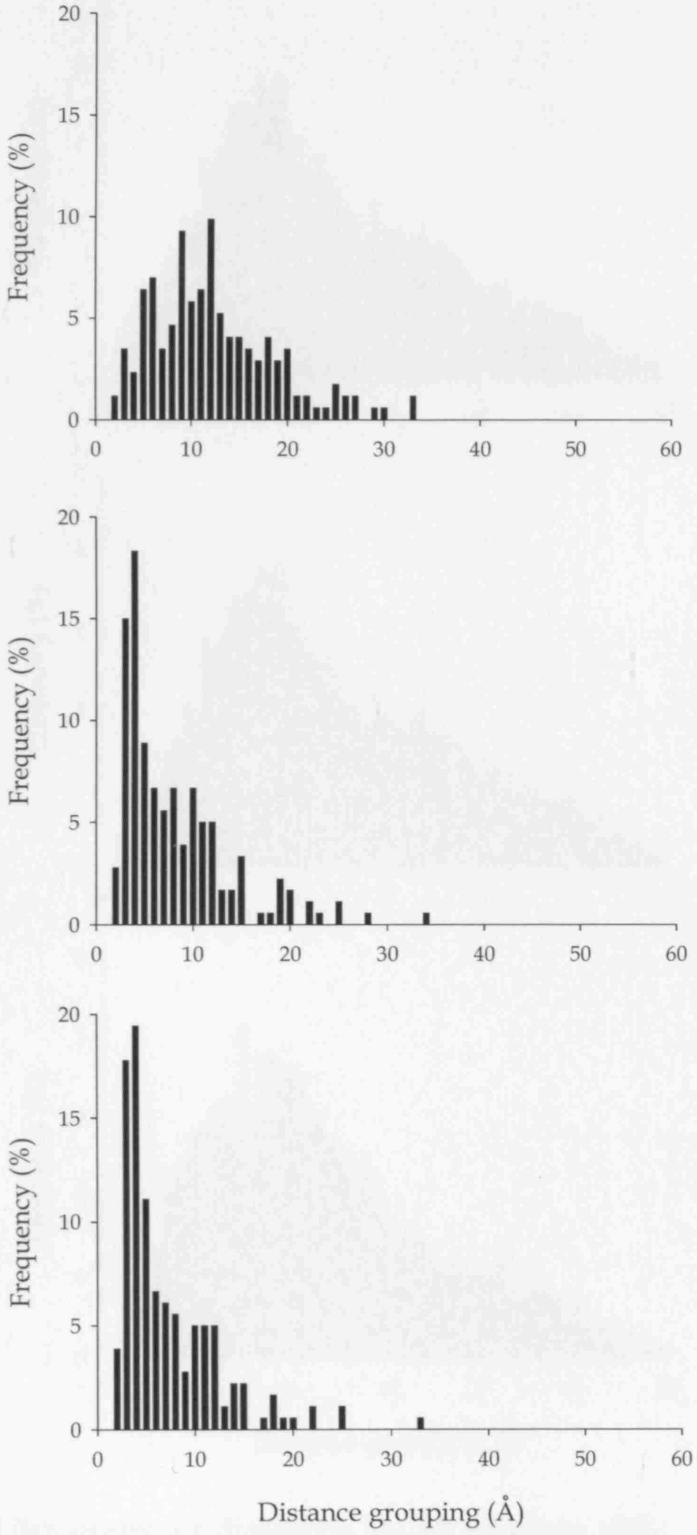

Figure 4-6. Histograms of distances for all target sites (RE). Top = distance method 1, middle = distance method 2, bottom = distance method 3. Each distance grouping spans 1 Å. Total number of target sites = 157.

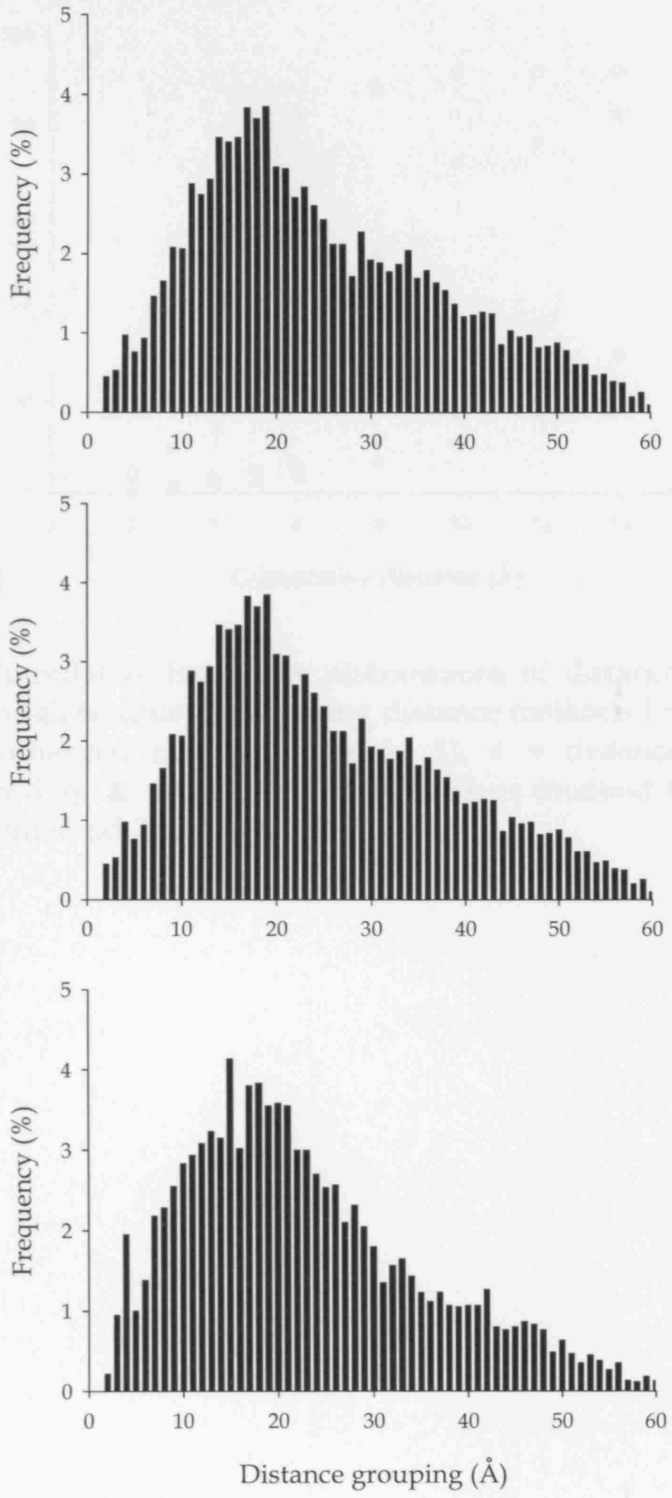

Figure 4-7. Histograms of distances for all residues (RE). Top = distance method 1, middle = distance method 2, bottom = distance method 3. Each distance grouping spans 1 Å. Total number of residue distances = 5564.

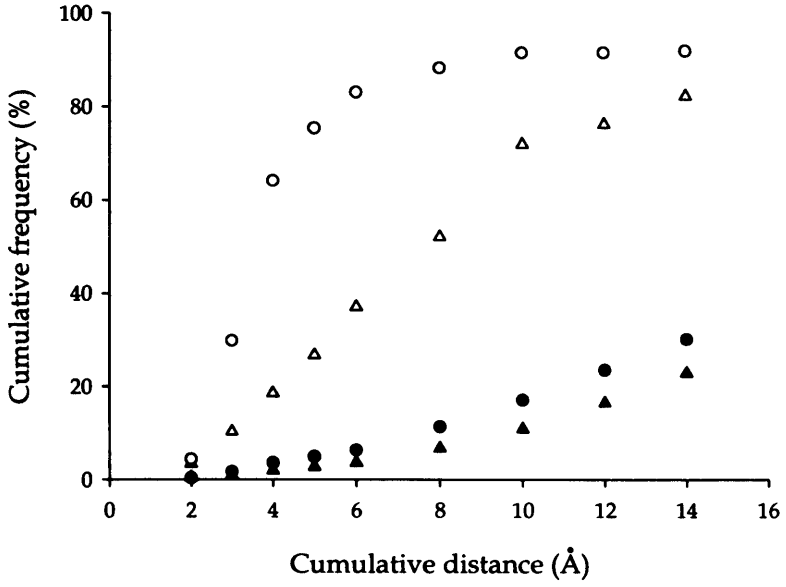

Figure 4-8. Cumulative frequency distributions of distances for enhanced mutants and for all residues (comparing distance methods 1 and 3) (RE). Δ = distances of enhanced mutants (method 1), ○ = distances of enhanced mutants (method 3), ▲ = distances of all residues (method 1), ● = distances of all residues (method 3).

4.3.3 Frequency distributions of entropies

4.3.3.1 Directed evolution study

Histograms were plotted showing entropy frequency distributions for each of the following: (1) activity-enhancing mutations; (2) specificity-enhancing mutations; (3) active-site-only residues; and (4) for all residues combined. Note that there were three versions of what constituted active-site-only residues depending on which distance method was used. Since there was no discernible difference between the plots in this instance, only method 1 is shown.

The histogram for activity-enhancing mutations (**Figure 4-9** - top) showed a fairly even distribution of entropies. The likelihood of a mutation resulting in enhanced activity was therefore regardless of the residue's degree of conservation. Entropies of specificity-enhancing mutations (**Figure 4-9** - bottom) did not have the same distribution as activity-enhancing mutations. There was a difference between mutations increasing non-natural specificity (shown as positive frequencies) and mutations increasing natural specificity (shown as negative frequencies). The first grouping of 0 to 0.1 entropy (the most highly conserved residues) did not register any improvements in specificity towards the natural substrate while improvements towards non-natural specificity registered the highest frequency in this grouping (~10%). This indicated that mutations at highly conserved sites were more likely to have increased specificity towards a non-natural substrate than towards a natural substrate supporting some previous

studies on this subject (Hibbert *et al.*, 2008; Hibbert *et al.*, 2007; Cochran *et al.*, 2006). The rationale behind this was that highly conserved residues in the active site had naturally evolved this way for efficient substrate and/or cofactor binding, and that mutating these would more likely be amenable to changes in non-natural activity. To support this, the conserved residues responsible for loss of function mutations were shown to be those in direct contact with a natural substrate or cofactor during catalysis. It can also be seen in this study that the remaining bins for the less conserved residues registered an equal spread of frequencies for improvements towards both natural and non-natural specificity.

The frequency distribution of active-site-only residues (**Figure 4-10** - top) showed an overwhelming bias of residues in the first bin of 0 to 0.1 entropy (> 20%) compared to the rest of the bins (< 10%). From this it can be concluded that the region within 10 Å of a key catalytic atom tended to be the most highly conserved with respect to the constituent residues. This was not unexpected, however, since the active site is the region of the enzyme most concerned with substrate and cofactor binding, and catalysis. Of all the structural elements comprising an enzyme and the purposes they serve, active site residues would be expected to have the most direct influence on (natural) substrate catalysis, which is deemed to be the primary function of the enzyme. It follows that these residues would need to remain largely unchanged across a range of homologous sequences.

The frequency distribution of activity-reducing mutations could not be determined as this study was confined only to *improvements* in enzyme activity. Generally, reductions in activity were not documented in the papers used for this study. However, there were instances in which activity fell following mutations at highly conserved sites (Hibbert *et al.*, 2008; Hibbert *et al.*, 2007; Cochran *et al.*, 2006), and whilst it is acceptable to conclude that mutating highly conserved sites was certainly the most likely group to improve activity, it should be noted that the effect can also be negative. Locating sites that resulted in a loss of activity upon mutation were not necessarily counter-productive, however. In many directed evolution studies, it has often been useful to locate residues that could "knock out" activity since these may be considered useful candidates for saturation mutagenesis. After the initial probing step to locate knock-out sites or sites important for catalysis, improved mutants were often found in the follow up stage (Wada *et al.*, 2003; Ni *et al.*, 2002), and in one study rational mutations were intentionally directed at important catalytic sites before a section of the gene was randomly mutated (Peimbert and Segovia, 2003). Of the examples used in this study, however, the various research groups proceeded with their studies differently. Some preferred to follow up initial library creation with second (and sometimes third) rounds of random mutagenesis (Funke *et al.*, 2005; Sacchi *et al.*, 2004; Leemhuis *et al.*, 2003; Williams *et al.*, 2003; Sio *et al.*, 2002) whilst others combined beneficial mutations (Jennewein *et al.*, 2006; Fujii *et al.*, 2005; Wilkinson *et al.*, 2004). The

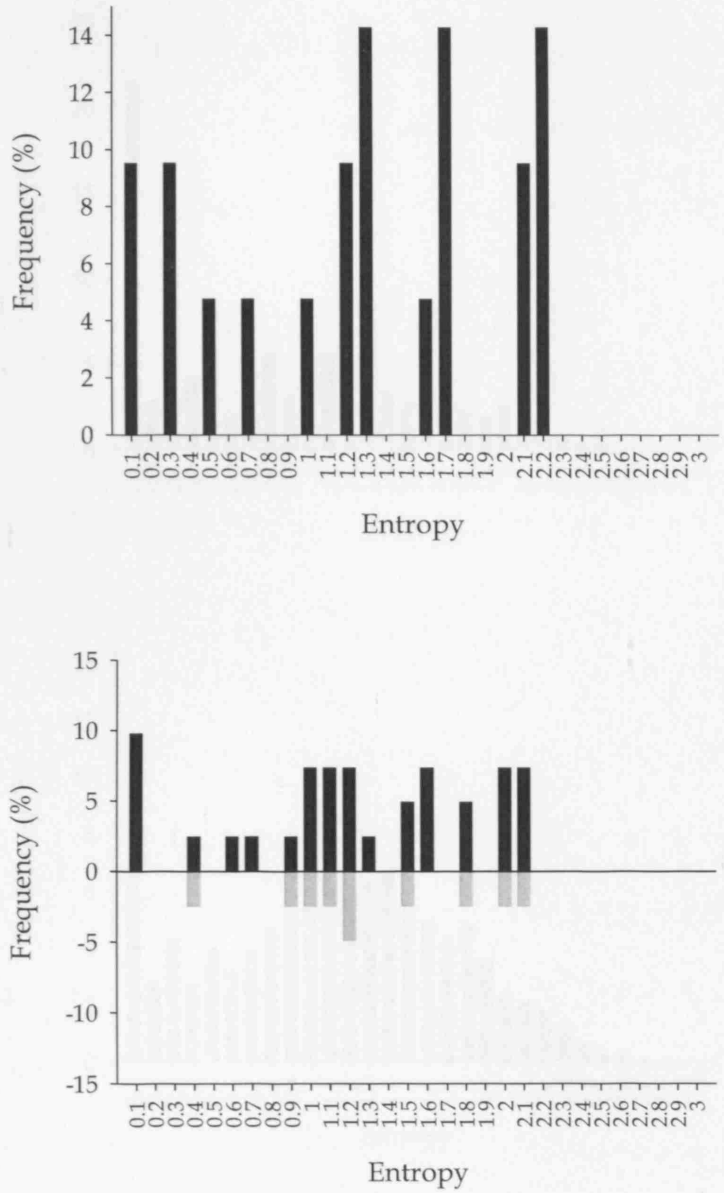

Figure 4-9. Histograms of entropies for activity-enhancing mutations (top) and specificity-enhancing mutations (bottom) (DE). Bottom: positive frequencies (black bars) indicate enhancements towards non-natural substrates while "negative" frequencies (grey bars) indicate enhancements towards natural substrates. Each entropy grouping spans 0.1 absolute entropy. Total number of activity-enhancing mutations = 21; total number of specificity-enhancing mutations = 41.

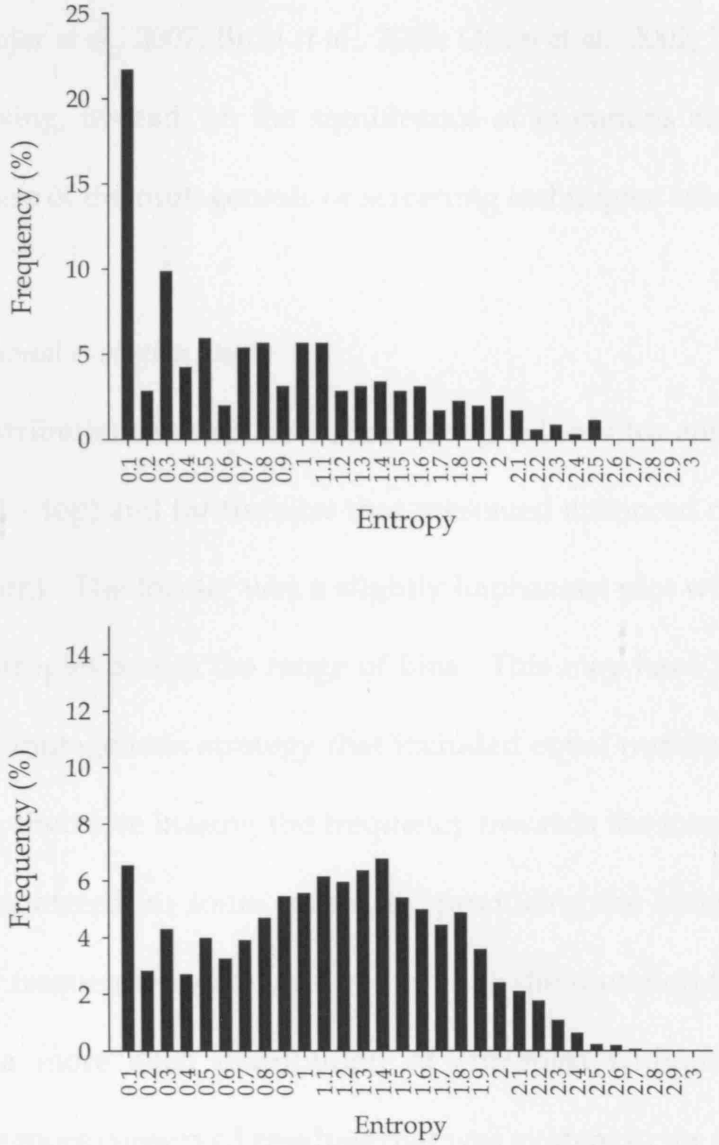

Figure 4-10. Histogram of entropies for active-site-only residues (distance method 1) (top) and for all residues (bottom) (DE). Each entropy grouping spans 0.1 absolute entropy. Total number of active-site-only residues = 355; total number of residues = 5114.

remaining groups did not proceed with any further mutagenesis step (Mueller-Cajar *et al.*, 2007; Broo *et al.*, 2002; Otten *et al.*, 2002; Delagrave *et al.*, 2001) focussing, instead, on the significance of mutations already found or the usefulness of the mutagenesis or screening techniques used.

4.3.3.2 Rational evolution study

Entropy distributions for RE examples were produced for enhanced mutants (**Figure 4-11** - top) and for the sites that presented enhanced mutants (**Figure 4-11** - bottom). The former was a slightly haphazard plot with equally high and low entropies across the range of bins. This may have been a result of not using a mutagenesis strategy that included equal numbers of mutations at each site, therefore biasing the frequency towards the most mutated sites. This was countered, to some extent, by producing the latter plot in which "good site" frequencies were used rather than the mutation frequency. This plot gave a more even distribution of entropies without any tendency towards the more conserved residues that was evident in the DE plot.

A plot showing the distribution of sites targeted for mutagenesis (**Figure 4-12** - top) revealed that highly conserved residues were by far the most targeted, and yet did not present enhanced mutants at a greater rate than the other sites generally. Active-site directed mutagenesis approaches, therefore, did not produce enhanced mutants primarily at highly conserved sites. As discussed before, this was likely to be due to the conflict of interest

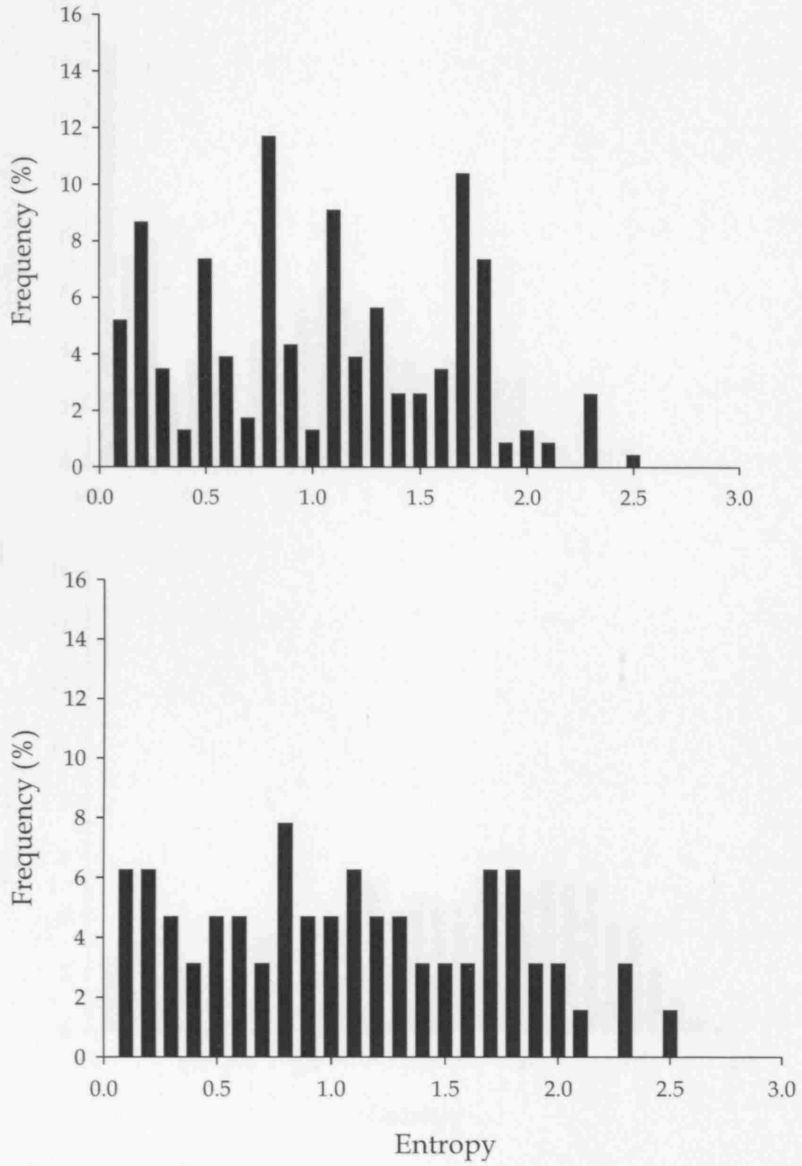

Figure 4-11. Histogram of entropies for enhanced mutants (top) and for enhanced mutant sites (bottom) (RE). Each entropy grouping spans 0.1 absolute entropy. Total number of enhanced mutants = 168; total number of enhanced mutant sites = 64.

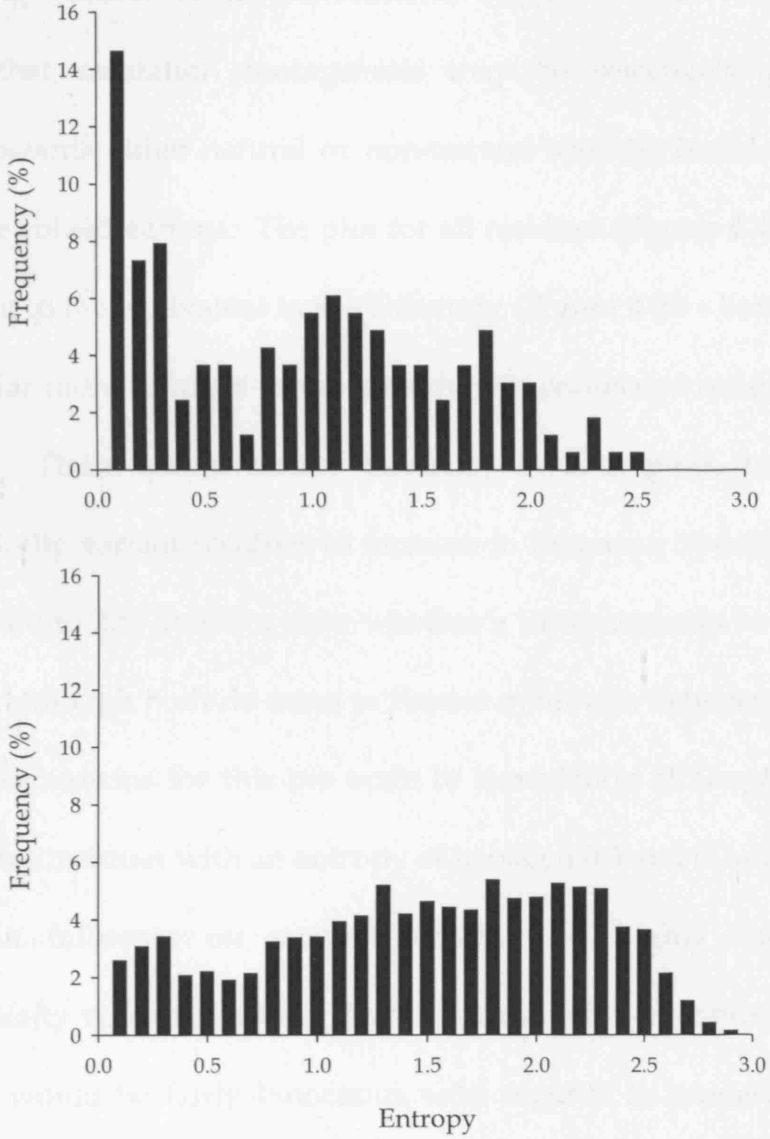

Figure 4-12. Histogram of entropies for target sites (top) and for all residues (bottom) (RE). Each entropy grouping spans 0.1 absolute entropy. Total number of target sites = 157; total number of residues = 5564.

in improving natural versus non-natural activity. This was a further indication that saturation mutagenesis may be selectively guided for evolution towards either natural or non-natural activity, based on entropy and distance considerations. The plot for all residues (**Figure 4-12** - bottom) was different to the equivalent in the DE study (**Figure 4-10** - bottom) in that there were far more residues in the most highly conserved residue bin with the former. There was a similar tendency in both plots, however, for phylogenetically variant residues to increase in frequency specifically in the range from 1.0 to 2.5. It is not clear whether a maximum can be considered in the plots although both do seem to have a minimum between 0.3 and 0.7 entropy. The reasons for this are open to speculation although it may be suggested that residues with an entropy of between 0.3 and 0.7 may not have as great an influence on enzyme function as highly conserved or phylogenetically variant residues. Such residues with entropies of between 0.3 and 0.7 would be fairly innocuous with regards to properties such as activity, specificity and thermostability, and therefore may not crop up as frequently as highly conserved or phylogenetically variant residues.

4.3.4 Energy change correlated with distance and entropy

$\Delta\Delta G$ is a measure of improvement in activity on a natural substrate and is defined as the change in activation free energy caused by a mutation. Negative $\Delta\Delta G$ values indicate a reduction in activity in which case data was excluded from figures. $\Delta\Delta\Delta G$ is a measure of improvement in specificity.

Mutant activity versus wild-type activity is compared specifically for a non-natural substrate versus a natural substrate accounting for the extra "Δ". Both positive and negative ΔΔΔGs may be perceived as an improvement in specificity whether directed towards the non-natural substrate (positive) or natural substrate (negative). Section 4.2.2 details the calculation methodology. Only DE examples were considered for correlations in this section since these mutations were spread across the whole genome rather than targeted on the basis of distance or entropy.

When ΔΔG and ΔΔΔG were plotted against entropy (**Figure 4-13**), there was a fairly even scatter across both axes, however, every point fell below an entropy of 2.1. The significance of this is not considered great since entropies greater than this did yield enhancements from the RE study. There was a clustering of specificity-enhancing mutations close to (or at) zero entropy, which again, highlighted the influence of highly conserved residues on specificity improvements. Interestingly, there appeared to be no correlation between the magnitude of improvement and entropy, and all but 7 mutations out of 88 (92%) boasted an energy change of less than 10 kcal per mol. Of these 7 special cases, all were specificity-enhancing with 4 of them having unusually large values of between 15 - 30 kcal per mol. Such large increases may have been possible since a change in specificity could have been the result of both an increase in activity on one substrate together with a decrease on another, whereas a change in activity could only have been the result of an increase on one substrate. The degree of enhancement was also

largely dependent on the experimental conditions for a particular study and the kinetic values that were presented in each paper, which were subject to variations. This makes it difficult to compare data across enzymes. Evolvability also depends on the state of evolution of the enzyme studied. Not all enzymes are fully evolved to the diffusion limit ($kcat/K_M \sim 10^9\ M^{-1}s^{-1}$) as this may not be necessary for cell survival.

Figure 4-14 is a plot of $\Delta\Delta G/\Delta\Delta\Delta G$ versus mutation distance using method 1. There appears to be a convergence of the points to zero energy change as the distance increased, although there were a few exaggerated points with large $\Delta\Delta G/\Delta\Delta\Delta G$ values outside this trend at the longer distances. This suggested that mutating closer to the active site typically resulted in larger enhancements. This would appear to be consistent with the hypothesis that the further away a mutation is, the smaller the effect is on the active site structure and its associated regions of substrate and cofactor binding. Smaller effects may be more likely to result in smaller improvements, as shown in the plot. Every point was below a distance of 28 Å. This may be considered a distance cut-off (using distance method 1) beyond which mutations enhancing either activity or specificity are far less likely.

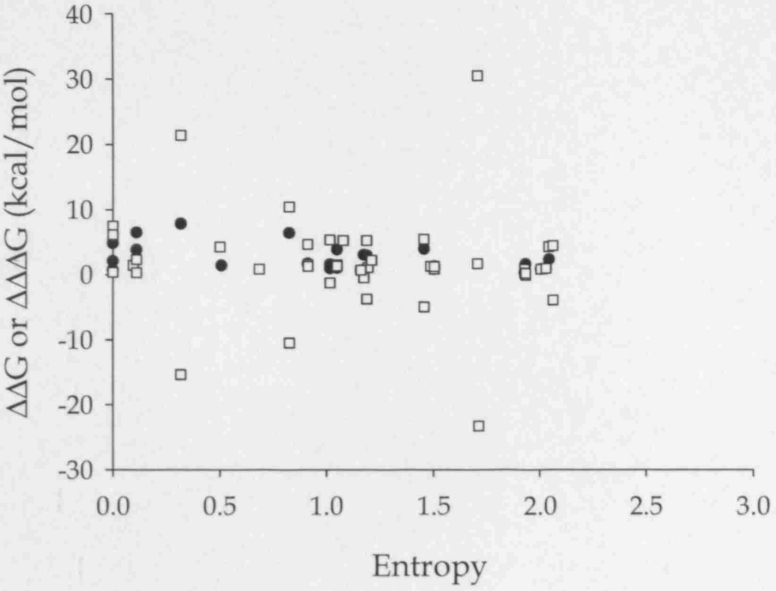

Figure 4-13. Scatter plot of activation free energy change versus entropy for activity-enhancing and specificity-enhancing mutations (DE). Legend: ▫ = specificity-enhanced mutants (ΔΔΔG); ● = activity-enhanced mutants (ΔΔG).

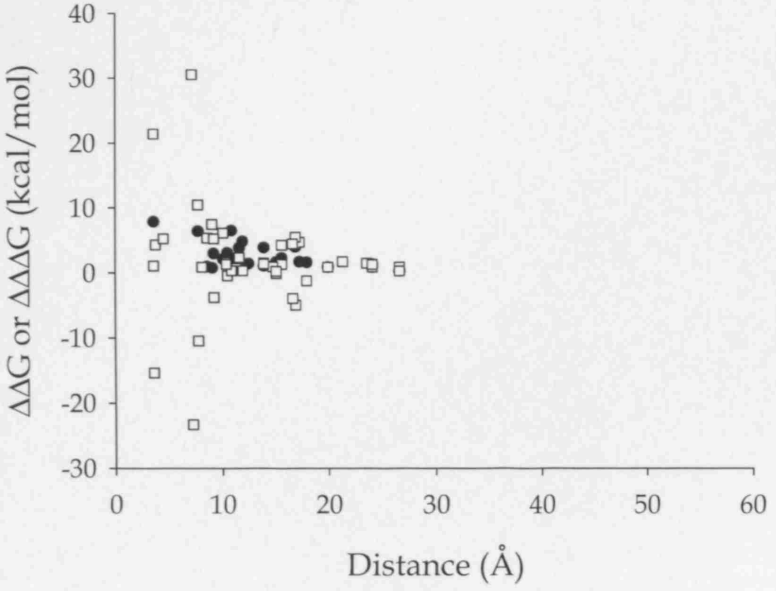

Figure 4-14. Scatter plot of activation free energy change versus distance (method 1) for activity-enhancing and specificity-enhancing mutations (DE). Legend: ▫ = specificity-enhanced mutants (ΔΔΔG); ● = activity-enhanced mutants (ΔΔG).

4.3.5 Entropy-distance correlations

A series of plots were produced examining the correlations between distance and entropy for the directed and rational evolution examples. The distance method used was 3, which was chosen on the basis that it was found previously to be the most useful for narrowing down sequence search space (see Section 4.3.2.2). **Figure 4-15** is a plot focussing on the activity-enhancing (Δ) and specificity-enhancing mutations (▫) as well as mutations that enhanced both activity and specificity (•) from the directed evolution study. All mutations appeared to be fairly evenly scattered, while falling below set limits for both distance and entropy. All entropies fell below 2.1 while all distances fell below 28 Å. **Figure 4-16** is a plot of distance versus entropy for all residues from the directed evolution examples. It can be seen that the upper limits extended to 2.7 for entropy and to 70 Å for distance, although most points clearly fell within the limits defined earlier for enhancing mutations (2.1 entropy and 28 Å).

In the case of rational evolution, **Figure 4-17** was plotted to show the correlation between distance and entropy focussing on target sites and enhanced mutant sites. The distribution of points appeared to be similar for both targeted sites and for sites that gave enhanced mutants. However, sites that were targeted tended to be biased within 20 Å of the reference point while sites that gave enhanced mutants were largely within 10 Å (x-axis). By contrast, entropies for either of these two groups (target sites and hit sites) did not appear to be biased towards high or low values (y-axis).

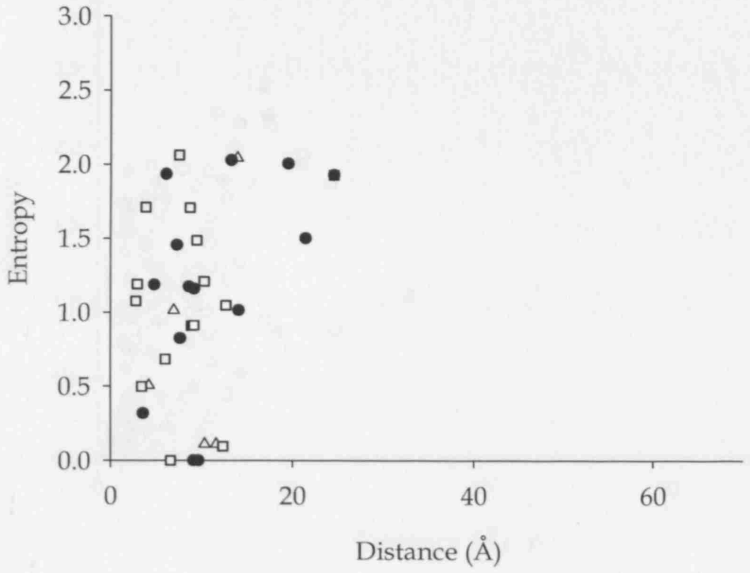

Figure 4-15. Scatter plot of distance (method 3) versus entropy for activity-enhancing and specificity-enhancing mutations (DE). Legend: Δ = activity-enhancing mutations, ▫ = specificity-enhancing mutations and ● = mutations in which both activity and specificity were enhanced.

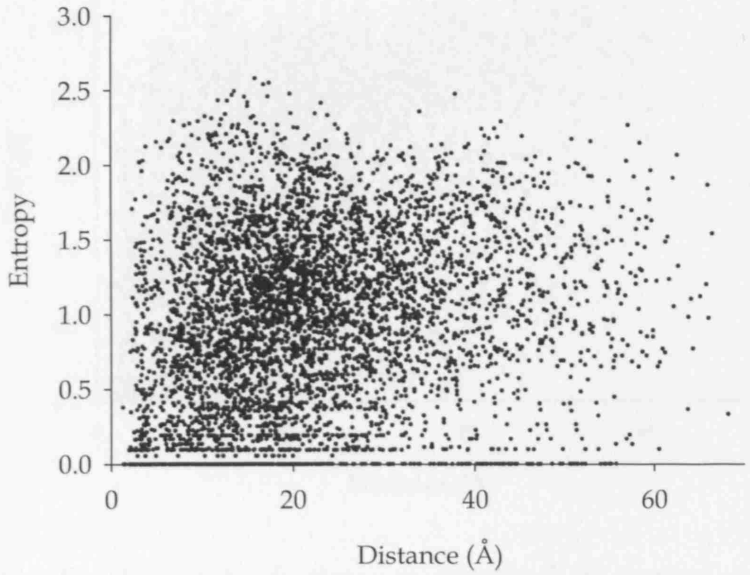

Figure 4-16. Scatter plot of distance (method 3) versus entropy for all residues (DE). Each black dot represents a single residue.

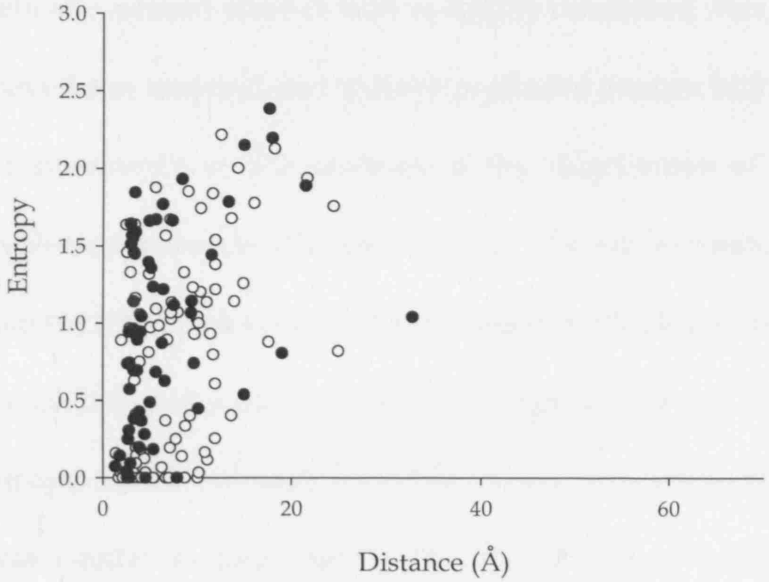

Figure 4-17. Scatter plot of distance (method 3) versus entropy for target sites and enhanced mutant sites (RE). Legend: ○ = target sites, ● = enhanced mutant sites.

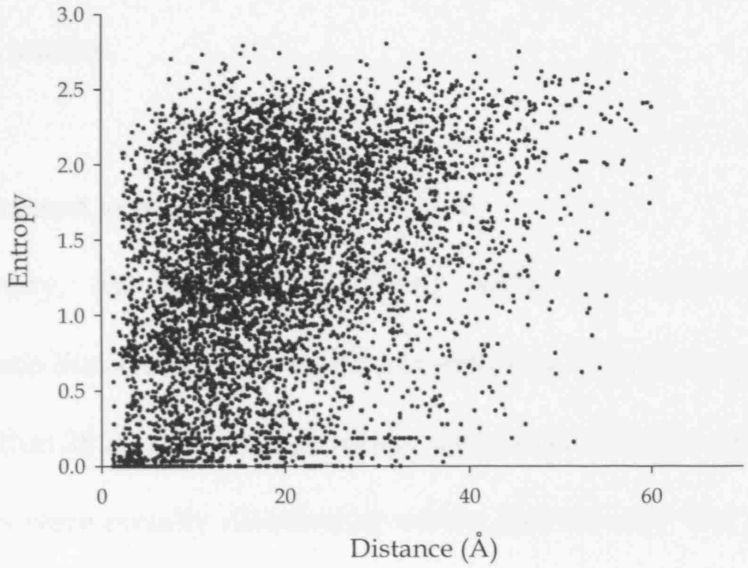

Figure 4-18. Scatter plot of distance (method 3) versus entropy for all residues (RE). Each black dot represents a single residue.

Phylogenetically variant sites as well as highly conserved sites were equally likely to have been targeted and to have produced mutant hits. This finding was noted previously in the analysis of the distribution of entropies for rational evolution examples (Section 4.3.3.2). The upper limits were slightly greater than for the directed evolution examples with all points falling below an entropy of 2.45 and a distance of 35 Å. **Figure 4-18** is a plot of distance versus entropy for all residues from the rational evolution examples. The pattern was similar to that shown for the directed evolution study with upper limits extending to 2.75 for entropy and 65 Å for distance. Comparison of the two plots indicated that there was no general bias in entropies or distances for enzymes selected in either the directed or rational evolution studies.

4.3.6 Directed versus rational evolution

In summary, the directed evolution study has shown that random mutagenesis libraries produced more mutations closer to a defined reference point (within 28 Å) than further away, whichever distance method was used. Mutations were equally distributed within this cut-off. The vast majority of mutations in the rational evolution study were targeted within 30 Å of the reference point indicating that distance was already deemed somewhat important in selecting mutation sites. Successful mutations, therefore, also occurred primarily within this cut-off although there was a particularly

strong representation within 10 Å of the reference point, especially when distance method 3 was applied.

In terms of entropy distributions the directed evolution study has highlighted the relationship between how conserved a residue is and the likelihood of it producing either an activity-enhancing or specificity-enhancing mutation (see Section 4.3.3.1). This type of analysis was not possible for the rational evolution study since residues were often pre-selected for mutagenesis based on the degree of conservation, which would have biased the results. The mutations were, instead, lumped into a general group of enhancing mutations whether on the basis of activity or specificity. The result was that although there was a strong representation of highly conserved target sites, the sites that gave good mutants were evenly distributed at all degrees of conservation. As mentioned previously, however, one comprehensive study (Hibbert *et al.*, 2008; Hibbert *et al.*, 2007) found that highly conserved residues were mainly responsible for improvements in non-natural activity while phylogenetically variant residues were mainly responsible for improvements in natural activity.

4.3.7 Bovine trypsin active site analysis

Bovine trypsin, as the subject of this project, has been separated for analysis from the rest of the examples in this chapter. Although no reported cases of trypsin enhanced by directed evolution were found (for any species), the findings of this chapter show that improvements to activity and specificity

occur in greater numbers when a mutation is closer to a key active site residue. The active site of bovine trypsin was analysed in detail with respect to the following (1) the constituent residues making up the active site; (2) the distance of active site residues from a chosen instance residue; and (3) the entropy of each residue. **Table 4-3** contains the data from this analysis as generated using PyMOL, PyScripter and BioEdit. The instance residue chosen was the lysine residue of the inhibitor molecule, T-Butoxy-ala-val-boro-lys 1,3-propanediol monoester, which was bound to the trypsin molecule in the PDB structure file (PDB ID: 1BTW). Lysine was chosen on the basis that residues with positively charged hydrophilic side chains (*i.e.* arginine and lysine) have been widely reported to be the primary specificity determinants in bovine trypsin either via structural studies (Bode and Huber, 1978; Bode and Schwager, 1975) or enzyme assay (Stewart and Dobson, 1965).

In total 61 out of a possible 223 residues (27%) were found to be within 10 Å of the instance residue. This narrowed down the amount of residues that may be mutated in a targeted approach to enhancing activity or specificity although realistically the amount of residues that may be investigated over the course of the project is limited. A closer look at the structure of trypsin revealed the sites that were directly in contact with the substrate (**Figure 4-19**).

The interaction between the lysine side-chain of the inhibitor and the active site is shown. The binding pocket of the active site has been divided

Position	Residue	Distance (Å)	Entropy
16	ILE	7.50	0.41
17	VAL	6.95	0.96
40	HIS	8.19	1.02
41	PHE	6.03	1.19
42	CYS	5.27	0.09
43	GLY	6.94	0.17
54	SER	9.20	0.50
55	ALA	6.94	0.30
56	ALA	8.64	0.42
57	HIS	1.91	0.13
58	CYS	5.60	0.16
59	TYR	8.84	1.17
94	TYR	8.78	0.42
99	LEU	5.77	1.61
102	ASP	6.15	0.06
138	ILE	7.28	0.80
141	TRP	9.84	0.23
142	GLY	8.35	0.06
143	ASN	8.02	1.12
145	LYS	9.54	2.00
146	SER	9.27	1.03
151	TYR	8.16	2.04
158	LEU	7.80	0.60
160	ALA	9.56	1.14
172	TYR	5.42	0.13
181	PHE	8.21	1.23
182	CYS	8.28	0.03
183	ALA	6.55	1.12
184	GLY	8.29	0.10
184	TYR	8.03	1.22
187	GLY	9.77	0.26
188A	GLY	9.19	0.30
188B	**LYS**	**7.00**	**0.49**
189	**ASP**	**2.44**	**0.21**
190	**SER**	**2.56**	**0.45**
191	CYS	3.66	0.03
192	**GLN**	**3.26**	**0.49**
193	GLY	3.48	0.48
194	ASP	4.62	0.08
195	SER	1.61	0.05
196	GLY	5.72	0.13
197	GLY	6.98	0.24
199	VAL	9.12	0.71
212	ILE	8.01	1.18
213	VAL	4.04	0.24
214	SER	1.98	0.10
215	**TRP**	**3.12**	**0.14**
216	**GLY**	**3.70**	**0.09**
217	**SER**	**5.12**	**2.02**
219	**GLY**	**3.48**	**0.93**
220	CYS	4.36	0.16

221A	ALA	3.96	0.39
221B	GLN	5.94	1.87
222	LYS	9.77	1.61
223	ASN	9.64	1.48
224	LYS	5.11	1.96
225	PRO	4.55	0.17
226	GLY	2.92	0.27
227	VAL	3.25	0.21
228	TYR	4.01	0.22
229	THR	8.64	0.95

Table 4-3. Active site residue properties of bovine trypsin. Residues within 10 Å of the lysine residue of the inhibitor T-Butoxy-ala-val-boro-lys 1,3-propanediol monoester and their entropies. Residues selected for MSSM in bold.

into two distinct regions; the first has been named BP1 (pink) and refers to residues Lys188, Asp189, Ser190 and Gln192 whilst the second has been named BP2 (yellow) and refers to residues Trp215, Gly216, Ser217 and Gly219. The two regions have been divided in this way to suit a directed evolution strategy that facilitates multiple-site saturation mutagenesis (MSSM). BP2 contains four consecutive residues while BP1 contains four residues with the Cys191 residue left unmutated. This residue, along with Cys220, form a disulphide bridge (green) in a region that is in contact with the binding pocket and is thought to provide crucial stability to the active site. Upsetting this key structural aspect would likely reduce the chances of obtaining useful mutants so it was decided not to mutate these residues. It may also be seen that the two residues together provide a point of connection between BP1 and BP2 further suggesting that the Cys191-Cys219 disulphide bridge is responsible only for stability while the surrounding active site residues may be responsible for specificity.

A study on the evolutionary divergence and conservation of trypsin has shown that secondary structure was strictly conserved across bovine, *Streptomyces griseus* (bacterium) and *Fusarium oxysporum* (fungus) trypsins (Rypniewski *et al.*, 1994). A more thorough alignment of all available sequences revealed that insertions and deletions occurred only in regions corresponding to loops between the secondary structure elements in the known crystal structures. Conserved residues, however, clustered around the active site. Almost all conserved residues could be associated with one of

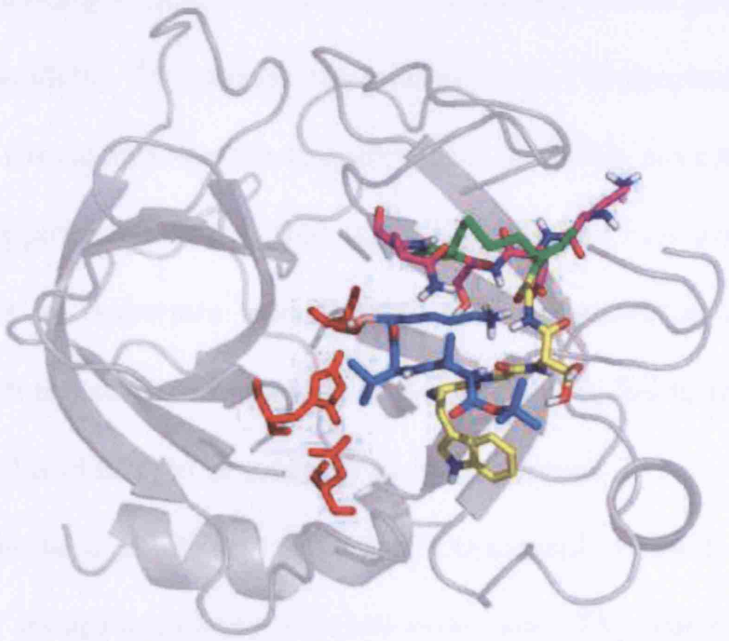

Figure 4-19. Active site of trypsin bound to inhibitor T-Butoxy-ala-val-boro-lys 1,3-propanediol monoester (created from PDB structure file 1BTW). Colours: Catalytic triad His57, Asp102 and Ser190 = red; binding pocket site 1 (BP1) = pink (backbone only); binding pocket site 2 (BP2) = yellow (backbone only); inhibitor = blue (backbone only); Cys191-Cys219 disulphide bridge = green. Atom colours: N = blue, H = white, O = red. The lysine side-chain of the inhibitor may be seen interacting with BP1 and BP2 sites.

the basic functional features of the protein: zymogen activation, catalysis and substrate specificity. In contrast, the residues of the hydrophobic core of the protein and the calcium ion binding sites were generally not conserved. This evidence supports the notion that the active site residues are the primary determinants of substrate specificity. Consideration of this evidence together with the structural analysis of the active site, led to the selection of two sites, BP1 and BP2, to be targeted for mutagenesis.

MSSM is a relatively unused mutagenesis method with limited examples of its application in directed evolution. The major advantage of this method is that it may generate much larger library sizes than traditional single-site saturation mutagenesis. For example, mutating either of sites BP1 or BP2 by MSSM would generate a library size of 20^4 = 160,000. The drawback is that this library size is far too large to be screened by traditional microplate-based methods. An alternative method for isolating mutants of interest would need to be developed. Success has been achieved in the past with rat anionic trypsin by using a nutritional selection method whereby a library is transformed into an auxotrophic strain of *E. coli* and then subjected to a selection pressure to test for substrate specificity (Evnin *et al.*, 1990).

4.3.8 Sites important for self-proteolysis

In addition to trypsins from various species, other notable examples of self-proteolysing enzymes include chymotrypsin (Kumar and Hein, 1970), subtilisin (Wells and Powers, 1986) and thermolysin (Fontana *et al.*, 1986).

The phenomenon appears to be compounded in digestive endoproteases with broad specificity. Endoproteases not used in digestion that have a more specialised use such as enteropeptidase (trypsin activator) and factor Xa (thrombin activator) require a greater degree of specificity for their role in metabolism. These proteases possess a high specificity towards their particular recognition sequence; DDDDK in the case of enteropeptidase and IEGR in the case of factor Xa. As a result these proteases must have a lower risk of self-proteolysis since exact matches of their recognition sequences are much less likely to occur in the protease's own sequence than straightforward single residue recognition sequences (R or K). Perhaps due to the less specific requirements of digestive proteases (to break down all dietary proteins for absorption in the gut), natural evolution may have dictated that these enzymes including trypsin and chymotrypsin retain a broad specificity. It follows that such enzymes should be more susceptible to self-proteolysis as a side-effect of their main function.

Figure 4-20 emphasizes the vulnerability to self-proteolysis faced by bovine trypsin. Arginine and lysine residues tend to be located close to the surface of a trypsin molecule rather than buried within. This may be due to their hydrophilicity and explains how trypsin molecules can access the arginine and lysine residues of another and begin proteolysis. Since some may be more susceptible than others, the important sites needed to be ascertained for an enzyme engineering strategy to improve the resistance of bovine trypsin to self-proteolysis.

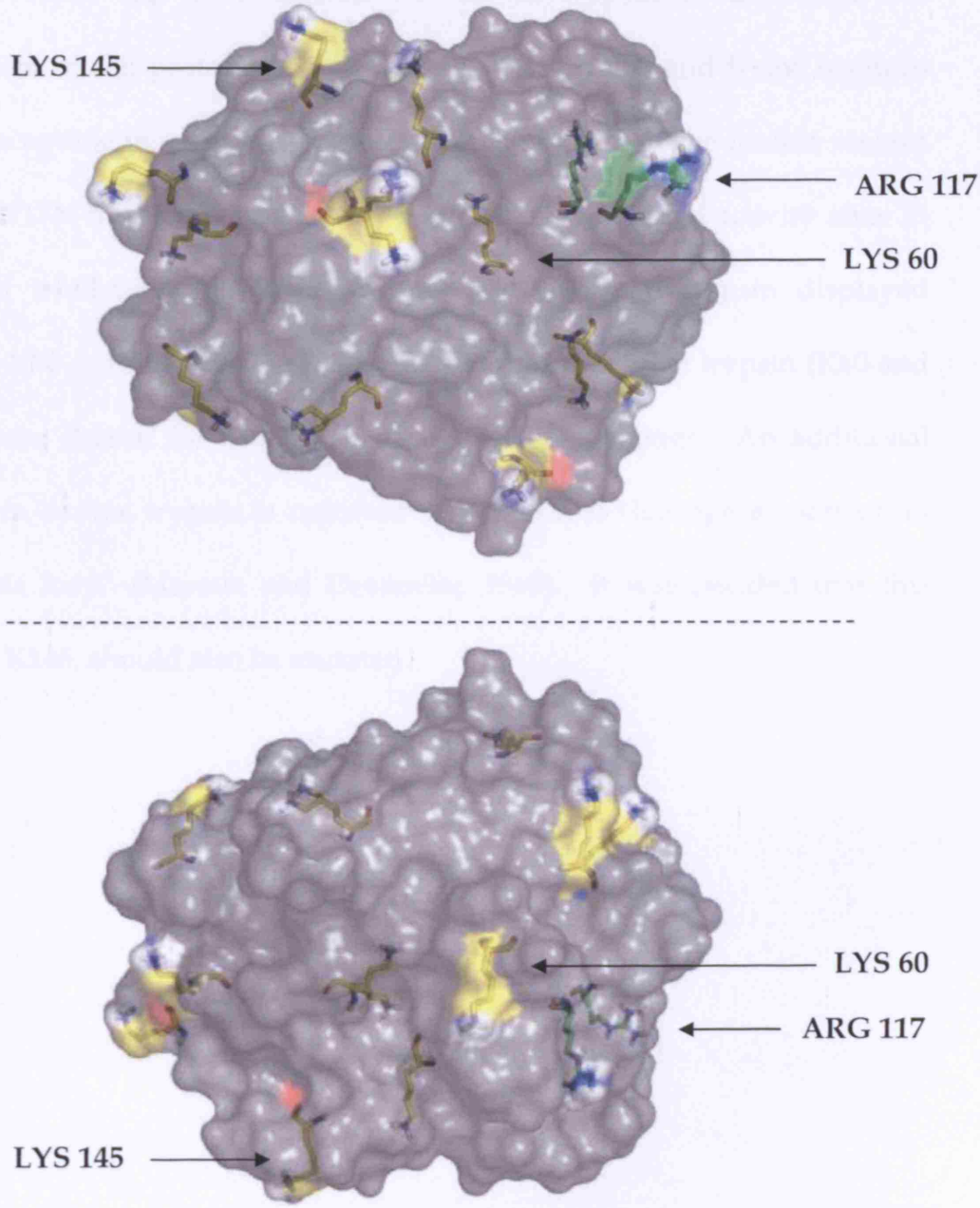

Figure 4-20. Trypsin molecule with arginine and lysine residues highlighted (created from PDB structure file 1BTW). The molecule has been rotated 180° about the x-axis. Arginine and lysine residues are coloured green and yellow respectively (backbone only). Atom colours: N = blue, H = white, O = red. The highlighted residues have a varying degree of surface-exposure shown by any surface colouring that is not grey.

A study has been carried out in rat trypsin to determine the mechanism of self-proteolysis and identify the arginine and lysine residues most susceptible in its sequence (Varallyay *et al.*, 1998). The double mutant K61N/R117N displayed activity close to 100% of the initial activity after 25 hours of incubation at 37 °C whereas the wild-type trypsin displayed roughly 10% activity. The homologous residues in bovine trypsin (K60 and R117) were chosen for mutation based on these findings. An additional residue in bovine trypsin is reported susceptible to cleavage as part of an "autolysis loop" (Maroux and Desnuelle, 1969). It was decided that this residue, K145, should also be mutated.

5 Targeted mutagenesis of trypsin

5.1 Introduction

The benefits of combining directed evolution and rational design approaches for improving properties, such as substrate specificity, have been reviewed in literature (Chica *et al.*, 2005). Examples of so called "semi-rational" approaches have been described whether they utilise computational methods (Dwyer *et al.*, 2004; Hayes *et al.*, 2002) or a purely physico-structural analysis (Hibbert *et al.*, 2007; Chockalingam *et al.*, 2005) to determine mutagenesis targets. MSSM may be considered a semi-rational approach in line with these examples. In this chapter two different protein engineering strategies aimed at improving trypsin properties were carried out. Firstly, a rational design strategy was followed to reduce the rate of self-proteolysis of the active enzyme. Following this, an active-site directed saturation mutagenesis approach was used with the aim of altering the substrate specificity of trypsin. The results of the following are described: (1) fully rational site-directed mutagenesis (SDM) followed by a stability screen and (2) semi-rational MSSM of specificity hotspots followed by a nutritional selection.

Targets for engineering both the specificity and autolysis of bovine trypsin were proposed in the last chapter. The two different approaches to enzyme engineering were chosen based on the characteristic to be enhanced and on available knowledge of the enzyme's structure. For specificity improvements, a directed evolution strategy (Section 1.3.2.2) was put

forward, which was to generate mutant libraries encompassing all possible mutants at target sites BP1 (K188, D189, S190 and Q192) and BP2 (W215, G216, S217 and G218). Target sites could have been extended to include the two cysteine residues C191 and C219 (**Figure 4-19**) which are neighbouring residues to both sites BP1 and BP2. However, the disulphide bridge that these residues give rise to, has been reported to provide essential structural support to the binding pocket (Wang *et al.*, 1997). As an integral feature of the binding pocket, removing the disulphide bridge was expected to reduce the chances of generating a useful mutant. It was therefore decided that these two residues should remain unchanged. For improving resistance to self-proteolysis, a rational design strategy (Section 1.3.2.1) was proposed. Structural analysis has previously shown that a single bovine trypsin molecule contains a number of residues vulnerable to proteolytic cleavage by other bovine trypsin molecules (Section 4.2.7).

New cases employing a rational design strategy to change substrate specificity continue to surface (Kristan *et al.*, 2007; Jourden *et al.*, 2007; Host *et al.*, 2006) although a rational design approach was not considered applicable to trypsin since there were no firm arguments on which to base single mutations. While past attempts have resulted in specificity enhancements in rat anionic trypsin of between 30 - 40 fold towards lysine over arginine, the overall activity ($kcat/K_M$) on both arginine and lysine substrates fell by a minimum of 300-fold (Craik *et al.*, 1985b). Traditional saturation mutagenesis (Wells *et al.*, 1985) was also not considered suitable. This

method would randomise each position independently rather than combinatorially. Each position would encompass 20 mutations (including the wild-type) at the four target positions in the binding pocket for a total library size of 80. Even though mutations would be targeted towards the substrate-binding site and may result in improvements to specificity, multiple-site saturation mutagenesis (MSSM) samples significantly more structural configurations of the binding pocket. Using such an approach, for example by combinatorial cassette mutagenesis (Reidhaar-Olson and Sauer, 1988), the library sizes for sites BP1 and BP2 were 160,000 (20^4) each; the calculation was based on each site consisting of four residues being mutated to any of the 20 possible residues. The majority of these were likely to be inactive or unfolded due to major structural modifications, however, the likelihood of obtaining mutants with different specificities was also greater. In comparison with single-site saturation mutagenesis, this represents a library coverage of 2000-fold ($160{,}000 \div 80$) higher permutations. The relative merits of the different mutagenesis methods is shown in **Table 5-1**.

The strategy of MSSM was considered feasible given that a viable selection method for searching large trypsin libraries has been previously demonstrated (Evnin *et al.*, 1990). A modified selection method based on this was developed here. Note that combining mutations from BP1 and BP2 together would create a "mega-library" of mutants (2.56×10^{10}) for which

	Rational design	Random mutagenesis	Saturation mutagenesis	MSSM
Library size	Few	< 4,000	80	160,000
High-throughput screening	Yes	Yes	Yes	No
Structural information	Comprehensive	Minimal	Moderate	Moderate
Sequence space coverage	Few residues	Entire gene	Few residues	Few residues
Residue accessibility	20	5.7	20	20
Probability of success	Moderate	Moderate	High	Highest

Table 5-1. Relative merits of using different mutagenesis methods for making specificity-enhancements to bovine trypsin.

screening or selection would be impractical over the duration of the project. For improving resistance to self-proteolysis only a single variant was required with mutations at locations K60, R117 and K145. This triple mutant is proposed in the last chapter via rational means.

The library size for improving specificity is very large in comparison to that for improving resistance to autolysis (160,000 versus a single mutant). The difference can be explained by the fact that sites BP1 and BP2 require mutations to all of the possible 20 residues at each position while only one triple mutant is sufficient to improve stability by mutating the three alleged problematic residues away from arginine or lysine. If the three positions were changed to all of the 20 possible residues, the combined library size would be $20^3 = 8{,}000$. Whilst such a library could have been screened it was not considered necessary. The activity of trypsin on residues other than arginine and lysine has been studied at length (Harris *et al.*, 2000) with the

finding that it did not possess any specificity towards other residues. Based on this, one suitably designed triple-mutant containing mutations at the sites susceptible to self-proteolysis was deemed sufficient to test for improved stability.

The new residue at the three positions was chosen on the basis that it would cause the least disruption to the molecule's overall structure and would not be at high risk of tryptic digestion. Asparagine was chosen for this purpose since it has been used successfully for a similar purpose elsewhere (Varallyay *et al.*, 1998). Due to the residue's superior hydrogen-bonding properties it is commonly found at the beginning and end of α-helices or at motifs in β-sheets (Wan and Milner-White, 1999) where it can readily bond with the peptide backbone. Since the residues proposed for mutation did not feature in any secondary structural elements, changing them to asparagine was expected to have a fairly innocuous effect on the overall structure of the molecule. In turn, this would result in a greater chance of catalytic activity remaining unaffected.

In a bid to improve the specificity of an enzyme, there are many more factors at play than there are to improve stability. The process of substrate binding involves many subtleties not seen with the relatively straight-forward task of removing a self-proteolysis route. Accordingly, a comprehensive library has been prescribed for specificity enhancements whilst only a single mutant was proposed for the rational design of trypsin stability.

5.2 Materials and methods

5.2.1 Site-directed mutagenesis

5.2.1.1 Primer design for SDM

Forward and reverse mutagenic primers were designed, in accordance with the guidelines described in Section 3.2.3.1, to encode the following substitutions: K60N, R117N and K145N. Since the sections of gene immediately surrounding the mutation sites were GC-rich, it was not possible to keep the melting temperatures below 65 °C. The melting temperatures of the primers (calculated in AnnHyb) were between 66.4 and 72.2 °C. **Table 5-2** shows the mutagenic primers designed.

5.2.1.2 PCR reactions for SDM

A variation of the QuikChange® SDM protocol (Stratagene Ltd., 2003) was followed. All reaction components were ordered from Stratagene Ltd except dNTP mix (Roche Diagnostics Ltd.). A 50 µL PCR reaction was made up in a sterile 0.5 mL centrifuge tube and incubated on ice. The reagents were added as shown in **Table 3-4**. Supplementation of 1 µL dimethyl sulfoxide (DMSO) was a variation of the standard protocol. The centrifuge tube was transferred to a TechGene thermal cycler (Techne Ltd.) and submitted to a program of temperature cycling (**Table 5-4**).

Mutagenic primer	DNA sequence (5' → 3')		
K60N forward	CACTGCTAC|AAC|TCCGGCATCCAGGTGC		
K60N reverse	GCACCTGGATGCCGGA|GTT|GTAGCAGTG		
R117N forward	CTGAACTCC|AAC|GTGGCCTCCATCTCTCTGC		
R117N reverse	GCAGAGAGATGGAGGCCAC|GTT|GGAGTTCAG		
K145N forward	GGCAACACT|AAC|AGCTCTGGCACCTCCTACCCAGACGTGC		
K145N reverse	GCACGTCTGGGTAGGAGGTGCCAGAGCT|GTT|AGTGTTGCC		
	T_m	GC content (%)	GC clamp
K60N forward	66.4	60.7	Yes
K60N reverse	66.4	60.7	Yes
R117N forward	66.6	58.0	Yes
R117N reverse	66.6	58.0	No
K145N forward	72.2	60.0	Yes
K145N reverse	72.2	60.0	Yes

Table 5-2. Mutagenic primers designed for SDM. The affected codons are shown between the vertical lines in each sequence. In all cases the asparagine codon, AAC, required the least number of base changes and was used on this basis.

Reagent	Volume (μL)	Final amount
Sterile pure water	40	-
10 × *PfuTurbo*® reaction buffer	5	-
dNTP mix	1	not given
Forward primer	0.5	250 ng
Reverse primer	0.5	250 ng
pET(T) plasmid template	1	150 ng
PfuTurbo® DNA polymerase	1	2.5 U
DMSO	1	2%
Total	50	

Table 5-3. The components of the QuikChange® SDM reaction. Forward and reverse primers are the mutagenic primers for K60N, R117N or K145N substitutions described in Section 5.2.1.1. dNTP mix consists of dATP, dCTP, dGTP and dTTP in concentrations not specified by Stratagene Ltd.

Phase	Cycles	Step(s)	Temperature (°C)	Duration (min)
1	1	Initial denaturation	95	1.0
2	24	Denaturation	95	1.0
		Annealing	55	2.0
		Extension	72	24.0
3	1	Final extension	72	10.0
4	1	Final hold	4	∞

Table 5-4. Program of temperature cycling used for QuikChange® SDM (modified from Stratagene Ltd., 2003).

The QuikChange® reaction products were digested with 10 U *Dpn* I (Stratagene Ltd.) for 2 hours at 37 °C. An agarose gel electrophoresis (Section 2.4.8) was carried out to confirm that the reaction had generated detectable reaction products. 1 µL of the digestion was electroporated (Section 2.4.6) into *E. coli* TOP10 cells. The transformed cells were spread on LB kan$^+$ agar plates. The plasmid DNA from ten discrete 5 mL overnight cultures (Section 2.4.1) was purified (Section 2.4.4) and sequenced (Section 2.4.9).

5.2.2 Screen for resistance to autolysis

To screen for increased resistance to autolysis, pET(T) and the newly designated pET(T3) plasmid were transformed into BL21-Gold(DE3) cells (Section 2.4.5). Ten discrete colonies were picked for cultures in 96 deep-square-well plates as described in Section 3.2.5. Cultures were stored at -80 °C and removed prior to screening. Lysate samples were incubated at room temperature for the following durations: 0, 5, 10, 24 and 72 hours. At the end of each incubation step, lysates were screened for activity on

L-BANA as described in Section 3.2.2 with the exception that assays were carried out over a 4 hour period.

5.2.3 Library creation by MSSM of target sites

5.2.3.1 Primer design for MSSM

Primers were designed according to the guidelines given in Section 3.2.3.1. Mutagenic primers encompassing random residues at sites BP1 and BP2 were designed (**Table 5-5**). Both regions were designed to contain NNS codons where N = A, G, T or C and S = G or C.

5.2.3.2 PCR reactions for MSSM

PCR reactions for MSSM were set up as described in Section 3.2.3.2. The components of the reaction for MSSM are shown in **Table 5-6**. The program of temperature cycling is shown in **Table 5-7**.

5.2.4 Electrocompetent cell preparation

Commercial arginine auxotrophic cells were not available at the time of this study and were therefore prepared in the laboratory. A strain of *E. coli* requiring supplementation of arginine and thiamine in the growth medium (ATCC® number: 23790) was supplied by LGC Promochem as a freeze-dried culture. The culture was rehydrated according to the supplier's instructions and glycerol stocked (Section 2.4.3) in 1 mL centrifuge tubes.

Primer name	DNA Sequence
BP1 forward	GCTACCTGGAGGGCGGC|NNSNNSNNSTGTNNS|GGTGATTCTGGTGGC
BP1 reverse	GCCACCAGAATCACC|SNNACASNNSNNSNN|GCCGCCCTCCAGGTAGC
BP2 forward	GCTCCAAGGCATCGTCTCC|NNSNNSNNSNNS|TGTGCCCAGAAGAACAAGCCTGGC
BP2 reverse	GCCAGGCTTGTTCTTCTGGGCACA|SNNSNNSNNSNN|GGAGACGATGCCTTGGAGC

Table 5-5. Primers for MSSM. Regions to be mutated are shown between vertical lines. N = G, C, T or A and S = G or C. BP1 contains the wild-type code for serine at position 190 (TGT) which was not intended for mutation. Reverse primers were reverse complements of the forward primers.

Reagent	Volume (µL)	Final amount/ concentration
Sterile pure water	40	-
10 × *Taq* reaction buffer	5	-
dNTP mix	1	not given
BP1/BP2 forward primer	0.5	250 ng
BP1/BP2 reverse primer	0.5	250 ng
pET(T) plasmid template	1	150 ng
Taq DNA polymerase	1	2.5 U
DMSO	1	
Total	50	

Table 5-6. The components of a PCR reaction for MSSM. Reactions used either BP1 forward and reverse primers or BP2 forward and reverse primers.

Phase	Cycles	Step(s)	Temperature (°C)	Duration (min)
1	1	Initial denaturation	95	0.5
2	24	Denaturation	95	0.5
		Annealing	50/52	1.0
		Extension	72	12.0
3	1	Final extension	72	10.0
4	1	Final hold	4	∞

Table 5-7. Program of temperature cycling for a PCR reaction for MSSM.

An M9 strep$^+$ arg$^+$ agar plate (Section 2.3.3) was streaked with a sterile inoculating loop and incubated for 48 hours at 37 °C. A negative control plate minus arginine supplement was set up in parallel. 5 mL of sterile M9 strep$^+$ arg$^+$ medium (Section 2.3.2) was prepared in a 50 mL Falcon tube, supplemented with 500 µM arginine, inoculated with a single colony from the arg$^+$ agar plate, and incubated for 27 hours at 37 °C. The culture was diluted in 500 mL of sterile M9 strep$^+$ broth in a 2 L conical flask. Cells were grown at 37 °C with 200 rpm shaking to an OD_{600} of 0.5 - 0.6 (typically 3 - 4 hours). The flask was transferred to an ice bath for 10 to 15 minutes before aliquoting the culture volume into 10 pre-chilled 50 mL centrifuge tubes. The tubes were centrifuged at 4000 rpm and 4 °C for 10 minutes using a Beckman GSA rotor. The supernatant was poured off and the cells in each tube were resuspended in 25 mL of pre-chilled sterile 10% glycerol solution. The tubes were centrifuged as described earlier. Supernatant was poured off and the cells were resuspended in the remaining solution in the tubes. Cells were aliquoted to pre-chilled 1.5 mL centrifuge tubes (50 µL per tube) and stored at -80 °C.

5.2.5 Nutritional selection for improved mutants

DNA from BP1 and BP2 libraries was combined and electroporated into electrocompetent arginine auxotrophic cells with a few notable deviations from the standard protocol described in Section 2.4.6. The incubation recovery period following the pulse was extended to four hours. At the half-

way point of the recovery stage, the cells were infected with bacteriophage CE6 (EMD Biosciences Inc.), which was added to a final concentration of 2×10^9 pfu. mL^{-1}. Magnesium sulphate was also added to a final concentration of 10 mM. A negative control was set up in parallel as above, with the transfection step omitted. Two Nunc™ bio-assay dishes (large square agar plates) (Fisher Scientific) were prepared containing M9 kan^+ agar medium (Section 2.3.2) supplemented with 500 µM arg-βNA. The transfected library mixture was spread onto one of the plates and the non-transfected mixture onto the other.

5.3 Results and Discussion

5.3.1 Rational mutations for stability improvements

5.3.1.1 Site-directed mutagenesis

In order to introduce the three mutations K60N, R117N and K145N into the trypsin gene, three QuikChange® SDM reactions (Section 5.2.1.2) were carried out in series. K145N was the first reaction carried out. Agarose gel electrophoresis (Section 2.4.8) confirmed that the PCR amplification had worked (**Figure 5-1**) but not whether the designated mutation was introduced. In order to verify the presence of the mutation the following three steps were carried out: (1) the candidate DNA was electroporated (Section 2.4.6) into TOP10 electrocompetent cells; (2) three discrete colonies were picked for inoculation of 5 mL overnight cultures (Section 2.4.1); (3) the

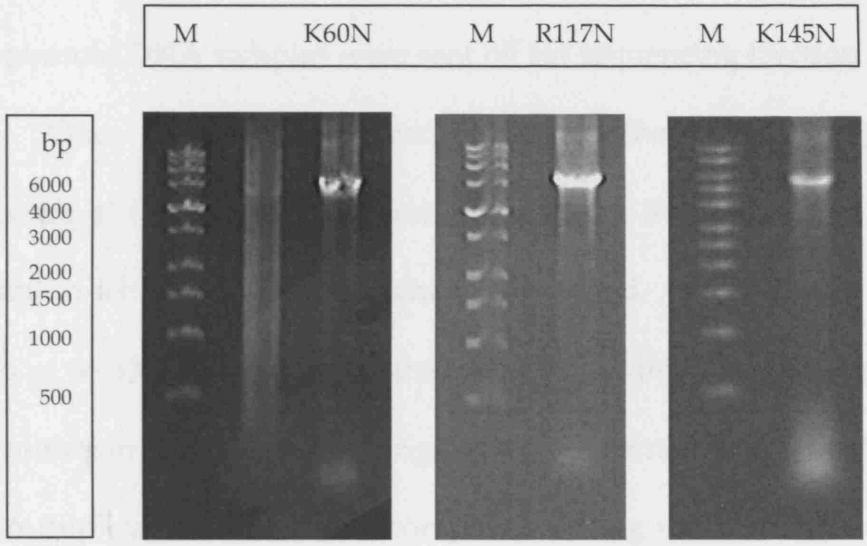

Figure 5-1. Agarose gel electrophoresis of QuikChange® SDM reactions. Clear bands can be seen at the 6000 bp mark in all three reactions which is consistent with the size of the pET(T3) plasmid. Below the 500 bp mark excess primer can be seen which was not used up in the reaction.

Sequence name	DNA Sequence
Wild-type trypsin	.....GGCTGGGGCAACACTAAGAGCTCTGGCACCTCCTACC.....
K145N 1	.....GGCTGGGGCAACACTAAGAGCTCTGGCACCTCCTACC.....
K145N 2	.....GGCTGGGGCAACACTAACAGCTCTGGCACCTCCTACC.....
K145N 3	.....GGCTGGGGCAACACTAACAGCTCTGGCACCTCCTACC.....
Wild-type trypsin	.....GCATCCCTGAACTCCCGCGTGGCCTCCATCTCTCTGCC.....
R117N 1	.....GCATCCCTGAACTCCAACGTGGCCTCCATCTCTCTGCC.....
R117N 2	.....GCATCCCTGAACTCCCGCGTGGCCTCCATCTCTCTGCC.....
R117N 3	.....GCATCCCTGAACTCCAACGTGGCCTCCATCTCTCTGCC.....
Wild-type trypsin	.....TCATCTCTGGCTGGGGCAACACTAAGAGCTCTGGCAC.....
K60N 1	.....TCATCTCTGGCTGGGGCAACACTAACAGCTCTGGCAC.....
K60N 2	.....TCATCTCTGGCTGGGGCAACACTAACAGCTCTGGCAC.....
K60N 3	.....TCATCTCTGGCTGGGGCAACACTAACAGCTCTGGCAC.....

Table 5-8. DNA sequences from QuikChange® SDM reaction. Candidate DNA from three discrete colonies was sequenced. Base substitutions are shaded in grey. In the reactions for K145N and R117N substitutions, two out of three sequences contained the desired mutation. In the reaction for K60N, all three sequences contained the desired mutation.

plasmid DNA was purified from the cultures (Section 2.4.4) and finally (4) the plasmid DNA samples were sent off for sequencing (Section 2.4.9).

Once the sequences were returned, they were analysed for the presence of the K145N mutation. Two out of the three sequences had the mutation whilst the other remained unmutated. This represented a success ratio of 66.6%. The sequence that remained wild-type was possibly due to the mutation reverting to the original base at some stage in further rounds of DNA duplication or due to incomplete priming in the PCR reaction leaving out the mutation.

One of the two successful K145N mutants was taken forward as a template in the next QuikChange® SDM reaction to introduce the R117N mutation. The same procedure was followed to verify the mutation as for K145N. The returned sequences showed again that two out of three had the mutation. One of the two successful R117N mutants was taken forward as a template for the next QuikChange® SDM reaction to introduce the last K60N mutation (after confirming that the K145N mutation was still present). This time all three returned sequences had acquired the mutation.

In all reactions DMSO was supplemented at a final concentration of 2%. This organic solvent helps to reduce the formation of secondary structures in either the template or primer and improves the chances of a successful reaction occurring (Pomp and Medrano, 1991).

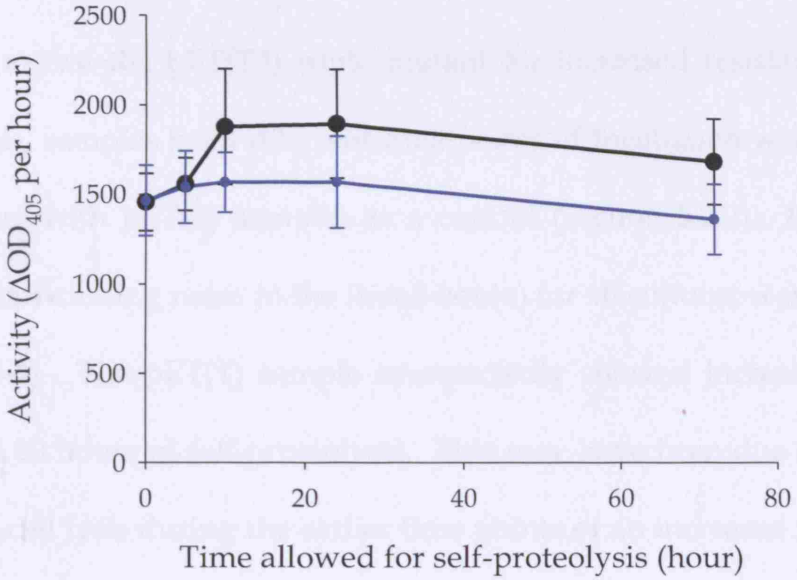

Figure 5-2. Stability screen for the effect of rational mutations on self-proteolysis. pET(T) = black; pET(T3) = blue. Assays were carried out after incubating lysate samples at RT for 0, 5, 10, 24 and 72 hours. Error bars correspond to ± one standard deviation about the mean.

5.3.1.2 Screen for resistance to autolysis

In order to test the pET(T3) triple mutant for increased resistance to self-proteolysis, samples from different time points of incubation were screened for 4 hours with pET(T) samples as a control (Section 5.2.2). Increases in OD_{405} (not including noise in the first 2 hours) for 10 cultures were averaged (**Figure 5-2**). The pET(T) sample unexpectedly showed increased activity from 5 to 10 hours of self-proteolysis. This may have been due to a lack of complete cell lysis during the earlier time points or an increased rate of self-proteolysis. From this point until the last time point of 72 hours there appears to have been only a modest deterioration in activity for both pET(T) and pET(T3) even accounting for error. This would indicate that the effects of self-proteolysis were not as pronounced as first thought even though Ca^{2+} was supplemented. The most likely explanation for this (apart from the stabilising effect of Ca^{2+}) is that the concentration of trypsin accumulating in the cell was not high enough for self-proteolysis to be significant. This is supported by the fact that the cultures were not IPTG induced and were therefore reliant on the low expression due to the leaky T7 promoter.

At this point, screening pET(T) and pET(T3) samples that had been IPTG induced might have been proposed. However, this was not deemed necessary since it was previously shown that trypsin activity (and therefore expression) was not dependent upon ITPG induction (**Figure 3-2**). Furthermore, inducing cultures for screening purposes would compound the problem of variable enzyme concentrations from well to well. This would

defeat the overall aim of improving the screen. It was decided therefore not to test IPTG induction for screening applications and to continue using the uninduced background protein stabilised with Ca^{2+} (see Section 3.3.1).

The pET(T3) plasmid was, nevertheless, selected to act as the template for the MSSM reactions over pET(T) on the basis of three reasons. Firstly, the vector proved to give as good a signal as pET(T) for the expression of active trypsin. Secondly and more crucially, pET(T3) activity was more consistent over the range from 0 - 10 hours, which is the period of interest for carrying out microplate assays. In contrast, pET(T) activity increased rapidly over the same time period indicating that certain factors at play including autolysis, misfolding and refolding of the enzyme, favour an increasing rate of activity up to the 10 hour period.

Finally, the three mutations present in pET(T3), may permit any future beneficial mutations to be displayed in an activity screen that would not show up otherwise. This may apply to circumstances in which beneficial mutations have enhanced either arginine or lysine activity. If present in the original pET(T) template, such an enhancement might not be identified in a screen since the newly expressed enzyme would supposedly self-proteolyse at a faster rate. Using pET(T3) as a template may avoid this pitfall potentially. An application of this strategy has been demonstrated for the evolutionary engineering of a β-Lactamase (Peimbert and Segovia, 2003) where SDM was carried out prior to random mutagenesis to create a biased

combinatorial library. Mutants were isolated with 10-fold higher cefotaxime resistance than the wild-type enzyme.

5.3.2 MSSM of specificity target sites

5.3.2.1 Primer design for MSSM

The residues at regions BP1 and BP2 were to be replaced by NNS codons (N = G, C, A or T; S = G or C). The NNS codon has the benefit of encoding all twenty naturally occurring amino acids with a reduced number of stop codons. The coding for UAA and UGA stop codons are both eliminated leaving only the chance of one stop codon, UAG, occurring. Stop codons signal the termination of translation and are not required at this point in the gene. The mutagenic primers used are shown in **Table 5-5**.

5.3.2.2 PCR reactions for MSSM

Separate MSSM reactions (Section 5.2.3.2) were carried out in order to introduce the mutagenic inserts for regions BP1 and BP2 into the pET(T) plasmid. The final reaction conditions used are given in **Table 5-6** (components) and **Table 5-7** (program of temperature cycling). Agarose gel electrophoresis (Section 2.4.8) confirmed the presence of DNA product consistent with the 6000 bp mark (**Figure 5-3**) albeit with some non-specific products of varying size. Annealing temperatures were tested from 50 - 55 °C. However, significant reaction products were found only at annealing temperatures of 50 and 52 °C for both BP1 and BP2.

Taq DNA polymerase does not have the "proofreading" ability required for the high fidelity replication of DNA. This quality gives an enzyme 3' to 5' exonuclease error-checking activity, meaning that it can work its way along the DNA from the 5' end to the 3' end and correct nucleotide misincorporations. Such a quality was not required for MSSM and so *Taq* DNA polymerase was favoured over the proofreading *PfuTurbo®* DNA polymerase.

5.3.2.3 *Verification of MSSM mutation rate*

In order to verify the mutation rate of successful reactions, candidate library DNA was treated similarly to fepPCR DNA as described in Section 3.2.3.4: Reaction products were electroporated into TOP10 cells, overnight cultures were grown from ten colonies and finally the DNA was purified and sent off for sequencing. Two out of ten sequences returned showed mutations in site BP1 (annealing temperature: 50 °C) and three out of ten in site BP2 (annealing temperature: 52 °C) as shown in **Table 5-9**. Such a poor uptake ratio of the mutagenic insert may be due to the relatively ambitious number of residues that were mutated collectively. Saturation mutagenesis is traditionally used to randomise a single residue position in a gene. MSSM of four residues requires the binding of a primer containing four times the number of bases required for single site saturation mutagenesis any of which may be non-complementary. Examples of libraries this vast have rarely been

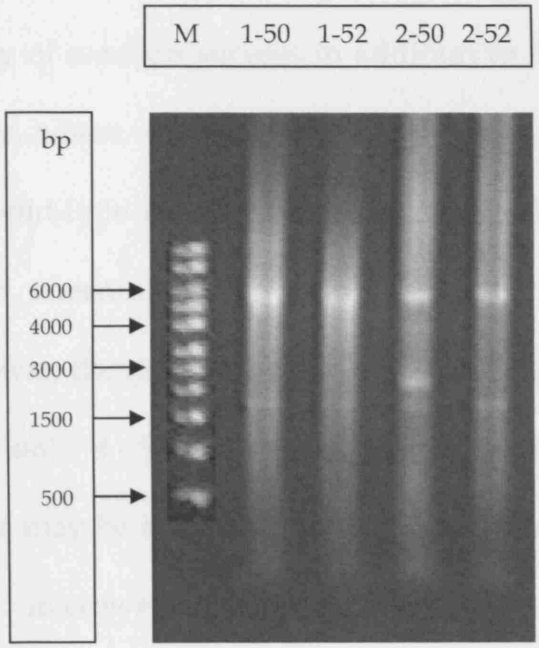

Figure 5-3. Agarose gel electrophoresis of MSSM sites. M = marker (alternate band sizes shown), 1-50 = binding pocket 1 with an annealing temperature of 50 °C, etc. In all lanes a band is present consistent with the 6000 bp mark. Some background intensity can be seen which may attributable to the formation of incomplete or erroneous DNA strands in the reaction.

Sequence name	DNA sequence
wild-type trypsin	... ... AAG GAT TCC TGT CAG ... ...
BP1-50 2	... ... GGG TGG AAG TGT CCG ... ...
BP1-50 9	... ... GCG GTG GCC TGT TCG ... ...
wild-type trypsin	... ... TGG GGT TCC GGC ... ...
BP2-52 4	... ... GAC ATG ACC ACC ... ...
BP2-52 7	... ... ATG TAG GGG GGG ... ...
BP2-52 9	... ... CCG GCC GGG TAG ... ...

Table 5-9. Sequences mutated by MSSM. BP1 refers to binding pocket 1 while BP2 refers to binding pocket 2. The digits following this (50 and 52) are the annealing temperatures (°C) for the MSSM reaction and the digits following this refer to the sequence number from 1 - 10. Only those sequences that contained a mutagenic insert are shown. The bases shaded in grey are those that differ from the wild-type sequence. TGT triplet in BP1 coding for cysteine was not designed to be mutated.

reported in literature (MacBeath *et al.*, 1998). This may be, in part, due to the low probability of reaction success in addition to the scale of the screening task afterwards, where no selection method exists. Given that the best ratio of library to wild-type DNA obtained in MSSM has been 20 - 30% of ten sequences, the screening task would appear to have been compounded further. However, the number of transformants required to fully encompass the library is only 4 - 5 times more than if a 100% mutation ratio was obtained. This may be addressed by using a robust clone selection method in favour of microwell screening and by optimising transformation efficiencies of library DNA.

Both sequences for site BP1 show the wild-type residue serine at position 191 to be unmutated as expected. Outside this, a possible twelve bases were susceptible to mutation of which nine were changed in the first sequence and seven in the second sequence. For site BP2, twelve consecutive bases were susceptible to mutation of which nine were changed in both the first and second sequences and ten in the third sequence. Sequences for 52 °C annealing with BP1 and 50 °C annealing with BP2 did not contain any mutations.

It appears that the optimal annealing temperature has been different for BP1 (50 °C) and BP2 (52 °C). This may be due to the fact that BP1 primers had an unmutated codon midway through the primer making it more conducive for template annealing at a lower temperature. However, the experiment would have to be repeated a number of times to confirm that this

can be a genuine conclusion. In reality, the two temperatures are close enough to explain the difference by chance. The same may be said of the fact that only 20% of BP1 sequences contained the mutagenic insert whilst it was 30% in the case of BP2.

5.3.3 Validation of nutritional selection method

In order to validate the selection method, both the electrocompetency and auxotrophy of prepared cells (Section 5.2.4) were tested in addition to the potential substrate to be used for the nutritional selection. To test for auxotrophy, arg^+ and arg^- agar plates were streaked separately with arginine-auxotrophic cells from a glycerol stock. One of the consequences of growing auxotrophic strains on minimal media is that the rate of cell doubling is considerably slower than for other commercial strains of *E. coli*. Following 48 hours incubation at 37 °C, dense cell growth was observed on the streaked regions of the arg^+ plate while no growth was observed on the arg^- plate. The arg^- plate did, however, show the development of colonies after 72 hours incubation indicating that plates should not be incubated beyond 48 hours to prevent false positives from appearing in a nutritional selection. Cell growth without the supplement was most likely the result of cells scavenging arginine from lysed dead cells in the vicinity, but may also have been due to some clones reverting to the original wild-type genotype, which can synthesize its own arginine.

Control electroporations (Section 2.4.6) were carried out to test the electrocompetency of the cells and the suitability of the substrate, arg-βNA. The pET(T3) plasmid was electroporated into arginine auxotrophic cells and spread onto two kan^+ agar plates (Section 2.3.3), one supplemented with 500 μM arginine and the other supplemented with 500 μM arg-βNA. After 48 hours of incubation at 37 °C, the plates were observed for colony formation. The plate supplemented with arginine contained a full plate of densely packed colonies indicating that the cells were suitably electrocompetent in the first instance. This equated to a transformation efficiency of 1×10^5 transformants per μg.

The plate containing only the substrate arg-βNA anomalously contained four colonies. There were a few possibilities for this occurrence: (1) the substrate had degraded during the course of the incubation and released sufficient arginine for minimal colony formation; (2) the cells had reverted to the original non-auxotrophic strain; and (3) the cells had scavenged arginine from dying cells in proximity. The relatively low occurrence of colonies, nevertheless, indicated that although the substrate arg-βNA was suitable for use as a growth selection nutrient, some false positives may be generated.

5.3.4 Nutritional selection for improved mutants

The nutritional selection method developed (Section 5.2.5) was only suitable for selecting novel exoprotease activity in bovine trypsin variants. It was not

thought possible to select for endoprotease activity since any growth selection using a peptide of the form -X-X-X-R-X-X-X- would not release free arginine. Consequently, a substrate of the form R-X had to be chosen where cleavage of the peptide bond would release free arginine for cells to use as an essential growth nutrient. Since this activity was novel, there was no available control plasmid to use that expressed an enzyme with such activity. While seemingly ambitious, some success has been had in the past using such a system for the modification of rat anionic trypsin specificity, although with a cost to overall activity (Evnin *et al.*, 1990).

Both libraries BP1 and BP2 were combined to make benefit of the vast throughput allowed by using a selection, rather than screening, method. This gave a combined total of 3.2×10^5 library members, which was comfortably within the transformation efficiency range quoted for electroporation with TOP10 cells (1×10^9 transformants per µg of DNA). Transfection with the bacteriophage CE6 was required to provide a source of T7 RNA polymerase for expression of the library DNA.

Large square agar plates were approximately ten times the size of a standard petri dish and were therefore used instead to increase the throughput per plate. Following electroporations as described in Section 5.2.5, plates were observed for colony growth after 48 hours of incubation at 37 °C. Colonies were observed on the two plates as follows: (1) library DNA with transfected CE6 phage gave 9 colonies; (2) library DNA minus transfection gave 5 colonies.

The colonies were picked from both plates and screened for activity (Section 3.2.4) on the substrate, arg-pNa, to check for novel activity. Unfortunately, the clones did not confer any activity above the base level of hydrolysis, indicating that the colonies were false positives (see Section 5.3.3). A repeat of the procedure produced similar results. The lack of identifying any positive clones may have been down to the possibility that the desired specificity may not have been an achievable goal and was certainly not achievable by mutating the selected hotspots. In addition, there was always the possibility that the successful mutations were simply not produced in the MSSM reactions despite having taken steps to maximise the chances of success. For a more thorough discussion of why this approach may not have yielded any useful mutants of bovine trypsin, see Section 7.2.

6 Characterising the self-proteolysis of trypsin

6.1 Introduction

While some gains in substrate specificity were obtained in Chapter 3, Chapter 5 has highlighted the inherent problems associated with auxotrophic selection using *E. coli* based expression systems. An alternative approach to generating variants was proposed based on the previously described ability of bovine trypsin to self-proteolyse into bovine α-trypsin yet remain active (Keil-Dlouha *et al.*, 1971a; Maroux *et al.*, 1967). Bovine α-trypsin, which contains an internal cut between Lys-131 and Ser-132, can self-proteolyse further into a "pseudotrypsin" with a cut between Lys-176 and Asn-177 (Smith and Shaw, 1969).

Since there are multiple arginine and lysine sites within a trypsin molecule, it follows that a mixture of pseudotrypsins could be generated when a sample is left to self-proteolyse over time. Such modified forms were expected to have specificities potentially different to that of native trypsin. Attempts were made to isolate the modified trypsins and characterise them with respect to specificity, molecular weight and activity on pro-insulin.

Trypsin variants generated by self-proteolysis differ only in the location and amount of internal cuts. Such a change from the native form may not affect the size, charge or hydrophobicity profile of a variant enough to form the basis for a chromatographic separation. However, when internal cuts are created by self-proteolysis, new C-termini must appear as either ...X-X-X-R or ...X-X-X-K. By virtue of its specificity, the addition of

Carboxypeptidase B (CpB) should remove any arginine or lysine residue from any new C-terminus. Internally cut trypsins would therefore be less positively charged under physiological conditions and separation on a cation exchange chromatography (CEC) column should be technically feasible.

This chapter describes the results of the following procedures: (1) over-activation of trypsinogen and trypsin as a function of time; (2) carboxypeptidase B digestion of variant mixture; (3) cation exchange chromatographic separation of variants with and without prior affinity chromatography (AC); and (4) characterisation of variants. The following techniques were used to characterise the variants: (1) colourimetric activity assays to determine kinetic properties; (2) reducing SDS-PAGE to determine molecular weights; and (3) HPLC analysis of enzymatically converted pro-insulin samples. Eli Lilly's KPB-HPI (Lys Pro chain B - human pro-insulin) was provided for the biotransformations. A summary of the strategy is shown in **Figure 6-1**.

A study of the pH optima and activity of trypsin on a range of natural substrates was carried out in an effort to define its specificity in addition to what is known from literature. Both bovine and recombinant bovine trypsin (r-Trypsin) were used for this study to measure any differences between them. From an industrial standpoint, this could lead to a better insight of what happens during the pro-insulin activation step. Currently the specificity of trypsin is defined as having proteolytic activity at the C-terminal residues of arginine or lysine residues alone unless a proline is

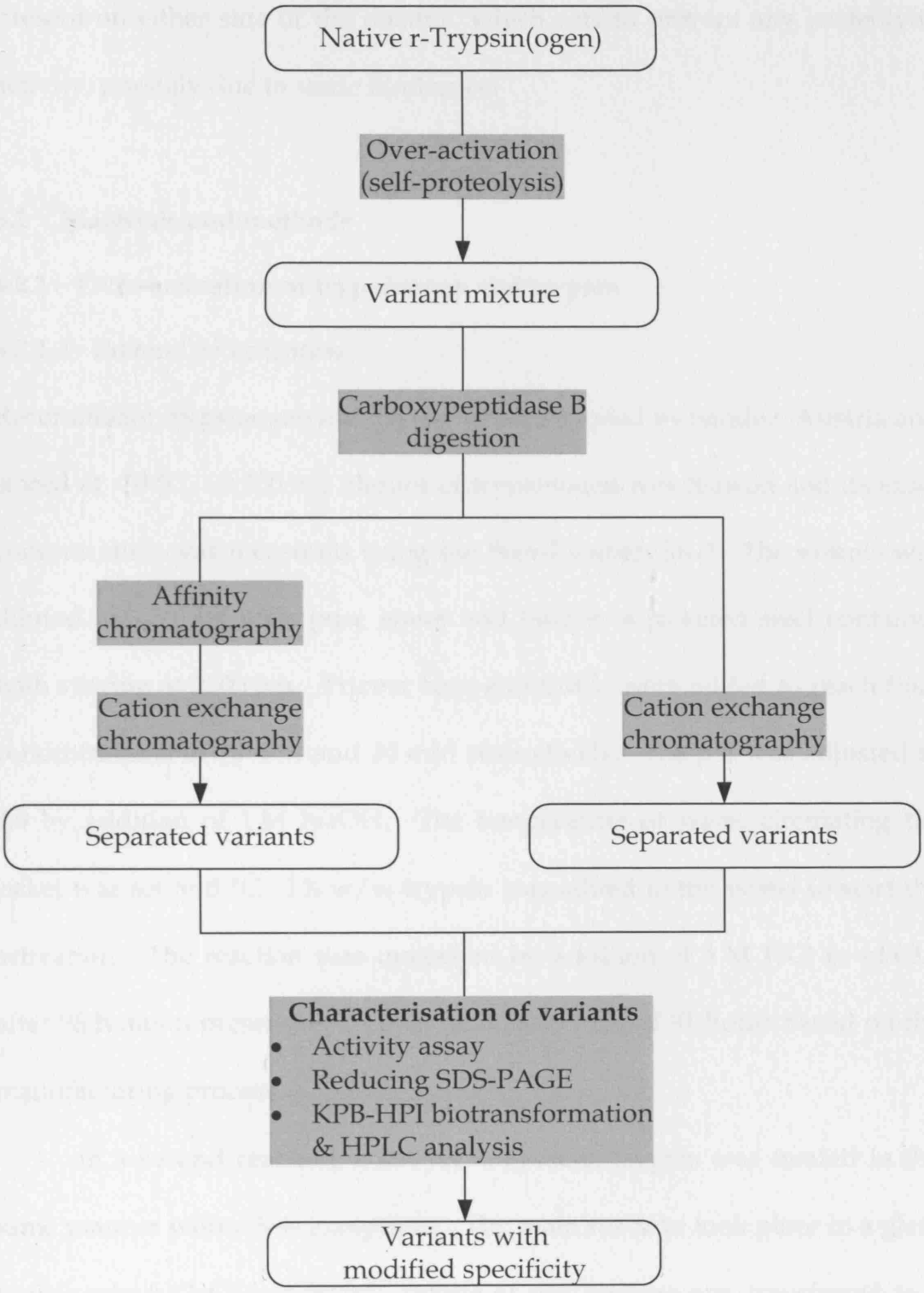

Figure 6-1. Strategy for isolating and characterising self-proteolysed variants of r-Trypsin. Two alternative methods of separation were tested; one involved direct CEC after CpB digestion whilst the other involved AC prior to CEC.

present on either side of the residue, which acts to prevent any proteolytic activity, possibly due to steric hindrance.

6.2 Materials and methods

6.2.1 Over-activation of trypsinogen and trypsin

6.2.1.1 Protocol for activation

Recombinant trypsinogen and trypsin were supplied by Sandoz, Austria and stored at -70 °C. A 250 mL aliquot of trypsinogen was thawed and its exact concentration was measured using the Beer-Lambert law[1]. The sample was diluted to 1.5 $g.L^{-1}$ with pure water and held in a jacketed steel container with stirring at 250 rpm. Trizma base and $CaCl_2$ were added to reach final concentrations of 70 mM and 30 mM respectively. The pH was adjusted to 8.0 by addition of 1 M NaOH. The temperature of water circulating the jacket was set to 8 °C. 1% w/w trypsin was added to the vessel to start the activation. The reaction was quenched by addition of 1 M HCl to pH 3.0 after 96 hours representing an over-activation time of 91 hours based on the manufacturing process.

In a second reaction, a 250 mL aliquot of trypsin was treated in the same manner with a few exceptions. The main reaction took place in a glass beaker minus $CaCl_2$. A 50 mL sample of this mixture was transferred to a new glass beaker and $CaCl_2$ was added to a final concentration of 30 mM.

[1] Beer-Lambert law: $Abs = \varepsilon c l$.

The reactions proceeded at room temperature (~23 °C) with stirring at 250 rpm.

6.2.1.2. TAME assay

At regular time intervals during the over-activations, trypsin activity was measured by spectrophotometric assay on the TAME substrate. Reaction components, their concentrations and the order in which they were added are given it in **Table 6-1**. A stock solution of 1.0 mM TAME was prepared by first dissolving it in 17.5 M glacial acetic acid. Trizma base and water were then added to achieve the desired concentrations. Two reaction components were added to a 3 mL cuvette. The reaction began upon addition of the substrate solution (2nd component) upon which the cuvette was placed in the spectrophotometer. The OD_{247} was measured every 20 seconds for 180 seconds via a kinetic programme on the spectrophotometer. The first reading was discarded when calculating ΔOD_{247} min^{-1}.

6.2.2 CpB digestion of variant mixture

CpB was added to the variant mixture to remove any C-terminal R or K residues generated by self-proteolysis. 5% w/w CpB was added to 200 mL of the variant mixture minus $CaCl_2$ and left for 2 hours at RT with stirring at 250 rpm. A 1 mL control sample was isolated from the main sample before the addition of CpB.

6.2.3 Separation of variants

6.2.3.1 Cation exchange chromatography

Both chromatography systems relied on strong cation exchangers (SO_3^-) for the basis of separation, at two different scales: (1) SP Sepharose™ Fast Flow (50 - 200 mL) and (2) pre-packed mono S 5/50 GL column (0.5 - 3 mL). Both of these and the ÄKTA Explorer protein purification equipment were provided by GE Healthcare, USA.

An empty column 1.6 cm in diameter was packed with SP Sepharose™ Fast Flow (SFF) medium to a height of 32 cm giving a column volume of 64.3 cm^3. The manually packed and pre-packed mono S columns were attached to an ÄKTA explorer via independent flow paths. Equilibration buffer consisted of 50 mM acetic acid and 25 mM NaCl; elution buffer consisted of 100 mM acetic acid and 1500 mM NaCl. Both buffers were approximately pH 3.5. The ÄTKA Explorer was run in manual mode in accordance with the manufacturer's standard protocol. Fractions were collected from start to finish of each peak and stored at 4 °C. The control sample was run in the mono S column and the CpB digested sample in the manually packed column.

6.2.3.2 Affinity chromatography

Benzamidine Sepharose™ 6B resin was provided by GE Healthcare. A 2.5 cm diameter column was packed to a height of 5 cm to give a column volume of 24.53 cm^3 and was then attached to an ÄKTA explorer.

Order of addition	Ingredient(s)	Concentration (mM)	Volume in cuvette (µL)	Final concentration (mM)
1st	Trypsin	5.25×10^{-4}	200	3.5×10^{-5}
2nd	Glacial acetic acid Trizma base TAME	175 160 1.0	3000	164 150 0.94
Total			3200	

Table 6-1. Components of the spectrophotometric TAME assay. The substrate solution (2nd component) was adjusted to pH 8.0 by addition of 1 M NaOH. The component named trypsin denotes mature trypsin with full activity as opposed to samples taken during trypsinogen activation; these samples were left at concentrations high enough for a clear signal to be given on the spectrophotometer (*i.e.* close to that of mature trypsin).

Order of addition	Ingredients	Concentration (mM)	Volume in cuvette (µL)	Final concentration (mM)
1st	Trypsin Acetic acid	5.25×10^{-4} 100	200	3.5×10^{-5} 6.67
2nd	Pure water Trizma base Calcium chloride	- 1000 184	See Table 6-3	- 0 - 466.67 12.25
3rd	Glacial acetic acid Trizma base pNa substrate	175 160 1.2	1200	70 64 0.48
Total			3000	

Table 6-2. Components of the colourimetric pNa assay. The 2nd component had variable amounts of Trizma base and water to achieve final pHs of 6.0, 7.0, 8.0 or 9.0 by titration ± 0.1 pH unit. These ingredients were pre-mixed with $CaCl_2$ to make up the various "pH adjuster" solutions with each being added as a single component of 1600 µL (**Table 6-3**).

Ingredient	pH adjuster solution			
	6.0	7.0	8.0	9.0
Pure water	1370	1345	1200	0
Trizma base (1000 mM)	30	55	400	1400
Calcium chloride (184 mM)	200	200	200	200
Total	1600	1600	1600	1600

Table 6-3. Make-up of pH adjuster solutions (2nd assay component).

Step	Buffer(s)	Flow rate ($ml.min^{-1}$)	Time (min)	No. of column volumes
1. Equilibration	Equilibration	6	20	2
2. Sample load	Crude trypsin	6	variable	variable
3. Post-load wash	Equilibration	6	10	1
4. Gradient elution	Equilibration & Elution	6	100	10
5. Re-equilibration	Equilibration	6	20	2

Table 6-4. CEC procedure for packed SFF column. Column volume = 64.3 cm^3. Buffers were degassed and lines were purged with the buffer that was to run through them before first use of the column. Crude trypsin denotes quenched over-activated trypsin(ogen) samples at a pH of between 3.0 - 3.5; typically between 25 - 200 mL was loaded per run. During gradient elution, 100% equilibration buffer was gradually displaced with 100% elution buffer at a constant linear rate over 100 minutes.

Step	Buffer(s)	Flow rate ($ml.min^{-1}$)	Time (min)	No. of column volumes
1. Equilibration	Equilibration	4	12	2
2. Sample load	Crude trypsin	4	variable	variable
3. Post-load wash	Equilibration	4	6	1
4. Step elution	Elution	4	12	2
5. Re-equilibration	Equilibration	4	12	2

Table 6-5. Affinity chromatography procedure. Column volume = 24.53 cm^3. Before first use of the column, the same pre-treatment was employed as for CEC. Typically between 25 - 150 mL of crude trypsin, adjusted to pH 5.8, was loaded per run.

Equilibration buffer consisted of 100 mM acetic acid and 100 mM trizma base adjusted to pH 5.8. Elution buffer consisted of 100 mM acetic acid adjusted to pH 3.0. The ÄTKA explorer was run in manual mode according to the procedure given in **Table 6-5**. Fractions were collected from start to finish of each peak and stored at 4 °C.

6.2.4 Characterisation of variants

6.2.4.1 Colourimetric assay

Isolated variants were tested for activity on the Z-RR-pNa substrate (Bachem, Switzerland). Reaction components, their concentrations and the order in which they were added are given in **Table 6-2**. A stock solution of 0.48 mM substrate was prepared by first dissolving it in 17.5 M glacial acetic acid. Trizma base and water were then added to achieve the desired concentrations. Three reaction components were added to a 3 mL cuvette. Assays were carried out at pH 8.0.

The reaction began upon addition of the substrate solution (3rd component) upon which the cuvette was placed in the spectrophotometer. The OD_{405} was measured every 20 seconds for 300 seconds via a kinetic programme on the spectrophotometer.

6.2.4.2 *Molecular weight determination of variants*

6.2.4.2.1 Reducing SDS-PAGE

Each fraction from CEC was concentrated in a 2.5 kDa microcon filter to > 0.1 $\mu g.mL^{-1}$ (Millipore, USA). The following NuPAGE® brand materials were ordered from Invitrogen, USA: Novex 10% bis-tris mini gels, LDS sample buffer, MES SDS running buffer, sample reducing agent, antioxidant and XCell *SureLock*™ mini-cell apparatus. Mark12™ unstained marker was also ordered from Invitrogen, USA. Reducing SDS-PAGE was carried out in accordance with the manufacturer's protocol.

6.2.4.2.2 Silver staining

Protein bands on reduced SDS-PAGE gels were visualised by silver staining. SilverXpress® silver staining kit was ordered from Invitrogen, USA. Staining was carried out in accordance with the manufacturer's protocol.

6.2.4.3 *KPB-HPI biotransformation*

This reaction converts the inactive precursor KPB-HPI into mature KPB-BHI. KPB-HPI was taken from Eli Lilly's insulin manufacturing process and stored at -20 °C. An aliquot was thawed and diluted to a concentration of 18 $g.L^{-1}$; 100 mL of sample was transferred to a glass beaker and held at 12 °C for the duration of the reaction with stirring at 200 rpm. A sample was collected for HPLC analysis by following these steps: (1) 0.5 mL of KPB-HPI mix was added to 0.5 mL of 1 M acetic acid and (2) this mixture was diluted

20-fold with dilution buffer (0.036 M sodium acetate trihydrate, 0.166 M glacial acetic acid). Prepared HPLC samples were stored at 4 °C.

The pH of KPB-HPI was adjusted to between 8.4 and 8.6. 1 M $CaCl_2$ solution was added to reach a final concentration of 5 mM and the pH lowered to between 7.3 and 7.4. A sample was collected for HPLC analysis. Recombinant CpB (Eli Lilly, USA) was added at a ratio of 1 $mg.g^{-1}$ KPB-HPI and left to stir for 20 minutes. rTrypsin was added at a ratio of 0.55 $mg.g^{-1}$ KPB-HPI; trypsin diluent (composition not revealed by Eli Lilly) was added at a rate of 1 $mL.mg^{-1}$ rTypsin. A sample was collected for HPLC analysis every 30 minutes after reaction initiation for 2.5 hours. The pH was maintained at 7.3 - 7.4 throughout.

6.2.4.4 HPLC analysis of enzymatically converted KPB-HPI

Quenched samples from a biotransformation reaction were diluted 1:20 in 0.2 M sodium acetate buffer (pH 4.0). The concentration of KPB-BHI in quenched samples was determined by HPLC. The HPLC system consisted of a refrigerated autosampler, UV detector, column oven and dual pumps (Dionex Corp., USA). Separation of the components was achieved using a Dupont Zorbax 300SB-C8 column.

A stock buffer was prepared from which both mobile phases were produced. The stock buffer consisted of 130 mM OSA (1-octanesulfonic acid, sodium salt), 5.6 mM sodium sulphate, 17.4 mM DEA (diethylamine) and 138 mM phosphoric acid (85% concentrated). Mobile phase A consisted of

25% stock buffer, 45% pure water and 30% acetonitrile. Mobile phase B consisted of 25% stock buffer, 25% pure water and 50% acetonitrile. A mobile phase gradient was programmed as shown in **Table 6-6**.

Time (minutes)	Mobile phase B (%)
0	41
10	56
18	67
18.1	41
25	41

Table 6-6. Mobile phase gradient at different time intervals.

Samples were placed in the autosampler and maintained at 5 °C. The temperature of the column was maintained at 26 °C by the oven. Flow rate was set to 0.8 mL per minute and the injection volume to 5 µL. The detector monitored the absorbance of the output stream from the column at 214 nm wavelength. The retention time of standard grade KPB-BHI was between 550 and 650 seconds.

6.2.5 Trypsin activity on a range of chromogenic substrates

The assay protocol is described in Section 6.2.4.1. Chromogenic substrates Z-R-pNa, Bz-R-pNa, R-pNa, K-pNa, Z-RR-pNa, Z-KR-pNa and Z-GR-pNa were ordered from Bachem, Switzerland. All assays were carried out at pH 8.0.

6.2.6 pH optima of selected trypsin substrates

The assay protocol is described in Section 6.2.4.1. Assays were carried out at pH 6.0, 7.0, 8.0, 9.0 and 10.0. Note that trizma base has a buffering capacity of up to pH 9.0; in order to achieve assay conditions of pH 10.0 ± 0.1, 20 µL of 1 M NaOH was added to the assay cuvette.

6.3 Results and discussion

6.3.1 Over-activation of trypsinogen and trypsin

Trypsinogen and trypsin samples were over-activated in order to generate a mixture of variants with internal cuts (Section 6.2.1.1). Activity of the variant mixtures was measured at regular time intervals by carrying out the TAME assay (Section 6.2.1.2). The activity levels for both trypsinogen and trypsin are shown in **Figure 6-2**.

Trypsinogen is an inactive zymogen converted to trypsin upon cleavage of the pro-site by the highly specific protease enteropeptidase (Maroux *et al.*, 1971). Given the natural broad specificity of trypsin, however, trypsinogen may also be activated by mature trypsin by cleavage at the pro-site. By seeding the trypsinogen sample with 1% trypsin (w/w), the activation was effectively initiated. The highest level of TAME activity was reached at the 9 hour time point indicating that the highest concentration of active trypsin would have been present here. By failing to quench the reaction at this point, it was thought there would be a steady drop in activity as self-proteolysis of the mature trypsin set in. It was proposed a drop in

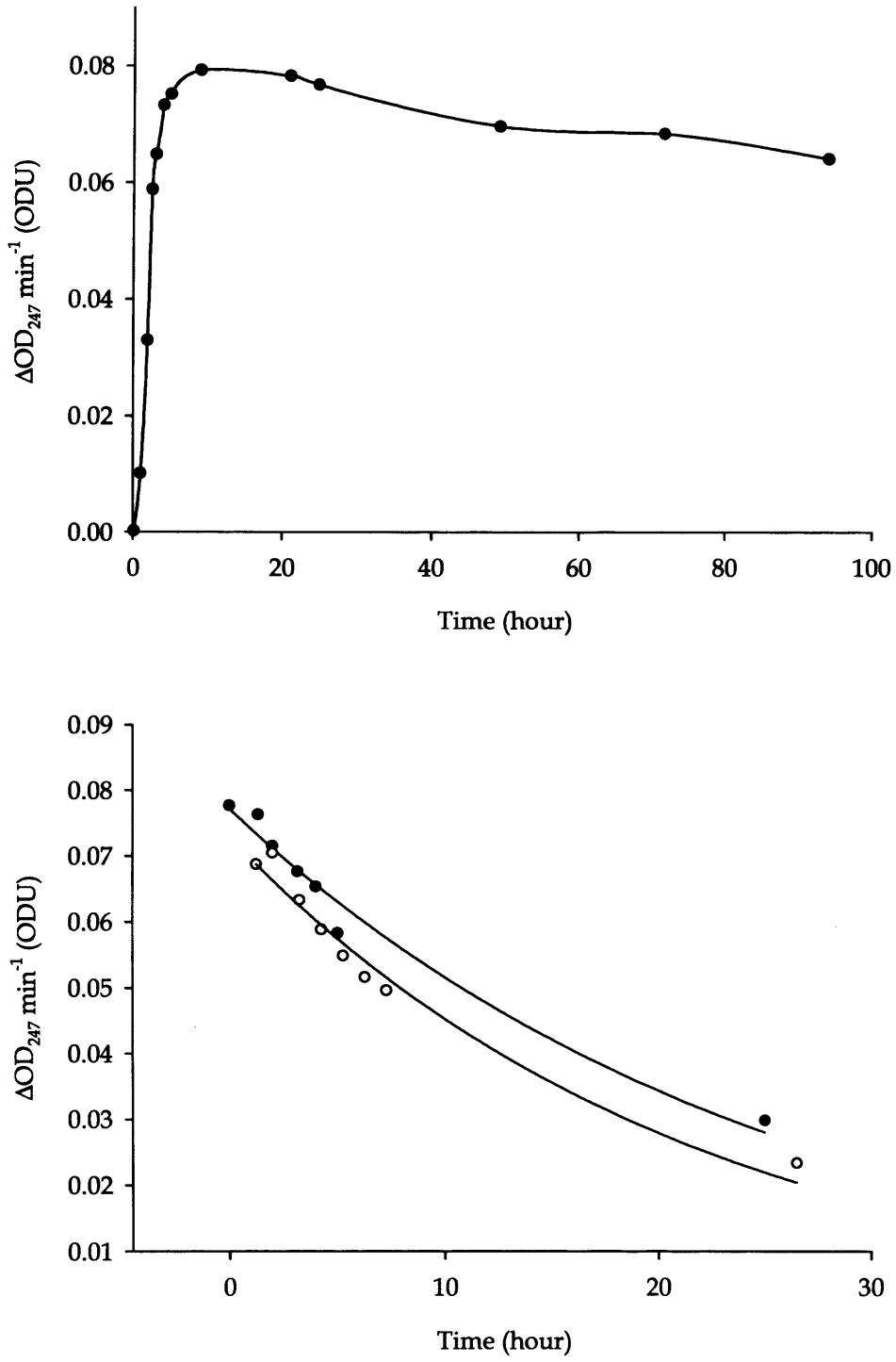

Figure 6-2. Over-activation of trypsinogen and trypsin. Top: Trypsinogen at 8 °C with $CaCl_2$ added to a final concentration of 30 mM. Bottom: Trypsin at RT with $CaCl_2$ added to a final concentration of 30 mM (•) and without $CaCl_2$ addition (◦). Activity levels at each time point were measured by carrying out the TAME assay (Section 6.2.1.2).

activity could represent both a complete inactivation of trypsins which have been fatally cut and, more importantly, changes to the specificity of trypsins which have been non-fatally cut.

The plot of trypsinogen activity after 9 hours shows an extremely slow and steady rate of activity reduction (**Figure 6-2** - top). After 94.5 hours the level of activity relative to the value at 9 hours was 80.8%. It can be concluded from this that the over-activation of trypsinogen was not conducive to a fast rate of self-proteolysis. It was decided that Eli Lilly's trypsin should be alternatively used as the starting material in an attempt to increase both the rate of self-proteolysis and the probability of yielding a positive variant. There was also evidence to suggest that in-house trypsin (purified by affinity chromatography following the activation of trypsinogen) contained a significant population of molecules digested at different locations (unpublished observations by Marc Ebtinger, Eli Lilly, USA). Using this instead of trypsinogen would supposedly lead to a better and faster rate of variant generation. Carrying out the over-activation at RT instead of 8 °C would also help to speed up the process. In addition to the changes mentioned, the effect of adding and omitting $CaCl_2$ to the trypsin over-activation was tested. Ca^{2+} ions have been shown to stabilise trypsin by slowing the rate of self-proteolysis (Sipos and Merkel, 1970).

Both plots for trypsin over-activation (± $CaCl_2$) show a similar profile for the drop in activity (**Figure 6-2** - bottom), which is much faster and greater when compared with trypsinogen. The activity levels for $CaCl_2$

addition and omission dropped to 38.4% (after 25 hours) and 34.0% (after 26.5 hours) respectively. The over-activations were quenched at these time points since the reductions in activity were deemed significant enough to have indicated that a large concentration of variants present.

6.3.2 Separation of variants

6.3.2.1 Cation exchange chromatography

A mixture of self-proteolysed trypsin variants was subjected to CpB digestion (Section 6.2.2) before separation by CEC (Section 6.2.3). The resulting chromatograms are shown in **Figure 6-3**.

Cation exchange chromatography separates molecules on the basis of differences in their net surface charge. Net surface charge can be divided into more specific features: overall charge, charge density and surface charge distribution of a molecule. It was proposed that these properties should be displayed in bovine trypsin variants generated by self-proteolysis and form the basis of separation.

A trypsin molecule with an internal cut inflicted by another trypsin molecule must have a new C-terminus ending in either an R or K residue by definition of the specificity of trypsin. By treating the internally cut molecule with CpB, the C-terminal R or K residue should be cleaved by definition of the specificity of CpB. The variant would then differ from native trypsin in sequence by the absence of one positively charged residue. Changes to

secondary and/or tertiary structure may occur depending on the location of the cleavage site eventually leading to a change in net surface charge.

The example described considers a single cut to a trypsin molecule. It follows that the extent of change from native trypsin's net surface charge should vary with the amount and location of cleavage sites on a variant. Native trypsin contains 13 K and 3 R residues each with a different level of susceptibility to lysis based on surface exposure (Section 4.3.8). If every possible variant was generated, the total library size would be greater than 10^{10} (basis for permutation calculation: 16 × R/K residue locations and up to 16 cleavage sites). However, the real number of variants likely to be generated is severely restricted by the fact that some R/K residues are not surface-exposed and therefore less susceptible to cleavage. Furthermore, those that are surface-exposed may be protected from cleavage by neighbouring residues blocking access to the R/K residue at risk. This indicates that favoured degradation paths must exist which would considerably limit library size to a practical level. Indeed, internally cut trypsins have been studied and the location of the cut(s) elucidated (Smith and Shaw, 1969; Schroeder and Shaw, 1968).

During the CEC elution step, raising the conductivity (NaCl concentration) at a constant rate would elute more weakly bound variants earlier than those with a greater affinity for the cation exchanger. Based on this, the last species to elute would be native trypsin which has not lost any R/K residues meaning it has not undergone any self-proteolysis and

therefore could not lose any net surface charge from CpB removal of R or K residues. Analytical CEC on a small-scale column was run first to act as control with no CpB digestion, and resulted in a minor peak between 7 - 8 mL separate from the main trypsin peak (**Figure 6-3** - top).

Self-proteolysed trypsin which had been CpB digested separated into several peaks (**Figure 6-3** - bottom). Initially there was a large amount of flow-through shown by a broad peak before elution began. This represented variant sample which has been modified by CpB so as to not retain any ability to bind to the resin. Such samples could be fragments of trypsin which lack any R/K residues following CpB digestion. As soon as the conductivity rose, peaks began to separate with poor resolution until a main peak is seen, which is also the last and was expected to have characteristics consistent with native trypsin. Seven fractions were collected starting from the flow-through peak (named S1) to the main peak (S7) with five peaks in between for separated variants (S2 - S6) all of which were characterised (Section 6.3.3).

The control sample of self-proteolysed trypsin minus CpB digestion separated into two peaks. The main peak must represent the several species of internally cut trypsins, which were not separated. A very small peak, which eluted directly before the main peak is seen at poor resolution, and must represent a trypsin variant (or mixture of) that has been modified adequately enough to allow for an earlier elution than the bulk of the sample. Crucially, however, there is a lack of flow-through peaks and major

elution peaks (apart from the main peak) indicating that the trypsin variants were not being separated without the CpB digestion step.

6.3.2.2 Affinity and cation exchange chromatography

An alternative strategy to variant separation was employed which included an affinity chromatography (AC) step prior to CEC. AC is an efficient tool for separating molecules on the basis of a reversible interaction between a protein and a specific ligand coupled to a chromatography matrix. Trypsin has a strong affinity for the active-site inhibitor para-aminobenzamidine (benzamidine) which contains a terminal NH_2^+ group and so acts as a substrate analogue. The protein can be eluted at a pH of 3.0. This may be used as a basis for trypsin purification (Hixson and Nishikawa, 1974).

Use of AC (Section 6.2.3.2) was intended to separate self-proteolysed variants into two batches: (1) those with an affinity for benzamidine and (2) those with no affinity for benzamidine. The batch with affinity for benzamidine would contain variants with any detectable activity (amongst other possible contaminants). The batch with no affinity may not necessarily have been completely redundant since it may have contained variants with specificity for two positively charged residues. A fraction was therefore collected and assayed, although it was later found to be completely inactive.

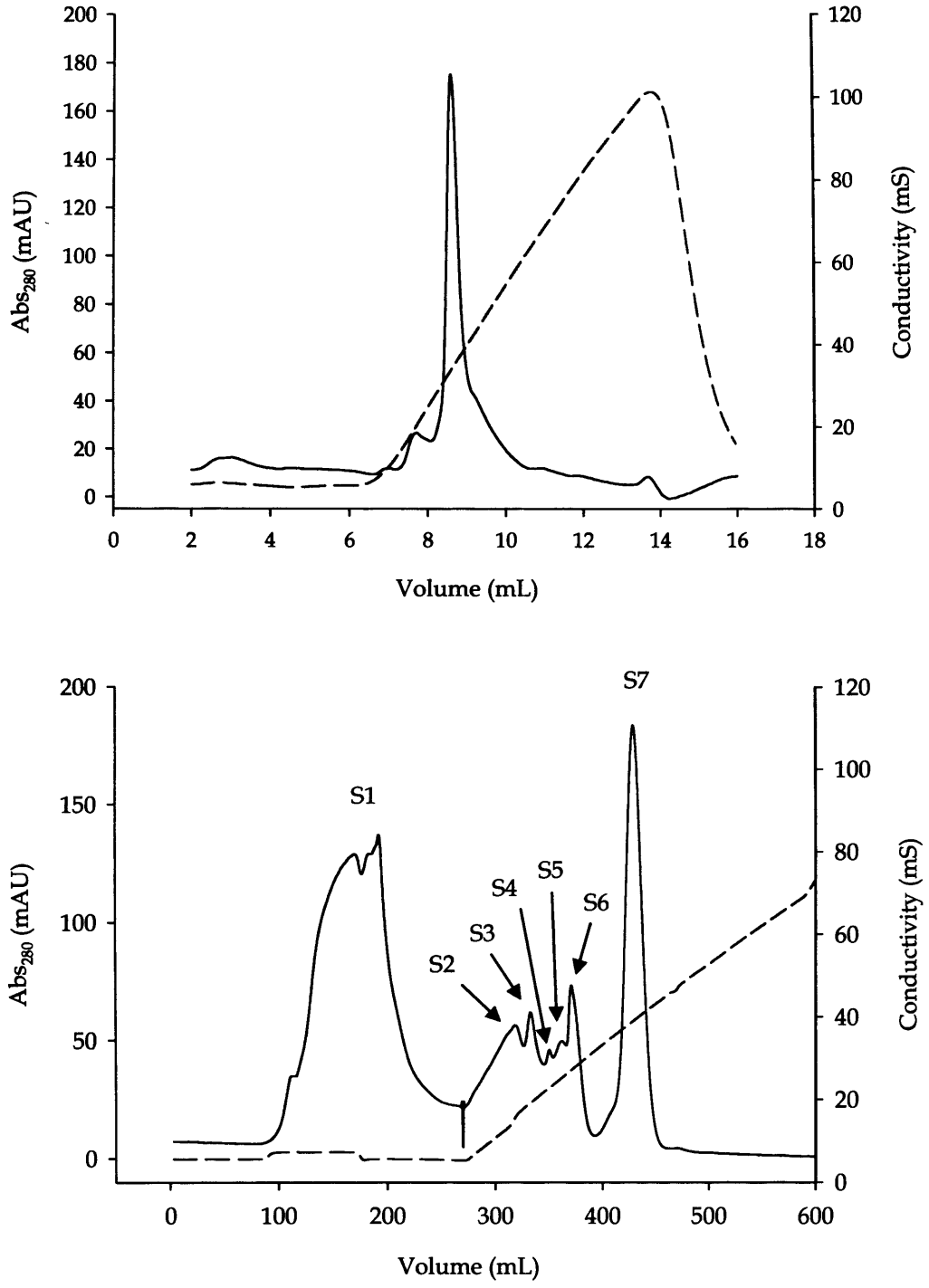

Figure 6-3. CEC chromatograms of trypsin variant mixture ± CpB digestion. Abs_{280} = solid line, conductivity = dashed line. Top: 0% CpB digestion on mono S column. Main peak elution began at 30 mS. Bottom: 5% CpB digestion on SFF column. Main peak elution began at 34 mS. Peaks eluting before this represent trypsin variants with a lower affinity for the cation exchanger.

AC was carried out both with and without CpB digestion of the variant sample, to test any effect CpB had on self-proteolysis (**Figure 6-4**). The main trypsin peaks from both chromatograms reached a similar peak height of ~ 1100 mAU disregarding any effect of CpB. Resolution in both chromatograms is very similar starting with a broad flow-through peak (variants with no affinity) followed by a sharp peak (variants with some affinity). The fraction containing variants with affinity for benzamidine (CpB digested) was put through CEC to separate the different variants (**Figure 6-5**). As expected, the main peak appears (named AS3) which signifies native trypsin. A small shouldered peak appears to the left-side of the main peak (AS2). There appears to be no other significant peaks, however, and two tiny peaks either side of the main peak (AS1 and AS4) which are only clear on the chromatogram once enlarged. This must mean that the flow-through fraction from AC contained the vast majority of variant population. Nevertheless, the four fractions were collected and characterised.

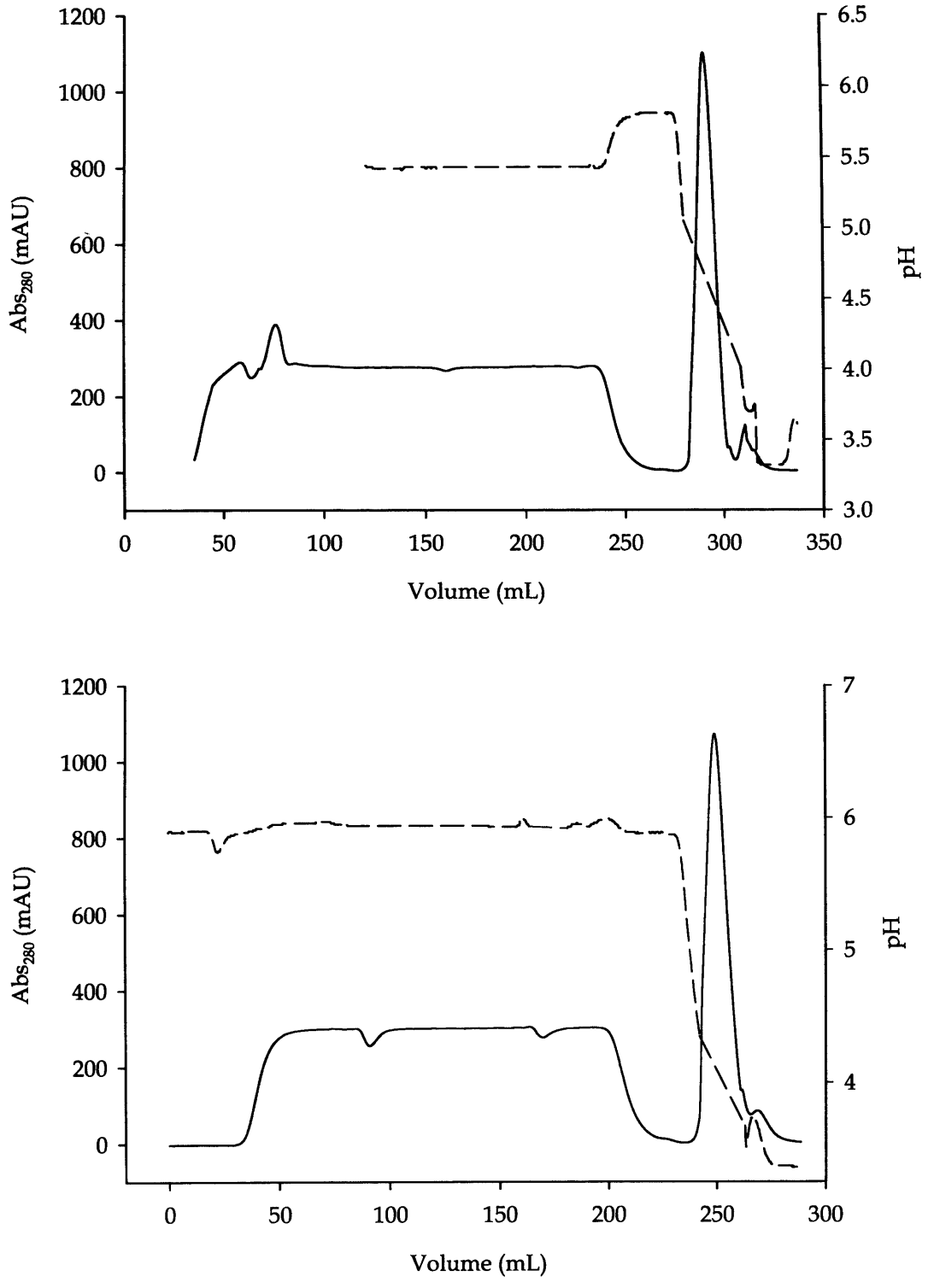

Figure 6-4. AC chromatograms of variant mixtures ± CpB digestion. Abs_{280} = solid line, pH = dashed line. Top: 0% CpB digestion. Bottom: 5% CpB digestion.

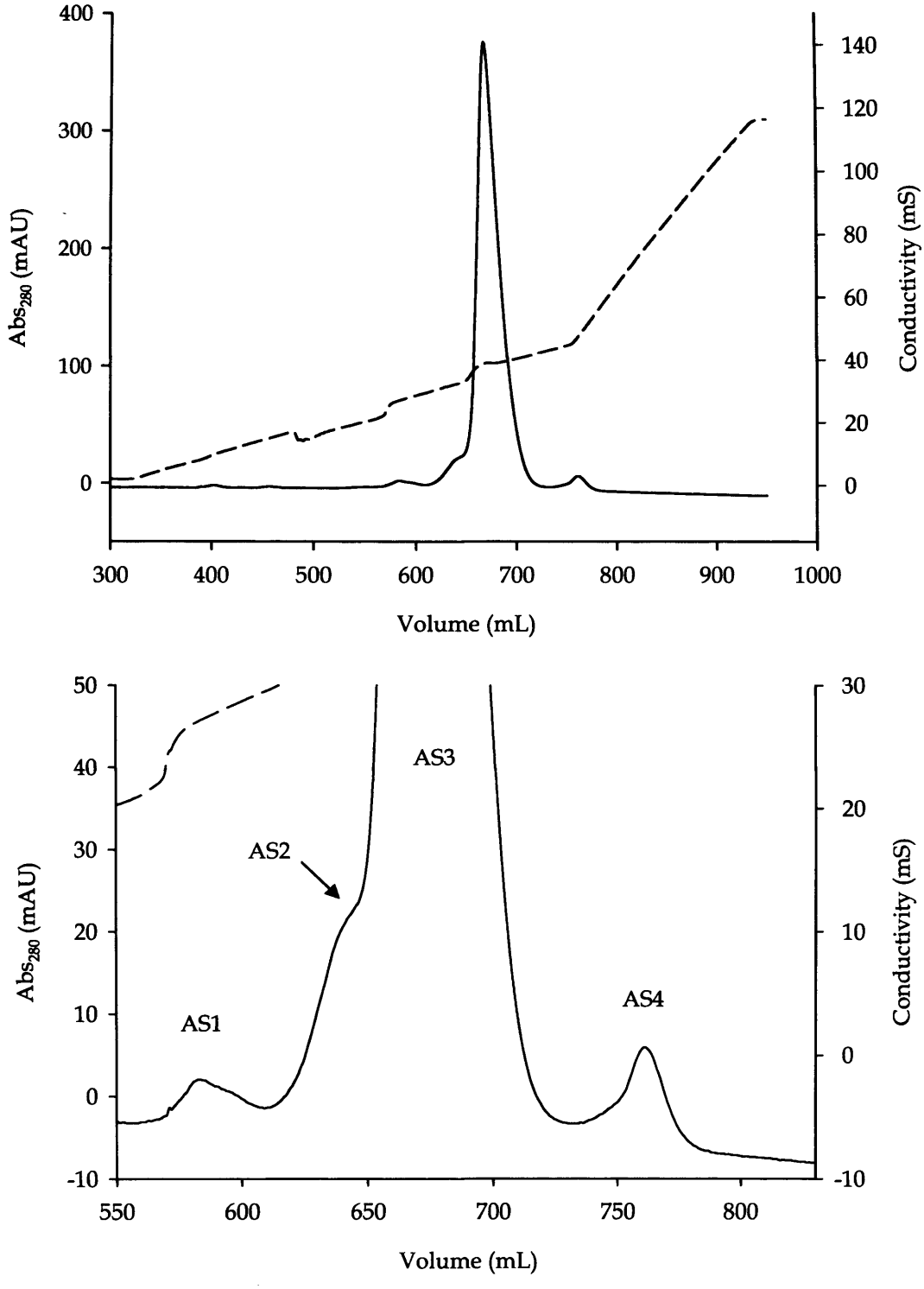

Figure 6-5. CEC chromatogram of CpB digested, AC purified variant mixture. Abs_{280} = solid line, conductivity = dashed line. The region of peak resolution has been enlarged (bottom).

6.3.3 Characterisation of variants

6.3.3.1 Colourimetric assays

Variants separated by CEC alone (S1 - S7) and by CEC following AC (AS1 - AS4) were assayed for activity on Z-RR-pNa substrate (Section 6.2.4.1). Reaction profiles are shown in **Figure 6-6**.

Variant fractions S1 - S6 showed a negligible increase in OD_{405} over the course of the assay. From this it can be concluded that all types of variant obtained from CEC were fully inactivated by self-proteolysis (when assayed over a five minute period). Fraction S7 showed activity consistent with that of r-Trypsin (Section 6.3.5). Trypsin that had escaped self-proteolysis during the over-activation may have been retained in this fraction in addition to internally cut trypsins that had not separated from the undegraded trypsin. Fractions AS1 - AS4 showed a greater variability in activity. AS3 showed activity consistent with r-Trypsin and is likely to have been composed of trypsin material similar to S7. Fraction AS1 had no detectable activity over the course of the assay whilst AS2 (shouldered to AS3) had a nominal increase in OD_{405} likely to be attributed to the spill-over of AS3 during fraction collection. Fraction AS4 is unusual in that it was the last species eluted from the CEC column yet only displayed ~ 25% of the increase in OD_{405} compared to AS3.

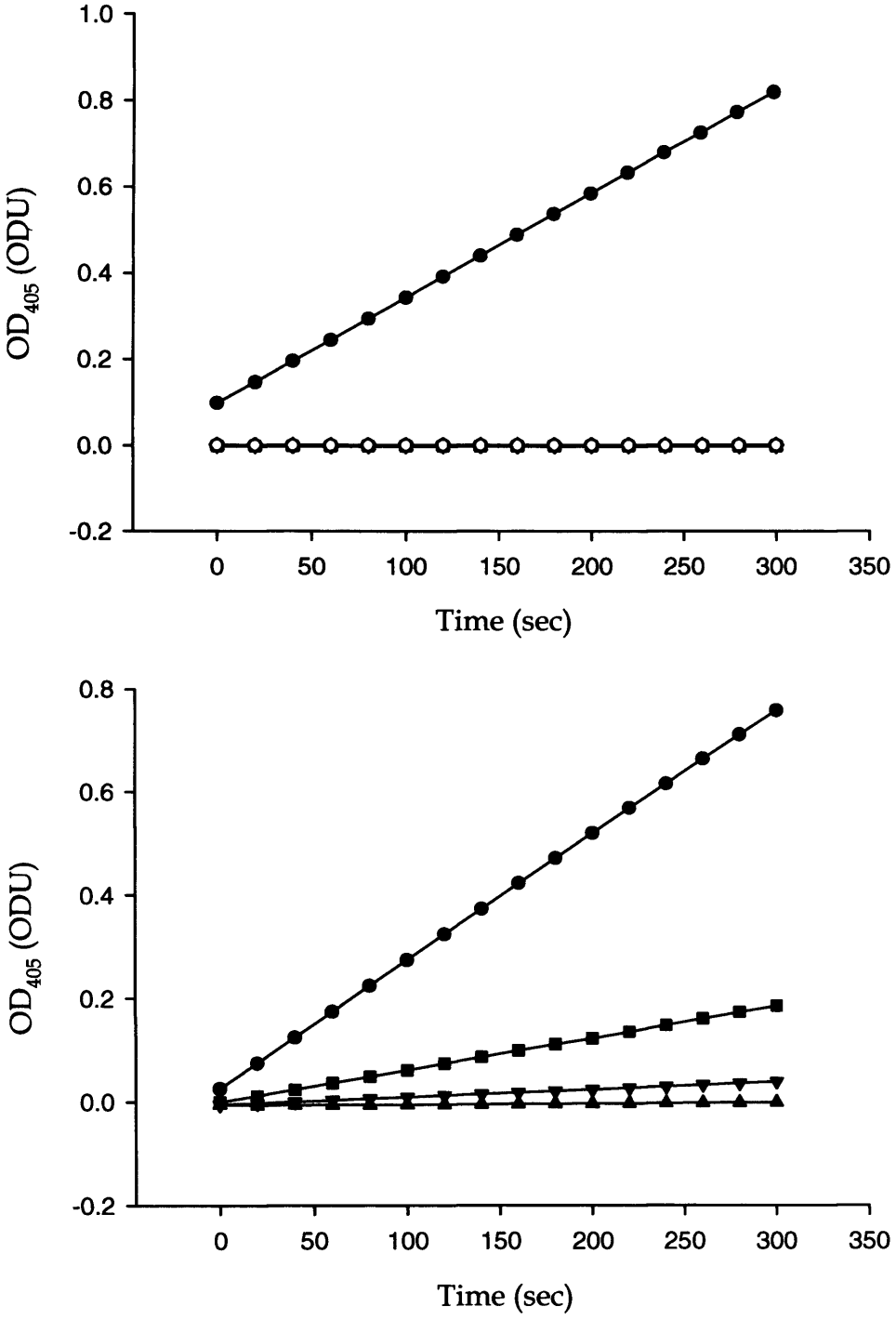

Figure 6-6. Colourimetric activity assays of fractions S1 - S7 and AS1 - AS4 on Z-RR-pNa substrate. Top: ▼ = S1, ▲ = S2, ▫ = S3, ▪ = S4, ◊ = S5, ○ = S6, ● = S7 (no significant signal given by any fraction except S7). Bottom: ▲ = AS1, ▼ = AS2, ● = AS3, ▪ = AS4.

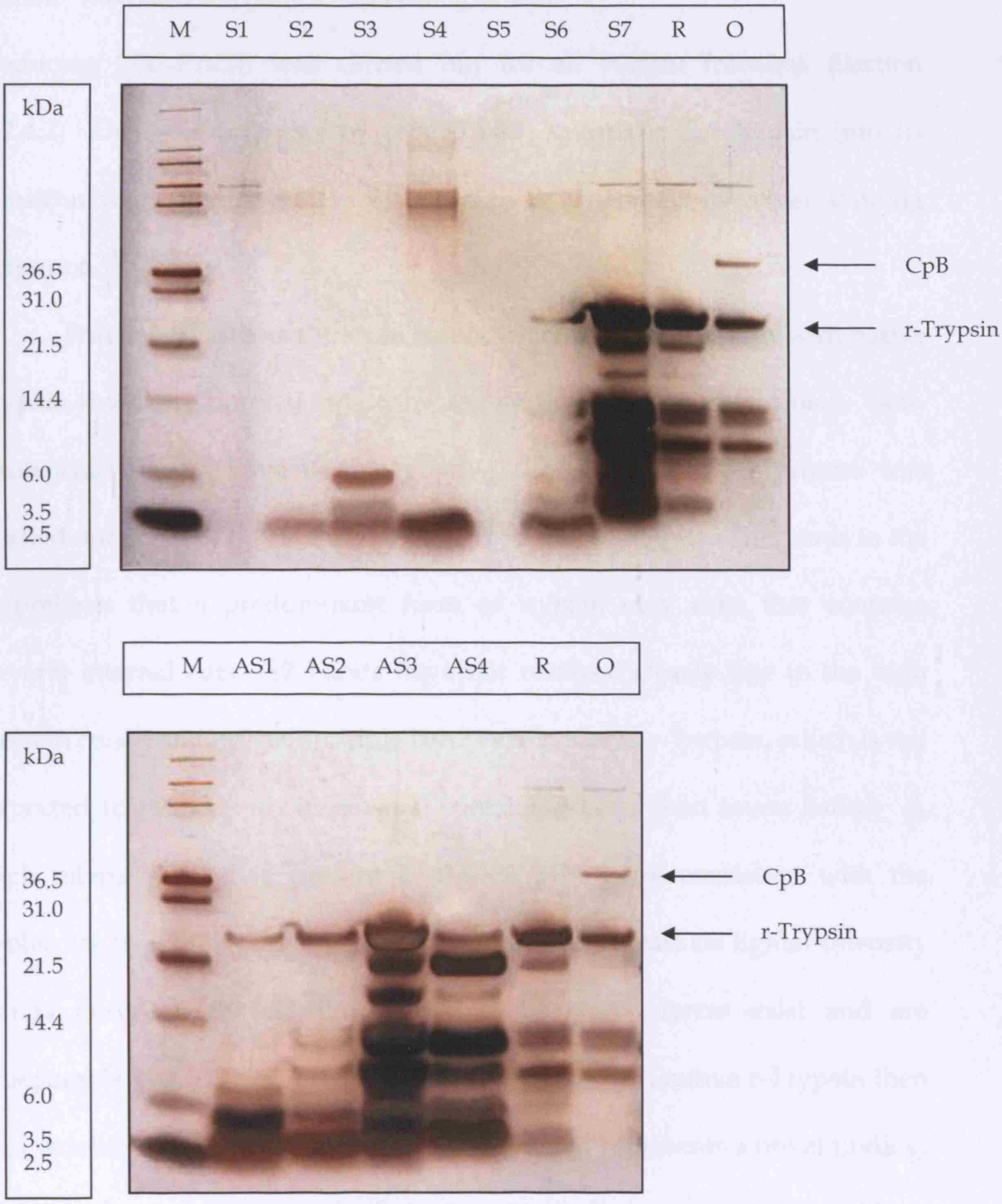

Figure 6-7. Reducing SDS-PAGE of fractions S1 – S7 (top) and AS1 – AS4 (bottom). M = Mark12™ unstained marker, R = r-Trypsin (not self-proteolysed), O = over-activated r-Trypsin + 5% CpB.

6.3.3.2 Molecular weight determination of variants

Reducing SDS-PAGE was carried out for all variant fractions (Section 6.2.4.2). This was intended to reduce each internally cut trypsin into its constituent fragments and to visualise these fragments by silver staining (**Figure 6-7**).

Fraction S7 shows multiple bands which is not consistent with native trypsin since any internal cuts were not expected to be present here. Self-proteolysis should have been negligible after CEC since the process was carried out at pH 3.0 - 3.5 and fractions were stored at 4 °C. This leads to the hypothesis that a predominant form of trypsin may exist that contains several internal cuts. S7 bands have not resolved clearly due to the high protein concentration yet multiple bands are present. r-Trypsin, which is not expected to reduce into fragments, contains greater than seven bands. A high intensity band is present at the 23 kDa mark consistent with the molecular weight of native trypsin in addition to at least six lighter intensity bands fairly evenly distributed. If such trypsin forms exist and are inseparable by CEC yet retain activity consistent with native r-Trypsin then this would not be detected by activity assays and represents a novel finding. The sample of over-activated r-Trypsin + 5% CpB shows a similar profile to r-Trypsin minus one band which is directly below the band at 23 kDa. The missing band may be due to action of CpB hastening the transition of this form to a different self-proteolysed form of trypsin.

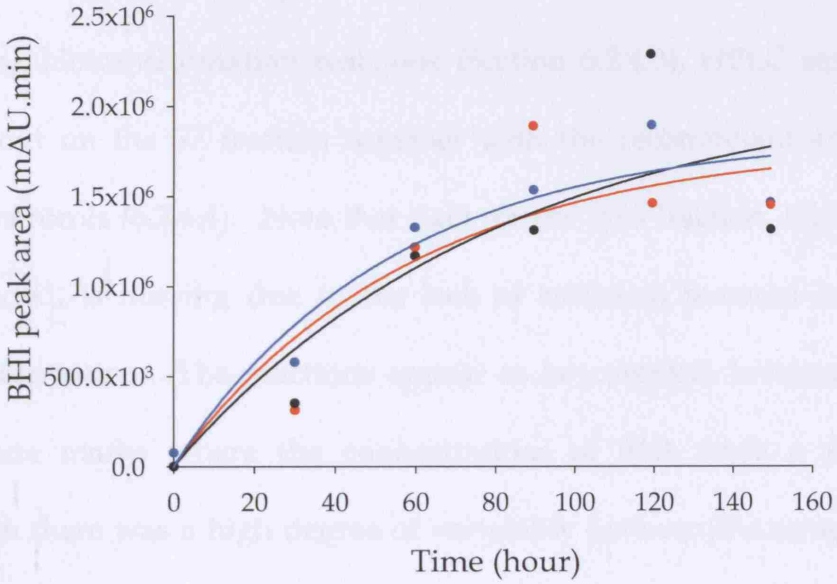

Figure 6-8. HPLC analysis of KPB-HPI biotransformation of S7 fraction, r-Trypsin and bovine trypsin samples. The y axis displays the peak area of KPB-BHI (or converted insulin) as measured by HPLC. Blue circles = bovine trypsin, red circles = r-Trypsin and black circles = fraction S7. The same colouring system has been used for fitted curves. The curves were fit to an exponential equation of the form $y = a(1 - e^{-bx})$. R^2 values: 0.98 for bovine tryspin; 0.95 for r-Trypsin; and 0.93 for fraction S7. Note that individual HPLC traces are not shown as BHI peak areas were taken as an indicator of product yield.

6.3.3.3 HPLC analysis of enzymatically converted KPB-HPI

Following biotransformation reactions (Section 6.2.4.3), HPLC assays were carried out on the S7 fraction together with the recombinant and bovine trypsin controls (6.2.4.4). Note that data for the AS3 fraction, which was to be included, is missing due to the lack of sufficient material left for the biotransformation. The reactions appear to be complete between the 90 - 120 minute marks where the concentrations of BHI reach a maximum. Although there was a high degree of variability between the sample points, all have a similar regression line fit, which follows a quadratic equation of the form $y = a(1 - e^{-bx})$ (**Figure 6-8**). The recombinant enzyme and S7 fraction have reaction profiles consistent with that of bovine trypsin. This indicates that the S7 fraction, obtained from over-activated r-Trypsin, is most likely r-Trypsin that had not yet been digested.

6.3.4 Trypsin activity on a range of chromogenic substrates

Activity assays were performed on a variety of pNa substrates (Section 6.2.5) in an attempt to better define the specificity of trypsin alongside current literature. Substrates available commercially varied in terms of the following characteristics: (1) presence or absence of an N-terminal blocker (Z, Bz or none); (2) presence or absence of a P2 residue (R, K, G, or none) and (3) the residue present at position P1 (R or K). All substrates featured the pNa marker group at the C-terminus which, upon liberation from a substrate, can be detected spectrophotometrically at 405 nm absorbance. Assays were

carried out on both commercial bovine trypsin and Eli Lilly's recombinant trypsin (production outsourced to Sandoz, Austria) as a point of comparison. The profiles of all reactions are shown in **Figure 6-9**. The first thing to note is that there was no significant discrepancy between the activity of bovine and r-Trypsin on all substrates tested. Activity of r-Trypsin (red lines) closely mimicked that of bovine trypsin (blue lines). All assays on R-pNa and K-pNa gave no significant increase in OD_{405} over the course of the five minute reaction showing that there was no discernible activity over this time frame. This is consistent with the specificity of trypsin being defined as posessing only endoprotease activity *i.e.* it cannot cleave at the C-terminal side of an R or K residue unless a residue or blocker is present on the N-terminal side of that R or K residue. Where this condition was satisfied, a range of activity levels were detected.

Dipeptide substrates gave greater absorbance increases (OD_{405}) than single residue substrates. Of the dipeptide substrates, Z-KR-pNa and Z-RR-pNa comfortably gave the highest assay signals followed by Z-GR-pNa. It may be concluded from this that trypsin had a higher preference for cleavage at the C-terminal side of two positively charged basic residues rather than just one. This was a novel finding that had not been reported previously in literature.

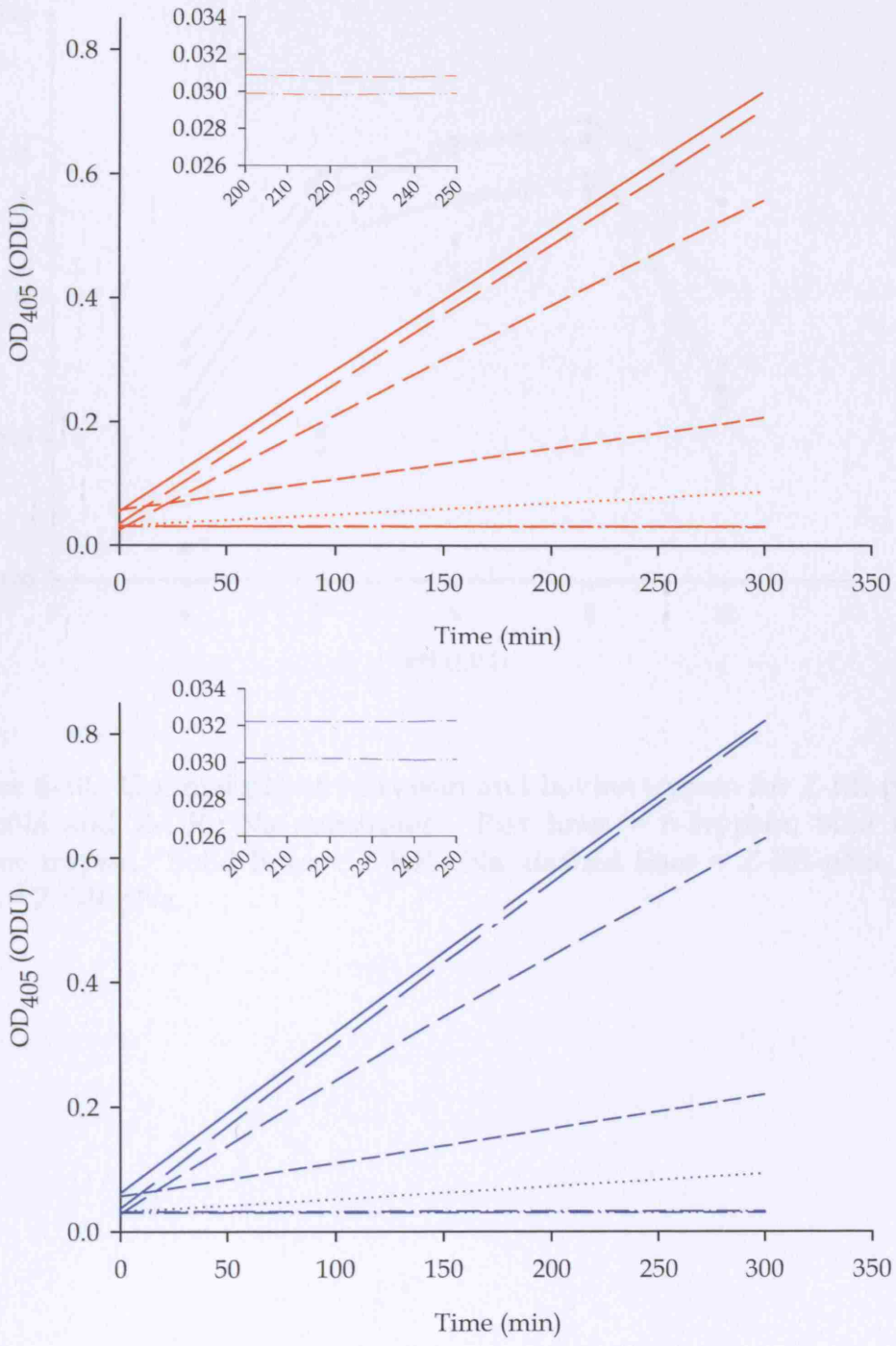

Figure 6-9. Reaction profiles for colourimetric assays of bovine and r-Trypsin on various pNa substrates. Red lines = r-Trypsin, blue lines = bovine trypsin. Solid lines = Z-KR-pNa, long dash = Z-RR-pNa, medium dash = Z-GR-pNa, short dash = Z-R-pNa, dotted line = Bz-R-pNa, dash-dot = R-pNa, dash-dot-dot = K-pNa.

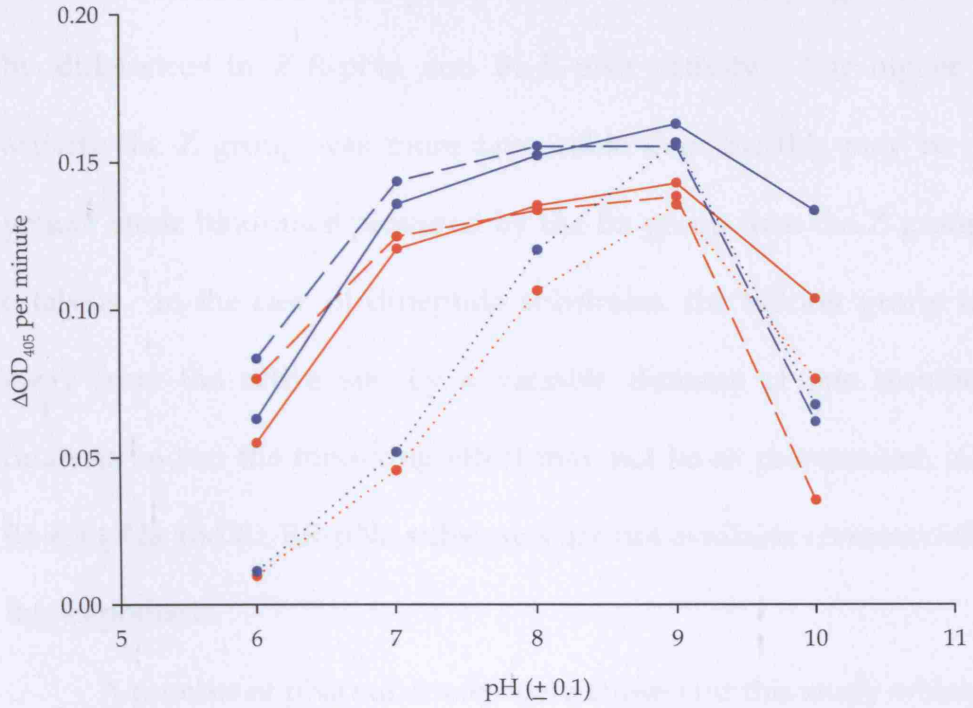

Figure 6-10. Optimal pH of r-Trypsin and bovine trypsin for Z-RR-pNa, Z-KR-pNa and Z-GR-pNa substrates. Red lines = r-Trypsin, blue lines = bovine trypsin. Solid lines = Z-KR-pNa, dashed lines = Z-RR-pNa, dotted lines = Z-GR-pNa.

The choice of blocker group can influence activity significantly as seen by differences in Z-R-pNa and Bz-R-pNa activity. For higher enzyme activity the Z group was more favourable than Bz; this may be due to a greater steric hindrance provided by the Bz group than the Z group during catalysis. In the case of dipeptide substrates, the blocker group is further away from the active site by a variable distance of one residue during catalysis and so the hindering effect may not be as pronounced. Currently Bz-KR-pNa and Bz-RR-pNa substrates are not available commercially to test this hypothesis.

A number of pNa substrates were chosen for this study which offer an assessment of trypsin's peptidase activity rather than its esterase activity as offered by the historically used TAME assay. A novel means to rationalise the specificity of trypsin has been revealed by comparing its activity on three dipeptide pNa substrates. The influence of positions P2 and P3 of a substrate may be studied in more detail using specially synthesized substrates. The use of adopting such a model may be similarly applied to other proteases for studies in relation to their substrate specificity.

6.3.5 pH optima of selected trypsin substrates

The substrates Z-RR-pNa, Z-KR-pNa and Z-GR-pNa were chosen for pH optimum studies since dipeptide substrates are closer to the recognition site of interest in KPB-HPI than single residue substrates. Activity assays were

performed at pH 6.0, 7.0, 8.0, 9.0 and 10.0 to an accuracy of ± 0.1 pH unit (Section 6.2.6).

The highest activity increases per minute were at pH 9.0 for all substrates (**Figure 6-10**). It may be seen from the plots that the activity of bovine trypsin closely resembles that of r-Trypsin. All plots feature a single maximum peak (with varying degrees of clarity) at which the optimal pH was found. Although the plots were consistent with a "bell shape", more data points would be required to confirm this trend. Plots for Z-GR-pNa gave the clearest indication that 9.0 was the optimal pH since the plots feature sharp peaks for this substrate. Plots for Z-KR-pNa and Z-RR-pNa also had optimal pHs at 9.0 although there was only a marginal difference between this pH and pH 8.0. This suggests that assays may be carried out between pH values of 8.0 and 9.0 to ascertain the optimal pH more accurately.

7 General discussion

7.1 Project summary

Bovine trypsin is an industrially useful protease. It catalyses hydrolysis reactions at the C-terminal end of arginine and lysine residues within a polypeptide. It has, therefore, found use as a biocatalyst in the conversion of human pro-insulin to active insulin. The process does not require such a broad substrate specificity, however, and a better process yield would result from directing the specificity of trypsin even more towards arginine residues. The overall aim of the project was to utilise methods within the field of enzyme engineering by which this target could be achieved.

A random mutagenesis strategy utilising fepPCR was employed as the first approach to modifying the specificity of trypsin in Chapter 3. This was chosen on the basis that there had been several reported successes using this method for the enhancement of enzyme properties including substrate specificity. Mutant trypsin libraries were generated, expressed in microwells, and screened for modified specificity. The following objectives and conclusions were reached in this chapter:

- Activity assays on pET(T) lysates demonstrated cleavage of the L-BANA substrate that was at least 6.5-fold greater than the baseline present with the pET26b(+) control. AEBSF incubated during cell growth after the IPTG induction inhibited the L-BANA cleavage by approximately 50%, indicating that the L-BANA cleavage was caused by a serine protease, *i.e.* the bovine trypsin. There was a lack of dependence of activity on IPTG

induction most likely due to a balance between the toxicity of trypsin being expressed and the leakiness of the T7 promoter.

- A total of 27 verified mutations were observed in 20 sequences after carrying out fepPCR successfully on a 522 bp stretch of the trypsin gene. This equated to a mutation rate of 1.35 per fragment (2.59×10^{-3} per nucleotide). 51.8% of the mutations were transitions while 48.2% were transversions. Given that there were twice as many possibilities for transversions as there were for transitions, this indicated a fairly strong bias towards transitions. The expression vector was successfully re-constructed around the fepPCR products by MEGAWHOP.
- Three mutants were identified from the screening stages with arginine to lysine preferences of between 2.31 and 2.56 compared to the wild-type preference of 1.54 . Two of these were double mutants (N79S, N115S and S146G, I180S), and had decreased overall activity on both arginine and lysine. The other was a single mutant (N79S), which had an increased overall activity on arginine (1.33 fold) and decreased overall activity on lysine (0.79 fold).

Building on the findings of chapter 3, the next goals of the project were to increase the enzyme's resistance to autolysis and enhance specificity beyond the levels obtained by fepPCR. In Chapter 4, mutagenesis targets for achieving these goals were proposed based on a study of enzymes previously enhanced by directed and rational evolution. Various distance

and entropy correlations were revealed, and structural analyses were carried out on bovine trypsin in order to narrow down the sequence search space. The following objectives and conclusions were reached in the process:

- A study of distance correlations revealed that selecting more regions of catalytic association (key catalytic atom, substrate and cofactors), reduced the total target residues required for successful enhancement. 84% of enhanced mutants occurred in only 7% of the total residues when a distance cut-off of 6 Å was used. However, to achieve the same level of success using only a the key catalytic atom as the reference point, a 14 Å cut-off was required and 23% of the total residues were within this distance.
- A study of entropy correlations revealed that mutations at highly conserved sites were more likely to have increased specificity towards a non-natural substrate rather than towards a natural substrate. Mutations at both highly conserved and phylogenetically variant sites were present for enhancements in natural activity.
- Active site residues of bovine trypsin were separated into two regions comprising the binding pocket. BP1 consisted of K188, D189, S190 and Q192, and BP2 consisted of W215, G216, S217 and G218. These residues were alleged to have the greatest bearing on substrate specificity, and proposed as targets for enzyme engineering.
- Following a structural analysis and literature review, K60, R117 and K145 were designated the residues most susceptible to self-proteolytic cleavage

within the bovine trypsin molecule, and therefore were proposed as mutagenesis targets.

The target sites identified were the subject of two mutagenesis strategies in Chapter 5: site-directed mutagenesis for the sites important to self-proteolysis and multiple-site saturation mutagenesis for the sites important to substrate specificity. The following objectives and conclusions were reached in this chapter:

- The triple-mutant plasmid, pET(T3), was successfully produced to encode a mutant with increased resistance to self-proteolysis. While it was found to be no more stable than the wild-type enzyme, it did have a more consistent level of activity over time by removing an initial increase in activity present for the wild-type. However, over a 72 hour period, the loss of activity was not shown to be significant indicating that self-proteolysis was not as significant as previously thought.
- A novel application of the MSSM technique was utilised successfully for the randomisation of two regions spanning four residues each. 20% of BP1 and 30% of BP2 sequences contained the mutagenic insert. A higher insert ratio could not be obtained due to the difficulties in trying to mutate such a long stretch of nucleotides.
- Electrocompetent cells were produced successfully from an arginine auxotrophic strain of *E. coli* with a transformation efficiency of 1×10^5 transformants per microgram of DNA.

- The nutritional selection method used did not identify any trypsin variants capable of novel activity. The target specificity may have been out of reach requiring a major re-design of the enzyme to achieve the target.

In Chapter 6, a new approach to generating trypsin variants was employed based on the enzyme's ability to self-proteolyse in solution. The following objectives and conclusions were reached:

- At the high concentrations tested (1.5 g/L), r-Trypsin rapidly lost activity over the course of an over-activation reaction due to autolysis, retaining only 38.4% (with Ca^{2+}) and 34.0% (without Ca^{2+}) of initial activity after 25 and 26.5 hours respectively.
- Following carboxypeptidase B digestion of the over-activated reaction products, cation exchange chromatography successfully separated bovine trypsin variants into seven fractions. All except one fraction were found to be inactive. The active fraction had a similar profile to that of r-Trypsin, and was presumably native trypsin unaffected by self-proteolysis.
- A modification to include affinity chromatography prior to cation exchange chromatography successfully separated the variants into four fractions. Only one fraction possessed activity similar to r-Trypsin; the remaining fractions were allegedly inactive variants, or the result of neighbouring peak contamination during the chromatography run.

- Characterisation of the active variant fractions using SDS-PAGE, biotransformation and activity assay methods revealed that they had traits comparable to that of r-Trypsin.
- Given the effectiveness of using cation exchange chromatography to separate active from inactive trypsin variants, this method may be considered a viable alternative to traditionally used affinity chromatography for the purification of industrial quantities of r-Trypsin. This method also has the added advantage of proceeding at low pH throughout the process, temporarily inactivating the trypsin and hence minimising the level of self-proteolysis.
- r-Trypsin and bovine pancreatic trypsin had very similar kinetic profiles when assayed on several substrates. Most revealing and previously unreported is the finding that trypsin preferentially cleaves at the C-terminal end of two positively charged basic residues (KR or RR) rather than after a single residue (GR).
- The Z blocker group on substrates was more amenable to trypsin activity than the Bz group with activities approximately 2-fold higher.
- The optimal pH across all substrates tested was found to be 9.0 under the reaction conditions used.

7.2 Project appraisal

Conclusions drawn from all of the results chapters led to a number of general points of interest. Considering mutagenesis strategies first, no

mutants of bovine trypsin were obtained with improvements of the magnitude typical for directed evolution studies (*i.e.* > 3-fold) indicating one of the potential limitations of using a random mutagenesis strategy such as fepPCR. An alternative explanation is that the enzyme bovine trypsin may already be so highly evolved towards arginine and lysine activity that no amount of alterations may produce the kind of improvement in specificity typical for directed or rational evolution approaches. With regards to specificity improvements, it is also important to consider that arginine and lysine share some key properties: both possess long side chains of similar length with positive charges on the ends and both are hydrophilic in nature. Evolving an enzyme towards a significant enhancement in specificity may not have been a realistic target in hindsight. Despite this, subtle improvements were made to both overall activity and specificity thereby achieving the initial project goal somewhat although further characterisation was needed.

The screening method used for libraries produced by fepPCR was valuable only in the context of a three-tier system designed to root out false positives at any stage. Indeed, false positives were present at each screening stage and this emphasised the importance of adopting such a screening system. Effects of autolysis, variable enzyme concentrations from well to well and the toxicity of the enzyme to the host cell may each have had some influence on the occurrence of false positives, however, the method proved

to be the most useful compared to the nutritional selection method tested subsequently.

In theory, increasing the library size by carrying out MSSM of target sites should have resulted in a better chance of success. However, the selection method that had to be adopted for such large libraries, was only suitable for selecting novel exoprotease activity. Unfortunately, no useful mutants were obtained using this approach. Ultimately, the large MSSM libraries may have yielded mutants with enhanced activity or specificity by screening in microwells on the same arginine and lysine substrates used in Chapter 3. This approach, however, was not practical given the time it would take to complete such a major screening effort. This suggests that smaller focussed libraries generated by random mutagenesis of a target region or saturation mutagenesis of selected sites may be the best means of exploring the project goal further.

Supplementary work carried out on distance and entropy correlations yielded some interesting findings, and allowed for a more targeted mutagenesis strategy which made use of MSSM. The technique was successful to a degree (20 - 30% insert ratio) and showed that it was possible to produce large combinatorial libraries (160,000). Such libraries have rarely been mentioned in the literature, especially from the perspective of enzyme engineering accomplishments. With current trends towards automation in the laboratory, particularly for the evolution of biocatalyst libraries (Ferreira-Torres *et al.*, 2005), high-throughput screening efforts are becoming more and

more able to process larger libraries. An alternative to narrowing down the sequence search space, therefore, is to embrace new technology and use it for the screening of large libraries. The advantage to this strategy is that it is more comprehensive and is likely to miss out fewer enhancing mutations. MSSM has been shown to be a viable option for the creation of large libraries, and optimisation to increase the insert ratio would be the next stage.

Some important observations were made by characterising the autolysis of bovine trypsin, which was initially intended for use as an alternative means to produce variants of trypsin. Having obtained several commercially available substrates, it was found that bovine trypsin had the greatest activity at the C-terminal ends of RR or KR peptide substrates. Historically, literature on bovine trypsin has conveyed that the enzyme is specific towards cleavage at the C-terminal end of single arginine or lysine residues without the dipeptide distinction. The new definition of bovine trypsin specificity must be considered in future efforts to engineer this property, which may lead to the design of novel assay and screening procedures.

A study on the effects of autolysis on the kinetics of bovine trypsin was not found in literature, and where constants are given a Michaelis-Menten model has been assumed. Kinetic values relating to trypsin were not sucessfully obtained, however, the need to consider an autolytic pathway in addition to the usual enzyme-substrate pathway would need to be

incorporated into a new model. This would provide more accurate kinetic constants important for the characterisation of wild type bovine trypsin as well as for other self-proteolytic enzymes.

7.3 Future work

Taking all of the discussion points into account, there are many directions in which to take the project:

- Mutations from the double mutants from fepPCR may be separated and produced as single mutants. All of these mutants may be purified and characterised further in terms of kinetic constants.
- Two smaller libraries may be created by carrying out fepPCR on sites BP1 and BP2 and following up with microplate screening for arginine and lysine activity.
- Microplate screening of old and new libraries may be carried out on the useful dipeptide substrates (Z-KR-pNa and Z-RR-pNa) identified in Chapter 6.
- In order to increase the likelihood of generating a useful mutant, computational methods may have a part to play. Computational protein design methods use algorithms to identify amino acid sequences that have low energies for target structures. This may be of use in predicting active site configurations that confer modified substrate specificity.
- Various statistical tools may also be used to narrow down the sequence search space and introduce rational mutations. Methods utilising

principle component analysis such as statistical coupling analysis, for example, may identify residues that are coupled by coevolution and motion.

- In order to circumvent expression problems experienced, *in vitro* translation or compartmentalisation techniques may be tested. These cell-free methods discount the need for a microbial system altogether. The advantage of this is that the background rate of substrate hydrolysis would be minimised during screening. There is also likely to be a better control of expression levels without variations of cell growth rates and plasmid copy numbers in microbial systems.
- In order to better characterise the autolysis of trypsin, a strategy may be attempted to deduce kinetic constants for the cannibalism of trypsin so that apparent *kcat* and K_M values obtained with typical substrates may be adjusted closer to their true values. Autolysis reaction profiles are needed with trypsin concentrations at several time points in order to ascertain the constants and produce a kinetic model different to that of the standard Michaelis-Menten model.

8 Economic and validatory implications of a process change

8.1 Introduction

A major part of the control of medicinal products relates to current good manufacturing practice (cGMP), which is defined as the 'component of quality assurance (QA) that ensures products are consistently produced and controlled to the quality standards appropriate to their intended use and as required by marketing authorisation or product specification'[1]. There are a number of professional bodies that regulate the industry and provide guidelines for all aspects of production. These are: (1) the Food and Drug Administration (FDA) in the USA; (2) the European Medicines Evaluation Agency (EMEA) in Europe; and (4) the Medicines and Healthcare Regulatory Agency (MHRA) in the UK.

This chapter studies the impact that incorporating a modified trypsin would have on validating the industrial bioprocess it is used in, as well as the general economic benefits that would follow. Recombinant bovine trypsin is currently used in the biotransformation step in the production of insulin (**Figure 1-7**), specifically for the conversion of pro-insulin to mature insulin. If the current trypsin is replaced by an engineered protease with enhanced substrate specificity, there would be several implications, which are discussed here.

---

[1] Rules and Guidance for Pharmaceutical Manufacturers and Distributors 2007 (the 'Orange Guide'), Medicines and Healthcare Regulatory Agency.

8.2 Revalidation of the biotransformation step

8.2.1 Scope of the task

In order to ensure that an industrial process is GMP compliant, it needs to be validated. To validate a process is to establish documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality characteristics[2]. Validation is not only a government regulation, it is also important for gaining both assurance of quality and cost reductions. Improvements in quality and better consistencies are frequently gained as a result of process validation (Bennet and Cole, 2003), as is process efficiency that leads to less wastage, less redesigning and fewer other financial setbacks.

A validation master plan needs to be drawn up, which is a formal document that describes the general philosophy, expectations, intentions and methods to be adopted for the study. For the purposes of this study, the modification being dealt with is that of the enzyme used in the biotransformation step. Tests will need to be developed to address the following queries:

- Proportion of active insulin produced relative to contaminant product. HPLC assays may determine this using a reference standard as point of comparison and mass spectrometry may determine the molecular weight of all species present.

[2] Guideline on General Principles of Process Validation, FDA.

- Stoichiometrics of transformation reaction as tested at laboratory and pilot-plant scale.
- Identification of any new contaminant species.
- Rate of degradation of the mutant trypsin and carboxypeptidase B enzymes.
- Process monitoring of reaction conditions such as pH and temperature.
- Time for completion of the reaction as monitored by HPLC analysis of samples at regular intervals.
- Stability of active insulin in the reaction vessel over prolonged periods.
- Removal of all product and reactant traces during cleaning.

Standard operating procedures (SOPs) must be written, which will include a description of the acceptance criteria. These are the product specifications that must be met, such as the acceptable and unacceptable quality levels necessary for making a decision on whether to accept or reject a batch. Any tests and assays need to be repeated to have confidence in the reproducibility of the results.

8.2.2 Design qualification and installation qualification.

Design qualification (DQ) is providing documented evidence that quality is built into the design. This includes such topics as the facility, environment, personnel flows, materials flows, equipment flows, general equipment design, maintenance and waste management. The requirements of the

process, product and user all need to be met as does the current GMP guidelines.

Installation qualification (IQ) studies establish confidence that the process equipment and ancillary systems are capable of consistently operating within established limits. Essentially, the design specifications need to be met during any installation of equipment or systems. This may include various static checks such as utilities connections, equipment inventories, instrument calibrations, materials qualifications and maintenance checks.

Both DQ and IQ will have been carried out with a regard for revalidation procedures for the original process, and as such will still apply since the change in the process is to a raw material rather than to any hardware or software. Any details given in the original documents regarding revalidation will need to be addressed although ultimately no additional equipment or system is envisaged as a result of the proposed process change.

8.2.3 Operational Qualification and Performance Qualification

Operational qualification (OQ) is the documented verification that equipment and ancillary systems perform as intended throughout the anticipated operating ranges. There are various functional checks on the equipment, generally performed using inert materials such as water or

compressed air, and in the absence of real product. Tests will have been designed to show that equipment performs as intended and to specification.

Performance qualification (PQ) is the documented evidence that the process operates within the established parameters, and performs effectively and reproducibly to produce a product meeting its predetermined specifications and quality attributes. As with OQ, the critical parameters and acceptance criteria of the system should be defined. PQ requires the examination of consecutive batches so that any variability can be demonstrated and shown to not affect product quality. With regards to the process change in question, this would require several biotransformations to be carried out. Various tests would be performed and product samples would be analysed as described in the validation master plan (see Section 8.2.1).

8.2.4 Cleaning validation

The creation and implementation of effective cleaning processes are essential for any biopharmaceutical production process. This is so to prevent the accumulation of dirt and microbial contamination which could affect product quality, and also to minimise the cross-contamination of one active product into a subsequent product. Since the process definition has changed, the new level of cleanliness that is achieved by the cleaning process must be measured. Samples may be taken to: (1) detect contaminants

following cleaning; (2) analyse and quantify the amount of contaminant; and (3) extrapolate the results to determine if the levels are acceptable.

8.3 Economic benefits

Several economic benefits may follow from replacing the original non-specific trypsin in the biotransformation step with an engineered mutant tighter in specificity. The most direct improvement would be to the yield of product (*i.e.* active insulin) produced. The ratio of active insulin produced is roughly 90% relative to an insulin-like contaminant which makes up the other 10%. At an industrial scale, such a loss translates to significant financial losses especially considering that insulin and its variants are the principle products of Eli Lilly's portfolio of therapeutics. These losses are not only felt at the biotransformation step, they will inevitably carry through to the remaining purification processes. Any improvement in yield is therefore likely to have a knock-on effect on the rest of the process by improving the overall yield of product that is produced finally.

An improvement in yield of 5% or greater, would potentially be significant enough to discard with the use of one of the purification steps that follows the biotransformation. The process currently utilises reverse-phase chromatography followed by size exclusion chromatography. The former is primarily used to remove the insulin-like contaminant that results from the non-specific cleavage of trypsin at the unwanted site in pro-insulin. If the concentration of this contaminant was significantly reduced, the final

chromatography step may be able to process the input material as efficiently as required. A separate re-validation study, investigating the overall impact of the change, would need to be carried out to confirm this. It should also be determined whether or not insulin manufactured via this alternative route, would be considered a new chemical entity that needs to undergo clinical trials. The cost of this, however, may negate the overall economic benefits of implementing the novel protease, in which case the protein engineering technology may be better served if transferred to the development of future bioprocesses.

References

Abecassis, V., Pompon, D. and Truan, G. (2000) High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2, *Nucleic Acids Res. 28*, e88

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J. Mol. Biol. 215*, 403 - 410.

Antikainen, N.M., Hergenrother, P.J., Harris, M.M., Corbett, W. and Martin, S.F. (2003) Altering substrate specificity of phosphatidylcholine-preferring phospholipase C of Bacillus cereus by random mutagenesis of the headgroup binding site, *Biochemistry 42*, 1603 - 1610.

Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Jr., Stoddard, B.L. and Baker, D. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity, *Nature 441*, 656 - 659.

Barfoed, H.C. (1987) Insulin Production Technology, *Chem. Eng. Prog. 83*, 49 - 54.

Barrett, A.J., Rawlings, N. and Woessner, JF. (2003) The Handbook of Proteolytic Enzymes, *2nd ed. Academic Press.*

Beckman, R.A., Mildvan, A.S. and Loeb, L.A. (1985) On the fidelity of DNA replication: manganese mutagenesis in vitro, *Biochemistry 24*, 5810 - 5817.

Bennet, B. and Cole, G. (2003) Pharmaceutical Production: An Engineering Guide, *The Institution of Chemical Engineers.*

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Res. 28*, 235 - 242.

Bode, W. and Huber, R. (1978) Crystal structure analysis and refinement of two variants of trigonal trypsinogen: trigonal trypsin and PEG (polyethylene glycol) trypsinogen and their comparison with orthorhombic trypsin and trigonal trypsinogen, *FEBS Lett. 90*, 265 - 269.

Bode, W. and Schwager, P. (1975) The refined crystal structure of bovine beta-trypsin at 1.8 A resolution. II. Crystallographic refinement, calcium binding site, benzamidine binding site and active site at pH 7.0, *J. Mol. Biol. 98*, 693 - 717.

Boychyn, M., Yim, S.S., Bulmer, M., More, J., Bracewell, D.G. and Hoare, M. (2004) Performance prediction of industrial centrifuges using scale-down models, *Bioprocess. Biosyst. Eng 26*, 385 - 391.

Broo, K., Larsson, A.K., Jemth, P. and Mannervik, B. (2002) An ensemble of theta class glutathione transferases with novel catalytic properties generated by stochastic recombination of fragments of two mammalian enzymes, *J. Mol. Biol. 318*, 59 - 70.

Buswell, A.M., Ebtinger, M., Vertes, A.A. and Middelberg, A.P. (2002) Effect of operating variables on the yield of recombinant trypsinogen for a pulse-fed dilution-refolding reactor, *Biotechnol. Bioeng. 77*, 435 - 444.

Cadwell, R.C. and Joyce, G.F. (1995) Mutagenic PCR. In: Dieffenbach C.W., Dveksler G.S. (eds.) PCR primer: a laboratory manual, *Cold Spring Harbor Laboratory Press.*

Carter, P. and Wells, J.A. (1988) Dissecting the catalytic triad of a serine protease, *Nature 332*, 564 - 568.

Chen, K. and Arnold, F.H. (1993) Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide, *Proc. Natl. Acad. Sci. U. S. A. 90*, 5618 - 5622.

Cheon, Y.H., Park, H.S., Kim, J.H., Kim, Y. and Kim, H.S. (2004) Manipulation of the active site loops of D-hydantoinase, a (beta/alpha)8-barrel protein, for modulation of the substrate specificity, *Biochemistry 43*, 7413 - 7420.

Chevalier, B.S., Kortemme, T., Chadsey, M.S., Baker, D., Monnat, R.J. and Stoddard, B.L. (2002) Design, activity, and structure of a highly specific artificial endonuclease, *Mol. Cell 10*, 895 - 905.

Chica, R.A., Doucet, N. and Pelletier, J.N. (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design, *Curr. Opin. Biotechnol. 16*, 378 - 384.

Chockalingam, K., Chen, Z., Katzenellenbogen, J.A. and Zhao, H. (2005) Directed evolution of specific receptor-ligand pairs for use in the creation of gene switches, *Proc. Natl. Acad. Sci. U. S. A 102*, 5691 - 5696.

Cochran, J.R., Kim, Y.S., Lippow, S.M., Rao, B. and Wittrup, K.D. (2006) Improved mutants from directed evolution are biased to orthologous substitutions, *Protein Eng. Des. Sel. 19*, 245 - 253.

Coco, W.M., Levinson, W.E., Crist, M.J., Hektor, H.J., Darzins, A., Pienkos, P.T., Squires, C.H. and Monticello, D.J. (2001) DNA shuffling method for generating highly recombined genes and evolved enzymes, *Nat. Biotechnol. 19*, 354 - 359.

Cohen, H.M., Tawfik, D.S. and Griffiths, A.D. (2004) Altering the sequence specificity of HaeIII methyltransferase by directed evolution using in vitro compartmentalization, *Protein Eng. Des. Sel. 17*, 3 - 11.

Copeland, R.A. (2000) Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis, *John Wiley & Sons (2nd Edition).*

Craik, C.S., Gardell, S., Roczniak, S., Fletterick, R. and Rutter, W.J. (1985a) Catalytic and Substrate-Specificity Studies of Trypsin and Carboxypeptidase A Via Site Specific Mutagenesis, *Abstracts of Papers of the American Chemical Society 190,* 100 - BIL.

Craik, C.S., Largman, C., Fletcher, T., Roczniak, S., Barr, P.J., Fletterick, R. and Rutter, W.J. (1985b) Redesigning trypsin: alteration of substrate specificity, *Science 228,* 291 - 297.

Crameri, A., Dawes, G., Rodriguez E Jr, Silver, S. and Stemmer, W.P. (1997) Molecular evolution of an arsenate detoxification pathway by DNA shuffling, *Nat. Biotechnol. 15,* 436 - 438.

Crameri, A., Raillard, S.A., Bermudez, E. and Stemmer, W.P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution, *Nature 391,* 288 - 291.

Cregg, J.M., Vedvick, T.S. and Raschke, W.C. (1993) Recent advances in the expression of foreign genes in *Pichia pastoris, Biotechnology (N. Y.) 11,* 905 - 910.

Cunningham, L.W.J. (1954) Molecular-kinetic properties of crystalline diisopropyl phosphoryl trypsin, *J. Biol. Chem. 211,* 13 - 19.

Dalby, P.A. (2003) Optimising enzyme function by directed evolution, *Curr. Opin. Struct. Biol. 13,* 500 - 505.

Delagrave, S., Murphy, D.J., Pruss, J.L., Maffia, A.M., III, Marrs, B.L., Bylina, E.J., Coleman, W.J., Grek, C.L., Dilworth, M.R., Yang, M.M. and Youvan, D.C. (2001) Application of a very high-throughput digital imaging screen to evolve the enzyme galactose oxidase, *Protein Eng 14,* 261 - 267.

Delano, W.L. (2002) The PyMOL Molecular Graphics System, URL - http://www.pymol.org [Accessed 01st Jun 2005].

Doi, N., Kumadaki, S., Oishi, Y., Matsumura, N. and Yanagawa, H. (2004) In vitro selection of restriction endonucleases by in vitro compartmentalization, *Nucleic Acids Res. 32,* e95

Doyle, S.A., Fung, S.Y. and Koshland, D.E., Jr. (2000) Redesigning the substrate specificity of an enzyme: isocitrate dehydrogenase, *Biochemistry 39,* 14348 - 14355.

Dwyer, M.A., Looger, L.L. and Hellinga, H.W. (2004) Computational design of a biologically active enzyme, *Science 304,* 1967 - 1971.

Eijsink, V.G., Bjork, A., Gaseidnes, S., Sirevag, R., Synstad, B., van den, B.B. and Vriend, G. (2004) Rational engineering of enzyme stability, *J. Biotechnol. 113*, 105 - 120.

Eijsink, V.G., Gaseidnes, S., Borchert, T.V. and van den, B.B. (2005) Directed evolution of enzyme stability, *Biomol. Eng. 22*, 21 - 30.

Evnin, L.B. and Craik, C.S. (1988) Development of an efficient method for generating and screening active trypsin and trypsin variants, *Ann. N. Y. Acad. Sci. 542*, 61 - 74.

Evnin, L.B., Vasquez, J.R. and Craik, C.S. (1990) Substrate specificity of trypsin investigated by using a genetic selection, *Proc. Natl. Acad. Sci. U. S. A 87*, 6659 - 6663.

Farid, S.S., Washbrook, J. and Titchener-Hooker, N.J. (2007) Modelling biopharmaceutical manufacture: Design and implementation of SimBiopharma, *Comput. Chem. Eng. 31*, 1141 - 1158.

Ferreira-Torres, C., Micheletti, M. and Lye, G.J. (2005) Microscale process evaluation of recombinant biocatalyst libraries: application to Baeyer-Villiger monooxygenase catalysed lactone synthesis, *Bioprocess. Biosyst. Eng 28*, 83 - 93.

Fersht, A. (2002) Structure and Mechanism in Protein Science, *W. H. Freeman & Co Ltd (4th Edition).*

Fontana, A., Fassina, G., Vita, C., Dalzoppo, D., Zamai, M. and Zambonin, M. (1986) Correlation between sites of limited proteolysis and segmental mobility in thermolysin, *Biochemistry 25*, 1847 - 1851.

Freshney, R.I. (2005) Culture of Animal Cells: A Manual of Basic Technique, *5th ed. Academic Press*

Friard, O.P. (2005) AnnHyb: A tool for working with and managing nucleotide sequences in multiple formats, URL - http://www.bioinformatics.org/annhyb/ [Accessed 17th Aug 2005].

Fujii, R., Nakagawa, Y., Hiratake, J., Sogabe, A. and Sakata, K. (2005) Directed evolution of Pseudomonas aeruginosa lipase for improved amide-hydrolyzing activity, *Protein Eng. Des. Sel. 18*, 93 - 101.

Funke, A.S., Reetz, M.T., Otte, N., Thiel, W., Van Pouderoyen, G., Dijkstra, B.W., Jaeger, K.E. and Eggert, T. (2005) Directed Evolution of an Enantioselective Bacillus subtilis Lipase, *Biocat. Biotrans 21*, 67 - 73.

Gabor, E.M. and Janssen, D.B. (2004) Increasing the synthetic performance of penicillin acylase PAS2 by structure-inspired semi-random mutagenesis, *Protein Eng Des Sel 17*, 571 - 579.

Gaytan, P., Osuna, J. and Soberon, X. (2002) Novel ceftazidime-resistance beta-lactamases generated by a codon-based mutagenesis method and selection, *Nucleic Acids Res. 30*, e84

Ghadessy, F.J., Ong, J.L. and Holliger, P. (2001) Directed evolution of polymerase function by compartmentalized self-replication, *Proc. Natl. Acad. Sci. U. S. A 98*, 4552 - 4557.

Goud, G.N., Artsaenko, O., Bols, M. and Sierks, M. (2001) Specific glycosidase activity isolated from a random phage display antibody library, *Biotechnol. Prog. 17*, 197 - 202.

Graf, L., Jancso, A., Szilagyi, L., Hegyi, G., Pinter, K., Naray-Szabo, G., Hepp, J., Medzihradszky, K. and Rutter, W.J. (1988) Electrostatic complementarity within the substrate-binding pocket of trypsin, *Proc. Natl. Acad. Sci. U. S. A. 85*, 4961 - 4965.

Greener, A., Callahan, M. and Jerpseth, B. (1996) An efficient random mutagenesis technique using an *E. coli* mutator strain, *Methods Mol. Biol. 57*, 375 - 385.

Griffiths, J.S., Cheriyan, M., Corbell, J.B., Pocivavsek, L., Fierke, C.A. and Toone, E.J. (2004) A bacterial selection for the directed evolution of pyruvate aldolases, *Bioorg. Med. Chem. 12*, 4067 - 4074.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis, URL - http://www.mbio.ncsu.edu/BioEdit/bioedit.html [Accessed 01st Nov 2003].

Hanquier, J., Sorlet, Y., Desplancq, D., Baroche, L., Ebtinger, M., Lefevre, J.F., Pattus, F., Hershberger, C.L. and Vertes, A.A. (2003) A single mutation in the activation site of bovine trypsinogen enhances its accumulation in the fermentation broth of the yeast *Pichia pastoris, Appl. Environ. Microbiol. 69*, 1108 - 1113.

Harris, J.L., Backes, B.J., Leonetti, F., Mahrus, S., Ellman, J.A. and Craik, C.S. (2000) Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries, *Proc. Natl. Acad. Sci. U. S. A. 97*, 7754 - 7759.

Harrison, R.G., Petrides, D., Todd, P.W. and Rudge, S.R. (2003) Bioseparations Science and Engineering, *Oxford University Press.*

Hayes, R.J., Bentzien, J., Ary, M.L., Hwang, M.Y., Jacinto, J.M., Vielmetter, J., Kundu, A. and Dahiyat, B.I. (2002) Combining computational and experimental screening for rapid optimization of protein properties, *Proc. Natl. Acad. Sci. U. S. A 99*, 15926 - 15931.

Hedstrom, L., Szilagyi, L. and Rutter, W.J. (1992) Converting trypsin to chymotrypsin: the role of surface loops, *Science 255*, 1249 - 1253.

Hengen, P. (1995) Purification of His-Tag fusion proteins from *Escherichia coli*, *Trends Biochem. Sci. 20*, 285 - 286.

Hibbert, E.G. (2003) Engineering and molecular biology approaches to improving trypsin-based bioprocesses, PhD thesis - University of London.

Hibbert, E.G., Senussi, T., Costelloe, S.J., Lei, W., Smith, M.E., Ward, J.M., Hailes, H.C. and Dalby, P.A. (2007) Directed evolution of transketolase activity on non-phosphorylated substrates, *J. Biotechnol. 131*, 425 - 432.

Hibbert, E.G., Senussi, T., Smith, M.E., Costelloe, S.J., Ward, J.M., Hailes, H.C. and Dalby, P.A. (2008) Directed evolution of transketolase substrate specificity towards an aliphatic aldehyde, *J. Biotechnol. 134*, 240 - 245.

Hixson, H.F. and Nishikawa, A.H. (1974) Affinity chromatography of bovine trypsin and thrombin, *Methods Enzymol. 34*, 440 - 448.

Host, G., Martensson, L.G. and Jonsson, B.H. (2006) Redesign of human carbonic anhydrase II for increased esterase activity and specificity towards esters with long acyl chains, *Biochim. Biophys. Acta 1764*, 1601 - 1606.

Ikawa, Y., Tsuda, K., Matsumura, S. and Inoue, T. (2004) De novo synthesis and development of an RNA enzyme, *Proc. Natl. Acad. Sci. U. S. A. 101*, 13750 - 13755.

Islam, R.S., Tisi, D., Levy, M.S. and Lye, G.J. (2007) Framework for the rapid optimization of soluble protein expression in *Escherichia coli* combining microscale experiments and statistical experimental design, *Biotechnol. Prog. 23*, 785 - 793.

Iyidogan, P. and Lutz, S. (2008) Systematic exploration of active site mutations on human deoxycytidine kinase substrate specificity, *Biochemistry 47*, 4711 - 4720.

Jennewein, S., Schurmann, M., Wolberg, M., Hilker, I., Luiten, R., Wubbolts, M. and Mink, D. (2006) Directed evolution of an industrial biocatalyst: 2-deoxy-D-ribose 5-phosphate aldolase, *Biotechnol. J. 1*, 537 - 548.

Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, C.F., III, Hilvert, D., Houk, K.N., Stoddard, B.L. and Baker, D. (2008) De novo computational design of retro-aldol enzymes, *Science 319*, 1387 - 1391.

Johnson, I.S. (1983) Human insulin from recombinant DNA technology, *Science 219*, 632 - 637.

Jourden, M.J., Clarke, C.N., Palmer, A.K., Barth, E.J., Prada, R.C., Hale, R.N., Fraga, D., Snider, M.J. and Edmiston, P.L. (2007) Changing the substrate specificity of creatine kinase from creatine to glycocyamine: evidence for a highly evolved active site, *Biochim. Biophys. Acta 1774*, 1519 - 1527.

Juillerat, A., Gronemeyer, T., Keppler, A., Gendreizig, S., Pick, H., Vogel, H. and Johnsson, K. (2003) Directed evolution of O6-alkylguanine-DNA alkyltransferase for efficient labeling of fusion proteins with small molecules in vivo, *Chem. Biol. 10*, 313 - 317.

Kast, P. and Hilvert, D. (1997) 3D structural information as a guide to protein engineering using genetic selection, *Curr. Opin. Struct. Biol. 7*, 470 - 479.

Keil-Dlouha, V., Zylber, N., Imhoff, J., Tong, N. and Keil, B. (1971a) Proteolytic activity of pseudotrypsin, *FEBS Lett. 16*, 291 - 295.

Keil-Dlouha, V., Zylber, N., Tong, N. and Keil, B. (1971b) Cleavage of glucagon by alpha- and beta-trypsin, *FEBS Lett. 16*, 287 - 290.

Kemmler, W., Peterson, J.D. and Steiner, D.F. (1971) Studies on the conversion of proinsulin to insulin. I. Conversion in vitro with trypsin and carboxypeptidase B, *J. Biol. Chem. 246*, 6786 - 6791.

Kristan, K., Stojan, J., Adamski, J. and Lanisnik, R.T. (2007) Rational design of novel mutants of fungal 17beta-hydroxysteroid dehydrogenase, *J. Biotechnol. 129*, 123 - 130.

Kumar, S. and Hein, G.E. (1970) Concerning the mechanism of autolysis of alpha-chymotyrpsin, *Biochemistry 9*, 291 - 297.

Latha, B., Ramakrishnan, K.M., Jayaraman, V. and Babu, M. (1997) Action of trypsin:chymotrypsin (Chymoral forte DS) preparation on acute-phase proteins following burn injury in humans, *Burns 23 Suppl 1*, S3 - S7.

Leemhuis, H., Rozeboom, H.J., Wilbrink, M., Euverink, G.J., Dijkstra, B.W. and Dijkhuizen, L. (2003) Conversion of cyclodextrin glycosyltransferase into a starch hydrolase by directed evolution: the role of alanine 230 in acceptor subsite +1, *Biochemistry 42*, 7518 - 7526.

Lutz, S., Ostermeier, M. and Benkovic, S.J. (2001a) Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides, *Nucleic Acids Res. 29*, E16

Lutz, S., Ostermeier, M., Moore, G.L., Maranas, C.D. and Benkovic, S.J. (2001b) Creating multiple-crossover DNA libraries independent of sequence identity, *Proc. Natl. Acad. Sci. U. S. A. 98*, 11248 - 11253.

MacBeath, G., Kast, P. and Hilvert, D. (1998) Probing enzyme quaternary structure by combinatorial mutagenesis and selection, *Protein Sci. 7*, 1757 - 1767.

Maroux, S., Baratti, J. and Desnuelle, P. (1971) Purification and specificity of porcine enterokinase, *J. Biol. Chem. 246*, 5031 - 5039.

Maroux, S. and Desnuelle, P. (1969) On some autolyzed derivatives of bovine trypsin, *Biochim. Biophys. Acta 181*, 59 - 72.

Maroux, S., Rovery, M. and Desnuelle, P. (1967) An autolyzed and still active form of bovine trypsin, *Biochim. Biophys. Acta 140*, 377 - 380.

Martoglio, B. and Dobberstein, B. (1998) Signal sequences: more than just greasy peptides, *Trends Cell Biol. 8*, 410 - 415.

Miller, O.J. (2004) Directed evolution of transketolase, a carbon-carbon bond forming enzyme, PhD thesis - University of London.

Miller, O.J., Bernath, K., Agresti, J.J., Amitai, G., Kelly, B.T., Mastrobattista, E., Taly, V., Magdassi, S., Tawfik, D.S. and Griffiths, A.D. (2006) Directed evolution by in vitro compartmentalization, *Nat. Methods 3*, 561 - 570.

Miozzari, G.F. and Yanofsky, C. (1978) Translation of the leader region of the Escherichia coli tryptophan operon, *J. Bacteriol. 133*, 1457 - 1466.

Miyazaki, K. and Arnold, F.H. (1999) Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function, *J. Mol. Evol. 49*, 716 - 720.

Miyazaki, K. and Takenouchi, M. (2002) Creating random mutagenesis libraries using megaprimer PCR of whole plasmid, *Biotechniques 33*, 1033 - 1038.

Morihara, K., Oka, T. and Tsuzuki, H. (1979) Semi-synthesis of human insulin by trypsin-catalysed replacement of Ala-B30 by Thr in porcine insulin, *Nature 280*, 412 - 413.

Morley, K.L. and Kazlauskas, R.J. (2005) Improving enzyme properties: when are closer mutations better?, *Trends Biotechnol. 23*, 231 - 237.

Mueller-Cajar, O., Morell, M. and Whitney, S.M. (2007) Directed evolution of rubisco in Escherichia coli reveals a specificity-determining hydrogen bond in the form II enzyme, *Biochemistry 46*, 14067 - 14074.

Mullegger, J., Jahn, M., Chen, H.M., Warren, R.A. and Withers, S.G. (2005) Engineering of a thioglycoligase: randomized mutagenesis of the acid-base residue leads to the identification of improved catalysts, *Protein Eng. Des. Sel. 18*, 33 - 40.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction, *Cold Spring Harb. Symp. Quant. Biol. 51 Pt 1*, 263 - 273.

Ni, J., Sasaki, Y., Tokuyama, S., Sogabe, A. and Tahara, Y. (2002) Conversion of a typical catalase from Bacillus sp. TE124 to a catalase-peroxidase by directed evolution, *J. Biosci. Bioeng. 93*, 31 - 36.

O'Maille, P.E., Bakhtina, M. and Tsai, M.D. (2002) Structure-based combinatorial protein engineering (SCOPE), *J. Mol. Biol. 321*, 677 - 691.

Olsen, J.V., Ong, S.E. and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues, *Mol. Cell Proteomics. 3*, 608 - 614.

Otten, L.G., Sio, C.F., Vrielink, J., Cool, R.H. and Quax, W.J. (2002) Altering the substrate specificity of cephalosporin acylase by directed evolution of the Beta -subunit, *J. Biol. Chem. 277*, 42121 - 42127.

Parikh, M.R. and Matsumura, I. (2005) Site-saturation mutagenesis is more efficient than DNA shuffling for the directed evolution of beta-fucosidase from beta-galactosidase, *J. Mol. Biol. 352*, 621 - 628.

Park, S., Morley, K.L., Horsman, G.P., Holmquist, M., Hult, K. and Kazlauskas, R.J. (2005) Focusing mutations into the P. fluorescens esterase binding site increases enantioselectivity more effectively than distant mutations, *Chem. Biol. 12*, 45 - 54.

Peimbert, M. and Segovia, L. (2003) Evolutionary engineering of a beta-Lactamase activity on a D-Ala D-Ala transpeptidase fold, *Protein Eng 16*, 27 - 35.

Perona, J.J. and Craik, C.S. (1995) Structural basis of substrate specificity in the serine proteases, *Protein Sci. 4*, 337 - 360.

Pomp, D. and Medrano, J.F. (1991) Organic solvents as facilitators of polymerase chain reaction, *Biotechniques 10*, 58 - 59.

Rao, M.B., Tanksale, A.M., Ghatge, M.S. and Deshpande, V.V. (1998) Molecular and biotechnological aspects of microbial proteases, *Microbiol. Mol. Biol. Rev. 62*, 597 - 635.

RaviKumar, T., Ramakrishnan, M., Jayaraman, V. and Babu, M. (2001) Effect of trypsin-chymotrypsin (Chymoral Forte D.S.) preparation on the modulation of cytokine levels in burn patients, *Burns 27*, 709 - 716.

Rawlings, N.D. and Barrett, A.J. (1993) Evolutionary families of peptidases, *Biochem. J. 290 (Pt 1)*, 205 - 218.

Reetz, M.T., Bocola, M., Carballeira, J.D., Zha, D. and Vogel, A. (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test, *Angew. Chem. Int. Ed Engl. 44*, 4192 - 4196.

Reetz, M.T. and Jaeger, K.E. (1999) Superior biocatalysts by directed evolution, *Top. Curr. Chem. 200*, 31 - 57.

Reidhaar-Olson, J.F. and Sauer, R.T. (1988) Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences, *Science 241*, 53 - 57.

Rost, B. (1999) Twilight zone of protein sequence alignments, *Protein Eng 12*, 85 - 94.

Rypniewski, W.R., Perrakis, A., Vorgias, C.E. and Wilson, K.S. (1994) Evolutionary divergence and conservation of trypsin, *Protein Eng. 7*, 57 - 64.

Sacchi, S., Rosini, E., Molla, G., Pilone, M.S. and Pollegioni, L. (2004) Modulating D-amino acid oxidase substrate specificity: production of an enzyme for analytical determination of all D-amino acids by directed evolution, *Protein Eng. Des. Sel. 17*, 517 - 525.

Salte, H., King, J.M., Baganz, F., Hoare, M. and Titchener-Hooker, N.J. (2006) A methodology for centrifuge selection for the separation of high solids density cell broths by visualisation of performance using windows of operation, *Biotechnol. Bioeng. 95*, 1218 - 1227.

Sambrook, J.E., Fristsch, F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.*

Santoro, S.W. and Schultz, P.G. (2002) Directed evolution of the site specificity of Cre recombinase, *Proc. Natl. Acad. Sci. U. S. A. 99*, 4185 - 4190.

Schechter, I. and Berger, A. (1967) On the size of the active site in proteases. I. Papain, *Biochem. Biophys. Res. Commun. 27*, 157 - 162.

Schmitzer, A.R., Lepine, F. and Pelletier, J.N. (2004) Combinatorial exploration of the catalytic site of a drug- resistant dihydrofolate reductase: creating alternative functional configurations, *Protein Eng. Des. Sel. 17*, 809 - 819.

Schroeder, D.D. and Shaw, E. (1968) Chromatography of trypsin and its derivatives. Characterization of a new active form of bovine trypsin, *J. Biol. Chem. 243*, 2943 - 2949.

Schroeder, W.A., Shelton, J.B. and Shelton, J.R. (1969) An examination of conditions for the cleavage of polypeptide chains with cyanogen bromide: application to catalase, *Arch. Biochem. Biophys. 130*, 551 - 556.

Shao, Z., Zhao, H., Giver, L. and Arnold, F.H. (1998) Random-priming in vitro recombination: an effective tool for directed evolution, *Nucleic Acids Res. 26*, 681 - 683.

Shinkai, A., Patel, P.H. and Loeb, L.A. (2001) The conserved active site motif a of Escherichia coli DNA polymerase I is highly mutable, *J. Biol. Chem. 276*, 18836 - 18842.

Sidhu, S.S., Lowman, H.B., Cunningham, B.C. and Wells, J.A. (2000) Phage display for selection of novel binding peptides, *Methods Enzymol. 328*, 333 - 363.

Sieber, V., Martinez, C.A. and Arnold, F.H. (2001) Libraries of hybrid proteins from distantly related sequences, *Nat. Biotechnol. 19*, 456 - 460.

Sio, C.F., Riemens, A.M., van der Laan, J.M., Verhaert, R.M. and Quax, W.J. (2002) Directed evolution of a glutaryl acylase into an adipyl acylase, *Eur. J. Biochem. 269*, 4495 - 4504.

Sipos, T. and Merkel, J.R. (1970) An effect of calcium ions on the activity, heat stability, and structure of trypsin, *Biochemistry 9*, 2766 - 2775.

Smith, R.L. and Shaw, E. (1969) Pseudotrypsin - A Modified Bovine Trypsin Produced by Limited Autodigestion, *J. Biol. Chem. 244*, 4704 - 4712.

Stemmer, W.P. (1994) DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution, *Proc. Natl. Acad. Sci. U. S. A 91*, 10747 - 10751.

Stewart, J.A. and Dobson, J.E. (1965) Trypsin-catalyzed hydrolysis of N-benzoly-L-arginine ethyl ester at low pH, *Biochemistry 4*, 1086 - 1091.

Stratagene Ltd. (2003) QuikChange® Site-Directed Mutagenesis Kit, catalogue number - 200519.

Strausberg, S.L., Ruan, B., Fisher, K.E., Alexander, P.A. and Bryan, P.N. (2005) Directed coevolution of stability and catalytic activity in calcium-free subtilisin, *Biochemistry 44*, 3272 - 3279.

Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. (1990) Use of T7 RNA polymerase to direct expression of cloned genes, *Methods Enzymol. 185*, 60 - 89.

Suenaga, H., Watanabe, T., Sato, M., Ngadiman and Furukawa, K. (2002) Alteration of regiospecificity in biphenyl dioxygenase by active-site engineering, *J. Bacteriol. 184*, 3682 - 3688.

Svendsen, A. (2003) Enzyme Functionality: Design, Engineering and Screening, *Marcel Dekker Ltd*

Tawfik, D.S. and Griffiths, A.D. (1998) Man-made cell-like compartments for molecular evolution, *Nat. Biotechnol. 16*, 652 - 656.

Tindall, K.R. and Kunkel, T.A. (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase, *Biochemistry 27*, 6008 - 6013.

Ting, A.Y., Witte, K., Shah, K., Kraybill, B., Shokat, K.M. and Schultz, P.G. (2001) Phage-display evolution of tyrosine kinases with altered nucleotide specificity, *Biopolymers 60*, 220 - 228.

Turner, J.M., Graziano, J., Spraggon, G. and Schultz, P.G. (2006) Structural plasticity of an aminoacyl-tRNA synthetase active site, *Proc. Natl. Acad. Sci. U. S. A 103*, 6483 - 6488.

Varallyay, E., Pal, G., Patthy, A., Szilagyi, L. and Graf, L. (1998) Two mutations in rat trypsin confer resistance against autolysis, *Biochem. Biophys. Res. Commun. 243*, 56 - 60.

Vasquez, J.R., Evnin, L.B., Higaki, J.N. and Craik, C.S. (1989) An expression system for trypsin, *J. Cell Biochem. 39*, 265 - 276.

Voet, D., Voet, G. and Pratt, C.W. (2004) Fundamentals of Biochemistry, *John Wiley & Sons*

Wada, M., Hsu, C.C., Franke, D., Mitchell, M., Heine, A., Wilson, I. and Wong, C.H. (2003) Directed evolution of N-acetylneuraminic acid aldolase to catalyze enantiomeric aldol reactions, *Bioorg. Med. Chem. 11*, 2091 - 2098.

Wan, W.Y. and Milner-White, E.J. (1999) A natural grouping of motifs with an aspartate or asparagine residue forming two hydrogen bonds to residues ahead in sequence: their occurrence at alpha-helical N termini and in other situations, *J. Mol. Biol. 286*, 1633 - 1649.

Wang, E.C., Hung, S.H., Cahoon, M. and Hedstrom, L. (1997) The role of the Cys191-Cys220 disulfide bond in trypsin: new targets for engineering substrate specificity, *Protein Eng 10*, 405 - 411.

Wang, L., Brock, A., Herberich, B. and Schultz, P.G. (2001) Expanding the genetic code of *Escherichia coli*, *Science 292*, 498 - 500.

Wang, L., Brock, A. and Schultz, P.G. (2002a) Adding L-3-(2-Naphthyl)alanine to the genetic code of E. coli, *J. Am. Chem. Soc. 124*, 1836 - 1837.

Wang, L., Zhang, Z., Brock, A. and Schultz, P.G. (2003) Addition of the keto functional group to the genetic code of Escherichia coli, *Proc. Natl. Acad. Sci. U. S. A 100*, 56 - 61.