# Engineering & molecular biology approaches to improving trypsin-based bioprocesses

A thesis submitted to the University of London

for the degree of

DOCTOR OF PHILOSOPHY

by

**Edward Graham Hibbert**

**BSc, MSc**

September 2003

Advanced Centre for Biochemical Engineering

Department of Chemical and Biochemical Engineering

University College London

Torrington Place

London WC1E 7JE





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

## Abstract

This project was designed, in partnership with Eli Lilly, Fegersheim, with the aim of developing a series of methodologies for the mutation and characterisation of trypsin variants with extended substrate specificities for certain amino acid combinations. This library of modified enzymes with enhanced specificity towards 4-aa 1° sequence motifs could then be used as efficient & 'clean' biocatalytic agents. The goal was thus to explore the tailoring of industrial bioprocess enzymes to better suit process criteria, thus lowering the overall cost. It was a principle aim of the project to design a practical method of directed evolution for the breeding of variants with novel or enhanced substrate specificities.

The organism on which this work was to be based was the Eli Lilly commercial recombinant bovine trypsinogen production strain, termed ELTRP-1. In order to devise a high-throughput assay suitable for screening a mutant library of $10^{3}$+ variants, it was essential to fully characterise a method of microwell fermentation, and also to engineer the recombinant protein for solubility, i.e render it available for assaying. A high-throughput screen such as this would have to involve the minimum number of process steps and be entirely microplate-based.

A microplate-scale solubilisation & refolding protocol for the r-trypsinogen inclusion bodies was developed that was successful for commercial enzyme but was impractical for use on recombinant inclusion bodies as the nature of the screen disallowed extensive purification, and thus the highly heterologous nature of the solutions appeared to inhibit refolding. Two variants of the enzyme, one with the prosequence removed, were cloned into a pET26b vector, behind a pelB leader, and all available fermentation parameters were experimented with. The construct plasmid was then subjected to a range of mutation rates using the XL1-Red mutator strain and variants were screened for increases in solubility. A series of mutants were obtained which demonstrated over 100μg/ml soluble mature enzyme after microplate fermentation, and enzyme translocation was monitored over 5ml & 100ml scale-up. Observed plasmid instability was due, in part, to the autotoxicity of the expressed enzyme and compounded by the fact that the T7 induction control mechanism of the strain appeared to have been disabled. These problems only manifested when the fermentation was scaled-up above microplate volumes and thus it appeared that the screening of the library for soluble mutant enzymes had been entirely successful by its own specific criteria, i.e. the isolated variants performed as required only when in well-culture.

## Acknowledgements

I'd like to thank my supervisors, John Woodley and Paul Dalby, for their invaluable guidance before each experiment, patience during its progress and damage control afterwards. Special thanks must also go to John Ward and Frank Baganz, both of whom contributed new ideas and insights to the work.

I would also like to thank Alain Vertes, from Eli Lilly, Fergersheim, whose cooperation made the project possible. I am also very grateful for financial support provided by the BBSRC.

Special thanks must go also to Rohnan O'Kennedy and Dunstan Cooke for their advice on the practicalities of molecular biology, and to all the upstanding citizens of Collanades office and the Macromolecular lab for their assorted virtues.

Thanks also to my parents for their support and for not labouring the fact that they'd both written up within three years.

And finally, many thanks must go to the vre Pepi who, despite everything, is still waiting to have a mutant named after her...

# Contents

## List of figures

## List of tables

## List of equations

## Abbreviations

| | |
|---|---|
| AFGP | Antifreeze glycoprotein |
| Ala (also A) | Alanine |
| Arg (also R) | Arginine |
| Asn (also N) | Asparagine |
| Asp (also D) | Aspartic acid |
| B | Benzyl |
| BHI | Biosynthetic human insulin |
| Cys (also C) | Cysteine |
| EpPCR | Error-prone PCR |
| Gln (also Q) | Glutamine |
| Glu (also E) | Glutamic acid |
| His (also H) | Histidine |
| Ile (also I) | Isoleucine |
| ITCHY | Incremental truncation for the creation of hybrid enzymes |
| L-BAPNA | $N_{\alpha}$-benzoyl-L-ariginine-p-nitroanilide |
| Leu (also L) | Leucine |
| Lys (also K) | Lysine |
| Met (also M) | Methionine |
| P (1-4) | Upstream protease binding sites |
| P (1-4)' | Downstream protease binding sites |
| PCR | Polymerase chain reaction |
| Phe (also F) | Phenylalanine |
| Pro (also P) | Proline |
| RACHITT | Random chimeragenesis on transient templates |
| S (1-4) | Upstream substrate binding sites |
| S (1-4)' | Downstream substrate binding sites |
| SBZI | thiobenzylester |
| SCRATCHY | ITCHY plus DNA shuffling |
| Ser (also S) | Serine |

| | |
|---|---|
| SHIPREC | Sequence homology independent protein recombination |
| TAME | p-toluenesulphonyl-L-arginine methyl ester |
| THIO-ITCHY | ITCHY plus use of α-phosphothiate nucleotides |
| Thr (also T) | Threonine |
| Trp (also W) | Tryptophan |
| Tyr (also Y) | Tyrosine |
| Val (also V) | Valine |
| wt | Wild-type enzyme (unmutated) |
| Z | Benzyloxycarbonyl |

# 1 Introduction

## 1.1 Protein engineering for bioindustry

The unwitting use of enzymes as biocatalysts has been a key feature in chemical transformation reactions throughout history, underlying all 'traditional biotechnology' (e.g. baking and brewing), and now, applied with increasing finesse, forms the basis of many modern bioindustrial processes. These range from the bulk manufacture of commodity chemicals such as fructose or citric acid, to the synthesis of complex pharmaceutical intermediates (Crueger and Crueger, 1990). Unsurprisingly, as most industrial biotransformation reactions seek to mimic cellular reactions, albeit on a vast scale, many enzyme families have proved to be excellent agents for industrial catalysis, with reaction rate accelerations of typically $10^5 - 10^8$ (Schwienhorst, 2002; Zaks, 2001). The chiral nature of enzymes allows for extremely high degrees of regio- and stereospecificity when acting on a preferred substrates in native conditions and the majority of biocatalytic reaction procedures have been designed and built with the aim of exploiting these properties by supplying a favourable chemical environment (Crueger and Crueger, 1990). Problems frequently arise in that the conditions required for activity in industrial applications are often sufficiently dissimilar to the enzymes' native environment as to cause a loss of functionality, be it activity, stability or specificity. As metabolic tools which have evolved over billions of years to perform precise functions within a self-regulating environment, many enzymes exhibit traits that further complicate their industrial usefulness, such as product inhibition, co-factor requirements or appreciable rates of autohydrolysis, which maintain the metabolic feed-back loops so essential for cellular homeostasis, and which can severely limit industrial product yield (Bornscheuer and Pohl, 2001; Zaks, 2001).

As the number of targets and potential markets for modern biotechnology expands, so does the urgent need for new biocatalytic agents and for technologies that may rapidly assess both an enzyme's activity and the suitability for industrial work. The natural world is, of course, the ultimate source of protein diversity, with current estimates of the number of microbial species in existence (at any one time) standing between $1\times10^6$ and $1\times10^8$, with currently only approximately 5,000 species being described in the literature (Daniel, 2002). The construction and subsequent screening of vast environmental libraries remains the classic method for searching for an enzyme activity or property to satisfy a new industrial need, but poses considerable problems for modern companies, because of the length of time and expense required, difficulty in culturing diverse

species and, finally, because many natural enzymes display the physiochemical 'restrictions' described above. A further consideration is that the substrate specificity of an enzyme may limit its synthetic utility, examples being the relatively loose specificity of many digestive proteases (Furlong *et al.*, 2002) or insufficient levels of enantioselectivity (Henke and Bornscheuer, 1999). In theory, a sufficiently large environmental library could provide an enzyme approximately suited for any synthetic reaction, but in practice, the optimisation of industrial processes frequently require activity under thoroughly unnatural conditions or specificities of a type that may not be selectively favourable in the wild. The ability to overcome these limitations and construct tailor-made biocatalysts, fine-tuned for optimal functionality under process conditions, is becoming increasingly possible with the advent of modern molecular biology. Knowledge of the amino acid sequence of an enzyme and its tertiary / quaternary structure from crystallisation, allows engineers to explore the physical basis of activities of interest through controlled substitution of selected residues. The success of a strategy of rational design hinged on relating structure to function, and thus relied on molecular modelling techniques to identify targets. Such a strategy could be described as a 'top-down' approach, predicting the structural disturbances caused by single amino acid substitutions, and is thus highly-information intensive. Although recent advances in the generation and solving of protein structures (Bornscheuer and Pohl, 2001; Rubingh, 1997) and significant improvements in the power of molecular modelling in such areas as free energy permutation and molecular dynamics calculations (Rubingh, 1997), now allow for considerable modern finesse in the modification of proteins, the complexity of the structure-function relationship and the lack of detailed data on most enzyme families means that rational design is not yet considered a practical method for the improvement of biocatalysts.

In contrast to this, the recent techniques of directed evolution (also called molecular or *in vitro* evolution), make no assumptions as to the nature of the required structural changes, beyond possibly delineating the region exposed to mutagenesis, and involves the construction of synthetic libraries of diversity from wild-type enzymes (Bornscheuer and Pohl, 2001; Tobin et al, 2000). The power of the technique is directly proportional to the size of the library generated, typically $\geq 10^6$ mutant variants, with the diversity being generated through variations of random mutagenesis (Chen and Arnold, 1993) and homologous (Stemmer, 1994a) or non-homologous recombination (Lutz and Benkovic, 2000). The enzyme libraries are then characterised by the screening of each individual clone for the desired activity, the quantity of variants meaning that the libraries are only

of practical use if they can be coupled to a system of small-scale fermentation and high-throughput screening (Schwienhorst, 2002). Wherever possible, an 'intermediate' selection step would be introduced, coupling the desired activity of the mutant enzyme to cell viability, avoiding the need to screen functionally wild-type or disabled enzymes, and thus dramatically enhancing the size of the library which may be processed. This evolutionary methodology could thus be seen as a 'bottom-up' approach to protein modification, relying on a precisely designed screen working in conjunction with the high-throughput handling of clones to provide a protein with the required properties, irrespective of which mutations were required to achieve this. A crass analogy would be that this technology works to transform protein engineering from an irregular and unpredictable art, to a form which replaces the 'brains' of the process with 'muscle', allowing careful mutation and mass screening to replace precise, considered modification, and creating a scalable method of biocatalyst discovery suitable for industry, one where investment of resources linearly increased the chances of success.

The following sections summarise current evolutionary theory and detail all techniques of random mutagenesis characterised to date. Variations of the classic methods of *in vitro* mutagenesis are continually being developed as researchers seek to mimic different aspects of natural protein evolution, and thus each exhibits different bias and mutation rates, as well as varying complexities in experimental procedure (Section 1.4). Each method, and potential combinations of protocols, was thus considered in terms of practicality and published effectiveness. The enzyme which was to be the subject of this work was a variant of bovine trypsinogen and the reasons for its scientific and economic suitability as a target, as well as current methods for its production, with reference to high-throughput microwell fermentation and screening, are discussed in Section 1.5-1.8.

## 1.2 Concepts of direct evolution

### 1.2.1 Evolution of novel protein-coding genes in nature

Darwin attributed the forms of all living organisms on earth, from the base components upwards, to be products of blind design (Patthy, 1999). Incremental changes, accumulating over successive generations, allow proteins to perform slightly better at old tasks or under changing conditions. Environmental stress acted on populations as a whole to select the fittest variants, thus

encouraging any number of solutions to arise from within the various systems providing that they enhance the total final activity. It is the *in vitro* simulation of this evolutionary adaptivness that has lead to the family of methodologies now termed 'directed evolution'. Although there is a steady influx of data regarding the structural effects of sequence mutations on protein functionality (for recent examples, please see Table 1.1), by far the bulk of current knowledge regarding enzyme engineering comes from the continued study of natural mechanisms (Lutz and Benkovic, 2000). Understanding the principles and motives whereby new protein structures and functions emerge, can provide guidelines for molecular modification, and gives insight into mechanisms, the effectiveness of which have been refined for literally billions of years.

Ancestral sequence studies (where a phylogenetic tree was constructed based on regions of conserved sequence homology) have shown that nature employs two fundamental approaches in the generation of novel proteins. The first involves the *de novo*, or spontaneous creation of a protein-encoding gene through the rearrangement of repetitive simple structural elements. The emergence of such proteins is difficult to prove, with evidence being frequently contested, as only extremely tenuous ancestral paths can be constructed in such cases. Homology studies do appear to show that certain 'structural-signatures' have arisen independently several times during microbial evolution, with the commonest type being proteins which consist almost entirely of a single, repeated, secondary structure element (Patthy, 1999). Such repetitive sequences frequently form tightly coiled structures (e.g. fibrous proteins) or account for regions enriched for a certain motif (e.g. leucine-rich-repeat proteins), however the best characterised examples of *de novo* generation of a gene are the serum antifreeze glycoproteins (AFGPs) of Artic and Antarctic white fish. In both cases, these proteins, which act to preserve the fluidity of the serum at temperatures as low as –2°C, consist entirely of an iterative tripeptide unit, Thr-Ala-Ala. Although structurally identical, the two forms of AFGP differ in their origins, with those of Antarctic fish having arisen via recruitment of the 5' & 3' terminal ends of an ancestral trypsinogen gene (obtaining secretory signal and untranslated regions) and then by amplification of the repeated motif. The origin of AFGPs from artic fish is known to be separate but resists identification.

*De novo* evolution of enzymes, while an undoubted natural occurrence and the clear (long-term) goal of *in silico* modelling studies, does not currently represent a useful technique for laboratory-based enzyme engineering as the probability of creating a functional enzyme through the random association of common sequence elements is inversely proportional to its complexity.

Essentially *de novo* synthesis requires the screening of every conceivable combination of elements, and with the distribution of functionality in sequence space being so sparse, would require libraries of unimaginable sizes (Daniel, 2002). The most fruitful path for study and simulation is the second approach that nature uses for the evolution of novel proteins, namely the remodelling of existing protein scaffolds for altered functionality (Nagano *et al.*, 2002; Todd *et al.*, 2002). The most common natural methods for the alteration of a protein-encoding gene are listed in Figure 1-1. Each involves the exploration of only local sequence space, the point of origin being a functional protein, and together comprise an effective set of methods for the slight alteration of structural or catalytic properties, or for the accommodation of environmental changes.

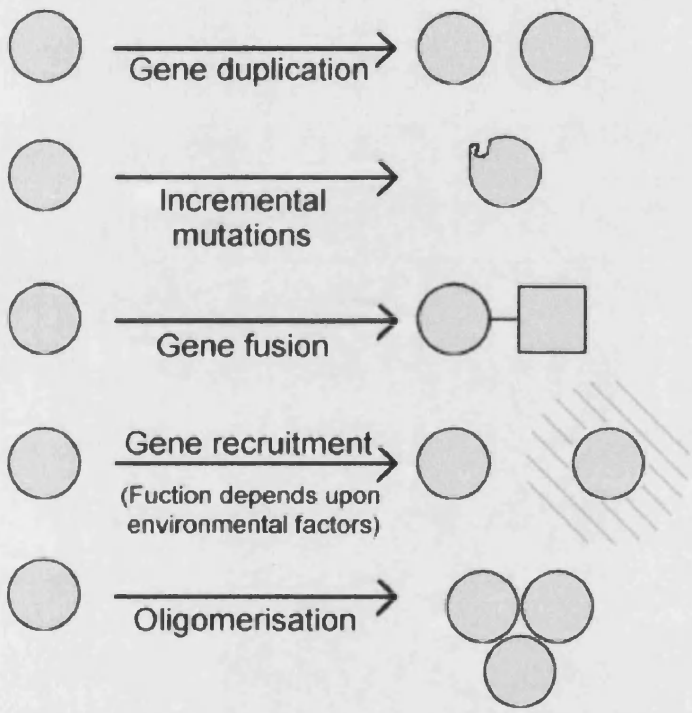



**Figure 1-1: Methods of altering protein functionality found in nature**
Directed evolution tends to refer to the technique of accumulating point mutations over successive generation, culminate changes leading to modification of the physical properties of an enzyme. Figure adapted from (Arnold, 1996; Lutz and Benkovic, 2000).

### 1.2.2 Examples and strategy

The individual methodologies of directed evolution are discussed in detail in subsequent sections (1.4.3 & 1.4.4), but an overview of a generic combined strategy for the mutation and screening of a protein-coding gene is presented in Figure 1-2. The principle underlying any whole protein system of directed evolution is that no assumptions are made as to regions of importance within the protein, essentially regarding the protein's sequence as a 'black box', and directing the bulk of the work into developing suitable environmental tests which would allow a variant with the desired fitness to emerge. Error-prone PCR, a low fidelity version of standard PCR that imposes an average mutation rate on the sequence being amplified, would commonly be used to ensure that protein mutants were generated at an accelerated rate and thus to produce a homologous gene library with a uniform, if low, level of diversity. DNA shuffling (also known as molecular recombination or breeding) is a recently developed technique designed to mimic the low level of *in vivo* gene shuffling which is believed to exist universally in biological systems (Minshull and Stemmer, 1999). *In vitro* recombination involves the random fragmentation of a library of gene and their reassembly by recombination with each other by 'primer-free' polymerase chain reaction (PCR) reaction. DNA shuffling acts as a complementary procedure to rounds of random mutagenesis and screening, allowing the concentration of beneficial mutations that have accrued on individual genes, while simultaneously providing a method by which mutations which impair protein function, and also 'neutral' amino acid mutations, can be identified and discarded from the final molecule. Neutral mutations are non-synonymous, i.e. result in amino acid substitutions (see Section 1.3.1), and can pose particular difficulty as they obscure functionally beneficial or co-operative mutations, and thus can require increased amounts of subsequent validation before the results can be accepted (validation would occur by targeted substitution of all mutated amino acids, in every possible combination, screening for improvements or interaction effects). The process of DNA shuffling may also be used to add further diversity to a library by forcing the PCR based reassembly to proceed under low fidelity conditions, but this is generally considered unnecessary in a combined strategy, and analysis of results can be performed far quicker if epPCR and recombination are designed as separate mechanisms. Table 1.1 gives a list of recent examples, where proteins have had their functionality improved towards diverse targets (usually for use in an industrial context) by means of varied methods of directed evolution.

A total strategy for improving protein functionality would most likely enlist both *in vitro* evolution and rational engineering of the targeted protein, using rational design to initially alter residues known to be important from previous studies, or using the information gained from the protein variants generated by directed evolution to adjust the primary sequence.

| Enzyme | Evolved property | Mutagenisis method | Result | Reference |
|---|---|---|---|---|
| Kanamycin nucleotidyltransferase | Thermostability | Mutator strain | >200-fold incr. in half-life @ 60-65°C | (Liao et al 1986) |
| Esterase | Stereoselectivity | Mutator strain | 25-fold incr. in stereoselectivity | (Bornscheuer et al 1998) |
| Subtilisin E | Activity in organic solvents | epPCR | 170-fold incr. for activity in –60%DMF | (Chen and Arnold 1993) |
| Subtilisin E | Thermostability | epPCR & shuffling | 500-fold incr. in half-life @ 60°C | (Miyazaki et al 2000) |
| Glutaryl acylase | Substrate specificity | epPCR | 5-fold incr. in activity on adipyl-chains | (Sio et al 2002) |
| Beta-Galactosidase | Substrate specificity | DNA shuffling | Creation of an efficient β-fucosidase | (Zhang et al 1997) |
| Paranitrobenzyl esterase | Substrate specificity and activity in organic solvents | epPCR and DNA shuffling | Up to 150-fold incr. in activity depending on %DMF | (Moore and Arnold 1996) |
| Green fluorescent protein | Increased activity | epPCR and DNA-shuffling | Up to a 42-fold improvement | (Crameri et al 1996) |

**Table 1.1: Examples of protein functionality altered by direct evolution**

Published directed-enzyme-evolution experiments that couple sequential generations of random mutagenesis and/or recombination with an appropriate selection or screening strategy.

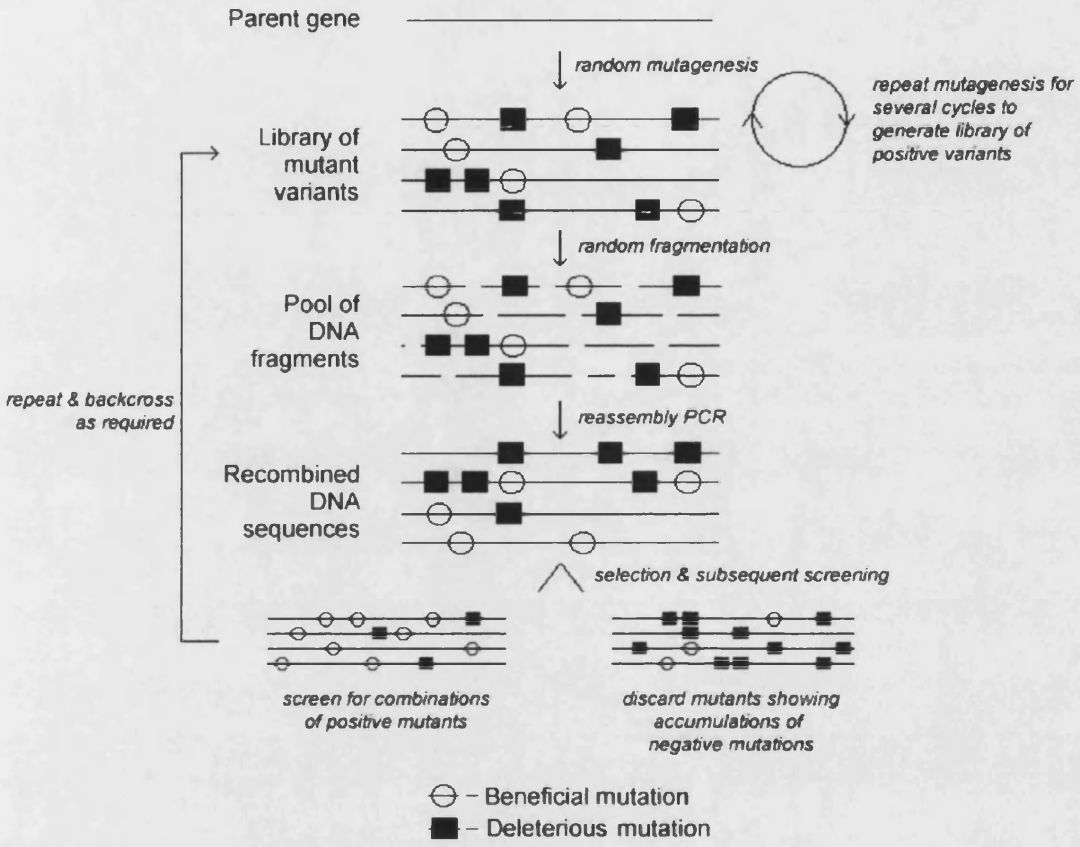



**Figure 1-2: Combining epPCR & DNA shuffling**

Illustrating the principle of a strategy of directed evolution, which combines random mutagenesis by error-prone PCR and DNA shuffling to generate and concentrate positive mutations. Adapted from (Kuchner and Arnold 1997). The techniques shown are discussed in detail in the following sections:

Random mutagenesis in section 1.4

Mutagenic PCR in section 1.4.3

DNA shuffling in section 1.4.4

## 1.3 Mutation: concepts and terminology

### 1.3.1 Mutation types

Almost all discussion of mutations in this work will refer to replication-dependant mutations, i.e. those arising through an error in the process of DNA replication. Although the methods used to generate a fixed rate of mutation may give rise to any type of error, only point substitution effects are desired. These are simply where a single nucleotide is replaced by another and, importantly, results in no change in the reading frame of the sequence. Other types of mutations, such as additions and deletions, which result in the insertion or subtraction of one or more nucleotides from a given point in a sequence, would inevitably alter the reading frame (Patthy, 1999). A frameshift mutation results in a change in all downstream amino acids, changing the nature of the protein and also the length as the position of the stop codon would now be randomised. This almost certainly leads to the translation of a truncated and crippled protein, i.e. non-functional and randomly folded.

Regarding single point substitutions, there are two types which may occur, with the frequency of each being determined by the bias of the *in vitro* or *in vivo* mechanism. The first type is a transition event where a purine (adenine or guanine) is replaced by another purine or, similarly, a pyrimidine (cytosine or thymine) is replaced by a pyrimidine. Transversion events may also be seen, were a purine is replaced by a pyrimidine or visa versa. Error-prone PCR is known to promote transitions over transversions (Kuchner and Arnold, 1997) and this is one example of how mutational bias, inherent to some degree in all 'random' mutation' methodologies, could limit the 'randomness' of the variants obtained and thus the range of sequence space explored in a library.

Further terminology of nucleotide mutations is based on how they impact on the resultant protein sequence, an important issue as degeneracy within the genetic code means that gene sequence mutations can frequently have no post-translational structural effect. If a mutation causes no amino acid change, mostly occurring at the third point of a codon, then it is termed a silent, neutral or synonymous mutation (Patthy, 1999). Nucleotide changes that alter the amino acid coded for are called non-synonymous substitutions and usually occur at the first and second positions of the codon. A missense mutation refers to a non-synonymous mutation that simply results in the substitution of one amino acid in the primary protein sequence for another. A mutation that changed a codon into a termination codon would be termed a nonsense mutation and leads to premature termination of the protein.

### 1.3.2 Spontaneous & induced substitution mutations

An induced mutation would be defined as one that was proven to result from the exposure of an organism to certain natural mutagenic agents. Such agents may act on the DNA directly to change its template properties or may affect the machinery of replication, lowering fidelity and allowing the introduction of error (Patthy, 1999). Examples of directly induced DNA mutations would be the chemical modification of bases by DNA reactive chemicals (e.g. nitrous acid) and ultraviolet radiation (Watson *et al.*, 1992). Nitrous acid acts by converting cytosine into uracil (a purine, which binds to adenine), thus altering the sequence on daughter strands from G $\rightarrow$ A, while intense doses of ultraviolet radiation are absorbed by the nucleotide chains, resulting in photochemical fusion of two adjacent pyrimidines.

Spontaneous mutations are frequently replication-dependant errors occurring (very rarely) as a result of the chemical nature of the nucleotides themselves. The most common form of naturally occurring spontaneous mutation results from the fact that amino and keto groups present on nucleotides can tautomerize (to imino and endol forms) and that these transient tautomers can give rise to base pairs possessing abnormal chemical properties (Watson *et al.*, 1992). An example of this would be an imino tautomer of adenine binding with cytosine instead of the standard pairing with thymine (Patthy, 1999). The spontaneous deamination of cytosine to form uracil would be an example of a non-replication dependant mutation, and presents a problem for the long-term storage of nucleotides as well as propagating a point substitution into daughter strands if copied.

### 1.3.3 Mechanisms of error-repair

The majority of spontaneous or induced mutations occurring *in vivo* do not become permanently incorporated into the DNA of an organism. The majority of error-control mechanisms are incorporated into the polymerisation complexes involved in the synthesis of the complementary strand and thus errors can be corrected fractions of a second after they occur. The DNA polymerases themselves proofread the outcome of an polymerisation step before proceeding to the next one, and can remove and replace mis-incorporated nucleotides by means of a 3'$\rightarrow$5' exonuclease function. Multiple enzyme complexes can be involved in large-scale excision-repair

processes, where entire regions of mutated sequence can be excised and re-synthesised (Watson *et al.*, 1992).

Despite the efficiency of the fidelity mechanisms, a steady trickle of mutation ensures a level of genotype (and phenotype) diversity within any given eukaryotic or prokaryotic population. Application of evolutionary theory leads to the tempting concept of an optimal mutation rate for all forms of life, determined by evolution and environmental stability. Too high a rate could lead to an excessive proportion of non-viable progeny, while perfect copying fidelity could lead to a static, homogenous population, vulnerable to new disease types and environmental upheaval. Of particular interest is the fact that many species of bacteria has been observed to upregulate their 'net' mutation rate in response to stress (chemical, starvation, etc), and certain types have been designated as mutators (Miller, 1998). A mutator strain is defined as a strain exhibiting a significantly increased, population wide mutation rate relative to closely related organisms. The first bacterial mutators were identified in the 1950s and were initially regarded as little more than genetically unstable curiosities, but are now recognised as essential tools for studies into evolutionary adaptation and mutation avoidance (Miller, 1998). Artificially engineered mutator strains, calibrated to impose a stable rate of mutation over their total genetic contents, are now finding increasing use as tool for direct evolution (Coia *et al.*, 1997; Greener *et al.*, 1996; Henke and Bornscheuer, 1999). Please see Section 6.1.2 for discussion of mutator strain application in directed evolution protocols.

## 1.4 Random mutagenesis methods

### 1.4.1 Optimal mutation rate

Random mutagenesis, as a tool for enzyme engineering, can be used in the absence of detailed structure-function knowledge of the protein, but with the uncertainty that the effect of positive mutations may be masked by non-specific disruption of the protein conformation caused by a second mutation (Kuchner and Arnold, 1997). The latter type is by far the most likely and so the probability that an enzyme will be improved in relation to wild-type activity decreases drastically as the mutation rate rises (Arnold and Moore, 1997). The practicalities of screening for functionality in local sequence space mean that directed evolution should only be used to attempt to develop properties, which are closely related to those of the parent protein, i.e. enhancement of existing

features, which could be achieved by sequential accumulation of point mutations within the primary sequence. Accepted wisdom is thus that the mutation rate should thus be carefully limited with the aim of producing enzyme libraries with primarily single amino acid substitutions per protein, so that the individual effect of each substitution may be assessed (Arnold and Moore, 1997; Kuchner and Arnold, 1997; Moore *et al.*, 1997). As approximately one-third of DNA base substitutions are silent, the mutagenic target should be in the region of 2-3 substitutions per gene. DNA shuffling provides a method by which factorial combinations of multiple mutations may be quickly assessed, and this has led to the development of an alternate school of thought based on the creation of gene libraries, with each clone containing multiple mutations. The rationale to this is that certain fitness solutions (i.e. mutational adaptations to a particular task) may only be accessible through a number of simultaneous mutations in the protein's structure. An extreme example of this would be the directed evolution of an aspartate aminotransferase with new substrate specificities, where an unscreened homologous library was recombined using the original error-prone protocol for DNA shuffling (mutation rate of 0.7-0.8% per base pair) (Stemmer, 1994b), with the result that the five clones identified to have the highest activity on the novel substrate (2-oxovaline) contained 23-28 point mutations each (Yano *et al.*, 1998). The nucleotide substitutions were distributed evenly over the gene and sequence analysis revealed that, on average, 10-14 were non-synonymous and that only 5 of these substitutions were conserved in all 5 mutants, suggesting functional importance. As shown by Equation 11.6, which charts the total possible number of sequence mutants arising from *n* simultaneous mutations, the limiting factor for the mutation rate should be the capacity of the screening 'infrastructure'. In a highly complex library (average maximum laboratory-scale size without selection is $10^6$ clones) incorporating multiple mutations per gene, it would be inevitable that only a tiny percentage of the total possible variants could be isolated and screened. A screen sensitive to slight increases in fitness (simplified if screening of novel substrate specificity as zero wild-type background activity) coupled with successive rounds of enrichment and backcrossing (recombining mutants with wild-type enzyme to help breed-out neutral mutations) would be essential if such a strategy were to succeed.

The required mutation rate would also be obviously dependant upon the length of the target gene and should take the bias of the method used into account (Matsumura and Ellington, 2001). As a general rule, the researcher should design a mutagenic strategy with the aim of

generating as much functional diversity as possible while maintaining a minimum threshold of fitness, and constrained at all times by the physical capacity of the screen.

### 1.4.2 Methods of random mutagenesis

A list of the major current methods for generating a library of randomly mutated clones is presented in Table 1.2. DNA shuffling and epPCR are the two methods most commonly employed, because of their relative technical simplicity and also the growing body of experience in their use. Both methodologies carry disadvantages however, as indicated in Table 1.2 and as discussed further in Sections 1.4.3 & 1.4.4. Figure 1-3 gives a schematic strategy of directed evolution of protein encoding genes by epPCR, DNA shuffling and DNA shuffling incorporating backcrossing with the wild type gene.

Propagation of the vector in a mutator strain can be an attractive option as it removes the need for excessive gene manipulation, but the mutation rate in low-copy number plasmids means that successive rounds of subcloning may be required, possibly pushing the genetically unstable strain past its 'window of viability'. This is discussed further in Section 6.1.2. Beyond these principle methods, specialist strategies of direct evolution have been developed for tackling specific research problems. Of these, cassette mutagenesis is the simplest, being simply a variation of saturation mutagenesis, performed on a region of a few codons only. The strategy is based on the addition of oligonucleotides comprising random codons to defined points in a gene by means of the design and synthesis of a primer library, followed by low-stringency PCR (Miyazaki and Arnold, 1999). It is perhaps debatable as to whether this can really be considered a strategy of random mutagenesis as the primer library is usually rationally synthesised to contain each possible amino acid at a certain position (hence saturation mutagenesis), but it provides a useful tool for introducing and evaluation multiple non-synonymous mutations within a very short region of the gene. Amino acid positions to be mutated by this method are frequently identified as conserved mutations resulting from an initial library where the entire gene was subjected to epPCR (Miyazaki and Arnold, 1999) A relevant study, involving the investigation of the substrate specificity of a eukaryotic trypsin expressed in *E. coli* using oligonucleotide directed mutagenesis, looked at the effect of substitutions of a region of 6 nucleotides (Evnin, Vasquez, and Craik, 1990). These corresponded to two amino acids at the base of the substrate binding pocket, known to be

significant from previous work (Graf *et al.*, 1988). The mutagenesis strategy was designed as illustrated in Figure 1-2 except that the substitutions were designed into cassettes instead of randomly produced, and the mutants with partially preserved function were selected from the resultant library using a highly effective method of genetic selection. It is important to note that this study was not concerned with improving the activity of the protein, merely on investigating the effects of neutral and deleterious mutations on two amino acids thought to be essential in defining substrate specificity. Oligonucleotide directed mutagenesis could thus be used to intensively mutate a small region of a protein's primary sequence and, when coupled to a selection strategy, allows every possible substitution to examined, but, by definition, limits the focus of the experiment to that region. Family shuffling refers to the technique of using DNA shuffling protocols to recombine a family of genes from diverse species, including functionally similar but structurally heterlogous 'convergently evolved' proteins (Crameri et al 1998). Please see Table 1.2 for further notes and discussion.

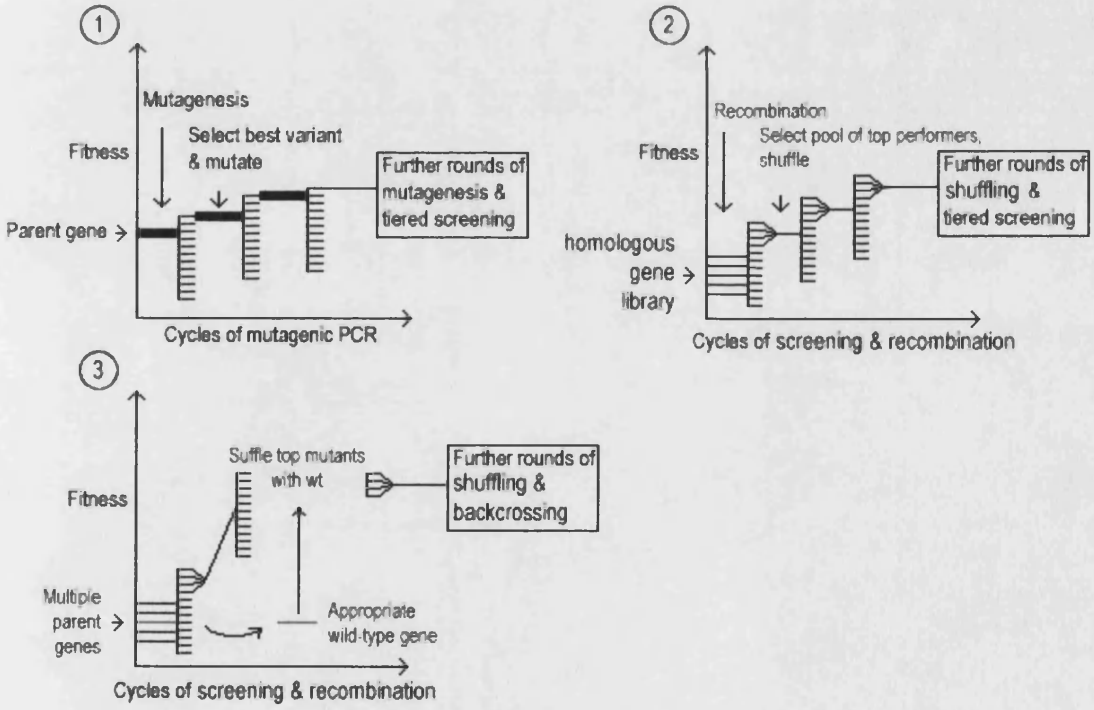



**Figure 1-3: Strategies of directed evolution**

Basic strategies of directed evolution of ① mutagenic PCR, ② DNA shuffling and ③ DNA shuffling incorporating backcross with the wild-type gene. Figure adapted from (Kurtzman *et al.*, 2001).

| Method | Principle | Mutagenesis rate (%) | Bias | Application | Disadvantages | Reference |
|---|---|---|---|---|---|---|
| epPCR ($Mn^{2+}$) | Low-fidelity PCR | Variable, documented at up to 0.9 | Transitions | Adjustable frequency of mutation | Ligation and digestion steps required. | (Fromant et al 1995; Moore and Arnold 1996) |
| epPCR ([dNTP]) | Low-fidelity PCR | Variable, documented at up to 0.9 | Transitions | Adjustable frequency of mutation | May require nucleotide analogues (if dPTP or 8-oxo-dGTP used) | (Zaccolo et al 1996) |
| DNA Shuffling | Fragmentation & reassembly | 0.7 - 0.8 | Transitions | Concentration of beneficial mutations | Use of Taq polymerase gives high basic mutation rate. | (Stemmer 1994b) |
| Family Shuffling | Shuffling of natural homologues | >0.7* | Transitions?* | Recombines divergent homologs | Technically difficult, high mutation rate. | (Crameri et al 1998) |
| Cassette mutagenesis | Extension of oligonucleotides with random dNTPs | 0 – 75 | None | Randomises 1-8 nearby codons | Requires structural model. | (Black et al 1996) |
| Mutator strain (XL1-Red) | *E. coli* lacking certain repair/proofreading pathways | 0.05 (dependant on plasmid copy no.) | Lack of published data | Technically simple method for generating diversity | Low mutation rate | (Greener, Callahan, and Jerpseth 1996) |
| Hypermutagenesis | Low-fidelity reverse transcription | 0.5 - 10 | Transitions | High mutation rate for up to 200bp | Rate too high for whole gene mutagenesis | (Matsumura and Ellington 2001) |

**Table 1.2: Comparison of modern methods of random mutagenesis.**

* The authors acknowledged the ambiguity of their data regarding a process of family shuffling and were thus unable to provide more than a rough estimate of mutation rate and bias. Mutational frequency is usually described as a percentage and target frequency can be calculated from the length of the coding sequence and the № of mutations desired. Table adapted from (Matsumura and Ellington, 2001).

### 1.4.3 Mutagenic PCR (epPCR)

#### 1.4.3.1 Theory

In mutagenic PCR, manganese ($Mn^{2+}$) ions, excess magnesium ($Mg^{2+}$) ions, unbalanced concentrations of nucleotides or the presence of synthetic nucleotide analogues in the reaction buffer are used to lower the fidelity of Taq polymerase (Matsumura and Ellington, 2001). This is thus a low-fidelity variant of standard PCR and contains the same basic steps, namely denaturation of the double-stranded template, annealing of sense & antisense primers and subsequent extension of the growing strand. It is during the extension step that mutations can occur when the polymerase allows non-complementary nucleotides to be incorporated into the growing chain. Taq is particularly suited for this work as it lacks a 3'-5' exonclease proof-reading activity and thus exhibits a base mutation rate significantly higher than comparable prokaryotic polymerises. Although certain guidelines have been established for use of each factor, it is important that the mutation rate rate be optimised empirically on the target gene as even the base levels of enzyme fidelity have been shown to be influenced by factors such as sequence (partially melted secondary structure or GC rich regions), purity and quality of the template plasmid DNA, and the manufacturer of the enzyme (Keohavong and Thilly 1989). Table 1.3 lists examples of various routinely used thermo-stable DNA polymerises for PCR. Only enzymes resistant to temperatures in excess of 95°C are used so to avoid the need for polymerase replenishment and, as indicated by the low error rate, Pfu polymerase is the clear choice for high fidelity amplification.

| Enzyme | Error Rate (errors/bp) in optimal conditions | Reference |
|---|---|---|
| Taq (*Thermus aquaticus*) | Btwn $2\times10^{-5}$ – $2.1\times10^{-4}$ | (Keohavong and Thilly 1989; Tindall and Kunlel 2003) |
| Pfu (*Pyrococcus furiosus*) | $1.6\times10^{-6}$ | (Lundberg et al 1991) |
| Vent (*Thermococcus litoralis*) | $2.4\times10^{-5}$ | (Cariello et al 1991) |
| Deep Vent (*Pyrococcus* species) | No published studied, but claimed to be higher than Vent | |
| KlenTaq (*Thermus aquaticus*) | $5.1\times10^{-5}$ | (Barnes 1992) |

**Table 1.3: Error rates for thermal resistant DNA polymerases**

As indicated by the bias shown in Table 1.2, epPCR does not create truly random nucleotide substitutions, favouring transitions over transversions and with a particularly high occurrence of A → G substitutions (Fromant *et al.*, 1995; Zhao *et* al., 1999). Bias affects thus the location of mutations (mutations introduced by Taq occur more often at AT base pairs than at GC pairs) and their type (Taq favouring purine transitions, A ↔ G). The significance of bias in an experiment can only usually be determined after the sequencing of a representative sample of genes. Lowering Taq fidelity by means of unbalanced concentrations of dNTPs clearly introduces bias in favour of the over-represented nucleotide and, while a well characterised method, is generally regarded as less useful for this reason (Zhao and Arnold 1997a).

Mutation frequencies ranging from 0.11 – 0.49% per base have been accomplished by alterations of the [Mg]:[Mn] ratio (Matsumura and Ellington 2001). The total concentration of salts in a PCR reaction buffer is usually 0.1 – 3mM, being empirically altered for each template in order to optimise reaction yields. In a reaction using [total salt] of 7mM, mutation rates of ~ 0.1% at 0mM $MnCl_2$ and ~ 0.5% at 0.5mM $MnCl_2$ were recorded. A sample protocol of salt-mediated epPCR, which the researchers claim can deliver an overall mutation rate of ~0.2% on purified plasmid DNA, is stated in Appendix B (Zhao *et* al., 1999). The overall mutational frequency in epPCR can thus be regarded as the product of three parameters: the error rate (fidelity) of the selected polymerase in the reaction conditions, the length of the gene being amplified, and the number of effective doubling cycles.

#### 1.4.3.2 Muller's ratchet & asexual reproduction

While epPCR provides a stable and controlled platform for the introduction of point mutations into a gene, its suitability for experiments likely to require successive rounds of mutagenesis decreases in inverse proportion to the number of cycles. The probability of epPCR continuously improving a gene through the addition of point mutations has been shown to decreases drastically with each round (Maranas, 1999), as the build-up of negative mutations typically masks beneficial substitutions. This phenomenon attends all forms of asexual reproduction and is referred to in population dynamics as 'Muller's ratchet'. Muller's ratchet is a theory put forward to explain the prevalence of sex as a means of reproduction. Asexual reproduction offers an efficient method for

any organism to reproduce itself (100% probability of each parent gene being expressed in offspring) and effectively doubles the reproductive strength of a species by eliminating the need for males. The theory describes a series of modelling experiments into the accumulation of deleterious mutations in lineages/individuals under asexual reproduction, and showed that, over time, the 'most fit' class of an asexual population would be gradually lost to the accumulation of negative mutations (Patthy, 1999). With each successive generation, i.e. with each 'turn of the ratchet', further mostly deleterious and silent mutations would accumulate, irreversibly lowering the mean fitness of a population. The model showed that sexual reproduction, while energy expensive (cost of meiosis = cost of producing males), provides a method for gene shuffling and recombination, allowing the least-mutated class of each generation to be regenerated, and thus effectively countering the ratchet.

The same model could equally be applied to the *in vitro* asexual generation of gene diversity by epPCR, showing that the accumulation of mutations on each gene would inevitably result in a lowering of fitness relative to the parent. DNA recombination is thus a naturally complementary method for epPCR, as well as being able to introduce diversity in its own right with certain protocols, the gene shuffling allows potentially beneficial mutations to be screened factorially and in isolation, while back-crossing with the wild-type gene ensures negative and silent mutations can be edited out of the clones without the need for sequencing or rational involvement. DNA shuffling is discussed in detail in Section 1.4.4.

#### 1.4.3.3 Modelling of mutagenic PCR

The modelling of mutagenic PCR requires initially a mathematical description of the polymerase chain reaction and, secondly, a robust method for the factoring in of the error-rate. Significant attention has been paid to developing a model of epPCR which can accommodate bias, primarily as such a model would provide a useful adjunct to the far more complex models of DNA recombination currently being debated. A simple equation for predicting the expected number of mutations of an epPCR protocol based only on the gross error rate is shown in Equation 1.1.

$$f = \frac{Np}{2}$$

**Equation 1.1**

Predicts overall error-rate (*f*) after *N* PCR cycles with a per cycle error rate of *p*. Equation proposed by (Eckert and Kunkel, 1990).

Equation 1.1 provides a useful rule-of-thumb for estimating the number of expected substitutions, but can provide no data beyond that as it only considers mutations as a group, with no reference to the bias of the method. For a more sophisticated model, one which treats individual mutations as nucleotide-dependant steps and thus would allow prediction of the types and locations of the expected mutations, it is first necessary to describe the PCR process more exactly. As the concept of lowering enzyme fidelity implies that the more frequently a sequence is copied, the more likely it is to acquire a mutation, it is important to note that after *N* cycles of PCR, not all sequences in the reaction mixture result from exactly *N* extensions of the original template (Moore and Maranas, 2000). Figure 1-5 illustrates how many cycles of expansion were required to create each individual strand. A particular DNA strand is obviously sequenced only once, but the higher the number of synthesis steps occurring before it is produced, the higher the chance of a mutation occurring. It is evident that a mutation which entered the sequence of a newly synthesised strand particularly early in the reaction (e.g. $N = 2$) would be propagated throughput a large section of the population, possibly masking subsequent substitutions and appearing as a conserved feature in the library. The relatively low mutation rates involved make this an unlikely situation however, and mutations would only be statistically likely to occur later in the reaction, when thousands of individual strands were being synthesised at once.

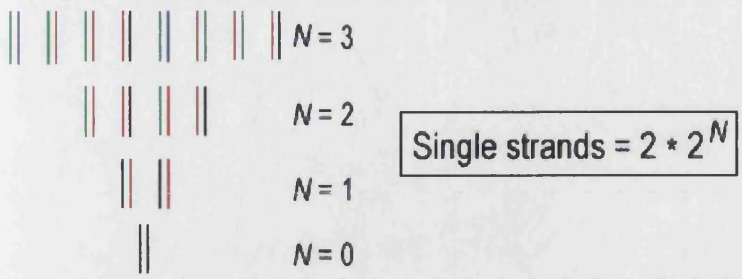



**Figure 1-4: Illustrating the exponential increase of DNA single strands using the polymerase chain reaction**

*N* refers to the number of extension steps (i.e. stand synthesis cycles) undergone. At *N* = 3 the mixture would contain 16 single strands, 2 of which would be parental template (black), 6 of which would be the result of one extension step (red), 6 of which would be the result of two expansion steps (green), and 2 would be the result of three expansion steps (blue). Total single strands increase exponentially with each step according to embedded equation.

There was a need to express the relationship of newly synthesised strands to the number of extension steps required to form them (Figure 1-5) mathematically so that the relative probability of generating a mutation during synthesis could be calculated. Moore and Maranas, researchers from Pennsylvania State University, proposed the model shown in Figure 1-5, which defined the concept of $Z_{N,n}$ as representing the number of single strands which have been through *n* extension steps after *N* cycles of PCR (Moore and Maranas, 2000). Central to the model were the statements that, as two parental template strands were present at the beginning of the reaction and were not the result of an extension step, $Z_{0,0} = Z_{N,0} = 2$, and similarly, after *N* PCR cycles are completed, no single strands will have been synthesised that are the result of more extension steps than *N*, i.e. $Z_{N,n} = 0$ for $n > N$. The authors took the combined statement of $Z_{N,n} = Z_{N-1,n} + Z_{N-1,n-1}$ Figure 1-5 and constructed a proof by induction for Equation 1.2, validating the relationship by solving for $n = 0, 1, 2 \& 3$.

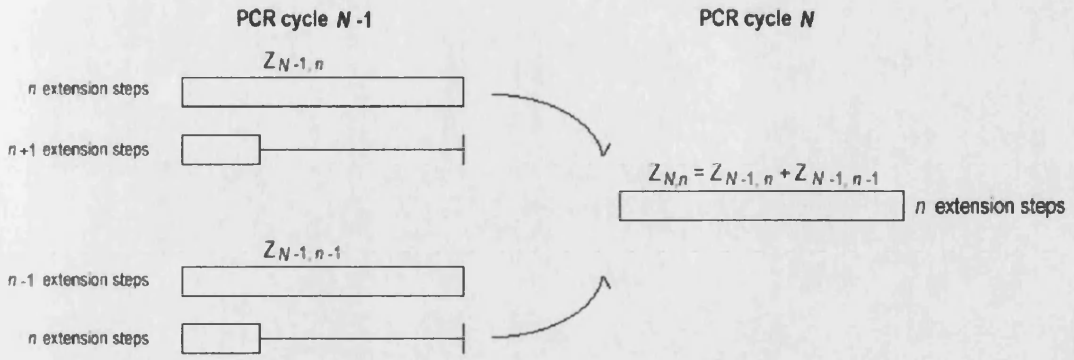



**Figure 1-5: Describing the number of strands that have been through *n* extension steps after *N* PCR cycles**

Each strand in the mix can be classed as originating from a template ($Z_{N,n}$) of one of two types: a strand synthesised prior to the PCR cycle *N*-1 and thus has been through *n* extension steps (template referred to as $Z_{N-1,n}$); or a comparatively new template strand, itself synthesised from a older template and thus having gone through *n*-1 extension steps (template referred to as $Z_{N-1,n-1}$). Figure adapted from (Moore and Maranas, 2000).

$$Z_{N,n} = 2\binom{N}{n}$$

**Equation 1.2**

The number of single strands ($Z_{N,n}$) that have been through *n* extension steps after *N* cycles of PCR. Equation from (Moore and Maranas, 2000)

Combining Equation 1.2 with the exponential statement for the synthesis of single DNA strands presented in Figure 1-5, results in Equation 1.3, which relates $Z_{N,n}$ to the total number of DNA strands synthesised at each stage of the reaction.

$$Z_{N,n} = \frac{1}{2^N}\binom{N}{n}$$

**Equation 1.3**

Fraction of sequences in mix after *N* PCR cycles that are the result of *n* extension steps. Equation from (Moore and Maranas 2000).

A crucial aspect of the model was that it would not just accurately factor the number of extension steps required to synthesis each strand and thus the 'window' for mutagenesis in each case, but it would also allow for the prediction of the type of mutation that would occur, thus enabling the model to compensate for any bias in the procedure. Mis-incorporation rates of polymerises were known to be nucleotide specific (e.g. Taq prefers A → G substitutions to any other type) (Fromant, Blanquet, and Plateau, 1995), and thus a per cycle mutation matrix (**M**) was proposed with elements $\mathbf{M}_{ij}$ representing the probability of nucleotide *i* being substituted with nucleotide *j* Equation 1.4 (Moore and Maranas, 2000). The probability associated with each individual element would need to be determined empirically and would be dependant upon the polymerase used, the sequence of the template and, crucially, the reaction conditions used.

$$\mathbf{M} = \begin{pmatrix} M_{AA} & M_{AT} & M_{AC} & M_{AG} \\ M_{TA} & M_{TT} & M_{TC} & M_{TG} \\ M_{CA} & M_{CT} & M_{CC} & M_{CG} \\ M_{GA} & M_{GT} & M_{GC} & M_{GG} \end{pmatrix}$$

**Equation 1.4**

Mutational matrix dealing with the individual probabilities of one nucleotide being replaced by another. Values to be determined empirically. Equation from (Moore and Maranas, 2000).

The per cycle mutation rate matrix **M** (Equation 1.4) was then be used to calculate the overall mutation rate matrix after *n* extension steps, known as $\mathbf{C}^n$ Equation 1.5. $\mathbf{C}^n$ could then used to determine the probable mutation rate of a sequence after *n* extension steps from the original parental template. A crucial point was that the occurrence of nucleotide substitutions introduced in each new strand during a synthesis step was independent of mutations that had occurred during previous cycles. $\mathbf{C}^n_{ij}$ (the probability of *i* being substituted by *j* after *n* extension steps) was calculated in a recursive manner, i.e. calculated by necessity for each value of *n*.

$$\mathbf{C}_{ij}^{n} = \begin{cases} \delta_{ij}, & n = 0, \\ M_{ij}, & n = 1, \\ \sum_{k=A,C,T,G} M_{kg} C_{ik}^{n-1}, & n \geq 2, \end{cases}$$

**Equation 1.5**

Matrix describing the mutation rates of a sequence obtained after *n* extension events, starting from the original template. $\delta_{ij}$ refers to the two parental templates at $n = 0$ and therefore will equal 0 unless $i = j$ (effectively no substitution) where it will equal 1. Equation from (Moore and Maranas, 2000).

The final form of the model combines matrix $\mathbf{C}^{n}_{ij}$ (Equation 1.5), with its combined measurements of the distinct nucleotide substitution rates, with Equation 1.3, that describes the number of single strands produced and the relative probability of finding a mutation on any particular one (proportional to *n* extension events). The combined statement is shown in Equation 1.6.

$$\mathbf{P}_{ij}^{N} = \frac{1}{2^{N}} \sum_{n=0}^{N} \binom{N}{n} C_{ij}^{n}$$

**Equation 1.6**

Describing the probability ($P^{N}_{ij}$) of nucleotide *i* being substituted with nucleotide *j* after *N* PCR cycles. Equation from (Moore and Maranas, 2000).

The relationship shown in Equation 1.6 allows a researcher to calculate, *a priori*, the probable number and type of mutations that could be expected on a gene of a certain length, after a certain number of complete PCR cycles and in calibrated reaction conditions. The equation also provides the mathematical basis for more involved forms that allow for calculation of the probability of obtaining a target sequence by mutagenesis of the original (Moore et al 2000; Moore and Maranas 2000). This is a particularly useful relationship to establish as, once calculated, extraction of the now desired mutation rate would allow the researcher to precisely tailor PCR reaction conditions so as to maximise the probability of the target sequence emerging in the screen.

### 1.4.4 DNA shuffling

#### 1.4.4.1 Theory

The importance of DNA recombination (i.e. sexual reproduction) in population dynamics has long been established as a means of generating variability while naturally eliminating deleterious mutations from an evolving population, thus avoiding the negative implications of Muller's ratchet (see Section 1.4.3.2). The first *in vitro* method for performing DNA recombination of a pool of double-stranded genes, and which has formed the template for subsequent specialised variations, was designed by Stemmer in 1994 and involved the enzymic digestion of the genes into short fragments by incubation with Dnase I, followed by reassembly of the overlapping fragments into full genes by means of a primer-free PCR reaction (Stemmer, 1994a). As the fragments are overlapping they are free during the PCR, to associate with complementary fragments from homologous genes, resulting in novel combinations of the mutations present in the initial library. A final high-fidelity PCR step can then produce large quantities of each variant for cloning and plasmid amplification in *E. coli*, and then transfer into a suitable expression strain if necessary (Arnold and Moore, 1997). The library would then be screened for enzymes with increased fitness towards a particular parameter, and a pool of variants, all of which would code for enzyme exhibiting fitness above a defined threshold, would then selected for the next round of recombination and screening (see Figure 1-3). In this way, DNA shuffling avoids the loss of different types of low-level beneficial mutations by selecting all to participate in the next level of the process. The mixing of mutations also acts to eliminate the risk of selecting a single active clone, which could in reality be stuck at a local fitness minimum, i.e. by the nature of the individual fitness-improving mutation acquired, it may be that no other single point substitutions could raise the fitness of the mutant any higher. In such a case, solutions could only be reached through multiple simultaneous amino acid substitutions, achievable by increasing the mutation rate of epPCR or through recombination of existing mutations (Lutz and Benkovic 2000)

An important factor to be considered when designing a strategy of directed evolution is that when Taq polymerase is used for the reassembly reaction, the act of forcing the polymerase to synthesise a full length gene from the small fragment DNA pool can cause the introduction of point mutations into the sequence, with a relatively high substitution rate of approximately 0.5-0.7% (see Table 1.2) (Stemmer, 1994b; Zhao and Arnold, 1997a). Successful improvements of enzyme

fitness have been reported using DNA recombination as a tool for both generating diversity and recombining the resultant mutations, for example brightness of a green fluorescent protein (Crameri *et al.*, 1996) and the substrate specificity of an aminotransferase (Yano *et al.*, 1998), but the unavoidably high mis-incorporation rate even under optimum conditions, means that the complexity and thus theoretical size of the library to be screened can deter researchers. Additional discussion as to optimum mutation rates can be found in Section 1.4.1. Replacement of Taq with the high fidelity Pfu polymerase (see Table 1.3 for error rates under standard PCR conditions) has been shown to lower the mutation rate of a shuffling protocol to 0.05% (Zhao and Arnold, 1997b). Use of this high fidelity recombination protocol means that initial diversity in a gene can be generated by means of epPCR with a carefully selected and controlled error rate, calculated to be optimal for the length of the gene and the capacity of the screen.

#### 1.4.4.2 Variations of the standard recombination protocol

Beyond the high-fidelity variant of DNA shuffling (discussed in Section 1.4.4.1), other protocols based on Stemmer's original technique have been developed, both for altered forms of recombination and for altered rates of mutagenesis. A list of the most significant types can be seen in Table 1.4.

The staggered extension primer process (StEP) is a technically simplified form of DNA recombination in that it eliminates the need for digestion of the strands with Dnase I and thus also enables the reaction to be performed in a single tube as the mix need not be purified after fragmentation (Zhao et al 1998). Essentially, StEP consists of the priming of the template sequences, followed by repeated cycles of denaturation and annealing, and then by extremely abbreviated polymerase catalysed extension cycles. The duration of the extension cycles is set so that the polymerase can only incorporate a set number of nucleotides into the growing strand per cycle (the number can be calibrated by the length of the gene and the amount of recombination required), before denaturation dissociates the double strand and polymerase complex. With each new annealing period, the growing fragments can anneal to different homologous templates based on sequence complementarity, and thus each PCR cycle can constitute a recombination event. Key to the success of the process is that annealing (hybridisation of growing strand to new template) must be performed under high-stringency conditions to minimise the incidence of

frameshift or indeed 'domain-wandering' mutation, and so, typically, $T_{anealing} \geq T_m - 25°C$, where 'm' equals the melting temperature of the initial primer (Zhao *et al.*, 1998).Taq polymerase has been reported to exhibit a nucleotide extension rate of 100-150 nucletides per second at its optimal temperature of 72°C, and 25-38 nucleotides per second at 55°C (the temperature normally used) (Zhao and Arnold, 1998).

Standard DNA shuffling reactions are limited by their need to work with homologous libraries of genes and thus the potential benefits to be gained through the recombination of highly diverse sequence sets or recombining an existing library in a sequence-independent manner cannot be explored. Evolutionary studies generally find a higher degree of structure, rather than sequence, conservation between distantly related proteins, and thus methods for combinatorial homology-independent protein engineering may be able to access regions of functional diversity previously denied (Lutz and Benkovic, 2000; Lutz and Benkovic, 2002). The method of incremental truncation for the creation of hybrid enzymes (ITCHY) facilitates this by providing a means by which randomly truncated genes could be brought together end-to-end, thus requiring no form of hybridisation. Essentially, ITCHY involves the truncation of both strands of the parent gene by incubation with exonuclease III (no internal sequence cuts), followed by the creation of a single cross-over library by means of ligation of the resultant N- and C- terminal fragments (Lutz and Benkovic, 2000). The hybrids generated would cover a range of sizes, both longer and shorter than the parent strands and this, combined with the naturally very high incidence of frameshift mutations occurring at the random 'junction points' means that a high proportion of the enzymes expressed would be inactive. A variant of ITCHY that involved the use of nucleotide analogues such as $\alpha$-phosphothiate nucleotides (THIO-ITCHY) was developed and involved a significantly simpler experimental protocol and faster library generation times (Lutz *et al.*, 2001a). Both ITCHY and THIO-ITCHY were limited in that only two parents could be joined at any one time, and thus each progeny hybrid carries only one junction. The high proportion of defective hybrids would also necessitate a selection scheme to avoid wasted screening work.

Sequence homology independent protein recombination (SHIPREC) is an alternative method of creating such single crossover libraries. The process creates an inverted fusion protein of the two parent strands (so as to protect the N- and C- terminus) and then subjects them to partial digestion with DNase I, before inverting them once more (Sieber *et al.*, 2001).

The creation of hybrid libraries with multiple sequence-independent crossovers though a combination of ITCHY and DNA shuffling is termed SCRATCHY (Lutz *et al.*, 2001b), and would appear to be the logical conclusion in the development of such techniques. The process involves the creation of two complementary ITCHY libraries composing of hybrid sequences screened to be of parental size and in the correct reading frame, and recombines the libraries by means of random fragmentation and reassembly PCR. The mixing of structural elements means that, again, a high proportion of inactive enzyme would be expected and the process should be coupled to a selection step in an auxotrophic expression host if at all possible.

Two remaining methods stand apart from the others, while still being based on the concept of genetic recombination. The first, *in vitro* exon shuffling, represents an attempt to mimic another other form of gene recombination observed in natural systems. This involves the amplification of eukaryotic exons that encode protein domains by means of chimeric oligonucleotides and their subsequent rearrangement and PCR reassembly (Kurtzman *et al.*, 2001). As this method involves the shuffling of entire structural domains, a high degree of fitness in the resultant clones would be expected. The second method is termed random chimeragenesis on transient templates (RACHITT) and is a technique based on DNA shuffling in which single-stranded parental gene fragments are annealed onto a full-length single-stranded template (Coco *et al.*, 2001). The 'reassembled' double-stranded template could then be treated with ligase to seal nicks in the sequence, before being amplified by high fidelity PCR.

| Method | Advantages | Disadvantages | Reference |
|---|---|---|---|
| StEP | No need for DNA digestion; fragment purification. | Reliant on sequence homology (similar to standard shuffling) | (Zhao, Giver, Shao, Affholter, and Arnold 1998) |
| ITCHY | Independent of sequence homology; random digestion allows distribution of junctions throughout gene. | Hybrids limited to 2 parents of equal length; parent length not conserved; high rate of frameshift events thus high proportion of low fitness hybrids. | (Lutz and Benkovic 2000) |
| THIO-ITCHY | Same pros as ITCHY but technically easier to perform. | Same cons as ITCHY; use of dNTP analogues may present problems for future cloning/manipulation steps. | (Lutz, Ostermeier, and Benkovic 2001a) |
| SHIPREC | Same pros as for ITCHY; allows selection of hybrids of parental size. | Cons as for ITCHY except that parent length may be conserved. | (Sieber, Martinez, and Arnold 2001) |
| SCRATCHY | Independent of sequence homology; allows multiple parents per hybrid. | High rate of frameshift events coupled with overall low fitness of hybrids requires a selection step. | (Lutz, Ostermeier, Moore, Maranas, and Benkovic 2001b) |
| Exon shuffling | Independent of sequence homology; based on structural homology thus tends to maintain fitness. | Restricted to intron-containing genes (eukaryotic origin); diversity is directly proportional to number of exons present. | (Kurtzman, Govindarajan, Vahle, Jones, Henrichs, and Patten 2001) |
| RACHITT | PCR –free method thus avoids bias & introduction of new mutations; only very short regions of sequence homology (0-5 bases) required. | Quality of DNA may be affected by process; process involves multiple cloning & preparatory steps. | (Coco, Levinson, Crist, Hektor, Darzins, Pienkos, Squires, and Monticello 2001) |

**Table 1.4: Variations on the standard protocol of DNA recombination**

All methods listed here are described in more detail in Section 1.4.4.2. Acronyms: StEP = Staggered Extension Process; ITCHY = Incremental Truncation for the Creation of Hybrid enzYmes; THIO-ITCHY = ITCHY with $\alpha$-phosphoTHIOate deoxynucleotide triphosphates; SHIPREC = Sequence Homology Independent Protein RECombination; SCRATCHY = ITCHY plus DNA shuffling (?); RACHITT = RAndom CHImeragenesis on Transient Templates.

#### 1.4.4.3 Modelling of DNA shuffling

Constructing a model of the standard DNA shuffling procedure is extremely complex, as it requires the description of three fundamentally separate processes if the precise mechanics and probabilities of the recombination are to be understood (Moore *et al.*, 2000). The first would be a model of the random fragmentation of the gene with DNase I. The model would have to factor in the distribution of digestion fragment sizes and the spread of mutations already present on the gene. The second model would describe the recombination of the fragments and the synthesis of complete genes, as well as the probability of mutations being introduced during synthesis (if Taq polymerase was used). The final section would involved a sequence matching model, which would use the recombination mechanics and fragment size distributions of the previous steps to calculate probability of a particular target gene (a product of *x* fragments from *n* sequences) arising so that oversampling of the library for screening could be adjusted accordingly. The three integrated models will not be discussed here, but the original calculations and full descriptions can be found in the year 2000 work of Moore and Maranas (Moore *et al.*, 2000; Moore and Maranas 2000).

A simple statistical method does, however, allow the calculation of the probability of generating a sequence through recombination which contains a certain number of mutations present in the original strand (Moore *et al.*, 1997). While this simplified model could not present a complete view of the process, it was sufficient to provide a guide for experimental design and to give an indication of the degree of oversampling required when screening clones. The model assumed that all mutations are unique and distributed close to randomly. Stating regular distribution would be a gross oversimplification, but here the assumption was that all mutations are far enough apart for recombination to occur freely between them.

The basic statistics of recombination can be stated as follows. Assuming a library of three genes, each containing a single different mutation, after fragmentation, the probability that any given mutation entering into a progeny sequence would be 1/3. Logically then, the chance of a progeny carrying a wild-type sequence at any given mutation point would be 2/3. For any mutations in the initial library, there is thus a statistical preference for the absence of the mutation in the progeny. The probability of the shuffling event resulting in a completely wild-type sequence (no mutations) would thus be calculated per mutation: $(2/3)^3 = 0.3$. Equally, the probability of generating a sequence with only one mutation through recombination (a regenerated parent

strand) at each mutation point would be (1/3) × (2/3)² = 0.15. As there are three parent sequences, combining the probability of their generation with that of the wild-type ( 0.3[0.15 × 3] = 0.75) leads to the conclusion that, of the shuffled library, fully 75% of sequences would be parental or wild-type, with only around 25% being novel recombinations of mutations (Moore *et al.*, 1997).

The above probabilities allow the construction of the factorial expression in Equation 1.7, which describes the number of ways that a sequence with $\mu$-mutations could be generated in a system with $M$ total mutations.

$$C_{\mu}^{M} = \frac{M!}{(M-\mu)!\mu!}$$

**Equation 1.7**

The number of ways that a $\mu$-mutation sequence may be generated in a recombination system containing $M$ total mutations. Equation from (Moore, Jin, Kuchner, and Arnold, 1997)

The logical progression was then to multiply the value from Equation 1.7 with the probability of generating any individual $\mu$-mutation sequence. A combined form of the equation is presented in Equation 1.8 and describes the probability ($P_{\mu}$) of generating a $\mu$-mutation sequence in a system with $n$ total sequences and $M$ total mutations per strand.

$$P_{\mu} = C_{\mu}^{M}\left(\frac{1}{N}\right)^{\mu}\left(\frac{N-1}{N}\right)^{M-\mu}$$

**Equation 1.8**

The probability of arriving, through DNA recombination, at a $\mu$-mutation sequence in a system with $n$ total sequences and $M$ total mutations per strand. Equation from (Moore, Jin, Kuchner, and Arnold, 1997).

In a frequency distribution of recombined sequences, those with a single mutation predominate ($N = M$), while the rarest sequences to be generated are those with the most number of concentrated mutations. The probability of screening the least likely sequence ($P^{M}$), containing all mutations ($\mu = M$), can be calculated as $1/N^{M}$ (Moore, 1999;Moore *et al.*, 1997).

## 1.5 Project background

### 1.5.1 Industrial uses of recombinant bovine trypsinogen

Bovine trypsin has been characterised as a robust, digestive enzyme, with high levels of activity towards a relatively broad range of substrate sequences (for a comprehensive discussion of the substrate specificity of this enzyme please see Section 1.6.7) and as such, has found use as a tool for biocatalysis in many laboratory-scale and industrial procedures. The protease is widely used in mammalian cell culture (trypsinization) and is frequently used for the processing of recombinant constructs within the pharmaceutical industry, examples being the cleavage of fusion proteins (solubility or affinity tags) (Colman, 1997) and, crucially for this work, the processing of insulin and insulin-analogues. The production of biosynthetic human insulin (BHI) comprises a major, relatively stable part of Eli Lilly's drug portfolio, with the current market being estimated at between 15,000 to 25,000 kg per year, growing at an annual rate of 5% to 6% (Petrides *et al.*, 1995). The related costs and mechanisms of insulin production, as well as the role of trypsin as a bioprocessing agent, are analysed in Section 1.8. Essentially, recombinant bovine trypsin is used to separate the C-peptide from the immature insulin proprotein (please see Figure 1-15), allowing the A- & B-chains to associate with each other and assume the native conformation. The relatively broad substrate specificity of the enzyme means that a variety of mis-cleavage or undesirable products also result from the enzymatic conversion and a comprehensive strategy of multimodal chromatography is required for BHI purification (see Figure 1-17).

Briefly, the purification steps necessitated by this enzymatic cleavage step are, following initial concentration by diafiltration, passage through an ion-exchange column (to remove C-peptide and modifying enzymes) and subsequent use of reversed-phase high-pressure liquid chromatography to remove structurally similar contaminants, e.g. miscut insulin/proinsulin. Figure 1-6 illustrates a breakdown of the process costs in terms of stage and equipment. Chemical modifications and intermediate purification steps (i.e. steps 3-6 on Figure 1-16), account for almost 60% of the total cost, while the purification resulting from step 6 (enzymatic cleavage) reaches approximately 30%. Significant savings could thus be realised if the process of enzyme conversion could be optimised into a more efficient, 'cleaner' reaction, and so the development of technology that would enable the rapid tailoring of trypsin substrate specificity would be of immediate use in a large-scale industrial context.

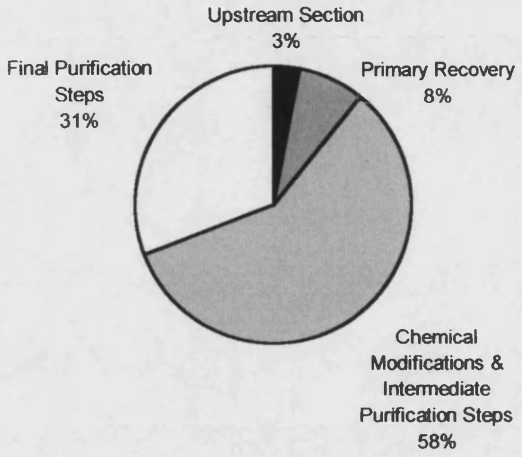



**Figure 1-6: Operating cost per process section of BHI production**
Operating cost of intracellular method of BHI production per process section and equipment category. Figure adapted from (Petrides, Sapidou, and Calandranis, 1995).

### 1.5.2 Project goals

The PhD project described in this thesis was proposed by Eli Lilly with the aim of developing a process for producing recombinant bovine trypsinogen (r-trypsinogen) in a manner that would allow the high-throughput screening of clones essential for a practical strategy of directed evolution. Such trypsin variants would then be screened against a range of novel peptide library designs, structured in such a way as to enable the identification and characterisation of specificity mutations, leading to the development of more precise and hence more cost effective biosynthetic protein processing tools. The goal was thus to explore methods by which a serine protease could have its substrate specificity tailored to optimise a particular biocatalytic reaction and to create, if possible, a 'tool-box' of mutant proteases with altered and characterised substrate specificities. Trypsin was selected for this work, as it is an well-studied protease with a fully solved structure

(Bode and Schwager, 1975) and it is currently used by Lilly in at least one large-scale bioconversion reaction (see Section 1.5.1) which could be used as a economic case study (see Section 1.8). Equally relevant was the fact that the activity and specificity of trypsin, while extensively explored through point mutation work (Evnin *et al.*, 1990; Hedstrom, 1996), have not previously been subjects of directed evolution experimentation. The reason for this was that mammalian trypsin, unlike related prokaryotic serine proteases, e.g. subtilisin (You and Arnold 1994), is extremely difficult to express in soluble, active form, whether in yeast or bacterial host, with only a few successful experimental systems being reported (Hanquier *et al.*, 2003; Perona *et al.*, 1993).

Three main requirements have been recognised as essential for the implementation of a successful strategy for directed evolution. These are: 1) the functional expression of the enzyme within a microbial host (for ease of genetic manipulation). The protein should preferably be expressed in an active state; 2) the choice of an appropriate screening or selection strategy sensitive to the desired properties of the enzyme; and 3) selection of a practical evolution strategy (Kuchner and Arnold, 1997; Zhao *et al.*, 1999). While condition 1) should be considered mandatory, points 2) & 3) are closely linked, the type of screen appropriate being dependent on the amount of mutation exhibited by each clone (Moore *et al.*, 1997).

## 1.6 Bovine trypsin: general data

### 1.6.1 Classification

Proteases are classified mechanistically in terms of the location of active sites, mechanism of action and three-dimensional structure. To date, four classes of proteases have been identified, and six families distinguished within them. The class of serine proteases are divided into two families (class I: Eukaryotic, class II: Bacterial) with bovine trypsin (EC 3.4.21.4) belonging to the first group along with other proteases of common ancestry such as chymotrypsin and elastase (Beynon and Bond, 1989). All members of the serine protease I possess common enzymatic mechanisms and active-site structures, and represent one of the most extensively characterised protease families. Although a degree of sequence variation exists within the family, the global conformation remains constant with all proteins exhibiting two similarly folded domains arranged around a two-fold axis of symmetry (Beynon and Bond, 1989; Lesk, 2001).

### 1.6.2 Sequence and secondary structure

A representative structure of a bovine trypsin is available on the public access Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Base (http://www.rcsb.org/pdb), file name: 1BJV or 1TGN. The sequence of the mature protease is identical to that of the recombinant Eli Lilly protein, however the propeptide (not shown here) differs by 3 amino acids at the 5' end.

Classification: Serine Protease
Source: Bos Taurus (bovine pancreatic extract)
Molecular weight (of mature trypsin): 23290
Molecular weight (trypsinogen): 23978
Number of Residues: 223 (+8 with prosequence)
Optimum pH: Approximately 8.0
Extinction coefficient: 14.3

| Letter | Structure Element |
|---|---|
| H | alpha helix (4-helix) |
| B | residue in isolated beta bridge |
| E | extended strand, participates in beta ladder |
| G | 310 helix (3-helix) |
| I | pi helix (5-helix) |
| T | hydrogen bonded turn |
| S | bend |

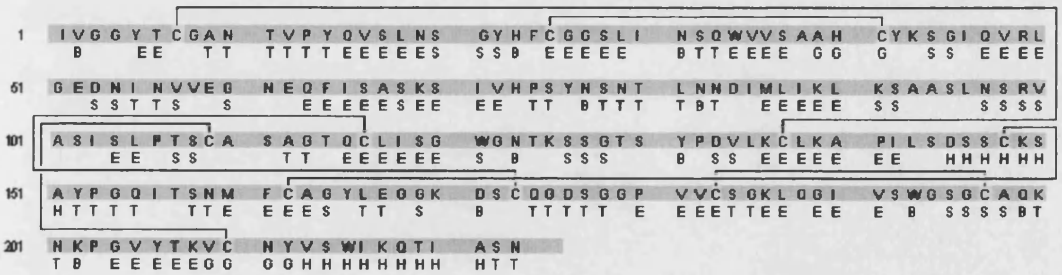


**Figure 1-7: Sequence and secondary structure annotation of mature bovine trypsin.**
Secondary structure nomenclature explained in embedded table. Black lines represent bonded cystine pairs. 3d structure of protein shown in Figure 1-8.

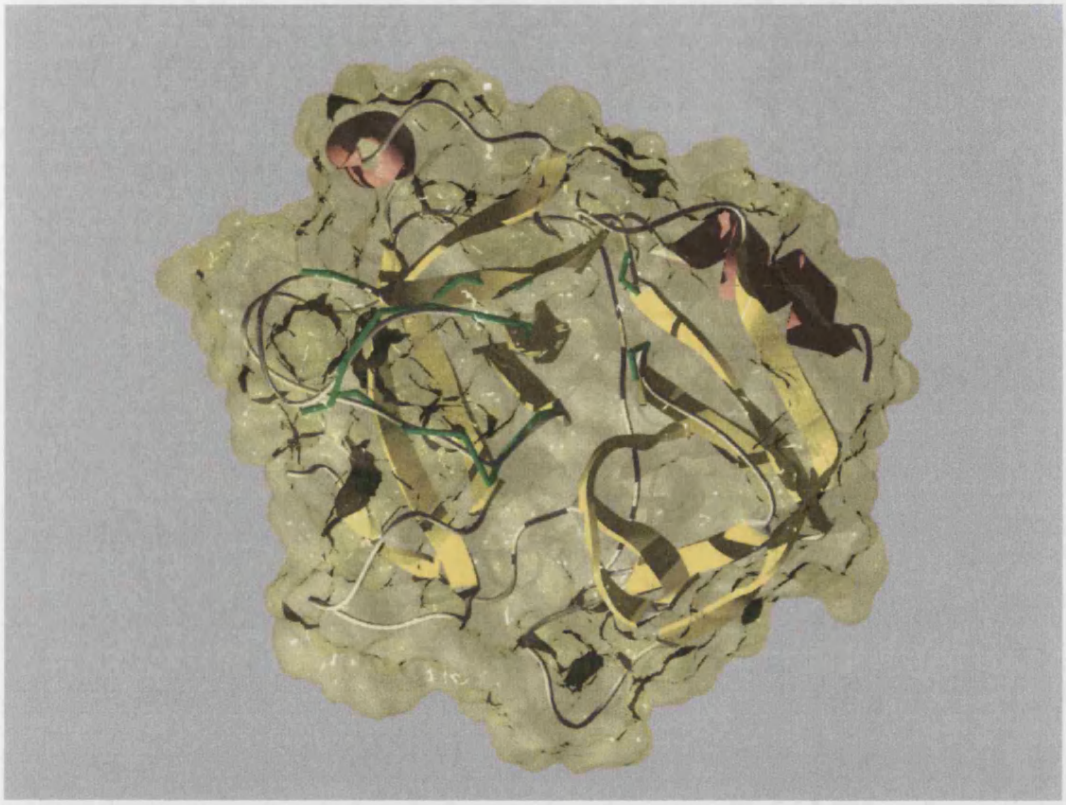


**Figure 1-8: Structure of bovine trypsin (PDB: 1TGN).**
Cartoon ribbon with transparent molecular surface overlay. Beta sheets shown in yellow (arrow indicates direction), helices shown in red. Amino acids forming part of the catalytic traid or substrate binding regions (Section 1.6.3, Figure 1-9) shown in green. Model produced using Swiss-PDB viewer 3.7 and PovRay 3.5.

### 1.6.3 Specificity and catalytic regions

Considerable work has been applied to crystallising and solving the structure of trypsin and means the structural basis for the enzyme's proteolytic activity is fairly well understood (Bode and Schwager 1975). Overall activity is usually considered as two separate components, enzymatic specificity and catalysis, with the various active portions of the enzyme being categorised according to type.

**Substrate Specificity regions:** The substrate binding pocket of bovine β-trypsin has been defined as residues 189-195, 214-220 and 224-228 (Hedstrom, 1996). The specificity of the binding pocket

for Arg/Lys is known to be primarily dictated by the presence of Asp189 at the base of the S1 site (Bode and Schwager, 1975; Hedstrom, 1996). The methylene groups of Arg & Lys (substrates) form hydrophobic interactions with the enclosing chains of the pocket (particularly with residues 214-216 & 190-191) and it has been shown that the terminal side-chain ammonium group of Lys forms a salt-link with the carboxylate group of Asp189 (Bode and Schwager, 1975).

**Catalytic regions:** A catalytic triad of residues has been characterised, namely Ser195, His57 & Asp102 (trypsin numbering) (Perona *et al.*, 1995) and are illustrated in Figure 1-9. These three amino acids comprise a charge relay system involving the aspartic acid residue being hydrogen-bonded to the histidine, which itself is hydrogen-bonded to the serine. This triad is found in all serine proteases, even those which are separated by vast evolutionary distance, such as the bacterial serine protease II, subtilisin (an example of convergent evolution).

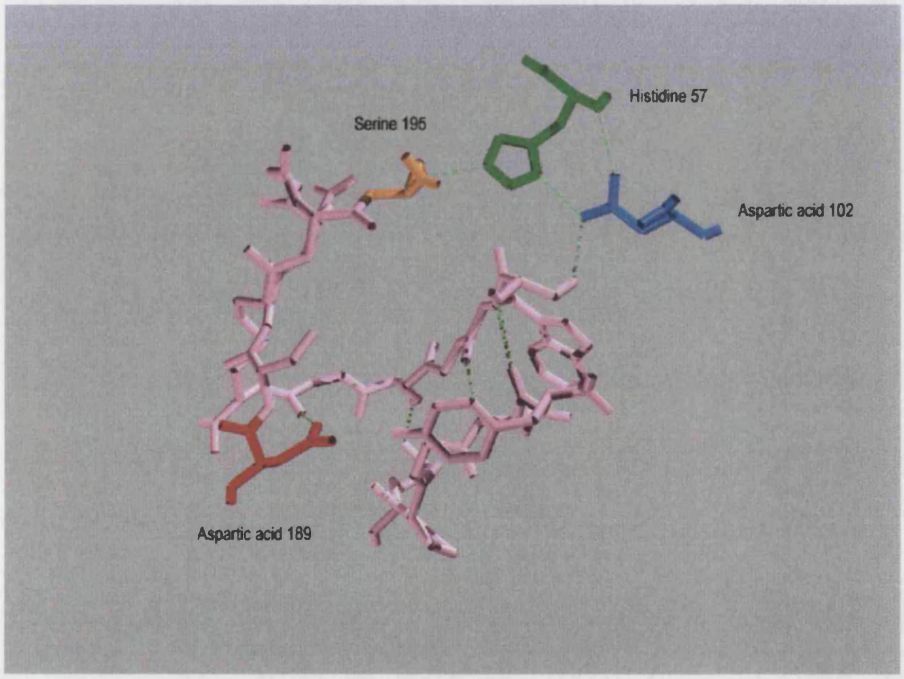



**Figure 1-9: Catalytic triad & substrate binding region of bovine trypsin**
Catalytic triad of amino acids (1.6.3, 1.6.6): Histidine (position) 57 shown in green; Aspartic acid 102. blue; Serine 195, orange. Primary specificity determinant (S1): Aspartic acid 189, red; remainder of substrate binding pocket shown in pink. Hydrogen bonds depicted as dotted green lines. Model produced using Swiss-PDB viewer 3.7 and PovRay 3.5.

### 1.6.4 Protease nomenclature

By convention, the binding site for the polypeptide substrate of an endopeptidase consists of 'protease-binding' subsites, with the point of hydrolysis being at either the N- or C-terminus of the P1, depending on the enzyme in question. The complementary binding sites on the protease are termed 'substrate-binding sites'. Note that substrate binding is frequently more complex than a simplistic one-to-one relationship, involving multiple interactions, especially at the extremes of the specificity determining regions (Fersht, 1999).

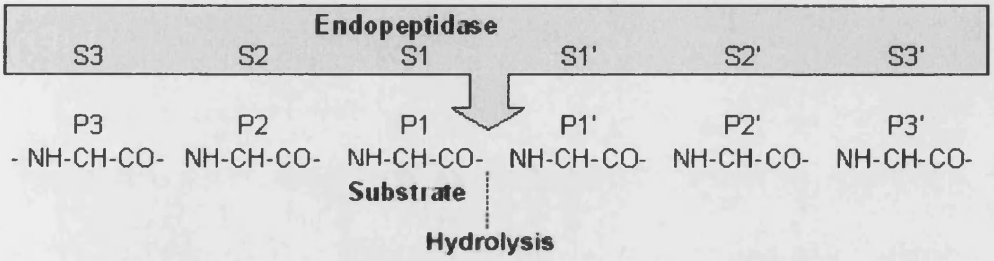



**Figure 1-10: Illustrating the Schechter & Berger numbering system for peptidase & substrate residues**

The scissile bond at which cleavage occurs is always located between residues P1-P1'. Cleavage site shown at C-terminus of P1 to illustrate action of trypsin.

### 1.6.5 Mechanism of substrate binding and catalysis

All serine proteases catalyse the hydrolysis of peptide or synthetic ester substrates by means of a conserved charge relay system of three amino acids and via the formation of an acyl-enzyme intermediate. The structures of the complexes involved have been particularly well characterised, in terms of atomic positions and thus bond strength, as it was shown to be possible to freeze the enzyme in each transition state by complexing with various natural small peptide inhibitors (Bode and Schwager, 1975). Co-crystallisation of these complexes revealed that the binding of inhibitors (and apparently also polypeptide substrates) did not distort the conformation of the enzyme, allowing for the refinement of highly accurate crystal structures. The active site of trypsin was complementary to the transition state of the reaction, a conformation that forces the proximity of the

Ser195 residue and the carbonyl carbon of the substrate (Bode and Schwager, 1975; Fersht, 1999).

The five steps of substrate binding & catalysis by trypsin are illustrated in Figure 1-11 & Figure 1-12 and are described in the sections below. The specificity of the binding pocket is the first influential effect in this process as a substrate of the correct topology associates with the enzyme to form the non-covalent Michaelis (or enzyme-substrate) complex, the two molecules being held together by mixtures of hydrophobic interactions, salt-bridges and hydrogen bonds (Fersht, 1999) (see Figure 1-11 panel 1). Crystal structures predict that, at this stage, a degree of 'steric strain' would exist between the P1' amino acid (indicated by R' on the figure) and the ser195 residue, acting to promote the distortion of the substrate chain and exposing the P1 carboxylic bond (Bode and Schwager, 1975).

After the physical binding of the substrate the reaction mechanism of the catalytic triad becomes crucial. For a full discussion of this please see Section 1.6.6. Steric distortion of the substrate results in the positioning of the oxygen atom of the 'targeted' carbonyl group between the backbone NH groups of Ser195 and Gly193 (Bode and Schwager, 1975; Copeland, 2000). The transition state forms by means of the attack of the hydroxyl of Ser195 on the carbonyl carbon. The resultant bond forces the reduction of the C=O bond to a single bond and the negative charge forced on the oxygen atom enhances the strength (and thus shortens the length) of the transient bonds then formed between this atom and the NH groups of Ser195 and Gly193 (see Figure 1-11 panel 2). It is the formation of these short hydrogen bonds which is thought to stabilise the high-energy transition state relative to the Michaelis complex (Fersht, 1999). The hydrogen atom transferred from Ser195 to His57 which allowed for the nucleophilic attack, now permits association of this NH group with the N'HR moiety, in preparation for bond scission.

The collapse of the tetrahedral intermediate to form the acyl-enzyme complex is triggered by the donation of the acquired hydrogen atom from His57 to the N'HR group and its replacement with a molecule of water (see Figure 1-11, panel 3). This completes the cleavage of the scissile bond and the leaving group (P1' onwards of substrate) is expelled from the acyl-enzyme complex. The bind energies of the P1'-S1', etc, interactions are apparently overwhelmed by the repulsive force generated by the proximity of the now complete amino group and the carbonyl carbon of the P1 residue (Bode and Schwager, 1975; Lesk, 2001).

Deacylation of the complex forms the second tetrahedral intermediate (See Figure 1-12, panel 1) occurs in a similar manner with the His57 residue activating its newly acquired water molecule so that the hydroxyl group performs a nucleophilic attack on the carbonyl carbon. The collapse of this intermediate state to the enzyme-product complex is caused by the donation of the hydroxyl group and the subsequent expulsion of Ser195 from the complex (See Figure 1-12, panel 2).

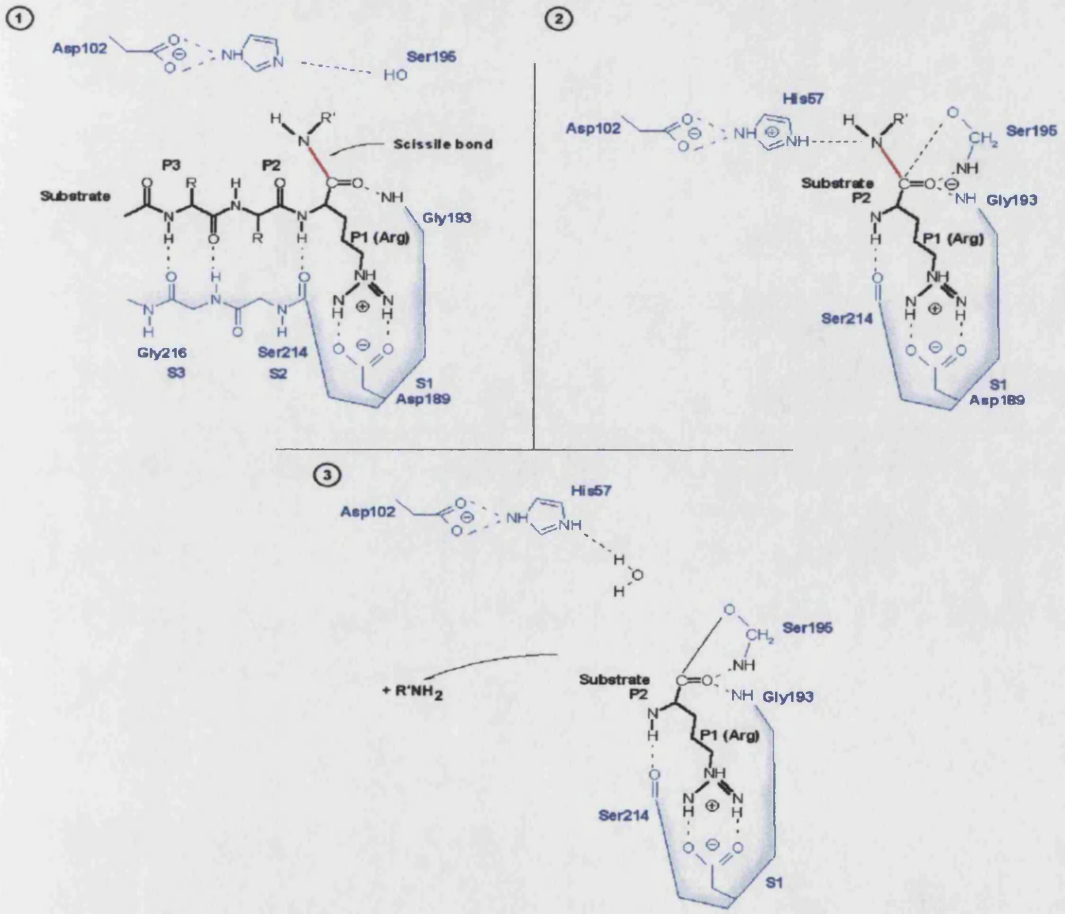



**Figure 1-11: Binding of substrate and formation of acyl-enzyme intermediate**

① Binding of substrate to binding pocket of enzyme, forming Michaelis complex; ② Transition state of complex following attack of the hydroxyl of Ser195 on the carbonyl carbon of the P1 substrate residue, 1st tetrahedral intermediate; ③ Collapse of tetrahedral intermediate to form acyl-enzyme and release of the leaving group ($R'NH_2$). Figure adapted from (Copeland, 2000; Fersht, 1999; Lesk, 2001).

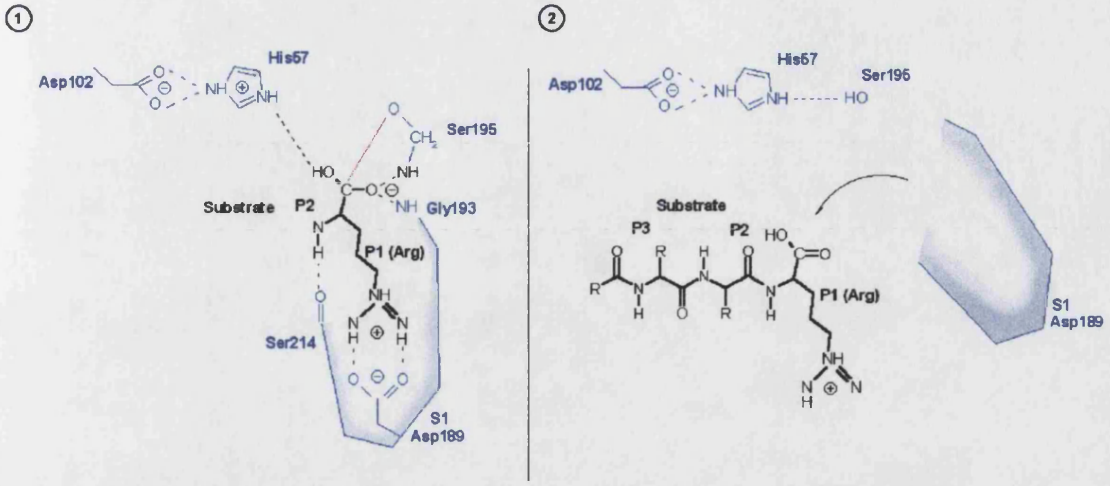



**Figure 1-12: Formation of 2nd tetrahedral intermediate and release of product**

① Hydrolysis of acyl-enzyme by nucleophilic attack by the hydroxy group of a water molecule bonded to His57, 2nd tetrahedral intermediate. ② Expulsion of Ser195 from complex, release of carboxylic acid product. ). Figure adapted from (Copeland, 2000; Fersht, 1999; Lesk, 2001).

### 1.6.6 Nature and action of catalytic triad

For a catalyst to successfully cause the hydrolysis of an amide bond, three chemical groups have been defined as essential (see Figure 1-13) (Price and Stevens, 1999):

1. A nucleophilic group, N, which attacks the carbonyl group on the P1 residue, leading to the formation of a tetrahedral intermediate.
2. A positively charged group, $X^+$, held near to the oxygen atom of the P1 carbonyl group. This acts to increase the susceptibility of the group to nucleophilic attack and may also help stabilise the tetrahedral intermediate.
3. A proton donor, Y-H, which would act on the -NH leaving group.

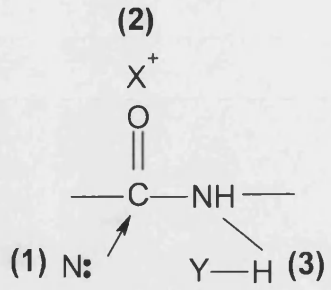



**Figure 1-13: Illustrating the essential chemical groups for a proteolytic catalyst.**

All three features listed above may be found within the catalytic triad of the serine proteases. Histidine is frequently found in the role of a catalytic residue as it imidazole side-chain (a tertiary amine) has a pKa of 6.5-7.4 when the amino acid is unattached, which changes to a pKa of 7 when it is incorporated into a protein. This means that histidine is the only amino acid capable of exchanging protons at a physiological pH. It is believed that the imidazole ring increases the nucleophilicity of the Ser195 (N:) hydroxyl group by acting as a general-base catalyst. This hisitidine residue thus represents the $X^+$ group (2) on figure 1.13. The importance of the Asp102-His57 hydrogen bond has not been conclusively established but is believed that it may be a low energy barrier hydrogen bond, which may possibly assist catalysis by helping to stabilise the transition state. By acting in this manner, Asp represents the Y-H group. An illustration of the interaction between the three amino acids using trypsin numbering can be seen in figure 1.14 below.

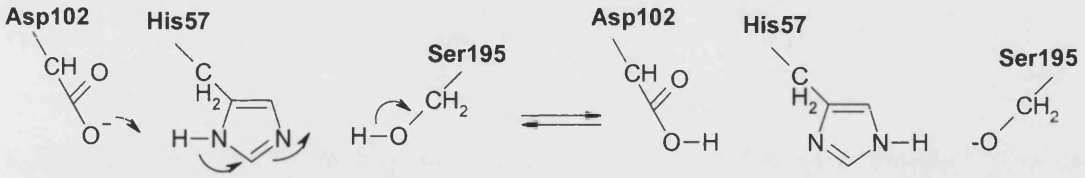



**Figure 1-14: Interactions between catalytic triad**

Illustrating the charge-relay system that exists between the catalytic triad of all serine proteases. This arrangement of the three side-chains allows for the partial ionization of the serine, resulting in a negatively charged oxygen atom that greatly enhances the nucleophilicity of the residue. Adapted from (Price and Stevens, 1999).

### 1.6.7 S & S' specificity of bovine trypsin

#### 1.6.7.1 Description of S subsite specificity

Trypsin specificity is primarily determined by the need of the S1 site at the bottom of the binding pocket for a basic amino acid, namely arginine or lysine (see Section 1.6.3), but it has been demonstrated that as for all proteases, the adjacent binding sites play significant roles in substrate hydrolysis. Past work has shown that they may also be modified to enhance specificity. Specificity in enzyme catalysis has been defined as "a consequence of optimal transition-state interactions made by cognate, but not non-cognate, substrates", the concept being that an enzyme binding pocket has evolved in such a way as to form sufficient electrostatic and apolar interactions with a desired substrate as to allow the energy barrier between basal and transition states to be overcome (Perona *et al.*, 1995). The energy released by substrate binding would then be turned to catalysis, the apparent rigidity of the binding pocket ensuring that little was lost in the process. Different portions of the binding pocket demonstrate varying degrees of specificity, raising questions as to the actual role of the components of the pocket and their interaction with each other and with more distal portions of the protein. The influence of the subsites has been probed through several studies, typically involving the calculation of the substrate-binding energy ($\Delta G_S$, calculated from $K_m$ values) and the activation energy ($\Delta G_T$, calculated from $k_{cat} / K_m$ values) from the digestion of libraries of peptide substrates. Recent work has shown that each subsite may characteristically favour either substrate binding or catalysis depending on whether it favours native substrate or transition state binding respectively (Marana *et al.*, 2001).

The S1 specificity of trypsin is well studied: chymotrypsin is the closest homologue, cutting by the same process, possessing an almost identical tertiary structure and preferring Phe, Tyr & Trp residues in the P1 site, and trypsin can discriminate between analogous substrates containing Lys & Phe by a factor $10^5$ (in terms of relative $k_{cat} / K_m$ values) (Hedstrom, 1996). Despite this, the specificity of the S & S' sites appears quite broad, allowing the binding of a wide range of residues while still greatly influencing the reaction kinetics.

#### 1.6.7.2 Description of S & S' subsite specificity

S2 & 3 Detailed work has not been performed in determining the preference of these subsites in bovine trypsin but binding energy studies of rat and cockroach trypsin present a picture of

relatively broad substrate acceptance, appropriate for an intestinal digestive enzyme (Marana *et al.*, 2001). The residues involved in binding the P2 & 3 sites on the substrate have long been identified as serine 214 and glycine 216, which form part of a loop that establishes a series of hydrophobic contacts with the substrate, anchoring it in position (Bode and Schwager, 1975). The P2 position appears to favour glutamic acid, lysine, leucine and threonine over other substrate amino acids by an approximate factor of two, while P3 demonstrates broad specificity overall, with a two to three fold preference for aspartic acid (Hedstrom, 1996; Marana *et al.*, 2001). In both rat and cockroach trypsin, P2 appeared to contribute more to substrate binding than catalysis, while for the P3 position, the reverse was true (Marana *et al.*, 2001).

S1' Molecular modelling studies have shown that two surface loops, loop 40 (residues 33-41) & loop 60 (residues 58-68) to be influential in determining S1' specificity, both in trypsin and chymotrypsin (Kurth *et al.*, 1998; Perona *et al.*, 1995). The side-chain of the basic amino acid Lys61 is believed to be the only group directly in contact with the P1' residue of the substrate. The S1' subsite has been shown to accept any aliphatic residues at P1' (possibly because the Lys side-chain would interact well with bulky, hydrophobic residues), aromatic residues and positively charged residues at least partially. Only the amino acid proline is not accepted in the P1' position, resulting in negligible hydrolysis, suggesting that a free NH group is a requirement.

S2' P2' appears to bind to the region of the binding pocket that is lined by His40 & Tyr151 (Schellenberger *et al.*, 1994). As the P2' side chain points in the opposite direction from the P1' & P3' chains, it has been shown that the pocket can accommodate fairly large and bulky residues at this position and the kcat/Km of the reaction can be seen to increase proportionately with the size of the residue. Another significant feature was that in all binding studies, a hydrogen bond was found to form between the NH of the P2' residue and the O of Phe41, once again ruling proline out of this position.

S3' Crystal structures suggest the P3' interacts with many of the same regions as P1' and the relationship between them has been clearly established by synthetic substrate studies. A large residue in either position (e.g. Tyr, Arg in P3') appears to 'crowd' the other, with favourable reaction kinetics only occurring if that position holds a small amino acid (e.g. Ala in P1') (Schellenberger *et al.*, 1994). It is believed that large amino acids in these

positions compete for contacts in the same region of the protein fold and that this explains the similarity between S'1 & S'3 specificities.

## 1.7 Production of r-trypsinogen

### 1.7.1 Current Eli Lilly methods for the production of recombinant bovine trypsinogen

Until recently, bovine trypsinogen, along with many other mammalian enzymes, was produced by extraction from animal organs, specifically cow pancreas. Ethical awareness, response to shareholder pressure and, crucially, the rise of uncharacterised new disease 'contaminants' such as bovine spongiform encephalopathies and genomic retroviruses, have raised concerns regarding the use of animal products in the pharmaceutical industry, with alternative production solutions being sought. Fermentation in a suitable microbial host provided the obvious solution, allowing for the reliable production of industrial quantities of product not derived form animal raw materials and also, in the case of trypsin, free of contaminating pancreatic enzymes, such as bovine chymotrypsinogen (Hanquier *et al.*, 2003).

The current production process for the variant of recombinant trypsinogen, and hence trypsin, used by Eli Lilly relies on its fermentation in *E. coli* and subsequent accumulation as inclusion bodies. The insoluble but concentrated zymogen (mature enzyme + pro-protein region) could then be separated from bacterial membranes and the cytosolic fraction by differential centrifugation (Kung *et al.*,1990). A mature trypsin is achieved by means of enzymatic modification with enterokinase (an endoprotease, recognition site of $[Asp]_4$-Lys↓) in the presence of suitable buffers. While this method of inclusion body isolation and solubilisation provides a precisely characterised and well-understood set-up for recombinant enzyme production, associated problems, particularly regarding the yield of native protein from the *in vitro* refolding reactions, mean that there may be economic advantages in developing a system for the expression of soluble enzyme. Process-scale refolding of the zymogen is initiated by drip-feeding concentrated, solubilised inclusion bodies (8M urea used as denaturing agent) into a refolding buffer containing an oxido-shuffling couplet (for controlled but dynamic redox conditions) and appropriate co-factors (for full protocols, please see Chapter 4) (Buswell *et al.*, 1995). Although precise recovery data may not be discussed here, internal company literature places the net yield of native (correctly folded enzyme, assessed by activity assay) protein at approximately 8-15% of the solubilised input.

Mis-folded protein issuing from the refolding vessel would be removed from the product stream by rounds of packed or expanded bed chromatography (Petrides, Sapidou, and Calandranis, 1995).

### 1.7.2 Expression of r-trypsinogen from *Pichia pastoris*

An academic collaboration was established five years ago (1998) in order to develop a trypsin production process in the methylotrophic yeast *Pichia pastoris*, with the aim of obtaining soluble and properly folded trypsin secreted into the growth medium. *P. pastoris* was an attractive target for development as progressive work over the last 10 years had permitted the construction of numerous host-vector systems for the heterologous protein expression of both prokaryotic and eukaryotic targets (Hanquier *et al.*, 2003). The presence of eukaryotic post-translational modification machinery and secretory pathways gives the yeast a relatively high success rate in the expression of soluble, correctly folded protein, frequently allowing the soluble expression of recombinant proteins that had proved intractably insoluble upon expression in *E. coli* due to a high number of disulphide bridges or other complicating factors (Venekei *et al.*, 1996). The strain has formed the basis of several economically-competitive industrial systems, with high cell densities being achieved in inexpensive media, and with no major complications regarding scale-up (Crueger and Crueger, 1990; Hanquier *et al.*, 2003). An additional reason for optimism over this approach was that an unrelated Californian group had engineered the soluble expression of rat trypsin from the strain *Saccharomyces cerevisiae* to levels of 10mg/L with no apparent complications (for a full sequence & structure comparison of bovine vs. rat trypsinogen, please see Section 1.7.4) (Venekei, Graf, and Rutter, 1996). The Lilly research group concentrated on the expression of the proprotein behind the strong methanol-inducible promoter, *aox1*, but only trace levels of the zymogen were detectable under a range of fermentation conditions, suggesting that that the expression of wild-type bovine trypsinogen was toxic to the cells. Resent work by the group allowed the apparent autotoxicity of the protease to be bypassed by means of engineering the proprotein activation site (pro-section comprises 8 N-terminal amino acids, terminating with a lysine residue, see Section 1.6.2) to prevent the auto-activation of the enzyme (Hanquier *et al.*, 2003). A single amino acid substitution proved sufficient, replacement of the position 8 lysine residue with asparagine removing the trypsin recognition site but still allowing controlled *in vitro* activation of the

enzyme by means of specific exopeptidases, and the yields of 40mg/L in a 6L fermentor were obtained with the now 'stable' zymogen (Hanquier *et al.*, 2003).

### 1.7.3 Autotoxicity issues associated with expression of mature proteases

A reported problem in the production of certain eukaryotic proteins in bacterial systems has been the toxicity of the proteins to the host, with the result of low expression rates and high levels of plasmid instability. While the expression of certain fusion proteins (e.g. TEM-β-lactamase) has been shown to severely poison an *E. coli* culture through an unknown mechanism, presumably interfering with cellular metabolism in a manner beyond simple accumulation of a recombinant protein (Mujacic *et al.*, 1999), the autotoxicity resulting from the high-level expression of an unregulated peptidase with relatively broad substrate range such as trypsin would appear easy to explain. The problem is exacerbated at high fermentation temperatures (~37°C) as, this the optimum temperature for trypsin activity, and also elicits the highest expression rates from many promoter systems. High levels of expression can act to 'saturate' translocation mechanisms and so cause protein intended for the periplasmic space to accumulate in the cytoplasm, enhancing any toxic effect and rapidly killing the culture (Perez-Perez *et al.*, 1994). Although the expression of trypsinogen should theoretically result in the accumulation of inactive zymogen, work in yeast cells has shown that autoactivation does occur, leading to progressively higher levels of active enzyme (Hanquier *et al.*,2003). This leads to the second issue in the expression of active enzyme, that of autodigestion, which adds to the already high levels of proteolytic sensitivity shown by soluble recombinant proteins in *E. coli* cytoplasm (Wickner *et al.*, 1991). Solutions must be determined empirically in each case, but proven experimental techniques include short periods (1-2hrs) of induction at low cell density followed by rapid harvesting, or by induction only at high cell densities at low temperatures (28°C) (Herber *et al.*, 1991) and use of a low-temperature *cspA*-promoter system for reduced expression levels (Mujacic, Cooper, and Baneyx, 1999). A general set of rules would appear to be that the expression of toxic and proteolytically-sensitive enzymes should be carried out at low temperatures, so as to reduce both protease activity and expression rates; use of low-level promoters, e.g. *trc* or *araBAD*; and tight control of promoter 'leakage', i.e. non-induced expression, as this can interfere with both plasmid stability and stock viability.

### 1.7.4 Comparison of bovine and rat trypsin

The structural basis of P1 binding were investigated in rat anionic trypsin by means of a strategy of nutritional selection and a series of rational oligonucleotides-directed mutations (Evnin *et al.*, 1990; Evnin and Craik, 1988). This involved the growth of an expression strain auxotrophic for arginine on a minimal media where all arginine was sequestered in nitroanilide compounds, sensitive only to hydrolysis by mammalian trypsin. The genetic selection thus relied on the ability of an engineered *E. coli* strain to secrete soluble, active rat trypsin into the periplasm. The vector used was a pBR322 derivative which expressed recombinant rat trypsin under the control of a phoA promoter, and studies had shown that it appeared to secrete active, mature protease to a concentration of 1mg/L (Evnin and Craik, 1988). This successful work, combined with the knowledge that teams from Eli Lilly had experimented unsuccessfully with a similar set-up, made it interesting to compare the structures of bovine and rat trypsin, to see if any blatant structural differences were evident which could provide an explanation for the differing solubilities of these closely related enzymes.

Bovine and rat trypsin display an 87% amino acid primary sequence identity, with the majority of these differences being at least 20% exposed to solution. Of these, the most significant would appear to be the substitution of a fully surface-exposed valine in rat trypsin for phenylalanine at postion 181 (bovine trypsin numbering) and similarly the substitution of an exposed glutamic acid for a tyrosine at position 151. While the first substitution is seemingly mirrored by a downstream (bovine)Y171F(rat) exchange and rat trypsin actually possesses a greater share of bulky, hydrophobic amino acids, these are all orientated to project their chains into the molecule. While not of any guaranteed use, these observations may provide a possible route for the rational engineering of bovine trypsin should problems with insolubility appear insurmountable by other means.

## 1.8 Evaluation of biosynthetic human insulin product with reference to potential improvements in enzymic conversions

### 1.8.1 Introduction

The following section details the current use of r-trypsin by Eli Lilly as a biocatalytic agent in the processing of mature human insulin, with reference to the potential economic gains that could be realised if the specificity of the enzyme could be tailored through directed evolution.

Diabetes mellitius, and its complications, is estimated to be the third largest cause of death in industrialised countries after cardiovascular diseases and cancer (Barfoed, 1987). Documented occurrences of the disease date back as far as the fifteenth century BC, and insulin, a mammalian hormone produced by the Islets of Langerhans in the pancreas, was first described in 1909, but it was not until 1922 that the link between the two became apparent. Insulin mediates the metabolism of carbohydrates and is thus essential for the flow of energy through an organism; impaired production cripples the process and leads to the formation of the disease.

As a therapeutically important and well-characterised small protein, insulin was an obvious target to be the first mammalian protein produced in bacteria using recombinant DNA technology. Two major techniques for the production of recombinant insulin in E. coli (the two-chain and the proinsulin methods) and a more recent yeast expression system have been developed with only the proinsulin prokaryotic method & the yeast system being commercialised to date (Petrides, Sapidou, and Calandranis, 1995). The three methods are detailed below.

1) Two-Chain Method: *Developed by Genentech. Inc. and scaled up by Eli Lilly and Co.* The two insulin chains are produced separately as β-galactosidase fusion proteins within inclusion bodies, recombined after purification.
2) Proinsulin method (intracellular): *Developed & commercialised by Eli Lilly and Co.* Relies on a single fermentation for the production of the inactive Trp-LE'-Met-Proinsulin, which is purified from inclusion bodies and subjected to cleavage and refolding treatments.
3) Proinsulin method (secreted): *Developed by Novo Nordisk A/S.* Uses genetically engineered yeast cells to secrete BHI as a single chain insulin precursor.

The proinsulin production was based on the expression of a precursor of proinsulin, a Trp-E fusion protein, and its subsequent digestion & activation by three additional enzymes, namely trypsin, carboxypeptidase I & a diamino-exopeptidase. The fermentation and enzymic conversion of the insulin molecule accounted for only half of the total industrial production process, with the remainder consisting of an extensive purification scheme based on multimodal chromatography. The extreme degree of purity required in a therapeutic product before it can progress to formulation, dictated that insulin-like products resulting from mis- or partial cleavage of the proprotein be reduced to trace levels. The nature of the enzymes used meant that these insulin-like products came in two main varieties, those produced by cleavage at undesirable points in the proprotein, which differed from the product by 8-10 amino acids and thus acted to reduce the yield of the process, and those which differed by only a single amino-acid residue and were considerably harder to separate as a result. The ability to tailor and tighten the substrate specificity of these modifying enzymes could improve the process by rationally minimising the number of undesirable cleavage products and thus improving the ratio of proinsulin to insulin conversion, and also by reducing the levels of near-identical products. An increase in the purity of the post-enzymic conversion product stream, particularly regarding the levels of single amino-acid variants, would allow for increased chromatographic yields (Kroeff *et al.*, 1989; Ngiam *et al.*, 2001), possibly leading to the simplification of the process.

### 1.8.2 Structure of insulin

Mature insulin is an extremely small protein consisting of 51 amino acids arranged in two chains: A, with 21 amino acids; and B, 30 amino acids in length. In the native host, these chains are cut from the proinsulin precursor, resulting in the loss of the dividing region, the 34 amino acid C chain. The A & B chains are held in position relative to each other by two disulphide bonds and are simple in structure, consisting of two short alpha helixes and an internal disulphide bond, and a single alpha helix respectively (see Figure 1-15). In 1996, Eli Lilly and Co. made use of recent research to improve the recombinant insulin into a more therapeutically potent variant. The exchanging of positions of pro28 & lys29 rendered aggregations of the molecule (now termed the KPB variant) considerably less stable. The individual insulin molecules of regular insulin clump together to form a hexamer. Only individual insulin molecules are biologically active, so the body must first break

the bonds that hold the six insulin molecules together. Individual insulin molecules become available in about 30 minutes. The amino acid rearrangement in KPB results in hexamers that bind together so weakly that they break apart much faster than regular insulin, and are thus active immediately. The KPB variant was alternatively called insulin lispro, and was marketed under the brand name of Humalog®.

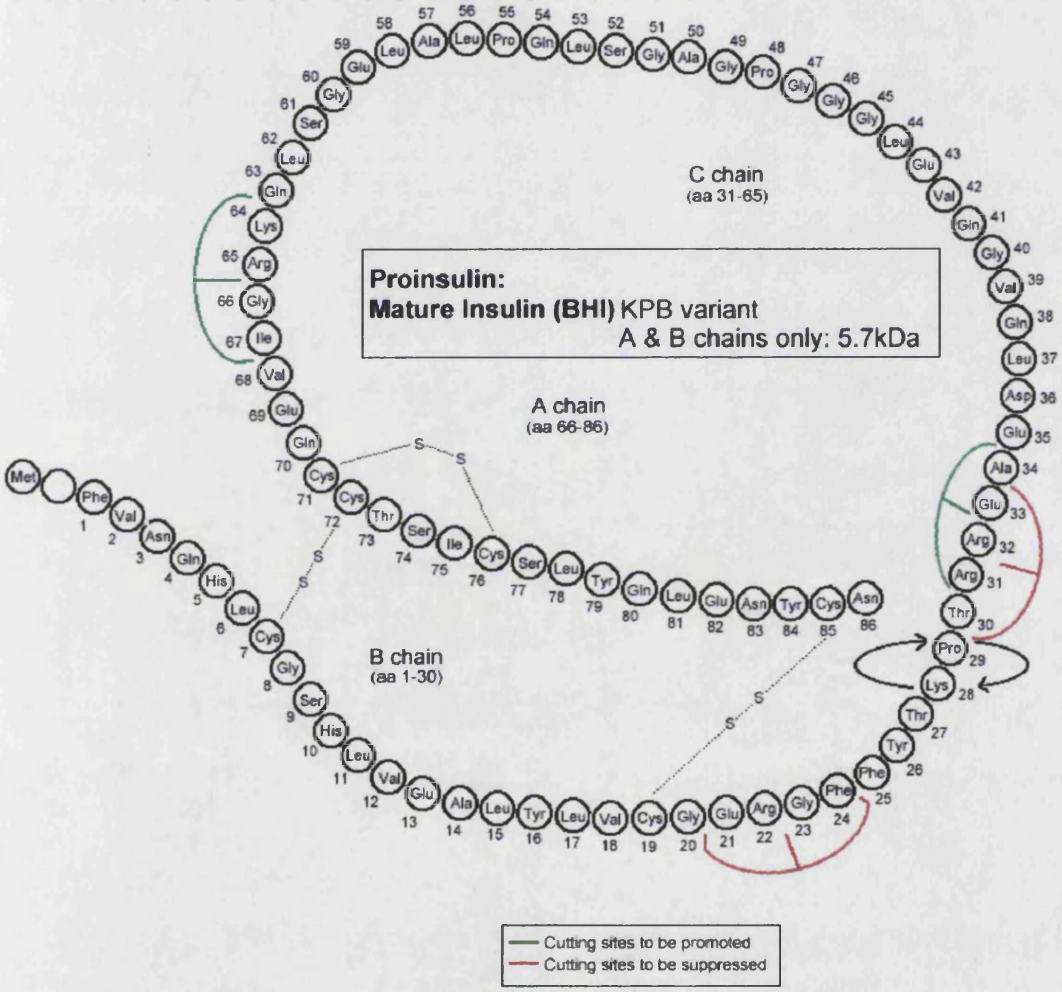



**Figure 1-15: Structure of the K28P proinsulin molecule (PDB ref: 1LPH)**

Cleavage of C-chain at amino acids 31 & 65 releases the A & B-chains which associate with each other to form mature BHI (biosynthetic human insulin) (PDB ref: 1LPH). The arrows at lys28 & pro29 denote where these two amino acids have been exchanged, swapping their positions from the native human insulin molecule to form the KPB variant. The legend indicating cutting sites to be either promoted or suppressed refer to specificity targets described in Section 1.8.5.

### 1.8.3 Description of Two-chain and yeast BHI production methods

The two-chain method was the first successful process for the synthesis of BHI, being developed by Genentech Inc. (San Francisco, CA) and subsequently scaled-up by Eli Lilly and Co. (Indianapolis, IN) (Petrides, Sapidou, and Calandranis, 1995). The A and the B chain were expressed as β-galactosidase fusion proteins within two separate recombinant strains of *E. coli*, under the control of the Trp operon. The two polypeptides accumulated as inclusion bodies in individual fermentations, were purified, and then combined in the appropriate redox conditions to yield mature human insulin. The principal drawback to this method was the obvious need for two parallel fermentations and primary purification runs.

The secretion of proinsulin from cultures of *Saccharomyces cerevisiae* was engineered by Novo Nordisk A/S. A single-chain precursor was secreted and purified, and then converted to mature insulin by means of a transpeptidation reaction in organic solvent in the presence of trypsin and a threonine ester (Petrides, Sapidou, and Calandranis, 1995). The product then underwent de-esterification and further purification. The clear advantage of this method was the unassisted solubility of the product, the reduced levels of host contamination and the option to reuse the cells by means of a continuous bioreactor-cell separator loop.

### 1.8.4 Description of the intracellular proinsulin method of BHI production

Eli Lilly currently produces BHI by means of expression of the product as a single proinsulin-fusion protein within an *E. coli* fermentation. An overview of this process is presented in Figure 1-16. The overexpressed product, a Trp-LE'-Met-proinsulin fusion protein accumulates in the cellular cytoplasm as inclusion bodies over the course of an 18hr fermentation, eventually comprising a level no less than 20% of the total dry cell mass, and with each inclusion body consisting of approximately 80% product (Petrides, Sapidou, and Calandranis, 1995). After cell distruption, the inclusion bodies are fully solubilised in a chaotrophic, reducing environment containing urea & 2-mercaptoehtanol. The Trp-LE'-Met signal sequence of the chimeric protein would be cleaved by the addition of CNBr (cyanogens bromide), yielding the denatured mature insulin. The proprotein is then fully unfolded by a process of oxidative sulfitolysis, which acts to break all disulphide bonds within the molecule that may have reformed since the initial denaturation, and adds $SO_3$ moieties to

all the sulphur residues on the cysteines, effectively preventing any uncontrolled dimerisation. Refolding of the proinsulin is accomplished at an undisclosed dilution factor at 8°C in the presence of mercaptoethanol, which acts to create a reducing environment, allowing the interchange of disulphide bond partners and thus the exchange of the $SO_3$ groups for the relevant cysteines (Winter *et al.*, 2002). The purified and native-state proinsulin would then be activated by means of an enzymatic reaction involving trypsin & carboxypeptidase B. Trypsin being used to create mid-sequence cuts at the C-terminii of arginine & lysine residues, and the exopeptidase carboxypeptidase B being then used to remove basic amino acids with exposed 3' termini, the result being active and structurally-native human insulin.

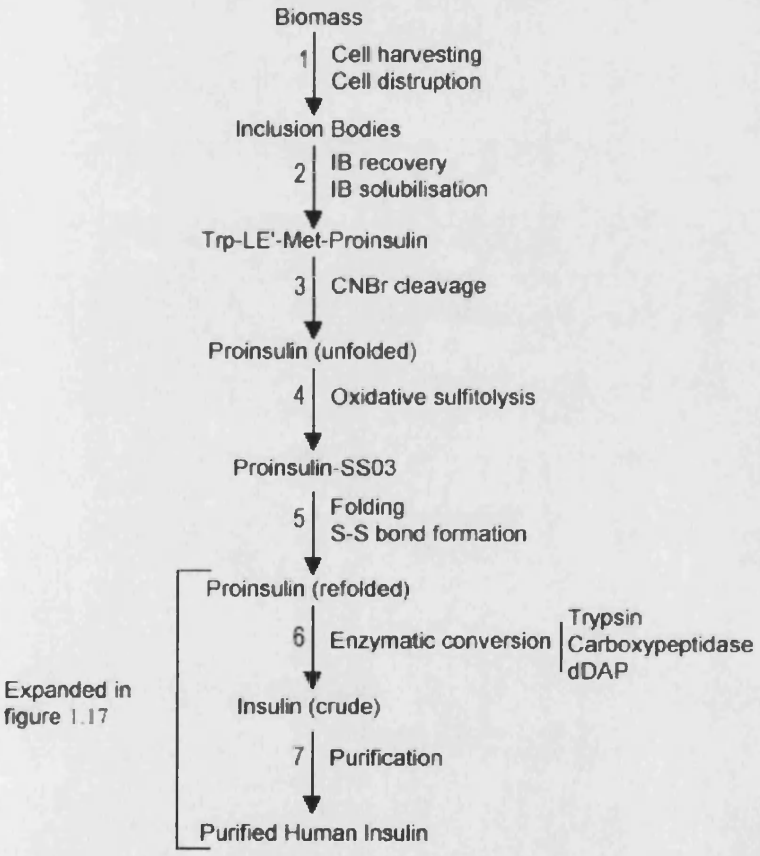



**Figure 1-16: Industrial production of intracellular BHI**

Schematic illustrating the current Eli Lilly process for the large scale production of BHI within *E. coli* by means of the fermentation and subsequent processing of an intracellular proinsulin molecule.

### 1.8.5 Proposed improvements to *E. coli* intracellular production process by means of specificity engineering of trypsin

1) Examination of the insulin molecule (see Figure 1-15) revealed that two particular species of contaminant are produced by trypsin digestion of proinsulin, these are a) an 8-aa peptide cut from the B-chain by undesirable cleavage at arg22/gly23; and b) the possibility that it might cleave at lys64/arg65 faster than at arg65/arg66, resulting in an A-chain 1aa too long, essentially reducing this fragment to an insulin-like contaminant. Eli Lilly's proposed that the following substrate specificities be tightened in trypsin to the exclusion of others. Certain specificities were to be evolved against, minimising cutting at these points. In all cases, a defined motif of four amino acids was stated, as the final goal of this project was to develop a system for the mutation of trypsin which would take advantage of its known substrate preferences at residues P2, P1' & P2' and breed highly selective cutting tools. To be improved: arg31-arg32/glu33-ala34 & lys64-arg65/gly66-phe67 (improve 'clean' release of A & B chains); to be reduced: glu21-arg22/gly23-phe24 (prevent unwanted cutting 2/3rds of the way through the B chain), thr30-arg31/arg32-glu33 and thr27-lys28/pro29-thr30(promote 'clean release of B chain).

2) With the use of directed evolution to explore the specificity of trypsin, it may be possible to dispense with the use of the enzyme carboxypeptidase, resulting in a significant lowering of the process costs. This would require the evolution of the following substrate specificity: pro29-thr30/arg31-arg32.

The feasibility and significance of these two goals will be examined in this chapter with reference t both the plausibility of engineering the required mutants and also to the impact that each alteration would have on the efficiency and process costs of BHI production.

### 1.8.6 Simulation of large-scale BHI production

The BHI fermentation process currently in use by Eli Lilly and Co. (see Figure 1-16) was analysed using the SuperProDesigner software, version 5.0 (Intelligen, Inc.)(Petrides, Sapidou, and Calandranis, 1995). This software allows a dynamic analysis of the progressive steps in a large-scale industrial biosynthetic process, modelling the efficiency of the reactions in each fermentation,

conversion or purification unit operation, and calculating an economic evaluation based on the yield and the pre-programmed operating costs and market prices. The complete BHI production process was modelled by the designer using published developmental literature from Eli Lilly and utilises cost & sales figures correct as of 2000 (Petrides, Sapidou, and Calandranis, 1995)(On-line documentation at www.intelligen.com).

The portion of the process which was of interest to this work, was that immediately downstream from the enzymatic conversion vessel, where proinsulin was converted to insulin plus 'protein contaminants' by the combined action trypsin and carboxypeptidase B (Figure 1-17). The program uses a set stoichiometric reaction to simulate the digestion of proinsulin, and then the crude protein contaminants and insulin-like molecules are progressively removed from the product stream by means of two packed-bed chromatography steps and a gel filtration. The levels of protein contamination entering the conversion vessel (diafiltered product stream from an HIC column) were as shown in Table 1.5. The simulation works on the assumption that proinsulin was converted into mature insulin and 'contaminating proteins' (namely the C-chain) in a ratio of 0.61:0.39 respectively, figures which exactly correlate with the size of the C-chain in proportion to the A & the B-chains combined. In other words, the model predicts that the enzymatic conversion acts at 100% efficiency and that all proinsulin fed into the vessel is converted into precisely cut A, B & C-chains. Insulin-like contaminants are factored to emerge from inclusion body processing and cellular disruption, but not from the marginal mis-cutting of the proinsulin by the modifying enzymes, resulting in chains differing in one or two amino acids. An attempt was made to introduce the insulin-like contaminants proposed in section 1.8.5 into the model in a variety of concentrations, evaluating their potential economic impact on the process and their significance on the downstream purification elements.

### 1.8.7 Economic analysis of a standard BHI production run

An economic analysis of the unaltered BHI production run was rendered using SuperPro Designer, version 5.0. In its entirety, the process consists of 30 principle units of operation, a summary of which was presented in Figure 1-16. In the enzymatic conversion step, proinsulin is converted to insulin and additional 'contaminating proteins' (i.e. the C-peptide). The general levels of contaminants present in the product stream as it is fed into the vessel are listed in Table 1.5. The

released C-peptide is added to the contaminating proteins category, which must be considered a fully heterologous fraction, containing as it does, insulin-like proteins, polypeptides of only a third the molecular weight, like the C-chain, and assorted low-weight degenerative products. Following enzymatic digestion of proinsulin, the majority of these proteins, including those contaminants that differed from mature insulin by more than a few amino acids and the bulk of the remaining proinsulin species, are removed by S-sepharose chromatography. After concentration of the feed by diafiltration, a reversed-phase chromatography step is used to accurately purify BHI from the remaining homologous solution of insulin-like contaminants. The key points of the economic report generated by the model are shown in Table 1.6.

| Substance | Flow Rate (kg/batch) | Conc. (g/L) | Origin |
|---|---|---|---|
| Cont Proteins | 0.321 | 0.0738 | At this stage, mainly inclusion body contaminants & degraded protein from CNBr treatment. |
| Insulin | 0.000 | 0.00 | |
| Proinsulin | 30.5 | 7.03 | Product from refolding vessel. |
| Denatured Proinsulin | 0.222 | 0.0511 | Majority is denatured protein from CNBr step which failed to respond to sulfitolysis. |
| Proinsulin-SSO3 | 1.01 | 0.232 | Proinsulin which failed to exchange –SO3 moiety during refolding reaction. |
| Trp-Proinsulin | 0.0188 | 0.00434 | Fusion-tagged proinsulin that failed to respond to CNBr treatment. |

**Table 1.5: The protein composition of the product stream entering the enzymic conversion vessel**

The process steps immediately up-stream comprised of an HIC column & a subsequent diafiltration. Additional elements in the product stream, not shown in the table were: acetic acid; sodium chloride & ultra-pure water.

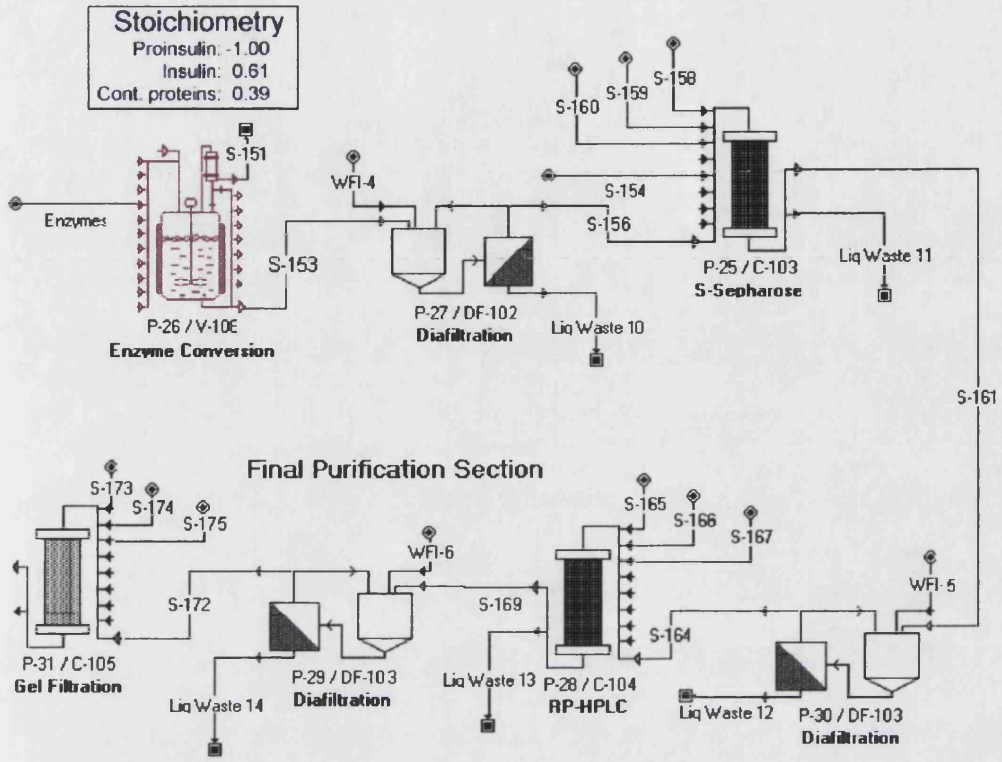



**Figure 1-17: Process steps in BHI purification**

Schematic of the enzymatic conversion step and down-stream chromatography elements from the Eli Lilly BHI production process. Produced using SuperPro Designer software, version 5.0 (Intelligen, Inc.).

| | |
|---|---|
| Total Capital Investment | 79410000 $ |
| Operating Cost | 73970000 $/year |
| Production Rate | 1810 kg/year of Product |
| Unit Production Cost | 40875 $/kg of Product |
| Total Revenues | 135697000 $/year |

**Table 1.6: Summary of the economic evaluation report generated from the current BHI production process (year 2000 prices)**

## 1.9 Alteration of enzymic conversion efficiency

As discussed in the previous section, the simulation assumes 100% efficiency in the enzymatic conversion step, with the complete mass of proinsulin being converted into BHI and C-peptide.

Figure 1-18 illustrates the potential economic impact of factoring in the emergence of the mis-cutting contaminant species proposed in section 1.8.5 in increasing quantities. The two types of contaminants predicted were a highly insulin-like product, differing only in the presence of an additional arginine at position 65, and a second two-chain product, but where the B-chain was truncated from arginine 32 onwards. For the purposes of Figure 1-18, both types have been classed together as 'Cutting-Byproduct 1' (CB1), incorrectly modified BHI molecules which lower the product yield without impacting on the quantity of C-peptide released.

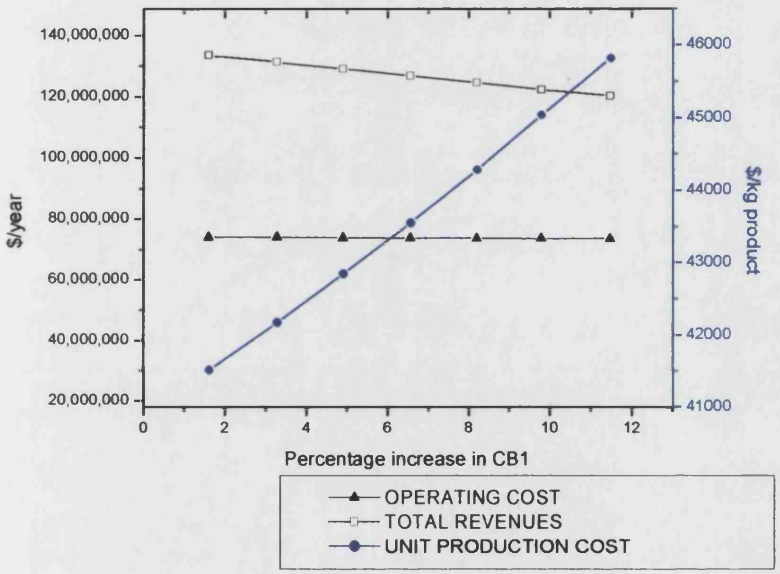



**Figure 1-18: Economic impact of trypsin mis-cutting**

Illustrating the economic impact of increasing the rate of trypsin mis-cutting of insulin into insulin-like contaminants differing by only 1 amino acid or chemical group Cutting Byproduct 1). The percentage scale denotes the percentage of insulin (cut from proinsulin, excluding mass of C-chain), which was considered to now be contaminant.

The simulation of the BHI production process showed that the production (per batch & per year) fell in a strictly proportional manner with the increase in the percentage of insulin forming CB1. For example, a 4% drop in insulin yield in the conversion step resulted in a final yield of 1737.6 kg/year, a 4% drop from the original.

### 1.9.1 Expanded project goals

In light of the project aim, which was, in short, to explore methods by which r-trypsinogen could be mutated for enhanced substrate specificities, the following technical goals were decided upon. The end result would be a practical and flexible strategy of directed evolution for the rapid adjustment of enzyme specificity, capable of being performed at a relatively small, laboratory scale while delivering an acceptable probability of success.

1. The initial part of the work would focus on developing an 'activatable', soluble trypsinogen in the periplasm of *E. coli*, as this would be essential for later experiments. This could possibly be achieved through the manipulation of the fermentation conditions of the Lilly production strain ELTRP-1 or through the use of an alternative strain. A proven strategy would be the attachment of a signal sequence into the plasmid at the N-terminus of the gene in order to allow the transport of the protein into the periplasm (Evnin and Craik, 1988). Mutagenic PCR would be used to alter the primary sequence if necessary to enhance solubility. See Chapters 3 & 4 for the majority of this work.
2. Rapid results could only be generated through the use of microwell scale assays, which would allow the performance of high-efficiency assays in screening for a variety of proteolytic activities. The detection test for such assays could be based on the principle of a spectrophotometric assay using various peptide-pNA substrates (cf. BAPNA/BANA/ TAME assays) and on the cleavage of non-nutritious substrates (e.g. arg-β-naphthylamide and lys-β-naphthylamide), proteolysis resulting in the release of the free amino-acid. Chapter 2 details high-throughput assay development, while the practicalities of nutritional selection are discussed in Chapter 6.
3. Mutagenesis would then be performed in *E. coli* with the aim of producing variants with increased site specificity for unique primary sequence motifs, e.g. RR, KR, or other sequences (of up to approximately 4-5 residues in length) to be determined. Techniques used would primarily involve mutagenic PCR (epPCR) or use of an *E. coli* mutator strain to generate the initial generations of mutants and then subsequent DNA-shuffling. The efficiency, practicality and mutational bias of all methods used would then be characterised. Mutational work is presented in Chapter 6.
4. Results generated would be used to construct a mutant library consisting of banks of recombinant trypsin genes and related proteases. The challenge would then be to devise a

complementary library of assays capable of screening for an increase in specificity (as opposed to simple activity) in a high-throughput manner. Library theory, design and synthesis methodologies comprise Appendix A.

# 2 High-throughput assay development

## 2.1 Introduction

### 2.1.1 General strategy

As discussed in Section 1.7.1, the current Eli Lilly commercial production procedure for recombinant bovine trypsinogen involves the solubilisation and refolding of the enzyme from insoluble inclusion bodies. Standard activity assay were available for trypsin purified in this manner, but an effective strategy of directed evolution demanded that soluble protein be assayable directly upon expression from a culture, so as to limit the processing steps required in the screen and thus enhance the practical maximum size of the library. Cloning work with the aim of achieving soluble enzyme from microplate cultures can be found in Chapters 5 & 6. An activity assay therefore needed to be developed which could reliably detect low levels of trypsin activity in the presence of live culture, cellular lysate and lysis agent.

A practical screening strategy for directed evolution should rely on assays that can be performed in microtitre plates (96 or 384 well) because of the large numbers of clones involved. Direct assays such as fluorimetric, colour change or turbidity measurements should be used if possible because of the ease of reading. A proven screen for the degree of trypsin activity involves digestion of 0.4mM *t*-Boc-Leu-Gly-Arg/Lys *p*-nitroanilide and then monitoring the release of *p*-nitroanilide at 405 nm (Evnin *et al.*, 1990). The activity assay required by Eli Lilly involves the use of p-toluenesulphonyl-L-arginine methyl ester (TAME). This substrate is known to have a high affinity for trypsin and involves monitoring the release of the ethyl ester @ 247nm. The TAME assay was to be used for this work in parallel with any others developed, principally so that the activity of variants generated could be directly compared against Eli Lilly batches, as this was the standard test used internally by Eli Lilly to assess the activity of their product. Trypsin variants generated by mutagenesis would have their specific activity rate on this substrate measured after promising isolates had been identified & characterized, but the method is impractical for a high-throughput screen because of the ultraviolet wavelength required (to which all acryl-plastic plates would be opaque). A gel overlay technique can be used to measure cleavage of *tos*-arginine methyl ester and *tos*-lysyl methyl ester substrates (Perona *et al.*, 1993)

The aim was to develop a basic assay for the activity of a trypsinogen protein expressed, in varying degrees of solubility, within the cytoplasm or periplasm of an *E. coli* cell and the assay would thus require the lysis of the cells and the possible purification of the soluble protein from the lysate. Activation of the zymogen would also be necessary and this could be accomplished either by addition of trace quantities of commercial trypsin (allowing gradual autoproteolysis and thus autoactivation) or via addition of the more efficient enterokinase (secreted by mammalian pancreatic cells for the activation of trypsinogen as it passes into the gut). As this assay should be suitable for use on thousands of clones, the key aspect was that it contained the minimum number of process steps, so as to speed up the generation of results and limit (fatigued) operator error. The initial gross activity assay would be based on the ability of the enzyme to catalyse the release of a chromogenic group from a single amino acid substrate containing only one scissile bond. Spectroscopic measurement would yield readings as to the kinetics of substrate cleavage and thus the activity of the variant. It was important to consider that any increase in this gross activity over the wild-type could be due to a number of factors namely: increase in cytoplasmic solubility; enhancement in periplasmic translocation (if targeted); increase in expression rate induced by transformation; or increase in enzyme activity, i.e. catalytically significant mutation in binding pocket or surface loop.

### 2.1.1 Choice of substrate

In a synthetic substrate for protease cleavage, the amino acyl residue (or peptidyl residues if a peptide is used) determines selectivity and specificity. In a single residue substrate, the substrate specificity screened for would be relatively broad, i.e. any protease capable of cleaving the bond at the carboxy terminus of the designated residue would release the marker. As explained before (see Section 1.4), trypsin is an endopeptidase and the design of a simple artificial substrate should naturally reflect this. The single residue (be it Arg or Lys) must be blocked at the N-terminus and with a suitable chromogenic group bound to the C-terminus.

The choice of substrates was restricted to those liberating a chromogenic residue upon enzymatic hydrolysis as equipment restrictions precluded the use of fluorogenic substrates. This was not considered a handicap as a particularly high-sensitivity assay was not necessary in this instance and might even have required the dilution of the expressed enzyme, forcing more process

steps. It is worth noting that while fluorogenic detection is, on average, many times more sensitive than spectrophotometric measurement, the relative sensitivity (rates of autohydrolysis when experiencing minor environmental fluctuations, e.g. temperature or pH) of certain substrates may at times negate this. For example, the extreme sensitivity of the thioester bond to enzymatic hydrolysis is such that lower concentrations of elastase have been assayed with a peptidyl-thioester substrate than with a fluorescent group (Keasling *et al.*, 2002).

Trypsin has the highest affinity for thiobenzylester (-SBzl) and 4-nitroanilide (-$NHPhNO_2$) substrates (see Table 2.1).

| Substrate | pH | $k_{cat}$($s^{-1}$) | $K_m$(μM) | $k_{cat}/K_m$ ($mM^{-1}s^{-1}$) |
|---|---|---|---|---|
| Z-Lys-SBzl | 8.0 | 75 | 50 | 1500 |
| Z-Arg-SBzl | 7.5 | 94 | 5.3 | 18000 |
| Tos-Gly-Pro-Arg-$NHPhNO_2$ | 7.5 | 69 | 17 | 4000 |

**Table 2.1: Showing three examples of chromogenic peptide substrates for trypsin**
Abbreviations: Z- = benzyloxycarbonyl-; -sBZl = -thiobenzylester; Tos- = tosyl-; -$NHPhNO_2$ = -4-nitroanilide. Adapted from (Sterchi and Stocker, 1999).

Previous work involving the molecular engineering of trypsin had used –nitroanilide substrates due to their low cost and relative stability. Thioester substrates, when hydrolysed, are released as free benzyl thiols which must react with either 5,5'-dithiobis(2-nitrobenzoicacid) (Ellman's reagent) or 4,4'-dithiopyridine to generate a signal at 412nm (Grassetti and Murray, 1967; Sterchi and Stocker, 1999). Nitroanilide substrates need no additional reagents to produce a signal and the freed nitroaniline group can be read directly at 410nm. If the resulting yellow colour is not sufficiently distinguishable from the background then the released chromophore may be diazotized and coupled to *N*-(1-naphthyl)-ethylene diamine to generate a red azo dye which may be read at 546nm (Sterchi and Stocker, 1999). This may be useful if the natural colour of the media (LB in particular) masks the signal.

### 2.1.2 Design of single peptide substrate

Endoprotease activity is dependant upon the absence of any free amino- or carboxy-termini and thus, assuming that the marker group is affixed to the C-terminal end of the target amino acid, a suitable $\alpha$-amino-terminal blocking group must be chosen unless the synthetic peptides extends for a least 5 amino acids up-stream of the cutting site (Corey *et al.*, 1995). The choice of blocking group is dependant on the aim of the assay as studies have shown them to have a pronounced effect on the sensitivity of the substrate to hydrolysis (see Table 2.2, Figure 2-1 for examples) (Somorin *et al.*, 1978a; Somorin *et al.*, 1978b).

Commonly used N-terminal blocking (protection) groups for Arg & Lys

| Name | Abbrev. | Formula | Residue wt. | Ref. |
|---|---|---|---|---|
| Acetyl | Ac | $C_2H_3O$ | 43.046 | (Somorin, Tokura, Nishi, and Noguchi 1978b) |
| t-Butoxycarbonyl | Boc | $C_4H_9O_2$ | 101.126 | (Perona, Evnin, and Craik 1993) |
| Benzoyl | B | $C_7H_5O_9$ | 105.117 | (Novabiochem 2000) |
| Benzyloxycarbonyl | Z | $C_8H_7O_2$ | 135.144 | (Somorin, Nishi, and Noguchi 1978a) |

**Table 2.2: Blocking groups**

Terminology and chemical structure of commonly used blocking groups for arginine and lysine containing substrates.

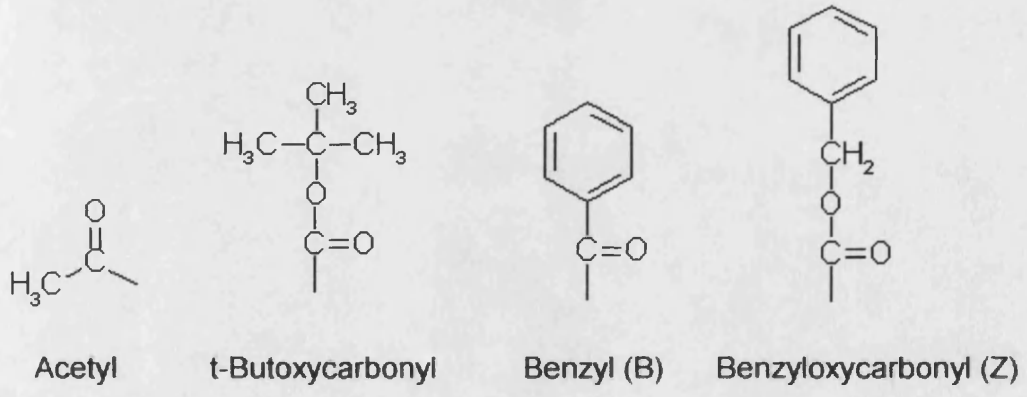



**Figure 2-1: Structure of blocking groups**

Illustrating the biochemical features of commonly used blocking groups for arginine & lysine containing substrates.

The relative rates of trypsin-catalysed hydrolysis of arginine-*p*-nitroanilide substrates, each incorporating a different α-amino-terminal blocking group can be seen below (Table 2.3).

| | Substrate | Relative rate |
|---|---|---|
| 1 | DL-BAPNA | 1.0 |
| 2 | L-BAPNA | 3.4 |
| 3 | L-AAPNA | 9.0 |
| 4 | L-ZAPNA | 10.0 |
| 5 | L-BPVAPNA | 1177 |
| 6 | L-ZPVAPNA | 1200 |

**Table 2.3: Describing how blocking groups can affect the relative cleavage rates of substrates**

Adapted from (Somorin *et al.*, 1978b). Abbreviations: 'DL' refers to the fact that racemization was allowed to occur during substrate synthesis. DL-BAPNA = *N*-benzoyl-DL-arginine-*p*-nitroanilide, L-AAPNA = *N*-acetyl-L-arginine-*p*-nitroanilide, L-BAPNA = *N*-benzoyl-L-arginine-*p*-nitroanilide, L-ZAPNA = *N*-benzyloxycarbonyl-L-arginine-*p*-nitroanilide, L-BPVAPA = *N*-benzoyl-L-phenylalanine-L-valine-L-arginine-*p*-nitroanilide, L-ZPVAPNA = *N*-benzyloxycarbonyl-L-phenylalanine-L-valine-L-arginine-*p*-nitroanilide. Reaction Conditions: Tests 1-4 were performed in 1% DMF (v/v). DMF has been shown to enhance the sensitivity of single residue substrates to hydrolysis by trypsin, with the greatest increase being the initial jump from 0-1% (v/v) and becoming inhibitory to enzyme activity above 10% (v/v) (Somorin *et al.*, 1978b). Reaction 1-4 contained 667mM substrate & 10μg/ml enzyme. Reaction 5 & 6 contained 15.6mM substrate & 0.99μg/ml enzyme.

The sensitivity of the synthetic substrates to hydrolysis clearly increase from 1→6. Note that in 5 & 6, additional amino acids were placed between the protecting group and the target residue and these were found to dramatically enhance the affinity of the peptide for trypsin. A highly labile substrate would, however, not be particularly advantageous for the latter stages of this project, where the assay would be directed towards measuring substrate specificity. A highly sensitive substrate would likely release its chromogen even if poorly bound by the enzyme, making an assay for determining relative substrate specificity's unworkable.

A comparison of $K_m$ & $k_{cat}$ values of substrates 4 and 6 (L-ZAPNA & L-ZPVAPNA) illustrates the higher level of enzyme-substrate complex dissociation with the single residue substrate (Table 2.4).

| | Substrate | Parameters | |
|---|---|---|---|
| | | $K_m$ ($M*10^{-4}$) | $k_{cat}$ $(M*S)^{-1}$ |
| 1 | L-ZAPNA | 2.1 | 1.31 |
| 2 | L-ZPVAPNA | 0.357 | 14.49 |

**Table 2.4: Hydrolysis rates related to substrate size**
Illustrating the high rates of activity obtained with peptide as opposed to single residue substrates. Adapted from (Somorin *et al.*, 1978b).

The benzoyl blocking group was selected for the initial work in this project, namely to be incorporated in the simple synthetic substrates which would form the base of the screening reactions. It was chosen on the grounds that it was well characterized and did not appear to overly enhance substrate sensitivity. The apparently highly labile nature of synthetic peptides is a factor that must be considered when designing the larger peptide substrates with which to screen the enzyme variants for increased substrate specificity. It is, however, perfectly possible that the heightened rate of catalysis with peptide substrates occurs because they present a more natural binding site to the enzyme, allowing it to demonstrate the true level of activity that it would show against a similar target site in a whole protein. The lower levels of activity seen with single residue substrates could thus indicate a higher $K_M$ in the enzyme-substrate complex or simply lower binding affinity.

### 2.1.3 $N_\alpha$-Benzoyl-L-arginine-p-nitroanilide

The single amino-acid substrate selected for the initial simple activity assays was $N_\alpha$-benzoyl-L-arginine-p-nitroanilide (Synonyms: L-BAPNA, L-BANA, BANI). This compound has been used for previous similar work involving the engineering of recombinant trypsin (Evnin, Vasquez, and Craik, 1990; Evnin and Craik, 1988) and was relatively inexpensive. There were no reported problems with specificity or stability, but gradual autolysis was known to occur at pH ≥ 8.

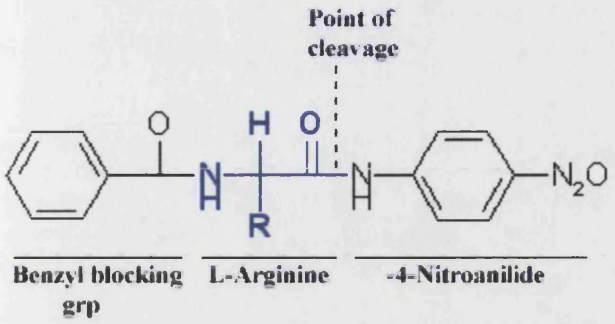



**Figure 2-2: Structure of L-BAPNA**

Molecular formula: $C_{19}H_{22}N_6O_4 \cdot HCl$. Molecular weight: 434.9. CAS number: 21653-40-7.

### 2.1.4 Other assay components

The assay would obviously only be capable of detecting trypsin that was both soluble & activated, and its success when working initially with the Eli Lilly commercial production strain, ELTRP-1, would rely on the assumption that not all of the recombinant trypsinogen expressed was bound as insoluble inclusion bodies within the cytoplasm. Data suggests that in such cases, at least a certain percentage would fold correctly within the interior of the cell and it is the catalytic activity of this fraction that would be assessed for the purposes of screening. The first significant aim of this project was to increase this fraction by means of altered fermentation conditions, molecular engineering or by random mutation.

The key to a successful screening assay is to use the smallest possible number of handling steps and, in this particular case, the problem lay in extracting and activating a cytosolic or periplasmic zymogen while avoiding the use of any complex purification steps. The first step was that the cells must be lysed. Ideally this step should rely on either the addition of a single compound or a fully automated process, and lyse the cells in as gentle a manner as possible. With this in mind, two possibilities were examined, namely the B-PER™ Bacterial Protein Extraction Reagent (Pierce), a mild, non-ionic detergent which has been developed specifically for soluble protein extraction, and the well-characterized freeze-thaw method. It was decided that assay development would first focus on the use of B-PER as it was by far the most rapid option (also requiring less handling of the plates by the operator) and had shown to be effective at extracting active soluble protein and inclusion body purification from total cell lysates. Other mechanical

methods of cell disruption, such as sonification, were not considered because of their impracticality on a microplate scale.

It was believed that LB medium might raise the baseline absorbance at 495nm (wavelength at which nitroaniline release is read) to unacceptable levels, masking signal from low quantities of protein. As the expected expression level was not known the sensitivity required in the assay could not be predicted, but it seemed prudent to develop the assay using a relatively optically clear medium in parallel with LB. The growth medium arrived upon consisted of defined M9 salts, complex casein amino acids and glucose and was termed M9 semi-defined (M9SD). The medium was adapted from M9 minimal media (Sambrook *et al*, 1989).

## 2.2 Materials & methods

### 2.2.1 Enzyme storage

A trypsin (Sigma) stock solution was of 1mg/ml was made in 20mM Tris-HCl, pH 7.5 & frozen in 1ml aliquots @ -20°C. Concentration was confirmed by measuring absorbance at $A_{280}$ applying the Beer-Lambert relationship:

$$c = \frac{A_{280}(blanked)}{\varepsilon * l}$$

$c$ = enzyme concentration ($mg*ml^{-1}$)

$\varepsilon$ = extinction coefficient (absorption coefficient) in units of $M^{-1}cm^{-1}$. $\varepsilon_{280}$ of trypsin = 1.41.

$l$ = optical pathlength of solution, 1cm for a standard cuvette.

### 2.2.2 Nα-benzoyl-L-ariginine-p-nitroanilide (L-BAPNA) activity assay

Basis for Assay:

In a single residue protease activity assay, an N-protected 4-nitroanilide is hydrolysed to release free 4-nitroaniline, the accumulation of which can be measured spectrophotometrically at 405nm. $A_{405}$ will increase when pH > 3.5. Nitroanilide compounds are known to autolyse gradually in solutions with pH ≥ 8, therefore the substrate solution (pH 7.5) must be made up immediately prior to the experiment.

Reagents:

Substrate Buffer (1X): 50mM Tris-HCl (Sigma), 20mM $CaCl_2$ (Sigma), 1% DMF (Sigma), pH 8.0, make as 10X concentrate.

Substrate Solution for Microplate Assay: 0.4mM L-BANA (Sigma) in final assay volume (200µl). Substrate soln. volume per well = 180µl, ∴ make 0.44mM soln. Made fresh on day of use, mixed well.

Substrate Solution for Cuvette Assay: 0.4mM L-BANA in final assay volume (1ml). Substrate soln. volume per cuvette = 500µl, make 0.8mM soln. Made fresh on day of use, mixed well.

Equipment:

Reader Setup: Microplate Assay. Reader (Dynatech MR7000) was set to measure at a single wavelength of 405nm and take 1 reading per minute for 10 minutes, with an initial 5s shake and immediate start to measurements.

Reader Setup: Cuvette Assay. Time-drive protocol was created on spectrophotometer (Uvikon 922), set to measure @ 405nm, taking 1 reading per minute for 10 minutes.

Procedure:

Microplate Assay: 180µl of substrate solution pipetted into all wells. 20µl of enzyme solution pipetted onto lip of each test well. Readings started immediately.

Cuvette Assay: 500µl of substrate soln. pipetted into both cuvettes (sample & blank). 4000µl of substrate buffer pipetted into both cuvettes. 100µl enzyme solution added to sample, mixed either by inversion (apply film to top & invert quickly three times) or with vial stirrer. Readings started immediately.

### 2.2.3 P-Toluenesulphonyl-L-Arginine Methyl Ester (TAME) Based Trypsin Activity

Basis for Assay:

In this single residue protease activity assay, the N-protected substrate is hydrolysed to release the methyl ester, causing an increase in absorbance when read at 247nm. Protein concentration is determined by following Beer's Law and reading at $A_{280}$. Due to the low wavelength at which the

reaction is followed (A247). it may only be performed at a cuvette scale (UV quartz cuvettes required) and not at microtitre scale.

Reagents:

Reaction Mixture: 0.001M TAME (Worthington Biochemicals) & 10mM $CaCl_2$ (Sigma), 40mM Tris-HCL (Sigma), pH 7.8 - 8.2. The reaction mixture is purchased as lyophilized aliquots from Worthington Biochemical corporation, with each vial being reconstituted with 6.5ml RO water. Mixture volume required = 2.9ml per vial.

Equipment:

Reader Setup: Cuvette Assay. UVIKON 922 Spectrophotometer was used. Time-drive protocol was created, set to measure @ 247nm, taking 1 reading per minute for 10 minutes. No calculation option was selected UV quartz cuvettes must be used for this assay.

Procedure:

Dissolved content of an appropriate number of vials with 6.5ml reagent grade water. Pipette 2.9ml into sample & blank cuvettes. 0.1ml trypsin solution added to the test cuvette and mixed rapidly, either by inversion (cover with film & invert quickly three times) or with vial stirrer. Readings started immediately.

### 2.2.4 Activity units

For the purposes of this work, one unit of enzyme activity was defined as the amount required to release 1μmol of chromophore per minute under the specific conditions used. Each set of experiments was represented by two graphs, the first illustrating the μmols of substrate released per litre as a function of time, the second showing the specific activity of the enzyme in the sample.

Calculation for substrate catalysis:

$$\mu mol * L^{-1} = \left(\frac{A_{410}}{\varepsilon * l}\right) * 10^6$$

A = blanked absorbance reading of sample at specific time, no units. Read @ 410nm for L-BAPNA or @ 247nm for TAME.

$\varepsilon$ = extinction coefficient (absorbtion coefficient) in units of $M^{-1}cm^{-1}$. 4-nitroaniline $\varepsilon_{410}$ = $8800L*moles^{-1}*cm^{-1}$; p-tolunenesulfonyd-L-arginine $\varepsilon_{247}$ = $540L*moles^{-1}cm^{-1}$.

l = optical pathlength of sample, units in cm. 1cm for standard cuvettes, empirically determines to be 0.63cm for 200µl of solution in a Starstedt microplate.

Calculation for specific activity of tested enzyme:

$$U * \mu g^{-1} = \frac{(\mu mol * ml^{-1}) * \min^{-1}}{c}$$

U = $\Delta\mu mol*ml^{-1}*min^{-1}$, µmol substrate released per minute.

c = enzyme concentration ($\mu g*ml^{-1}$).

### 2.2.5 Media preparation

Luria-Bertani (LB) media: (per litre of deionised $H_2O$) 10g bacto-tryptone (Merck); 5g bacto-yeast extract (Merck); 10g NaCl (Sigma). Adjust pH to 7.0 with $5_N$ NaOH (Signa). Sterilized by autoclaving for 20 minutes at 15 lb/sq. in. on liquid cycle.

M9 Semi-Defined (M9SD) media: (per litre of deionised $H_2O$) 100ml 10x concentrate M9 salts (see below); 10g casein hydrolysate (Merck). Sterilized by autoclaving for 20 minutes at 15 lb/sq. in. on liquid cycle. After autoclaving add: 2ml 1M $MgSO_4$ (Sigma) and 0.1ml 1M $CaCl_2$ (Sigma) (both autoclaved separately as stocks); 20ml 20% glucose (Merck)(prepared on day, sterilized by filtration); 1ml 0.1% thiamine (Sigma)(filter sterilized in solution).

M9 salts: (10x concentrate, per litre of deionised $H_2O$) 128g $Na_2HPO_4{\cdot}7H_2O$; 30g $KH_2PO_4$; 5g NaCl; 10g $NH_4Cl$. Sterilized by autoclaving for 20 minutes at 15 lb/sq. in. on liquid cycle. All salts purchased from Sigma.

### 2.2.6 Microplate centrifugation

A series of experiments was performed to assess fraction separation & cell pelleting in a microwell format and to determine whether these procedures could be used as a step in a high-throughput screen. All work was done with a Sorvall RC-5C centrifuge using a ST-H750 rotor (radius from centerline of rotor to sample = 15.02cm) and standard round-bottomed 96-well microplates (Starstedt). Microplates with all 96 wells filled with infected LB media ($A_{600}$ = ~ 1.0) were loaded

according to manufacturer's protocols, and were centrifuged at the maximum permissible speed for the rotor: 4300rpm, generating a relative centrifugal force (RCF) value of 3102. Assessment of cellular and lysate precipitation was performed by comparison of the turbidity ($A_{600}$) of the well supernatant before and after centrifugation.

### 2.2.7 Test of B-PER lytic efficiency at increasing dilution

In order to test for the lowest effective dilution of the B-PER detergent (Pierce) (essential for cost purposes if the reagent was to be applied to possibly thousands of well-cultures), it was necessary to monitor the release of soluble protein into the media from treated cells. Initially, it was attempted to follow the extent of lysis by measuring the absorbance of the media (clarified by 15min centrifugation @ 13,000rpm) at $A_{280}$, but the errors involved proved unacceptably high. The Lowry assay provided a simple method for estimating the quantity of soluble protein released, with the added advantage that the reagents could be prepared in-house. The buffer compositions & assay protocol are described in Appendix 2. Experiments were performed with the M9SD medium as this was known to demonstrate a lower background with the assay. Cell cultures of *E. coli* DH5$\alpha$ were grown overnight in M9SD at 37°C, 200rpm and then diluted to an $A_{600}$ (blanked) of 1. 500$\mu$l of culture was then added to a 1ml cuvette, with B-PER then being added to the appropriate dilution and the remainder being filled with MilliQ water. Cell suspensions were then vortexed for 5 seconds. All experiments were performed at room temperature and, after vortexing, cultures were gently agitated (shaker platform at ~ 20rpm) for 10min to prevent settling and encourage dissociation of cellular fragments and debris.

## 2.3 Results

### 2.3.1 Basic Kinetics of Trypsin on Single Residue Substrates in Optimal Buffered Conditions

Initial standard curves for the unit activity and the specific unit activity of commercial trypsin on both the L-BAPNA (Figure 2-3) and TAME (Figure 2-4) substrates were prepared in order to assess the lower detection limit offered by each substrate. Buffer conditions were optimal, prepared as recommended by the manufacturer (See Sections 2.2.2 & 2.2.3). For each activity assay, the specific unit activity ($U \times \mu g^{-1}$) was calculated as future mutation work may impair (or conceivably enhance) enzyme activity, and thus records of 'wild-type' performance were crucial.

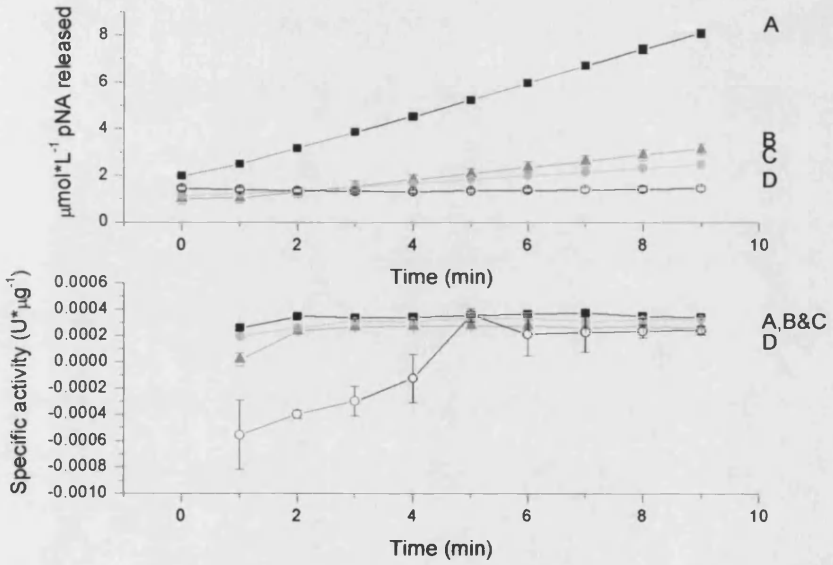



**Figure 2-3: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in 100mM Tris-HCl, pH 8.0.**

Enzyme concentrations denote amount of enzyme in 1ml reaction volume. Key: A (–■–) = 2µg trypsin; B (–▲–) = 1µg; C (–●–) = 0.5µg; D (–○–) = 0.1µg. Reactions were blanked against substrate and reaction buffer without enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg = 0.703; 1µg = 0.262; 0.5µg = 0.154; 0.1µg =

0.001. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

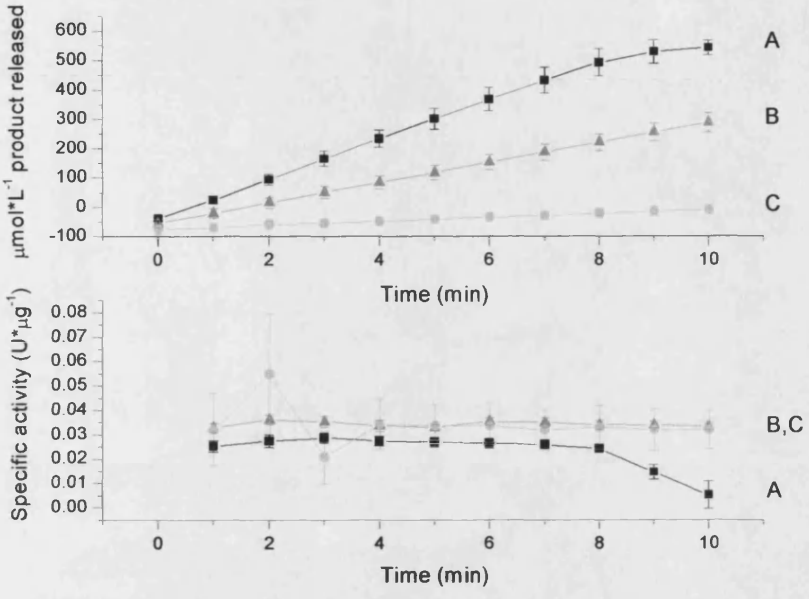



**Figure 2-4: Kinetics of commercial bovine trypsin on 0.001M TAME in 0.04M Tris buffer, pH 7.8 - 8.2**

Enzyme concentrations denote amount of enzyme in 1ml reaction volume. Key: A (—■—) = 2.5µg trypsin; B (—▲—) = 1µg; C (—●—) = 0.2µg. Reactions were blanked against substrate and reaction buffer without enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2.5µg = 66.6; 1µg 34.3; 0.2µg = 6.23. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

### 2.3.2 Incorporation of media into the assay

A high-throughput assay would require the activity of soluble enzyme to be assessed without purification or processing of the microwell cell culture. A series of tests were thus run to examine the sensitivity and reproducibility of the L-BAPNA assay in sterile (Figure 2-5, Figure 2-7) and inoculated ($A_{600}$ = 1) (Figure 2-6, Figure 2-8) LB and M9SD (optically clear) media.

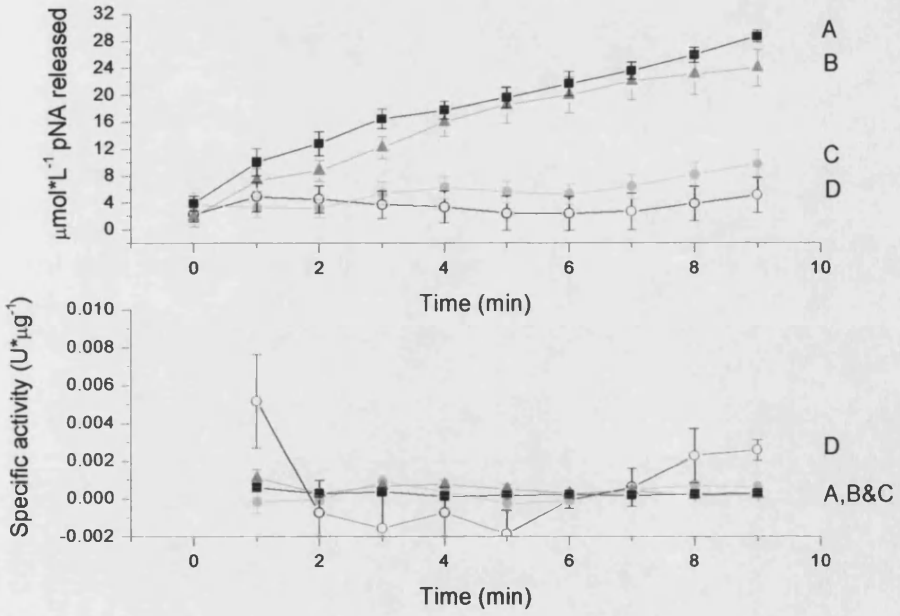



**Figure 2-5: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in uninoculated LB medium**

Key: A (—■—) = 2µg trypsin; B (—▲—) = 1µg; C (—●—) = 0.5µg; D (—○—) = 0.1µg. Concentrations denote the amount of enzyme added to 200µl, ∴ concentrations: 2µg = 10µg/ml, 1µg = 5µg/ml, 0.5ug = 2.5µg/ml & 0.1µg = 0.5µg/ml. Reactions were blanked against substrate and uninoculated medium without the addition of enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg = 2.54; 1µg = 2.44; 0.5µg = 0.52; 0.1µg = -0.36. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

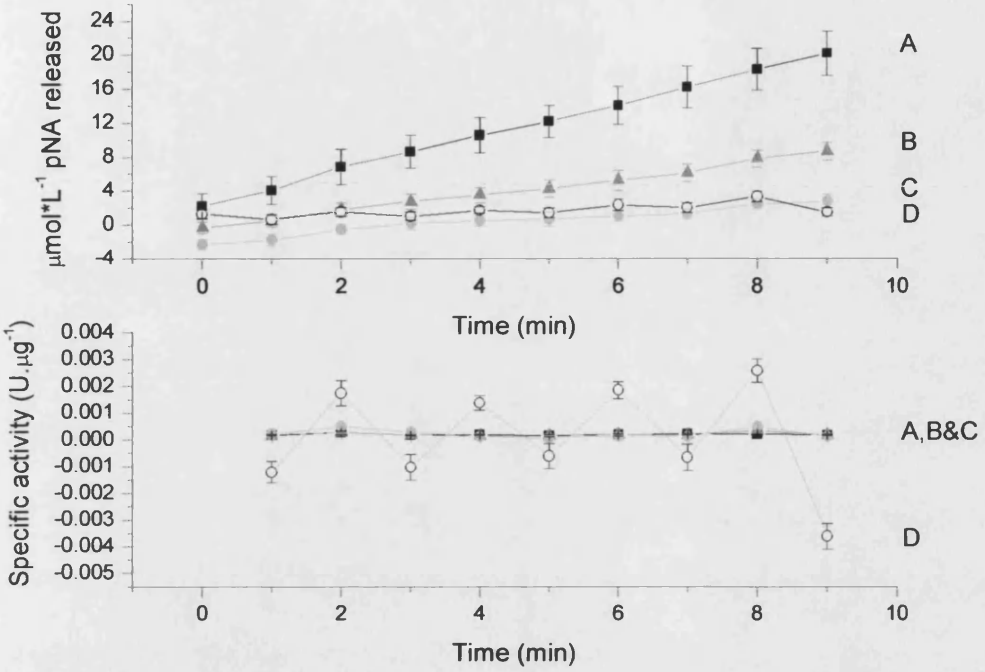



**Figure 2-6: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in inoculated LB medium (O.D. 1 @ 600nm)**

Key: A (—■—) = 2µg trypsin; B (—▲—) = 1µg; C (—●—) = 0.5µg; D (—○—) = 0.1µg. Concentrations denote the amount of enzyme added to 200µl, ∴ concentrations: 2µg = 10µg/ml, 1µg = 5µg/ml, 0.5ug = 2.5µg/ml & 0.1µg = 0.5µg/ml. Reactions were blanked against substrate and inoculated media without the addition of enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg = 2.02; 1µg = 0.942; 0.5µg = 0.506; 0.1µg = 0.225. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

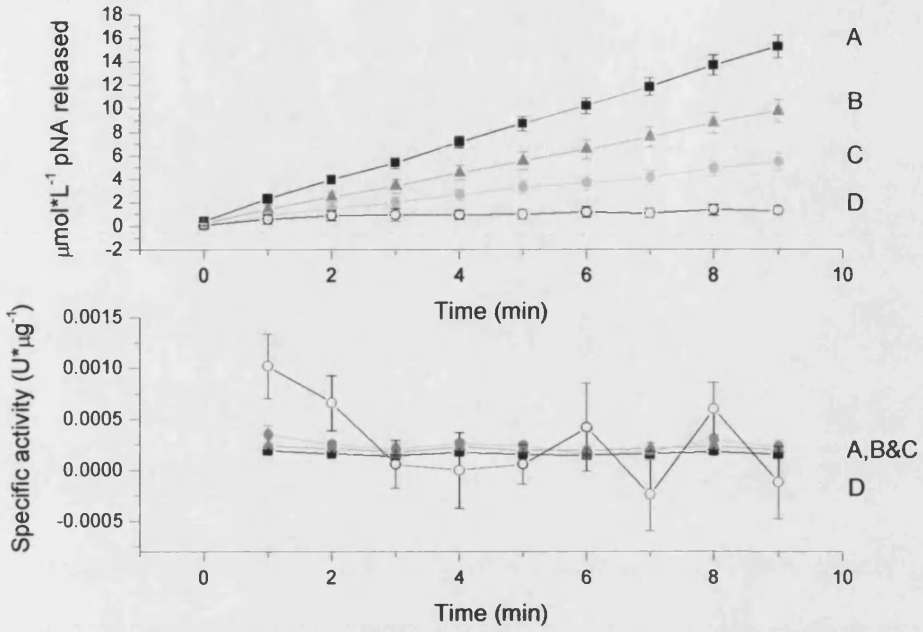



**Figure 2-7: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in uninoculated M9SD medium**

Key: A (—■—) = 2µg trypsin; B (—▲—) = 1µg; C (—●—) = 0.5µg; D (—○—) = 0.1µg. Concentrations denote the amount of enzyme added to 200µl, ∴concentrations: 2µg = 10µg/ml, 1µg = 5µg/ml, 0.5ug = 2.5µg/ml & 0.1µg = 0.5µg/ml. Reactions were blanked against substrate and inoculated media without the addition of enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg = 1.58; 1µg = 1.02; 0.5µg = 0.541; 0.1µg = 0.08. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

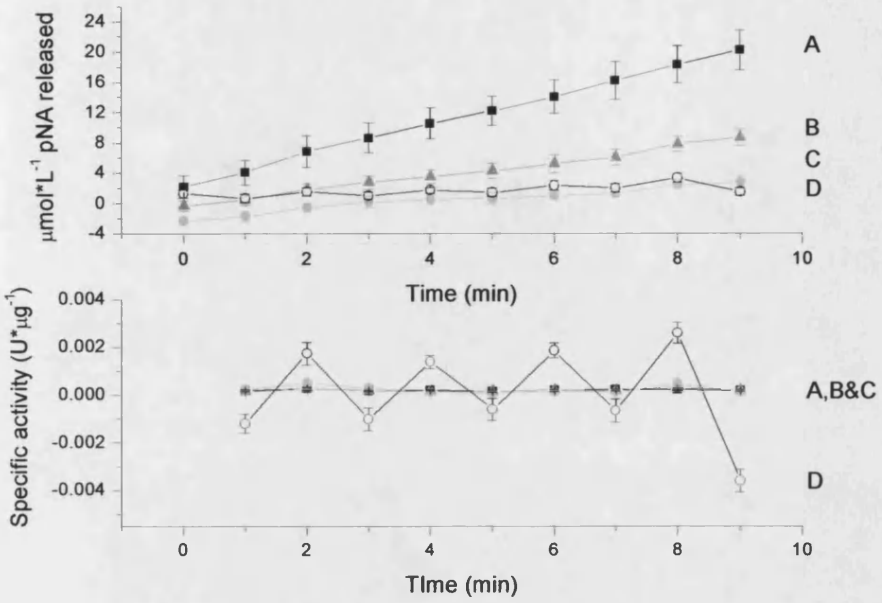



**Figure 2-8: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in inoculated M9SD medium (O.D. 1 @ 600nm)**

Key: A (—■—) = 2µg trypsin; B (—▲—) = 1µg; C (—●—) = 0.5µg; D (—○—) = 0.1µg. Concentrations denote the amount of enzyme added to 200µl, ∴concentrations: 2µg = 10µg/ml, 1µg = 5µg/ml, 0.5ug = 2.5µg/ml & 0.1µg = 0.5µg/ml. Reactions were blanked against substrate and inoculated media without the addition of enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg = 2.02; 1µg = 0.941; 0.5µg = 0.506; 0.1µg = 0.225. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

### 2.3.3 Assessing the impact of incorporating a cell lysis step using the B-PER reagent

The sensitivity and stability of L-BAPNA was assessed in the presence of a chemical lysis agent, the B-PER detergent, and the resultant cellular lysate, so that the practicality of lysing the well-cultures to access soluble cytoplasmic protein in a single step could be established.

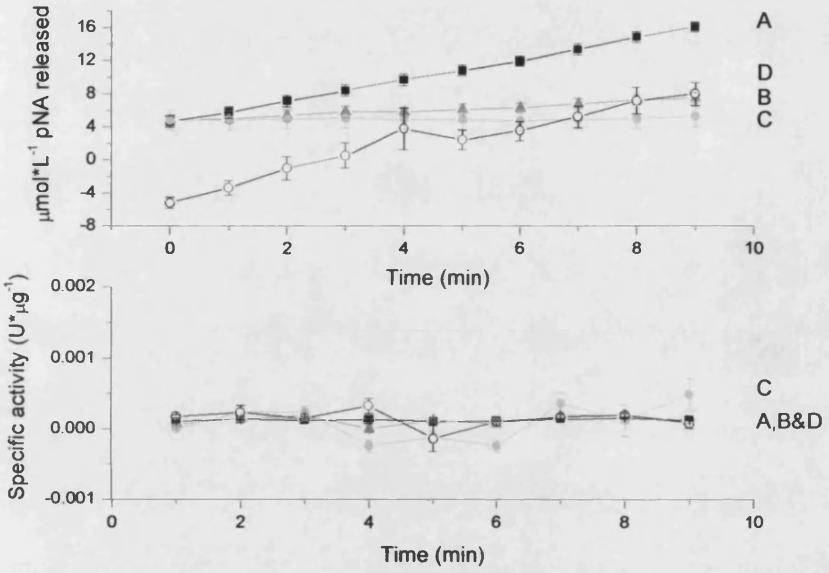



**Figure 2-9: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in the presence of inoculated LB media (O.D. 1) and B-PER (1 in 4 soln)**

To examine effect of lysis agent dilution on enzyme activity, 2ug of enzyme was also tested in the presence of neat B-PER. Key: A (–■–) = 2µg trypsin; B (–▲–) = 1µg; C (–●–) = 0.5µg; D (–○–) = 2µg trypsin in neat B-PER. Concentrations denote the amount of enzyme added to 200µl, ∴concentrations: 2µg = 10µg/ml, 0.5ug = 2.5µg/ml & 0.1µg = 0.5µg/ml. Reactions were blanked against substrate and inoculated media without the addition of enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg (1:4 B-PER) = 1.27; 2µg (neat B-PER) = 1.43; 0.5µg = 0.3; 0.1µg = 0.02. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

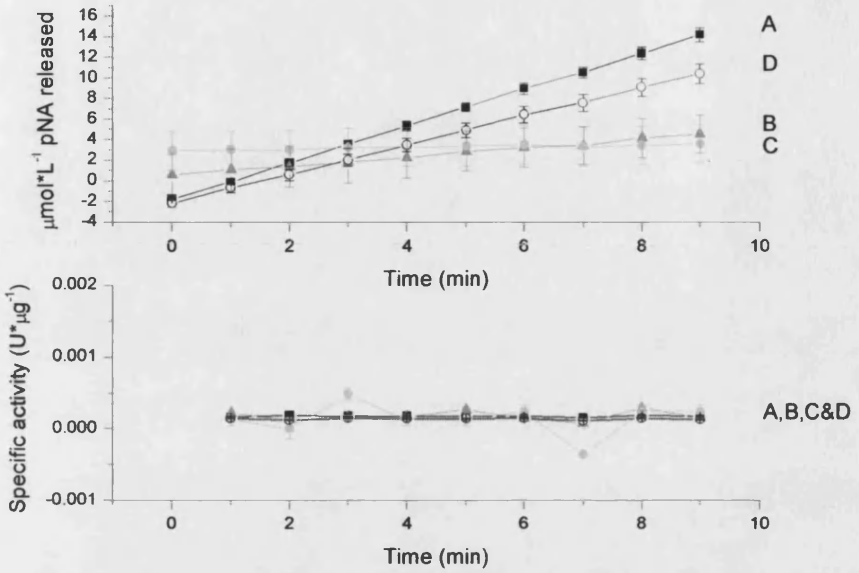



**Figure 2-10: Kinetics of commercial bovine trypsin on 0.4mM L-BANA in the presence of inoculated M9SD media (O.D. 1) and B-PER (1 in 4 soln).**

To examine effect of lysis agent dilution on enzyme activity, 2ug of enzyme was also tested in the presence of neat B-PER. Key: A (—■—) = 2µg trypsin; B (—▲—) = 1µg; C (—●—) = 0.5µg; D (—○—) = 2µg trypsin in neat B-PER. Concentrations denote the amount of enzyme added to 200µl, ∴concentrations: 2µg = 10µg/ml, 0.5ug = 2.5µg/ml & 0.1µg = 0.5µg/ml. Reactions were blanked against substrate and inoculated media without the addition of enzyme. Upper graph: µmols per litre of product released (pNA group) at each enzyme concentration. Units (µmol pNA released per litre per min) for each concentration (determined from linear region): 2µg (1:4 B-PER) = 1.77; 2µg (neat B-PER) = 1.37; 0.5µg = 0.391; 0.1µg = 0.05. Lower graph: Specific molecular activity (Units of trypsin activity per µg of enzyme) detected.

### 2.3.4 Assessing the lytic efficiency of B-PER

The manufacturer's protocols for the use of B-PER state that, for optimal results, pelleted cells must be resuspended in neat solution and subsequently shaken for 1 minute. As lytic efficiency was unlikely to be based on a 'trigger' concentration, a series of experiments were performed using increasing dilutions of B-PER and extended incubation times (cultures now shaken for 10 minutes) in order to establish whether use of the agent a higher dilution might be equally effective.

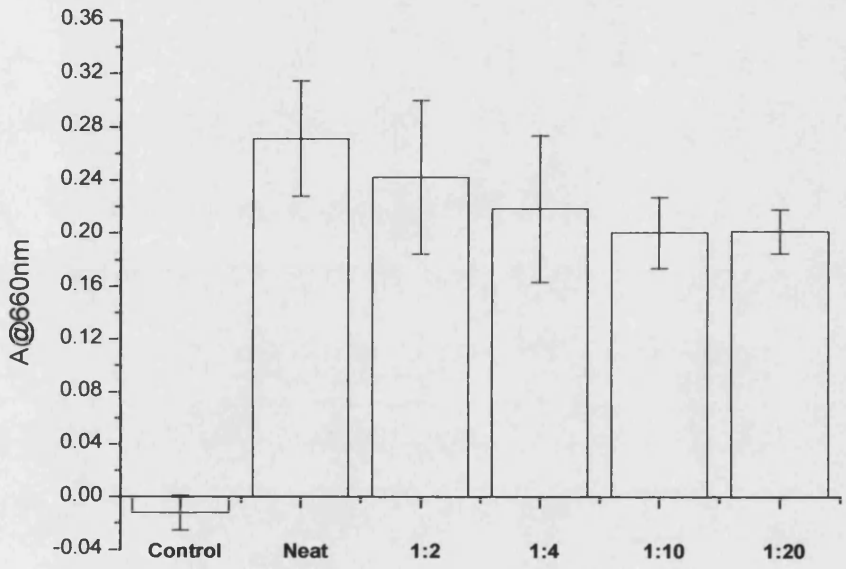



**Figure 2-11: Lytic efficiency of B-PER**

Illustrating the absorbance obtained from the Lowry protein concentration assay after addition of B-PER solutions of decreasing concentration to inoculated M9SD medium (O.D. 1 @ 600nm, then diluted 1:10 for assay). Control was inoculated medium without B-PER. B-PER solutions were diluted with RO water and each reaction was blanked against a tube containing a 1:10 diluted mixture of uninoculated medium & B-PER soln (@ appropriate diln) and which had undergone the full Lowry assay. The control was blanked against a reaction tube in which the protein soln was uninoculated media + water. No standard curve was run and thus no attempt was made to apply concentration values to absorbance, results were indicative only.

### 2.3.5 Use of microplate-centrifuge to pellet cells & insoluble fractions

A series of experiments were carried out to assess the usefulness of the Sorvall RC-5C centrifuge with the ST-H750 microplate carrier in pelleting cells from liquid culture and also in the separation of the insoluble fraction of the cells from the soluble fraction after lysis. Efficiency was assessed by filling wells with 300μl of culture, taking absorbance readings at A600 & A280, centrifuging for a specified time and speed, and then extracting the uppermost 200μl and re-measuring the absorbance at both wavelengths.

The ST-H750 rotor limited the centrifuge to a maximum of 4000 revolutions per minute (rpm). The arm length (radius) of the rotor was 15.02cm and thus the maximum relative centifugal force that could be generated was 2684. A series of tests involving infected LB and M9SD media showed that centrifugation for 20min @ 4000rpm served to pellet approximately 70% of the cells from suspension in both media. Centrifugation for a further 20min increased this value to approximately 92%, with a clear pellet forming at the bottom of the wells. After lysis of a cell culture, even an hour of maximum speed centrifugation produced no significant alterations in the readings. It was important to note that the Starstedt microplates suffered from increasing physical stress when centrifuged at an rcf of 2684 for 30min or longer. Plates frayed severely around the edges and faint discolorations began to appear in the main body of the plate, presumably due to stressing of the plastic.

## 2.4 Discussion

### 2.4.1 Assay development

Figures 2.9 & 2.10 show the sensitivity achievable using the final form of the proposed high-throughput assay in infected LB & M9SD media respectively (blanked $A_{600}$ = 1) and in the presence of the lysis reagent. The optically-clear M9SD medium offered obvious benefits in terms of minimum detection levels, with the signal ($A_{405}$) evolved from enzyme concentrations below 10μg/ml presumably being masked by the yellow colouration of the LB culture, but initial tests (see Chapter 3) showed that ELTRP-1 growth was significantly slowed in the semi-defined medium. This suppression of growth rate would not, in itself, present a problem, but it was not known whether it may impact on r-trypsinogen overexpression, possibly requiring altered time of induction away from the Eli Lilly standard fermentation protocols, or whether it would pose a problem for the

microwell fermentations which were to form the backbone of the high-throughput screen. It was decided that testing of the production strain, ELTRP-1, was to continue using both media in parallel.

Use of a microplate centrifuge proved to be effective in pelleting cells out of suspension in both media, despite the limited rcf generated. A potential improvement to the sensitivity of the assay could thus involve centrifuging the cells and resuspending in MilliQ water prior to lysis. This would enable the researcher to benefit from the high-growth rate of LB medium within well cultures and being able to separate the cells in each plate in a sigle, simple process step. Centrifugation was however a lengthy step, requiring at least 40min, and thus a certain bottleneck when processing a library of isolates, and also appeared to mildly damage the microplates. There always existed the potential that such damage could result in spillage of infected medium into the apparatus, the subsequent cleaning of which could result in severe delay and the waste of possibly hundreds of well cultures. As the sensitivity of the L-BAPNA assay with even LB media was still relatively high, it was decided that such a centrifugation step was not necessary unless detection levels became clearly critical.

The proposed form of the high-throughput assay would thus involve the overnight fermentation of isolates within shallow or deep 96-well plates, the transfer of perhaps 100μl of mixed culture to an assay plate and the immediate reading of activity after the addition of enterokinase. Such an assay would require only one transfer of material and thus a clear minimum of process steps and researcher intervention. Activity would be read over 10 minutes if possible, but an adequate estimate of enzyme concentration could be established from only two readings if they lay within the linear region of the reaction slope. The 2$^{nd}$ & 6$^{th}$ minute would be appropriate in this case.

### 2.4.2 Comparison of substrates

Nα - benzoyl - L - Arg p-nitroanilide (L-BANA) and p - toluenesulphonyl - L - Arg methyl ester (TAME) differ in both their N-terminal blocking & C-terminal chromogenic groups. Release of the methyl ester following hydrolysis of the scissile bond results in a signal which is, on average, 3-4 times stronger than that generated by the nitroanilide group. The reaction rate of the enzyme for TAME appeared to be in the region of 110x higher than for L-BAPNA, with substrate depletion

effects occurring after 9min for 2,5μg/ml trypsin (Figures 2.3 & 2.4). Blanks remained constant over time, showing that signal was not caused by substrate instability, and so this heightened affinity of the enzyme for TAME shows why the assay was adopted as an industrial standard.

Although the basic TAME assay had to be assessed so that the protease variants produced could be compared against the trypsin data held by Eli Lilly, it was known that it would be impractical for use in a high-throughput assay because of its UV absorbance spectrum (max A @ 247nm). Only the L-BANA substrate was therefore tested for compatibility with the other elements of the screen.

### 2.4.3 Examining the lytic efficiency of B-PER

The standard Pierce protocol calls for B-PER to be used neat on a resuspended pellet of cells. As mentioned before, this was unacceptable, both in terms of the requirements of the assay and in terms of cost, and so it was desirable to establish a curve of dilution vs. activity. All assay development work was done using B-PER at a 1:4 diln, added 1:1 to a cell suspension, i.e. a 1:8 diln of the protocol recommended concentration. Figures 2.9 & 2.10 examine the impact of adding neat B-PER to the cellular suspension, as opposed to the 1:4 soln. In LB medium, this control demonstrated higher variability and a decreased absorbance profile while maintaining the virtually the same rate of reaction. The neat B-PER soln thus appeared to interfere with the precision of the assay while not retarding enzyme activity over the effect already observed.

Figure 2-11 used the Lowry assay to attempt to establish the effectiveness of the reagent in terms of the assay and appears to show a reproducible curve, demonstrating decreasing efficiency with increasing dilution. No trypsin was added as the test was purely to monitor the increase in soluble protein in the media upon lysis and the blanks used were thus uninoculated. This may explain the negative value obtained with the control (inoculated media without B-PER) as it is possible that the cells and cell debris sequesteres much of the free protein within the media, rendering it inaccessible to the Lowry assay. The results appear to show a tailing off of efficiency after the 1:4 dilution, but visual inspection of the samples provided ample proof of lysis (slight clarification of media, presence of large insoluble aggregations). It was decided that a 1:8 dilution of the lytic detergent would prove a adequate compromise providing that the well cultures were

gently shaken for 10min after the addition of the agent, to allow full lysis of the cells and the gradually disassociation of the cellular components.

## 2.5 Conclusions

In optimized conditions the catalysis of L-BAPNA and TAME could be detected clearly and reproducibly at minimum levels of 0.5μg/ml by the standard activity assays tested in this chapter. In uninfected LB media, this sensitivity rose to 2.5μg/ml for L-BAPNA (TAME was not tested in microwell format) and to 5μg/ml in infected media. The specific activity of the commercial trypsin remained identical in infected & uninfected medium, indicating that the enzyme was not to any significant degree inhibited by extracellular proteins or lysis products.

Experimentation with the optically clear medium M9SD resulted in a considerable improvement in sensitivity. Minimum detection levels were reduced to 0.5μg/ml in uninfected medium and 2.5μg/ml for infected medium. The specific activity of the controls was not affected.

The addition of the selected lysis agent, B-PER, appeared to depress trypsin activity by approximately 0.5 units on average, both diluted and neat, and in both media types. This was not considered to be a significant effect.

While the errors involved in assessing B-PER activity at progressive dilutions were high (Figure 2-11) and the assay involved tracked cellular lysis only indirectly, the data suggested that using B-PER at a 1:4 dilution with a 10min standing time provided an acceptable rate of lysis.

Centrifugation of the microplates using available equipment could be used as an effective step for the pelleting of cells from well cultures following a microwell fermentation in a standard, shallow Starstedt 96-well plate. The maximum rcf that could be generated was, however, clearly insufficient to precipitate inclusion bodies from suspension after culture lysis.

# 3 Initial ELTRP-1 characterization & fermentation work

## 3.1 Introduction

This section details the preliminary work done with the Eli Lilly r-trypsinogen production strain ELTRP-1 after it was received on the 28th September 2000. The questions that needed to be addressed before formulating any mutagenic or protein-engineering strategy were related to extending our knowledge of the strain's characteristics and thus suitability for the proposed mutagenesis and subsequent high-throughput screen. Growth and expression levels in complex medium (LB) were first to be assessed along with the degree of fine control available over induction. It was then useful to test whether growth on the optically clear M9SD growth medium significantly affected either growth or expression. M9SD had been tested as an alternate growth medium to LB using *E. coli* DH5α as a test strain, as this media allowed for greater sensitivity in the L-BAPNA activity assay, where aliquots of cell culture were added directly to the assay reagent (see Chapter 2). Eli Lilly had worked with the assumption that, following induction and overnight (16hr) incubation, 100% of the recombinant protein had aggregated as inclusion bodies with no residual native activity remaining. For the sake of fully characterizing the 'wild-type' strain and to establish the background for the future work, it was necessary to assay for trace levels of trypsin activity within the various fractions of the cell (insoluble, soluble cytosolic, periplasmic & supernatant) during the course of a standard fermentation.

Preparatory molecular work on the pHKY603 plasmid involved enzymatic digestion to confirm identity and the optimisation of PCR conditions both for confirmation of gene sequence and as preparation for the planned gene modification work (prosequence deletion, leader sequence attachment, site-specific mutations) (see Chapter 5 & 6). Standard PCR conditions also needed to be established so as to provide a basis for future work with mutagenic PCR, if it was decided that this method of directed evolution was most suitable for developing the planned mutant libraries.

### 3.1.1 Characteristics of the ELTRP-1 production strain

The host-expression vector system used by Eli Lilly for the production of (recombinant trypsin) r-trypsin) consisted of a derivative of *E. coli* K12, RV308, as the host, and a derivative of the plasmid pBR322, termed pHKY603, which codes for the production of methionyl-phenylanalyl-bovine

trypsinogen (biosynthetic Met-Phe-bovine trypsinogen) as the expression vector. The resulting strain RV308:pHKY603 was called ELTRP-1. For an abbreviated plasmid map of pHKY603, please refer to Figure 3-1.The strain carried plasmid encoded tetracycline resistance and genome encoded streptomycin resistance, both genes expressed constitutively.

The product gene (*met-phe-trypsinogen*) was under the control of a temperature sensitive promoter. The identity and type of this promoter was not released by the company, but the behaviour of the strain indicated that it was a repressor system, where an inhibitory factor, linked to the temperature-sensitive promoter, was expressed at high levels and blocked the binding of transcription factors to *met-phe-trypsinogen*. Temperatures below 36.5°C were classed as 'permissive' and allowed the expression of the inhibitor, while 36.5°C and above were 'restrictive', preventing inhibitor expression and thus allowing the transcription of *met-phe-trypsinogen*. Company literature suggested that the induction temperature should be considered to have a range of 35°C-38°C, with the accepted temperature for non-induced growth being 32°C.

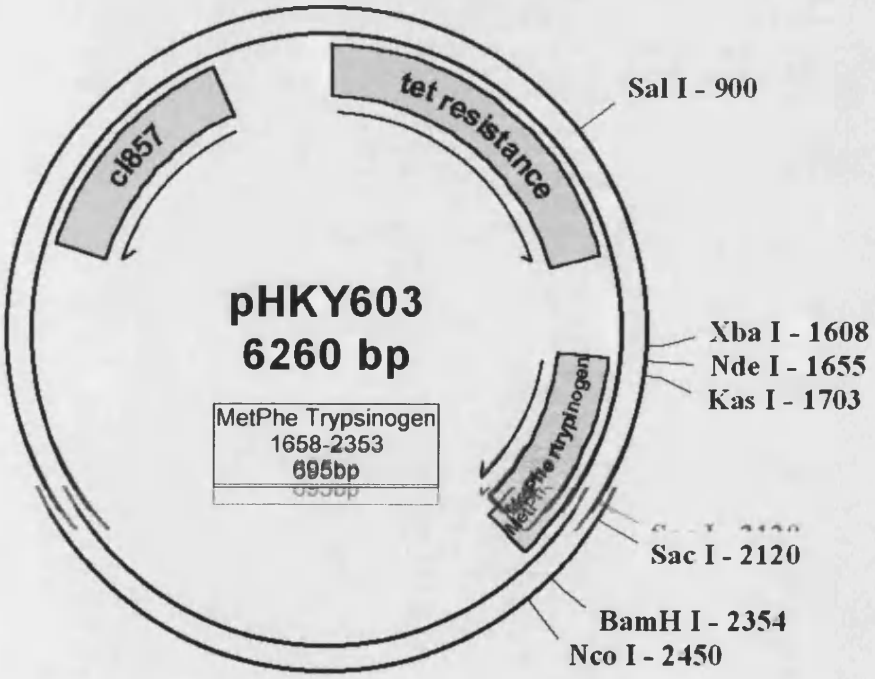



**Figure 3-1: Abbreviated structure of the pHKY603 plasmid**
Source Eli Lilly company literature.

### 3.1.2 Industrial fermentation conditions

Detailed conditions & feeds, as well as precise descriptions of the vessels used could not be stated here for reasons of company confidentiality. No data was provided for small scale fermentation & expression work and available protocols detailed a 2000L production run of ELTRP-1. Fermentation volumes were scale-up from 500ml to 250L after inoculation with frozen master-stock, with the latter volume forming the inoculum for the 2000L vessel. Growth conditions were held at 29-31°C, 6.9-7.1pH with the induction point being determined by monitoring of glucose depletion rates. Glucose concentration could be measured either on-line using a continuous glucose analyser or by analysis of off-line samples at the appropriate frequency. Induction would be set to occur when the glucose concentration fell below 10g/l and involved the raising of the vessel temperature to 36.5°C. Under the specific conditions used, the culture was induced on average 8-10 hours after inoculation.

The end point of the fermentation was typically determined by time, with harvesting occurring approximately 8 hours after induction. The length of the peak production period of the fermentation was known to be influenced by the operating parameters. At the completion of the fermentation, the broth was inactivated (no viable *E. coli* detectable) by standard measures. The broth was heated *in situ* to a minimum of 58°C, and this temperature was held until 18 $F_{58}$ had been accumulated.

Protocols for the recovery and purification of Met-Phe-trypsinogen were not released, but it was surmised that they followed the standard form for the treatment of an insoluble, inclusion-body bound recombinant protein. Centrifugation of batches of the media would allow the dense inclusion bodies to be recovered in the heavy phase, with the resultant slurry (which would contain approximately 20% solids w/w) being washed with sterile water and detergent prior to solubilisation in a suitably reducing, chaotrophic environment. An in-depth discussion of the refolding steps available and an analysis of the Eli Lilly method may be found in Chapter 4.

## 3.2 Materials & methods

### 3.2.1 ELTRP-1 genotype & phenotype

Strain is a derivative of *E. coli* K12, RV308. Plasmid is termed pHKY603, a derivative of pBR322. Phenotype: $Tc^R$, $Sm^R$, $Lac^-$.

### 3.2.2 Upon receipt of strain

A stab culture of ELTRP-1 was received from Eli Lilly (Fergersheim plant) on 28.09.'00. Immediately on arrival, the culture was used to inoculate 10mls of sterile selective LB media (tetracycline @ 10μg/ml, streptomycin @ 100μg/ml: Antibiotics were at stringent plasmid concentration as recommended in Eli Lilly literature). The liquid culture was grown for 16hrs at 30°C. 200rpm and was used to create streak plates (LB agar: tet. 10μg/ml, strep. 100μg/ml) and 10 aliquots of 15% glycerol stocks which were then stored at –80°C. Agar plates were grown for 20hrs at 32°C and then single colonies were picked to inoculate 10ml cultures of selective media which was grown in the same conditions as before. Small-scale plasmid extractions were performed on each of the overnight cultures using the QIAprep Spin Miniprep kit (Qiagen). The plasmid preparations were measured at $A^{260}$ using a quartz ultra-micro cuvette (Bellman) and an ATI Unicam UV/Vis Spectrometer (Model UV2). The DNA samples were diluted appropriately to give readings of between 0.1 and 1.0. A 0.8% agarose gel was run at 80V, 50mA for 2hrs to verify the plasmid size and purity.

### 3.2.3 Testing relative growth rates on LB & M9SD media

Single ELTRP-1 colonies were picked and used to inoculate 3× 100mls of LB and M9SD in 500ml baffled shake-flasks (tet. 10μg/ml, strep. 100μg/ml). These were then grown at 32°C, 200rpm for 16hrs with 1ml samples being taken for the first 7 hours. Readings were taken on the ATI Unicam UV/Vis Spectrometer (Model UV2) and were blanked against uninfected media of either type. Media temperature was monitored rigorously by means of a thermometer within a control flask of uninfected media to ensure that it did not approach 35°C. Eli Lilly literature stated that induction occurred at temperatures between 36.5-37.5°C and that 35°C must be considered with the range of error. So as to achieve comparative data for growth rates after induction, the experiment was

repeated but the temperature of each culture was raised to 37°C at the end of exponential growth (in LB: 4hrs, M9SD: 6hrs).

### 3.2.4 Testing for soluble rTrypsinogen

Single ELTRP-1 colonies were picked and used to inoculate 10mls of LB and M9SD in 50ml falcon tubes (tet. 10μg/ml, strep. 100μg/ml). These seed cultures were grown at 32°C, 200rpm overnight and then 1ml of each was used to inoculate 2x 100mls of LB and M9SD in 500ml baffled shake-flasks (tet. 10μg/ml, strep. 100μg/ml). These were grown at 32°C, 200rpm until the end of exponentional growth (in LB: 4hrs, M9SD: 6hrs) and then one set was raised to the induction temperature (37°C) and then grown for a further 16 hours. The second set (uninduced control) was held at 32°C for the identical time.

The aim of the experiment was to test for the presence of soluble (and thus activatible) trypsinogen in the supernatant or the cytoplasm (released with the lysis agent, B-PER) and so the experiment was set up in the following manner using standard, shallow 96-well microplates (Starstedt) and a microplatereader (Dynatech MR7000). Data was collected and analysed using BioLinx software, version 2.0 (Dynatech laboratories, inc).

For each media type the experimental wells below were repeated with both uninduced and induced cell cultures. Each well was filled with 300μl and activation of the recombinant trypsinogen was effected by the addition of 1μg enterokinase (to a final well conc. of 3.3μg/ml) and allowing the plate to stand at room temperature for 5min. The substrate solution was added last, immediately prior to reading.

| | |
|---|---|
| Blank | 120μl induced/uninduced culture + 140μl MilliQ-water + 40μl 0.4mM L-BAPNA |
| Activity Control | 120μl induced/uninduced culture + 120μl MilliQ-water + 40μl 0.4mM L-BAPNA + 20μl trypsin stock 1 (@ 50μg/ml ∴ 1μg added) |
| Supernatant Activity | 120μl induced/uninduced culture + 120μl MilliQ-water + 40μl 0.4mM L-BAPNA + 20μl trypsin stock 2(@ 500ng/ml ∴ 10ng added) |
| Cytoplasmic Activity (trypsin) | 120μl induced/uninduced culture + 120μl B-PER + 20μl MilliQ-water + 40μl 0.4mM L-BAPNA |
| Cytoplasmic Activiy (trypsinogen) | 120μl induced/uninduced culture + 120μl B-PER + 40μl 0.4mM L-BAPNA + 20μl trypsin stock 2 (@ 500ng/ml ∴ 10ng added) |

Experimental setup: Blank – used to measure the baseline absorbance reading with intact cells, also would indicate if active rTrypsin was present in supernatant or if mix was acting to hydrolyse the substrate. Control 1 – used to provide a standard signal from 1μg trypsin in the presence of intact cells. Supernatant Activity – shows whether rTrypsinogen was present in the supernatant using 10ng trypsin as an activator (the uninduced culture wells will provide a standard signal for the added enzyme). Cytoplasmic Activity (trypsin) – shows whether rTrypsin is present in the cytoplasm. Cytoplasmic Activity (trypsinogen) – tests whether rTrypsinogen is present in cytoplasm (with commercial trypsin as activator). The substrate used was L-BAPNA with a final assay concentration of 0.4mM. For activity profiling of this substrate, buffer composition and the calculation of activity units see Chapter 2, sections 2.2.2, 2.2.4 & 2.3.3).

Reading conditions: Wavelength = 420nm; Shaking = 10s initially, 5s prior to each subsequent reading; Temperature control = off (ambient); Intervals = 1 reading/min for 10min.

### 3.2.5 Comprehensive testing of ELTRP-1

The following protocol was designed to test for the presence of soluble, active trypsin in an ELTRP-1 fermentation under a variety of conditions. The experiments were extended variants of those described in 3.2.4, and refer to the results shown in Figure 3-11.

Cell culture was prepared as described in 3.2.4 and a 100μl aliquot was transferred to an assay plate, where additional reagents were added (total well volume of 200μl). Substrate (L-BAPNA, with a final well concentration of 0.4mM) was added immediately prior to the start of measurement in each case. Assay components were as follows:

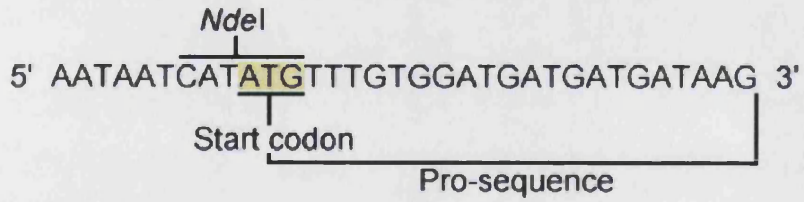
| | |
|---|---|
| 1 | Basic activity (no lysis, no denaturant): Induced culture + L-BAPNA |
| 2 | Periplasmic activity: Induced culture + L-BAPNA + 0.5μg/ml lysozyme |
| 3 | Total soluble activity: Induced culture + B-PER + L-BAPNA. |
| 4 | Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 0.1M urea + L-BAPNA |
| 5 | Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 0.5M urea + L-BAPNA |
| 6 | Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 1M urea + L-BAPNA |
| 7 | Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 2M urea + L-BAPNA |
| 8 | Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 5M urea + L-BAPNA |
| 9 | Controls: Uninduced culture + L-BAPNA +0.1, 0.5, 1, 2μg commercial trypsin. |

All experiments were blanked against their exact counterparts, run with uninduced medium. The purpose of this was to firstly generate the clearest results possible in terms of genuine r-trypsin activity, and also to probe for possible hydrolysis of the substrate in uninduced cultures under particular conditions. A positive result in the latter case could indicate a degree of leakage from the repressor system, allowing very low levels of trypsin to be expressed without condensing as inclusion bodies. Activity units as well as calculations for substrate depletion can be found in section 2.2.3.

### 3.2.6 Establishing PCR conditions

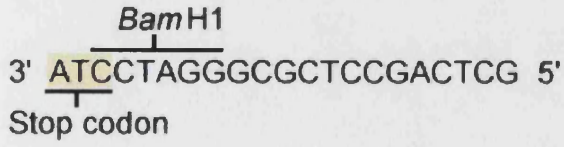
Positive and negative sense primers (referred to as P & N respectively) were designed for pHKY603, flanking *rtrypsinogen* and providing 'protection' of the start & stop codons and the pro-sequence of the zymogen. In this case 'protection' means that these aspects of the sequence are included in the primers and thus are safe from mutagenesis/miscopying. Oligonucleotides were ordered via custom synthesis from Amersham Pharmacia.

**Positive sense primer (P)**




Properties: 33nt, $T_m$= 64.2°C

**Negative sense primer (N)**




Properties: 21nt, $T_m$= 67.0°C

**Figure 3-2: Amplification primers for *rtrypsinogen* on pHKY603**

For the initial test, PCR reaction conditions were set as standard for the amplification of a 1kb gene (Sambrook *et al.*, 1989). Thermocyler settings: Initial denaturation (5min @ 94°C); [ Denaturation (60s @ 94°C); Annealing (90s @ 60C); Extension (60s @ 72°C) ] cycle x28; Final extension (10min @ 72°C); Holding temperature (4°C). Temperature ramping was set to max and the heated lid was set to 100°C.

100 pmol required for a typical reaction. Oligonucleotides (Amersham Pharmacia Biotech) supplied at the following levels:

P: 27.8nmol/285.9μg ∴ resuspend in 500μl MilliQ-water ⇒ 2μl = 110pmol

N: 62nmol/396.0μg ∴ resuspend in 500μl MilliQ-water ⇒ 1μl = 120pmol

Reagents used: Template = approx. 20-50ng in 1μl; reaction buffer 10X = 10μl; dNTP (10mM each nucleotide stock) = 2μl (0.2mM/nucleotide); primer P = 2μl; primer N = 1μl; $MgCl_2$ (25mM stock) = 6μl; MilliQ-water = 78μl. Total volume = 100μl. All PCR reagents purchased from Promega.

When the initial reaction yielded no signal, the reaction conditions were varied in a number of ways, namely: Stringency of the reaction was lowered by decreasing annealing temp. to 55°C.

[$MgCl_2$] was varied to low (0.8mM) & high (2mM).

[dNTP] increased to 0.8mM/nucleotide (8μl from 10mM/nucleotide stock).

[Template] increased: Miniprep from initial work @46ng/μl ∴ concentrations used: 46ng = 1μl; 92ng = 2μl; 138ng = 3μl.

Other reagents & conditions were held as before and each reaction was made up to 100μl with MilliQ-water. Exact conditions used for each experimental variant are listed in the Results section.

### 3.2.7 Cellular fractionation, inclusion body purification & washing and SDS-PAGE

SDS-PAGE gels were used as a direct visual test for the presence of overexpressed protein. The procedure for the separation of ELTRP-1 cultures into soluble & insoluble protein fractions was adapted from the standard Pierce B-PER Mini-Scale Bacterial Protein Extraction protocol. This was designed for use on 1-2mls of overnight cultures ($OD_{600}$ 1.5-3.0) and allows for the fractionation of the cell into soluble (total cytosolic & periplasmic) protein and inclusion bodies. For increased purification of the inclusion bodies it was recommended that lysozyme be used to further digest cellular debris, but that was considered unnecessary for this relatively crude assay. Protease inhibitors were also omitted.

The method was as follows: 1ml of overnight culture (induced & uninduced) was spun down for 10min @ 13,000rpm so as to pellet the cells. Supernatant was retained for further analysis, cells were resuspended in 200µl neat B-PER and vortexed vigorously for 1 minute until the solution appeared homogeneous. The solution was centrifuged again for 5min @ 13,000rpm and supernatant was retained as 'soluble fraction'. Pellet was resuspended in 200µl diluted B-PER (1:10 dilution with MilliQ water), vortexed for 1 minute and then centrifuged for a further 5 minutes. Pellet 'washed inclusion body' was resuspended in 200µl MilliQ water. Samples were either analysed on same day or stored at –20°C. All SDS-PAGE reagents were purchased from National Diagnostics and gels were prepared and run according to manufacture's protocols.

## 3.3 Results

### 3.3.1 Characterization of ELTRP-1 in various media

After the initial culturing of the Eli Lilly 'stab culture' and the creation of glycerol stocks from single colonies, the growth rate of the strain was tested in both LB and M9SD. Results shown overleaf in Figure 3-3.

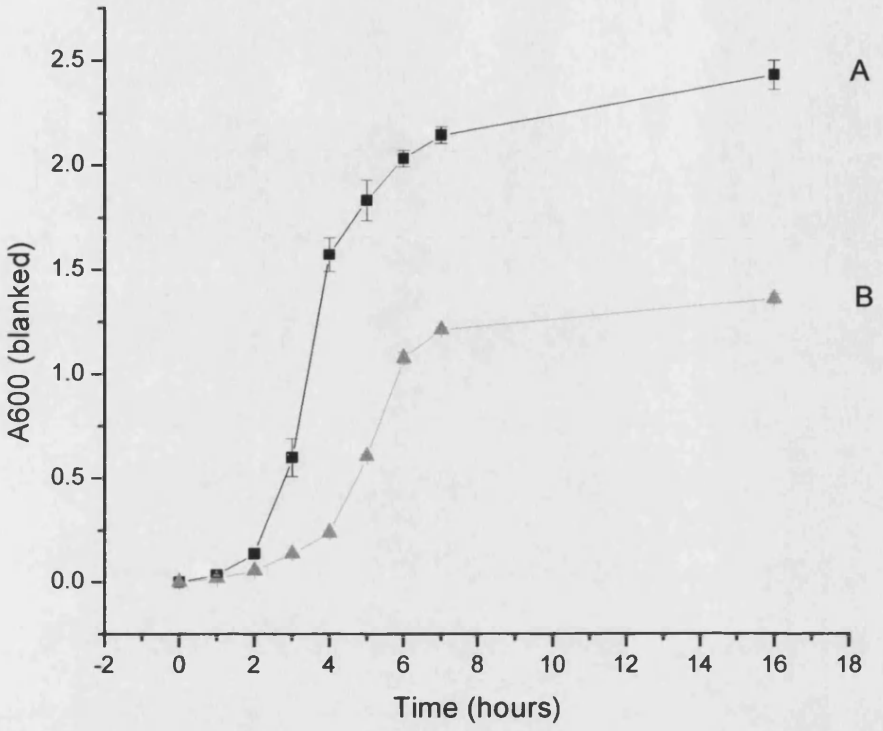



**Figure 3-3: Growth of ELTRP-1 in LB & M9SD (both Tc+, St+) without induction**
Key: A (—■—) = ELTRP-1 grown in LB; B (—▲—) = ELTRP-1 grown in M9SD. Reactions performed in triplicate & conditions held @ 32°C, 200rpm throughout.

Plasmid DNA was purified from the overnight culture in LB using the Qiagen MiniPrep procedure and run on an agarose gel against a supercoiled ladder in order to confirm the presence and size of the pHKY603 plasmid. The purified plasmid appeared to approximately 1Kb smaller than the documented size of pHKY603 (6260 bp), but despite this the strain appeared to exhibit the correct phenotype ($Tc^r$, $St^r$). The growth rates were compared against those where the cultures were raised to the restrictive temperature of 37°C at the end of exponential growth (in LB: 4hrs, M9SD: 6hrs) and then at the beginning of exponential growth (in LB: 2hrs, M9SD: 4hrs). These two induction points were designed to test the Eli Lilly assertion that maximum levels of rTrypsinogen would be achieved if the cultures were induced at the beginning of the stationary phase.

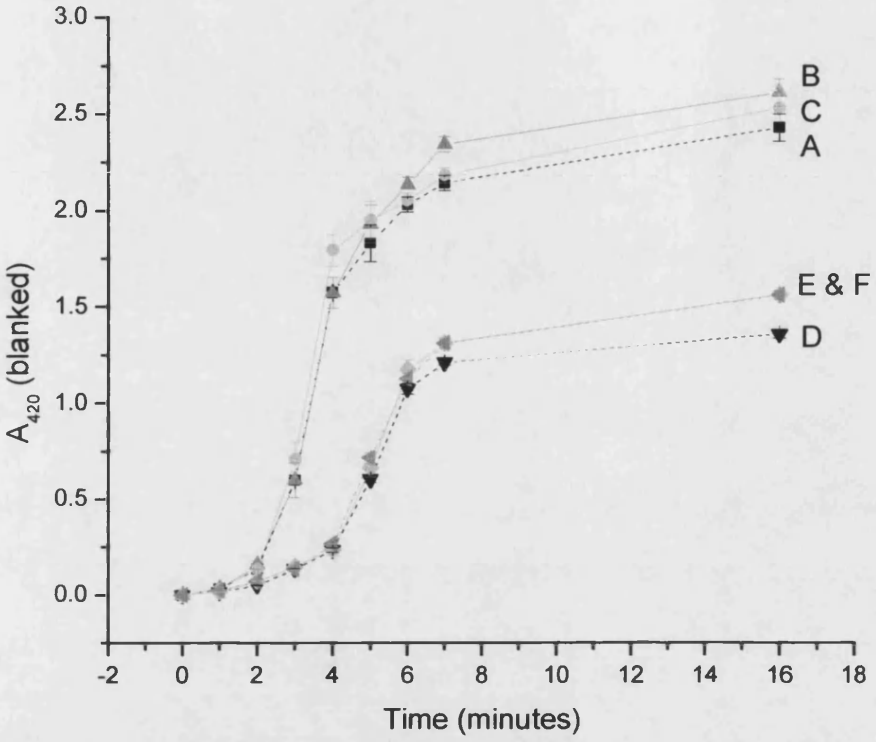



**Figure 3-4: Growth of ELTRP-1 in LB & M9SD with induction times before and after exponentional phase**

Key: ···■··· = A; —▲— = B; —●— = C; ···▼··· = D; —◆— = E; ···◀··· = F.

Plots A & D show the uninduced growth in both media as drawn previously in figure 1; B & E chart growth with late induction (in LB: 4hrs, M9SD: 6hrs) in LB & M9SD respectively; C & F show growth with early induction (in LB: 2hrs; M9SD: 4hrs) in LB & M9SD respectively. All reactions were performed in triplicate with 100mls infected media within 500ml baffled flasks, initial growth @ 32°C, 200rpm, induced by raising temperature to 37°C, 200rpm.

It seemed evident that no great metabolic load was being placed upon the cultures after induction, irrespective of the point at which applied. A transient dip in the growth rate would have provided indirect evidence for recombinant protein overexpression and thus provided both further proof of the strain's identity and a convenient method for roughly testing the lower range of the heat-inducible promoter. The logical next step was to assay for specific trypsin activity in induced cultures using the final assay method described in Chapter 2.

### 3.3.2 Testing for soluble r-trypsinogen

Replica flasks of LB & M9SD cultures were grown until the onset of stationary phase and then raised to induction temperature and grown for a further 16hrs. Various tests and controls were prepared for microplate assay (as described in Methods) with the aim of determining whether any soluble recombinant trypsinogen (or indeed trypsin) was present in either the supernatant or soluble fraction of the cell. A positive signal at this stage would have been highly useful as it would have served to confirm the identity of the strain and would also have shown whether any physical or genetic manipulation of the culture would be necessary, or whether it was already a suitable platform for directed evolution (i.e. if soluble protein be obtained from the fermentation in a manner appropriate for a high-throughput screen). Please refer to section 3.2.4 for exact protocol.

No trypsin activity was detected in the soluble fraction of either induced or uninduced culture grown in either media, and thus only PCR and plasmid digestion remained as paths for strain confirmation.

### 3.3.3 Amplification & digestion of *r-trypsinogen*

With the identity of the strain still unconfirmed, in so far as there was no apparent metabolic load upon induction, direct confirmation seemed essential and there were three obvious routes for achieving this, namely: amplification of the target gene (*r-trypsinogen*) via the polymerase chain reaction (PCR); digestion of the pHKY603 plasmid with appropriate restriction enzymes to produce a specific digest pattern; and SDS-PAGE of the cellular fractions (supernatant, periplasm, soluble cytoplasm & insoluble cytoplasm) to visualize if a protein of the correct molecular weight was being overexpressed. Sequencing would have provided the most direct method of inspection, but was delayed until PCR results demonstrated that the primers used could generate a signal from pHKY603 as the sequencing process would be dependant upon this reaction unless universal pBR322 primers were acquired.

The initial reaction conditions were set as standard for a 1kb gene (Sambrook *et al.*, 1989): Template (from frozen plasmid preparation run on a 0.8% agarose gel previously to confirm quality) = approx. 20-50ng in 1μl. The PCR product was left unpurified and was run on a 1.2% agarose gel & visualized using EtBr. Using the custom positive & negative sense primers described in Methods, 3.2.6, a product of approximately 700bp (723bp exact calculated length) was expected.

No signal was detectable from the first two reactions but initial failure in the amplification of a gene is not uncommon, as variables such as the nature & source of the target gene, the supplier of the various reagents and the sequence (length & predicted secondary structure) of the primers can all require that the optimal reaction conditions be determined empirically. Easily adjusted elements include the total salt concentration in the reaction, the annealing temperature and the concentration of the template plasmid and so each of these was varied in turn with simple continued aim of verifying the *rtrypsinogen* gene. In each experiment the reaction tubes were collected immediately after cooling from the final cycle and the putative product was then purified via spin column. Subsequent PCRs thus examined template concentration and salt levels in a combinatorial manner, with [template] ranging from 55ng to 0.5ng, and an [$MgCl_2$] of between 0.5mM and 3mM. Still no signal was obtained, despite the effectiveness of all reagents being confirmed by the amplification of a control plasmid. The final experiment involved the alteration of the annealing temperature for low stringency in the hybridisation step, the result of which can be seen in Figure 3-5.

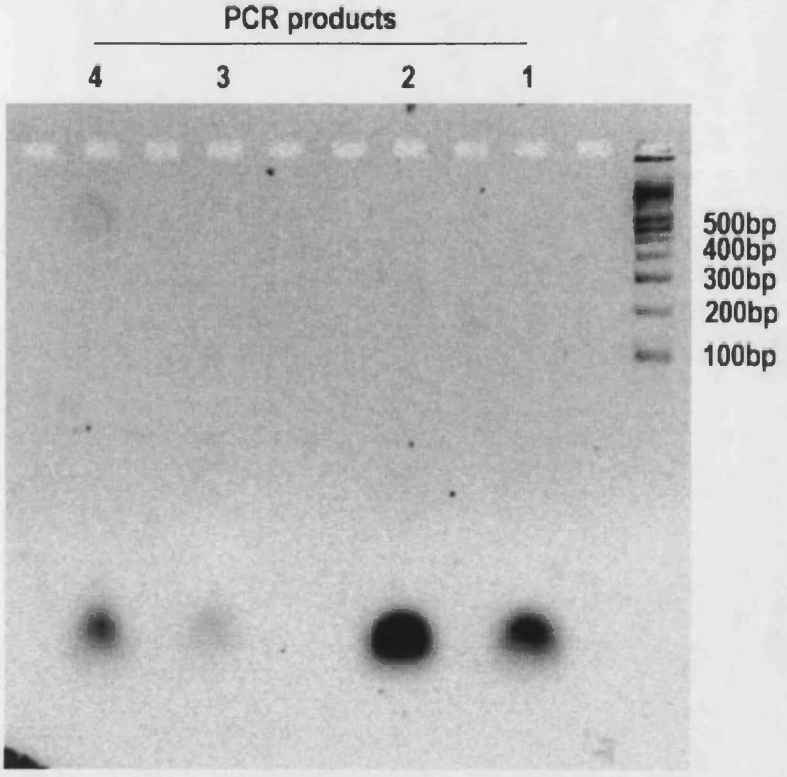



**Figure 3-5: PCR products from *r-trypsinogen* on pHKY603 amplified using varying annealing temperatures**

1.2% agarose gel was run at 100V, 120mA, 1.5hrs. Primers P & N have predicted $T_m$s of 64.2 & 67 respectively. Lanes: 1 – Annealing temp. of 50°C; 2 – $T_A$ of 55°C; 3 – $T_A$ of 60°C; 4 – $T_A$ of 64°C.

Thus no alteration in reaction conditions resulted in amplification of the *r-trypsinogen* gene and attention turned towards the confirmation of plasmid identity by means of digestion at unique sites to create a signature fragment pattern.

Cutting of pHKY603 using two unique restriction sites, at least one of which should lie within the target gene would constitute a definitive test for strain identity. The enzymes selected were *Sal* I (approximately midway through the tetracycline resistance gene, see Figure 3-1) and *Sac* I (398 nucleotides into *rtrypsinogen*). A plasmid-free version of the ELTRP-1 strain, termed RV308, was received from Eli Lilly at this time for future use as a transformation host. The strain was grown in LB @37°C, 200rpm for 16 hours and then the culture was subjected to the MiniPrep procedure in order to confirm that no plasmid was present.

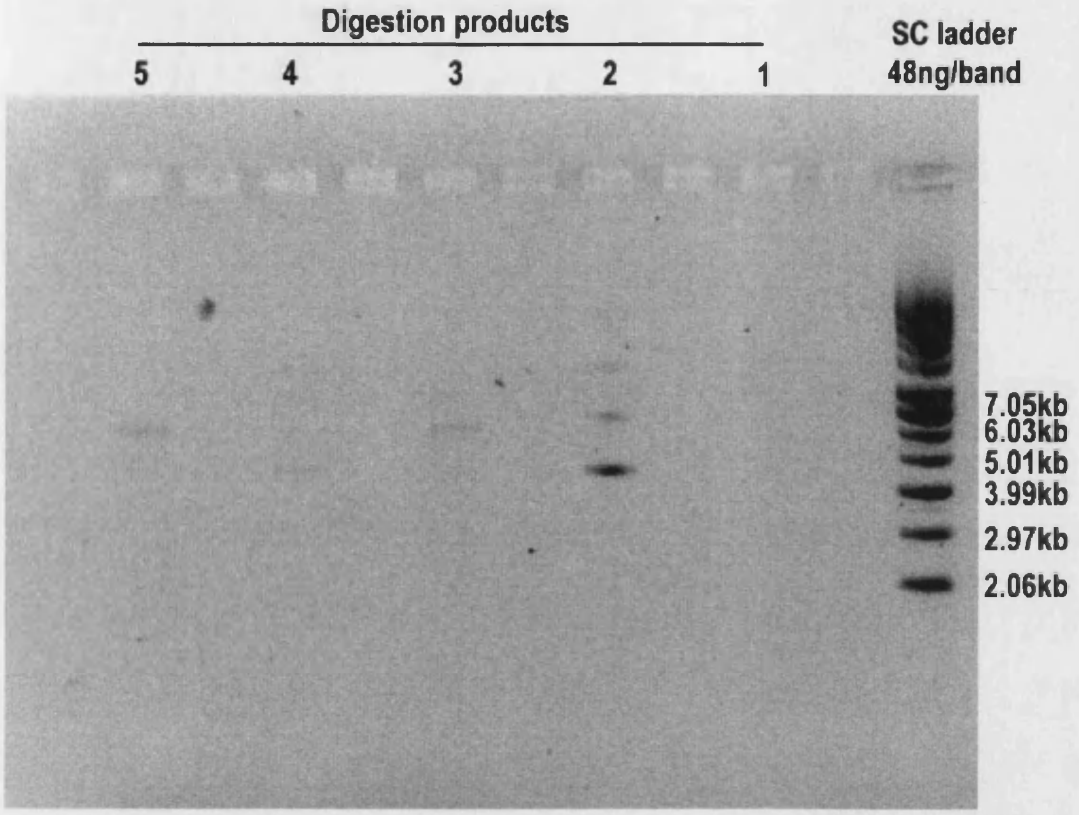



**Figure 3-6: Double digestion of pHKY603**

Digestion of pHKY603 with *Sal* I & *Sac* I and also plasmid purification from RV308. 0.8% agarose gel was run at 100V, 120mA, 1.5hrs. lanes: 1 – plasmid purification from RV308; 2 – 90ng supercoiled pHKY603; 3 – pHky603 digested with *Sal* I; 4 – pHKY603 digested with *Sac* I; 5 – pHKY603 digested with *Sal* I & *Sac* I.

From Figure 3-6, it seemed evident that while digestion with *Sal* I had linearised the plasmid (easily visible when compared to the supercoiled signal), digestion with *Sac* I had no effect and the double digestion produced only linearised plasmid instead of the two bands (1220bp & 5040bp) which were predicted. The cutting of *Sal* I within the plasmid as expected seems to confirm the presence of the tetracycline resistance gene, although this had already be proved by the resistant phenotype of the strain. The failure of *Sac* I to cut, although according to Eli Lilly documentation it is situated within *r-trypsinogen*, provided further evidence that the false strain had been received.

### 3.3.4 SDS-Page analysis of protein expression

An SDS-PAGE gel of the total cellular protein at various stages of an ELTRP-1 fermentation was used to directly check for overexpression upon induction. The culture was grown in LB and induced at the end of exponential growth. Samples were taken hourly for seven hours after the restrictive temperature was reached and the weight of the constituent proteins were analysed by comparison with commercial trypsin. No protein bands were visible that were not also present in the uninduced lanes and thus the experiment cast further doubt onto the identity of the strain.

The final characterization experiment involved a total protein SDS-PAGE comparison between ELTRP-1 and RV308 cultures in LB, both of which were raised to induction temperature at the end of logarithmic growth. The RV308 strain had been shown to carry no plasmid of any size and was documented by Eli Lilly to produce no genome encoded comparable proteases to trypsin.

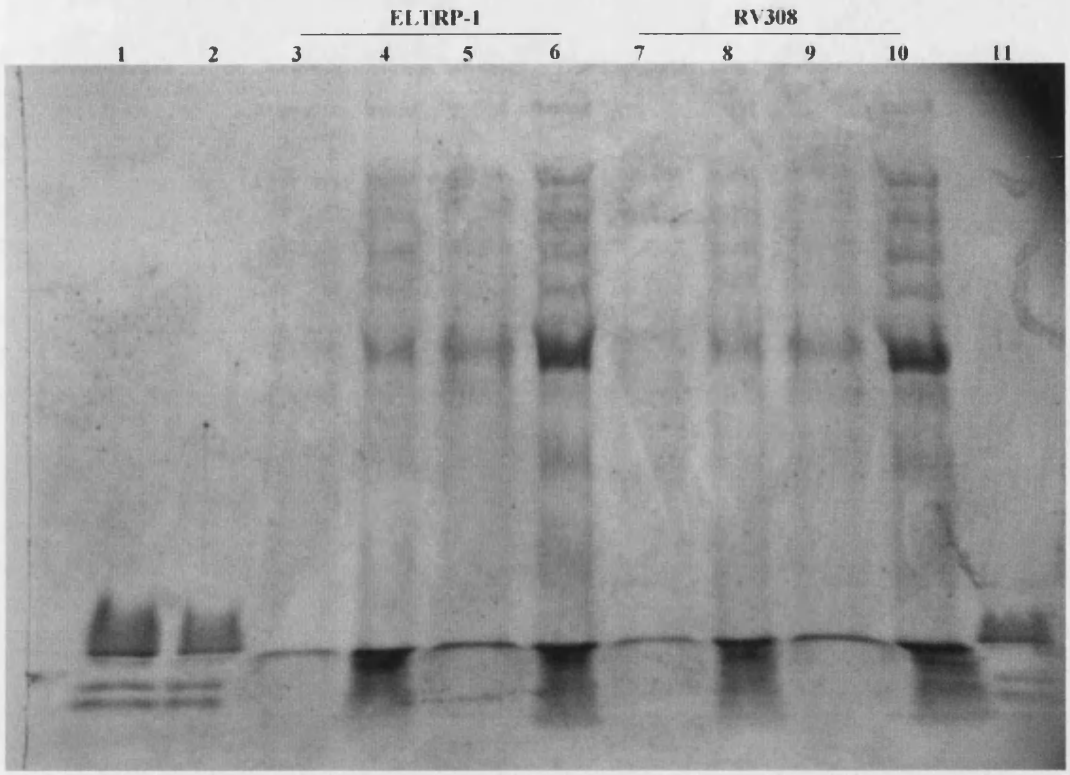



**Figure 3-7: SDS-PAGE of total protein from induced ELTRP-1 & RV308 cultures**
15% acrylamide gel with 4% stacking gel, run @ 80V for 40min then @ 150V for 120min. Both cultures grown in LB for 4hrs @ 32°C, 200rpm, then raised to 37°C. Lanes: 1 – 15µg commercial trypsin; 2 – 7.5µg commercial trypsin; 3 – ELTRP-1 ($T_0$ – immediately prior to induction); 4 – ELTRP-1 ($T_2$ – 2 hours after induction); 5 – ELTRP-1 ($T_4$); 6 – ELTRP-1 ($T_6$); 7 – RV308 ($T_0$); 8 – RV308 ($T_2$); 9 - RV308 ($T_4$); 10 - RV308 ($T_6$); 11 – 7.5µg commercial trypsin.

The size markers in lanes 1, 2 & 11 failed to run correctly, most likely due to overlong storage of the trypsin stock solution and thus the occurrence of advanced autoproteolysis. Nevertheless, no unique bands were visible in the ELTRP-1 samples, giving an identical banding profile to that of RV308.

The strain of ELTRP-1 received on 28.9.'00. could thus be said to have failed every conveniently available identity test for the presence of the *r-trypsinogen* gene. While exhibiting tetracycline (plasmid encoded) and streptomycin (genome encoded) resistance, the plasmid was approximately a kilobase below the expected 6kb, it then failed to cut in a manner consistent with

pHKY603 and the strain did not appear to overexpress any protein specifically upon induction. A further direct identity check would have involved attempting to sequence the region of the plasmid believed to contain *r-trypsinogen* using the P & N oligonucleotides. As these closely flanked the gene even a successful reaction would only have achieved good resolution over the center of *r-trypsinogen*, but this would have enough to confirm its presence. However, as the PCR experiments had failed to produce a product under any condition, it was believed that attempting to optimise a sequencing reaction would prove futile. A full copy of all details was sent to Eli Lilly, Fergersheim, and, after a review period, a second ELTRP-1 stab culture was dispatched.

### 3.3.5 Receipt of new ELTRP-1 culture

On the 23rd February 2001 a new stab culture of ELTRP-1 was received from Eli Lilly. The new stock was immediately cultured & preserved in the same manner as the original. From cultured isolates from selective plates ($tc^+$, $st^+$) the identity of this strain was queried by means of agarose gel electrophoresis of the purified plasmid, double-digestion with *Sal* I & *Sac* I and SDS-PAGE of induced culture. Product can be seen in Figure 3-8 overleaf).

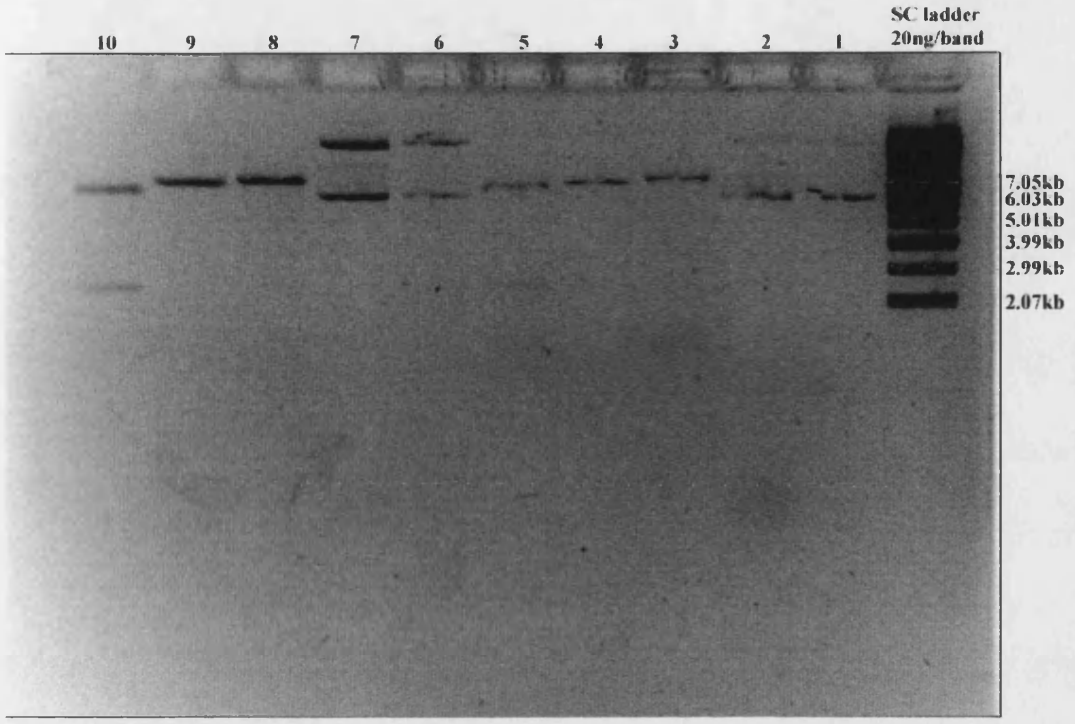



**Figure 3-8: Agarose gel of pHKY603 from new isolates plus *Sal* I & *Sac* I digest products**
Lanes: 1- 5µl pHKY603 (isolate 1a); 2 - 5µl pHKY603 (isolate 1b); 3 – 20µl digestion mix: pHKY603 from '1b' with *Sal* I; 4 - 20µl digestion mix: pHKY603 from '1b' with *Sac* I; 5 - 20µl double digestion mix: pHKY603 from '1b' with *Sal* I & *Sac* I; 6 – 5µl pHKY603 (isolate 2a); 7 – 5µl pHKY603 (isolate 2b); 8 – 20µl digestion mix: pHKY603 from '2b' with *Sal* I; 9 - 20µl digestion mix: pHKY603 from '2b' with *Sac* I; 10 - 20µl double digestion mix: pHKY603 from '2b' with *Sal* I & *Sac* I

The supercoiled purified pHKY603 samples were of the correct molecular weight (approximately 6kb) with the relaxed form visible immediately above. The plasmid appeared to be linearised after digestion with *Sal* I or *Sac* I and the double digestion yielded 2 bands of approximately the correct size (direct sizing not possible due to the linear nature of the cut DNA and the supercoiled marker). The DNA from isolate 2b appeared of comparatively high quality and so this isolate was selected for further work. To search for overexpressed protein after induction the culture was grown and induced as before and fractionated 4 & 6hrs after induction.

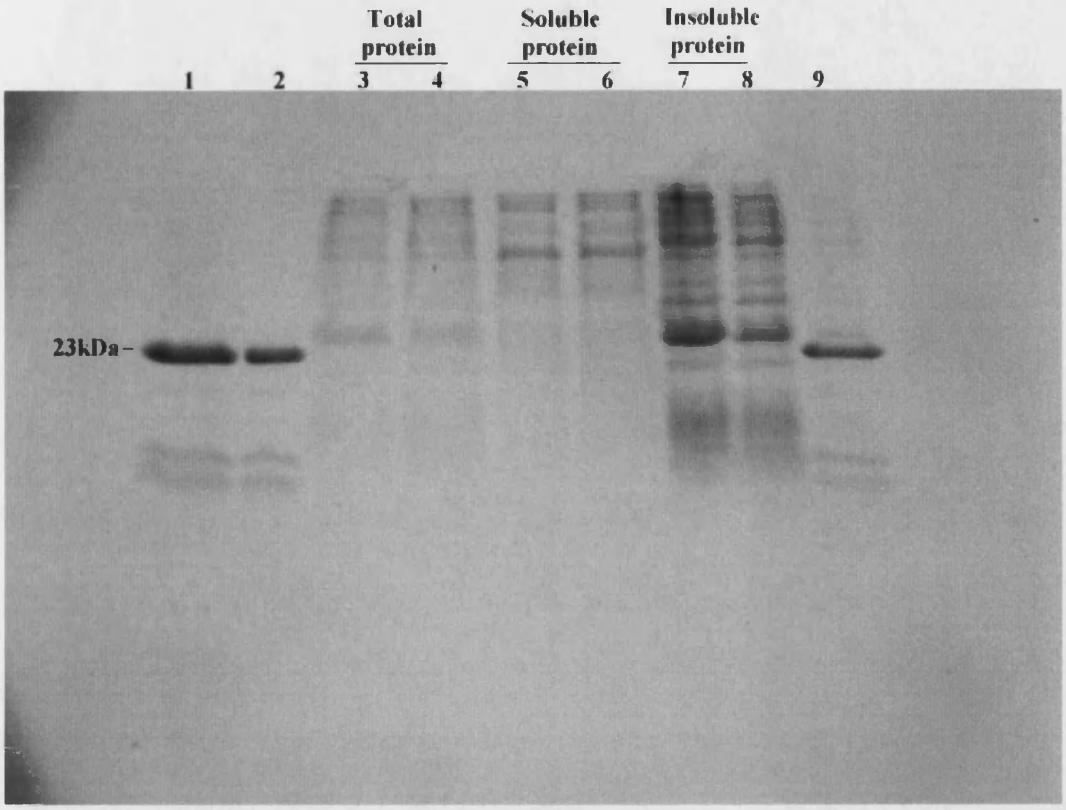



**Figure 3-9: SDS-PAGE of fractionated ELTRP-1 culture 6hrs after induction**

15% acrlyamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 90min. Lanes: 1 – 7.5ug commercial trypsin; 2 – 3.75µg commercial trypsin; 3 – 15µl total protein soln. ($T_6$ – 6hrs after induction); 4 - 15µl total protein soln. ($T_4$ – 4hrs after induction); 5 – 15µl soluble fraction ($T_6$); 6 - 15µl soluble fraction ($T_4$); 7 – 15µl insoluble fraction ($T_6$); 8 - 15µl insoluble fraction ($T_4$); 9 – 3.75µg commercial trypsin.

From Figure 3-9, overexpressed protein, of approximately the correct molecular weight (24.2 kDa) could clearly be seen in the insoluble fraction of the induced cells.

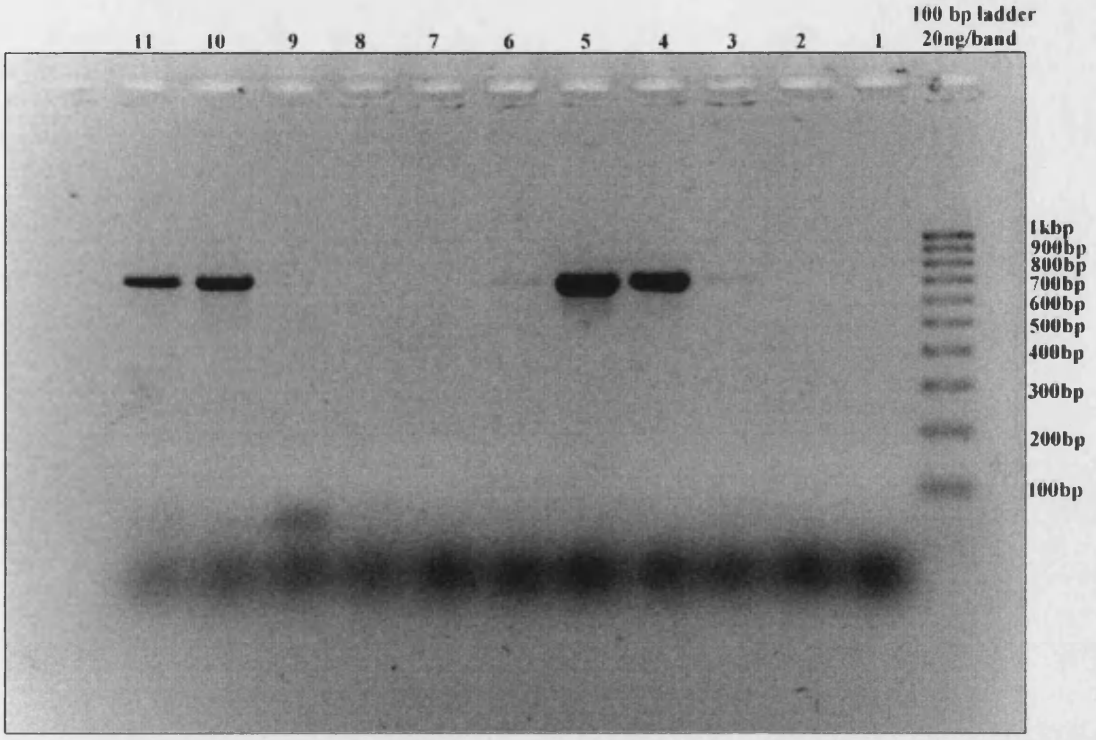



**Figure 3-10: Agarose gel showing PCR products from pHKY603 under conditions of increasing [$MgCl_2$]**

1.2% agarose gel run @ 100V, 120mA for 1hr. 17.7ng Plasmid MiniPrep used for each reaction. Annealing temp. of 55°C used. In each lane the concentration of $MgCl_2$ was altered (25mM stock, final reaction volume of 100µl). Lanes: 1 – 0mM $MgCl_2$; 2 – 0.2 $MgCl_2$; 3 – 0.4mM $MgCl_2$; 4 – 0.6mM $MgCl_2$; 5 – 1mM $MgCl_2$; 6 – 1.2mM $MgCl_2$; 7 – 1.6mM $MgCl_2$; 8 – 2mM $MgCl_2$; 9 – 2.4mM $MgCl_2$; 10 – 2.6mM $MgCl_2$; 11 – 3mM $MgCl_2$.

From Figure 3-10, strong signal of the correct size (723bp expected weight of product) was visible at two distinct salt optima. A comprehensive activity test measuring signal in the presence of lysis & denaturing agents was performed on induced ELTRP-1 culture. Readings were blanked against those from uninduced culture in the presence of the identical reagents. Results can be seen in Figure 3-11, refer to Section 3.2.5 for exact protocol.

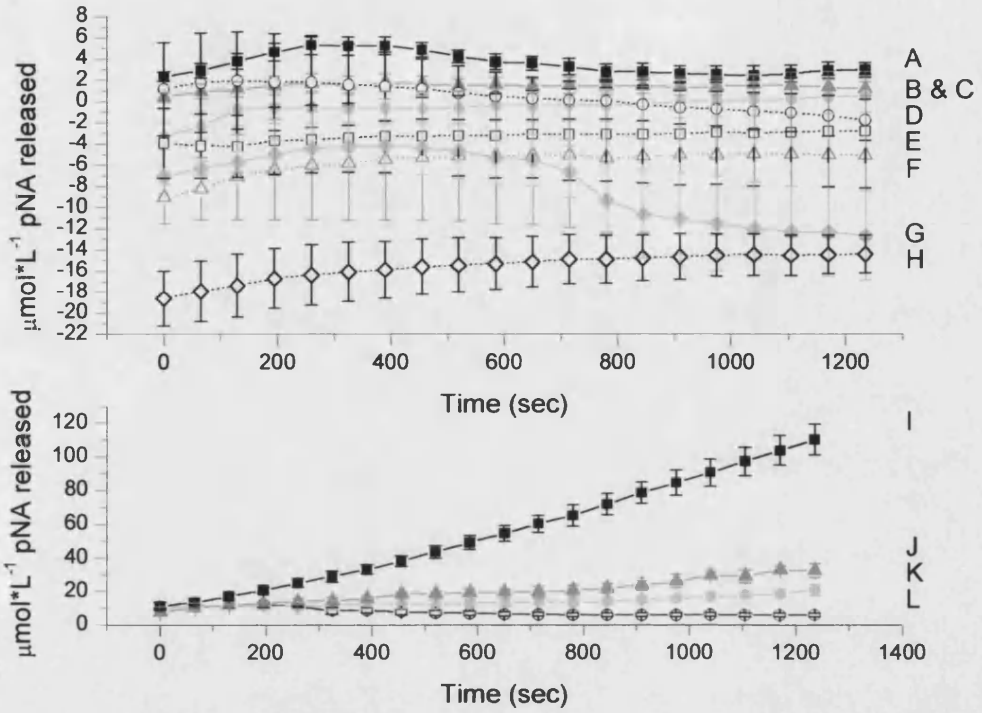



**Figure 3-11: Activity of ELTRP-1 on 1mM L-BAPNA in various conditions, 4hrs after induction**

Cultures were grown at 37°C. 200rpm, and induced after 3hrs ($A_{600}$ = 0.85, blanked against uninfected medium). All samples were incubated at room temperature with the various reagents & gently shaken to prevent settling. Addition of substrate was performed immediately prior to reading. All activities were calculated after blanking against the specific uninduced sample blank. All experiments were performed at least in triplicate.

| Experimental Key (From Figure 3-11). | | U (µmol pNA released*min$^{-1}$ /ml$^{-1}$ |
|---|---|---|
| Upper graph: | | |
| Basic activity (no lysis, no denaturant): Induced culture + L-BAPNA | **A** (—■—) | $1.91 \times 10^{-6}$ |
| Periplasmic activity: Induced culture + L-BAPNA + 0.5µg/ml lysozyme | **B** (—▲—) | $2.48 \times 10^{-5}$ |
| Total soluble activity: Induced culture + B-PER + L-BAPNA. | **C** (—●—) | $1.58 \times 10^{-4}$ |
| Insoluble activity: Induced culture + B-PER + 200µg/ml lysozyme + 0.1M urea + L-BAPNA | **D** (--○--) | $-1.83 \times 10^{-4}$ |
| Insoluble activity: Induced culture + B-PER + 200µg/ml lysozyme + 0.5M urea + L-BAPNA | **E** (--□--) | $8.20 \times 10^{-5}$ |
| Insoluble activity: Induced culture + B-PER + 200µg/ml lysozyme + | **F** (--△--) | $1.87 \times 10^{-4}$ |

| | | |
|---|---|---|
| 1M urea + L-BAPNA | | |
| Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 2M urea + L-BAPNA | G (—·—) | $-3.57 \times 10^{-4}$ |
| Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 5M urea + L-BAPNA | H (··◇··) | $2.05 \times 10^{-4}$ |
| Lower graph: | | |
| Control: 0.5μg commercial trypsin. | I (—○—) | $-3.37 \times 10^{-4}$ |
| Control: 1μg commercial trypsin. | J (—·△·—) | $4.15 \times 10^{-4}$ |
| Control: 2μg commercial trypsin. | K (—▲—) | $1.34 \times 10^{-3}$ |
| Control: 5μg commercial trypsin. | L (—■—) | $5.31 \times 10^{-3}$ |

## 3.4 Discussion

### 3.4.1 Strain identity

Eli Lilly provided no details that could explain the mis-performance of the first strain received. If it is assumed that the culture was correct when dispatched then the only explanation would be that the plasmid picked up a number of critical mutations before the glycerol stocks were made. A point mutation within the gene could distort the *Sac* I site and, if either an insertion or deletion could disrupt the reading frame of the protein, likely leading to early termination. A mutation forming a premature stop codon would have the same consequence, although neither of these theories would explain why the plasmid appeared approximately 1kb smaller than expected and also why the PCR consistently failed even under low stringency conditions. The spontaneous deletion of a significant portion of the gene could account for both results but would seem highly unlikely given the short time elapsed between receipt of the culture and the creation of stocks.

The two most likely explanations for the behaviour of the behaviour of the first ELTRP-1 strain were: a) that the strain received was a control strain used by the Eli Lilly molecular biology department, i.e. the plasmid was in its 'pre-alteration' state, correct in all essentials but without the addition of *met-phe-trypsinogen.* Such a strain (RV308:pBR322 variant) could have been shipped due to a confusion about the nature of the project.

b) That the glycerol stab cultures shipped were heterologous in composition, containing control variants and deletion mutations in suspension along with ELTRP-1. One of these could thus have been selected for the basis for this work due to identical resistances & (presumably) growth rates. It should be noted that, before returning to the company with the request for a second shipment of ELTRP-1, the original stocks were cultured three times, with

random colonies selected for propagation, in an attempt to culture a strain with the correct characteristics.

### 3.4.2 Testing for soluble r-trypsin

The Eli Lilly assumption that all expressed r-trypsinogen was sequestered as inclusion bodies, with none reaching either the cytoplasm or periplasm, or leaking into the supernatant upon individual cell death was shown to be correct. The metabolic load placed upon the culture on induction was shown to temporarily depress growth, but overnight fermentation allowed the cell density to achieve the same level as the uninduced control, showing that no autotoxicity was evident. A degree of autotoxicity may possibly be expected if even minute quantities of correctly folded and thus soluble & active trypsinogen were expressed into the soluble fraction. In mammalian cells, the route of activation of the zymogen is via enterokinase secreted from the surfaces of the pancreatic cells, however autoproteolysis, and thus autoactivation will occur at a comparatively slow but appreciable rate (Fersht, 1999).

The inflexibility in the expression vector (essentially a simple on-off switch, with no gradation possible) relates to its function as an industrial production strain. The heat-activated promoter allows a simple, cost effective mechanism of induction without the addition of reagents to the primary vessel, but was designed for maximal yields of an inactive & insoluble protein.

The total lack of trypsin activity present in the soluble fractions of the ELTRP-1 cultures made it seem unlikely that any increase in solubility could be effected by simple alterations in fermentation conditions. The nature of the promoter meant that, essentially, the only parameter that could be altered was the point of induction, but initial work in this direction (induction after 2hrs in LB, the beginning of exponential growth) had shown this to be an unpromising approach. Two lines of research presented themselves for the problem of generating soluble, active protein by simple methods suitable for a high-throughput screen, these were: a) the design of a microplate-based procedure for the solubilisation & refolding of r-trypsinogen (see Chapter 4); and b) the use of molecular methods to attempt to engineer translocation to the oxidising environment of the periplasm by means of a signal sequence, or transfer of the gene to a vector with a more flexible promoter system (see Chapter 5).

## 3.5 Conclusions

The now successful amplification, coupled with the plasmid size & digestion and, crucially, the apparent overexpression of a 24kDa protein upon induction, all suggested that the strain was a) the Eli Lilly commercial production strain ELTRP-1 and; b) was genetically dissimilar to the first strain tested.

Fractionated SDS-PAGE (Figure 3-9) indicated that protein was accumulating as insoluble inclusion bodies in the course of a standard fermentation. Comprehensive activity testing revealed no active enzyme in any fraction, and the rigid induction controls (linked to growth conditions) gave no indication that this could be altered.

# 4 High-throughput refolding of r-Trypsinogen from ELTRP-1

## 4.1 Introduction

### 4.2.1 Aims

The failure to achieve active r-trypsin from the soluble fraction of ELTRP-1 (Chapter 3) left only two choices available: the use of an alternative expression strain which would allow for comprehensive manipulation of fermentation conditions; or the development of a high-throughput methodology for the solubilisation and refolding of the insoluble protein. Eli Lilly had specified that they would prefer all options with the original host strain to be exhausted before alternative vector:host systems were explored, and so work proceeded with the aim of standardising a microplate-based refolding screen for trypsin activity.

### 4.2.2 Inclusion body formation by recombinant proteins

In the early nineteen seventies, C.B. Anfinsen published a series of experiments which established as fact one of protein sciences' then most promising theories (Anfinsen, 1973). He demonstrated the reversible denaturation of two small proteins, namely ribonuclease A and *staphylococcal* nuclease, specifically their spontaneous refolding into an active conformation after chemical or thermal denaturation in the absence of any additional cellular protein. This proved that the information for both the activity and folding pattern of these simple structures was encoded within their amino-acid sequence, seemingly providing a neat path from the native protein back through the primary sequence to the gene itself. The folding of native and recombinant proteins into desired conformations is of major significance to the modern bio-industries, and over two decades of work has shown that, if a set of general principles does exist by which proteins encode their individual folding pathways, then its complexity and interactions with the native environment defy a simple, universal set of guidelines (Dobson, 2000).

In a prokaryotic expression system, the formation of inclusion bodies by a recombinant protein is a precipitation reaction indicating extreme mis-folding of the nascent polypeptide. Simply put, recombinant proteins enter such a state as they have evolved to fold into their correct conformation within a specific set of native conditions, and these conditions are frequently

incompletely met in their new host. Examples of such conditions would be the absence of specific chaperone groups (Leroux and Hartl, 2000), the absence of required modifying or translocation pathways (Kohno, 1990), the overexpression of the protein at a temperature higher than that of its native strain (Kohno, 1990), and attempting to fold an extracellularly-targeted protein within the reducing environment of the cytoplasm (Creighton, 2000). In certain circumstances, conditions may be deliberately manipulated to produce an accumulation of insoluble, inactive enzyme, as eukaryotic proteins can frequently prove toxic to bacterial systems in overexpressed concentrations. From an engineering perspective the formation of inclusion bodies can prove beneficial as these highly concentrated and relatively pure protein granules can be easily purified as the insoluble fraction from a fermentation culture.

In designing an industrial strategy for isolating & refolding a recombinant protein that has precipitated into inclusion bodies, there are three steps to consider. Firstly, the protein aggregates must be at least partially purified by centrifugation, and then solubilised in a suitably denaturing solution. The second step required that the solubilised protein be diluted into an environment that favours the formation of the native, active structure. Finally, as no biological process, particularly an artificially induced one involving many thousands of potentially stable transition states, can ever be 100% efficient, a method is required by which the residual, partially incorrectly folded protein is removed. An example detailed in this thesis would be the refolding and purification of biosynthetic human insulin, discussed in Section 1.8.4.

Cytoplasmic solubility of recombinant proteins has been achieved via random mutagenesis where the substituted amino acid was a hydrophobic and/or bulky surface-exposed loop residue but this approach is uncertain by definition. Successful rational approaches include the use of fusion proteins, e.g. thioredoxin (enhance solubility and ease of purification) or signal-sequence tags (e.g. hisJ, pelB) which cause the polypeptide to be translated into the oxidative environment of the periplasm; stimulating co-overexpression of chaperone proteins via a particular stress-response (e.g. heat, starvation); and lowered rates of expression (Kim, 1999; Kohno, 1990; Kung, 1990).

### 4.2.3 Strategy required for processing of r-Trypsinogen from ELTRP-1 culture

In the ELTRP-1 strain, the inflexible two-state promoter precluded any improvement in the solubility of r-Trypsinogen by means of reducing expression or fermentation temperature (as discussed in Chapter 3) and thus a simple solubilisation and refolding protocol was sought, based upon that already in use by Eli Lilly (Buswell *et al.*, 1995). While a method comprising of all the three steps mentioned in the above section posed no problem for large-scale industrial cultures, a high-throughput microplate-screen demanded both a simple and reproducible assay, and dilution steps into secondary plates increase both experimental errors and processing time. The limitations imposed upon the method by the need for a high-throughput screen were that a minimum of process steps and dilution-replica plates be used and, equally significantly, that all processing of the sample be done within a microwell format. This severely limited the number of techniques by which the inclusion bodies could be purified from the other cellular fractions and fragments, effectively blocking the use of filtration and limiting centrifugation to the force that could be withstood by a microplate. The sensitivity of the L-BAPNA assay was also an issue, with the production levels of r-Trypsinogen from ELTRP-1 not yet established and a possible progressive series of dilutions necessary to allow correct folding, it was essential to monitor protein concentration at each step to ensure that it remained within detectable limits.

### 4.2.4 Eli Lilly protocol for rTrypsin purification, solubilisation & refolding

The industrial protocols by which Eli Lilly isolated & cleaned the inclusion bodies from an ELTRP-1 fermentation, and detailing the solubilisation and subsequent refolding of rTrypsin were not released for reasons of confidentiality. The composition of the solubilisation & refolding buffers could be inferred from a paper written by a chief Eli Lilly engineer in association with the University of Cambridge (Buswell *et al.*, 1995). Inclusion bodies formed of recombinant trypsinogen were solubilised in a urea buffer (5.5M urea, 10mM EDTA, 100mM cysteine, pH 9.5, 8°C) for 1hr. Refolding was initiated by injection of solubilised material into a 100ml reaction vessel with a 4-blade Rushton impeller containing the refolding buffer (50mM $CaCl_2$, 5mM Tris, 3mM cysteine, 1mM cystine, pH 9.0, 8°C). The study was designed to test the importance of mixing as a parameter in the refolding process, varying the impeller between 60 and 300rpm. Their results showed that the speed of the impeller did not appear to affect the yield of native trypsinogen, but

that the yield was improved by increasing the concentration of solubilised trypsinogen in the initial buffer. Significantly, it was not mentioned in the paper to what degree the inclusion bodies were purified prior to solubilisation, and also the dilution factor used for the refolding work was not stated. In these apparently optimised conditions, the researchers achieved a yield of native trypsinogen of 11.0 - 11.6% (on stirring range of 60-300rpm) from a solubilisation buffer of 1.4mg/ml pure trypsinogen. When the concentration in the solubilisation buffer was increased to 2.8mg/ml, the refolding yield rose to 13.6 - 14.5%. The dilution factor was altered according to the concentration of solubilised protein so that the final concentration of enzyme in the refolding buffer was held at 60μg/ml. The team concluded that the denaturant (urea) had possibly a negative impact on the refolding of the zymogen and that an increased dilution factor was thus beneficial.

### 4.2.5 Solubilisation & refolding strategies

Insoluble aggregates of overexpressed proteins can be solubilised by a variety of denaturing agents. The use of elevated temperatures, detergents and high concentrations of inorganic salts or organic solvents have been proven as effective agents in breaking down most forms of secondary or tertiary organisation in a protein, resulting in a soluble, disorganised polypeptide (Fersht, 1999). Low or high pH has also been used, possibly acting to increase electrostatic repulsion between protein molecules, thus discouraging aggregates (Kohno *et al.*, 1990). The 'denaturing strength' of an agent varies with the individual nature of the molecule, but as a rough guide, the following may be arranged in order of decreasing effectiveness: Guanidinium chloride (GdmCl)>urea>SDS>high temp>pH shock>high hydrostatic pressure. The presence of a strong reducing agent is also necessary to ensure full disassociation of all disulphide bonds, and DTT or β-mercaptoethanol is usually employed in millimolar quantities.

The key requirements for the refolding of a protein to be initiated are that aggregates have been dissolved and that all disulphide bonds have been oxidised. The denaturing agent would then be removed by means of dialysis or progressive dilution and the redox conditions of the buffer changed so as to encourage dynamic interactions between cysteine residues. Disulphide bonds act to stabilise tertiary structure and thus conditions should allow them to form as the protein passes through transition states, but each cysteine could conceivably bond with any other, holding the protein in mis-folded state, and thus a couplet of reduced & oxidised thiols would be used to

promote the exchange of bond partners. The appropriate ratio of the disulphide pairs and their oxidised counterpart must be determined empirically. Commonly used conjugate thiols include oxidised glutathione (GSSG) & glutathione, cysteine & cystine, and reduced & oxidised DTT.

**Solubilisation methodologies**

Initial work focused on using GdmCl as the denaturing agent for the inclusion bodies of rTrypsinogen derived from ELTRP-1 culture. Standards of commercial trypsin added to the buffers then demonstrated near total loss of activity, indicating that GdmCl was a potent inhibitor of bovine trypsin. The reason for this effect was obvious upon study of the molecules' structure, as it possess a strongly basic guanidinyl group, identical to the side-chain of arginine, and it is this moiety which binds to the base of the substrate pocket of the enzyme. While extensive dialysis could possibly have removed the GdmCl from the enzyme-inhibitor complex, depending on the strength of the salt-bridge formed, for a microplate-based assay the molecule had to be considered an irreversible inhibitor and subsequent work focused on the use of urea as the denaturant.

The solubilisation step was not believed to be temperature sensitive, but the procedure was done at 8°C so as to imitate the work done by Buswell *et. al.* (Buswell *et al.*, 1995). Factors to be investigated in this step were the optimal concentration of solubilised protein and the effectiveness of the buffer at solubilising relatively crude inclusion bodies from bacterial culture.

Although the final form of the assay was to be fully microplate-based, and could thus involve no purification steps, i.e. the inclusion bodies would not be able to be separated from the cellular lysate before addition of the solubilising buffer, for the initial experiments it seemed sensible to work with cleaned inclusion bodies. Successful optimisation of solubilisation and refolding protocols for purified and washed inclusion bodies would be taken as proof of concept for the refolding of partially purified recombinant trypsinogen.

**Refolding methodologies**

Temperature was known to be a critical factor in the kinetics of protein folding and the work was thus performed at 8°C and room temperature for comparison. Additional factors, which needed to be investigated in this step, were optimal dilution factor for the solubilised protein and also the rate of protein loss at low concentrations. Early work had shown a slight, but steady, loss in protein in

the microwells over time. This problem was insignificant prior to the dilution step, but after dilution, when protein concentrations were near the lower detection limits for the activity assay, any loss needed to be minimised. Two alternative plates were thus investigated in addition to the work done in the standard plastic microplate. One was the standard Starstedt plate which was blocked (in a manner similar to the blocking of a plate prior to an ELISA) by incubation overnight at room temperature with a 1mg/ml powered milk solution. The second plate tested was an Ultra Low Binding plate (Colstar), the wells of which were coated with a high-water content gel so as to minimise protein adsorption to the plastic.

### 4.2.6 Importance of co-factors in protein refolding

A successful *in vitro* refolding process is one in which the researcher successfully provides the solubilised protein with conditions which mimic those of its native environment. While in particularly intractable cases this may involve the addition of chaperone proteins, the literature shows that, in the main and when coupled with a suitable dilution step, the approximation of the appropriate chemical environment and careful manipulation of the redox balance is sufficient to refold a proportion of the solubilised protein. Simple and inexpensive methods for improving the yield by approximating the host environment can include adding trace amounts of the enzymes' natural substrate to the buffers, empirical adjustment of pH, ensuring that any incorporated metal or mineral ions are present in excess and adding simple sugars and amino acids to the mix. The presence of a high-affinity calcium ion binding loop in bovine trypsinogen and the ion's documented importance for the conformation of the molecule (Bode and Schwager, 1975) dictate that calcium chloride, or an alternative source, be present in a refolding buffer. Optimal quantity appears to have been determined empirically in each case but is always in concentrations of 50mM or higher.

The second strategy for the improvement of *in vitro* refolding yields is to experiment with the addition of certain small organic and inorganic molecules, solvents and surfactants, which, while not an attempt to predict the cellular environment, have been reported to interact favourably with solubilised large molecules. The effects involved are frequently ill defined but it seems likely that most act to negate charges on molten loops & domains, overcoming repulsion effects in secondary and tertiary structure which could bar the molecule from certain transition states, and possibly minimising attractions between molecules, thus limiting aggregation. Literature suggests

that all these chemicals are highly specific in which protein refolding reactions they assist, acting as retardants with structurally similar proteins, and also that interactions between co-factors can dramatically enhance or depress effects Favourable co-factors must thus be determined empirically for the molecule of interest, with a factorial screen, in which the effects of each potential combination are analysed, being both the simplest and the most comprehensive method available.

### 4.2.7 Fractional & fully-factorial analysis of co-factor interaction

This section of the work was performed with James Myers, another EngD student from the Department of Biochemical Engineering, UCL, and built upon similar co-factor studies that he had performed using commercial lysozyme.

Design of the fractional and fully factorial screens was performed using Minitab, version 13.32 (Minitab Inc.). 11 potentially beneficial co-factors were identified from the literature and their individual effects and combined interactions were evaluated using a 2-level fractional factorial screen, i.e each factor was screened at 2 levels of concentration (1 & 0): a molarity which had proved to be beneficial in similar work and in their absence. A full factorial analysis of the interactions of the 11 factors would require $2^{11} = 2048$ separate experiments, a prohibitive number of runs even for a microplate based assay. A fractional factorial screen can reduce the number of runs in two ways, the first being to exclude certain factor-level combinations. This leads to certain results being approximated from analysis of related results sets and thus not being able to be evaluated independently. Co-dependant results such as these are said to be aliased, and the overall term is that certain aspects of the data set are confounded. The second method involves limiting the resolving power of the design, by exploring only the lowest orders of interaction, i.e. only interactions of two or three separate factors at a time would be tested. The potential beneficial effects brought from a combination of 4 or 5 factors would not be detected by such a screen, but the lower order data would be free of confounding effects.

For the 11 factor initial 2-level screen used here, only first order interactions were examined, i.e. no more than two factors tested in each experiment and, as the two-level design simply analysed the presence or absence of each factor, many of the potential interactions could be deduced from the control runs, further condensing the number of runs required. The final design

required 72 well experiments, performed as per section 4.3.6. Main effect & interaction plots against enzyme activity were generated using the Minitab programme.

## 4.3 Methods

### 4.3.2 Fermentation & induction of ELTRP-1

Conditions were held identical to those used in Chapter 3. Briefly, these were as follows. A single discrete colony was used to inoculate a 100mls of LB and M9SD in 500ml baffled shake-flasks (tet. 10μg/ml, strep. 100μg/ml). These were then grown at 32°C, 200rpm until the end of exponential growth (LB: 4hrs, blanked $OD_{600nm}$ of approximately 1.5; M9SD: 6hrs, blanked $OD_{600nm}$ of approximately 1.0), whereupon the temperature was raised to 37°C, 200rpm. The fermentation was then allowed to proceed overnight for approximately 16hrs.

### 4.3.3 Solubilisation protocol

The solubilisation buffer composition was derived from previous Eli Lilly work and also from additional literature (Buswell *et al.*, 1995; Fersht, 1999). Solubilising buffer (1x): 5.5M Urea, 100mM DTE, 10mM EDTA, pH 9.5. All reagents purchased from Sigma. The inclusion bodies were obtained and washed as follows. Cells from 1ml of induced overnight cultures of ELTRP-1 were pelleted by centrifugation at 13,000rpm for 15min at room temperature. The cells were then lysed by resuspension in 100μl neat B-PER and, after being allowed to stand for 10min with periodic agitation, were centrifuged again at 13,000rpm for 10min. For the initial work (Figures 4.1 - 4.3), this pellet was then resuspended in solubilisation buffer and diluted accordingly. For the second set of work (Figures 4.4 - 4.7), an attempt was made to purify the inclusion body fraction more thouroughly by means of repeated washings. The supernatant was disposed of and the pellet was resuspended in 100μl B-PER (at a 1:4 dilution with MilliQ water) and then re-pelleted under identical conditions. Finally the pellet was resuspended in either 100μl or 50μl of solubilising buffer and allowed to stand for 1hour at 8°C, with occasional vortexing. Protein concentration was measured by reading at $A_{280}$, blanking against sterile buffer.

### 4.3.4 Refolding protocol

The refolding buffer was derived from the same sources as the solubilising buffer (Buswell *et al.*, 1995; Fersht, 1999) and was composed as follows. Refolding buffer (1x): 5mM Tris, 50mM $CaCl_2$, 1mM GSSG, 10mM Glutathione, pH 9.0. All reagents purchased from Sigma. Prior to refolding, the low-binding plates (Colstar) were allowed to soak in refolding buffer for 2hrs at room temperature. To initiate refolding, an aliquot of the solubilised protein would be injected into the refolding buffer. The dilutions used for this work were 1:30 & 1:60 and so typically, 20μl of protein solution would be added to 580μl or 1180μl of refolding buffer. Care must be taken that, after dilution, the concentration of protein remained within the minimum detection level of the L-BAPNA activity assay. In an optically clear solution, which did not act to impair trypsin activity, enzyme concentrations of 1μg/ml could be easily detected over a 10 minute period on 1mM L-BAPNA (see Chapter 2). The temperature at which refolding occurred was typically held at 8°C, but higher temperature were also experimented with (indicated at specific points within Results).

Refolding was allowed to proceed for a number of hours, with samples being taken for assaying usually after 2, 4 & 16 hours. Protein concentration was recorded at each step by reading at $A_{280}$ as before.

### 4.3.5 Activation and assaying

Samples of possibly refolded r-trypsinogen needed to be activated by removal of the prosequence prior to assaying. The prosequence was removed by addition of either commercial trypsin (to a final concentration of 0.5μg/ml) or enterokinase (to a final concentration of 5μg/ml). Both enzymes were purchased from Sigma. Samples were then allowed to stand for 5 minutes at room temperature before mixing with a 2x substrate solution of 2mM L-BAPNA (Sigma). Reading proceeded immediately at $A_{405}$, blanking against substrate buffers with activating enzymes added. In the experiments depicted in Figure 4-2 - Figure 4-5, concentration of enzyme in sample was determined by running a range of standards of commercial trypsin in the refolding buffers, but it is important to note that these standards were not subjected to the refolding process. The concentration of active enzyme detected in these experiments has been multiplied by the dilution factor to allow direct comparison, in terms of refolding yield, with the initial solubilised protein concentration.

### 4.3.6 Experimental procedure for co-factor evaluation

The effects of the following cofactors were evaluated on the refolding of pure, commercial trypsin (Sigma): L-Arginine (Sigma), stock solution = 0.574M, used at 0.250M; Ammonium sulphate (Sigma), stock solution = 1.136M, used at 0.5M; Sucrose (Sigma), stock solution = 0.876M, used at 0.250; Glucose (Sigma), stock solution = 0.832M, used at 0.5M; Glycerol (Sigma), stock solution = 0.217, used at 0.05M; PEG300 (Sigma), stock solution = 0.067M, used at 0.0201M; PEG3350 (Sigma), stock solution = 0.006M, used at 0.00201; Ethanol (Sigma), neat stock, used at 20% v/v; L-BAPNA (substrate)(Sigma), stock solution = 3.33M, used at 0.333M; 2-Pentanol (Sigma), stock solution = 0.114M, used at 0.01M; Cyclohexanol (Sigma), stock solution = 0.101M, used at 0.01M.

For each experiment, commercial trypsinogen (1mg/ml in MilliQ water) was solubilised in the solubilisation buffer (Section 4.3.3 above) @ 17°C and was allowed to stand for two hours. 30µl of the enzyme solution was added to each buffer well (filled to a volume of 300µl) and enterokinase (Eli Lilly) was added to a final concentration of 1mg/ml. The wells were allowed to stand for 5½ hours @ 17°C with occasional gentle agitation to prevent settling. Activity assays were performed by the addition of 100µl of well solution to 100µl of 4mM L-BAPNA and commencing reading absorbance at 405nm immediately.

## 4.4 Results

### 4.4.2 Protein concentrations & activity after refolding

The interactions of initial inclusion body concentration, dilution factor (1:30, 1:60, 1:90 & 1:120) and temperature (8°C, 20°C) were investigated using methodologies described in Sections 4.3.3 & 4.3.4. Care was taken that, after dilution into refolding buffer, protein concentration did not fall below 6µg/ml (3µg/ml being minimum detection limit of L-BAPNA assay). This limit was thought necessary as the 'purity' of the inclusion bodies was not known and this would allow trypsin activity to remain detectable if 50% of the aggregate was comprised of heterologous protein. One problem that became immediately apparent was the progressive drop in protein concentration over the course of the solubilisation and refolding incubations, due presumably to adhesion of protein to the plates, and thus the effects of using 'blocked' microplates and low-binding plates were investigated (See Section 4.2.5). Enzyme activity was calibrated by means of commercial enzyme standards

run in refolding buffer, but these standards were not subjected to the solubilisation/refolding process, i.e. were added only immediately prior to assay.

No activity was detected from the inclusion-body bound recombinant enzyme under any of the solubilisation/refolding conditions tested. The concentration of solubilised inclusion-body of a 1ml aliquot from a 100ml, 37°C fermentation, induced overnight, ranged from 1.9-2.7mg/ml when solubilised in 50µl of buffer. The use of 'blocked' and low-binding plates did prevent protein loss from solution (see Figure 4-1) and the use of the blocked plates did not result in an increase in background noise of the assay (a possible consideration).

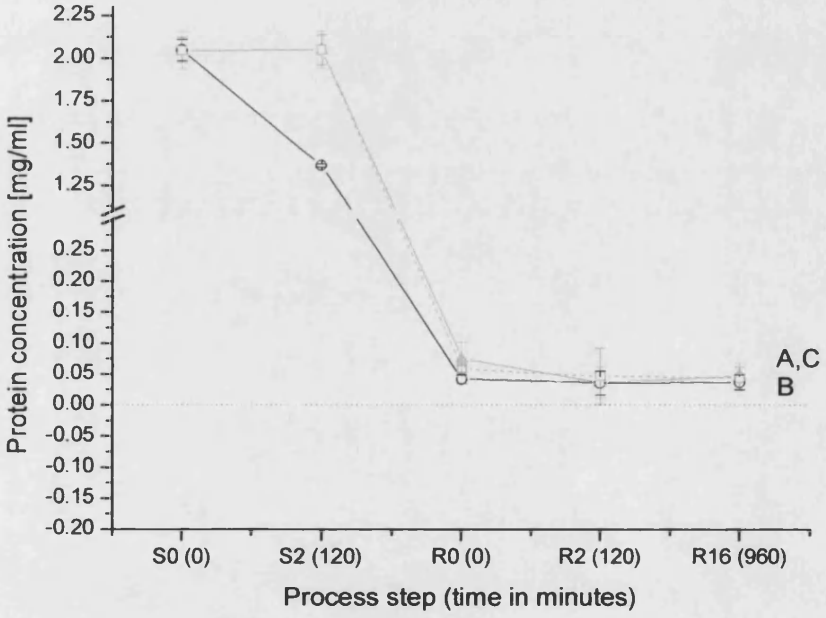



**Figure 4-1: Total protein concentration over 60-fold diln.**

Investigation of protein concentration during a solubilisation & 60-fold dilution experiment at a variety of protein concentrations, and using blocked & un-blocked plastic microplates and the Costar low-binding plate. Key: 'S' refers to incubation time in solubilisation buffers, 'R' refers to incubation time in refolding buffers. A (—▲—) = refolded in blocked plates, activated by enterokinase; B (—o—) = refolded in un-blocked plates, activated by enterokinase; C (···□···) =

refolded in low-binding plates, activated by enterokinase. The experiment was performed at 8°C. All expts. used the insoluble fraction from a ml of LB culture, resuspended in 100μl of solubilising buffer. Aliquots from all expts. activated by treatment with enterokinase before assaying. No activity was detected from any sample in this experiment and commercial trypsinogen added to blank (uninduced samples) and similarly processed also failed to generate signal.

In Figure 4-1, the average final protein concentration was in the region of 5.7μg/ml, sufficient to be detected by the activity assay. No activity, however, was recorded from any of the samples taken at Refold 0 (approximately 2 minutes after dilution), Refold 2 (2 hours after dilution) & Refold 16 (16 hours after start of refolding). Commercial enzyme standards run in the relevant buffers were unaffected.

### 4.4.3 Analysis of temperature, dilution & activation agent on process

Following the failure of this earlier work to generate any detectable signal from the recombinant refolded & activated trypsinogen, a semi-factorial series of experiments was planned to attempt to discover a viable set of operational conditions. Factors screened against each other were: the temperature of the process; the dilution factor of the refolding step; the concentration of the initial solubilised inclusion body sample; and the choice of activating enzyme. In graphs Figure 4-2, Figure 4-3, Figure 4-4 & Figure 4-5, the process steps referred to are as follows: Initial solubilisation of the washed inclusion bodies (S); the sample immediately upon dilution into the refolding buffer (in realistic terms, at least 2-3 minutes would pass before assaying) (R0); then the time in hours from R0, i.e. R1, R2, etc. The activity in the four graphs represents the blanked activity detected, multiplied by the dilution factor. As only levels of trypsin activity fractionally above the baseline were detected, it was assumed that no significant levels of substrate hydrolysis were occurring above background noise, a fact confirmed by the fact that correcting for dilution factor resulted in a staggered profile in each case (Figure 4-2).

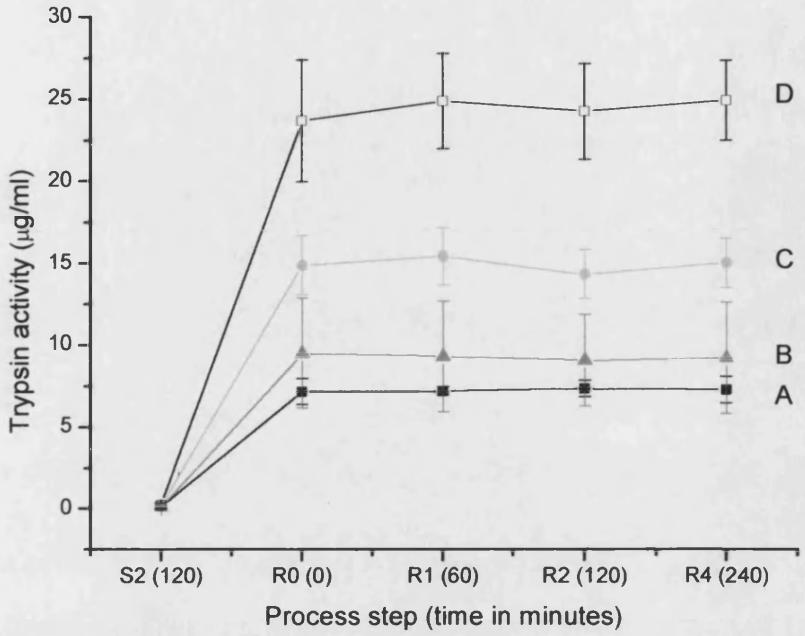



**Figure 4-2: Concentration of soluble inclusion bodies in a range of dilution factors for the refolding step**

Key: A (—■—) = solubilised concentration of 1-4mg/ml total protein, refolding dilution of 1:30; B (—▲—) = solubilised concentration of 1-4mg/ml total protein, refolding dilution of 1:60; C (—●—) = solubilised concentration of 1-4mg/ml total protein, refolding dilution of 1:90; D (—○—) = solubilised concentration of 1-4mg/ml total protein, refolding dilution of 1:120. All runs were held at 8°C for the duration of the experiment. Activity of active trypsin in the standards was established by comparison with a range of commercial trypsin standards added to the appropriate process buffers at the point of assay, i.e. the standards were not subjected to the solubilisation & refolding process. Dilution factors were corrected for in the determination of the enzyme concentration.

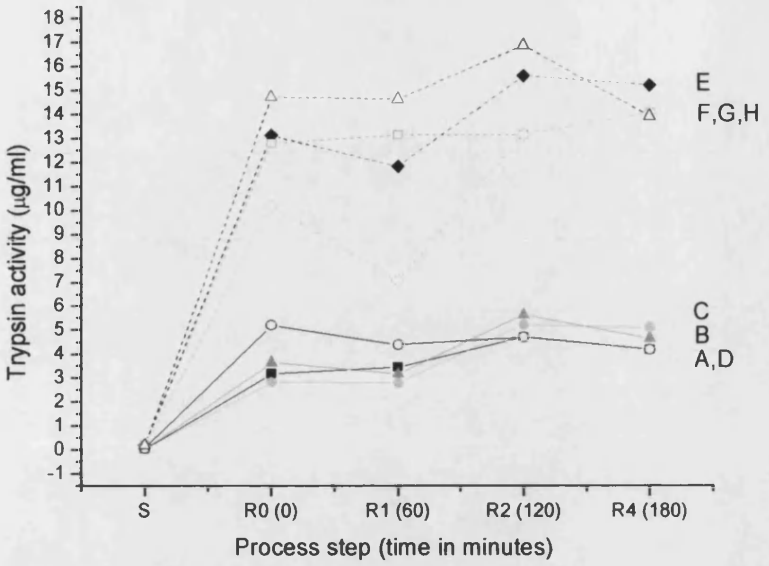



**Figure 4-3: Concentration of solubilised inclusion bodies vs. temperature of process**

Key: A (–■–) = solubilised protein concentration of 1mg/ml, refolded at room temperate (~ 20°C); B (–▲–) = concentration of 2mg/ml, refolded at room temperature; C (–●–) = concentration of 3mg/ml, refolded at room temperature; D (–○–) = concentration of 4mg/ml, refolded at room temperature; E (··◆··) = concentration of 1mg/ml, refolded at 8°C; F (··□··) = concentration of 2mg/ml, refolded at 8°C; G (··◇··)= concentration of 3mg/ml, refolded at 8°C; H (··△··) = concentration of 4mg/ml, refolded at 8°C. Concentration of active trypsin in the standards was established by comparison with a range of commercial trypsin standards added to the appropriate process buffers at the point of assay, i.e. the standards were not subjected to the solubilisation & refolding process. Dilution factors were corrected for in the determination of the enzyme concentration.

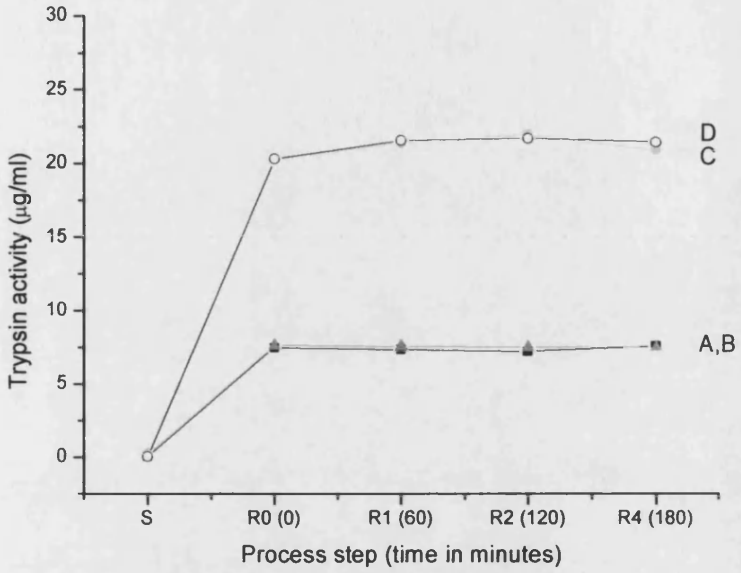



**Figure 4-4: Assessment of the effects of activation agents on the refolding of trypsinogen**
Key: A (–■–) = 1:30 refolding dilution, activated by trypsin; B (–▲–) = 1:30 dilution, activated by enterokinase; C (–●–) = 1:120 dilution, activated by enterokinase; D (–○–) = 1:120 dilution, activated by enterokinase. All experiments were held at 8°C for the duration of the process. Concentration of active trypsin in the standards was established by comparison with a range of commercial trypsin standards added to the appropriate process buffers at the point of assay, i.e. the standards were not subjected to the solubilisation & refolding process. Dilution factors were corrected for in the determination of the enzyme concentration.

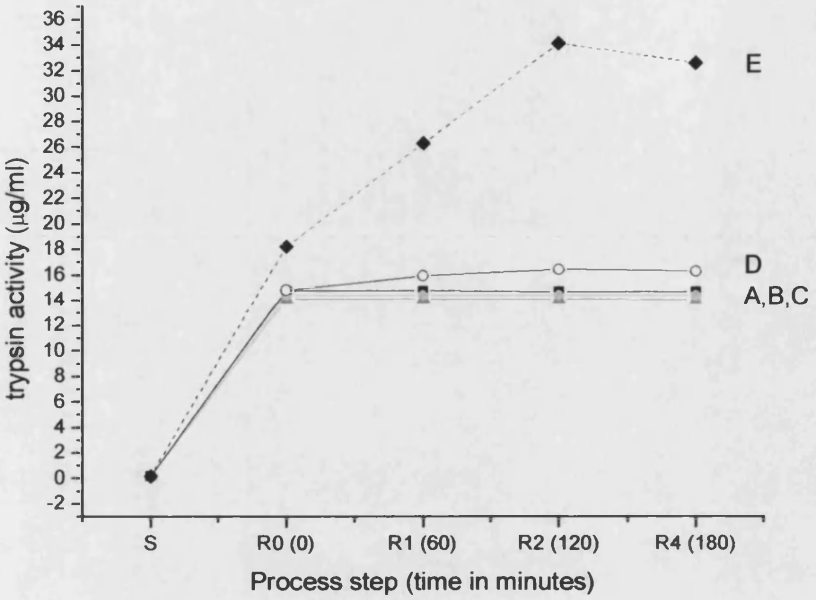



**Figure 4-5: Solubilisation & refolding of commercial trypsin**

Comparison of commercial trypsin standards subjected to refolding process with those added to the appropriate buffers at the point of assay. Key: A (—■—) = commercial trypsin solubilised at 30µg/ml; B (—▲—) = commercial trypsin solubilised at 60µg/ml; C (—●—) = commercial trypsin solubilised at 90µg/ml; D (—○—) = commercial trypsin solubilised at 120µg/ml; E (··◆··) = commercial trypsin solubilised at 300µg/ml. All experiments were held at 8°C and the dilution factor used for the refolding step was 1:60. Dilution factors were corrected for in the determination of the enzyme concentration

### 4.4.4 Refolding of recombinant trypsinogen

The series of experiments aimed to evaluate the process conditions required for the refolding of active recombinant trypsinogen from partially washed inclusion bodies purified from ELTRP-1 fermentation met with very little success. Figure 4.1 was intended to demonstrate that the protein concentration in each experiment remained above the lower detection limit for the assay after the various dilution steps, despite generating no detectable signal. Figures 4.2 – 4.4 describe a second set of experiments, conducted with extensively washed inclusion bodies, where the interactions of temperature, one-step dilution factors and initial soluble concentration were extensively examined

for any positive effect on the yield of refolded, active enzyme. Although trace activity levels, relative to unactivated trypsinogen, were recorded, these were in the order of 2.5-5% of initial inoculum (working on the simplistic assumption that recombinant trypsinogen comprised 100% of the inclusion bodies). The activity detected did not appear to be reproducibly affected by any single factor and extremes of dilution & temperature failed to produce any significant trends in the results. The conformity of the results (in the mean range of 0.2 – 0.5µg/ml), particularly in regard to their apparent insensitivity to the various dilution factors used, would seem to indicate that they are not significant, and do not represent the genuine refolding of any of the recombinant enzyme. Similarly, the activity apparently detected from the refolded commercial trypsin standards was only barely detectable above the baseline and should not be regarded as truly indicative of enzyme regeneration.

### 4.4.5 Factorial analysis of folding co-factors


**Figure 4-6: Co-factor main effect plot**

Main effects plot for activity of the 11 co-factors evaluated for the *in vitro* refolding of commercial trypsinogen. Graph shows individual effects of the presence (1) and absence (0) of each factor on the activity (and thus yield) of refolded enzyme. The dotted red line indicates the overall mean of the response data.

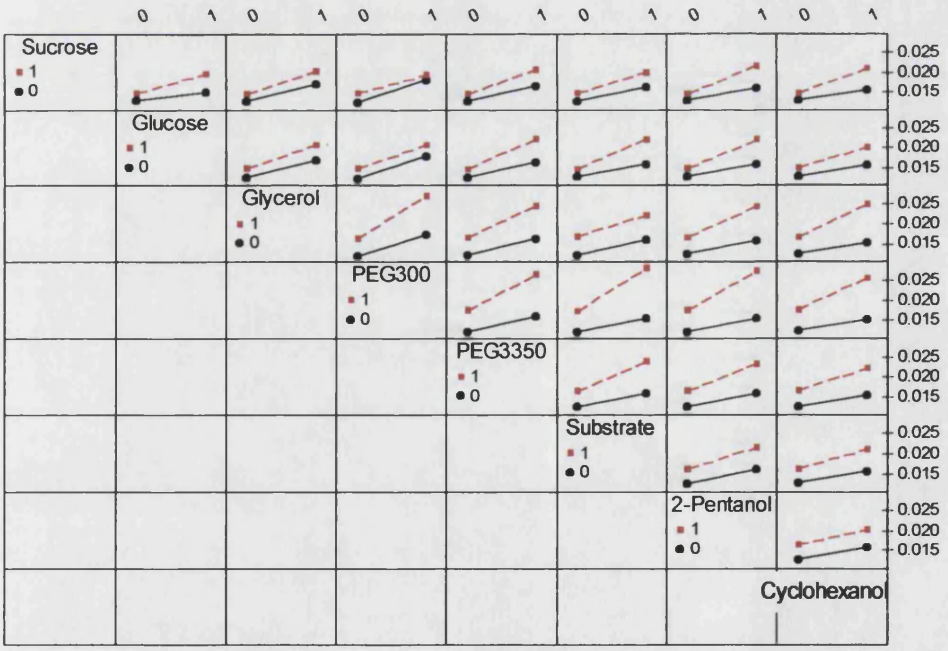



**Figure 4-7: Co-factor interactions**

Interaction effects of each potential co-factor plotted against the mean activity. An interaction plot shows the impact that changing the settings of one factor has on another factor. Interpretation is simple - for each cell: the name of the principle factor being evaluated is shown at the bottom, the

concentrations of which are shown on the X-axis at the top of the cell; the name of the interacting factor is shown to the left, with the concentrations being shown in the key immediately below it. A significant interaction between two factors would be revealed by two lines of differing slope.

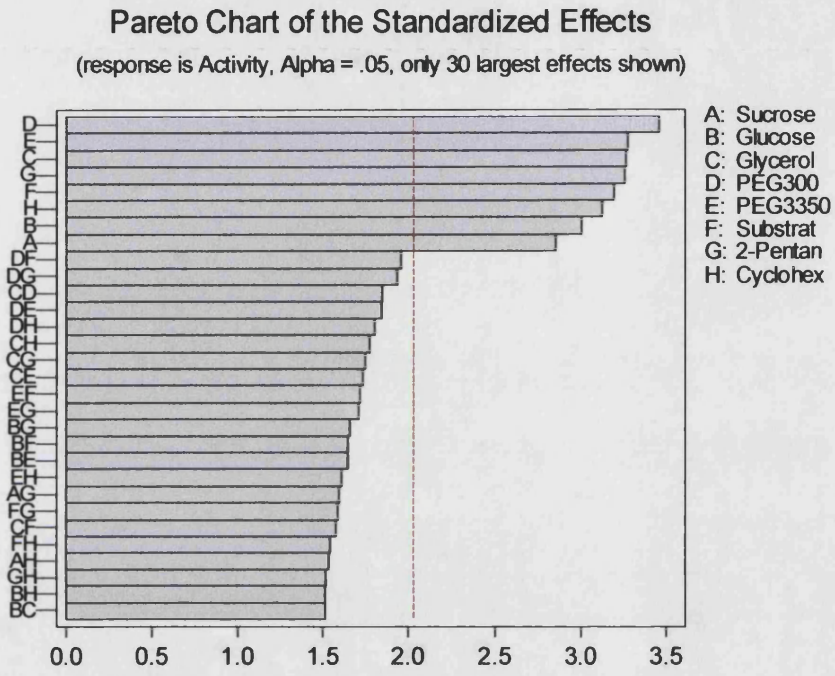



**Figure 4-8: Pareto chart of standardised effects**

A Pareto chart used to compare the relative magnitude and statistical significance of both the main and interaction effects. The dotted red line indicates the alpha value of 0.05m, all effects that extend past this reference line were held to be significant by the model. It is important to note that weighting and normal distribution of the model has been unbalanced by the exclusion of several large (negative) effects (see section 4.4.6).

### 4.4.6 Evaluation of refolding co-factors

The main effect plot illustrates the blanked activity of refolded & activated commercial trypsinogen in the presence of each of the 11 co-factors. Data means of the activity in the presence and absence of each factor were plotted and provided a rough guide as to which factors influence the response and the relative strength of the effects. A significant effect was present when the mean

response changed across the levels of a factor, and thus the pitch of the slope indicated the magnitude of the effect. The graph indicated that the presence of glycerol, PEG300 & PEG3350, substrate (L-BAPNA), 2-pentanol and cyclohexanol all had slight beneficial effects on the yield, but the magnitude of these enhancements were small when measured against the inhibitory effects of L-arginine, ethanol and ammonium sulphate. Caution should be taken when attempting to apply precise figures to each event due to the small number of replications involved and the fact that an individual set of controls was not run in each refolding buffer solution, and these results should only act to serve as a pointer for further work.

All results were fitted to a model that used an analysis of variance to determine which sets of effects were statistically significant. For the initial screen, the $\alpha$-level value (the probability of making a type I error, i.e. incorrect rejection of the null hypothesis) was set to 0.1 and only the three inhibitory effects were calculated to be significant in relation to the data set. Upon reduction of the model to discount those negative effects and an increase in the $\alpha$-level selectivity to 0.05 to compensate for the decreased range in the data, sensitivity was increased such that all the single factor main effects except for sucrose & glucose were deemed to be significant interactions.

An interaction plot was used to determine whether any two of the tested co-factors interacted in their effects on activity and to compare the relative strengths of the effects. The two lines of the graph illustrate the change in the activity mean from the low to the high level of a factor depending on the level of a second factor. The greater the degree of departure of the lines from being parallel, the stronger the interaction effect.

The Pareto graph (Figure 4-8) shows that the following 6 factor-factor interactions had the greatest impact on activity, ranked from highest downwards: PEG300 – Substrate (L-BAPNA); PEG300 – 2-Pentanol; PEG300 – Glycerol; PEG300 – PEG3350; PEG300 – Cyclohexanol; Glycerol – Cyclohexanol. From Figure 4-8, the co-factors which produced the strongest interactive effects when combined with any other factor (excluding L-arginine, ammonium sulphate or ethanol) were PEG300 and glycerol. With the $\alpha$-level discrimination set at 0.05, these effects lay just below the significance threshold, but it should be remembered that this was set at an arbitrarily high level in order to attempt to rank the data coming from an unbalanced design. The screening of selected beneficial co-factors in a full-factorial model, with suitable controls in each buffer, would achieve far more reliable discrimination between effects.

## 4.5 Discussion

### 4.5.2 Refolding of recombinant trypsinogen

The persistent failure of the refolding work, using inclusion bodies of recombinant trypsin from ELTRP-1 fermentation, to return an unambiguous signal was puzzling, particularly as it was designed to closely mimic a proven methodology (Buswell *et al.*, 1995). Deviations from the published methods were: 1) the use of the use of glutathione and GSSG as the redox exchange couplet in the refolding buffer in the place of cystine and cysteine; 2) the extraction and purification protocols for the inclusion bodies (which were not known, although if these follow the Eli Lilly standard for the primary recovery of recombinant enzymes then the steps involve high-pressure homogenization and then centrifugation, with subsequent washing of the inclusion body slurry with a 0.66% w/w Triton-X-100 detergent solution, 3) the addition of DTE to the solubilisation buffer in order to act as a reductant; and 4) undoubted differences in the dilution factor of the refolding step and the impracticality of implementing an injection process. To clarify, an injection process was used by the previous research team to initiate refolding by progressively 'injecting' small volume of solubilised enzyme into the refolding buffer, up to a set concentration. Such a method was not suitable for use in a high-throughput microplate screen as it would place a considerable demand on both operator precision and work load, and was thus not tested. The nature of screen required placed added limitations on the work, namely that potentially refolded protein could not be concentrated for assay purposes.

The results generated in those experiments based on partially purified & washed inclusion bodies (Figures 4.2 – 4.4) appeared to show slight activity from refolded enzyme, but should not be held as a proof of concept. Enzyme concentration data was multiplied by the dilution factor so as to provide an at-a-glance yield comparison with the initial inoculum, but only extremely small signals were generated by all experiments, frequently only barely distinguishable from the blank (refolded inclusion bodies without the addition of activating enzyme). The results generated were thus open to considerable doubt and could not be regarded as the basis for an assay.

The solubilisation and refolding protocols used in this work were, in themselves, unsuitable for a high-throughput screen, involving, as they did, a multiple high-speed centrifugation steps and a series of washes. The final form of the screen would have, of necessity, involved cellular lysis followed by the solubilisation of the inclusion bodies without any separation from the soluble

fraction (microplate centrifugation using available equipment having been shown to be unable to generate the required rcf to pellet inclusion bodies, see Section 2.3.5). It was predicted that additional problems would arise before the screen could be fully implemented in the screening of a mutant library, the most significant of which would be whether the specific level of insoluble enzyme produced by a 1ml microplate well-culture would be sufficient to be diluted and remain within the detection limits of the assay.

### 4.5.3 Potential alterations to methodology

The incompleteness of the bulk of this work is evident and it is possible that a workable refolding strategy for recombinant trypsinogen could have been developed with further careful testing. Additional work would include the optimization of the process using just commercial trypsinogen (instead of adding to an inclusion body mix), the testing of other oxidizing & reducing agents (notably cysine/cysteine) and experimentation into the optimal redox balance, and experimentation with steped dilutions. Whether any of these strategies would have resulted in a method suitable for the high-throughput screen described was not known, but time restraints on the work encouraged the exploration of other avenues to protein solubility, particularly the use of N-terminal signal sequences, with which promising work had been done with analogous enzymes (Clare *et al*, 2001; Evnin *et al.*, 1990) (see Chapter 5).

### 4.5.4 Evaluation of folding co-factors

While the lack of controls and an accurate standard curve in these screens (Figure 4-6 - Figure 4-7) mean that accurate values should not be assigned to the effects, the broad results of the tests appear valid and significant. Suitable controls would indicate whether ethanol & ammonium sulphate actually retarded the refolding process, but it is far more likely that the negative impact of L-arginine at least was due to binding to Asp189 of trypsinogen and thus the blocking of the binding pocket. The positive effects of PEG300 and glycerol on the refolding process were pronounced and were worthy of further study. The researcher with whom this work was performed had plans for a 3 level, full factorial screen of the effects of PEG300, glycerol, 2-pentanol and substrate on the refolding of commercial trypsinogen, coupled with a full suite of controls and standards run in each buffer. Such work will aim to fully elucidate the effects and interactions of

each factor, and may even make a microplate-based assay of recombinant trypsinogen from well-cultures a practical possibility.

## 4.6 Conclusions

Under the conditions tested, no signal could be detected from recombinant trypsinogen solubilised from unwashed inclusion bodies produced by a 16hr ELTRP-1 fermentation. Total protein concentration in the refolding buffers remained within the limits of activity assay detection at all points.

Use of the Costar Ultra Low-Binding plate or the blocked standard microplate demonstrated advantages over the use of an un-blocked standard microplate in that they acted to prevent the gradual loss of soluble protein by absorption during refolding incubations.

Work with inclusion bodies purified from the bulk of the insoluble fraction by repeated washing with B-PER solutions appeared to generate marginal signal from the refolded & activated trypsinogen, but the activity was slight and indistinct, barely distinguishable from the baseline, and, if genuine, would represent only a 2.5-5% refolding yield of active enzyme.

Refolding yield appeared sensitive to temperature, with work at 8°C producing results 2-3 fold higher than the near-undetectable signal generated by work conducted at room temperature. No other factors (initial soluble concentration, dilution of factor of refolding step, choice of activating enzyme) appeared to influence results in a reproducible fashion.

The addition of quantities of commercial trypsinogen to low-concentration inclusion body solutions, which were then solubilised and refolded under standard conditions, showed that the refolding efficiency of this process was extremely poor, even when increasing quantities of purified protein were added to the system. Slight signal was detectable from the added standards, but represented an increasingly low yield of the solubilised enzyme, with only 5% of the 0.3mg/ml standard apparently being regenerated.

The fraction factorial experiment revealed that purified, commercial trypsinogen appeared considerably more amenable to refolding, responding to a variety of co-factors with a clear increase in signal. PEG300 and glycerol were shown to be highly beneficial to the reaction, producing the greatest effects when used singularly, but with significant effects when combined either with each other or with the bulk of the remaining factors.

Ethanol, L-arginine and ammonium sulphate were shown to be either detrimental to the refolding reaction or inhibitory to the action of enterokinase or trypsin.

# 5 Signal sequence work – Strain construction

## 5.1 Introduction

### 5.1.1 General strategy

As it did not appear possible to achieve correctly folded recombinant trypsinogen from a culture of ELTRP-1 by means of a microplate-scale solubilisation & refolding protocol (Chapter 4), three remaining routes of manipulation remained open. The attachment of a signal sequence (a short N-terminal polypeptide which directs the translocation of the unfolded protein into the periplasm) to *r-trypsinogen* in pHKY603, the ligation of the entire target gene into a vector more suitable for experimental expression work (the pET series from Novagen would allow variation of expression rates independent of temperature, which could be used as an additional variable), or the subjection of the entire gene to random mutagenesis with the aim of screening via activity for any increase in solubility. Eli Lilly had expressed the preference that the gene remained in the original strain and plasmid until all other options had been exhausted, so as to simplify their evaluation and possible use of any enzyme mutants generated. The construction and insertion of a suitable leader sequence into pHKY603 was thus presented as the next logical step.

This chapter details the synthesis and testing of three plasmid constructs: pHKY603(T) & pHKY603(M), which consisted of a pelB leader sequence inserted at the N-terminus of the genes for the zymogen and mature protease respectively in the original plasmid; and pET(T), where the gene for mature trypsin was inserted into a pET26b(+) plasmid and expressed in the BL21(DE3) lysogenic strain. The aim in all cases was to achieve soluble, active enzyme from a conventional fermentation with a minimum number of processing steps. Stable expression of detectable levels of mature trypsin would represent a significant benefit, as it would remove the need for an activation step and could allow for the implementation of a nutritional selection strategy.

### 5.1.2 Signal sequence theory

A signal (or leader) sequence could be defined as a short polypeptide which directs the translocation of a nascent or completed protein into or across the plasma membrane (Martoglio & Dobberstein, 1998). These relatively simple peptides are specific enough to target a protein

towards a number of distinct membrane bound transport mechanisms and contain the information of whether the protein was to be inserted into the membrane or fully translocated across it. While signal sequences show a high degree of sequence diversity, they are characterised as a group by possessing a common pattern of charge and hydrophobicity . A generic leader sequence is illustrated below in Figure 5-1. Signal peptidases have a tripartite structure consisting of: a central hydrophobic h-region; and hydrophyllic N- & C-terminal regions (Martoglio & Dobberstein, 1998), referred to as the n- & c-regions respectively The sequences may range between 15 to over 50 residues in length, with the polar and positively charged n-region being responsible for most of the variation. If the precursor protein were targeted for translocation across the membrane then the polar c-region would contain a further conserved motif at its point of contact with the protein targeted for export. The -3 and -1 positions always appear to contain a small, uncharged residue, such as alanine or lysine, with a larger residue, e.g. methionine, at the -2 position. This 'small-large-small' pattern forms the recognition site for the signal peptidase, which allows the release of the attached protein. Immediately upstream from this motif lie varying numbers of proline and glycine residues which serve to break the helical conformation of the chain, presumably to provide a more accessible site for the signal peptidase.

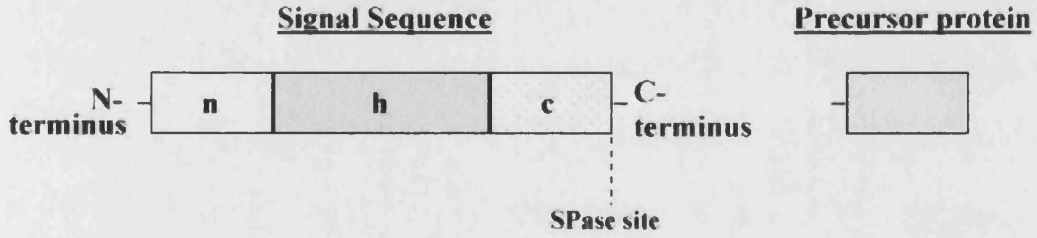



**Figure 5-1: Illustrating the tripartite structure common to all signal sequences**
The hydrophobic h-region is flanked by the hydrophilic, potentially membrane translocated n- & c-regions. The SPase cutting site is located at residues -1 → -3 of the c-region.

### 5.1.3 Expression of recombinant proteins into the periplasmic space

Work done by Evnin *et al.* (Evnin & Craik, 1988) demonstrated that the attachment of a signal peptide to the N-terminus of rat anionic trypsinogen would allow the transport of the recombinant protein across the plasma membrane and subsequent folding and accumulation in the periplasm in sufficient levels to allow a selection strategy and activity screens. The signal sequence used was *hisJ* and a controlled level of expression of the protein was achieved by use of a *tac* promoter within a pT3 plasmid. The work proposed in this project would attempt to duplicate these results using, initially, a *pelB* signal sequence and a pET-20b (+) ribosome binding site, both inserted into the Eli Lilly production strain vector pHKY603. If similar results were obtained the formation of insoluble inclusion bodies could be greatly reduced and the protease pre-cursor could be harvested from the periplasm using a 25% sucrose/10mM Tris-HCl, pH8.0/5mM EDTA solution as described by Vásquez *et al* (Evnin *et al.*, 1990) or by established variants of this extraction protocol (Clare *et al.*, 2001; Cooke *et al.*, 2001).

The periplasm of gram-negative bacteria, such as *E. coli*, differs profoundly in its chemical composition from the rest of the cell in that it is intimately linked to the external environment, due to the permeability of the outer membrane to small organic molecules. As a direct result of this, the periplasmic space cannot be said to have a regulated pH or maintained ionic balance, and contains no thiol or disulphide compounds (Creighton, 1992). Such a system tends towards an oxidizing environment, thus encouraging the formation and dynamic interchange of disulphide bonds and making it ideal for the establishment of the correct conformation of densely bonded, robust extracellular catalytic enzymes. A class of periplasmic chaperones, termed DsbA to G, all containing the relatively unstable bond-forming thioredoxin fold, have recently been characterized, and may also interact with translocated recombinant proteins (Van Straaten *et al.*, 1998). In their 1988-90 work Evnin & Craik (Evnin and Craik, 1988; Evnin *et al.*, 1990) supposed that it was the reducing environment of the cytoplasm that retarded the correct folding of rat trypsin by preventing the formation of disulfide bonds. Allowing the protein to accumulate in the cytoplasm would also leave it vulnerable to the wide varieties of cytoplasmic proteases and, conversely, may help to limit any autotoxic effects arising from the high-level expression and proteolytic action of the mature enzyme. Added to this, the fact that protein localised in the periplasm would be considerable easier to extract, it seemed clear that, for the purposes of this work, a similar strategy should be pursued.

### 5.1.4 Design of constructs: pelB, pHKY(M), pHKY (T)

The signal sequence chosen for this work is pelB from the pET-20b (+) plasmid (Novagen). This particular leader sequence was selected simply on the basis that its action was well characterised within *E. coli* and there was considerable in-house experience with its use. The signal sequence was to be inserted as a sticky-ended oligonucleotide between the *Xba* I & *Nde* I sites of the *r-trypsinogen* gene. The oligonucleotide insert was designed so that, upon ligation, the signal peptidase recognition site of the leader sequence will lie immediately upstream of the first methionine codon of the gene, thus preserving the reading frame and avoiding the addition of any 'foreign' amino acids at the N-terminus of the protein after signal sequence cleavage. The ribosome-binding site to be used was copied from the pET-20b (+) plasmid and preserved in the same position relative to the start codon of the signal sequence.

The pelB insert was formed by the slow annealing of 6 oligonucleotides to form a 85bp 'sticky ended' construct, as shown below (Figure 5-2). The sequence of the leader sequence & of the ribosome binding site were taken from pET20b (Novagen).

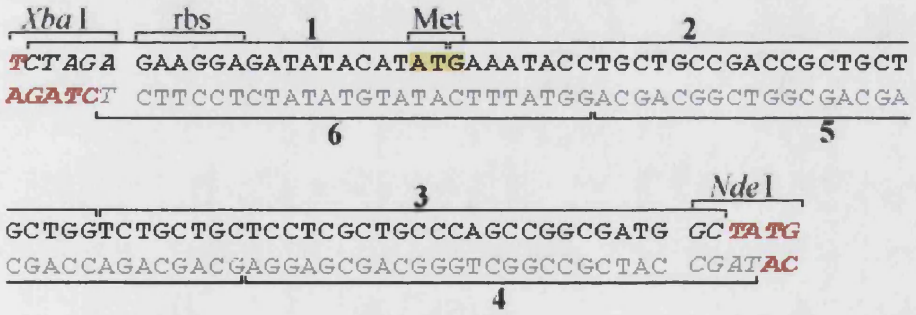



**Figure 5-2: Showing the design of the pelB insert**

Numbers show length & alignment of component oligonucleotides: 1 (22mer); 2 (30mer); 3 (33mer); 4 (26mer); 5 (30mer); 6 (26mer). Sequence shown in red was not part of construct. 'rbs' denotes the ribosome binding site, 'Met' shows the location of the methionine start codon for pelB transcription. *Xba* I = T'CTAG_A, *Nde* I = CA'TA_TG.

One of the considerations at this point was the possible advantage of engineering the removal of the pro-sequence from *r-trypsinogen*. Expression of mature trypsin would remove the need for the activation of the zymogen, thus simplifying the assay, and possibly allowing for some form of

nutritional selection when screening the future libraries. Overexpression of a highly active, unregulated and 'foreign' protease into the cytoplasm of an *E. coli* would doubtless have a severe autotoxic effect, however, the introduction of a signal sequence means that (in theory) the enzyme could be sequestered in the periplasm, allowing it to accumulate to assayable levels away from the bulk of cellular processes. A group attempting to probe serine protease specificity by means of rationally- designed mutants (Evnin and Craik, 1988) managed to express active rat trypsin into the periplasm up to an approximate concentration of 1mg/L without any visible autotoxicity effects on the cell. The cutting site for enterokinase is DDDDK↓ (Asp4-Lys) and removal of the pre-cursor would cut 8 amino-acids from the mature sequence. If the signal sequence insert were altered so as to fit immediately upstream of the first residue of the mature sequence (Ile), then it was possible that a similar level of expression could be achieved. No data appeared to exist on the level of autotoxicity that bovine trypsin caused in prokaryotic cells, but it seemed reasonable to assume that high-level expression would have a significant impact on the health of the culture. It was proposed that initial work with the system would rely on the 'leakiness' of the pHKY603 temperature sensitive promoter. It was not known to what degree the promoter leaked, i.e. what level of protein was expressed at a non-activation, i.e. repressor permissive, temperature (<36.5°C), but if an assayable level of protein was produced then it would eliminate the need for an activation step after purification. An expression level of 1mg/L may not be useful for industrial purposes (such a reduction in yield may not allow a profitable process, despite removing the need for enterokinase), but would be sufficient for the selection and screening tests proposed in this work.

Two constructs were thus proposed: pHKY (T) – which would code for the complete *r-trypsinogen* gene and have the leader sequence inserted between the *Xba* I & *Nde* I (Figure 5-3); and pHKY (M) – where the 8aa prosequence was removed and the leader sequence ligated directly on the N-terminus of the mature protein gene (Figure 5-5). The plasmid (pHKY603) would be cut with *Nde* I & *Kas* I, and the 'bridging' oligonucletides would replace the sequence between the *Kas* I site and the lysine of the activation site (Figure 5-4). A single nucleotide substitution was needed to convert this lysine into a methionine, and the signal sequence could then be inserted in the same manner as in pHKY (T).

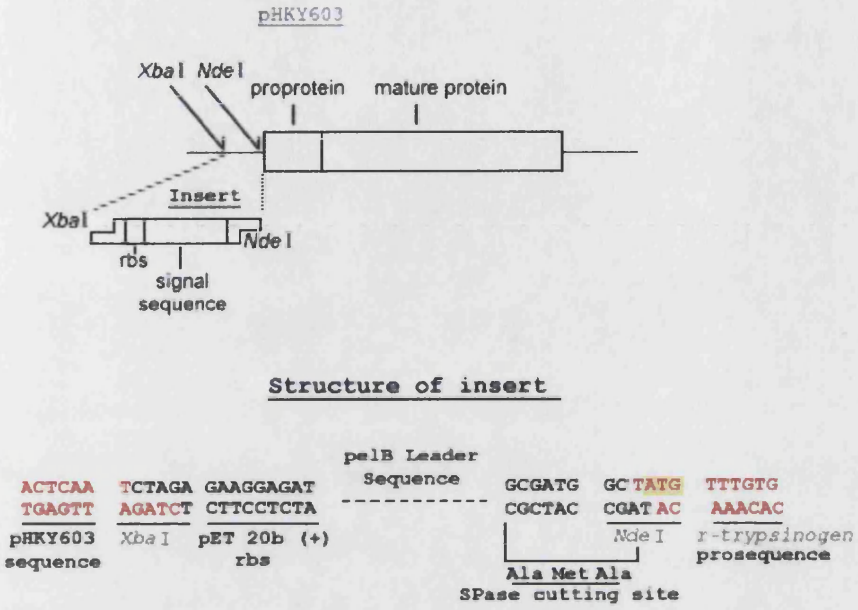



**Figure 5-3: Design of pHKY (T)**

Sequence shown in red was original plasmid pHKY603, sequence in blank comprised the leader sequence insert. *Xba* I = T'CTAG_A, *Nde* I = CA'TA_TG.

```
                                NdeI                              KasI
Oligo 1  5'-3' Sticky:    TAtGATCGTTGGCGGCTACACCTGTG
Oligo 2  3'-5' Sticky:      aCTAGCAACCGCCGATGTGGACACCGCG

 Properties:  1: 26nt, Tm = 74.6°C
              2: 28nt, Tm = 84.1°C
```

**Figure 5-4: Design of *Nde* I/*Kas* I 'bridging' sequence**

Nucleotide letters in small case indicate where a deliberate mutation was introduced – here an adenine (5'-3') was replaced by a thymine to convert a lysine residue into a methionine. Restriction sites marked show only the portion of the site contained within the construct: *Nde* I = CA'TA_TG, *Kas* I = G'GCGC_C.

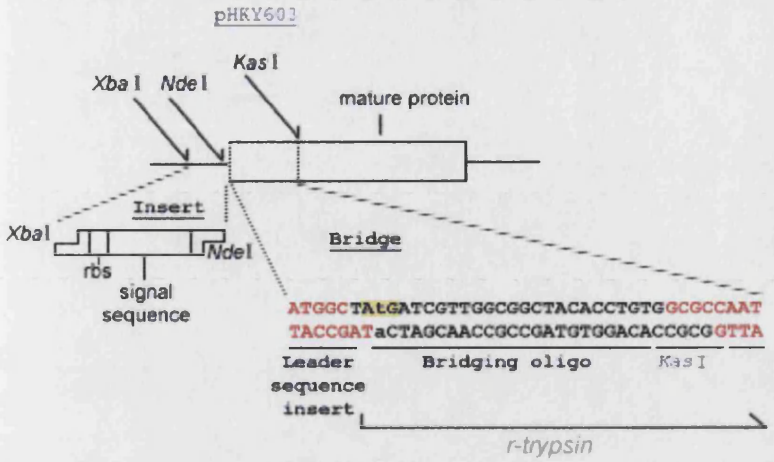



**Figure 5-5: Design of pHKY (M)**

Sequence shown in red was original plasmid pHKY603, sequence in black comprised the *Nde* I/*Kas* I bridging insert. *Nde* I = CA'TA_TG, *Kas* I = G'GCGC_C. Note that although the bridging insert was designed with an N-terminal overhang complementary to an *Nde* I cut site, this restriction site would be destroyed upon ligation with the leader sequence.

### 5.1.5 Construction of pET (T)

As the work with pHKY (T) & pHKY (M) was to prove inconclusive, and the general unsuitability of the ELTRP-1 strain for molecular work became evident, the decision was made to take the next logical step and transfer the target gene into a vector which would allow for flexible, temperature independent, control of expression rates. The pET series from Novagen are based upon a viral T7 promoter tied to the lac operon, in which the T7 polymerase is genome encoded within the lysogenic *E. coli* strain BL21(DE3) (Studier and Moffatt, 1986). This T7 polymerase can only be expressed when the constitutively expressed lac repressor is bound by IPTG (the inducing molecule – analogue of allolactose in original system). Unlike systems based on *E. coli* promoters (*e.g., lac, tac, pL*), the bacteriophage T7 promoter is used by the pET system to direct the expression of target genes. Since *E. coli* RNA polymerase does not recognize the T7 promoter, there is virtually no transcription of the target gene in the absence of a source of T7 RNA polymerase and the cloning step is thus effectively uncoupled from the expression step. The advantages of using such a strain over ELTRP-1 were thus: temperature independent control of expression rates; easy establishment of maintenance strains (transforming the modified plasmid

into a host other than BL21(DE3) will prohibit any expression; the host strain is protease deficient (both *lon* and *ompT*) and bred specifically for molecular work; tight control over expression. The relative 'leakiness' of ELTRP-1 was never established, but temperature based promoters are notoriously difficult to use reproducibly at lab scale (effective restrictive temperature cut-off point, poor heat transfer within incubators, etc). A pET vector should provide for tighter control of expression and, if an uninduced basal expression rate proves a problem, protocols exist for inexpensive enhancement of lac operon repression (e.g. enriching the medium with 1-2% glucose), use of pLysS/E 'double repressor systems (Novagen).

It was decided that the plasmid pET26b(+) would be used for this work (T7 promoter, kanamycin resistance, f1 origin, pBR322 origin, pelB coding sequence immediately upstream of the multiple cloning site, see Figure 5-6) and that only the section of *r-trypsinogen* coding for mature trypsin would be cloned, creating a construct called pET (T). There were two steps to this work, the first involved the introduction of an *Nco* I site at the activation site of *r-trypsinogen* by means of a mutagenic primer (Figure 5-7), and the second exploited this new restriction site, cutting out the trypsin gene by means of *Nco* I & *Bam* HI and ligating it into pET26b(+) directly behind the pelB sequence in such a way that the reading frame was conserved and no extra amino acids were added between the Spase site and the methionine of r-trypsin (Figure 5-8). For the correct formation of the *Nco* I site, a non-synonymous mutation had to be introduced and Ile2 was converted into Val2. This was considered acceptable, as both were non-polar aliphatic amino acids of similar molecular weight (113.16Da to 99.133Da respectively) and, while partially surface exposed, were sufficiently distant from the active site & binding pocket for any impact on activity to be unlikely.

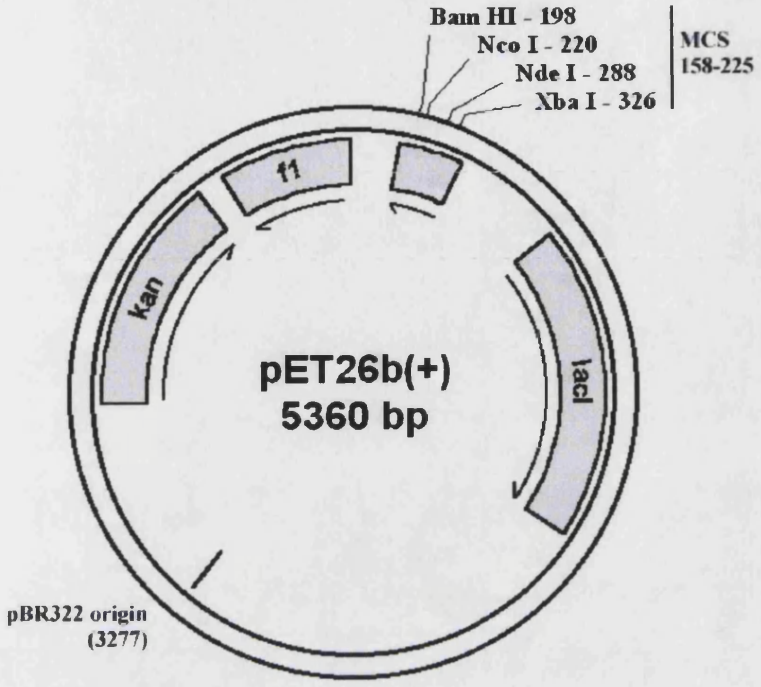



**Figure 5-6: Simplified structure of pET26b(+)**

Plasmid supplied by Novagen. Characteristics: T7*lac*, His•tag, f1 ori, pelB, kan$^r$

```
                                        Nco I
Forward primer 5'-3'      GATGATGATGccAtGgTCGTTGGCGGCTACACC
Properties: 33nt, Tm = 82.0°C                mature enzyme

                                        Bam I
Reverse primer 5'-3'      GCTCAGCCTCGCGGGATCCTAATTGGAAGC
Properties: 30nt, Tm = 78.5°C                stop
                                             codon
```

**Figure 5-7: *r-trypsinogen* mutagenic primers**

The mutagenic primers designed to prepare *r-trypsinogen* on pHKY603 for insertion into pET26b(+). Nucleotides in small case indicate deliberate mutation, in this case required to form the *Nco* I site. Methionine (ATG) indicated in yellow and stop codon (CTA) highlighted in pink. *Nco* I = C'CATG_G, *Bam* HI = G'GATC_C.

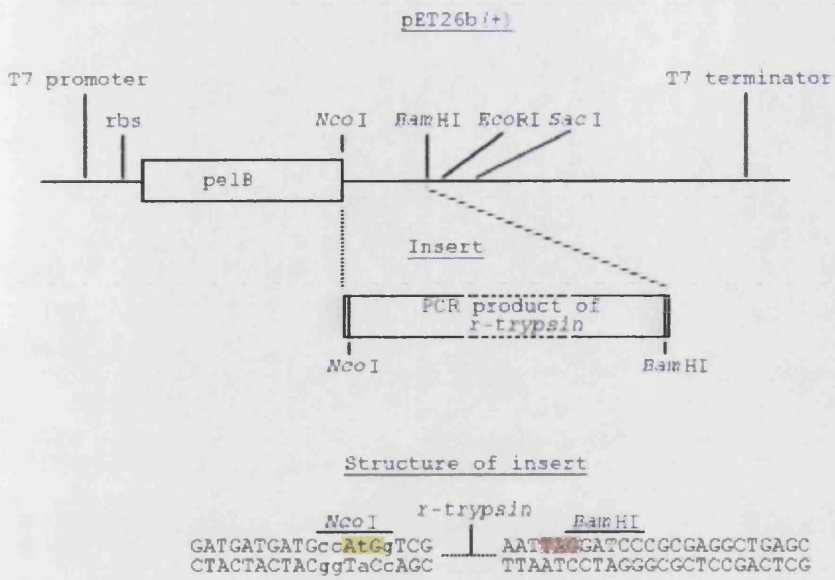



**Figure 5-8: Design of pET (T)**

Nucleotides in lower case denote nucleotide substitutions from original. Methionine (ATG) indicated in yellow and stop codon (CTA) highlighted in pink. *Nco* I = C'CATG_G, *Bam* HI = G'GATC_C.

## 5.2 Materials and methods

### 5.2.1 Strains & vectors used

RV308 (Eli Lilly): (ELTRP-1 = RV308 + pHKY603) Strain was a derivative of *E. coli* K12. Fully genotype not released by company. Resistance: $Sm^r$.

BL21(DE3) (Novagen): derivative of B strain, $F^-$, *ompT* $hsdS_B$ ($r_B^-$ $m_B^-$) *gal dcm* (DE3)

XL1-Blue (Stratagene): *recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F' *proAB* $lacI^qZ\Delta M15$ Tn*10* ($tet^r$)]

pHKY603 (Eli Lilly): a derivative of pBR322. Fully sequence not released by company. Resistance: $Tc^R$,

pET26b(+) (Novagen): T7*lac*, His•tag, f1 ori, pelB, $kan^r$

### 5.2.2 Enzymes and oligonucleotides

All digestive and modifying enzymes along with appropriate buffers were purchased from Promega. Oligonucleotides were synthesised to order by Amersham Pharmacia Biotech.

### 5.2.3 Construction of pelB insert & *Nde* I/*Kas* I 'bridging' sequence

Plasmid (pHKY603) prepared by miniprep (Qiagen) from a 5ml LB culture grown at 37°C, 200rpm for 16hrs. Purified plasmid eluted into 50μl of sterile MilliQ water and ~70ng was digested with *Nde* I and *Xba* I for both pHKY(T) & pHKY603(M), and *Nde* I and *Kas* I for pHKY(M), for 4hrs, according to manufacture's protocols. Cut plasmid would then be separated from the excised fragment by means of spin-column (Qiagen). The 6 oligonucleotides comprising the pelB insert (Amersham Pharmacia Biotech) were resuspended in 500μl MilliQ water to form stock solutions of 78-164μM. For the annealing reaction, all oligonucleotides were pooled into a 250μl solution so that each was present at 10μM. A water bath was heated to 95°C, then, with the vial containing the solutions suspended within it, was sealed shut and turned off. The water was allowed to return to room temperature over the course of 4-5hrs and the solution of annealed fragments was retrieved. The mixture of fragments was run on a 0.8% agarose gel to determine the degree of purity and a region corresponding to the approximate molecular weight of the insert was excised and purified by gel-extraction kit (Qiagen). The insert would then be ligated with the purified plasmid overnight at 4°C in the appropriate buffer.

### 5.2.4 Fermentation of 5ml or 100ml BL21(DE3):pET (T) cultures

A 5ml seed-culture of BL21(DE3):pET (T) was grown overnight in selective media (LB or M9SD, $Kan^+$ 30μg/ml unless stated otherwise) at the temperature required for the subsequent induction run (28°C, 32°C or 37°C). A 1/100 inoculation of this culture was made into either 5mls or 100mls of selective broth and the cells were grown to the induction point of $OD_{600nm}$. Induction of rTrypsin expression was achieved by adding IPTG to the culture to a final concentration of 1mM. The cells were then grown overnight at the appropriate temperature, 200rpm, with 1ml samples being taken throughout the course of the fermentation. Samples were either frozen immediately at –20°C or

fractionated and then frozen in order to prevent excess cross-contamination of the soluble fractions due to freeze-thaw lysis.

### 5.2.5 Fractionation of *E. coli* culture

In order to obtain a detailed picture of the levels of recombinant protein expressed, the solubility & point of accumulation of this protein and the subsequent percentage of that which was folded into an active state, it was necessary to fractionate the cultured cells and test each segment individually. As a coarse test, cell culture would be divided into three fractions, namely the supernatant, the soluble fraction (periplasm & cytoplasm) and the insoluble fraction (composed of inclusion body proteins, nucleic acids & lipids). More detailed work required the soluble fraction to be resolved into the individual periplasmic and cytoplasmic samples. If suitable controls were used to monitor the activity of commercial trypsin in the buffers at each stage, then the sum of the activity in each test fraction should equal the total trypsin activity of the cell. The protocol for this work was adapted from two sources, one in-house team working on a related system (Cooke *et al.*, 2001) and from a team who had developed and tested a strain for the expression of a soluble trypsin-streptavidin fusion protein (Clare *et al.*, 2001).

#### 5.2.5.1 Preparation of 'soluble' & 'insoluble fractions from BL21(DE3):pET (T) culture

Samples intended for fractionation were processed immediately in order to minimize soluble fraction contamination due to lysis of the cells. 1ml of culture was centrifugation for 10min at 13,000rpm, room temperature. The supernatant was drained and frozen at –20°C, labeled as 'Culture supernatant'. The pellet was resuspended in 200µl neat B-PER and allowed to stand at room temperature for 10min, with mild agitation to prevent settling. The sample was centrifuged again for 15min at 13,000rpm and the supernatant was removed, frozen, and labeled as 'Soluble fraction'. The pellet was resuspended again in a washing solution of 1:4 B-PER to MilliQ-water, and centrifuged for a further 15min. The supernatant was discarded and the pellet was resuspended in MilliQ-water, subsequently frozen and labeled as 'Insoluble fraction'.

#### 5.2.5.2 Preparation of 'periplasmic', 'cytosolic' & 'insoluble' fractions from BL21(DE3):pET (T) culture

The 'Supernatant' sample was collected as described in 5.2.5.1. The resultant pellet was resuspended in 200µl of the periplasmic extraction buffer, containing: 20% sucrose, 50mM Tris, 1mM EDTA, 50µg/ml lysozyme. Buffer was chilled on ice prior to addition and resuspended sample was allowed to stand for 15min with mild agitation to prevent settling.
The cells were then centrifuged for 10min at 13,000rpm, and the supernatant was carefully drained, frozen, and labeled as 'Periplasmic fraction'. The pellet was resuspended in 200µl neat B-PER and allowed to stand at room temperature for 10min, with mild agitation to prevent settling. The sample was centrifuged again for 15min at 13,000rpm and the supernatant was removed, frozen, and labeled as 'Cytosolic fraction'. The 'Insoluble fraction' was prepared in an identical manner to that in 5.2.5.1.

### 5.2.6 Transformation protocols

Competent *E. coli* cells were prepared for $CaCl_2$ transformation in the manner described by Sambrook (Sambrook *et al.*, 1989). Transformation of DNA (typically 50ng or less in 10µl) was added to 200µl of ice-chilled competent cells and allowed to stand for 45 seconds. The tubes were then heatshocked for 45 seconds and then incubated on ice for 5 minutes before 45 minute incubation in non-selective media at 37°C. For low stringency protocols, non-selective incubation time was increased to 90 minutes and the initial quantities of DNA were substancially increased (see individual Results sections for specific quantities).

## 5.3 Results

### 5.3.1 Construction of pHKY (T) & pHKY (M)

A pelB signal sequence, constructed from 6 complementary nucleotides and adapted, along with control elements form pET20b(+), was to be inserted into pHKY603, at the immediate N-terminus of *r-trypsinogen*. This plasmid would be termed pHKY (T). A second construct, pHKY603 (M)

(Mature protease) would have the prosequence of the zymogen edited out and the pelB sequence cloned in as before (see Figure 5-3 & Figure 5-5).

In order to establish the pHKY (M) construct, approximately 75ng of plasmid was then digested with *Nde* I & *Kas* I according to standard protocols. The product was then purified by spin column and analysed by electrophoresis. Direct sizing was impossible due to supercoiled structure of ladder, but the plasmid had clearly been linearised if compared to supercoiled pHKY603 and a faint double cut fragment was visible. The digested plasmid appeared, however, to be far smaller than it should be considering that only 40nt were excised. The digestion was redone using approximately 120ng of DNA and allowed to proceed for only 1 hour before heat inactivation & spin column purification (see Figure 5-9 below).

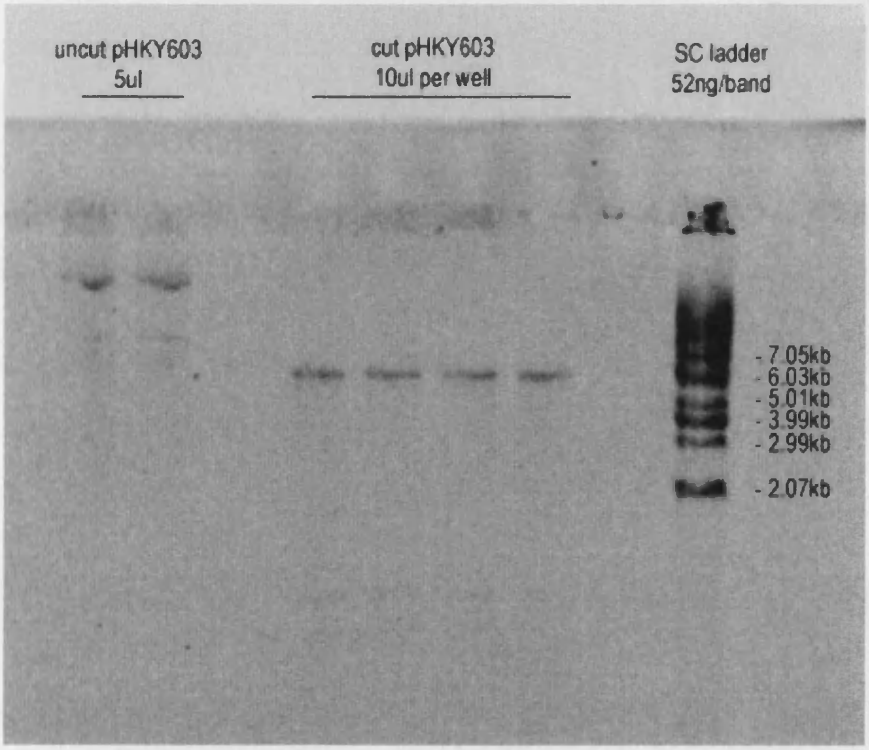



**Figure 5-9: Gel showing products from second *Nde* I/*Kas* I digestion**
0.8% agarose gel, run at 100V, 120mA for 1hr.

In Figure 5-9, the cut plasmid again appeared to be far smaller than expected when compared to the supercoiled pHKY603 (considering its linearised nature and the relatively small section of 40nt

excised), but the reaction was performed with all care from fresh stocks of both pHKY603 and enzyme and so it appeared to be a genuine result. Reexamination of the pHKY603 plasmid map provided by Eli Lilly revealed no design error on the experiment, in that the *Nde* I & *Kas* I sites were indeed unique.

The oligonucleotides forming the *Kas* I/*Nde* I bridge fragment (see Figure 5-4) were allowed to slowly hybridize by incubation at approximately 5°C below their $T_m$s (69°C for 2hrs, then 60°C, allowed to cool to room temperature – see Methods, Section 5.2.3), and then approximately 2µg was then ligated into 170ng of cut pHKY603. The product was analyzed by agarose gel electrophoresis.

The digestion reaction was run for a final time, using fresh plasmid, purified in the morning of work, and was allowed to proceed for only 30 minutes before heat inactivation and column purification. 200ng of this digestion product was then ligated with approximately 3µg of hybridized insert, with the reaction running overnight at 4°C. Digestion & ligation products shown in Figure 5-10 below).

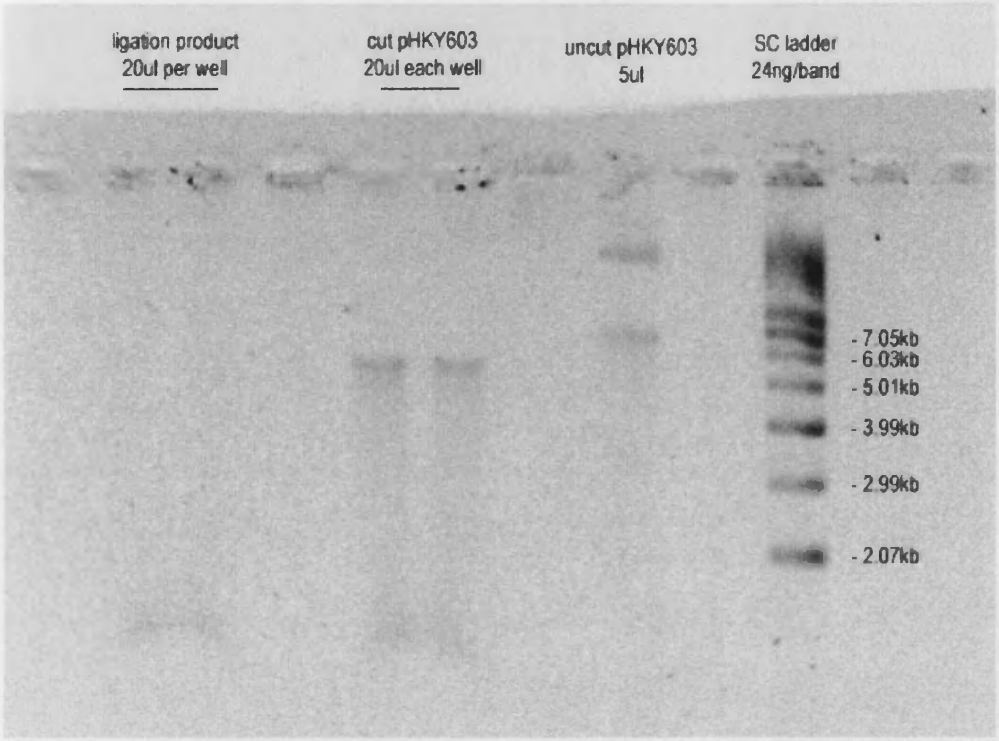



**Figure 5-10: pHKY603 *Nde* I/*Kas* I digestion & ligation products**

Gel showing a miniprep of pHKY603 and the second ligation attempt of the *Nde* I/*Kas* I pHKY603 digestion product & ligation with the hybridized bridge insert. 0.8% agarose gel run at 100V, 120mM, 1hr.

Figure 5-10 showed a similarly weighted digestion product as before and no trace of a ligated plasmid. The complete absence of even cut, linear plasmid in the ligation lanes (recircularization would be theoretically impossible due to the asymmetric digestion) could be explained.

The pHKY (T) construct was being developed in parallel with this work and the initial step required the digestion of pHKY603 with *Xba* I & *Nde* I. The digestion reaction was calculated to release a 52nt fragment. The reaction was run for an hour before heat inactivation, spin-column purification and concentration. Product shown in Figure 5-11.

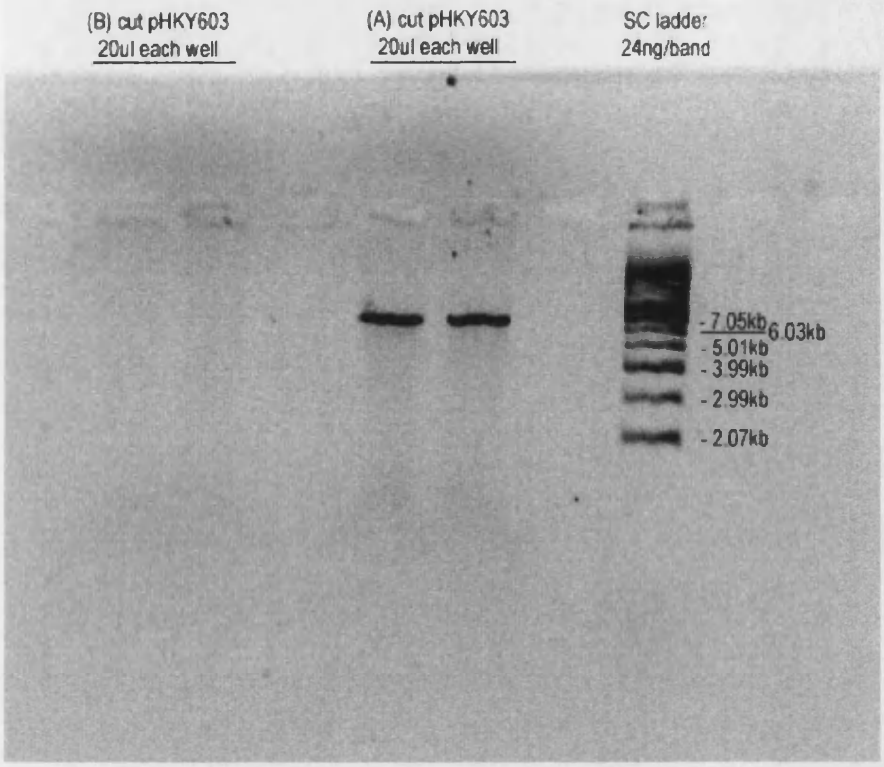



**Figure 5-11: Showing the pHKY603 *Xba* I/*Nde* I digestion product**
0.8% agarose gel run at 100V, 120mM, 1hr.

The *Xba* I/*Nde* I digestion fragment appeared to be approximately the expected size in that it ran slightly higher than the 6kb of the supercoiled plasmid.

The pelB insert was prepared from the 6 complementary oligonucleotides (see Figure 5-2) by means of the slow hybridization of a solution containing 10µM of each (see Methods, Section 5.2.3). The resultant mix of hybridized fragments was analysed by electrophoresis. The size of the correctly hybridized construct would be 85bp.

Approximately 170ng of the pHKY603 *Xba* I/*Nde* I digestion product (the basis for pHKY (T)) was ligated with 200ng of the hybridized pelB insert. A new pHKY603 *Nde* I/*Kas* I digestion was prepared from stocks, cleaned & concentrated and then immediately ligated with the *Nde* I/*Kas* I bridge fragment. Products from both reactions can be seen in Figure 5-12.

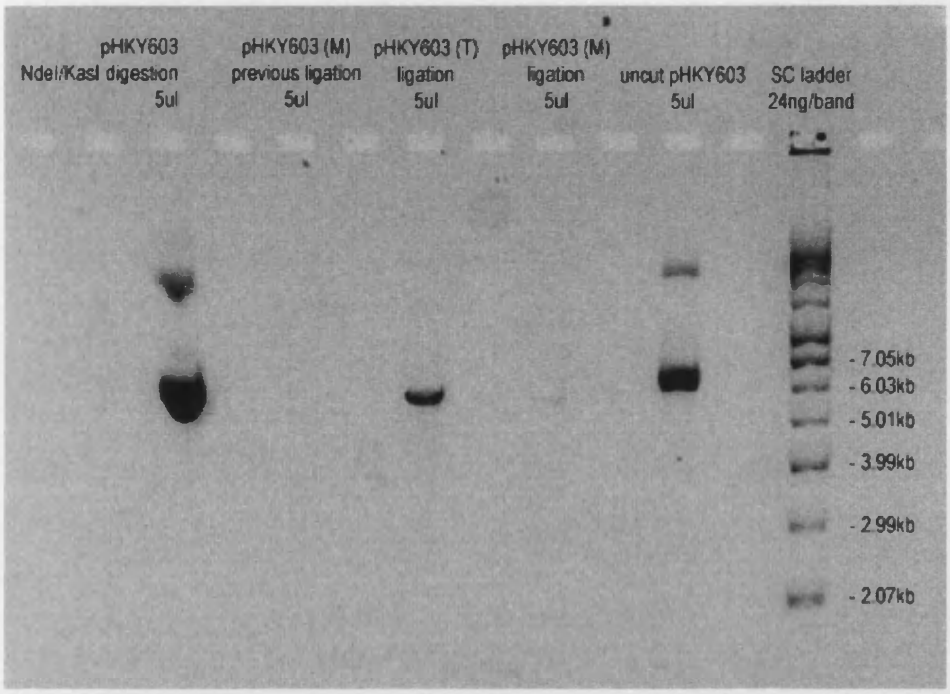



**Figure 5-12: pHKY(M) & pHKY(T) ligations**

Gel showing the new pHKY (M) & pHKY (T) ligations versus a previous pHKY (M) ligation, pHKY603 cut with *Nde* I & *Kas* I and a pHKY603 miniprep. 0.8% agarose run at 100V, 120mM for 1hr.

Figure 5-12 showed that the pHKY (T) ligation resulted in a clear product of approximately the correct size (although a portion of it could still remain as linearised plasmid). The pHKY (M) ligation (pHKY603 cut with *Nde* I & *Kas* I, ligated with the bridge insert) showed as a faint band of a similar size and no conclusions as to the success of the ligation could be made at this stage. This ligation mix was cleaned & concentrated and then cut with *Xba* I & *Nde* I and ligated with an approximately 3 fold molar excess of the pelB insert. The resultant product was then tested, along with the pHKY (T) ligation product in the transformation of competent RV308.

Stocks of RV308 (the ELTRP-1 *E. coli* strain without the pHKY603 plasmid) were prepared for $CaCl_2$ competency by means of the standard protocol (see Methods). The cells were then transformed with approximately 20ng of each cleaned ligation and subjected to heat shock @ 42°C for 90s. No commercial transformation control was used but, in its place, 10ng of minipreped supercoiled pHKY603 was used, and a further control involved a vial in which no DNA was added to competent cells although they were subjected to heat-shock. The importance of this was that no

transformants were visible after 24 hours growth on the selective plates. The 'blank' aliquot of cells also failed to grow on non-selective agar and thus it appeared that RV308 did not survive the transformation protocol. As a commercial production strain, designed for bulk expression, its unsuitability for molecular work and its sensitivity to certain strenuous protocols was to be expected. The transformation protocol was thus progressive altered to reduce its stringency and thus impact upon the strain. The temperature of heat-shock was seen as unalterable, but the period of heat-shock was reduced from 90s to 60s and then finally 30s. Similarly, the period of recovery in non-specific SOB media was increased from 30min to 1h. The concentration of tetracycline & streptomycin in the plates was reduced to stringent plasmid levels (10µg/ml each) and the transformants were gently concentrated prior to plating.

After three attempts at increasing lower stringency, and with the cells being prepared fresh from glycerol stocks each time, viable transformants were achieved. By this point XL1-Blue was also being used as an alternative host due to its documented higher competency. The projected aim of this was to use XL1-Blue as a transformation intermediate strain, receiving modified, linear plasmid and yielding supercoiled plasmid for transformation into the lower competency RV308. A problem existed in that XL1-Blue carried genome encoded tetracycline resistance, but it was hoped that by increasing the concentration of the antibiotic in the plates to 100µg/ml, there should be a selective pressure favoring those cells with two copies of the gene, thus enriching for the plasmid. Competent DH5α cells, prepared in house, were also used as a further test. Transformation efficiencies resulting from XL1-Blue & DH5α transformation with ligation products, and the subsequent transformation of RV308 with plasmid from XL1-Blue are shown in Table 5.1 below.

| Host Strain | Vector | Result | Approximate transformation efficiency (cfu/μg) |
|---|---|---|---|
| XL-Blue | pHKY (M) | Approx. 300 very small, white colonies. | $6 \times 10^4$/μg DNA |
| | pHKY(T) | Approx. 300 very small, white colonies. | $6 \times 10^4$/μg DNA |
| | Control (pHKY603) | Approx. 300 very small, white colonies. | $6 \times 10^4$/μg DNA |
| | Blank (no DNA added) | No growth | N/A |
| DH5α | pHKY (M) | Faint lawn. | N/A |
| | pHKY(T) | Faint lawn. | N/A |
| | Control (pHKY603) | Faint lawn. | N/A |
| | Blank (no DNA added) | No growth. | N/A |
| RV308 | pHKY (M) | Approx. 40 small, white colonies. | $8 \times 10^3$/μg DNA |
| | pHKY(T) | Approximately 210 small, white colonies. | $4.2 \times 10^4$/μg DNA |
| | Control (pHKY603) | Almost confluent growth. | N/A |
| | Blank (no DNA added) | No growth | N/A |

**Table 5.1: Transformants of pHKY (M) & pHKY (T)**

Concentration and type of antibiotics in media was strain dependant. RV308: plates contained tetracycline & streptomycin @ 10μg/ml; DH5α: plates contained tetracycline @ 10μg/ml; XL1-Blue: plates contained tetracycline @ 100μg/ml.

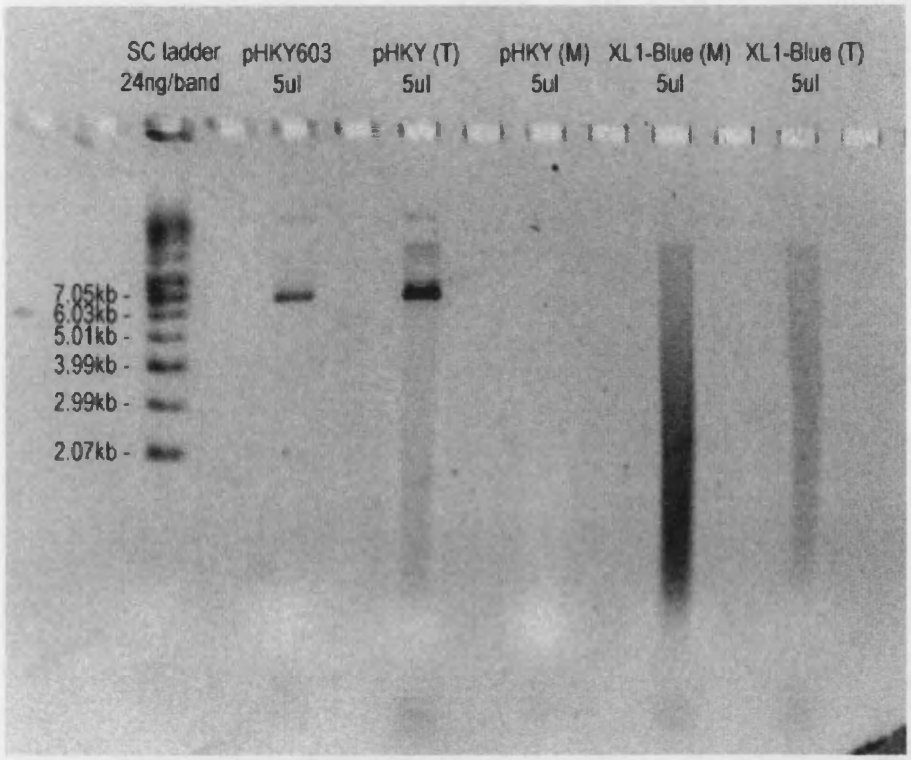



**Figure 5-13: Plasmid isolated from successful transformants**
Showing plasmid purified from single colonies picked from the transformants in Table 5.1. 0.8% agarose run at 100V, 120mM for 1hr.

Figure 5-13 shows the purified plasmid from isolates grown overnight in liquid media from the previous transformation. The DH5α colonies failed to grow in selective media and the RV308:pHKY (M) grew only very poorly. From the gel, it was evident that only the RV308:pHKY (T) clone had produced a plasmid of approximately the correct weight (6345bp as opposed to the 6260 of unmodified pHKY603). The broad molecular weight smears from the XL1-Blue transformants were indicative of Dnase activity, despite the genotype of the strain, and appeared to show that the attempt to select for a gene held on in both the vector and the host had been unsuccessful.

Confirmation of the identity of the constructs was sought by PCR using the standard trypsin amplification primers & primers that originated within the pelB sequence. Amplification was attempted at the two salt optima determined by work on unmodified pHKY603 (see Chapter 3, Section 3.3.5), results can be seen in Figure 5-14.

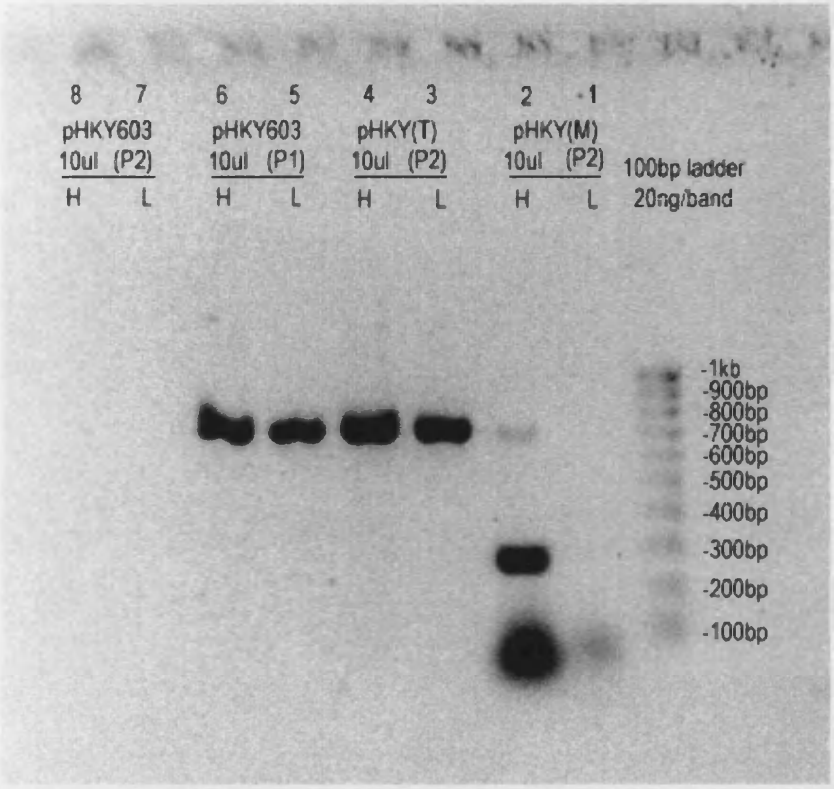



**Figure 5-14: PCR products from pHKY (M), pHKY (T) & pHKY603**

1.2% agarose gel run at 100V, 120mA for 50min. P1 indicates the use of the standard trypsin amplification primers (see Figure 3-2), while P2 indicates the use of the pelB-rtrypsinogen amplification set (see Figure 5-2). A newly digested & ligated pHKY (M) construct was used for this reaction. H & L refer to high and low [salt], 2.8mM & 0.8mM respectively. Lanes: 1 – pHKY (M) @ 0.8mM $MgCl_2$ using P2; 2 – pHKY (M) @ 2.8mM $MgCl_2$ using P2; 3 – pHKY (T) @ 0.8mM $MgCl_2$ using P2; 4 – pHKY (T) @ 2.8mM $MgCl_2$ using P2; 5 - pHKY603 @ 0.8mM $MgCl_2$ using P1; 6 – pHKY603 @ 2.8mM $MgCl_2$ using P1; 7 – pHKY603 @ 0.8mM $MgCl_2$ using P2; 8 – pHKY603 @ 2.8mM $MgCl_2$ using P2.

The plasmid purification & PCR reaction apparently proved that the pHKY (T) construct had been stably replicated and was a structure distinct from pHKY603 in that it was shown to contain the pelB coding sequence. The existence of a faint product of pHKY (M) at 2.8mM $MgCl_2$ was confusing as the ligated construct was particularly difficult to form and consistently failed to transform competent cells.

The presence of the gene thus being confirmed, the nature of the recombinant protein was then determined by means of a 5ml 32°C, 200rpm fermentation, followed by cell-fractionation, activity testing and SDS-PAGE of the fractions (see Methods, section 5.2.5). Induction was achieved by raising fermentation temperature to 37°C at $A_{600}$ = 0.61. Fermentation samples were collected and process 2 & 4 hours after induction, with the total soluble & insoluble fractions being shown in Figure 5-15, and the soluble fraction being separated in supernatant, periplasm & cytoplasm in Figure 5-16.

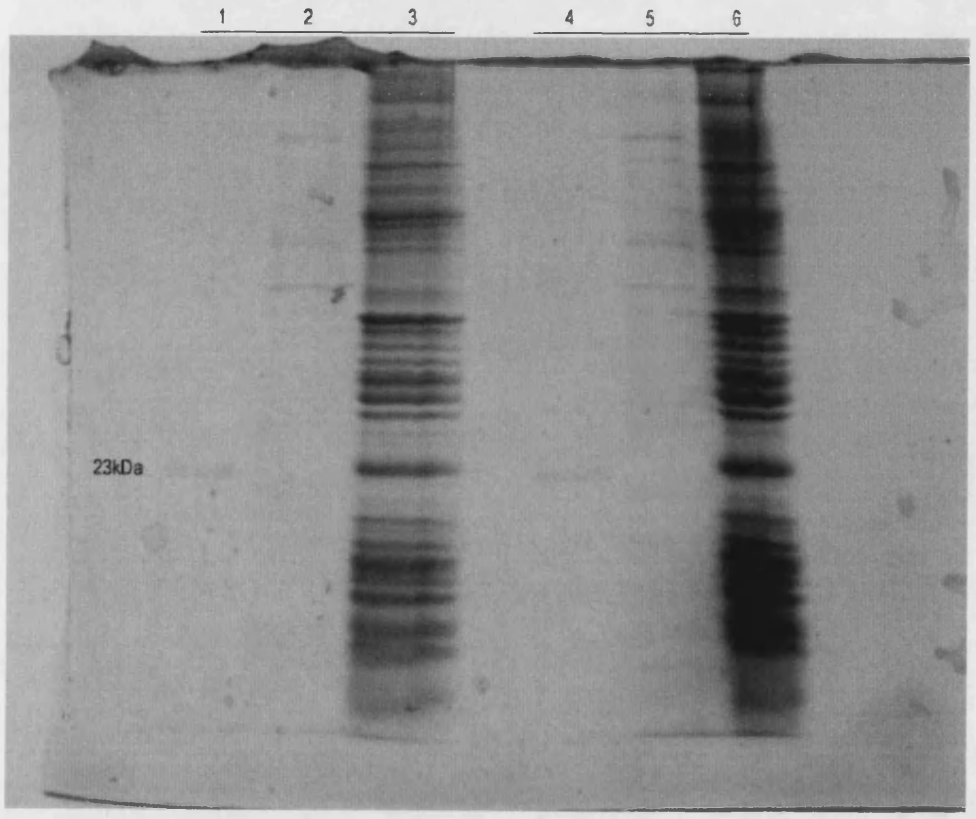



**Figure 5-15: Soluble & insoluble fractions from RV308:pHKY(T) 2 & 6 hours after induction**
SDS-PAGE run as 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5µg commercial trypsin; 2 – ($T_2$ – 2hrs after induction) soluble fraction; 3 – ($T_2$) insoluble fraction; 4 – 7.5µg commercial trypsin; 5 – ($T_6$ – 6hrs after induction) soluble fraction; 6 – ($T_6$) insoluble fraction).

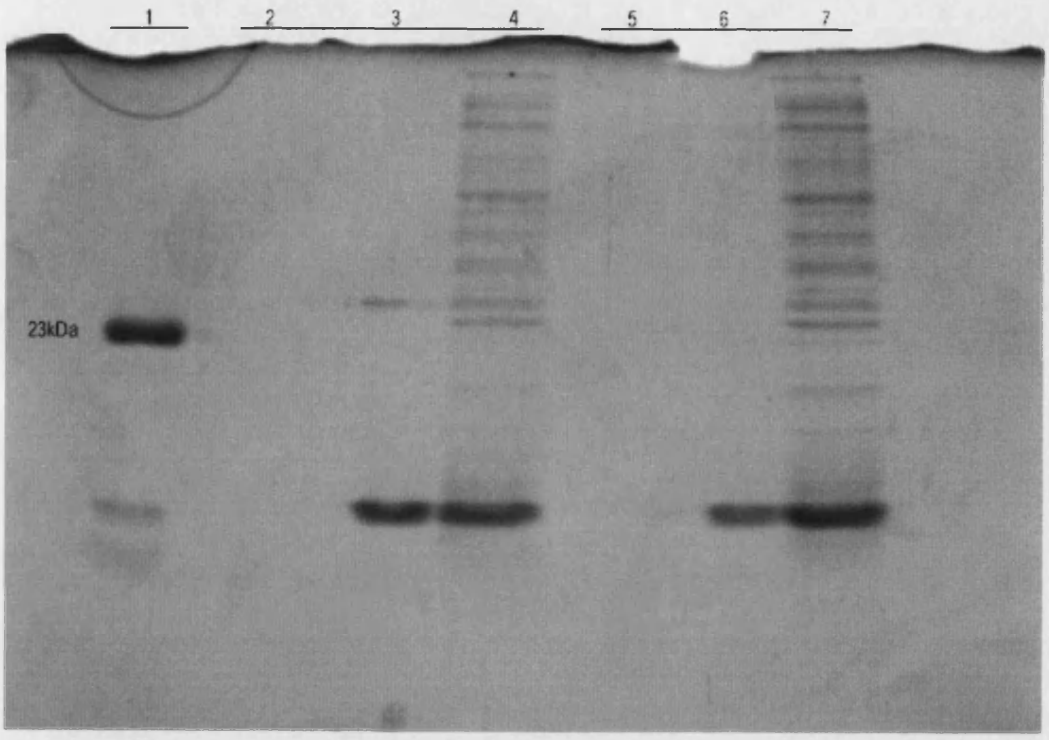



**Figure 5-16: Soluble fractions (supernatant, periplasm, cytoplasm) from RV308:pHKY(T) 2 & 6 hours after induction**

SDS-PAGE of the supernatant and periplasmic & cytosolic fractions from a RV308:pHKY (T) fermentation. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5µg commercial trypsin; 2 – ($T_2$ – 2hrs after induction) supernatant; 3 – ($T_2$) periplasm; 4 - ($T_2$) cytoplasm; 5 – ($T_6$ – 6hrs after induction) supernatant; 6 – ($T_6$) periplasm; 7 – ($T_6$) cytoplasm.

SDS-PAGE analysis of the fractions (see Figure 5-15, Figure 5-16) appeared to show moderate expression at the correct molecular weight in the insoluble fraction, but activity testing with L-BAPNA revealed no detectable signal in any of the fractions (for activity testing protocols, see Methods, Section 5.2.6. Activation of trypsinogen was by treatment with enterokinase). Although more rigorous testing of the pHKY (T) & pHKY (M) was possible at this stage (namely DNA sequencing and a better designed series of protein gels), the limitations placed upon the work by the heat-inducible promoter system and the apparent failure of the signal sequence to influence

protein accumulation, meant that it seemed prudent to move swiftly with a more flexible vector system.

The manipulation of *r-trypsin* required for insertion into the pET26b(+) vector is detailed in Figure 5-7 & Figure 5-8. Briefly, the gene segment corresponding to the mature protein only was to be amplified by PCR with mutagenic primers for the insertion of an *Nco* I site at the N-terminus (results shown in Figure 5-17), the ligation into pET26b(+) cut with *Nco* I & *Bam* HI (initial product shown in Figure 5-18) and the subsequent transformation of BL21(DE3).

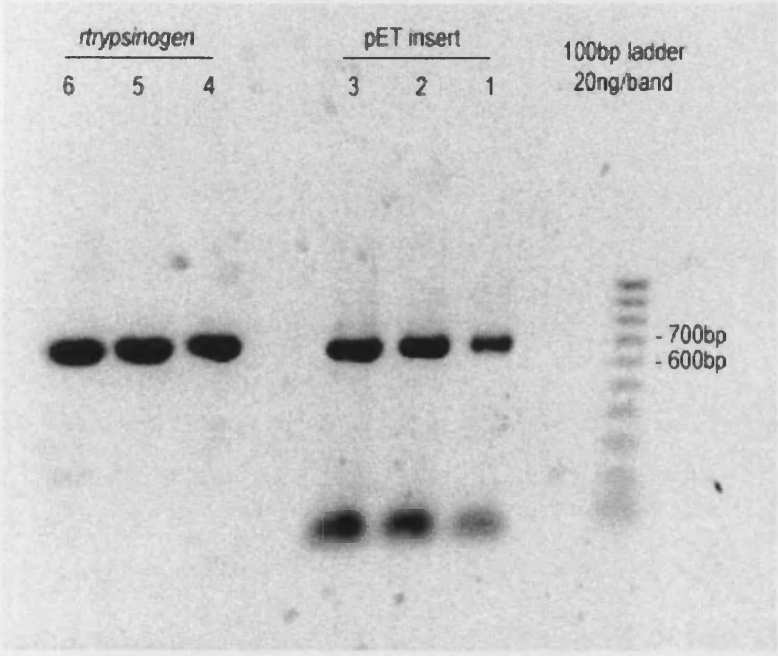



**Figure 5-17: Amplification of *r-trypsinogen* with standard & mutagenic primers**
1.2% agarose gel run at 100V, 120mM for 70min. Lanes 1-3 hold *r-trypsinogen* amplified 'pET26b(+) insert' mutagenic primers (see Figure 5-7), lanes 4-6 hold *r-trypsinogen* amplified with standard primers (see Figure 3.1). PCR products of 715bp (lanes 1-3) & 723bp (lanes 4-6) expected. Lanes: 1 – *r-trypsinogen* amplified @ 0.8mM $MgCl_2$; 2 – *r-trypsinogen* amplified with @ 1.5mM $MgCl_2$; 3 – *r-trypsinogen* amplified @ 2.8mM $MgCl_2$; 4 – *r-trypsinogen* amplified @ 0.8mM $MgCl_2$; 5 – *r-trypsinogen* amplified @ 1.5mM $MgCl_2$; 6 – *r-trypsinogen* amplified @ 2.8mM $MgCl_2$.

The PCR product was cleaned & concentrated via spin column and was then digested with *Nco* I & *Bam* HI. Commercial pET26b(+) was also double digested with the enzymes and, after spin-

column purification of the reactions, approximately 410ng of the gene insert was ligated with 1.04µg of vector (Product shown in Figure 5-18 below).

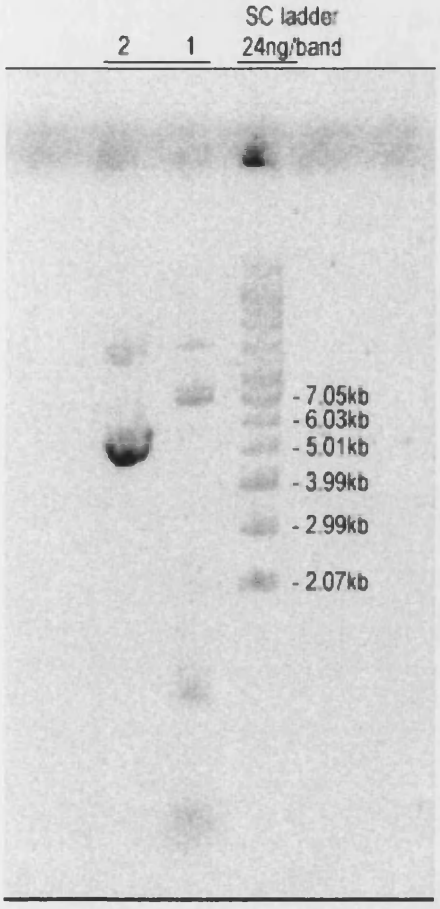



**Figure 5-18: pET (T) ligation product**

0.8% agarose gel run at 100V, 120mA for 1hr. Lanes: 1 – 5µl ligation product; 2 – supercoiled pET26b(+).

The ligation product visible in Figure 5-18 was of approximately the correct size (pET26b(+) + 1kb) in relation to the unmodified plasmid. Direct sizing was not possible due to the supercoiled nature of the ladder & plasmid. Following this work, the product was purified by spin-column and then transformed into commercially competent BL21(DE3) cells in parallel with the induction control/transformation control plasmid pET20b(+). Transformation was performed at medium stringency (see Methods, Section 5.2.6) and, while the induction control yielded strong growth on

the appropriate selective plates, BL21(DE3):pET (T) grew only as faint lawn of tiny (0.5-1mm diameter) white colonies. Single colonies were picked from each, grown overnight in LB and subsequently used as seed for a series of 100ml fermentations to assess the growth rate of uninduced cultures. Novagen literature recommended induction with IPTG to a final concentration of 0.4mM at $A_{600}$ 0.6-1 (blanked against uninfected media) and the initial induction work followed this guideline. Initial growth profiles shown in Figure 5-19.

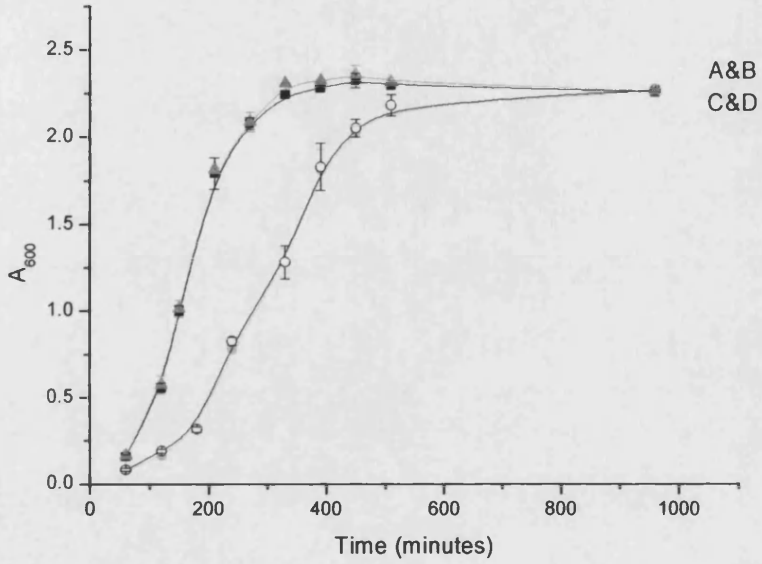



**Figure 5-19: Growth rates of induced & uninduced cultures of putative BL21(DE3):pET (T) strain in LB at 30°C & 37°C**

Key: A (–■–) = uninduced culture, grown at 37°C; B (–▲–) = induced culture grown at 37°C, 200rpm; C (–●–) = uninduced culture grown at 30°C; D (–○–) = induced culture grown at 30°C. Induction with 0.4mM IPTG at 150min (37°C) & 240min (30°C). Error bars indicate standard error over three individual fermentation runs for each condition.

No increased metabolic load upon induction was evident and standard activity assays using L-BAPNA revealed no enzyme activity in either the supernatant or soluble fractions of the induced cultures. The separated cellular fractions were then examined by SD-PAGE in order to attempt to establish whether r-trypsin was at least accumulating on the insoluble fraction of the cells or

whether the cloned vector had been flawed in a more fundamental manner. An induced culture of the induction control (BL21(DE3):pET20b(+)) was used as the blank for this work (instead of an uninduced experimental culture) so as to provide a clear contrast for the trypsin, avoiding any issues arising from the documented slight leakiness of the T7 promoter. The pelB + trypsin construct had a calculated weight of 25.5kDa. The supernatant, soluble & insoluble fractions of the cultures can be seen in Figure 5-20.

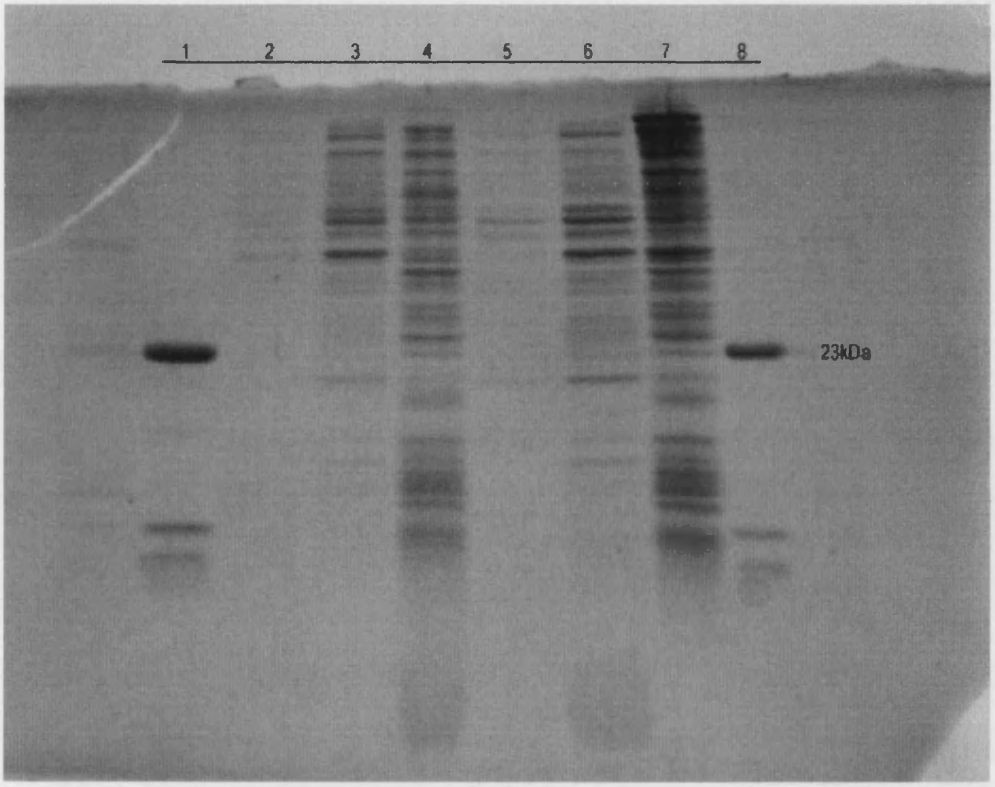



**Figure 5-20: Cellular fractions of the putative BL21(DE3):pET (T) culture**

SDS-PAGE run as 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes 2-4 hold fractions from induced BL21(DE3):pET (T) and lanes 5-7 hold fractions from BL21(DE3):pET20b(+). Both cultures were grown @ 37°C, 200rpm, and were induced with 0.4mM IPTG at approximately $A_{600}$ = 0.6 (blanked against uninfected medium), sampling & fractionation occurred 4hrs after induction. Lanes: 1 – 7.5µg commercial trypsin; 2 – supernatant; 3 – soluble fraction; 4 – insoluble fraction; 5 – supernatant; 6 – soluble fraction; 7 – insoluble fraction; 8 – 7.5µg commercial trypsin.

From the protein fraction gel in Figure 5-20, no overexpression of any 25.5kDa protein in lanes 2-4 was visible and, indeed, there appeared to be no differential in the banding patterns between

BL21(DE3):pET(T) and the induction control in lanes 5-7. The only conclusion possible was that the experimental strain was not expressing recombinant trypsin in any form. Plasmid was extracted from an overnight culture and linearised by digestion with *Nco* I in order to assess the weight in comparison with similarly linearised, but otherwise unmodified, pET26b(+) and thus determine whether *r-trypsin* was present. A double digestion with *Nco* I & *Bam* HI also performed in order to release the inserted gene, if present. All products were compared by agarose gel electrophoresis and it was evident that linearised pET (T) had the same molecular weight as linearised pET26b(+) and that digestion of pET (T) with *Nco* I & *Bam* HI seemed not to alter the weight of the plasmid and also did not release a second fragment (insert was 715bp). The initial ligation reaction was thus be seen as a failure and the transformation of BL21(DE3) must have occurred with recircularised plasmid despite the asymmetric cutting sites.

In order to dismiss the chance that the colony picked for the expression gels (Figure 5-20) was not representative of all the transformants, i.e. that a few authentic BL21(DE3):pET (T) clones had been successfully created but were effectively masked by the preponderance of those containing recircularised plasmid, 36 more colonies were picked and tested. The testing took the form of growing the colonies in overnight cultures, followed by plasmid extraction & size comparison with commercial pET26b(+).All supercoiled plasmids appeared to be of approximately the same molecular weight as the supercoiled wild-type (~5300bp) and thus it was concluded that the flaw lay in the ligation product rather than in under-sampling of the transformants.

The ligation reaction was twice repeated using fresh material grown from glycerol stocks and agarose gel extraction of the digested plasmids, but the ligation product always appeared similar to that in Figure 5-18, i.e. similar in size to linearised commercial pET26b. Transformations into BL21(DE3) were performed and in each case yielded poor growth as compared to the transformation control with only a faint lawn of tiny (0.5-1mm in diameter) white colonies grown after 20 hours. Although the colony physiology appeared abnormal and similar to that achieved in the previous transformation, approximately 30 colonies from each reaction being tested by plasmid size comparison, but, in each case, the extracted plasmid proved to be probable recircularised wild-type.

The repeated failure of the experiment appeared to indicate that the ligation step was being inhibited by one or more of the following factors: incorrect ratio of vector to insert (1:3 used previously); poor quality enzymes (no ligation control used to test ligase efficiency); conditions

appeared to favor plasmid recircularisation. The work was repeated using fresh material & enzymes, variations of the ligation reaction and dephosphorylation of the digested vector using calf intestinal alkaline phosphatase to prevent recircularisation.

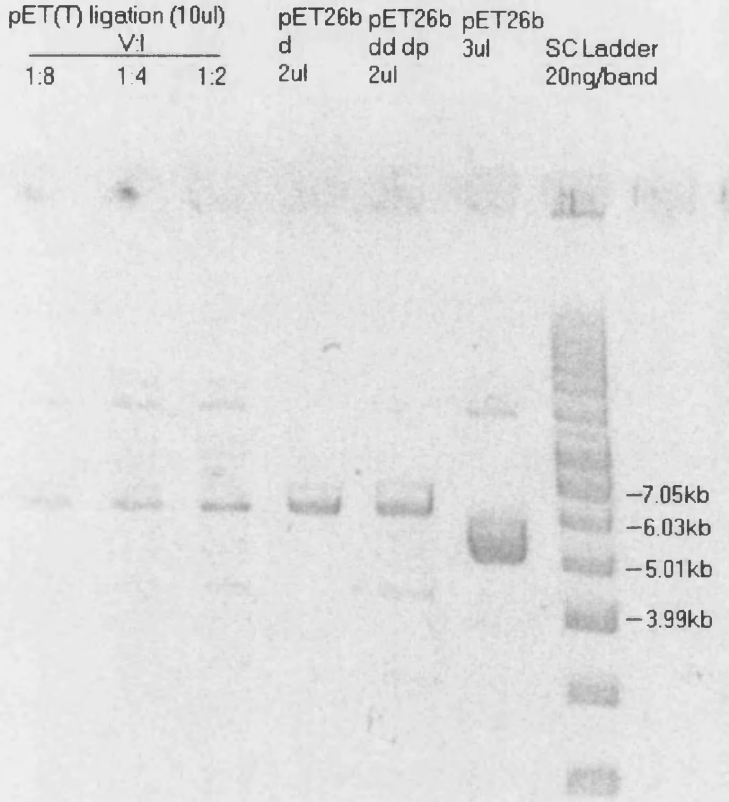



**Figure 5-21: Vector to insert molar ratio variation**

0.8% agarose gel run at 50V, 80mA for 40min. Lanes 4-6 contain pET (T) ligation product at different ratios of vector (V) to gene insert (I), lanes 2 & 3 contain digested pET (T) for weight comparison. Lanes: 1 - 2μl supercoiled pET26b(+); 2 - 5μl pET26b(+) digested with *Nco* I / *Kas* I, dephosphorylated; 3 – 5μl pET26b(+) linearised with *Bam* HI; 4 – 10μl pET (T) ligation product, V:I of 1:2; 5 - 10μl pET (T) ligation product, V:I of 1:4; 6 - 10μl pET (T) ligation product, V:I of 1:8.

From Figure 5-21, all ligation products appeared to be approximately the same weight as the linearised vector, despite the dephosphrylation of the 5' termini, which should prevent recircularisation. The presence of faint bands of a higher molecular weight suggested that larger

products may have present in small quantities and so a gel extraction was performed on the regions of the lanes from (as measured by supercoiled ladder) 7 – 11kb). 50ng of the purified plasmid was used to transform commercially competent BL21(DE3) cells. The result was poor growth (as compared to the transformation control) with a faint lawn of small (0.5-1mm) white colonies grown after 20hrs. 20 colonies were picked and grown overnight in LB @ 37°C, 200rpm. 9 isolates failed to grow on liquid culture entirely. Plasmid was extracted using the miniprep method and analysed by electrophoresis. Of the remaining 11, 10 showed no evidence of plasmid, indicating that these colonies were either contaminants or perhaps severely poisoned BL21(DE3) cells. The sigle plasmid band obtained was of approximately 5.3kb and was thus considered to be 'unmodified' or recircularised pET26b(+).

The persistent failure of the method for pET (T) construction appeared to indicate some terminal flaw in either concept or execution. While the successful creation of the ligation product was far from assured, an additional step that was proven to assist the uptake & replication of heavily modified or 'damaged' plasmids was that of a cloning intermediate strain. XL1-Blue was selected due to its easy availability, proven use in such methods & high transformation efficiency. The ligation reaction was repeated in the same form as before, involving the 1hr digestion & subsequent gel purification of vector & insert; the 30min dephosphorylation of the vector followed immediately by heat & EDTA inactivation; and a 16hr ligation reaction. The resultant ligation product can be seen in Figure 5-22.

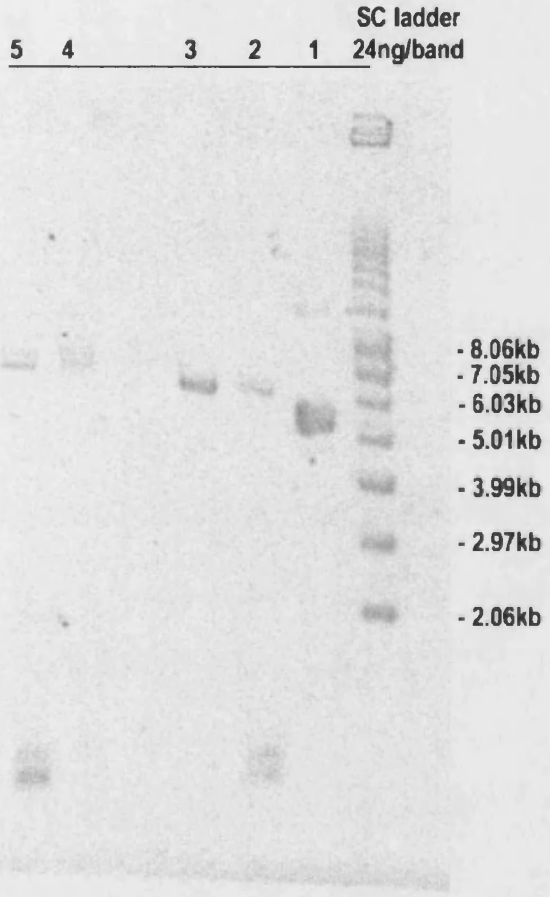



**Figure 5-22: New pET (T) ligation product vs linearised plasmid**

0.8 % agarose gel run at 50V, 80mA for 40min. Lanes: 1 – 2μl supercoiled pET26b(+); 2 - 5μl pET26b(+) digested with *Nco* I & *Bam* HI; 3 – 5μl pET (T) linearised by digestion with *Bam* HI; 4 & 5 - 10μl pET (T) ligation product.

In Figure 5-22, the ligation product did appear to be approximately 500+bp larger than the linearised unmodified pET26b(+) plasmid. The bands were purified by gel extraction & 50ng were transformed into commercially competent XL1-Blue cells. Growth was weak compared to the transformation control, with approximately 400cfu/ml. Colonies were distinct, white & 1-2mm in diameter. 20 discrete colonies were picked and grown overnight in 5ml LB @ 37°C, 200rpm. Plasmid was extracted & analysed by electrophoresis (results shown in Figure 5-23).

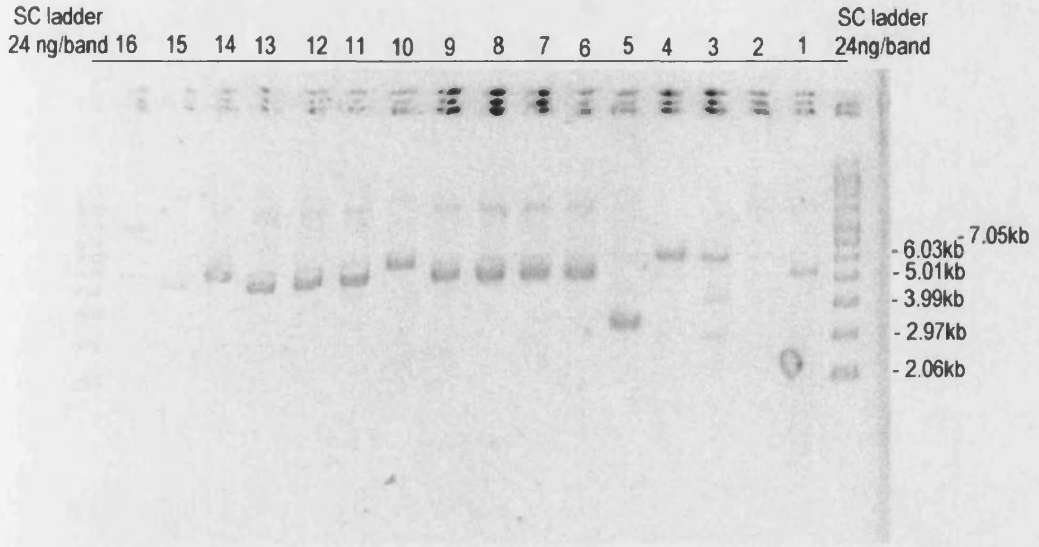



**Figure 5-23: XL1-Blue:pET(T) minipreps**

Multiple minipreps from XL1-Blue:pET (T) isolates vs pET26b(+) & pET (T) ligation product. 0.8% agarose gel run @ 50V, 80mA for 40min. Lanes: 1 – 2μl supercoiled pET26b(+); 2 - 5μl pET (T) ligation product; 3 - 5μl pET26b(+) digested with *Nco* I & *Bam* HI; 4 – 5μl pET (T) linearised by digestion with *Bam* HI; 5-16 - 5μl pET (T) transformant minipreps.

Although the majority of plasmids appeared to be the same weight as unmodified pET26b(+), 2 isolates (lanes 10 & 14) were sufficiently closer to the target size of 6.1kb to be worthy of further study. Both isolates were replated and 4 colonies from each isolate were cultured separately. Extracted plasmid was analysed by gel electrophoresis and direct sizing of the extracted plasmids revealed that they were approximately 6kb in weight, which corresponded well with the 6.1kb calculated size. As a conclusive test for the presence of the *r-trypsin* gene, a PCR designed to amplify only the target gene was used on the 2 prospective successful isolates, 2 isolates determined by sizing to have failed and the original pHKY603 plasmid. As the primers (see Figure 5-7) overlapped the gene at both termini and thus contained elements of sequence specific to pHKY603, the reaction was performed at low stringency to allow annealing with the putative pET (T) plasmid. Isolates '4' & '13' from the previous transformation (see Figure 5-23) were used as negative controls for this reaction. A product of 693bp was expected. Results shown below in Figure 5-24.

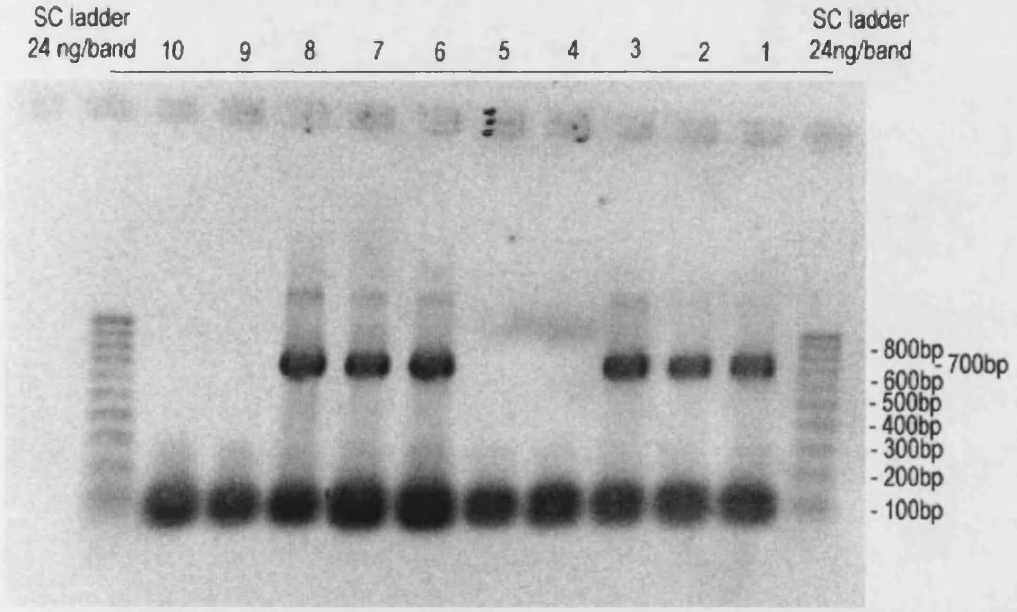



**Figure 5-24: PCR products from pET (T), pET26b(+) & pHKY603**

1.2% agarose gel run at 50V, 80mA for 40min. Lanes 1-5 were amplified using 1.5mM $MgCl_2$. Lanes 1 - pHKY603 PCR product; 2 – isolate '10' PCR product; 3 – isolate '14' PCR product; 4 - isolate '4' PCR product; 5 – isolate '13' PCR product; 6 - pHKY603 PCR product; 7 – isolate '10' PCR product; 8 – isolate '14' PCR product; 9 - isolate '4' PCR product; 10 – isolate '13' PCR product;

The successful amplification of the gene confirmed the identity of isolates '10' & '14' as XL1-Blue:pET (T). Both isolates were propagated as maintenance strains and 50ng of purified plasmid from both was used to transform commercially competent BL21(DE3) cells. Both transformations resulted in good growth in relation to the transformation control, with transformation efficiencies of approximately $2.4 \times 10^4$cfu/μg. Colonies were white & 1-3mm in diameter. 6 discrete colonies from each isolate plate ('10' & '14') were picked and grown overnight in LB media @ 37°C, 200rpm. Plasmid from each was extracted and purified, then analysed by electrophoresis with all supercoiled extracts proving to be of the correct size (6.1kb).

Analysis of the now confirmed BL21(DE3):pET (T) strain was continued with growth & expression profiling and fraction activity testing. Colonies of isolate '10' were picked, grown overnight in 5mls LB @ 37°C, 200rpm, then 1ml of this seed culture was used to inoculate 100mls

of LB in baffled 500ml shake flasks. Experimental fermentations were divided into those grown at 37°C, 200rpm and those grown at 30°C, 200rpm, and subdivided into those induced early ($A_{600}$ of 0.6, blanked against uninfected media) and those induced late ($A_{600}$ of 1.0). Growth profiles shown in Figure 5-25.

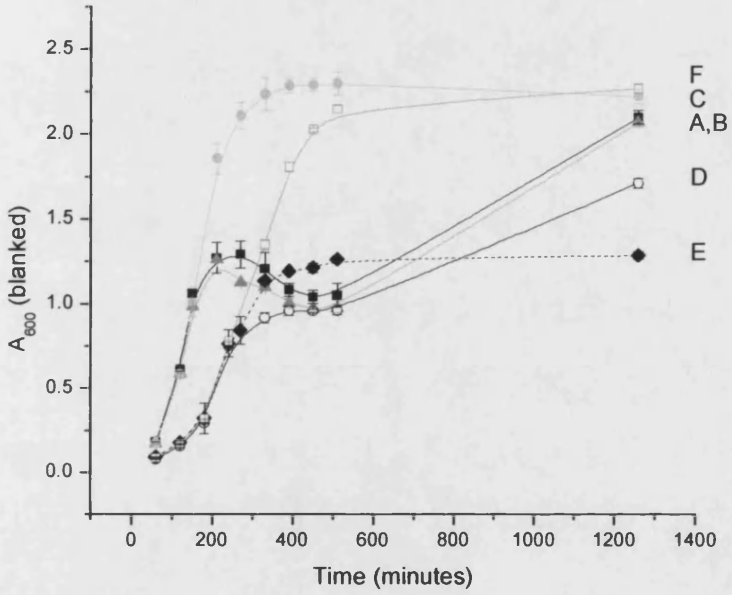



**Figure 5-25: Growth of putative BL21(DE3):pET(T) cultures**

Growth of induced & non-induced BL21(DE3):pET (T) cultures in 100ml LB @ 30°C & 37°C, early & late induction, 200rpm. Induction was with 0.4mM IPTG. Key: A (–■–) = early induction (120min, $A_{600}$ = 0.61), grown at 37°C; B (–▲–) = late induction (150min, $A_{600}$ = 0.98), grown at 37°C; C (–●–) = non-induced culture, grown at 37°C; D (–○–) = early induction (240min, $A_{600}$ = 0.71), grown at 30°C; E (··◆··) = late induction (270min, $A_{600}$ = 0.85), grown at 30°C; F (··□··) = non-induced culture, grown at 30°C. All $A_{600}$ values blanked against uninfected media. Error bars represent standard error from three separate fermentations, run concurrently.

The cultures in Figure 5-25 incurred an obvious metabolic load upon induction, but overnight incubation showed comparable cell densities, which argued against the expression of an autotoxic product. In order to visually test for the accumulation of an enzyme of the correct molecular weight (23Kda), a series of SDS-PAGE gels were performed, which ran the total cellular protein of induced and uninduced cultures against a commercial trypsin marker. Cultures were grown at 30°C

and 37°C to investigate the influence of fermentation temperature on enzyme localization, with the effects of early & late induction ($A_{600}$ = 0.6, beginning of exponential growth, $A_{600}$ = 1.0, mid to end of exponential growth) also being factored. The annotated test sample gel for growth at 30°C & 37°C with early in induction is shown in Figure 5-26, growth at both temperatures with late induction being shown in Figure 5-27. The gel of uninduced cultures at both temperatures is shown in Figure 5-28.

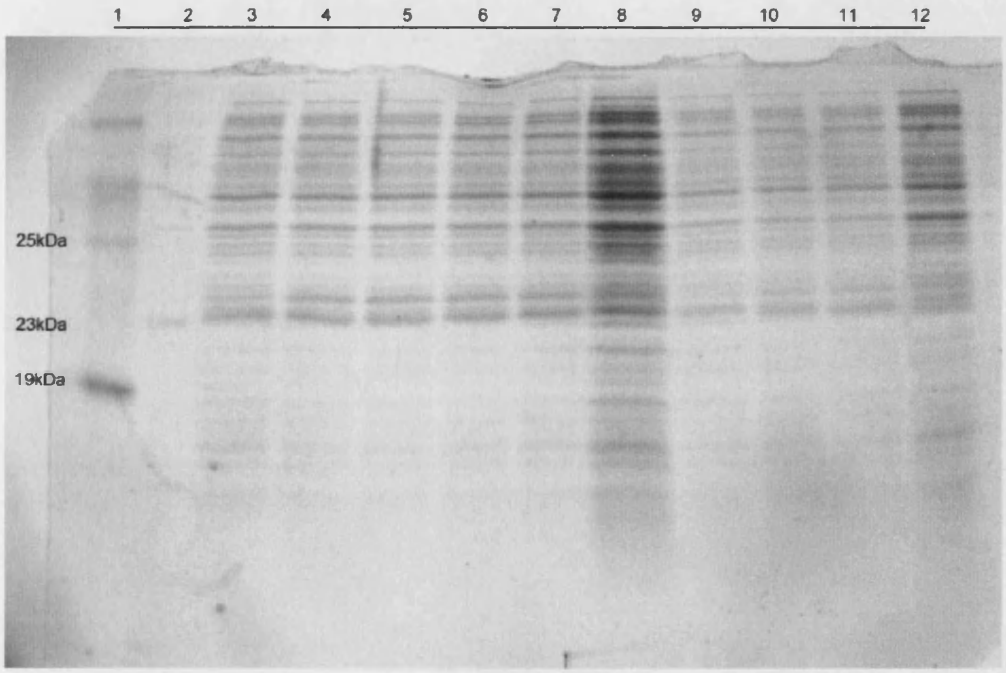



**Figure 5-26: Total protein from BL21(DE3):pET(T) grown at 30°C & 37°C, early induction**
Total protein SDS-PAGE of BL21(DE3):pET (T) cultures induced early with 0.4mM IPTG at approximately $A_{600}$ = 0.6, grown at 37°C & 30°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes 3-8 contain the total protein of culture grown at 37°C, lanes 9-11 contain the total protein of culture grown at 30°C. Lanes: 1 – commercial protein mid-weight size ladder; 2 – 7.5μg commercial trypsin; 3 -.$T_1$ (1hr after induction), 4 – $T_2$; 5 – $T_3$; 6 – $T_4$; 7 – $T_5$; 8 – $T_6$; 9 – $T_1$; 10 – $T_2$; 11 – $T_3$; 12 – $T_4$.

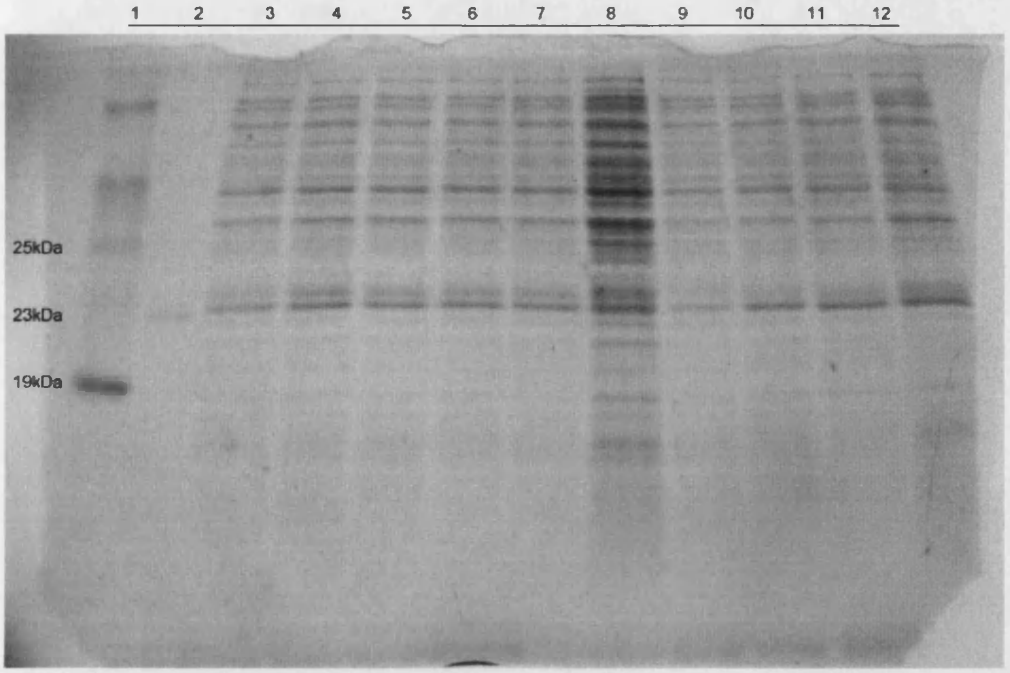



**Figure 5-27: Total protein from BL21(DE3):pET(T) grown at 30°C & 37°C, late induction**
Total protein SDS-PAGE of BL21(DE3):pET (T) cultures induced late with 0.4mM IPTG at approximately $A_{600}$ = 1.0, grown at 37°C & 30°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes 3-8 contain the total protein of culture grown at 37°C, lanes 9-11 contain the total protein of culture grown at 30°C. Lanes: 1 – commercial protein mid-weight size ladder; 2 – 7.5μg commercial trypsin; 3 -.$T_1$ (1hr after induction), 4 – $T_2$; 5 – $T_3$; 6 – $T_4$; 7 – $T_5$; 8 – $T_6$; 9 – $T_1$; 10 – $T_2$; 11 – $T_3$; 12 – $T_4$.

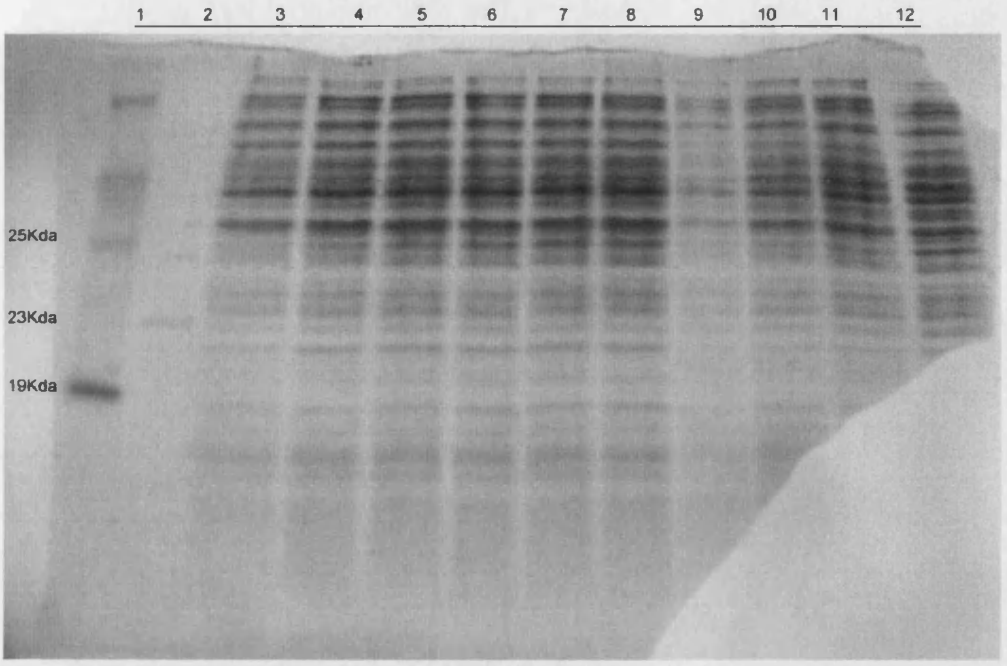



**Figure 5-28: Total protein from BL21(DE3):pET(T) grown at 30°C & 37°C, no induction**
Total protein SDS-PAGE of BL21(DE3):pET (T) non-induced cultures, grown at 37°C & 30°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes 3-8 contain the total protein of culture grown at 37°C, lanes 9-11 contain the total protein of culture grown at 30°C. Lanes: 1 – commercial protein mid-weight size ladder; 2 – 7.5μg commercial trypsin; 3 -.$T_1$ (1hr after induction), 4 – $T_2$; 5 – $T_3$; 6 – $T_4$; 7 – $T_5$; 8 – $T_6$; 9 – $T_1$; 10 – $T_2$; 11 – $T_3$; 12 – $T_4$.

Although, in the absence of a specific test for r-trypsin concentration (e.g. a suitable ELISA), the 23.5KDa band visible in the induced cultures could not be positively identified as r-trypsin, its density, as compared with the faint band visible in the uninduced cultures, was interpreted as a positive sign. The logical next step involved exploring the location of the overexpressed protein within the cell by means of fractionation and subsequent SDS-PAGE. The insoluble and soluble fractions of the culture were separated by the methodology described in Methods, section 5.2.5. As with the previous gels, cultures were grown at 30°C and 37°C and with early and late induction. Figure 5-29, Figure 5-30 & Figure 5-31 show samples run from cultures grown at 37°C and induced early, induced late and non-induced respectively. Figure 5-32, Figure 5-33 & Figure 5-34 show samples grown at 30°C and induced early, induced late and non-induced respectively.

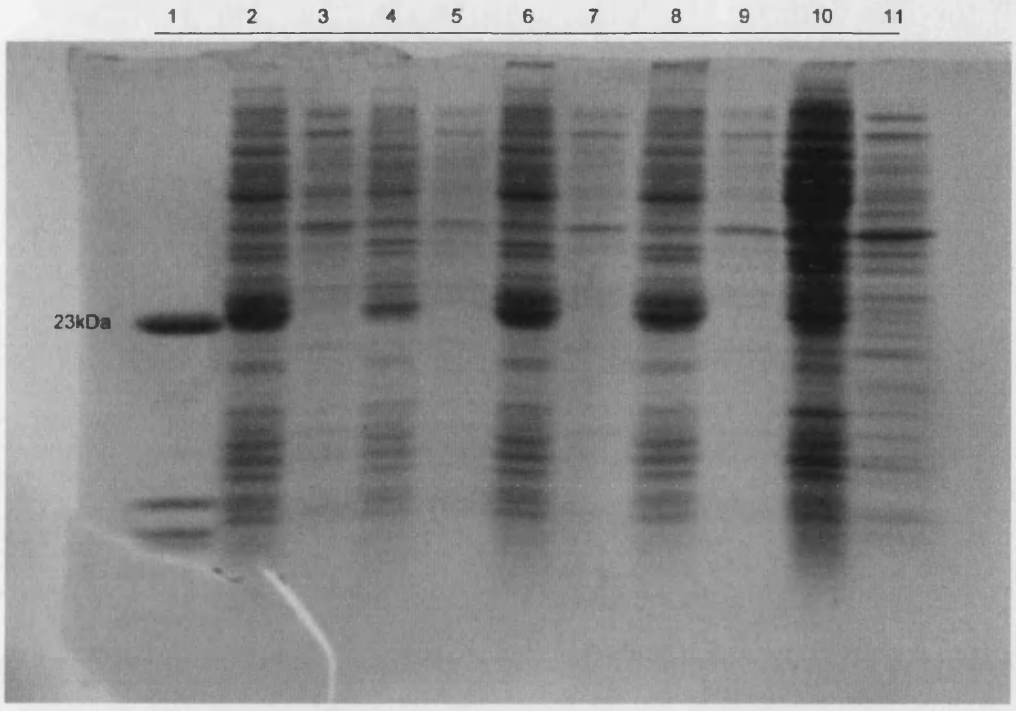



**Figure 5-29: Fractionated BL21(DE3):pET(T), grown at 37°C with early induction**

SDS-PAGE of fractionated induction run of BL21(DE3):pET (T) induced early with 0.4mM IPTG at approximately $A_{600}$ = 0.6, grown at 37°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5μg commercial trypsin; 2 – ($T_2$ – 2hrs after induction) insoluble fraction; 3 – ($T_2$) soluble fraction; 4 – ($T_3$) insoluble fraction; 5 – ($T_3$) soluble fraction; 6 – ($T_4$) insoluble fraction; 7 – ($T_4$) soluble fraction; 8 – ($T_5$) insoluble fraction; 9 – ($T_5$) soluble fraction; 10 – ($T_{16}$) insoluble fraction; 11 – ($T_{16}$) soluble fraction.

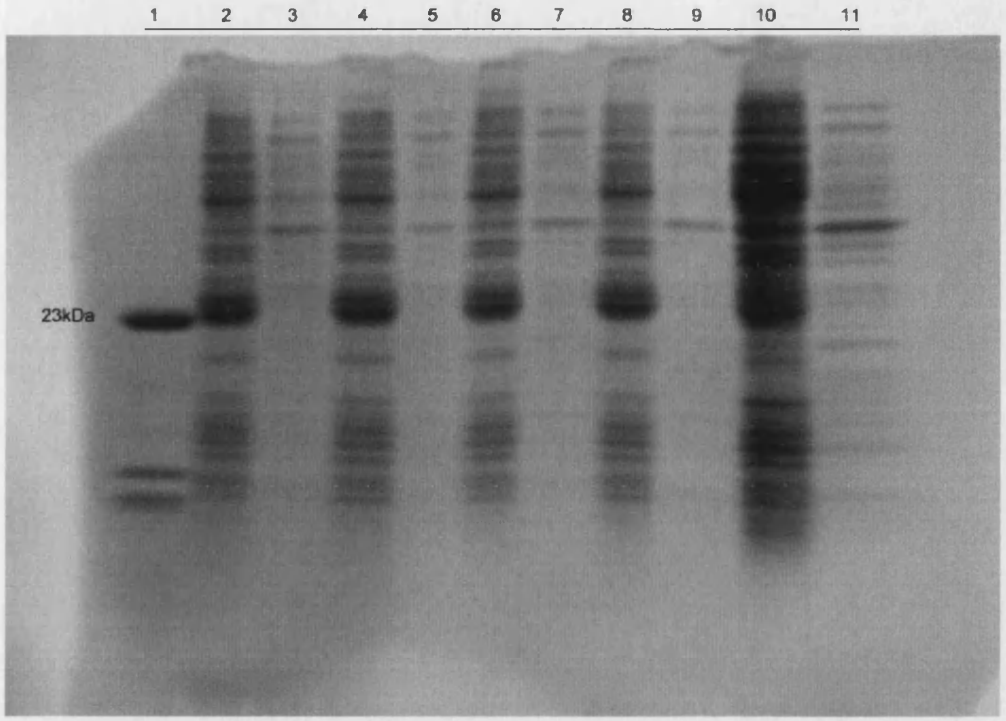



**Figure 5-30: Fractionated BL21(DE3):pET(T), grown at 37°C with late induction**

SDS-PAGE of fractionated induction run of BL21(DE3):pET (T) induced late with 0.4mM IPTG at approximately $A_{600}$ = 0.1, grown at 37°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5μg commercial trypsin; 2 – ($T_2$ – 2hrs after induction) insoluble fraction; 3 – ($T_2$) soluble fraction; 4 – ($T_3$) insoluble fraction; 5 – ($T_3$) soluble fraction; 6 – ($T_4$) insoluble fraction; 7 – ($T_4$) soluble fraction; 8 – ($T_5$) insoluble fraction; 9 – ($T_5$) soluble fraction; 10 – ($T_{16}$) insoluble fraction; 11 – ($T_{16}$) soluble fraction.

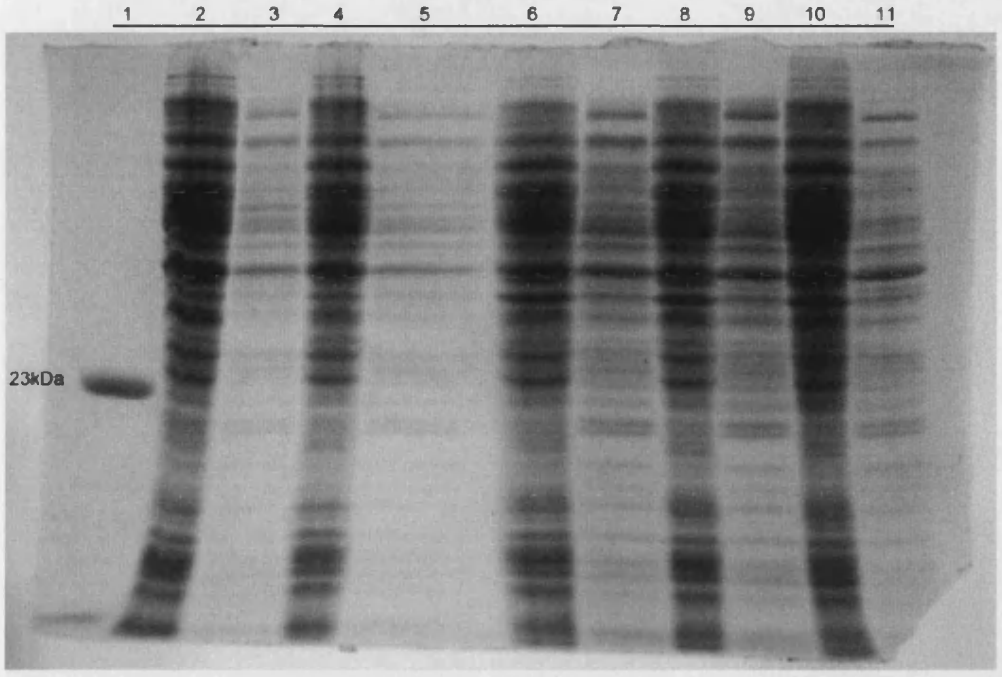



**Figure 5-31: Fractionated BL21(DE3):pET(T), grown at 37°C without induction**

SDS-PAGE of fractionated BL21(DE3):pET (T), non-induced, grown at 37°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5μg commercial trypsin; 2 – ($T_2$ – 2hrs after induction of cultures in Figure 5-25) insoluble fraction; 3 – ($T_2$) soluble fraction; 4 – ($T_3$) insoluble fraction; 5 – ($T_3$) soluble fraction; 6 – ($T_4$) insoluble fraction; 7 – ($T_4$) soluble fraction; 8 – ($T_5$) insoluble fraction; 9 – ($T_5$) soluble fraction; 10 – ($T_{16}$) insoluble fraction; 11 – ($T_{16}$) soluble fraction.

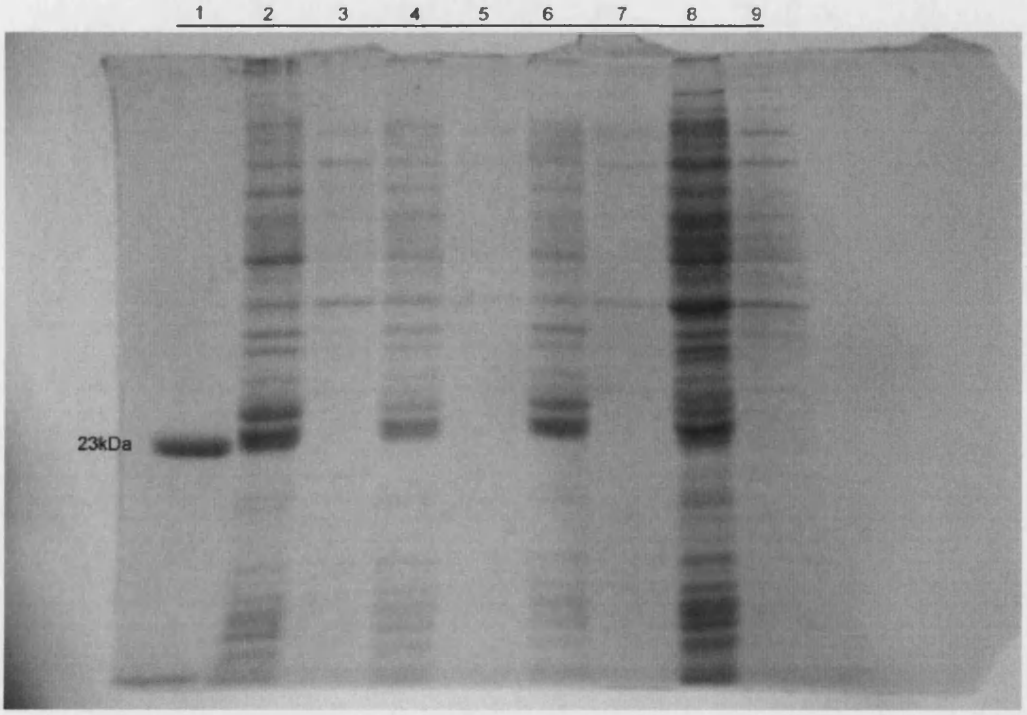



**Figure 5-32: Fractionated BL21(DE3):pET(T), grown at 30°C with early induction**

SDS-PAGE of fractionated induction run of BL21(DE3):pET (T) induced early with 0.4mM IPTG at approximately $A_{600}$ = 0.6, grown at 30°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5μg commercial trypsin; 2 – ($T_1$ – 1hr after induction) insoluble fraction; 3 – ($T_1$) soluble fraction; 4 – ($T_2$) insoluble fraction; 5 – ($T_2$) soluble fraction; 6 – ($T_3$) insoluble fraction; 7 – ($T_3$) soluble fraction; 8 – ($T_{16}$) insoluble fraction; 9 – ($T_{16}$) soluble fraction.

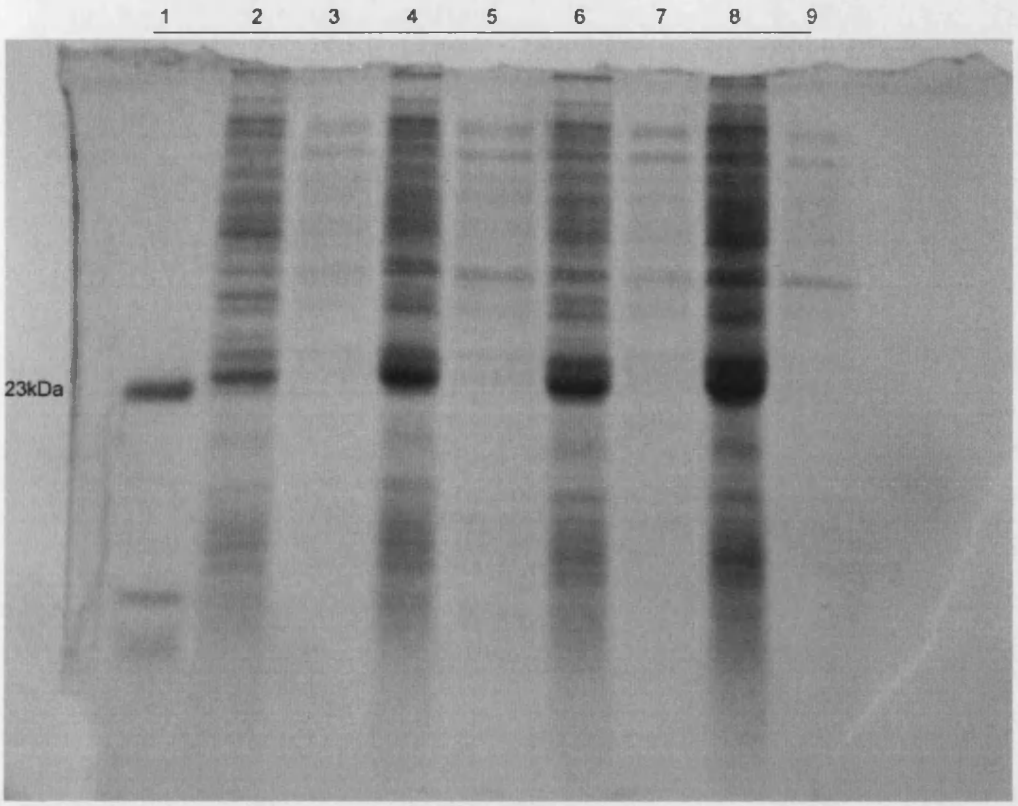



**Figure 5-33: Fractionated BL21(DE3):pET(T), grown at 30°C with late induction**

SDS-PAGE of fractionated induction run of BL21(DE3):pET (T) induced late with 0.4mM IPTG at approximately $A_{600}$ = 1.0, grown at 30°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5μg commercial trypsin; 2 – ($T_1$ – 1hr after induction) insoluble fraction; 3 – ($T_1$) soluble fraction; 4 – ($T_2$) insoluble fraction; 5 – ($T_2$) soluble fraction; 6 – ($T_3$) insoluble fraction; 7 – ($T_3$) soluble fraction; 8 – ($T_{16}$) insoluble fraction; 9 – ($T_{16}$) soluble fraction.

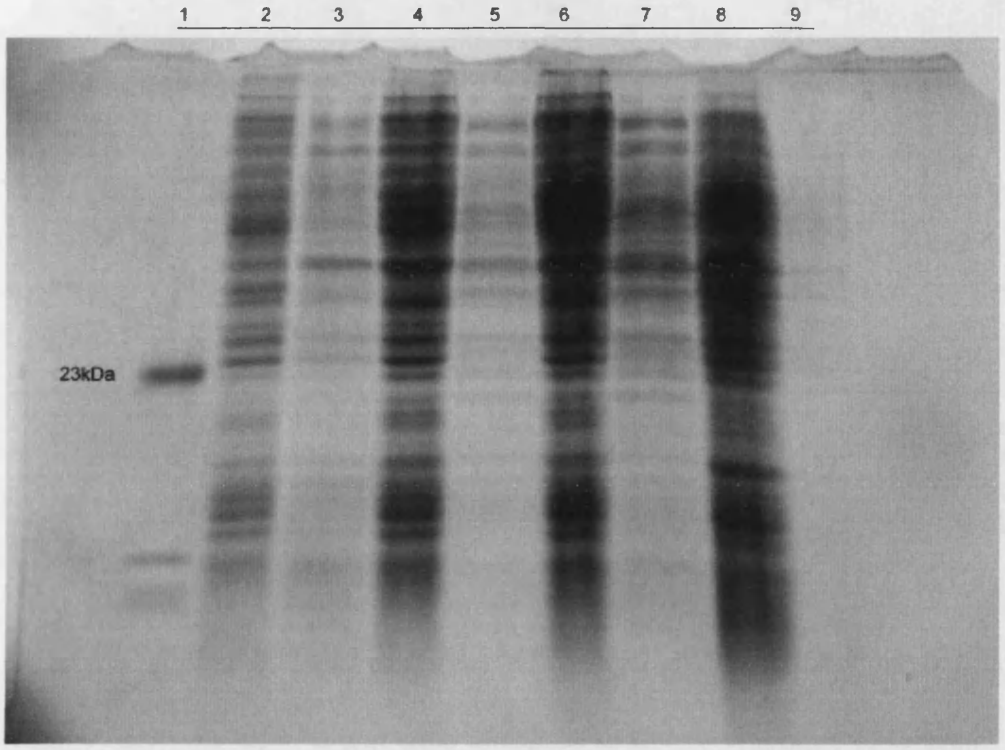
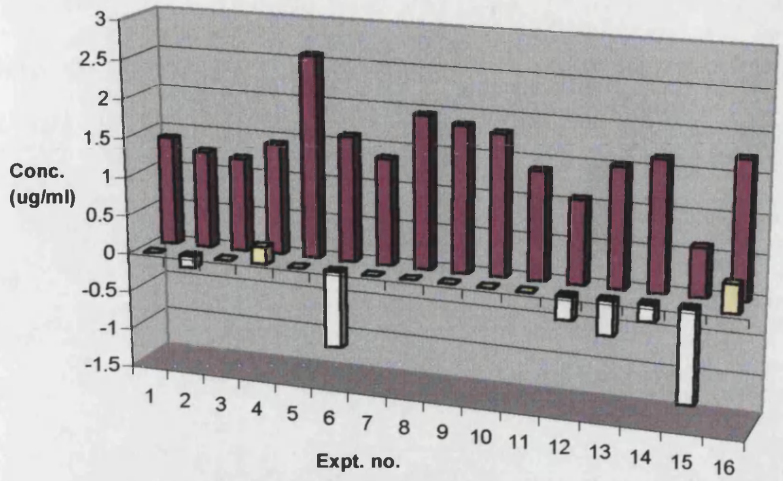



**Figure 5-34: Fractionated BL21(DE3):pET(T), grown at 30°C without induction**
SDS-PAGE of fractionated BL21(DE3):pET (T) non-induced, grown at 30°C. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes: 1 – 7.5μg commercial trypsin; 2 – ($T_1$ – 1hr after induction of culture in Figure 5-25) insoluble fraction; 3 – ($T_1$) soluble fraction; 4 – ($T_2$) insoluble fraction; 5 – ($T_2$) soluble fraction; 6 – ($T_3$) insoluble fraction; 7 – ($T_3$) soluble fraction; 8 – ($T_{16}$) insoluble fraction; 9 – ($T_{16}$) soluble fraction.

The preceding fraction gels appeared to indicate that r-trypsin was accumulating within the cytoplasm as inclusion bodies after induction, but it was necessary to run a comprehensive series of activity assays against L-BAPNA in order to ascertain whether any soluble protein present in the cytoplasm, reaching the cytoplasm, or being released into the supernatant upon cell lysis under any of the fermentation conditions tested. Methodology for these experiments is stated in 5.2.5. All experiments with induced culture were blanked against an uninduced culture processed in exactly the same manner, with the sample reagents. Blank (uninduced) values for this strain on L-BAPNA were relatively high (as compared to ELTRP-1) and so unblanked & blanked results have both been shown in Figure 5-35 (culture grown at 37°C, 4hrs after induction) and Figure 5-36 (culture grown at 37°C, 16hrs after induction).

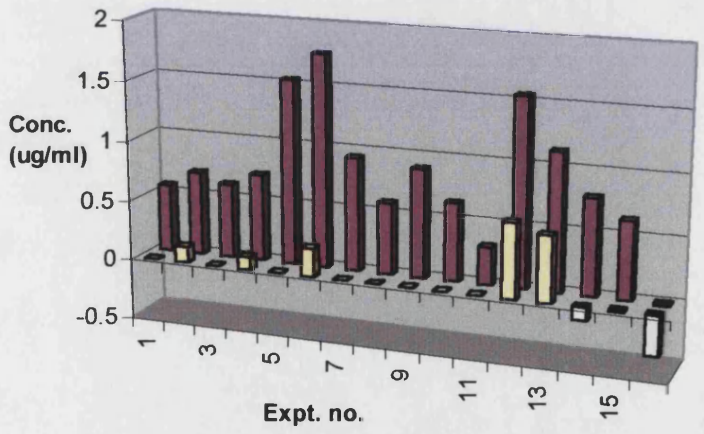



| | StError (in μg/ml) |
|---|---|
| 1 | 0.0732 |
| 2 | 0.101 |
| 3 | 0.112 |
| 4 | 0.0900 |
| 5 | 0.326 |
| 6 | 0.110 |
| 7 | 0.0487 |
| 8 | 0.187 |
| 9 | 0.0748 |
| 10 | 0.176 |
| 11 | 0.0631 |
| 12 | 0.506 |
| 13 | 0.284 |
| 14 | 0.454 |
| 15 | 0.555 |
| 16 | 0.107 |

**Figure 5-35: Activity of BL21(DE3):pET (T) in various conditions, 4hrs after induction**
Cultures were induced with 1mM IPTG after 3.5hrs ($A_{600}$ = 0.62, blanked against uninfected medium). All samples were incubated at room temperature with the various reagents & gently shaken to prevent settling. Addition of substrate was performed immediately prior to reading. All activities were calculated after blanking against the specific uninduced sample blank. All experiments were performed at least in triplicate and standard error refers to deviation between experiments from a single fermentation.

Experiment key for Figure 5-35 & Figure 5-36: **1** - Basic activity blank (no lysis, no denaturant): Uninduced culture + L-BAPNA; **2** - Basic activity (no lysis, no denaturant): Induced culture + L-BAPNA. Blanked against (1); **3** - Periplasmic activity blank: Uninduced culture + L-BAPNA + 0.5μg/ml lysozyme; **4** - Periplasmic activity: Induced culture + L-BAPNA + 0.5μg/ml lysozyme. Blanked against (3); **5** - Total soluble activity blank: Uninduced culture + B-PER + L-BAPNA.; **6** - Total soluble activity: Induced culture + B-PER + L-BAPNA. Blanked against (5); **7** - Insoluble activity blank: Uninduced culture + B-PER + 200μg/ml lysozyme + 0.1M urea + L-BAPNA; **8** - Insoluble activity blank: Uninduced culture + B-PER + 200μg/ml lysozyme + 0.5M urea + L-BAPNA; **9** - Insoluble activity blank: Uninduced culture + B-PER + 200μg/ml lysozyme + 1M urea +

L-BAPNA; **10** - Insoluble activity blank: Uninduced culture + B-PER + 200μg/ml lysozyme + 2M urea + L-BAPNA; **11** - Insoluble activity blank: Uninduced culture + B-PER + 200μg/ml lysozyme + 5M urea + L-BAPNA; **12** - Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 0.1M urea + L-BAPNA. Blanked against (7); **13** - Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 0.5M urea + L-BAPNA. Blanked against (8); **14** - Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 1M urea + L-BAPNA. Blanked against (9); **15** - Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 2M urea + L-BAPNA. Blanked against (10); **16** - Insoluble activity: Induced culture + B-PER + 200μg/ml lysozyme + 5M urea + L-BAPNA. Blanked against (11).




| | StError (in ug/ml) |
|---|---|
| 1 | 0.0732 |
| 2 | 0.101 |
| 3 | 0.112 |
| 4 | 0.0900 |
| 5 | 0.326 |
| 6 | 0.110 |
| 7 | 0.0487 |
| 8 | 0.187 |
| 9 | 0.0748 |
| 10 | 0.176 |
| 11 | 0.0631 |
| 12 | 0.506 |
| 13 | 0.284 |
| 14 | 0.454 |
| 15 | 0.555 |
| 16 | 0.107 |

**Figure 5-36: Activity of BL21(DE3):pET (T) in various conditions, 16hrs after induction**
Cultures were induced with 1mM IPTG after 3.5hrs ($A_{600}$ = 0.62, blanked against uninfected medium). All samples were incubated at room temperature with the various reagents & gently shaken to prevent settling. Addition of substrate was performed immediately prior to reading. All

activities were calculated after blanking against the specific uninduced sample blank. All experiments were performed at least in triplicate and standard error refers to deviation between experiments from a single fermentation.
Experiment key: As above for Figure 5-35.

## 5.4 Discussion

### 5.4.1 Construction of pHKY (M) & pHKY (T)

Although the construction of pHKY (T) plasmid (oligonucleotides coding for pelB inserted at N-terminus of *r-trypsinogen*) appeared to be confirmed by successful transformation & colony propagation, and by PCR (Figure 5-14) originating in the signal sequence, the reasons for the repeated failure of the pHKY (M) construct (pro- sequence excised, pelB insert at N-terminus gene for mature enzyme) were unclear. SDS-PAGE of insoluble & soluble fractions of pHKY (T) (Figure 5-15, Figure 5-16) showed that protein of the correct size was accumulating as inclusion bodies upon induction in approximately similar concentrations as with 'wild-type' ELTRP (see Chapter 3). Single residue activity assays performed on whole cell culture and in the presence of lysis and denaturing agents yielded no signal above the baseline hydrolysis rate of L-BAPNA in the specific conditions. The addition of the denaturing agent in particular (urea) appeared to increase the lability of the substrate approximately three-fold. As insoluble, non-active inclusion bodies appeared to be the default destination for all recombinant protein overexpressed in RV308:pHKY (T) in the course of a standard fermentation, and as no alteration of fermentation environmental conditions and thus expression rate were possible, the decisions was made to proceed with a vector more tailored for this work.

### 5.4.2 Construction of pET (T)

The repeated failures in the ligation of the pET (T) plasmid and the subsequent transformation did not appear to be attributable to any one cause. It was likely that a number of factors contributed, namely: initially relatively low quantities of material were used for the reaction; the purity of the double-digested vector was far from optimal, allowing easily recircularised once-digested plasmids to dominate the sample; the early stage failure to dephosphorylate the cut vector, again allowing

for increased levels of recircularisation. These issues were overcome by means of extended digestion reactions, precise agarose gel extraction of digested vector and the brief use of calf intestinal alkaline phosphate in treating the cut, purified vector. The introduction of XL1-Blue as a cloning intermediate was critical in achieving viable transformants and thus stable, supercoiled pET (T) for subsequent BL21(DE3) transformation.

The cultures did appear to sustain an increased metabolic load upon induction (Figure 5-25). Fractionated, induction SDS-PAGE gels (Figures 5.26-5.34) revealed the accumulation of a protein of approximately the correct molecular weight (25.5kDa) in the insoluble fraction of each culture, irrespective of whether the culture was induced early ($A_{600}$ = 0.6) or late ($A_{600}$ = 1.0), or whether grown at 30°C or 37°C. L-BAPNA activity assays performed on whole cell culture and in the presence of lysis and denaturing agents yielded no signal above the baseline hydrolysis rate of the substrate (Figure 5-35 & Figure 5-36).

The insolubility and thus inactivity of the expressed protein and the lack of an assay that would test specifically for trypsin in a quantitative manner (e.g. a commercially available anti-trypsin IgG to form the basis for an ELISA) means that the 'leakiness' of the T7 promoter system in uninduced cultures could not be accurately assessed. Faint bands of the correct molecular weight were visible in uninduced BL21(DE3):pET(T) cultures (Figure 5-31, Figure 5-34), but this could have been to due to the expression of native chromosomal proteins of a similar molecular weight. Comparison of uninduced RV308:pET (T) cultures with the insoluble fraction of RV308:pET20b(+) (Figure 5-20) supports this hypothesis.

The failure of the BL21(DE3):pET(T) to yield soluble, active r-trypsin, apparently irrespective of fermentation conditions, meant that only two lines of research remained open for the goal of generating a viable high-throughput screen. These were:

a) the processing of the inclusion bodies produced by BL21(DE3):pET(T) under various conditions and subjecting them to a factorial screen of the refolding parameters and co-factors identified in Chapter 3. The aim would be to attempt to establish a simple, microplate-based solubilisation & refolding protocol, similar to that tested in Chapter 3 but based around a more flexible promoter system.

b) the use of error-prone PCR or a suitable mutator strain to generate a mutant library of BL21(DE3):pET(T) with subsequent screening protocols based on the basic L-BAPNA activity assay. As any signal above the baseline would indicate the presence of soluble enzyme, false

positives should be rare and a large number of variants could be generated and tested based on their performance. This procedure was explored in Chapter 6.

## 5.5 Conclusions

The construction of pHKY (T) was shown to be successful, but resulted in no detectable active r-trypsinogen, which continued to accumulate as inclusion bodies within the cytoplasm. No alterations in fermentation conditions proved effective in altering the solubility of the recombinant enzyme.

# 6 Mutagenesis & microwell fermentations

## 6.1 Introduction

### 6.1.1 Aims

The establishment of the BL21(DE3):pET (T) strain was verified in Chapter 5. Expression and subsequent state of the *pelB-r-trypsin* gene fusion was tested by means of fractionated SDS-PAGE and fraction activity assay under a variety of fermentation conditions, involving early & late induction and altered temperature (30°C & 37°C). The recombinant mature bovine trypsin was shown to be being expressed in high concentrations upon induction, but with all detectable enzyme accumulating as inactive inclusion bodies within the cytoplasm, a state that did not appear to be amenable to improvement by alteration in either induction level or temperature of the fermentation. As rational methods had thus failed to solubilise the enzyme, the logical next step was to construct a mutant library based on a simple activity screen. The current baseline of no detectable enzyme activity added to the simplicity and sensitivity of this process, with the result that any signal from a library variant tested on L-BAPNA implied directly that a certain percentage of the protein would be present in soluble, active form.

The screening of even a small, low-success probability library (e.g. $\geq$ 2000 mutants) demanded the use of both microwell fermentations (for the individual culturing of the variants) and a microplate based activity assay (developed in Chapter 2). Micro-scale fermentation work, on-going within the department, had shown that small-scale fermentations (1L) could be could be simulated, with various degrees of accuracy, by fermentations with a 96 deep, square-welled microplate. Such plates could be filled to a maximum of one fifth of their total volume with media and the crucial limiting factor in micro-scale down, that of oxygen transfer, was improved by means of high agitation rates, in the region of 900rpm on an orbital shaker, with the corners of the square wells acting as baffles and thus creating turbulence and enhanced mixing. For BL21(DE3):pET (T) mutants to be evaluated in this way, it was essential to first chart the growth, expression levels & enzyme activity of the un-mutated 'wild-type' over the course of a microwell fermentation.

### 6.1.2 Use of mutator strains

The efficiency and flexibility of PCR-based methods for introducing mutations into a defined region of DNA, had been demonstrated by numerous research teams (Arnold and Volkov ,1999; Minshull and Stemmer, 1999)). The process was moderately labour intensive however, as it required that the mutated gene be cloned back into the vector. *E. coli* mutator strains have the potential to simplify the process of library generation by accumulating mutations over the uncut plasmid, requiring simply a pooled lysis step and the subsequent transformation step into the expression strain. The XL1-Red mutator strain, a derivative of XL1-MRA (Stratagene) was developed by Greener and Callahan (Greener *et al.*, 1996) with the aim of being able to generate single-base mutation in a gene of interest over the course of a single over-night fermentation. To achieve this, 3 well defined DNA repair pathways were irreversibly disabled. These corresponded to the genes: *mutS* – which contributes to the mismatch repair pathway; *mutT* – part of the oxo-dGTP repair pathway; and *mutD* – repairs base substitutions and frameshifts by forming, in conjunction with *dnaQ*, the 3'-5'exonuclease epsilon subunit of DNA polymerase III (Miller, 1998). The researchers claim that this gave the strain a base mutation rate approximately 5000 times higher than that of wild-type *E. coli* and that the spontaneous mutation rate on a pBluescript plasmid was now measured at 1 mutation/2kb cloned DNA after 30 generations of exponential growth. As a possible side effect to the extensive genetic manipulation of the strain, the doubling time of XL1-Red was now increased to 90 minutes at 37°C. In order to allow the plasmid to be propogated through at least 30 generations with an overnight fermentation, colonies were first plated, then pooled (see Methods, section 6.2.3).

The test plasmid pET(T) was far from optimal for this work as *r-trypsin* constituted only an approximate sixth of the plasmid's total sequence, and the percentage chance of a single mutation landing within this region was correspondingly low. An additional worry was the relatively low copy-number of the pET(T) plasmid (pET vectors have a published copy number of 10-15, while the pBluescript on which the developers' calculations were based has a copy number of 25) (Greener, Callahan, and Jerpseth, 1996). It was decided to attempt to construct a mutation calibration curve using XL1-Red:pET (T), continually rediluting the test culture for continued exponential growth and thus propagating the plasmid through 30, 40, 50, 60 generations. The aims were to calculate the overall mutation rate of the strain, to compare the quantity of mutations within the target region with

the theoretical predictions and to attempt to calculate the impact of plasmid copy-number in such a process.

### 6.1.3 Microwell fermentations

An effective strategy of colony screening would require standardizing the growth rate of the well-cultures so that all wells could be induced simultaneously at the identical cell density. Novagen literature and the previous characterization work with RV308:pET (T) (Chapter 5) had established an induction OD of approximately $A_{600}$ = 0.6 (blanked against uninfected media) to be optimal for this strain. The physical dimensions and the opacity of the 96 square deep-well plates (□-plate) prevented direct measurement of the cultures by any of the available spectrophotometers. Growth rates would thus be initially determined by means of sacrificial wells and, on experimental plates, by means of a single sacrificial 'marker' well inoculated with 'wild-type' RV308:pET (T). A uniform growth profile across the plate would be dependant upon a standardized inoculum being given to each well. It has been proven that the concentration of cells transferred by a manual or automated colony pick can vary by up to two orders of magnitude, resulting in well cultures entering logarithmic growth at intervals separated by up to several hours (unpublished departmental work). A further problem arose in that the colony picker was unable to inoculate deep-well plates unless they were filled with over 1.5mls of media, with1ml being known to be the maximum fill level possible before splashing would occur at 900rpm. Both problems could be solved by means of an overnight incubation in a standard 'shallow' 96-well plate. Experimental transformants would be picked and inoculated into 100μl of media in a standard 96-well plate which would then be covered by a gas-diffusible membrane and incubated overnight to allow all well cultures to achieve an approximately equal end-density.

Opportunities for cross contamination of wells were predicted to occur at various points in the fermentation/screening protocol and could have a significant impact on the efficiency of the process in terms of the masking of less fit mutants and the generation of false-positives. The impact of errors accumulating in this way would be exacerbated with any increase in library size and thus it was essential to attempt to model the various forms of contamination using initial control plates. The shallow-well 16-hour incubation plates would be covered with a gas-permeable membrane following colony picking and thus could be agitated at 900rpm without fear of cross-

contamination. Removal of the membrane inevitably resulted in a slight splash effect across the plate. This was minimized by the use of only limited volume (100μl) in each well. The primary window for well cross-contamination was during the open square 96 deep-well fermentation. It had been shown that the use of the gas-permeable membrane during a fermentation involving a ml or more of culture severely retarded the rate of oxygen transfer, and thus growth (unpublished departmental work). Temperature control in the current generation of open thermomixers was also proven to be unreliable, with a significant gradient between the top & bottom of a well, the only suitable solution involved installing the thermomixer within a sterilized dedicated incubator. The fermentation would then be run with the incubator could providing uniform temperature control and the thermomixer providing the high agitation rates necessary. Although no splashing was visible above the lip of the wells when 1ml of culture was mixed at 900rpm, the aerosolization due to high agitation and evaporation made a degree of contamination inevitable.

## 6.2 Materials and methods

### 6.2.2 Strains & vectors used

RV308 (Eli Lilly): (ELTRP-1 = RV308 + pHKY603) Strain is a derivative of *E. coli* K12. Fully genotype not released by company. Resistance: $Sm^r$.

BL21(DE3) (Novagen): derivative of B strain, $F^-$, *ompT* $hsdS_B$ ($r_B^- m_B^-$) *gal dcm* (DE3)

XL1-Red (Stratagene): *endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac mutD5 mutS mutt*, Tn10($tet^r$)

XL1-Blue (Stratagene): *recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F' *proAB* $lacI^qZ\Delta M15$ Tn*10* ($tet^r$)]

pHKY603 (Eli Lilly): a derivative of pBR322. Fully sequence not released by company. Resistance: $Tc^R$,

pET26b(+) (Novagen): T7*lac*, His•tag, f1 ori, pelB, $kan^r$

### 6.2.3 Use of XL1-Red to generate *r-trypsin* mutant library

Transformation of pET(T) into XL1-Red and subsequent growth was performed as specified in the manufacture's protocol with additional reference to the developers' notes (Greener, Callahan, and

Jerpseth, 1996). The critical points are summarized here. The Epicurian coli commercially competent cells were thawed on ice, with β-mercaptoethanol being added to a final concentration of 25mM. 10-50ng of mini-prep quality plasmid DNA was added, allowed to stand on ice for 30min, and then transformed by a 45sec heat-shock at 42°C. The mix was then rested on ice for 2min and then allowed 1hr of non-selective incubation in sterile SOC at 37°C. The mix was then plated on selective LB agar ($kan^+$ 30μg/ml) and incubated overnight at 37°C. The following day, a suitable plate (containing ≥ 200 colonies) was used to create the starter culture for the overnight liquid media fermentation. The colonies were pooled by washing the plate with a ml of sterile LB, which was then added to a further 9ml of LB broth ($kan^+$ 30μg/ml). The culture was grown for an initial 15hrs to achieve the target of 30 XL1-Red generations and was then rediluted 1:1000 with sterile broth and grown for a further 15hrs to propagate the plasmid through a further 10 generations (90 minute doubling time × 10 generations). This sub-culturing step was repeated as desired. Plasmid was recovered via standard mini-prep and transformed into XL1-Blue and subsequently plated (for propagation & analysis) and BL21(DE3) (for expression & testing of r-trypsin). In order to obtain plasmid isolates for sequencing and thus calibration of mutation rate, random discrete XL1-Blue isolates were selected after plating and used to infect an overnight 5ml culture. Plasmid DNA was isolated and concentrated via mini-prep and quality was assessed prior to sequencing by means of 0.8% agarose gel electrophoresis & PCR using T7 promoter & T7 terminator primers (Novagen) under standard conditions.

### 6.2.4 Microwell fermentations

For the experiments involving the microwell fermentation of BL21(DE3):pET (T) and its library variants, the following protocol was devised, with occasional slight variations indicated at the relevant points in the Results section.

A sterile standard shallow 96-well microplate (Starstedt) had each test well filled with 100μl sterile medium (either LB or M9SD) with 70μg/ml kanamycin. A colony picker (QPix) was used to inoculate each well with the appropriate strain from an agar spread plate infected no longer than 1 week previously and stored at 4°C. The plate would then be covered with Breeth Easy membrane (Sigma Aldrich) and secured within a standard orbital shaker-incubator. Conditions would be set at 37°C, 200rpm and fermentation would be allowed to occur for approximately 16-20hrs. Growth of

the well cultures in the plate would then be analysed in a microplate reader at $A_{600}$, blanking against wells containing uninfected medium. The BreethEzee film had to be removed gently to avoid the creation of a partial vacuum above the wells, leading to aerolization of droplets. For the primary induction run of the cultures, a sterile square, 96-deep-well plate (□-plate) would be prepared with 1ml media (70μg/ml kan) in each blank, control & test well. 20μl of inoculum from the appropriate well of the previous overnight plate would be transferred by multipipette and the □-plate would be secured on a thermomixer, which would then be placed within a sterilized, dedicated incubator. The orbital shaker on the thermomixer would be set to deliver 900rpm with the heating plate disabled, and the incubator would be set to 37°C. As these were 'open fermentations' the door of the incubator was sealed with lab tape. The fermentation was allowed to proceed until the cultures reached a blanked $A_{600}$ of approximately 0.6-0.7 (time point determined empirically by means of sacrificial wells in control plates) and they were then induced with 10μl 100mM IPTG (final conc. of 1mM). The fermentation then continued under the same conditions for approximately 16hrs. At the termination of the run, 500μl of each of the well cultures was processed (cellular fractionation, addition of lysozyme or reducing agents) and then either subjected to activity testing against L-BAPNA or denatured and run on an SDS-PAGE. The remaining 500μl of each test culture would either be transferred into individual, sterilized eppendorf tubes and stored at –20°C (considered viable for up to one week after freezing), or have sterilized glycerol added to a final concentration of 20% and stored at -80°C.

### 6.2.5 5ml fermentation procedure

For the 5ml fermentation work, which was intended to characterize in greater detail the favorable BL21(DE3):pET (T) mutant variants identified through the microplate screening work, the following protocol was used.

A sterile 50ml falcon tube would be filled with 5ml sterilized media (kan 70μg/ml) and capped with a porous foam bung. This would be inoculated with the appropriate strain either from glycerol stocks or from a recently infected agar spread plate. This was placed within an orbital shaker-incubator, securing all tubes at an angle of approximately 45° and allowed to grow overnight at 37°C, 200rpm. For the primary induction run of the cultures, an identical tube would be prepared and inoculated with 50μl of the overnight culture. This fermentation would proceed at

200rpm and a set temperature (either 28°C, 32°C or 37°C) until in reached a blanked OD of approximately 0.6-0.7. The test & control cultures would then be induced by addition of 50μl 100mM IPTG (final conc. of 1mM) and the fermentation would proceed under the same conditions for approximately 16hrs. Upon completion the cultures were processed and stored in a manner identical to that of the deep-well cultures (Section 6.2.4).

### 6.2.6 100ml fermentation procedure

100ml fermentations of the potentially interesting BL21(DE3):pET (T) variants indicted by microplate screening were performed in order to obtain greater quantities of material for analysis and explore the possibility of scale-up work. All 100ml fermentations were performed according to the following protocol.

A sterile 50ml falcon tube would be filled with 5ml sterilized media (kan 70μg/ml) and capped with a porous foam bung. This would be inoculated with the appropriate strain either from glycerol stocks or from a recently infected agar spread plate. This was placed within an orbital shaker-incubator, securing all tubes at an angle of approximately 45° and allowed to grow overnight at 37°C, 200rpm. For the primary induction run of the cultures, a sterilized 500ml baffled flask would be filled with 100mls sterilized medium (kan 70μg) and capped with a foam bung. This would be inoculated with 500μl of the appropriate overnight culture. These flasks would then be secured within an orbital shaker-incubator and the fermentation would proceed at 200rpm and at a set temperature (either 28°C, 32°C or 37°C). Induction would occur when the cultures achieved a blanked OD $A_{600}$ of approximately 0.6-0.7 and by the addition of 1ml 100mM IPTG (final conc. of 1mM). The fermentation would then proceed for a further 16hrs under the identical conditions. Upon completion the cultures were processed and stored in a manner identical to that of the deep-well cultures (Section 6.2.4)

### 6.2.7 Cellular fractionation

The protocols for the separation of the *E. coli* cultures into periplasmic, cytosolic & insoluble fractions were the same as described in chapter 5.

### 6.2.8 Activity assays

Each enzymic activity assay performed on the BL21(DE3):pET (T) cellular fractions used the FLUOstar optima plate-reader (BMG). The assay was set up & controlled via the proprietorial software version 1.10-0. All tests were run in standard Starstedt shallow, 96-well microplates. The assay read the absorbance of the well solution at $A_{405}$, with the reading period defined as a single kinetic window with each of the 20 readings consisting of 3 flashes, and a total of 65s to process a complete cycle. The plate was initially agitated (internal orbital shaker, 5mm displacement) for 10s prior to the first reading. Each reaction plate was set up according to the individual needs of the experiment, but in each case final well volume was held at 200µl. Substrate concentration (L-BAPNA) was kept at 1mM, added immediately prior to reading as a 50µl 4mM aliquot.

## 6.3 Results

### 6.3.2 Microwell fermentations of un-mutated BL21(DE3):pET (T)

Initially, it was critical to establish the growth rate & induction point of un-mutated BL21(DE3):pET(T) in a 96 square, deep-well microplate, and investigate general compatibility with the QPix colony picker. Growth of the cultures was measured in both LB and M9SD media and was assessed by means of sacrificial wells. The initial overnight inoculum plate was prepared by using the Qpix to pick BL21(DE3):pET(T) colonies from recently infected LB agar plates (kan 70µg/ml) to inoculate 48 wells of LB media (kan 70µg/ml) and 48 wells of M9SD media (kan 70µg/ml) on a standard shallow, 96-well plate. Readings were taken at $A_{600}$ once per hour with 3 wells being used for each data point.

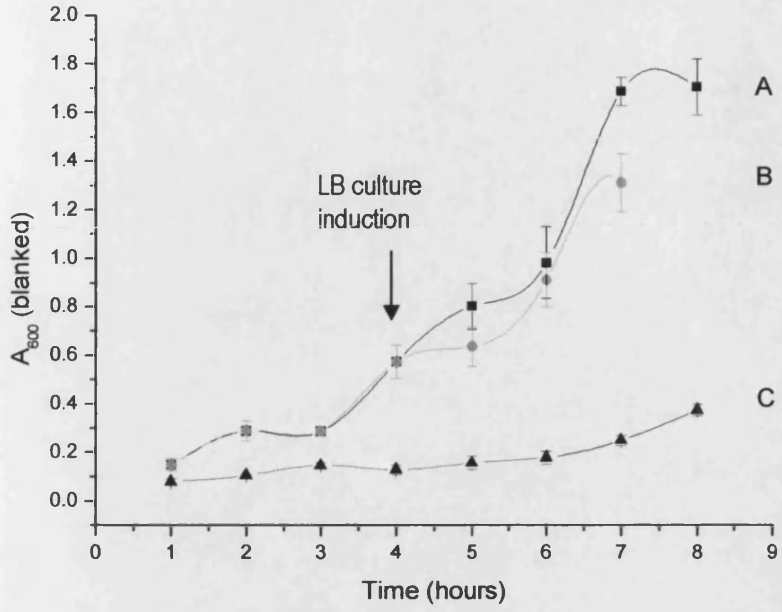



**Figure 6-1: Growth of BL21(DE3):pET(T) in a □-plate in LB and M9SD media at 37°C, 900rpm**
Key: A (—■—) = Induced LB media; B (—●—) = uninduced LB media; C (—▲—) = uninduced M9SD. 20μl from the overnight inoculum plate was used to inoculate 1ml of media. Density of inocula well cultures (blanked against uninfected media) at point of induction: LB = $A_{600}$ of 0.579, StDev of 0.215; M9SD = $A_{600}$ of 0.425, StDev of 0.152. LB cultures were induced with 1mM IPTG at 4hrs ($A_{600}$ = 0.582), M9SD cultures failed to achieve the required density over the course of the experiment and were not induced. Error bars represent standard error from three separate well-cultures.

Activity profiles of the fractionated LB cultures were performed in the presence of the same lysis & denaturing agents as those performed on the 100ml fermentations of un-mutated BL21(DE3):pET(T) (see Figures 5-35 & 5-36). The assay was performed on samples harvested 4hrs after induction and samples collected at the termination of the experiment, approximately 16hrs after induction. Blank (uninduced) values for this strain on L-BAPNA were relatively high (as compared to ELTRP-1) and so both unblanked & blanked results have both been shown in Figure 6-2 & Figure 6-3 so as to illustrate this fact.

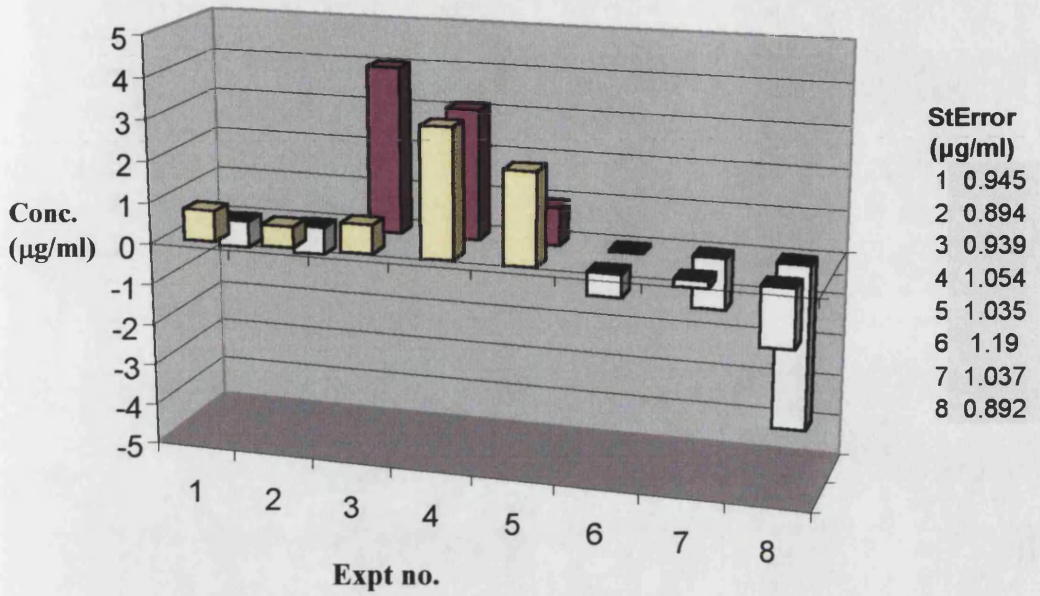



**Figure 6-2: Activity of BL21(DE3):pET(T) 4hrs after induction**

Activity profile (in terms of µg/ml trypsin) of BL21(DE3):pET(T) on 1mM L-BAPNA, 4hrs after induction of a microplate fermentation held at 37°C, 900rpm (see Figure 6-1). Key: Yellow bars = blanked positive results; purple bars = unblanked positive results; negative results shown in white. Cultures were induced with 1mM IPTG after 4hrs (blanked $A_{600}$ of approximately 0.579, based on previous work: Figure 6-1). Each test was blanked against a series of wells containing identical conditions and uninduced culture. All experiments were performed at least in triplicate, results pooled from 2 fermentations under identical conditions. Experiments: 1 - Basic activity (no lysis, no denaturant): Induced culture + L-BAPNA); 2 - Periplasmic activity (Induced culture + L-BAPNA + 0.5mg/ml lysozyme); 3 - Total soluble activity (Induced culture + B-PER + L-BAPNA); 4 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 0.1M urea + L-BAPNA); 5 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 0.5M urea + L-BAPNA); 6 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 1M urea + L-BAPNA); 7 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 2M urea + L-BAPNA); 8 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 5M urea + L-BAPNA).

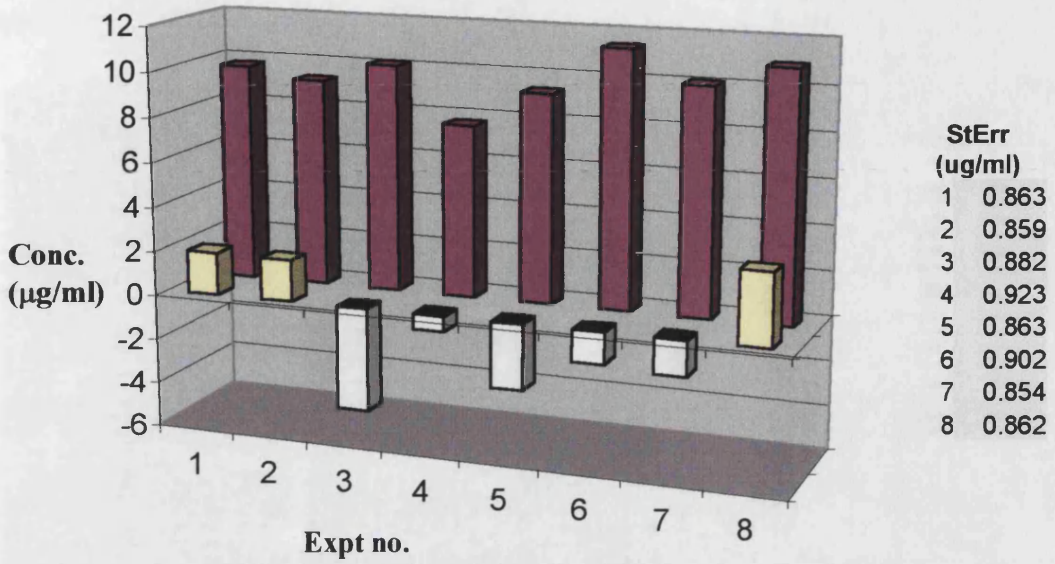



**Figure 6-3: Activty of BL21(DE3):pET(T) 16hrs after induction**

Activity profile (in terms of μg/ml trypsin) of BL21(DE3):pET(T) on 1mM L-BAPNA, 16hrs after induction of a microplate fermentation held at 37°C, 900rpm (see Figure 6-1). Key: Yellow bars = blanked positive results; purple bars = unblanked positive results; negative results shown in white. Cultures were induced with 1mM IPTG after 4hrs (blanked $A_{600}$ of approximately 0.579, based on previous work: Figure 6-1). Each test was blanked against a series of wells containing identical conditions and uninduced culture. All well experiments were performed at least in triplicate, results pooled from 2 fermentations under identical conditions. Experiments: 1 - Basic activity (no lysis, no denaturant): Induced culture + L-BAPNA); 2 - Periplasmic activity (Induced culture + L-BAPNA + 0.5mg/ml lysozyme); 3 - Total soluble activity (Induced culture + B-PER + L-BAPNA); 4 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 0.1M urea + L-BAPNA); 5 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 0.5M urea + L-BAPNA); 6 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 1M urea + L-BAPNA); 7 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 2M urea + L-BAPNA); 8 - Insoluble activity (Induced culture + B-PER + 200mg/ml lysozyme + 5M urea + L-BAPNA).

An SDS-PAGE gel of the culture in both media, 4 & 16hrs after induction was prepared in order to provide direct evidence for the overexpression and accumulation of r-trypsin in either the soluble or insoluble fraction.

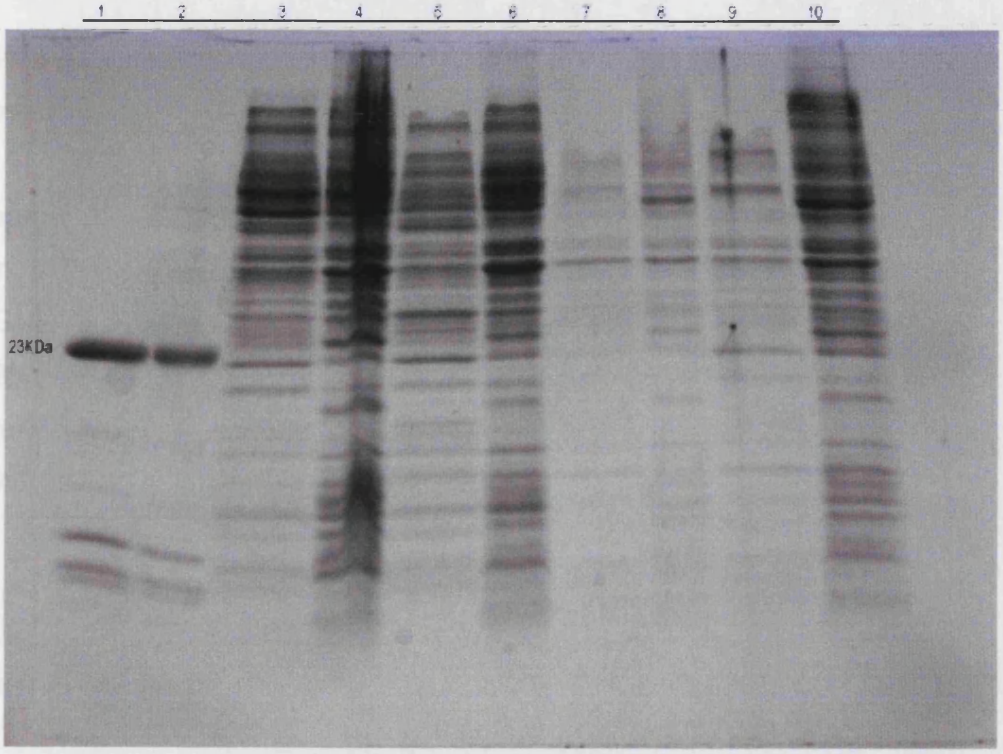



**Figure 6-4: Soluble & insoluble fractions from BL21(DE3):pET(T) fermentation**
SDS-PAGE of soluble & insoluble fractions from un-mutated BL21(DE3):pET (T) microwell fermentations in LB & M9SD medium. 15% acrylamide gel with 4% stacking gel, run @ 80V for 40min, then @ 150V for 120min. Lanes 3-6 hold fractions from culture grown in LB medium, lanes 7-10 hold fractions from culture grown in M9SD medium. Both cultures were grown in square, 96 deep-well plates (1ml culture, 37°C, 900rpm), with the LB culture being induced with a final conc. of 1mM IPTG 4hrs after inoculation. Lanes: 1 - 15μg commercial trypsin; 2 – 7.5μg commercial trypsin; 3 – soluble fraction from LB culture, 4hrs after induction; 4 - insoluble fraction from LB culture, 4hrs after induction; 5 - soluble fraction from LB culture, 16hrs after induction; 6 - insoluble fraction from LB culture, 16hrs after induction; 7 - soluble fraction from M9SD culture, 4hrs after induction; 8 - insoluble fraction from M9SD culture, 4hrs after induction; 9 - soluble fraction from M9SD culture, 16hrs after induction; 10 - insoluble fraction from M9SD culture, 16hrs after induction.

### 6.3.3 Generation of first BL21(DE3):pET (T) mutant library

#### 6.3.3.1 Activity of library on L-BAPNA after microwell fermentation

Plasmid was propagated through 30 generations of XL1-Red growth as described in Section 6.2.3. After lysis of the mutator culture, pET(T) was transformed into BL21(DE3) and spread onto selective agar plates (Kan$^+$ 70μg/ml).

The initial whole cell activity test was performed by picking distinct colonies from the spread plates constituting the library and inoculating a shallow, 96-well microplate. This was then incubated overnight and subsequently used to infect a □-plate (as described in Section 6.2.3). Growth rate and thus induction time were based on those determined empirically for un-mutated BL21(DE3):pET (T) (Figure 6-1). For each screening run, two 96-well plates were infected, a restriction necessary due to the availability of only two thermomixers which were essential for the induction fermentation.

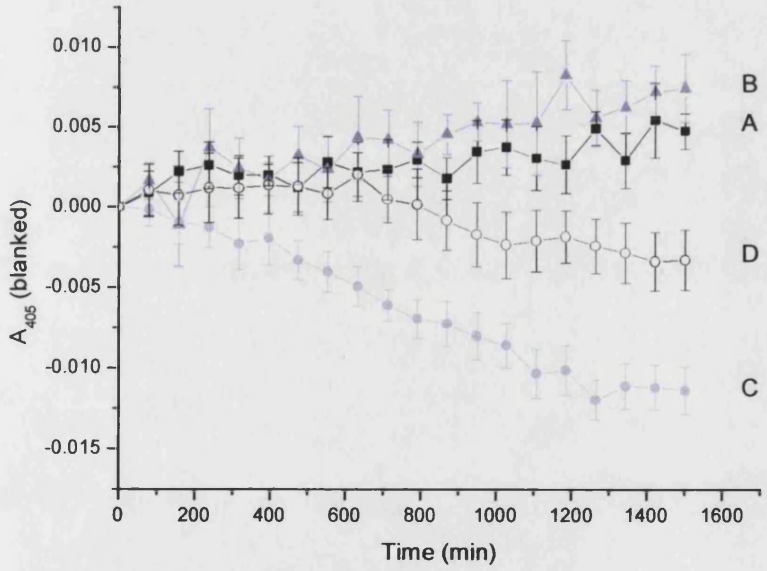



**Figure 6-5: Top 20 isolates from plates A & C, 4hrs after induction**

Average activity of top 20 performing isolates from plates A & C with un-mutated strain as control, 4hrs after induction. Key: A (—■—) = uninduced 'wild-type' pET(T):BL21(DE3); B (—▲—) = induced 'wild-type' pET(T):BL21(DE3); C (—●—) = top 20 isolates from plate A; D (—○—) = top 20 isolates from plate C. Deep-well cultures were inoculated from overnight plate and induced to a final conc.

of 1mM IPTG after 4hrs, conditions held at 37°C, 900rpm. Whole cell culture was mixed with lysis agent (B-PER) for 10min and then assayed immediately upon addition of L-BAPNA to a final concentration of 1mM. Error bars on activity of un-mutated strain indicate standard error calculated over 3 separate fermentation runs; error bars on mutant activity refer to deviation between isolates in the initial screen.

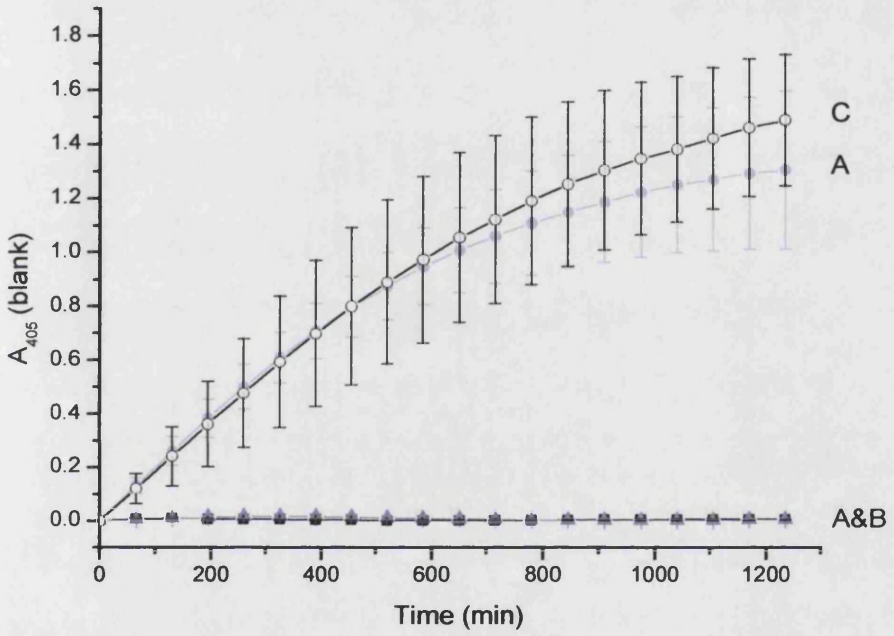



**Figure 6-6: Top 20 isolates from plates A & C, 16hrs after induction**

Top 20 performing isolates from plates A & C, with un-mutated strain as control, 16hrs after induction. Key: A (—■—) = uninduced 'wild-type' pET(T):BL21(DE3); B (—▲—) = induced 'wild-type' pET(T):BL21(DE3); C (—●—) = top 20 isolates from plate A; D (—○—) = top 20 isolates from plate C. Deep-well cultures were inoculated from overnight plate and induced to a final conc. of 1mM IPTG after 4hrs, conditions held at 37°C, 900rpm. Whole cell culture was mixed with lysis agent (B-PER) for 10min and then assayed immediately upon addition of L-BAPNA to a final concentration of 1mM. Error bars on activity of un-mutated strain indicate standard error calculated over 3 separate fermentation runs; error bars on mutant activity refer to deviation between isolates in the initial screen. Isolates selected for both plates are indicated in bold in Table 6-1.

**Plate A (μg/ml trypsin)**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 1.75 | 1.98 | 4.36 | 1.32 | 5.47 | 2.82 | -0.12 | 3.04 | 3.16 | 22.06 | **117.58** | 75.23 |
| **B** | 4.28 | 0.86 | 1.13 | 5.00 | 3.47 | 7.12 | 5.95 | 21.99 | 21.80 | 71.38 | **147.77** | **99.32** |
| **C** | 11.32 | 3.82 | -0.29 | -1.97 | 3.73 | 36.35 | 41.34 | 56.18 | 56.01 | 28.59 | 90.46 | **96.71** |
| **D** | 7.75 | 6.71 | 1.68 | 75.00 | 14.99 | **112.35** | **113.26** | 71.77 | 67.64 | 18.18 | 5.65 | 68.84 |
| **E** | 17.09 | 20.53 | 47.06 | 4.97 | 91.77 | **135.32** | 32.38 | 96.38 | **114.79** | 44.67 | 12.99 | 8.89 |
| **F** | 12.25 | 61.55 | 59.71 | 41.17 | 96.36 | **102.95** | **154.21** | **130.11** | **124.90** | **101.92** | 61.93 | 7.26 |
| **G** | 5.20 | 11.16 | **106.41** | 94.45 | 89.16 | **96.53** | **101.14** | **109.16** | **97.95** | **134.97** | 61.49 | 39.76 |
| **H** | 5.90 | 36.49 | 66.40 | 68.23 | 92.14 | 65.93 | 67.06 | 92.03 | **99.84** | 29.09 | 74.21 | 34.17 |
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |

**Plate C (μg/ml trypsin)**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 7.40 | 8.52 | 7.53 | 16.85 | 10.12 | 53.53 | 1.60 | 3.07 | 4.90 | 3.12 | 8.49 | 2.24 |
| **B** | 5.88 | 10.92 | 7.78 | 8.10 | 9.33 | 48.41 | 2.60 | **59.98** | **92.71** | 5.52 | 7.74 | 3.32 |
| **C** | 12.50 | 5.32 | 7.47 | **160.31** | 29.63 | **143.69** | **149.95** | **158.80** | 53.15 | 10.66 | 8.22 | 8.13 |
| **D** | **157.11** | **156.12** | 8.05 | **161.32** | **104.18** | **56.00** | 98.35 | **117.51** | 13.51 | 7.79 | 8.13 | 3.10 |
| **E** | 5.80 | 48.39 | 6.35 | **70.90** | 8.98 | 6.42 | **123.12** | **172.10** | 4.39 | 4.80 | 3.37 | 5.53 |
| **F** | 55.09 | 44.11 | **92.90** | 8.02 | 8.15 | 6.69 | 8.43 | 5.30 | 7.31 | -1.62 | 7.45 | 2.95 |
| **G** | **112.93** | **112.18** | 4.69 | 8.42 | 7.74 | **68.54** | 41.72 | 7.05 | 8.70 | 1.43 | 8.95 | 2.28 |
| **H** | 6.73 | 4.38 | 8.16 | 7.46 | 7.42 | 17.07 | 11.14 | 3.93 | 8.36 | 3.95 | 8.10 | 1.54 |
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |

**Table 6-1 :Results from plates A & C from first library, 16hrs after induction of well cultures.**
Results in μg/ml of trypsin activity, measured against standards of commercial trypsin run on a separate plate. Isolates in bold represent the twenty top performing mutants from each plate.

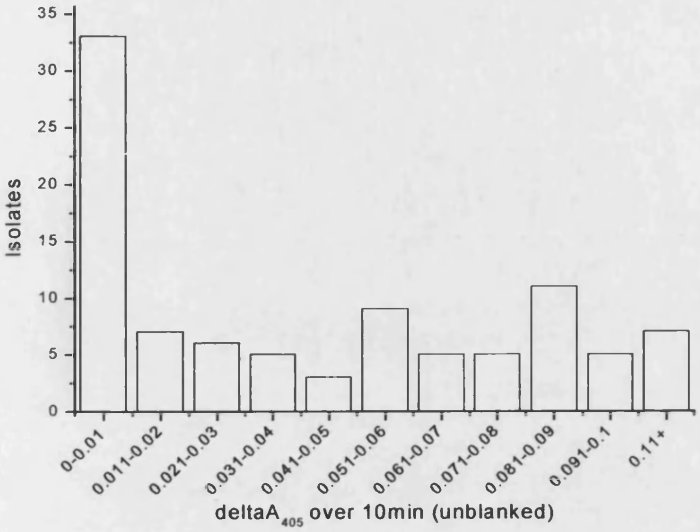



**Figure 6-7: Activity histogram from 'A' plate**

Activity distribution of the 96 screened BL21(DE3):pET (T) variants from the 'A' library plate.

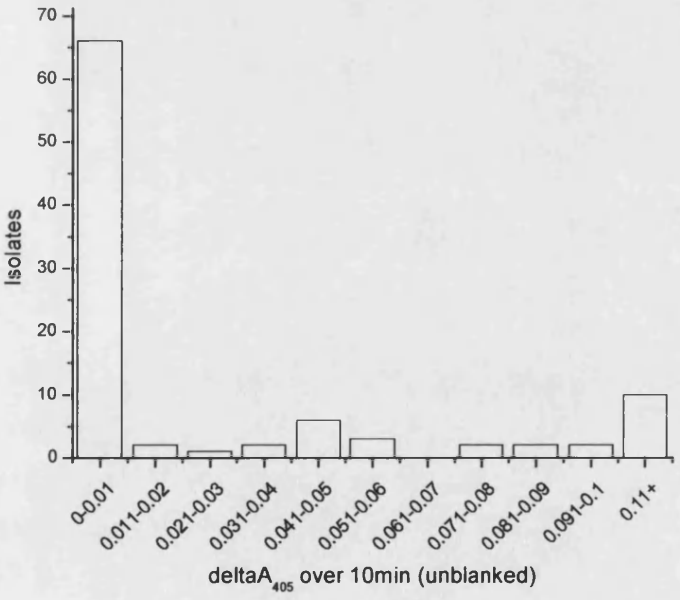



**Figure 6-8: Activity histogram from 'C' plate**

Activity distribution of the 96 screened BL21(DE3):pET (T) variants from the 'C' library plate.

Stringent testing of the best performing isolates was necessary in order to distinguish between genuine soluble mutants and false positives arising from the open deep-well fermentation. 10 from each plate were selected for further testing, including both the strongest performers and a sample of those in close proximity to them. Retesting involved the culturing of each clone from streak plates grown from original glycerol stocks, and growth in 6 individual deep-wells, separated from other well-cultures by at least a column of uninfected wells. Mutants were blanked against induced & uninduced 'wild-type' pET(T):BL21(DE3) and activity was calibrated by the supplementation of uninduced wild-type cultures with commercial trypsin.

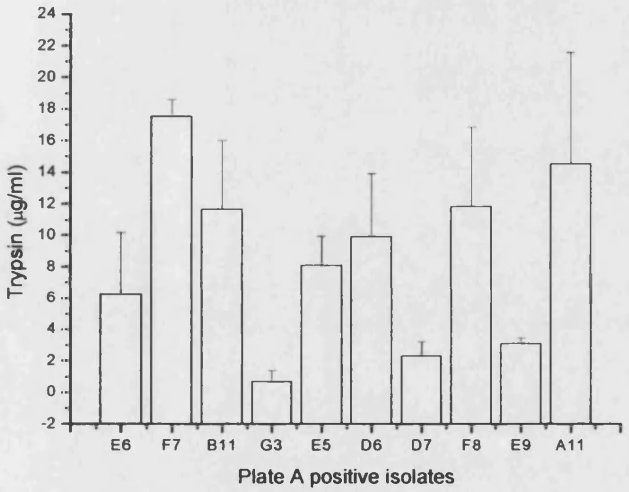



**Figure 6-9: 'A' plate Isolate retesting**

Retesting of positive isolates from 'A' library plate by microwell fermentation in presence of appropriate controls. Activity read over 20min @ $A_{405}$, blanked against induced activity of un-mutated BL21(DE3):pET (T). Whole cell culture was mixed with lysis agent (B-PER) for 10min and then assayed immediately upon addition of L-BAPNA to a final concentration of 1mM. X-axis label indicates the well position of the isolate on the initial library plate. Error bars refer to standard error from 6 individual well-cultures.

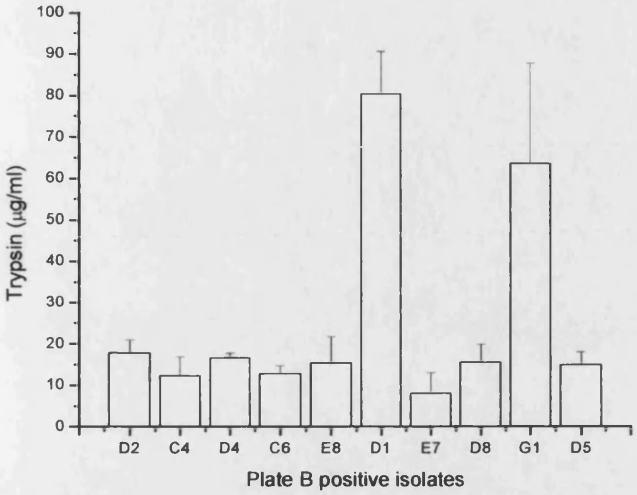



**Figure 6-10: 'C' plate isolate retesting**

Retesting of positive isolates from 'C' library plate by microwell fermentation in presence of appropriate controls. Activity read over 20min @ $A_{405}$, blanked against induced activity of un-mutated BL21(DE3):pET (T). Whole cell culture was mixed with lysis agent (B-PER) for 10min and then assayed immediately upon addition of L-BAPNA to a final concentration of 1mM. X-axis label indicates the well position of the isolate on the initial library plate. Error bars refer to standard error from 6 individual well-cultures.

### 6.3.3.2 Expression profile & activity of library on L-BAPNA after 5ml fermentation

Based upon the results from the further analysis of the positive isolates from library plates A & C (see Figure 6-9 & Figure 6-10 respectively), 5 variants from plate C were selected for further analysis. These were, in no order: G1, D1, D2, D4, C4. So as to better characterize these variants in terms of expression levels, plasmid stability & growth rate with regard to the potential autotoxicity of the recombinant mature trypsin, a series of 5ml fermentations was performed. A profile of r-trypsin expression levels and activity in LB at growth temperatures of 28°C, 32°C & 37°C, 200rpm over the course of a 5ml fermentation was built by means of growth plots (Figures 6.12-14), L-BAPNA activity testing of cellular fractions (Figure 6-15: 28°C; Figure 6-20: 32°C; Figure 6-25 &Figure 6-26: 37°C) & SDS-PAGE gels (Figures 6.16-19: 28°C; .Figures 6.21-24: 32°C; Figures 6.27-30: 37°C).

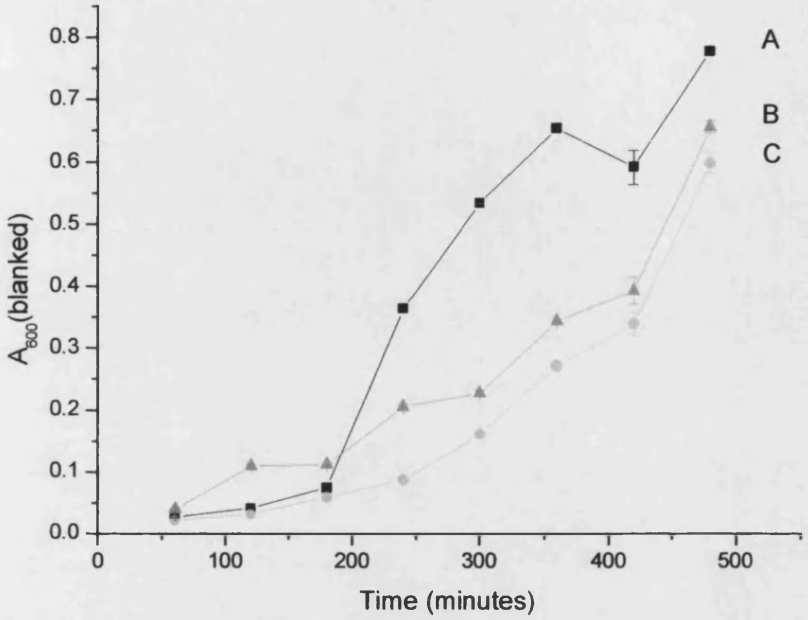



**Figure 6-11: Growth of un-mutated BL21(DE3):pET (T) in a 5ml LB culture at 28°C, 32°C & 37°C**

Key: A (—■—) = culture grown at 37°C; B (—▲—) = grown at 32°C; C (—●—) = grown at 28°C. Agitation held at 200rpm. Induction of cultures grown at 37°C with a final concentration of 1mM IPTG occurred after 360 minutes (blanked $A_{600}$ of approximately 0.6); cultures grown at 28°C & 37°C were induced 480 minutes after inoculation. Error bars indicate standard error from 8 individual well-cultures.

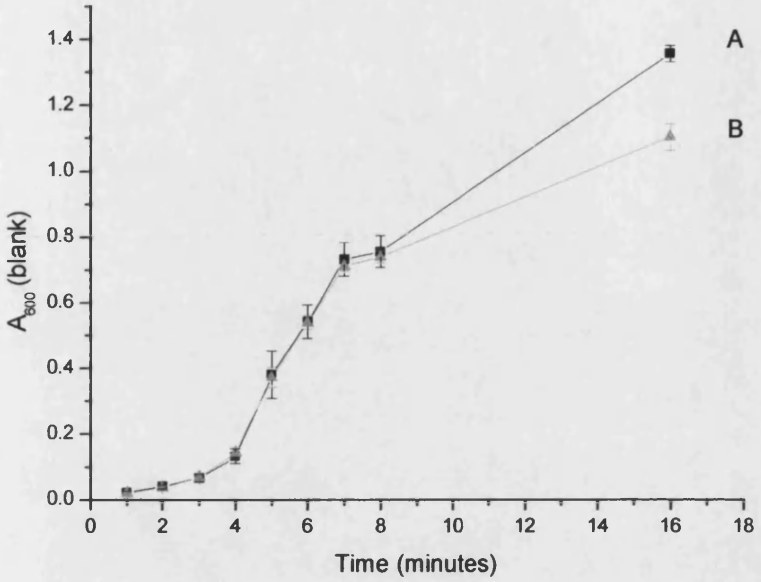



**Figure 6-12: Growth rates of the 5 selected variants in 5ml LB cultures @ 28°C, 200rpm**
Key: A (—■—) = isolates G1 (induced & uninduced), D1 (induced & uninduced); B (—▲—) = isolates D2 (induced & uninduced), D4 (induced & uninduced), C4 (induced & uninduced). Blanked against uninfected media. Isolates grouped for clarity, error bars refer to standard deviation between them.

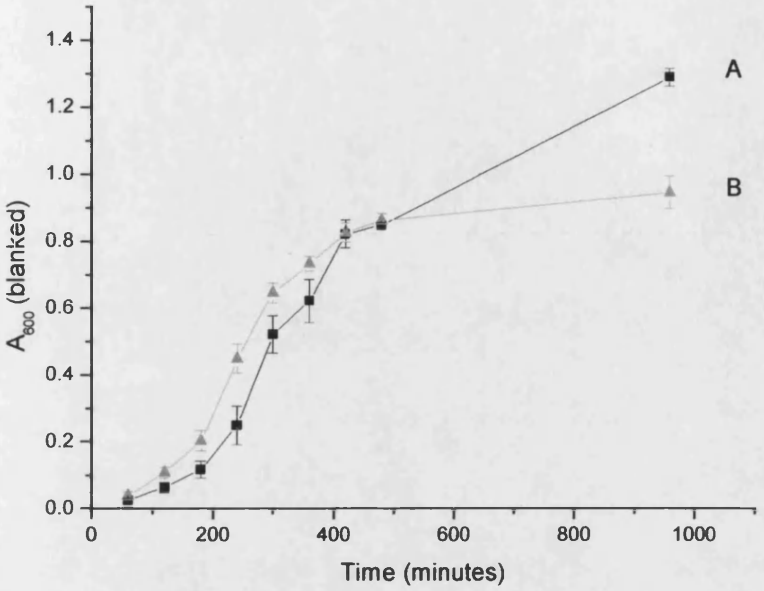



**Figure 6-13: Growth rates of the 5 selected variants in 5ml LB cultures @ 32°C, 200rpm**

Key: A (—■—) = isolates G1 (induced & uninduced), D1 (induced & uninduced); B (—▲—) = isolates D2 (induced & uninduced), D4 (induced & uninduced), C4 (induced & uninduced). Blanked against uninfected media. Isolates grouped for clarity, error bars refer to standard deviation between them.

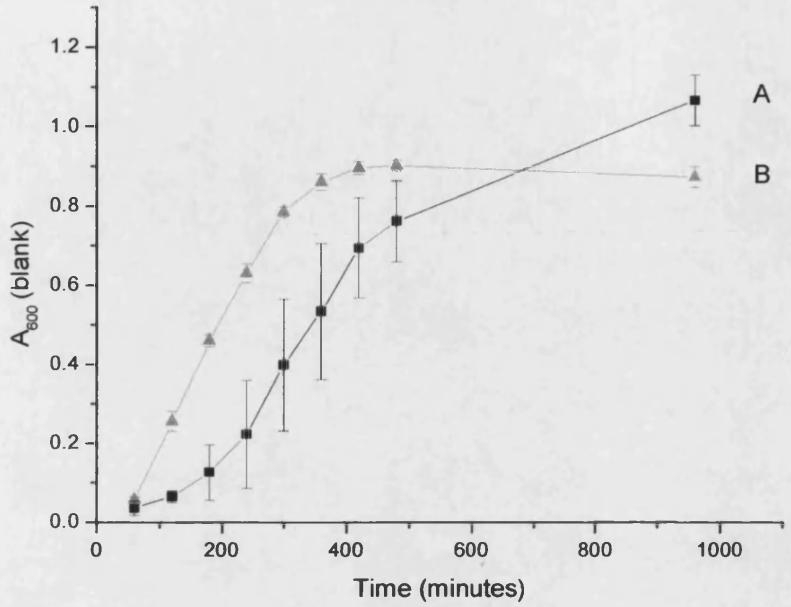



**Figure 6-14: Growth rates of the 5 selected variants in 5ml LB cultures @ 37°C, 200rpm**

Key: A (—■—) = isolates G1 (induced & uninduced), D1 (induced & uninduced); B (—▲—) = isolates D2 (induced & uninduced), D4 (induced & uninduced), C4 (induced & uninduced). Blanked against uninfected media. Isolates grouped for clarity, error bars refer to standard deviation between them.

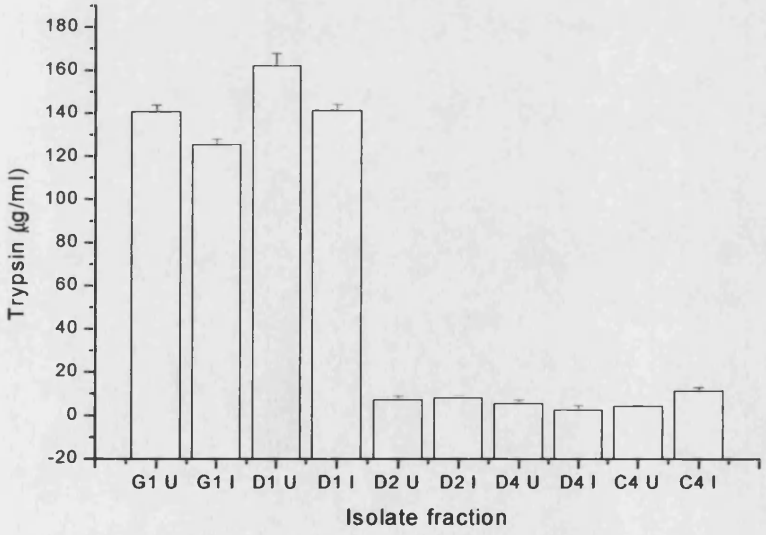



**Figure 6-15: Activity of 5ml cultures grown in LB @ 28°C, 200rpm**

The terms U & I refer to uninduced and induced cultures respectively.

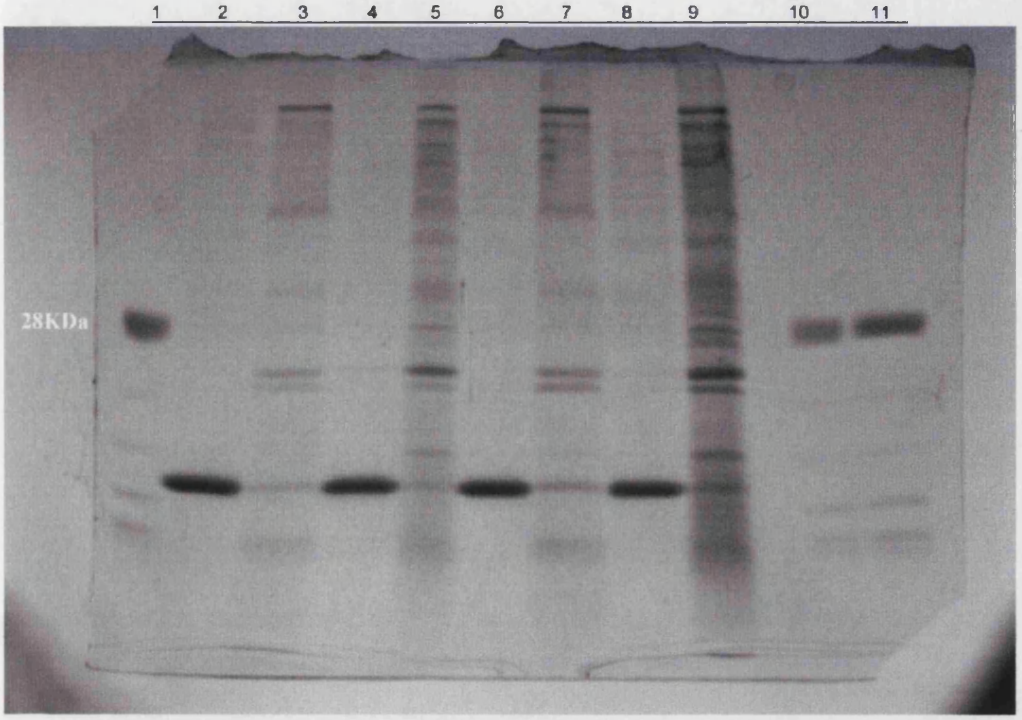



**Figure 6-16: G1 fractions, 5ml, 28°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant G1 during the course of a 5ml fermentation in LB, held at 28°C, 200rpm. Lanes 2-5 contain culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 hold culture which was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – 7.5μg commercial trypsin; 11 - 15μg commercial trypsin.

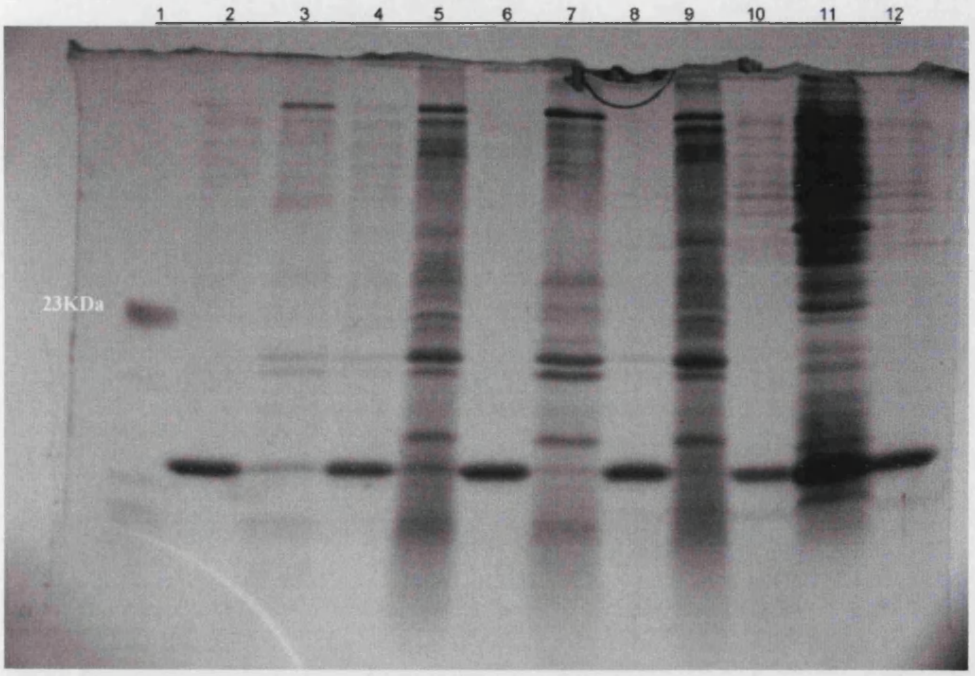



**Figure 6-17: D1 & C4 fractions, 5ml, 28°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variants D1&C4 during the course of a 5ml fermentation in LB, held at 28°C, 200rpm. Lanes 2-9 hold samples from D1, lanes 9-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) soluble fraction of culture, 3hrs after induction; 11 – (C4) insoluble fraction of culture, 3hrs after induction; 12 – (C4) soluble fraction of culture, 3hrs after induction of test culture.

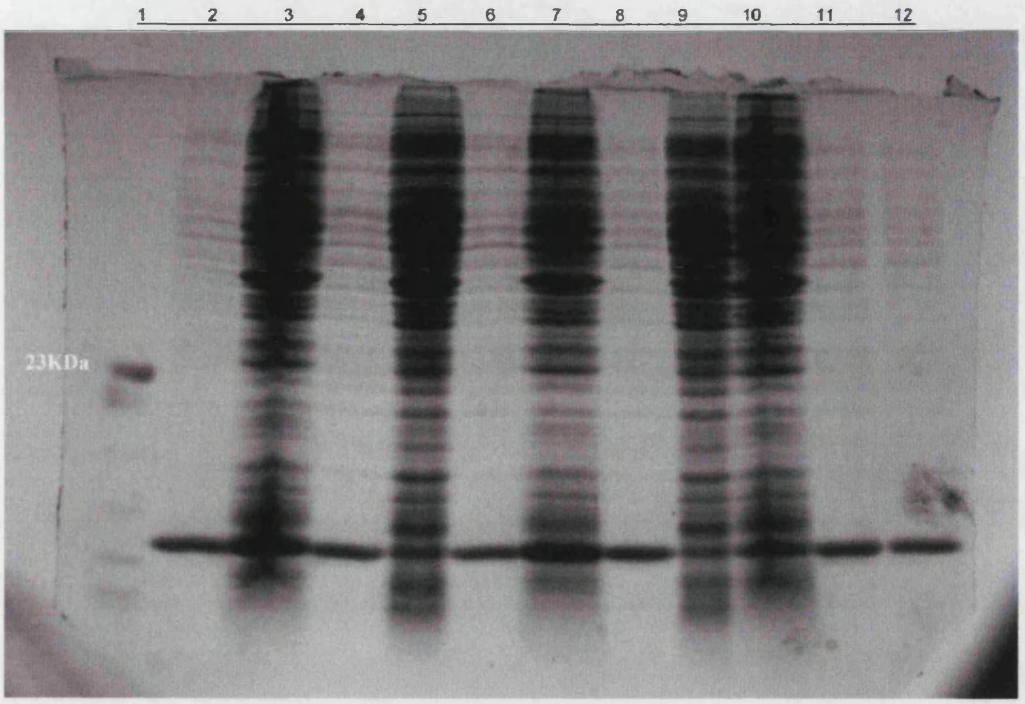



**Figure 6-18: D2 & C4 fractions, 5ml, 28°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variants D2 & C4 during the course of a 5ml fermentation in LB, held at 28°C, 200rpm. Lanes 2-9 hold samples from D2, lanes 9-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) insoluble fraction of culture, 3hrs after induction of test culture; 11 – (C4) soluble fraction of culture, 16hrs after induction; 12 – (C4) soluble fraction of culture, 16hrs after induction of test culture.

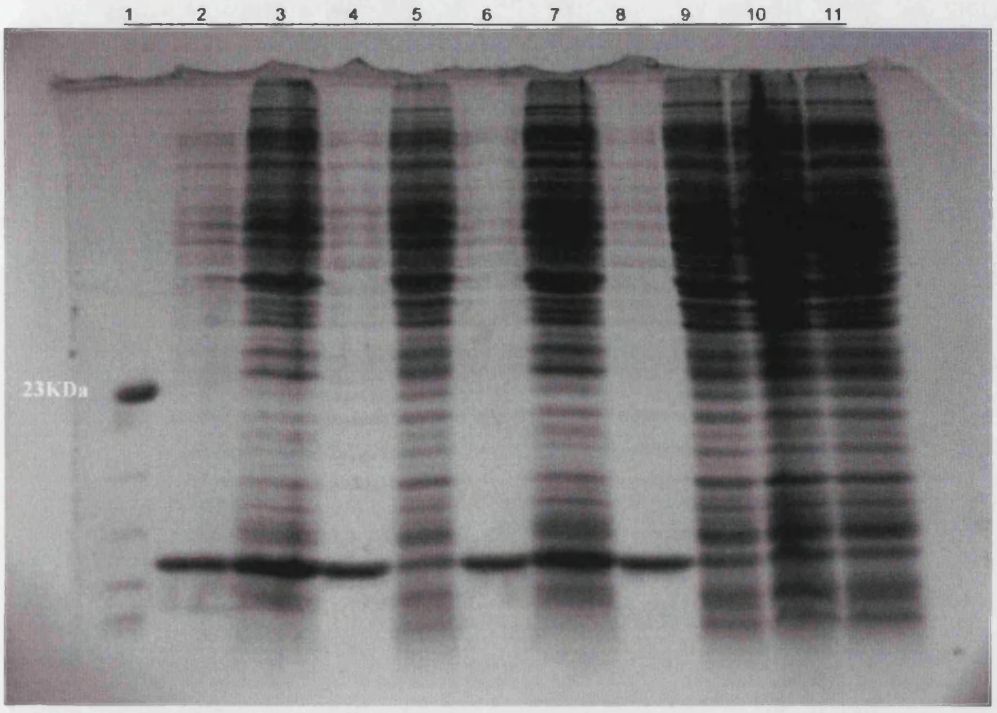



**Figure 6-19: D4 & C4 fractions, 5ml, 28°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D4 & C4 during the course of a 5ml fermentation in LB, held at 28°C, 200rpm. Lanes 2-9 hold samples from D4, lanes 9-11 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) insoluble fraction of culture, 16hrs after induction; 11 – (C4) insoluble fraction of culture, 16hrs after induction of test culture.

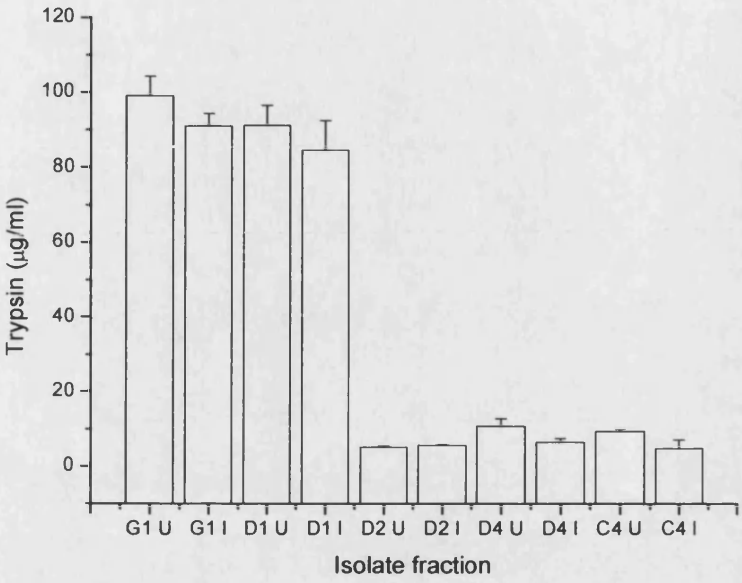



**Figure 6-20: Activity of 5ml cultures grown in LB @ 32°C, 200rpm**

The terms U & I refer to uninduced and induced cultures respectively.

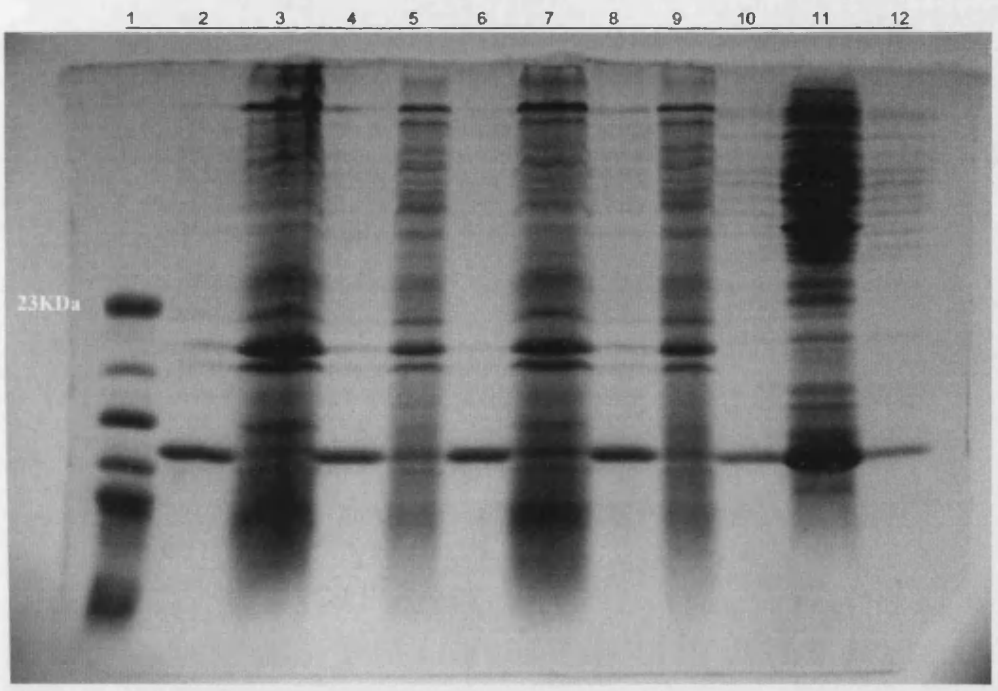



**Figure 6-21: G1 & C4 fractions, 5ml, 32°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant G1 & C4 during the course of a 5ml fermentation in LB, held at 32°C, 200rpm. Lanes 2-9 hold samples from G1, lanes 9-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) soluble fraction of culture, 3hrs after induction; 11 – (C4) insoluble fraction of culture, 3hrs after induction; 12 – (C4) soluble fraction of culture, 3hrs after induction of test culture.

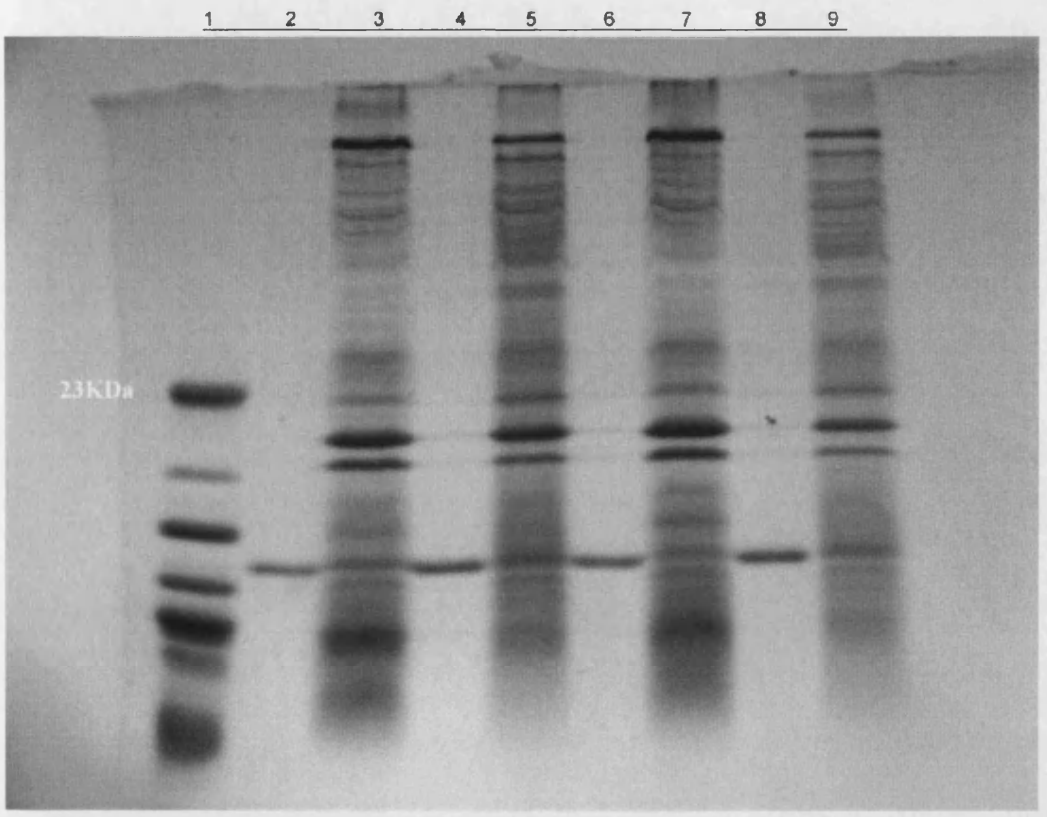



**Figure 6-22: D1 fractions, 5ml, 32°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D1 during the course of a 5ml fermentation in LB, held at 32°C, 200rpm. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture.

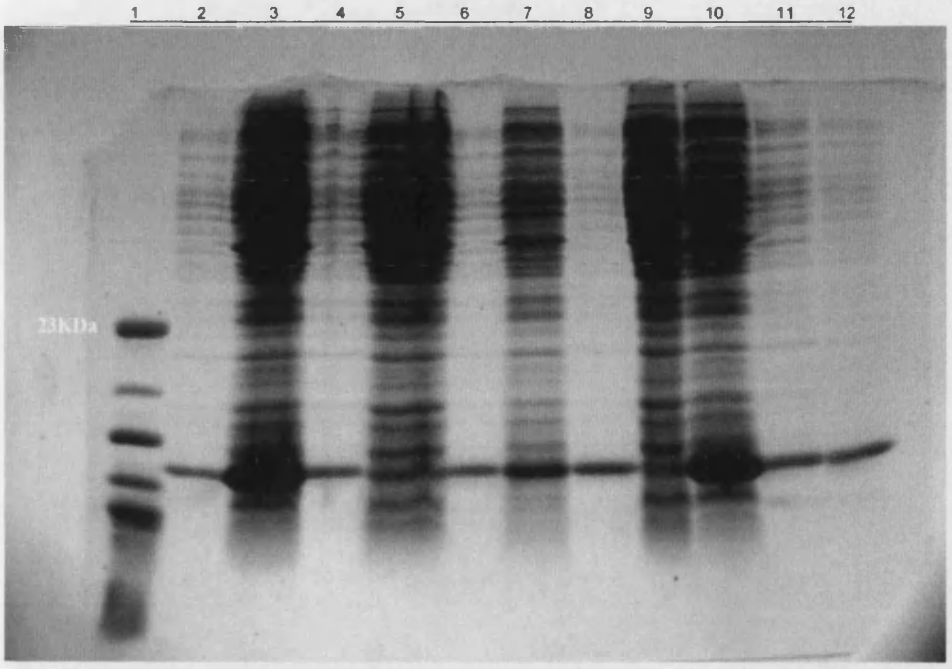



**Figure 6-23: D2 & C4 fractions, 5ml, 32°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D2 & C4 during the course of a 5ml fermentation in LB, held at 32°C, 200rpm. Lanes 2-9 hold samples from D2, lanes 9-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) insoluble fraction of culture, 3hrs after induction of test culture; 11 – (C4) soluble fraction of culture, 16hrs after induction; 12 – (C4) soluble fraction of culture, 16hrs after induction of test culture.

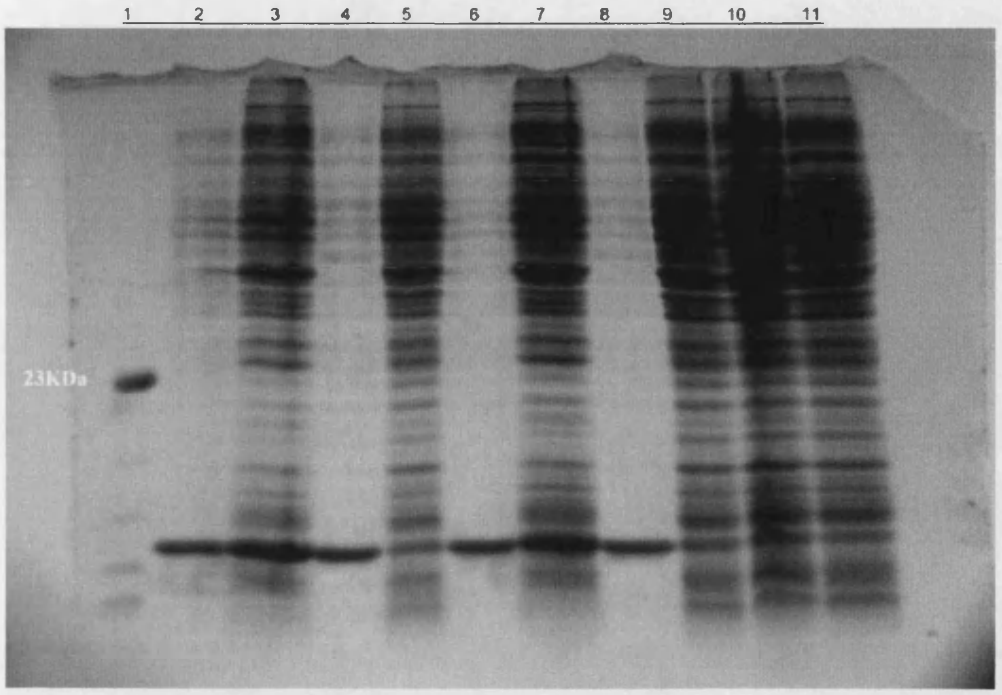



**Figure 6-24: D4 & C4 fractions, 5ml, 32°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D4 & C4 during the course of a 5ml fermentation in LB, held at 32°C, 200rpm. Lanes 2-9 hold samples from D4, lanes 9-11 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) insoluble fraction of culture, 16hrs after induction; 11 – (C4) insoluble fraction of culture, 16hrs after induction of test culture.

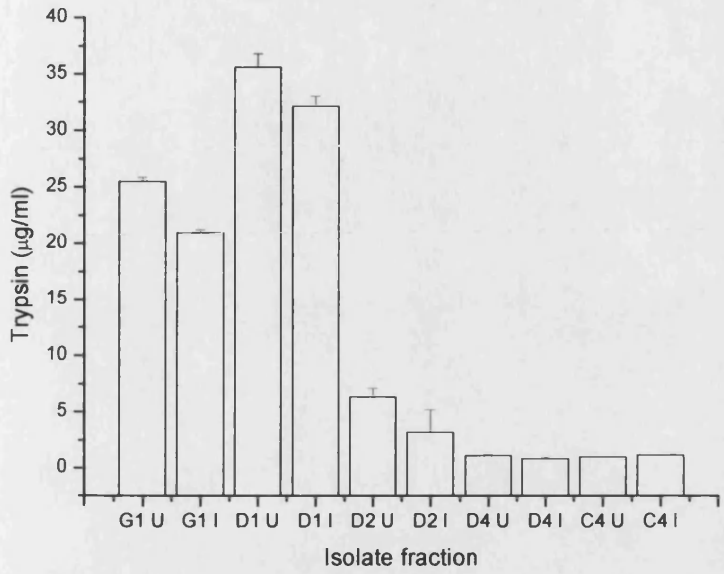
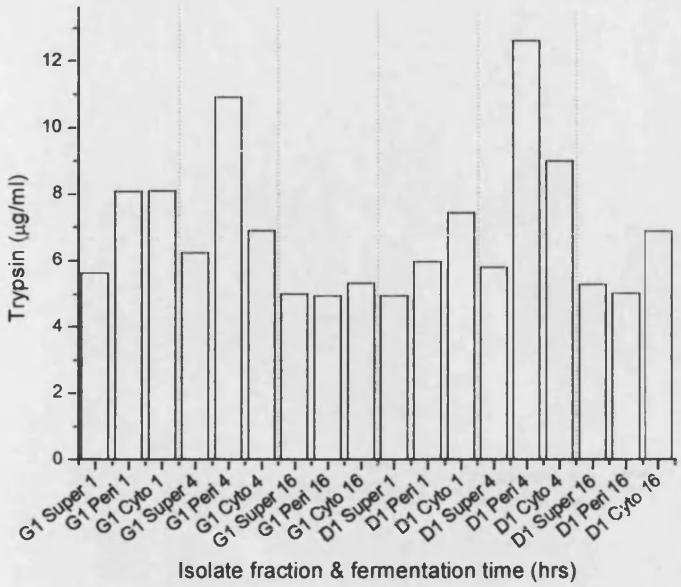



**Figure 6-25: Activity of 5ml cultures grown in LB @ 37°C, 200rpm**

The terms U & I refer to uninduced and induced cultures respectively.

Figure 6-25 illustrates the total soluble enzyme activity obtained from each of the five isolates. In order to determine the location of the active enzyme within the cell, all soluble fractions (supernatant, periplasm & cytoplasm) were then purified and assayed. Results shown in Figure 6-26.



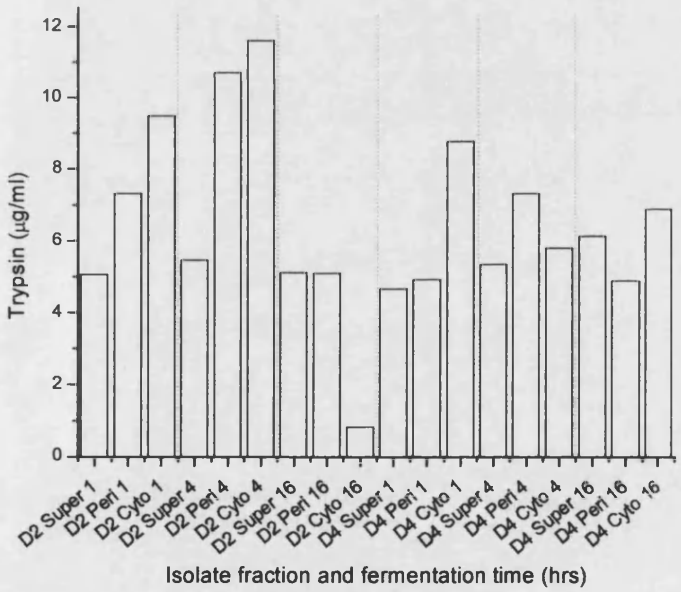




**Figure 6-26: Activity of fractionated 5ml culture grown in LB @ 37°C, 200rpm**

All results blanked against the activity of the appropriate fraction from an uninduced culture. Fractions tested were: supernatant (Super); periplasm (Peri); cytoplasm (Cyto). The terms 1, 4 & ON (overnight) refer to cultures tested 1, 4 & 16 hours after induction respectively. Results were

not blanked against uninduced culture as such cultures exhibited almost identical levels of activity and thus could not be considered background.

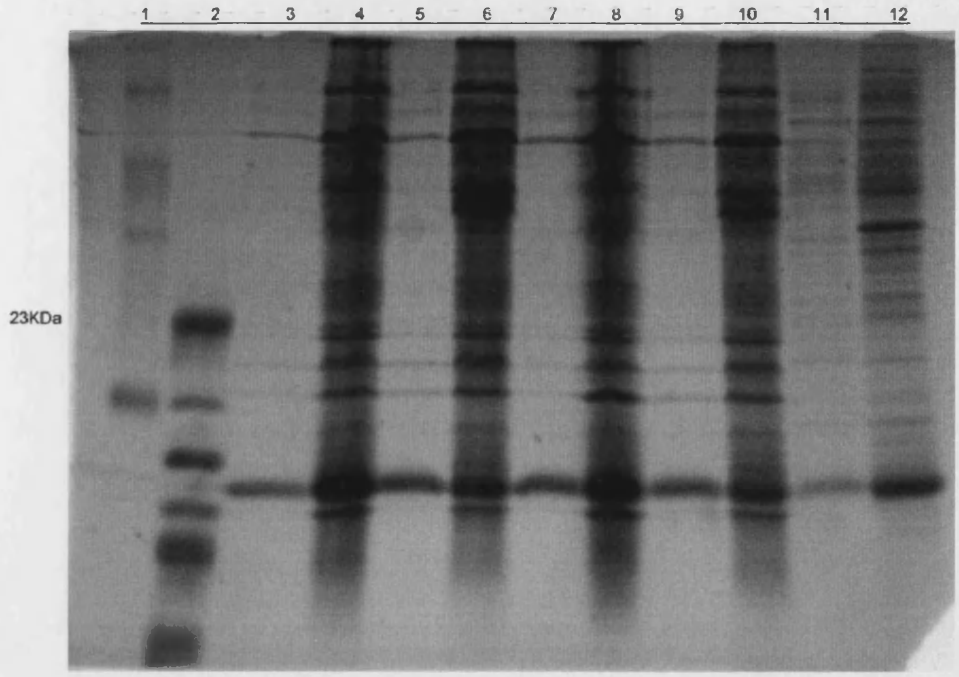



**Figure 6-27: G1 & C4 fractions, 5ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant G1 & C4 during the course of a 5ml fermentation in LB, held at 37°C, 200rpm. Lanes 3-10 hold samples from G1, lanes 11-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – commercial protein ladder A; 2 – 7.5μg of commercial trypsin; 3 – soluble fraction of culture, 3hrs after induction; 4 – insoluble fraction of culture, 3hrs after induction; 5 – soluble fraction of culture, 16hrs after induction; 6 – insoluble fraction of culture, 16hrs after induction; 7 – soluble fraction of culture, 3hrs after induction of test cultures; 8 – insoluble fraction of culture, 3hrs after induction of test culture; 9 – soluble fraction of culture, 16hrs after induction of test culture; 10 – insoluble fraction of culture, 16hrs after induction of test culture; 11 – (C4) soluble fraction of culture, 3hrs after induction; 12 – (C4) insoluble fraction of culture, 3hrs after induction.

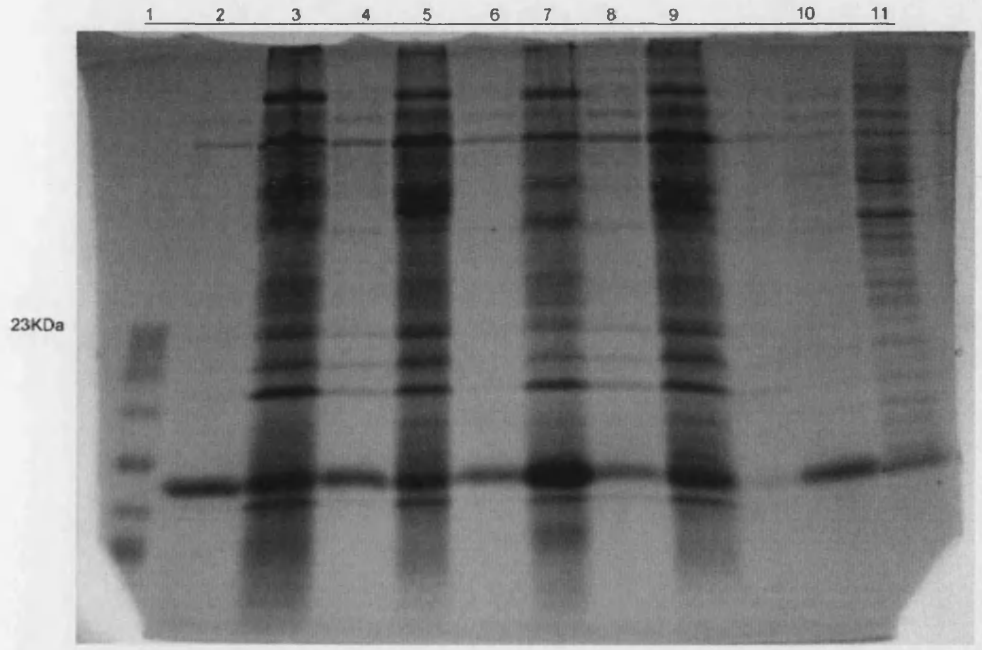



**Figure 6-28: D1 & C4 fractions, 5ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D1 & C4 during the course of a 5ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-9 hold samples from D1, lanes 9-11 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) soluble fraction of culture, 3hrs after induction of test culture; 11 – (C4) insoluble fraction of culture, 3hrs after induction of test culture.

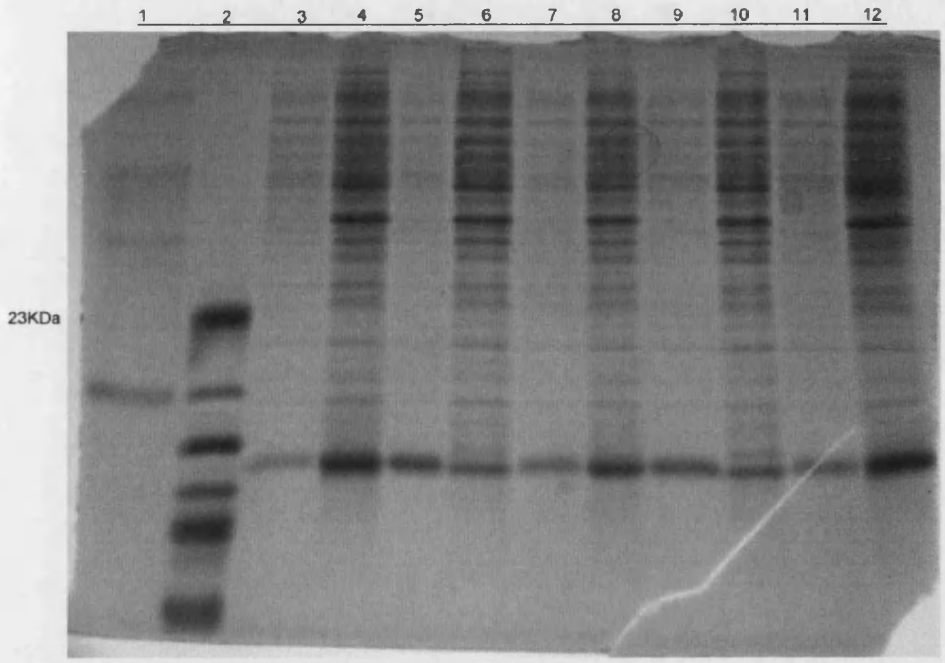



**Figure 6-29: D2 & C4 fractions, 5ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D2 & C4 during the course of a 5ml fermentation in LB, held at 37°C, 200rpm. . Lanes 3-10 hold samples from D2, lanes 11-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – commercial protein ladder A; 2 – 7.5μg of commercial trypsin; 3 – soluble fraction of culture, 3hrs after induction; 4 – insoluble fraction of culture, 3hrs after induction; 5 – soluble fraction of culture, 16hrs after induction; 6 – insoluble fraction of culture, 16hrs after induction; 7 – soluble fraction of culture, 3hrs after induction of test cultures; 8 – insoluble fraction of culture, 3hrs after induction of test culture; 9 – soluble fraction of culture, 16hrs after induction of test culture; 10 – insoluble fraction of culture, 16hrs after induction of test culture; 11 – (C4) soluble fraction of culture, 16hrs after induction; 12 – (C4) insoluble fraction of culture, 16hrs after induction.

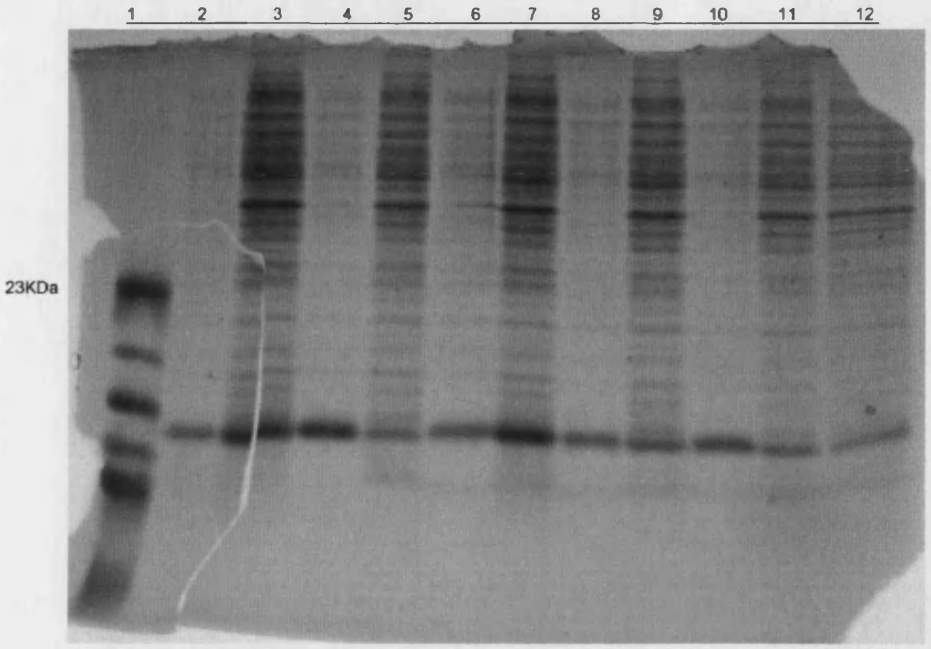



**Figure 6-30: D4 fractions, 5ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D4 during the course of a 5ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-9 hold samples from D4, lanes 9-12 contain samples from C4. Lanes 2-5 hold culture that was induced with IPTG to a final conc. of 1mM, lanes 6-9 contain culture that was not induced. Lanes: 1 – 7.5μg of commercial trypsin; 2 – soluble fraction of culture, 3hrs after induction; 3 – insoluble fraction of culture, 3hrs after induction; 4 – soluble fraction of culture, 16hrs after induction; 5 – insoluble fraction of culture, 16hrs after induction; 6 – soluble fraction of culture, 3hrs after induction of test cultures; 7 – insoluble fraction of culture, 3hrs after induction of test culture; 8 – soluble fraction of culture, 16hrs after induction of test culture; 9 – insoluble fraction of culture, 16hrs after induction of test culture; 10 – (C4) soluble fraction of culture, 16hrs after induction of test culture; 11 – (C4) insoluble fraction of culture, 16hrs after induction of test culture; 12 – (C4) repeat of lane 11.

#### 6.3.3.3 Expression profile & activity of library on L-BAPNA after 100ml fermentation

In order to expand upon the results obtained from the 5ml fermentation of the 5 selected isolates, an attempt was made to scale the fermentations up to 100ml. The obvious advantage gained from such a move would be the increase in material available for protein analysis and the increased

ease by which the growth profile of a larger culture could be measured (i.e. the possibility of on-line turbidity or dissolved oxygen measurements). Such refinements could prove crucial if plasmid stability or recombinant protein autotoxicity were indeed proving to be a problem, requiring control of the fermentation. Once again a profile of trypsin expression levels and activity in LB at 37°C, 200rpm over the course of a 100ml fermentation was built by means of growth plots, L-BAPNA activity testing of cellular fractions & SDS-PAGE gels.

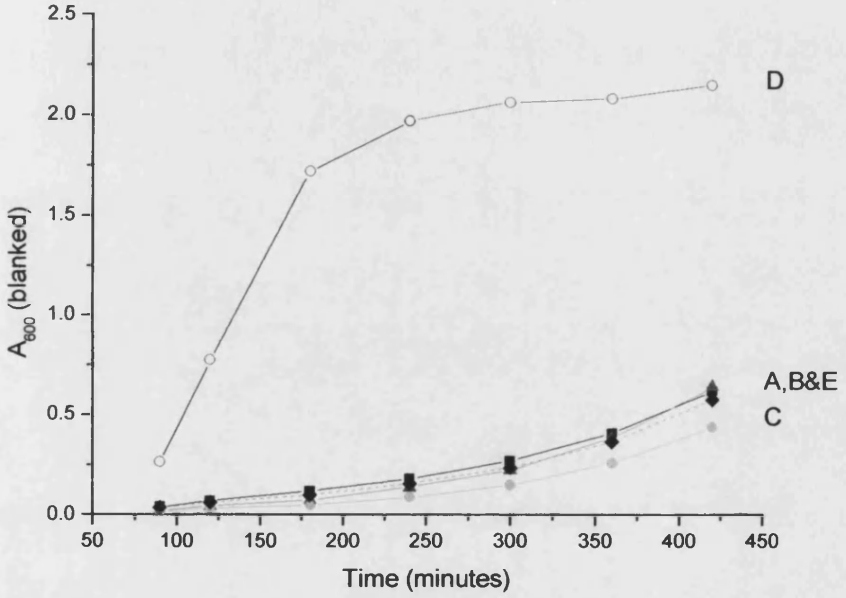



**Figure 6-31: Growth rates of the 5 selected variants in 100ml LB cultures @ 37°C, 200rpm**
Key: A (–■–) = isolate G1; B (–▲–) = D1; C (–●–) = D2; D (–○–) = D4; E (··◆··) = C4. Blanked against uninfected media.

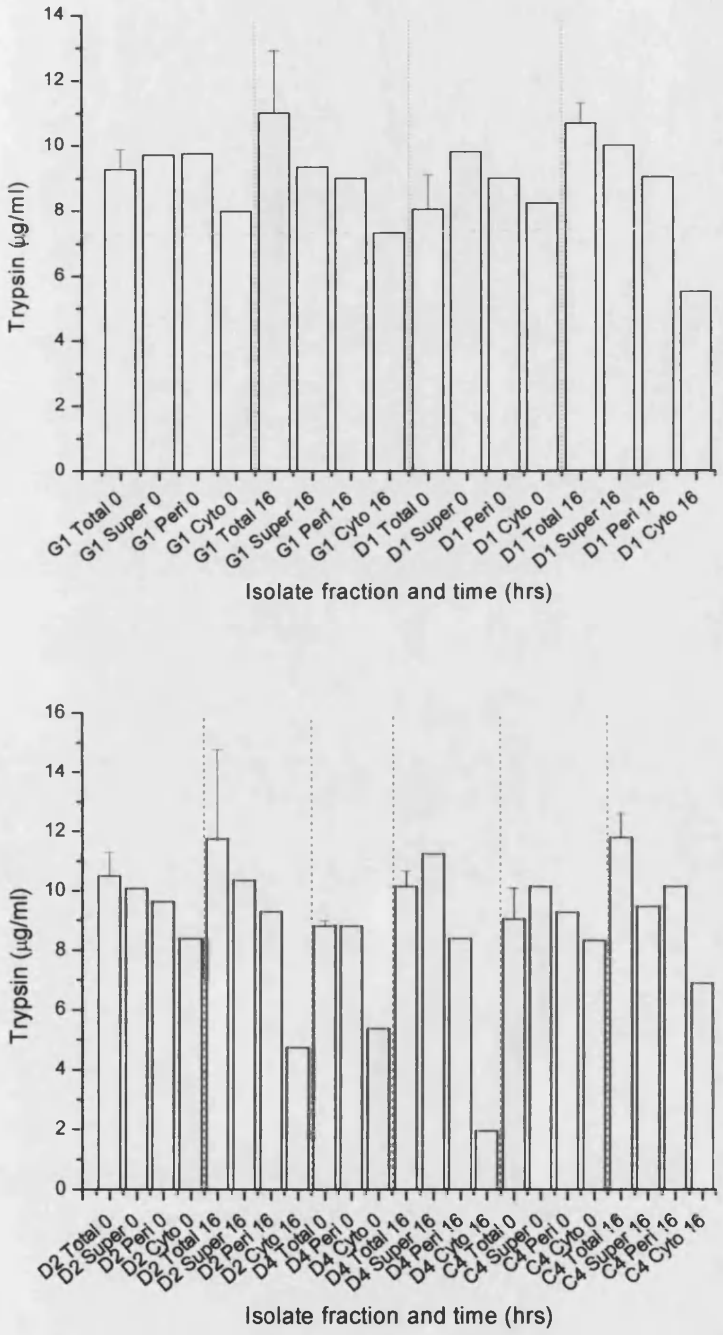



**Figure 6-32: Activity of fractionated 100ml cultures grown in LB @ 37°C, 200rpm**

Fractions tested were: supernatant (Super); periplasm (Peri); cytoplasm (Cyto). All culture tested was harvested 16hrs after induction.

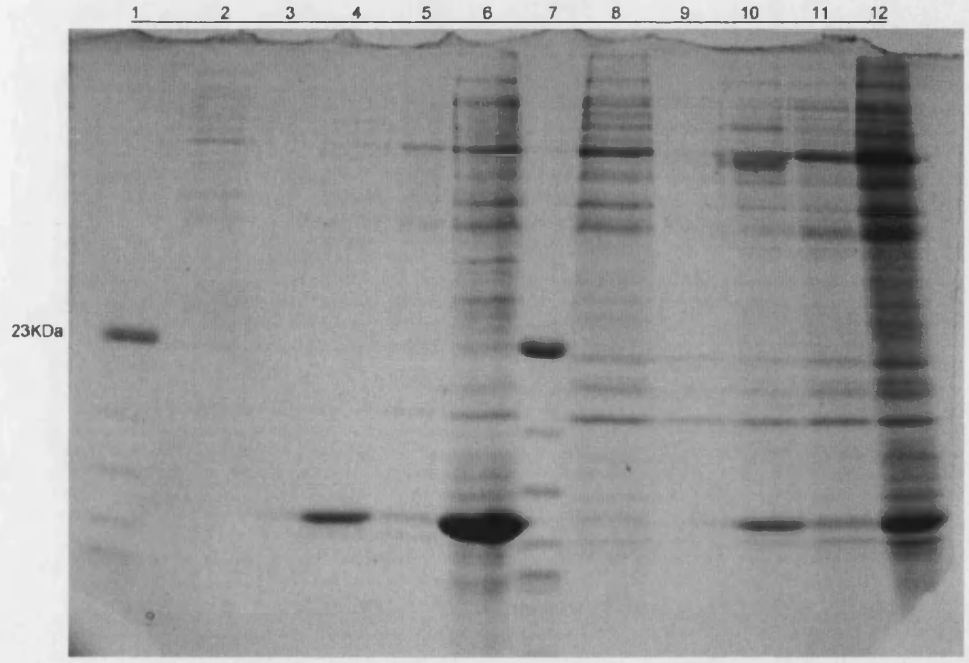



**Figure 6-33: G1 fractions, 100ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant G1 during the course of a 100ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-6 hold uninduced culture, lanes 8-12 contain induced culture. Lanes: 1 – 7.5μg commercial trypsin; 2 – total protein from culture, 3hrs after induction of test culture; 3 – supernatant from culture, 3hrs after induction of test culture; 4 – periplasmic fraction, 3hrs after induction of test culture; 5 – cytoplasmic fraction, 3hrs after induction of test culture; 6 – insoluble fraction, 3hrs after induction of test culture; 7 – 7.5μg commercial trypsin, 8 – total protein from culture, 16hrs after induction; 9 – supernatant from culture, 16hrs after induction; 10 – periplasmic fraction, 16hrs after induction; 11 – cytosolic fraction, 16hrs after induction; 12 – insoluble fraction, 16hrs after induction.

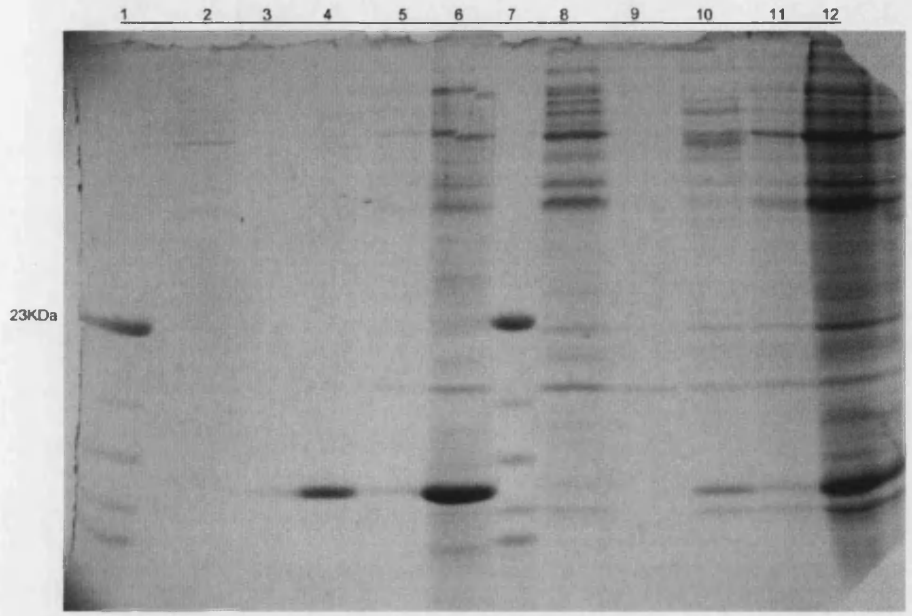



**Figure 6-34: D1 fractions, 100ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D1 during the course of a 100ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-6 hold uninduced culture, lanes 8-12 contain induced culture. Lanes: 1 – 7.5μg commercial trypsin; 2 – total protein from culture, 3hrs after induction of test culture; 3 – supernatant from culture, 3hrs after induction of test culture; 4 – periplasmic fraction, 3hrs after induction of test culture; 5 – cytoplasmic fraction, 3hrs after induction of test culture; 6 – insoluble fraction, 3hrs after induction of test culture; 7 – 7.5μg commercial trypsin, 8 – total protein from culture, 16hrs after induction; 9 – supernatant from culture, 16hrs after induction; 10 – periplasmic fraction, 16hrs after induction; 11 – cytosolic fraction, 16hrs after induction; 12 – insoluble fraction, 16hrs after induction.

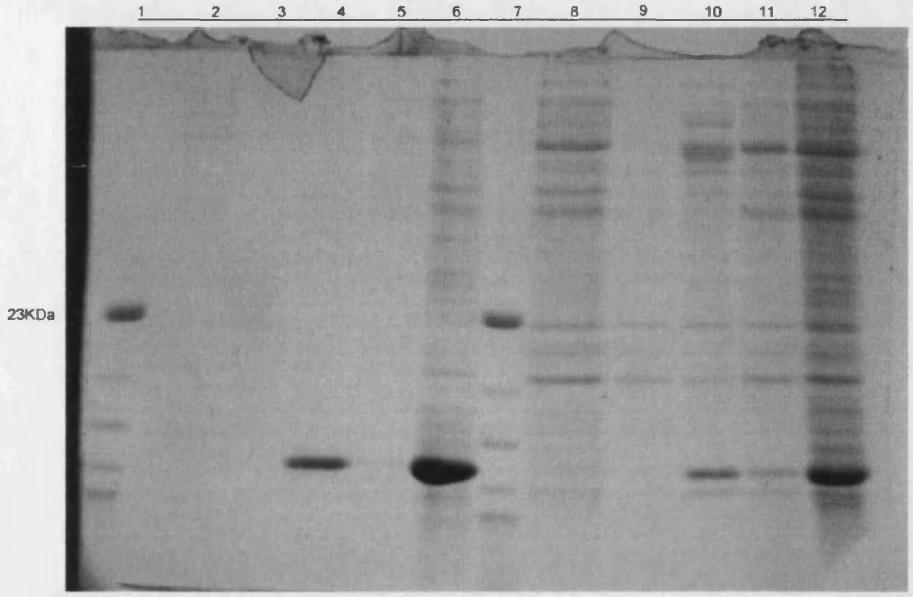



**Figure 6-35: D2 fractions, 100ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D2 during the course of a 100ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-6 hold uninduced culture, lanes 8-12 contain induced culture. Lanes: 1 – 7.5μg commercial trypsin; 2 – total protein from culture, 3hrs after induction of test culture; 3 – supernatant from culture, 3hrs after induction of test culture; 4 – periplasmic fraction, 3hrs after induction of test culture; 5 – cytoplasmic fraction, 3hrs after induction of test culture; 6 – insoluble fraction, 3hrs after induction of test culture; 7 – 7.5μg commercial trypsin, 8 – total protein from culture, 16hrs after induction; 9 – supernatant from culture, 16hrs after induction; 10 – periplasmic fraction, 16hrs after induction; 11 – cytosolic fraction, 16hrs after induction; 12 – insoluble fraction, 16hrs after induction.

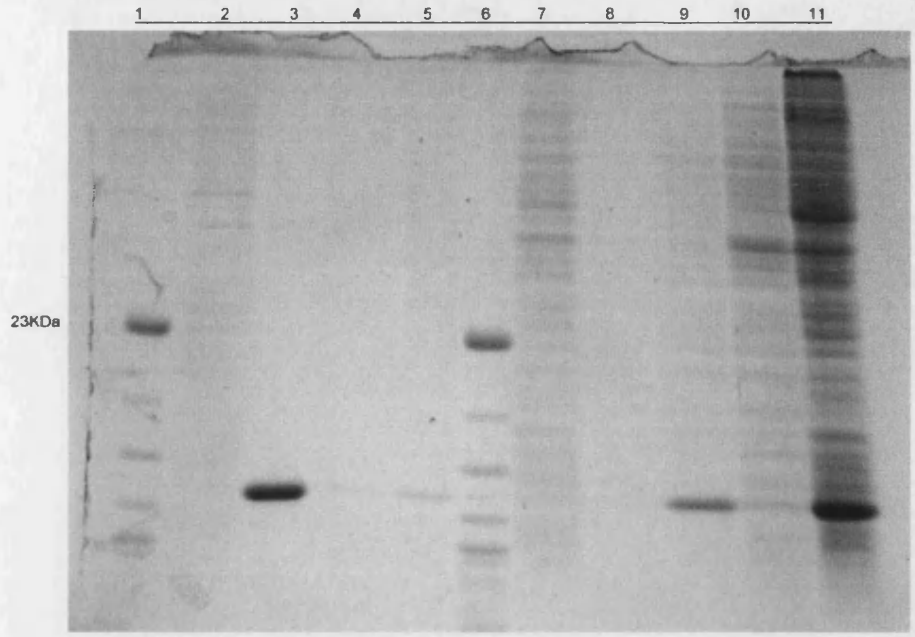



**Figure 6-36: D4 fractions, 100ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant D4 during the course of a 100ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-6 hold uninduced culture, lanes 8-12 contain induced culture. Lanes: 1 – 7.5μg commercial trypsin; 2 – total protein from culture, 3hrs after induction of test culture; 3 – supernatant from culture, 3hrs after induction of test culture; 4 – periplasmic fraction, 3hrs after induction of test culture; 5 – cytoplasmic fraction, 3hrs after induction of test culture; 6 – insoluble fraction, 3hrs after induction of test culture; 7 – 7.5μg commercial trypsin, 8 – total protein from culture, 16hrs after induction; 9 – supernatant from culture, 16hrs after induction; 10 – periplasmic fraction, 16hrs after induction; 11 – cytosolic fraction, 16hrs after induction; 12 – insoluble fraction, 16hrs after induction.

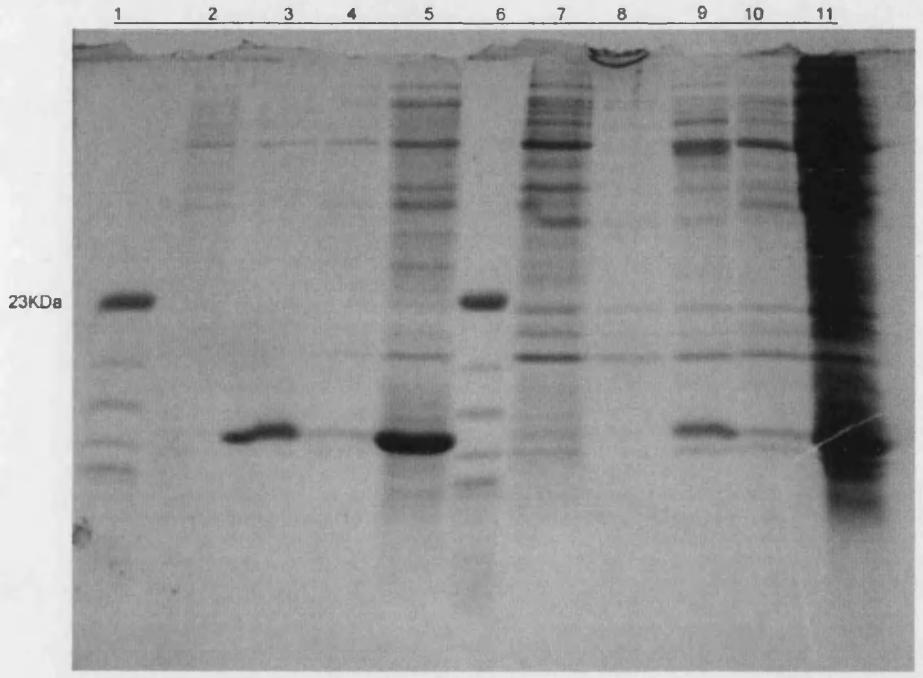



**Figure 6-37: C4 fractions, 100ml, 37°C**

Soluble & insoluble fractions of BL21(DE3):pET (T) variant C4 during the course of a 100ml fermentation in LB, held at 37°C, 200rpm. Lanes 2-6 hold uninduced culture, lanes 8-12 contain induced culture. Lanes: 1 – 7.5μg commercial trypsin; 2 – total protein from culture, 3hrs after induction of test culture; 3 – supernatant from culture, 3hrs after induction of test culture; 4 – periplasmic fraction, 3hrs after induction of test culture; 5 – cytoplasmic fraction, 3hrs after induction of test culture; 6 – insoluble fraction, 3hrs after induction of test culture; 7 – 7.5μg commercial trypsin, 8 – total protein from culture, 16hrs after induction; 9 – supernatant from culture, 16hrs after induction; 10 – periplasmic fraction, 16hrs after induction; 11 – cytosolic fraction, 16hrs after induction; 12 – insoluble fraction, 16hrs after induction.

### 6.3.4 Sequencing of variants

The *rtrypsin* gene in all variants was sequenced using universal T7 promoter & terminator primers and the region immediately upstream was sequenced using an oligonucleotide complementary to the terminal portion of the pelB leader sequence (oligo 6 from Figure 5-2). The sequencing failed or produced poor results on the first four attempts with the apparent cause being poor plasmid quality.

Variants D1, D2, D4 & C4 were eventually resolved but revealed no mutations from the original cloned pET(T) construct, either within the coding region or in the control elements immediately upstream. Variant G1 continually failed to be sequenced, despite the isolation of mini-prep quality plasmid and the successful optimization of PCR conditions by this researcher.

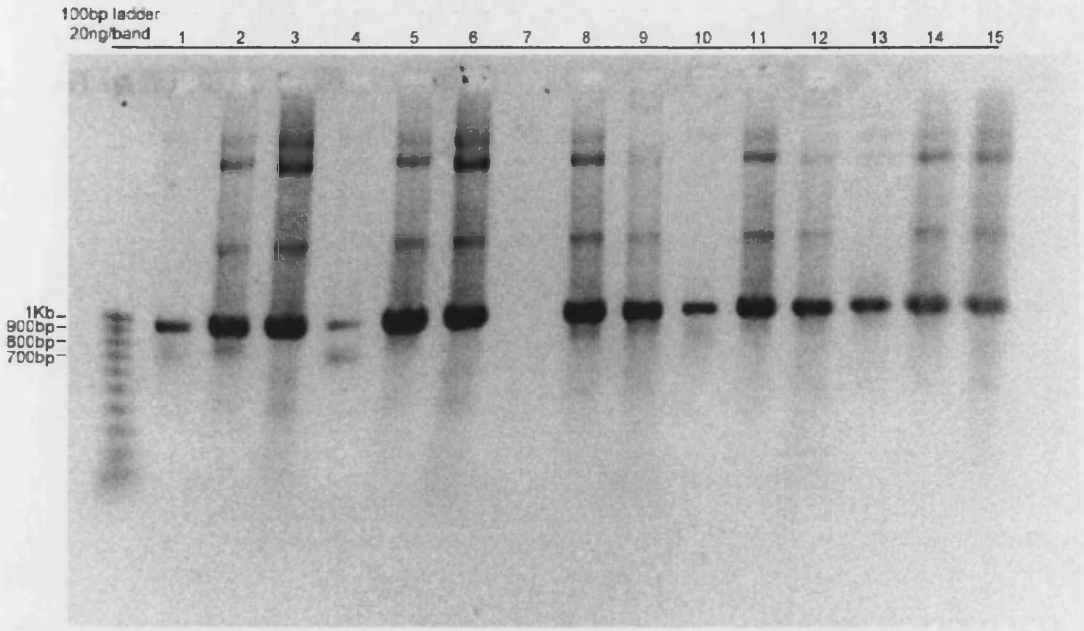



**Figure 6-38: PCR of the 5 *r-trypsin* variants in conditions of varying [$MgCl_2$]**

120ng of Plasmid pET(T) was amplified with T7 promoter & terminator oligonucleotides: expected product = 980 nucleotides. 1.2% agarose gel run at 50V, 80mA for 1hr. Lanes: 1 - G1 with 0.8mM $MgCl_2$; 2 – G1 with 1.5mM $MgCl_2$; 3 – G1 with 2.8mM $MgCl_2$; 4 – D1 with 0.8 mM $MgCl_2$; 5 - D1 with 1.2mM $MgCl_2$; 6 – D1 with 2.8mM $MgCl_2$; 7 – D2 with 0.8mM $MgCl_2$; 8 – D2 with 1.5mM $MgCl_2$; 9 – D2 with 2.8mM $MgCl_2$; 10 – D4 with 0.8mM $MgCl_2$; 11 – D4 with 1.5mM $MgCl_2$; 12 – D4 with 2.8mM $MgCl_2$; 13 – C4 with 0.8mM $MgCl_2$; 14 – C4 with 1.5mM $MgCl_2$; 15 – C4 with 2.8mM $MgCl_2$.

### 6.3.5 Testing of mutation rate of XL1-Red

In order to calibrate the mutation rate generated by XL1-Red, pET(T) plasmid was propagated through increasing generations of growth (growth and subculturing protocols in Section 6.2.3). Sequencing was performed using universal T7 primers and results are shown in Table 6-2.

| XL1-Red:pET(T) Generation | Mutations |
|---|---|
| G40A, B, C | None within the sequenced range. |
| G50A | None within the sequenced range. |
| G50B | Possible single 'T' insertion @ nt position 943. |
| G50C | Possible substitution 'A→G' @ nt 15 |
| G60A | None within the sequenced range. |
| G60B | Possible substitution 'T→C' @ nt 325. |
| G60C | None within the sequenced range. |

**Table 6-2: XL1-Red induced mutations**

Showing results from the XL1-Red:pET (T) growth experiment, listing mutations accumulated in the region between the T7 promoter & terminator (incl. *r-typsin*) over the course of 60 generations. The test region sequenced was 1006nt in length and was sequenced with in both directions, but the limitations of process meant that good resolution was only achieved over an average of 850nts. Total size of pET(T) = 6031bp. Plasmid from three random isolates was purified & sequenced after 40, 50 & 60 generations. Possible mutations refer to those detected at a non-overlapping end of the sequence trace and thus may be signal artefacts.

## 6.4 Discussion

### 6.4.2 Microwell fermentations of un-mutated BL21(DE3):pET (T)

Comparison of the growth of BL21(DE3):pET (T) in 1ml of the two media types (LB & M9SD) over the course of a 20hr fermentation @ 37°C, 900rpm in a □-plate showed that only LB was capable

of supporting growth at an acceptable rate (see Figure 6-1). In LB, the induction point of $A_{600}$ 0.6-0.7 (blanked against uninfected medium) was reached after 4hrs.

Activity testing of the cellular fractions was performed on samples taken 4hrs & 16hrs after induction and blanked against those taken from uninduced □-plate cultures of a similar age (see Figure 6-2,Figure 6-3). The 4hr results yielded approximately 0.7μg/ml trypsin from the soluble fraction as a whole, based on an assay involving the simple lysis of the cell, while the periplasmic fraction yielded 0.51μg/ml and the unlysed cell culture (supernatant plus cells) gave 0.76μg/ml. Work with low levels of urea appeared to give an elevated signal, possibly reflecting the release of r-trypsin from the insoluble fraction, but this could not be confirmed. 16hr results (Figure 6-4) showed an elevated basic activity of the unlysed cell culture to 1.9μg/ml and a similar rise in periplasmic activity to 1.86μg, but contrasted this with a severe drop in the activity of the total soluble fraction as opposed to that of the uninduced control.

The SDS-PAGE of the soluble and insoluble fractions from the deep-well fermentations of both media (Figure 6-4) clearly showed protein bands of approximately the correct molecular weight in the soluble & insoluble fraction, but, without an appreciable degree of overexpression or a activity-independent specific concentration assay (e.g. a dual layer ELISA), the identity of these bands could not be clarified.

### 6.4.3 Activity of library on L-BAPNA after microwell fermentation

The percentage of positive results from the first two plates screened from the initial XL1-Red mutant library was extremely high, with a full 47% of isolates on the 'A' plate and 27% of isolates on the 'C' plate exhibiting higher activity levels than the 'wt' BL21(DE3):pET (T) after a 20hr □-plate fermentation. The apparent statistical impossibility of such a result is mollified by the fact that a significant proportion of results were false positives arising form the open deep-well fermentation. From the isolates selected for further testing in tables 6.5 & 6.6, (A)G3, (A)D7 & (A)E9 were clearly false positives with a number of other isolates exhibiting only slight increases over wild-type activity. The still high numbers of positive isolates would appear to indicate that a significant mutation occurred within one of the first few generations of XL1-Red growth and was then propagated through the subsequent generations, eventually being represented in a high fraction of the library.

### 6.4.4 Expression profile & activity of library on L-BAPNA after 5ml fermentation

Growth of the five selected isolates (G1, D1, D2, D4, C4) in 5ml LB cultures demonstrated a slightly enhanced rate of growth over that of the un-mutated strain in identical conditions (Figure 6-11). During a fermentation held at 28°C (Figure 6-12), all five isolates achieved the induction point (blanked $A_{600}$ of 6-0.8) after 420 minutes as opposed to the 480 minutes required by the wild-type. A 32°C fermentation (Figure 6-13) demonstrated the isolates reaching induction point after 360 minutes, with the wild-type again requiring 480 minutes. During the course of a 37°C fermentation (Figure 6-14), both the variants and the wild type achieved the required density after 360 minutes. After induction at 360 minutes during the 37°C fermentation, the wild-type appeared to suffer a transient drop in cell growth rate & density, suggesting that induction placed a metabolic load upon the culture. No similar effect was visible in the growth kinetics of the mutant variants.

Activity testing of the cellular fractions from a 5ml culture held at 37°C (see Figure 6-26) appeared to reveal a particular pattern, namely that the concentration of r-trypsin in the periplasm 1 hour after induction was approximately 2-4 fold that found in the supernatant, while the concentration in the cytoplasm read as less than that in the uninduced cultures. These levels, while remaining constant relative to each other, increased 3-4 fold in the fractions harvested 4 hours after induction. The fractions taken from the overnight cultures (approximately 16 hours after induction) were considerably lower, usually beneath the level of 1-hour readings.

The collection of whole cells from 5ml fermentations grown at 28°C, 32°C, 37°C and the subsequent testing of the total soluble protein (Figure 6-15, Figure 6-20, Figure 6-25) respectively) revealed a 30-50% drop in r-trypsin concentration with each increase in temperature. As with the previous results, there was minimal difference in protein concentration between the induced and uninduced cultures, although the levels in the uninduced culture did tend to exhibit a higher titre of approximately 1-5μg/ml.

Analysis of the SDS-PAGE gels run with extracts of the soluble & insoluble fractions of culture grown at each temperature (Figures 6.16-6.30) did not reveal any evidence of recombinant protein over-expression in either fraction of any of the isolates. There appeared to be no significant difference in the protein profile between the induced & uninduced cultures of the mutated variants and no visible relationship between expression levels and the activity readings obtained was evident.

### 6.4.5 Expression profile & activity of library on L-BAPNA after 100ml fermentation

All 5 variants performed abnormally in 100ml cultures with a severely depressed growth rate as compared to that of the wild-type strain. Variants G1, D1, D2 & C4 only achieved the cell density required for induction (blanked $A_{600}$ of 0.6-0.8) after 420 minutes (Figure 6-31), while un-mutated BL21(DE3):pET (T) reached the induction point after only 120 minutes (Figure 5.38). 16 hours after the induction of these isolates, the cultures were a muddy brown in colour, with a relatively low cell density ($A_{600}$ of 1.2-1.36) and a high degree of aggregated insoluble matter in suspension, suggesting that a considerable degree of cellular lysis had occurred. Isolate D4 grew at a comparable rate to the wild-type, yet demonstrated no transient reduction of growth rate upon induction. After 16hrs, this culture was the usual turbid, deep-yellow colour.

Activity readings of the cellular fractions after a 20-hour 100ml fermentation held at 37°C (Figure 6-32) appeared relatively similar to those recorded after the 37°C 5ml fermentations of the 5 variants, with the exception that slightly elevated levels of r-trypsin were present in the supernatant. Once again, there was no appreciable difference in the concentration of r-trypsin between the induced & uninduced cultures.

The SDS-PAGE gels of the fractionated 37°C 100ml fermentation (Figures 6.33-6.37) show fractions from cultures at the point of induction and 16 hours afterwards, but show no significant over-expression or translocation of any protein of the correct molecular weight.

### 6.4.6 General

The results show a clear but puzzling progressive drop in activity in all mutant variants since their initial isolation. The second □-plate fermentation demonstrated a 2-10 fold drop in protein concentration from the first screening. This trend was exacerbated in un-shown work that recorded subsequent □-plate fermentations tending towards almost untraceable concentrations. Scale-up of the test cultures brought further complications as cultures became increasingly prone to failure (no growth or heavily retarded), although the reduction in metabolic rate caused by holding the temperature of a 5ml fermentation down to 28°C did bring activity levels back up to half of their initially screened value. The apparent loss or suppression of enzyme after the procedure of fractionation was most likely due to the increase in dilution factor, a statement difficult to quantify as no attempt was made to estimate the volume of soluble protein in the cytoplasm and the

periplasm. A suppression of activity was observed but compensated for by the use of commercial trypsin controls in each of the buffers. Attempts to scale-up the cultures by a further factor of 20 increased their apparent instability, resulting in severely retarded growth rates and an apparent high degree of lysis in the cultures, although recombinant protein remained at low but detectable levels, with a predictably increased supernatant titre.

The five isolate tested appeared to fall into two distinct groups: G1 & D1 showing comparably high levels of activity in the 5ml low-temperature fermentations; and D2, D4 & C4 exhibiting consitently lower levels of activity. The reproducibility of these results would appear to point towards the presence of two disintinct mutations, but this could only be confirmed by sequencing. For detailed discussion of the the characteristics of the five isolates, please see Chapter 7.

### 6.4.7 Solubility of rtrypsin following mutation & scale-up issues

The presence of no mutations within *r-trypsin* and its immediate upstream control elements, coupled with the clear reduction in specific protein yield shown by the various SDS-PAGE gels when compared to the fractionated gels made from un-mutated BL21(DE3):pET (T) cultures, implied that either elements influencing plasmid stability had been altered or that expression was being influenced by a remote mutation. Selective pressures brought about by the autotoxicity of the now soluble recombinant mature trypsin were also clearly compromising plasmid stability. Although it was not possible to observe this directly, due to the apparent disabling/deregulating of the induction mechanism, indirect evidence was presented by the increasingly poor growth performance of the strain as fermentation conditions improved as a result of scale-up. On the basis of the evidence presented here, it would therefore seem reasonable to assume the following: 1) that plasmid variants tested had accrued through XL1-Red fermentation one or more identical or functionally identical nucleotide mutations, none of which were detected in the gene itself or in the upstream ribosome binding site, -35 or –10 motifs; 2) these mutations suppressed T7 based expression to such a degree that induction of the strain was now functionally irrelevant; 3) despite this, the apparent degree of expression cassette 'leakage' from these variants appeared to be significantly higher than that of un-mutated BL21(DE3):pET (T), while remaining at a low enough level to allow the recombinant enzyme to fold naturally into an active conformation in the soluble

fraction of the cell; 4) partial translocation of the now soluble pelB sequence-tagged protein between the cytoplasm & periplasm was hinted at in the fractionation results at all scales, but could not be accurately assessed, or indeed regarded as a fully genuine result, as no fraction specific protein 'markers' were available as thus the integrity of both fractions collected was in doubt. While a series of tests was run using periplasmic extraction buffers containing progressively less lysozyme, there could be no confirmation in this system that a certain percentage of cells were simply being lysed and thus contaminating the sample.

Although no protein bands of the same molecular weight as r-trypsin were detected in the gels, there did appear to be a range of new lower molecular weight bands in both the soluble and insoluble fractions. The possible significance of these bands is discussed in Section 7.1.2.

### 6.4.8 XL1-Red mutation rates

For any mutator strain, the rate of spontaneous mutagenesis would be directly proportional to the number of generations that the DNA of interest was propagated and thus, although strain doubling time is significant, the copy number of the target plasmid would be equally important. This fact is ignored in Stratagen company literature, which states only the mutation rate determined by the developers using a colE1 vector (Greener *et al.*, 1996). pBluescript colE1 variant (*rop* deletion) cloning vectors exhibit a copy-number of ~100, total size = 4kb, while pBR322 vectors (on which pET(T) was based) have a copy-number of 15-25 per cell, total size of pET(T) = ~6.1kb (Sambrook, 1989). To generate the same mutation rate per kb, an experimental protocol would thus require pET(T) to be propagated for four to five times the number of generations as stated in the manufacture's notes, with a rate of 2nt substitutions per kb requiring approximately 150 generations. As 30 generations can be accomplished through a pooling and overnight growth step, but each successive sub-culturing (1:1000 dilution and allowed to grow to stationary phase: approximately 10 generations) involves a 15 hour fermentation step, any advantages in speed and ease of manipulation gained through the use of the mutator strain are negated. The size of the vector possess an additional problem as any mutation has only a 1:6 chance of lying within the region of interest and thus a high mutation rate must be pursued to avoid a library composed almost entirely of wild-type enzyme. pBR322 vectors can thus not be recommended for use with the XL1-Red strain unless a strategy of selection could be implemented. Viable alternatives would

be any of the pUC-based plasmids which are based on the same mutated colE1 origin as the vector used by the developers and would thus exhibit a similar copy number.

## 6.5 Conclusions

High levels of r-trypsin activity achieved in microwell fermentation of multiple mutants generated through XL1-Red mutagenesis.

Subsequent testing revealed approximately 80% of the positive mutations generated to be false positives. Such false positives would have arisen through cross-contamination of the microwells during open fermentation.

The five isolates tested performed increasingly poorly when scaled up to 5ml and then 100ml fermentations, exhibiting raised levels of plasmid instability and culture lysis.

Activity appeared to be temperature and scale sensitive, with a 5ml fermentation held at 28°C being the only experiment to show comparable levels of activity with the initial microwell fermentations.

No protein bands of the correct molecular weight as trypsin were detected in fractionated protein gels made from each fermentation, but new bands of a range of lower molecular weights were apparent.

XI1-Red should be considered an inefficient method of mutagenesis for the majority of pET vectors due to their size and copy number.

# 7 Discussion

## 7.1 Further analysis of pET(T):BL21(DE3) behaviour

### 7.1.1 Linking activity & plasmid stability to culturing steps

All five selected clones from the first pET(T) library, see Section 6.3.3, demonstrated high levels of plate to plate variation in terms of activity and plasmid instability (difficulty in obtaining good quality plasmid or culture failure), with sequential repeats of an experiment yielding progressively poorer results. The principle example of this was the failure to replicate the initial high activity levels obtained from the clones in the first screen (Figure 6-6) in subsequent rounds of microplate fermentation and testing, and in the repeated failure of 5ml cultures to reach an acceptable cell density (un-shown work). Upon examination of the sequence of experiments it became clear that experiments demonstrating the highest levels of enzyme activity (any fraction) and the lowest level of well variability were those where the fermentation was inoculated directly from the original glycerol stocks (replica fermentation plates from 'A' & 'B', see Section 6.2.4) and that the decreasing fitness of each clone could be approximately correlated with the degree of propagation away from these stocks.

A significant point was that for each non-microplate fermentation run (5ml, 100ml), in the interests of a standardised inoculum and thus limited variability in growth rates and induction point, an initial 5ml overnight fermentation was performed. While this would be good practise in an established system when attempting to maximise run reproducibility, in an evidently unstable experimental strain such as pET(T):BL21(DE3), forcing cultures to grow to saturation an additional time before commencing the monitored fermentation and activity assay could have greatly added to the problems encountered. The second major issue was that, with the exception of the final round of 5ml fermentations, experimental inoculum overnight cultures were themselves inoculated from 'live culture' spread plates as oppose to directly from glycerol stock. The agar (LB) spread plate cultures were refreshed once per month and held at 4°C throughout, with the concept being that the use of such plates would prevent continual accessing of the stocks, with the possibility of contamination and the continual slight degradation of quality that this would bring. Analysis of the scheduling of particularly the replicates of the 5ml experimental runs showed that the highest levels of activity, coupled with low in-test variation, were obtained from cultures inoculated from spread-

plates only 1-4 days old. Clearly, the autotoxicity of whatever soluble protein was being produced, coupled with the documented 'leakiness' of the T7 promoter (Komai *et al.*, 1997), was sufficient to rapidly degrade the fitness of any standing culture, even with metabolism being repressed at 4°C.

### 7.1.2 Location of expressed r-trypsin

As variable activity readings were detectable from all fermentations it was reasonable to conclude that some levels of soluble trypsin was being expressed in the culture (despite the comparable activity levels between induced and uninduced strains, see Section 7.1.5), but the fractionated SDS-PAGE gels run after every fermentation (see Chapter 6) showed very little expressed enzyme of the correct molecular weight. Comparison of induced and uninduced cellular fractions proved futile as there was no significant difference in the banding profile between the two, but analysis of gels run from the fermentation products of the wild-type strain under identical conditions (Chapter 5) showed that several new protein bands were evident, encompassing a range of molecular weights from ~ 20kDa to 2kDa. The most significant addition, present in the soluble and insoluble fractions of all variants, at each scale and temperature, was a dense low molecular weight band, which appeared to contain increasing amounts of material as the fermentation temperature increased. The band was also usually stronger in the soluble fraction for the early part of the fermentation (1-4hrs), while the four other new bands (ranging in size from slightly below the 23kDa of mature trypsin, to just above the dominant trace) were uniformly stronger in the insoluble fraction. Comparisons of growth temperature and clones also revealed that these four 'middle molecular weight' bands were strongest in fractions run from variants G1 and D1 at 28° and 32°C, which correlates with the high activity readings obtained from these two particular clones. As cytosolic and periplasmic trypsin activity was confirmed by a comprehensively blanked assay and yet no mature enzyme was visible, it was reasonable to assume that the new bands were digestion products of the mature enzyme, resulting from attack of native proteolytic enzymes or through autohydrolysis. An attempt to examine the autohydrolysis products of trypsin was made by allowing the slight degradation of the marker protein used (storage at 4°C overnight before freezing as stock). The resultant digestion bands from the commercial trypsin approximately match those seen in the experimental lanes although certain of them appear slightly smaller, a fact which could point to failure in the release of the pelB leader sequence from the recombinant protein. The pelB

sequence contains 20 amino acids with a molecular weight of 2.026kDa and thus could account for a slight visible size discrepancy in certain fragments, and would indicate that those enzyme digestion products contained the N-terminal region of the protein.

Bovine trypsin contains 16 potential sites for autohydrolysis (15 lysine residues and 2 arginine), excluding the lysine found in the cleavage site of the trypsinogen prosequence (please see Figure 1-7 for an annotated trypsin sequence). If auto-digestion were allowed to run to completion then 17 fragments would result, ranging in length from 2 amino acids to 63 (extended section of beta-sheet plus pelB leader), the accessibility of each cutting site would significantly influence the rate and order of product formation, but, if allowed to proceed for long enough, all sites would eventually become accessible through the gradual dissolution of the molecule. The potential theoretical number of partially digested enzyme fragments would be extremely large, but it was likely, that within a limited time frame, a few forms would predominate based on certain exposed arginine or lysine residues. Potentially favourable cutting sites were hard to predict however as only one arginine residue and two lysine residues are completely hidden below the surface of the molecule, while the rest participate in surface helices, sheets ands loops.

### 7.1.3 Linking activity readings to autohydrolysis

The apparently partially digested state of the expressed enzyme raises the question of how any trypsin activity was recorded from the cultures, particularly as care was taken that samples intended for SDS-PAGE were processed (denatured) immediately upon collection, and thus should accurately reflect the internal state of the cells. At an IPTG concentration of 0.4 - 1mM, a T7 promoter system is capable of mediating rapid expression of a recombinant enzyme, such that after 2-4hrs it can comprise ~50% of the total cellular protein (Studier ,1991; Studier and Moffatt, 1986), and thus even a high rate of autohydrolysis would leave a steady-state level of active enzyme in the cell.

An additional point would be that partial autohydrolysis need not destroy the functionality of the protease as, depending on the location of the digestion point, the forces involved in secondary structure may often prove sufficient to maintain a protein in the correct conformation. In possible confirmation of this, of the surface exposed cutting sites, 6 of the lysine are structural components of anti-parallel beta-sheets, where strong lateral hydrogen bonding can provide a stabilising

'scaffold' against chain breakage (Creighton, 2000), and 2 lysine residues and the sole exposed arginine are present in short loops between anti-parallel sheets, which could again compensate for any loss in stability due to cleavage. As a robust digestive-tract enzyme, trypsin possesses six disulphide bonds, four of which span at least half the sequence length of the protein (see Figure 1-7), and these may also provide a stabilising influence, allowing a peptidase with one or two primary sequence cuts to remain active, assuming that they were able to form correctly in the cytoplasm or periplasm. If the above suggestions were in some way accurate, and that primary sequence cuts were accumulating on the protease molecules but that forces of secondary and tertiary structure were acting to maintain the integrity of the substrate binding pocket and catalytic sites, then the proteins would appear as multiple small fragments under SDS-PAGE, as sample preparation conditions are sufficiently denaturing to dissolve all bonds except the covalent peptide bond.

### 7.1.4 Cleavage of pelB

Fractionation activity tests of the five variants gave ambiguous results as to whether translocation of the expressed enzyme into the periplasm was occurring at an appreciable rate, but gel electrophoresis appeared to show trypsin autodigestion products still bonded to the pelB leader sequence in the soluble fraction (see Section 7.1.2). As a 'signal-recognition particle' (SRP) independent signal sequence translocated through the Sec system (Martoglio and Dobberstein, 1998), pelB is known to be cleaved by the membrane bound Spase I and released back into the cytoplasm. It could thus be reasonable to assume that those fragements that are still connected to the leader sequence did not undergo active, Sec-mediated transport through the cytoplasmic membrane. There are multiple documented cases of failure in the translocation of fusion proteins, with commonly cited reasons being a propensity of the recombinant protein to aggregate, possibly leading inclusion body formation, or high rates of proteolysis, with the added factor that high expression rates can exacerbate both effects (Mujacic *et al.*, 1999; Wickner *et al.*, 1991). Improved expression levels do not necessarily translate into higher yields of periplasmic protein as it has been demonstrated that export efficiency decreases significantly when protein overexpression is sufficient to saturate elements of the system (Mujacic, Cooper, and Baneyx, 1999). It is entirely possible that the failure to sufficiently reduce the expression rate (even the 'enhanced leakage' rate

evident in the mutants, see Section 7.1.5) caused a bottleneck at the level of SecB binding and thus allowed the rate of proteolytic digestion (including autodigestion) of the enzyme to increase to an effective $V_{max}$. Upregulated secretion rates, with an accompanying drop in levels of insoluble cytoplasmic recombinant protein, have been achieved by the supplementation of additional copies of the genes *secY* & *secE*, elements of the sec translocation machinery, showing that periplasmic transport can indeed act as a rate-limiting step (Perez-Perez *et al*., 1994).

Signal sequences that remain exposed within the cytoplasm have been shown to be processed by a number of prokaryotic peptidases, principally membrane-bound protease IV and cytosolic oligopeptidase A (Martoglio and Dobberstein, 1998). These peptidases can cleave the majority of signal peptides at certain conserved regions and can also act to degrade the sequence while still fused to a recombinant enzyme if the hybrid is retained within the cytoplasm. The fact that a number of r-trypsin fragments appear to remain bonded to fragments possessing approximately the correct molecular weight as the complete pelB sequence could be explained in two ways. Firstly, the attached leader sequences have been unevenly degraded at a number of sites, so as to produce several distinct species of r-trypsin N-terminal fragments with pro-sequences of varying lengths, some still bonded to undigested pelB. The fact that only four conserved new bands were visible on the protein gels, rather than a smear across a range of molecular weights which would be expected to occur if a population of r-trypsin N-terminal autodigestion products were bonded to a number of pelB fragments of different lengths, would indicate that this would be an unlikely explanation. The second possibility would be that the leader sequence had become associated with the bound enzyme, folding inwards and possibly becoming involved in the r-trypsin tertiary structure binding interactions, and thus was protected from rapid digestion.

### 7.1.5 Loss of induction control

Activity profiles and protein gels run from induced and uninduced fermentations of the five selected isolates clearly demonstrate that the mutants constitutively express r-trypsin, irrespective of the addition of IPTG. The fact that no nucleotide mutations were found in either *r-trypsin*, the prosequence or the immediate upstream transcription elements (-10 & -35 consensus sequences, RBS, lac operator region or the T7 promoter binding site) would seem to indicate that the

mutation(s) is influencing some other aspect of the T7 control mechanism on a remote region of the plasmid. The mutation must be carried on the plasmid (both plasmid and XL1-Red genome were mutated, but the vector was transformed into BL21(DE3) for propagation and expression), and the only other plasmid encoded T7 control element is the *lac I* sequence which spans positions 764-1843 (see Figure 5-6). The T7 system is based on the standard lac operon, a system of negative promoter control, and thus dependant upon the constitutive expression of the lac repressor from the *lac I* gene, which binds to the lac operator region and sterically hinders the binding of the T7 polymerase. A mutation which disabled the *lac I* promoter and thus prevented or lowered expression, or within the body of gene, acting to lower its affinity for the lac operator would have the effect of reducing negative control of T7 polymerase binding, thus allowing expression levels to reach inducer-saturated conditions. This clearly has not occurred as r-trypsin expression levels were seen to be considerably lower than those of 'wt' pET(T):BL21(DE3), while still higher than the slight leakage of protein visible under uninduced conditions. The most convincing explanation would be that a mutation had occurred within the body of the *lac I* gene, which alters the structure of the protein so as to lower its affinity towards an inducer (allolactose or IPTG) and thus shifting the equilibrium of the reaction so that the majority of repressor molecules remain bound to the operator, with only a small percentage at any one time binding to the inducer and altering their conformation to release the operator. As induction does not appear to result in any increase in recombinant protein production (conclusion arrived at by examination of fractionated SDS-PAGE gels and activity readings) then the repressor has been mutated so as to posses no affinity for the inducer molecule at all and this structural mutation (assuming the presence of just a single substitution) has an additional effect of distorting the conformation of the DNA-binding region, resulting in a higher basic rate of disassociation, i.e. lowering the stability of the repressor-DNA complex. This would allow 'enhanced leakage' through the T7 system while still ensuring a low level of protein expression.

### 7.1.6 Possible unified theory

Viewing the effects discussed in Sections 7.1.3– 7.1.4 in the light of the probable *lac I* mutation (Section 7.1.5) it becomes possible to link the apparent expression rates of the five mutants with the r-trypsin activity detected in each cellular fraction. The model proposed in Figure 7-1 has the

expression rate of the system acting as a motive force, driving unfolded wild-type enzyme into inclusion body aggregates. The reduction in 'pressure' caused by the *lac I* mutation causes a variety of different target locations to become accessible, including periplasmic translocation, but the presence of soluble cytoplasmic enzyme also allows for protease-related autotoxicity to emerge and equally exposes the enzyme to cell-mediated degradation and autohydrolysis. The rate of degradation of r-trypsin was clearly reduced at the lower fermentation temperatures, hence the high activity levels recorded when grown at 28°C, and this could be explained in terms of the reduced expression rate at this temperature leading to a lower overall quantity of enzyme in the cytoplasm, and thus reduced rates of proteolysis. The relationship between enzyme concentration and the rate of degradation could possibly be seen as directly proportional up to a point, with high level of r-trypsin saturating native degradative proteases and also with each molecule of trypsin have an increased chance of contact with a second, thus enhancing the rate of autohydrolysis. As G1 & D1 showed consistently high activity during low-temperature fermentations but possibly exhibited lower concentrations of r-trypsin digest fragments on the fractionated SDS-PAGE gels, it is possible that these two isolates possessed a mutation which depressed their expression rates even below that of the others (D2, D4 & C4). The resultant 'trickle' of enzyme would thus remain soluble and active longer than if expressed at a higher rate, effectively allowing a strong signal to be assayed from a significantly autotoxic and labile recombinant product.

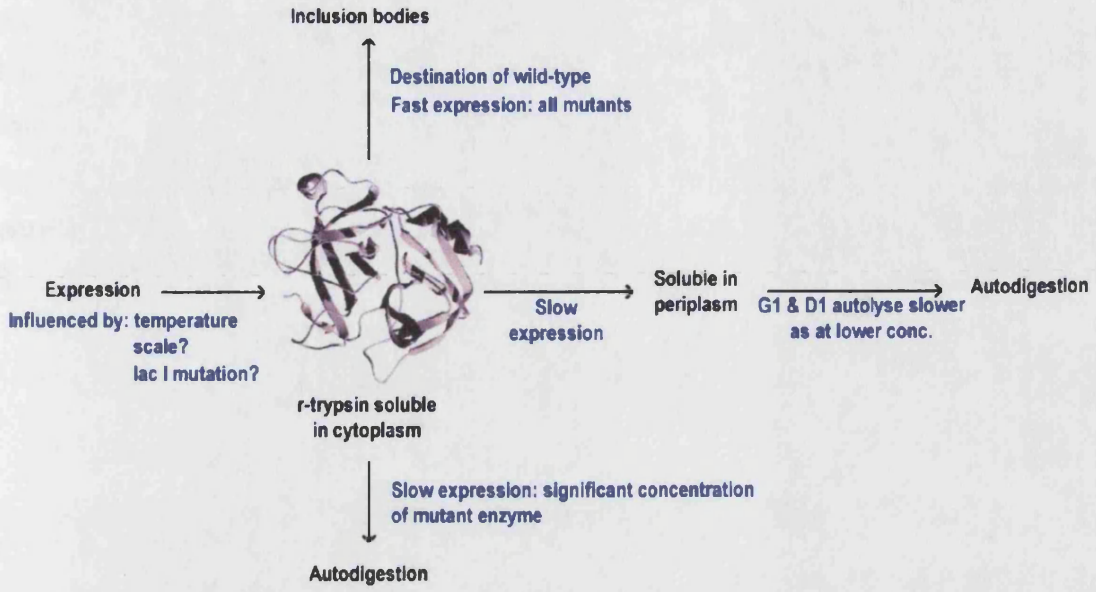



**Figure 7-1: Possible model for behaviour of mutants**

Relating activity and concentrations of r-trypsin mutants to expression rates.

### 7.1.6 General

The generation of pET vector mutants as opposed to trypsin mutants was a severe blow in terms of the project goals, but it should be noted that they provide a perfectly valid solution to the problem of r-trypsin insolubility. They also illustrate both the principle danger and advantage of random mutagenesis, namely that unexpected solutions are actively encouraged to arise, and it is dependant upon the researcher to design a mutagenic strategy and screen which will most focus the possibilities in his or her favour. The use of a mutator strain lays open the possibility of a favourable mutation occurring outside the target gene and, with protein solubility being influencd by a wide variety of factors, an expression-rate suppressing mutation should not, in hindsight, have been unexpected.

The first rule of directed evolution is "you get what you screen for" (Arnold, 1996), and refers to the fact that the researcher is essentially mining sequence space for all possible solutions to a specific question posed by the screen. In the case of the work described in Chapter 6, the question was "what mutation would allow soluble, active trypsin to be expressed in detectable levels during a deep-well microplate fermentation?" and we were, in the classic sense of making

wish, given exactly what we asked for but in an unexpected form. The supposed *lac I* mutation allowed the autotoxic, labile mature protease to be expressed at detectable levels during deep-well fermentation, but proved intolerant to scale-up and, of course, not an enzyme mutation at all. The system does, however, present a potentially useful platform for the expression of toxic recombinant enzymes within *E. coli* and methods for characterisation and improvement are discussed in Section 8.

# 8 Conclusions

A practical methodology for the high-throughput culturing and screening of cells expressing recombinant trypsin was developed and tested, with the capacity of the system limited only by the lack of more than two microplate thermomixers.

A high-throughput screening protocol involving the microplate solubilisation and refolding of inclusion bodies could not be established due to the inability to sufficiently purify the insoluble fraction, dilution factors being limited by the need to screen for activity without a means of concentrationg the enzyme, and the complexity of the trypsin structure. The use of certain additional co-factors, particularily PEG300 and glycerol, was shown to be effective in enhancing the regenerated yield of commercial trypsin and thus may reward further work in this area.

A mutant library of the pET(T) construct was prepared through propogation for 30 generations with XL1-Red. Two mutations would appear to be present in the *lac I* gene (unconfirmed by sequencing as yet) which disable induction control and retard expression, while still allowing the production of assayable levels of protease. Instability of the mutants prevented full characterisation.

Despite this result, XL1-Red has been shown to be statistically a poor choice for a mutagenic strategy involving a mid-sized, medium to low copy-number plasmid, with an attempted mutation versus doubling generations calibration failing to generate sufficient hits to be a useful tool.

# 9 Future work

The essential next step in the work would be the sequencing of the *lac I* gene (nucleotides 764-1843) and its immediate upstream control elements on the five pET(T) mutants. The expression characteristics of the mutants would point to this region as being the most logical position for the putative mutation (see Section 7.1.5), but, should sequencing fail to show any abnormalities, then sequencing would be extended over the entire plasmid. It is quite likely that there are only two distinct mutations, one found in clones G1 & D1, and the other in clones D2, D4, C4 based on their differing small fermentation volume activities (see Section 6.3). After a mutation had been identified, its influence would be confirmed by using point mutagenesis to introduce it into a 'wild-type' pET(T) vector.

The problems encountered in characterising the variants, namely their apparently variable growth characteristics and r-trypsin yields, are a direct result of the autotoxicity of the mature protease and, conversely, the susceptibility of the enzyme to autohydrolysis and possible attack by native cytosolic proteases. The 'enhanced leakage' of the pET(T) mutants could provide a useful system for the soluble expression of difficult recombinant enzymes, however, in the case of r-trypsin, ways must be found to disable the autotoxicity of the protease and decrease its susceptibility to autohydrolysis. The following solutions could be attempted in parallel as each would present a slightly different set of advantages and would need to consider the industrial or research context of the enzyme.

- Work on optimising the expression of r-trypsinogen in the yeast *Pichia pastoris* found that autoactivation of the protease, and the subsequent autotoxic effect on the host, could be prevent by alteration of the pro-sequence cleavage site ($[Asp]_4$ – Lys↓). The lysine residue was replaced by an additional aspartic acid, with the result that no mature protease was detectable within the celluar soluble fraction. Activation of the pro-enzyme was then accomplished through the use of dDAP. As the pET(T) mutation would appear to be within the *lac I*, excision of *r-trypsin* and replacement with *r-trypsinogen* would allow the low-level expression of the zymogen. The construction of an *r-trypsinogen* strain with a Lys8Asp mutation, thus making it inactivatable, and parallel testing with the 'wild-type' *r-trypsinogen* and *r-trypsin* strains would allow precise determination of the degree of active autotoxicity presented by the mature

protease and any deleterious metabolic effects arising from T7 expression of a recombinant protein.

- The interactions of the catalytic triad of His57, Asp102 & Ser195 are key to the nucleophilic attack at the C-terminus of an Arg or Lys residue and this charge relay system can present a target for protein engineering for the precise control of catalysis. . It has been established through comparative studies of diverse serine proteases that, while the presence of all three amino acids is essential, their relationship with each other does not depend on being rigidly held in constant positions relative to each other, i.e. the charge relay system between the three (see Section 1.6.6) can be established as long as they are in close proximity with each other, with considerable flexibility apparently being allowed in terms of bond angles and relative alignment. The researchers in question removed the catalytic histidine residue from the protease and substituted a histidine in the substrate near the cutting site (P2 & P1' sites) so that the act of binding that region of the substrate completed the catalytic triad (Corey *et al.*, 1995). The subtilisin H64A variant constructed (subtilisin numbering, corresponds to the trypsin His 57 residue) showed very high specificity for hydrolysis at HX↓ and X↓H peptide linkages, where X corresponded to a small hydrophobic amino acid, typically alanine. Similar engineering in r-trypsin could eliminate all types of protease-based autotoxicity and autohydrolysis, and complete specificity in substrate cleavage sites. Mutation of the enzyme for enhanced specificity towards extended sequence motifs in the now tightly regulated protease would however be restrained in terms of sequence diversity.
- Expression levels, cytoplasmic solubility and also periplasmic translocation of r-trypsin could all possibly be improved by certain combinations of fusion proteins and commercially available specialist expression strains. Fusion proteins (alternative signal sequences, N- or C- fusions of small native proteins) would be tested in a factorial manner with alternative expression strains. Certain tags would also enable rapid purification of the recombinant protein from either media or cell lysate through the use of microplate wells coated with an appropriate antibody or ligand. Such methods were not explored in this work but would present an effective high-throughput purification methodology, enabling both a more sensitive screening procedure and concentration of the enzyme.

| Tag or fusion partner | Size | Location | Function |
|---|---|---|---|
| Disulphide oxidoreductase (DsbA) | 208aa | N- | Enhanced expression & periplasmic translocation. Increased copy № of foldase genes may increase folding yields. |
| Disulphide isomerase (DsbC) | 236aa | N- | As above. |
| Streptavidin | 15kDa | N- or C- | Possibly enhanced secretion, effective purification. |
| Omp A or T | 21aa/22aa | N- | Sec, SRP independent Secretion tags, possibly improved action over pelB. |
| PapD | 28.5kDa | N- | Periplasmic folding chaperone. Chaperone activity believed to be specific to pili subunits but may enhance translocation. |
| AD494 (Novagen) | N/A | Host strain | Mutation in thioredoxin reductase (*trxB*) allows for a more oxidising cytoplasmic environment to encourage folding of proteins with multiple Cys-Cys. |
| Origami (Novagen) | N/A | Host strain | Double mutant in thioredoxin reductase (*trxB*) and glutathione reductase (*gor*), effect as for AD494 above. |

**Figure 9-1: Possible approaches to r-trypsin expression/folding**

Fusion tags and proteins that may enable enhanced expression and/or secretion of r-trypsin while repressing autotoxicity, and two engineered strains designed to encourage folding of low-levels of eukaryotic proteins in *E. coli* cytoplasm. All references taken from (Gething, 1997).

- Use of an alternative promoter with a lower base expression rate could allow greater levels of translocation as well as lowering levels of autohydrolysis by decreasing the concentration of peptidase in the cytoplasm. Figure 6-15 showed that higher yields of active enzyme were obtained at 28°C, presumably due to both the depression in expression rate and the reduced activity of r-trypsin at this growth temperature (enzyme activity from fractions assayed at room temperature), and with the possible benefit of up-regulation of cold-shock chaperones at this temperature (Georgiou and Valax, 1996). The best characterised *E. coli* cold-shock promoter is *cspA*, which has been previously used for the successful low-temperature expression of both toxic and proteolytically sensitive fusion proteins (Mujacic *et al.*, 1999). An alternative low-level *E. coli* system would be the nutritionally inducible arabinose promoter (*ara*BAD), which allows graded levels of protein expression down to trace levels by means of titration of the L-arabinose inducer (Keasling *et al.*, 2000). Investigation of both systems could

provide a means of low-level enzyme expression which, coupled with molecular engineering to reduce autohydrolysis, could present a stable base for further rounds of mutagenesis.

# Nomenclature

| | | |
|---|---|---|
| $\varepsilon$ | Extinction coefficient | $M^{-1}cm^{-1}$. |
| $\mu$ | Number of mutations on individual strand | - |
| A | Absorbance | - |
| $C^n$ | Overall mutation rate matrix after *n* extension steps from parent | - |
| $c_p$ | Protein concentration | $\mu g*ml^{-1}$ |
| *f* | Overall expected error frequency (equation 1.1) | - |
| *i* | Number | - |
| *j* | Number | - |
| *l* | Optical pathlength of solution | cm |
| M | Matrix of individual nucleotide substitution probabilities | - |
| *M* | Total number of mutations throughout population | - |
| *n* | Number of extension steps required to produce a DNA strand | - |
| *N* | Number of PCR cycles | - |
| *p* | Error rate per PCR cycle | - |
| $T_{annealing}$ | Annealing temperature of double-stranded oligonucleotide | °C |
| $T_m$ | Melting temperature of double-stranded oligonucleotide | °C |

# References

Anfinsen, C. Principles that govern the folding of protein chains. Science 181, 233. 1973

Antal,J., Pal,G., Asboth,B., Buzas,Z., Patthy,A., Graf,L. Specificity Assay of Serine Proteases by Reverse-Phase High-Performance Liquid Chromatography Analysis of Competing Oligopeptide Substrate Library. Analytical Biochemistry 288, 156-167. 2001.

Arnold,F., 1996. Directed Evolution: Creating Biocatalysts for the Future. Chemical Engineering Science 51, 5091-5102.

Arnold,F., Moore,J., 1997. Optimization of industrial enzymes by directed evolution. Advanced in biochemical engineering 58, 1-14.

Arnold,F., Volkov,A., 1999. Directed evolution of biocatalysts. Current Opinion in Chemical Biology 3, 54-59.

Auld,D.S., 1999. Fluorometric Assays. In: Sterchi,E.E., Stocker,W. (Eds.), Proteolytic Enzymes: Tools and Targets. Springer-Verlag, Heidelberg, pp. 30-48.

Backes,B., Ellman,J., 1999. An Alkanesulfonamide "Safety-Catch" Linker for Solid-Phase Synthesis. J. Org. Chem. 64, 2322-2330.

Backes,B., Harris,J., Leonetti,F., Craik,C., Ellman,J., 2000. Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. Nature Biotechnology 18, 187-193.

Barfoed,H.C., 1987. Insulin Production technology. Chemical Engineering Progress 83, 49-54.

Barnes,W.M., 1992. The fidelity of Taq ppolymerase catalysing PCR is improved by an N-terminal deletion. Gene 112, 29-35.

Birkett,A.J., 1997. Determination of enzyme specificty in a complex mixtures of peptide substrates by N-terminal sequence analysis. Analytical Biochemistry 196, 137-143.

Black,M., Newcomb,T., Wilson,H.-M., Loeb,L., 1996. Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy. Proc. Natl. Acad. Sci. USA 93, 3525-3529.

Bode,W., Schwager,P., 1975. The Refined Crystal Structure of Bovine beta-Trypsin at 1.8A Resolution. Crystallographic Refinement, Calcium Binding Site, Benzamidine Binding Site and Active Site at pH 7.0. J. Mol. Biol. 98, 693-717.

Bornscheuer,U., Altenbuchner,J., Meyer,H., 1998. Directed Evolution of an Esterase for the Stereoselective Resolution of a Key Intermediate in the Synthesis of Epothilones. Biotechnology and Bioengineering 58, 554-559.

Bornscheuer,U., Pohl,M., 2001. Improved biocatalysts by directed evolution and rational protein design. Current Opinion in Chemical Biology 5, 137-143.

Buswell,AM, Ebtinger,M, Vertes,AA & Middelberg,APJ. 2002. Effect of operating variables on the yield of recombinant trypsinogen for a pulse-fed dilution-refolding reactor. *Biotech & Bioeng 77*:435-444

Cariello,N.F., Swenberg,J.A., Skopek,T.R., 1991. Fidelity of Thermococcus Litoralis DNA Polymerase (Vent) in OCR determined by denaturing gradient gel electrophoresis. Nucleic Acids Research 19, 4193-4198.

Chen,K., Arnold,F., 1993. Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc. Natl. Acad. Sci. USA 90, 5618-5622.

Clare,D.A., Valentine,V.W., Catignani,G.L., Swaisgood,H.E., 2001. Molecular design, expression and affinity immobilization of a trypsin-streptavidin fusion protein. Enzyme and Microbial Technology 28, 483-491.

Coco,W.M., Levinson,W.E., Crist,M.J., Hektor,H.J., Darzins,A., Pienkos,P.T., Squires,C.H., Monticello,D.J., 2001. DNA shuffling method for generating hgihly recombined genes and evolved enzymes. Nature Biotechnolgy 19, 354-359.

Coia,G., Ayres,A., Lilley,G., Hudson,P., Irving,R., 1997. Use of mutator cells as a means for increasing production levels of a recombinant antibody directed against Hepatitis B. Gene 201, 203-209.

Colman,R.F., 1997. Affinity labelling. In: Creighton,T. (Ed.), Protein Function. Oxford University Press, Oxford, pp. 155-184.

Copeland,R., 2000. Enzymes: A Practical Introduction to Structure, Mechanism and Data Analysis, 2nd edition. Wiley-VCH, Inc..

Corey,D., Willett,W., Coombs,G., Craik,C., 1995. Trypsin Specificity Increased through Substrate-Assisted Catalysis. Biochemistry 34, 11521-11527.

Crameri,A., Railard,S.A., Bermudez,E., Stemmer,W., 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391, 288-291.

Crameri,A., Whitehorn,E., Tate,E., Stemmer,W., 1996. Improved Green Fluorescent Protein by Molecular Evolution Using DNA Shuffling. Nature Biotechnology` 14, 315-319.

Creighton,T.E., 2000. Protein folding coupled to disulphide-bond formation. In: Pain,R.H. (Ed.), Mechanisms of protein folding. Oxford University Press, Oxford, pp. 250-279.

Crueger,W., Crueger,A., 1990. A Textbook of Industrial Microbiology, 2nd edition. Science Tech Publishers, Munchen.

Daniel,R., 2002. Construction of Environmental Libraries for Functional Screening of Enzyme Activity. In: Brakmann,S., Johnsson,K. (Eds.), Directed Molecular Evolution of Proteins: Or How to Improve Enzymes for Biocatalysis. Wiley-VCH Verlag GmbH, Weinheim, pp. 63-78.

Dobson, C.M., 2000. The nature and significance of protein folding. In: Pain, R.H. (Ed.), Mechanisms of protein folding: Frontiers in Molecular Biology, 2nd edition. Oxford University Press, pp 1-33.

Eckert,K., Kunkel,T., 1990. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. Nucleic Acids Research 18, 3739-3744.

Evnin,L., Craik,C., 1988. Development of an Efficient Method for Generating and Screening Active Trypsin and Trypsin Variants. Annals New York Academy of Sciences 542, 61-74.

Evnin,L., Vasquez,J., Craik,C., 1990. Substrate specificity of trypsin investigated by using a genetic selection. Proc. Natl. Acad. Sci. USA 87, 6659-6663.

Fersht,A., 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding, 3rd edition. W.H. Freeman and Company.

Fromant,M., Blanquet,S., Plateau,P., 1995. Direct Random Mutagenesis of Gene-Sized DNA Fragments Using Polymerase Chain Reaction. Analytical Biochemistry 224, 347-353.

Furlong,S.T., Mauger,R.C., Strimpler,A.M., Liu,Y., Morris,F.X., Edwards,P.D., 2002. Synthesis and Physical Characterization of a P1 Arginine Combinatorial Library, and Its Applications to the Determination of the Substrate Specificity of Serine Peptidases. Bioorganic and Medicinal Chemistry 10, 3637-3647.

Georgiou,G., Valax,P., 1996. Expression of correctly folded proteins in *Escherichia coli*. Current Opinion in Biotechnology 7, 190-197.

Gething,M.-J., 1997. Guidebook to Molecular Chaperones and Protein-Folding Catalysts, 1st edition. Sambrook and Tooze Publishing Partnership, Oxford University Press, Oxford.

Graf,L., Jancso,A., Szilagyi,L., Hegyi,G., Pinter,K., Naray-Szabo,G., Hepp,J., Medzihradszky,K., Rutter,W., 1988. Electrostatic complementarity in the substrate binding pocket of trypsin. Proc. Natl. Acad. Sci. USA 85, 4961-4965.

Greener,A., Callahan,M., Jerpseth,B., 1996. An Efficient Random Mutangenesis Technique Using an *E. coli* Mutator Strain. Methods in Molecular Biology 57, 375-385.

Haddix,P.L., Paulsen,E.T., Werner,T.F., 2000. Measurement of Mutation to Antibiotic Resistance in *Serratia marcescens*. Bioscene 26, 17-21.

Hanquier,J., Sorlet,Y., Desplancq,D., Baroche,L., Ebtinger,M., Lefevre,J.-F., Pattus,F., Herschberger,C.L., Vertes,A.A., 2003. A single mutation in the active site of bovine

trypsinogen enhances its accumulation in the fermentation broth of the yeast *pichia pastoris*. Applied and Environmental Microbiology 69, Not yet published.

Heath,D., 1998. An Introduction to Experimental Design and Statistics for Biology, 1st edition. UCL Press Ltd., London.

Hedstrom,L., 1996. Trypsin: A Case Study in the Structural Determinants of Enzyme Specificity. Biological Chemistry 377, 465-470.

Henke,E., Bornscheuer,U., 1999. Directed evolution of an esterase from *Pseudomonas fluorescens*. Random mutagenesis by error-prone PCR or a mutator strain and identification of mutants showing enhanced enantioselectivity by a resorufin-based fluorescence assay. Biological Chemistry 380, 1029-1033.

Herber,W.K., Bailey,J., Carty,C.E., Heimbach,J.C., Maigetter,R.Z., 1991. Production of cytotoxic proteins in *Escherichia coli*: a fermentation process for producing enzymatically active HIV-1 protease. Appl. Microbiol. Biotechnol. 36, 149-152.

John,R.A., 1992. Photometric assays. In: Eisenthal,R., Danson,M.J. (Eds.), ENzyme Assays: A Practical Approach. Oxford University Press, New York, pp. 59-92.

Keasling,J.D., Jones,K.L., Van Dien,S., Pramanik,J., Carrier,T.A., 2002. New Tools for the Metabolic Engineering of *Escherichia coli*. Tools for Metabolic Engineering of *E. coli*. pp. 97-111.

Kenner,G.W., McDermott,J.R., Sheppard,R.C. Development of an alkansulfonaminde linker for solid phase peptide synthesis. J.Chem.Soc. 636. 1976.

Keohavong,P., Thilly,W.G., 1989. Fidelity of DNA pol;ymerases in DNA amplification. Proc Natl Acad Sci U S A 86, 9253-9257.

Kim,S.K., Kim,K.H., Choi,S.E., Yoon,J-W., Kang,Y., 1999. Soluble Expression in *Escherichia coli* of Murine Endogenous Retroviral Transmembrane Envelope Protein having Immunosuppressive Activity. Protein Expression and Purification 16, 19-26.

Kohno,T., Carmichael,D.F., Sommer,A., Thomson,R.C., 1990. Refolding of Recombinant Proteins. Methods in Enzymology 185, 187-195.

Komai,T., Ishikawa,Y., Yagi,R., Suzuki-Sunagawa,H., Nishigaki,T., Handa,H., 1997. Development of HIV-1 protease expression methods using the T7 phage promoter system. Appl. Microbiol. Biotechnol. 47, 241-245.

Kroeff,E.P., Owens,R.A., Campbell,E.L., Johnson,R.D., Marks,H.I., 1989. Production scale purification of biosynthetic human insulin by reversed-phase high-performance liquid chromatography. Journal of Chromatography 461, 45-61.

Kuchner,O., Arnold,F., 1997. Directed evolution of enzyme catalysts. Trends in Biotechnology 15, 523-530.

Kung, H., Wei, C., Bekesi, E., Lowy, D., and Willumsen, B. 1990 Solubility of Recombinant Proteins Expressed in *Escherichia Coli*. Academic Press, Inc. Current Research in Protein Chemistry.

Kurth,T., Grahn,S., Thorman,M., Ullmann,D., Hofmann,H., Jakubke,H., Hedstrom,L., 1998. Engineering the S1' Subsite of Trypsin: Design of a Protease Which Cleaves between Dibasic Residues. Biochemistry 37, 11434-11440.

Kurtzman,A., Govindarajan,S., Vahle,K., Jones,J., Henrichs,V., Patten,P., 2001. Advances in directed protein evolution by recursive genetic recombination: applications to therapeutic proteins. Current Opinion in Biotechnology 12, 361-370.

Lam,K.S., Lebl,M., 2003. Synthesis of a one-bead one-compound combinatorial peptide library. Methods in Molecular Biology 87, 1-6.

Leroux, M.R., Hartl, F.U., 2000. Cellular functions of molecular chaperones. In: Pain, R.H. (Ed.), Mechanisms of protein folding: Frontiers in Molecular Biology, 2nd edition. Oxford University Press, pp 364-405.

Lesk,A.M., 2001. Introduction to protein architecture, 1 edition. Oxford University Press, Oxford.

Liao,H., McKenzie,T., Hageman,R., 1986. Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. Proc. Natl. Acad. Sci. USA 83, 576-580.

Lin,Z., Thorson,T., Arnold,F., 1999. Functional Expression of Horseradish Peroxidase in *E. coli* by Directed Evolution. Biotechnol. Prog. 15, 467-471.

Lundberg,K.S., Shoemaker,D.D., Adams,M.W., Short,J.M., Sorge,J.A., Mathur,E.J., 1991. High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. Gene 108, 1-6.

Lustig K.D., 1997. Small pool expression screening: identification of genes involved in cell cycle control, apotosis and early development. Methods in Enzymology 283, 83-99.

Lutz,S., Benkovic,S., 2000. Homology-independant protein engineering. Current Opinion in Biotechnology 11, 319-324.

Lutz,S., Benkovic,S., 2002. Engineering Protein Evolution. In: Brakmann,S., Johnsson,K. (Eds.), Directed Molecular Evolution of Proteins: or How to Improve Enzymes for Biocatalysis. Wiley-VCH Verlag GmbH, Weinheim, pp. 177-214.

Lutz,S., Ostermeier,M., Benkovic,S., 2001a. Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides. Nucleic Acids Research 29, 1-7.

Lutz,S., Ostermeier,M., Moore,G., Maranas,C., Benkovic,S., 2001b. Creating multiple-crossover DNA libraries independant of seqence identity. Proc Natl Acad Sci U S A 98, 11248-11253.

Marana,S.R., Lopes,A.R., Juliano,L., Juliano,M.A., Ferreira,C., Terra,W.R., 2001. Subsites of Trypsin Active Site Favor Catalysis or Substrate Binding. Biochemical and Biophysical Research Communications 290, 494-497.

Martoglio,B., Dobberstein,B., 1998. Signal Sequences: more than just greasy peptides. Trends in Cell Biology 8, 410-415.

Matsumura,I., Ellington,A., 2001. Mutagenic PCR of Protein-Coding Genes for *In Vitro* Evolution. In: Braman,J. (Ed.), Methods in Molecular Biology, vol. 182: In Vitro Mutagenesis. Humana Press Inc, Totowa, NJ, pp. 261-269.

Matthews,D.J., Wells,J.A., 1993. Substrate phage: selection of protease substrates by monovalent phage display. Science 260, 1113-1117.

Miller,J., 1998. Mutators in *Escherichia coli*. Mutation Research 409, 99-106.

Minshull,J., Stemmer,W., 1999. Protein evolution by molecular breeding. Current Opinion in Chemical Biology 3, 284-290.

Miyazaki,K., Arnold,F., 1999. Exploring Nonnatural Evolutionary Pathways by Saturation Mutagenesis: Rapid Improvement of Protein Function. J Mol Evol 49, 716-720.

Miyazaki,K., Wintrode,P., Grayling,R., Rubingh,D., Arnold,F., 2000. Directed Evolution Study of Temperature Adaptation in a Psychrophilic Enzyme. Journal of Molecular Biology 297, 1015-1026.

Moore,G., Maranas,C., Gutshall,K., Brenchley,J., 2000. Modeling and optimization of DNA recombination. Computers & Chemical Engineering 24, 693-699.

Moore,G.L., Maranas,C.D., 2000. Modeling DNA Mutation and Recombination for Directed Evolution Experiments. J. theor. Biol. 205, 483-503.

Moore,J., Arnold,F., 1996. Directed evolution of a *para*-nitrobenzyl esterase for aqueous-organic solvents. Nature Biotechnology 14, 458-467.

Moore,J., Jin,H., Kuchner,O., Arnold,F., 1997. Strategies for the *in vitro* Evolution of Protein Function: Enzyme Evolution by Random Recombination of Improved Sequences. Journal of Molecular Biology 272, 336-347.

Mujacic,M., Cooper,K.W., Baneyx,F., 1999. Cold-inducible cloning vectors for low-temperature protein expression in *Escherichia coli*: application to the production of a toxic and proteolytically sensitive fusion protein. Gene 238, 325-332.

Nagano, N., Orengo, C.A., Thornton, J.M., 2002. One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based in their Sequences, Structures and Functions. J. Mol. Biol. 321, 741-765.

Ngiam,S.H., Zhou,Y.H., Turner,M.K., Titchner-Hooker,N.J., 2001. Graphical method for the calculation of chromatographic performance in representing the trade-off between purity and recovery. Journal of Chromatography 937, 1-11.

Novabiochem, 2000. 2000 Catalog. CN Biosciences.

Ostresh,J.H., Winkle,J.H., Hamashin,V.T., Houghten,R.A., 1994. Peptide Libraries: Determination of Relative Reaction Rates of Protected Amino Acids in Competitive Couplings. Biopolymers 34, 1681-1689.

Patthy,L., 1999. Protein Evolution, 1 edition. Blackwell Science Ltd, Budapest.

Perez-Perez,J., Marquez,G., Barbero,J., Gutierrez,J., 1994. Increasing the Efficiency of Protein Export in *Eschericia coli*. Biotechnology 12, 178-180.

Perona,J., Evnin,L., Craik,C., 1993. A genetic selection elucidates structural determinants of arginine versus lysine specificity in trypsin. Gene 137, 121-126.

Perona,J., Hedstrom,L., Rutter,W., Fletterick,R., 1995. Structural Origins of Substrate Discrimination in Trypsin and Chymotrypsin. Biochemistry 34, 1489-1499.

Petrides,D., Sapidou,E., Calandranis,J., 1995. Computer-Aided Process Analysis and Economic Evaluation for Biosyntheitc Human Insulin Production - A Case Study. Biotechnology and Bioengineering 48, 529-541.

Price,N., Stevens,L., 1999. Fundamentals of Enzymology, 3rd edition. Oxford University Press Inc..

Proteolytic Enzymes: A Practical Approach. Beynon, R. J. and Bond, J. S. 1. 1989. Oxford, Oxford University Press. Practical Approach. Rickwood, D. and Hames, B. D.

Rubingh,D., 1997. Protein engineering from a bioindustrial point of view. Curent Opinion in Biotechnology 8, 417-422.

Sambrook,J., Fritsch,E.F., Maniatis,T., 1989. Molecular Cloning: A Laboratory Manual, 2nd edition. Cold Spring Harbour Laboratory Press.

Schellenberger,V., Siegel,R.A., Rutter,W. Analysis of Enzyme Specificity by Multiple Substrate Kinetics. Biochemistry 32, 4344-4348. 1993a.

Schellenberger,V., Turck,C., Rutter,W., 1994. Role of the S' Subsites in Serine Protease Catalysis. Active-Site Mapping of Rat Chymotrypsin, Rat Trypsin, alpha-Lytic Protease, and Cercarial Protease from *Schistosoma mansoni*. Biochemistry 33, 4251-4257.

Schellenberger,V., Turck,T., Hedstrom,L., Rutter,W., 1993b. Mapping the S' Subsites of Serine Proteases Using Acyl Transfer to Mixtures of Peptide Nucleophiles. Biochemistry 32, 4349-4353.

Schwienhorst,A., 2002. Advanced Screening Strategies for Biocatalyst Discovery. In: Brakmann,S., Johnsson,K. (Eds.), Directed Molecular Evolution of Proteins: Or How to Improve Enzymes for Biocatalysis. Wiley-VCH Verlag GmbH, Weinheim, pp. 159-175.

Sichler,K., Hopfner,K.-P., Kopetzki,E., Huber,R., Bode,W., Brandstetter,H., 2002. The influence of residue 190 in the S1 site of trypsin-like serine proteases on substrate selectivity is universally conserved. FEBS letters 530, 220-224.

Sieber,V., Martinez,C.A., Arnold,F., 2001. Libraries of hybrid proteins from distanly related sequences. Nature Biotechnolgy 19, 456-460.

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

Smith,A.J., 1992. Techniques in Protein Chemistry, 3 edition. Academic Press, Orlando, Orlando.

Somorin,O., Nishi,N., Noguchi,J., 1978a. New Synthetic Methods for Benzyloxycarbonyl-L-arginine-*p*-nitroanilide, and Acetyl-L-arginine-*p*-nitroanilide. Bulletin of the Chemical Society of Japan 51, 1255-1256.

Somorin,O., Tokura,S., Nishi,N., Noguchi,J., 1978b. The Action of Trypsin on Synthetic Chromogenic Arginine Substrates. J. Biochem. 85, 157-162.

Stemmer,W., 1994a. DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. Proc. Natl. Acad. Sci. USA 91, 10747-10751.

Stemmer,W., 1994b. Rapid evolution of a protein *in vitro* by DNA shuffling. Nature 370, 389-391.

Sterchi, E. and Stocker, W. 1. 1999 Proteolytic Enzymes: Tools and Targets. Springer.

Studier,F.W., 1991. Use of Bacteriophage T7 Lysozyme to Improve an Inducible T7 Expression System. J. Mol. Biol. 219, 37-44.

Studier,F.W., Moffatt,B.A., 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol. 189, 113-130.

Tindall,K.R., Kunlel,T.A., 2003. Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. Biochemistry 27, 6008-6013.

Tobin,M.B., Gustafsson,C., Huisman,G.W., 2000. Directed evolution: the 'rational' basis for 'irrational' design. Current Opinion in Structural Biology 10, 421-427.

Todd, A.E., Orengo, C.A., Thornton, J.M., 2002. Plasticity of enzyme active sites. TRENDS in Biochemical Sciences 27, No. 8, 419-426.

Venekei,I., Graf,L., Rutter,W., 1996. Expression of rat chymotrypsin in yeast: a study on the strucural and functional significance of the chymotypsinogen propeptide. FEBS letters 379, 139-142.

Watson,J.D., Gilman,M., Witkowski,J., Zoller,M., 1992. Recombinant DNA, 2nd edition. Scientific American Books, New York.

Wickner,W., Driessen,A., Hartl,F., 1991. The Enzymology of Protein Translocation Across the *Escherichia coli* Plasma Membrane. Annu. Rev. Biochem. 60, 101-124.

Winter,J., Hauke,L., Rudolph,R., 2002. Renaturation of human proinsulin - a study on refolding and conversion to insulin. Analytical Biochemistry 310, 148-155.

Yano,T., Oue,S., Kagamiyama,H., 1998. Directed evolution of an aspartate aminotransferase with new substrate specificities. Proc. Natl. Acad. Sci. USA 95, 5511-5515.

Yano,T., Oue,S., Kagamiyama,H., 1998. Directed evolution of an aspartate aminotransferase with new substrate specificities. Proc. Natl. Acad. Sci. USA 95, 5511-5515.

You,L., Arnold,F., 1994. Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. Protein Engineering 9, 77-83.

Zaccolo,M., Williams,D.M., Brown,D.M., Gherardi,E., 1996. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. J. Mol. Biol. 255, 589-603.

Zaks,A., 2001. Industrial Biocatalysis. Current Opinion in Chemical Biology 5, 130-136.

Zhang,J.-H., Dawes,G., Stemmer,W., 1997. Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. Proc. Natl. Acad. Sci. USA 94, 4504-4509.

Zhao,H., Arnold,F., 1997a. Functional and nonfunctional mutations distinguished by random recombination of homologous genes. Proceedures of the National Academy of Science USA 94, 7997-8000.

Zhao,H., Arnold,F., 1997b. Optimisation of DNA shuffling for high fidelity recombination. Nucleic Acids Research 25, 1307-1308.

Zhao,H., Giver,L., Shao,Z., Affholter,J., Arnold,F., 1998. Molecular evolution by staggered extension process (StEP) *in vitro* recombination. Nature Biotechnology 16, 258-261.

# 11 Appendix A

## 11.1 Design of a specificity screen

## 11.2 Introduction

In the context of enzymology, the term 'specificity' has a highly precise meaning in that it refers to both the activity of an enzyme and its affinity to a substrate, and it is inherently competitive, intended to describe the discrimination of protein between alternate targets. Specificity is thus a function of both the competitive substrate binding ($K_M$) and catalytic rates ($k_{cat}$) used in describing the associations of an enzyme in a solution with a desirable and less-desirable substrates. Improvements in one process can balance out reductions in the other and thus the ratio of $k_{cat}/K_M$ is used as the kinetic constant in describing enzyme specificity. The crucial point is that true $k_{cat}/K_M$ values can only be by means of competitive assays rather than *in vitro* testing against single substrates species.

A goal of this project was to devise a practical strategy by which the specificity of bovine trypsin mutant variants could be rapidly and accurately evaluated in a high-throughput manner. As the value of this screen would lie in its capacity to detect both unnatural substrate specificities (P1 amino acids other than Arg or Lys) and enhanced extended specificities (evaluation of optimal amino acids in P2 and P3 positions) it was clear that the screen must be based on a library of tripeptides. The synthesis of hundreds (at the very least) of defined tripeptides would prove costly and immensely time consuming and so a combinatorially generated library designed expressly for positional scanning was the obvious solution. Combinatorial approaches have been used to probe the optimal consensus sequences for binding to the substrate-recognition regions of several proteases and the process involves three fundamental steps: 1) the generation of the library (by biological or synthetic methods); 2) proteolysis of the substrates; and 3) identification of optimal substrate sequence (Backes *et al*, 2000). These will be discussed here in turn.

Biological methods for generating a combinatorial peptide library include classic filamentous phage display (Matthews and Wells, 1993), for which a number of partially randomised cDNA libraries with defined P1, P2 or P3 positions must be produced, and the use of such cDNA libraries in *in vitro* expression systems (Lustig K.D., 1997). Both methods would allow for effective and flexible library production with the potential for easy scale-up of material, although the cost

effectiveness of *in vitro* transcription/translation reagents would require evaluation against synthetic methods for each positionally defined peptide produced. There would be several apparent disadvantages of the biological methods for this work, the first being the cost and technical difficulty involved in preparing the cDNA library consisting of defined and randomised positions. This would involve either the bulk manufacture of the partially randomised oligonucleotides by a contract manufacturer or the preparation of a heterologous primer 'mini-library' for each peptide mix, which could then be amplified by PCR. Each oligonucleotide would contain 9 base pairs in length coding for the relevant amino acids, with N- & C-terminal extensions containing appropriate restriction sites and also promoter elements if required. The second issue that arises when employing any host expression system is that of potential discrepancies in expression levels as well as the inability to incorporate unnatural amino acids into the peptide if required. Expression level variations could be influenced by codon usage of the strain (impact depending on the promoter strength of the gene, but virtually certain to be encountered at some point in a randomised library) and the possibility of insolubility (unlikely when seeking the expression of such small peptides). The use of *in vitro* systems can compensate for variations in expression levels by means of balanced pools of available amino acids and would also allow multiple solutions for dealing with insolubility (Lustig K.D., 1997). The third problem is critical, in that it concerns the detection of substrate hydrolysis with the emphasis being on speed. Proteolytic cleavage from a display library can most accurately be determined by means of chromatography and mass spectrometry (Backes *et al.*, 2000), but these methods, while providing accurate qualitative data as to substrate acceptability, require time-consuming work after the event, forming a potential bottle-neck for a high-throughput assay and also fail to capture the kinetics of the reaction, essential for specificity analysis.

Synthetic methods of peptide library construction involve support-bound peptides generated by the sequential addition of $\alpha$-amino and side-chain protected amino acid residues to an initial residue bound to an insoluble, polymeric support (Figure 7.1). When the design of the assay allows for a chromo- or fluorophore to be attached to the N-terminus of the chain then the generation of signal, and thus the kinetics of the reaction, can be measured on-line. The problem arising in this case is that trypsin cuts at the C-terminus of strongly basic amino-acids and thus would require a chromophore to be located at this position (see L-BAPNA design, Chapter 2). Work has been done involving a fluorescence-quenched support resin which evolved a signal upon proteolysis of the displayed peptide (Lam and Lebl, 2003), but the synthesis of such a library was

technically complex. An additional problem with small, resin-bound substrates is that research has shown that the digestion kinetics of such substrates can differ from those of sequence-identical soluble molecules and, although the two data sets were related in a linear manner, thus allowing for the support-bound substrates to yield useful qualitative data, the knowledge of such effect does diminish the usefulness of the library. In order benefit directly from the enzymic assay data generated from the initial testing work on L-BAPNA (Chapter 2) and the screening of 'enhanced solubility' mutants (Chapter 6), it seemed evident that a variation of the standard solid phase peptide synthesis method was required which would allow for the synthesis of a similar soluble peptide with a C-terminal pNA group.

### 11.3 Solid phase peptide synthesis

As mentioned previously, the classical method of solid phase peptide synthesis (SPPS) involves the sequential addition of amino acids, blocked at the N-terminus and with protected side-chains, to a an initial amino acid which is bound to a support resin by a C-terminal linker (Figure 7.1). The N-$\alpha$ blocking group is removed from the newly coupled amino acid by means of a specific deprotection procedure (dependant on the nature of the blocking group) and the reaction is repeated until the desired length is reached. The side-chain protecting groups and the blocking group are then removed and the mature peptide may be uncoupled from the resin by means of cleavage of the linker region or held if the resin is to form a part of the display library. The choice of the linker agent used determines whether cleavage yields a C-terminal peptide acid or an amide (Novabiochem, 2000).

The chemistry of the N-$\alpha$ blocking group determines the reagents required for the extension of the peptide chain (steps 3 & 4 in Figure 7.1) and thus the bulk of the chemical steps as the resin can be easily bought with the first amino acid attached (thus ensuring the integrity of the linker region). There are two main variants of blocking group chemistry, namely the older Butoxycarbonyl (Boc) scheme and the now more common protocol utilising a 9-Fluroenylmethoxycarbonyl (Fmoc) group. Cleavage of a Boc-protection group requires trifluroacetic acid and side-chain deprotection requires a strong acid such as hydrogen fluoride (HF) or trifluoromethanesulfonic acid (TFMSA). Fmoc chemistry requires the use of less hazardous reagents, needing only treatment with the mild base piperidine for un-blocking the leading residue

and only a single, final treatment with HF for side-chain deprotection. Purity comparisons of short peptides synthesised using the two strategies showed comparable results, but with slightly higher levels of side-chain degradation and truncation in samples prepared using Boc chemistry (Smith, 1992). For these reasons only Fmoc chemistry was considered when preparing the strategy of combinatorial library synthesis.

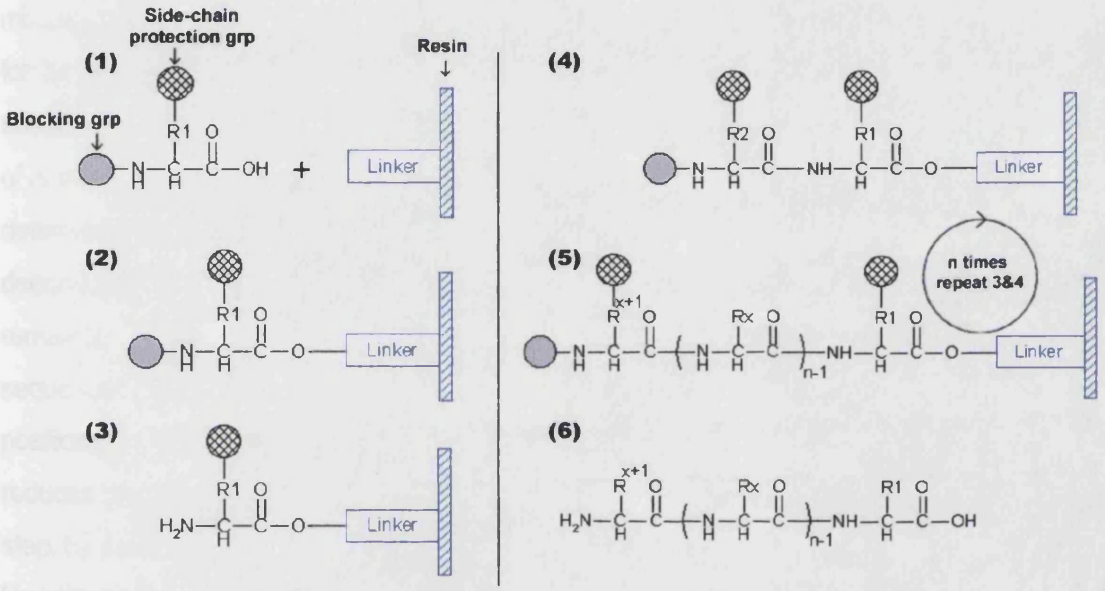



**Figure 11-1: General scheme of solid phase peptide synthesis**

Side-chain protecting groups and blocking groups refer to those used in both Boc and Fmoc chemistry. (1) Attachment of C-terminal of initial peptide to linker, usually by esterification; (2) Formation of ester bond; (3) Removal of N-$\alpha$ blocking group; (4) coupling of 2nd activated amino acid; (5) repeat of un-blocking and coupling, steps 3 & 4 until all chain reaches desired length; (6) Deprotection of side-chains and cleavage of linker-peptide bond. Figure adapted from (Novabiochem, 2000).

## 11.4 Positional-scanning synthetic combinatorial libraries

Synthetic combinatorial libraries are a tool by which vast numbers of closely related peptides may be generated for screening, either in mixtures or as automated parallel reactions, in the time usually required to prepare a handful of analogues by conventional methodologies. All combinatorial libraries based on SPPS involve the principle of portioning-mixing, which simply

describes the splitting of the resin and growing chain into aliquots and then the coupling and combining with the various activated amino acid mixtures. A full combinatorial library of all possible tripeptides from a complete pool of the 20 amino acids would require $3^{20}$ peptides, a huge number but not beyond the capacity of an automated setup. The complexity of each peptide mixture in such an experiment must be carefully planned to balance ease of synthesis with the resolution by which a positive result could be pinpointed, i.e. how much was known about the composition of each mixture and so how much data could be gained from a positive signal. When generating a library for the purpose of determining the optimum substrate sequence for an enzyme it can be advantageous to use a step-wise synthetic combinatorial library (SCL), as this reduces the number of analogous peptides involved and its more controlled, incremental nature, allows for accurate determination of the P1, P2, P3, etc positions. The library is based on the principle of serial deconvolution, by which groups of related mixtures, each with one or two defined positions and the remainder randomised, are generated and screened in solution (Novabiochem, 2000). Those sequences generating the highest signal are carried to the next level of the screen and the defined positions are used as the base for further portioning and mixing. A good variation of this technique reduces the need for the continual resynthesis of the various sub-libraries used in each mixing step, by using a technique known as recursive deconvolution (Figure 11-2). By this method, sub-libraries composed of peptides at each stage of synthesis, and thus at each length, are retained and used in subsequent synthesis steps. A true positional scanning synthetic peptide library (PS-SCL) eliminates the need for iterative synthesis and screening completely as it may be synthesised in a single step. A tripeptide PS-SCL would be composed of 3 positional sublibraries, with each library acting to define one position in the peptide. Each sublibrary would be separated into well containing tripeptides with one position occupied by a known amino acid (thus a full library would require 19 or 20 divisions for each sublibrary, depending if troublesome residues such as cysteine or methionine were excluded) and with the other positions filled with randomised amino acids.

It is important to note that a PS-SCL libraries is not a full factorial peptide library, i.e. while it must contain an equimolar mixture of every possible peptide, these peptides are not grown in discrete reactions allowing for defined quantities to be added to each mix, but rather are synthesised by the addition of a solution of amino acids at each coupling step. Work into the kinetics of SPPS has shown that the relative reaction (coupling) rates of each amino acid varies by type up to a factor of 10 (Ostresh et al 1994). The initial resin-bound amino acid was also found to

exert a significant influence on the coupling rates of incoming amino acids, but this appeared to be a constant effect, acting with the same proportional effect on all types, and thus unlikely to present a problem for a library synthesised from a defined P1, P2, etc (Figure 11-2) (Ostresh *et al.*, 1994). Thus, although the relative coupling rates are largely independent of the resin-bound amino acid, a mixture of protected amino acids must be carefully composed if the iterative competitive coupling of the mix to a growing peptide is to produce an un-biased, randomised library, containing approximately equimolar concentrations of each peptide. For position X in Figure 11-2 & Figure 11-3 to be truly random, the mix of single, activated residues used in each coupling step must be present in the proportions indicated in Table 11.1, the ratios were determined by competitive, fractional-factorial coupling experiments compensating for resin effects (Ostresh *et al.*, 1994). The proportions within the mixture would seem intuitive as those amino acids requiring greater concentrations if they are to be represented equally in the peptide library are those with bulky sidechains, which would act to sterically hinder the formation of complexes.

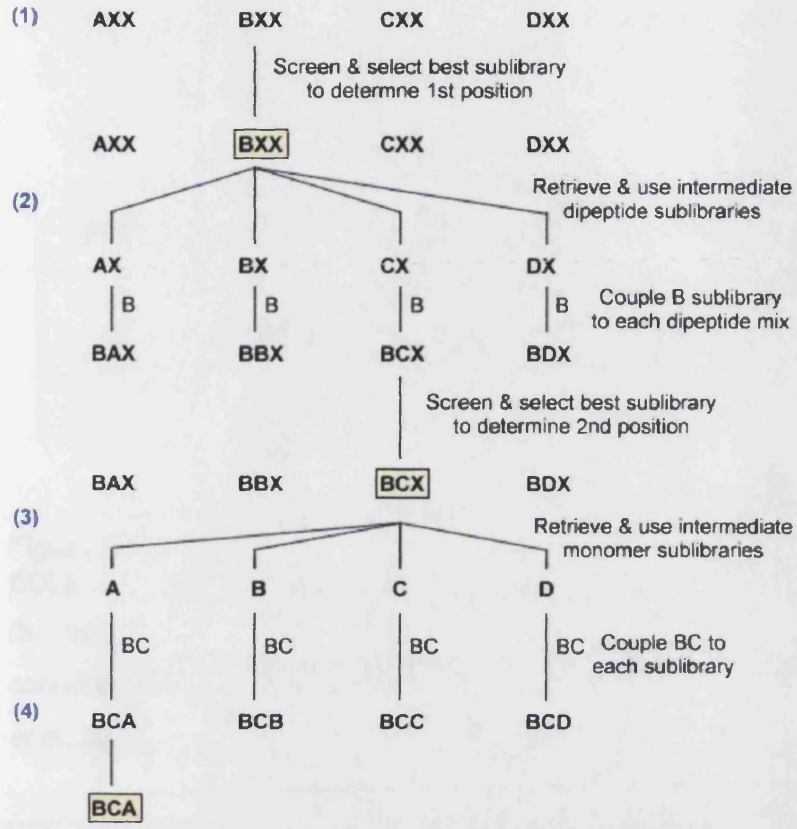



**Figure 11-2: General scheme of a synthetic combinatorial tripeptide library (SCL) using recursive deconvolution**

X = randomised position containing an equimolar mixture of amino acids (in this example A, B, C & D). Steps: 1) screening of initial tripeptide library with only the P1 position defined, selection of 'best' substrate to form base of 2nd round of screening; 2) Coupling of intermediate dipeptides, saved from midway through the synthesis of the sublibrary in step 1, to the empirically determined optimum P1; 3) determination of optimal P1 and P2 sequence by screening, resynthesis of peptide and coupling to monomer sublibrary retained from step 1; 4) screening for the optimal complete sequence. Figure adapted from (Novabiochem, 2000).

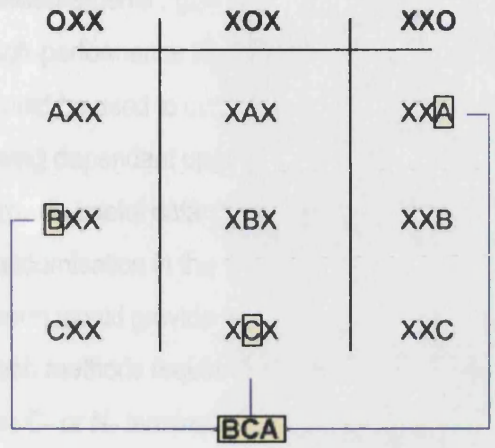



**Figure 11-3: General scheme of a positional scanning synthetic combinatorial library (PS-SCL).**

Illustrates a tripeptide library from a mixture of 3 amino acids (A, B, C), X = randomised position containing an equimolar mixture of amino acids, O = defined position. Figure adapted from (Backes *et al.*, 2000).

| Ala (A) | Arg (R) | Asn (N) | Asp (D) | Gln (Q) | Glu (E) | Gly (G) | His (H) | Ile (I) |
|---|---|---|---|---|---|---|---|---|
| 3.58 | 6.87 | 5.64 | 3.70 | 5.61 | 3.84 | 3.04 | 3.76 | 18.3 |
| **Leu (L)** | **Lys (K)** | **Met (M)** | **Phe (F)** | **Pro (P)** | **Ser (S)** | **Thr (T)** | **Tyr (Y)** | **Val (V)** |
| 5.23 | 6.56 | 2.42 | 2.66 | 4.56 | 2.93 | 5.04 | 4.36 | 11.9 |

**Table 11.1: Isokinetic rates if individual amino acids**

Showing the mole percentage of amino acid derivative required for equimolar coupling using a 10-fold excess of protected amino acid mixture to resin-bound residue. Table adapted from (Ostresh *et al.*, 1994).

### 11.5 Analysis of hydrolysis products

In order for it to be possible to determine the kinetic constants of substrate consumption ($k_{cat}/K_M$), essential if a specificity profile of enzyme variants is the target, then a method must obviously be in place to follow substrate concentration over time. This could be accomplished via continuous adsorption (Yano *et al.*, 1998) or fluorescent (Backes *et al.*, 2000; Kurth *et al.*, 1998)

measurements , gas chromatography, N- or C- terminal sequence analysis (Birkett, 1997) or by high-performance liquid chromatography (Antal *et al.*, 2001; Schellenberger *et al.*, 1994). The data would be used to calculate the relative second-order rate constants with the power of the analysis being dependant upon the technique, i.e. quantitative amino acid analysis from HPLC would provide useful data on the digested states of a mix of peptides, depending on the degree of randomisation in the library, while monitoring the release of an absorbent or fluorescent leaving group would provide information on the hydrolysis of only the tagged peptide. A further limitation of such methods requiring the release of a reporter group is that the tag must be attached to either the C- or N- terminal end of the cutting site, depending on the activity of the tested enzyme, meaning that only the influence of the $P_1$, $P_2$,...,$P_n$ or the $P'_1$, $P'_2$,...,$P'_m$ sites respectively could be monitored. These methods also leave open the possibility for the leaving group to interact with the substrate-binding region of the enzyme in some manner, skewing results away from those based on the hydrolysis of larger peptides.

To fully assess the contribution to specificity of each amino acid within the protease-binding site of the substrate, methods which allow for the evaluation of positions $P_4 - P'_4$ (dependant upon size of protease-binding region, determined for a wide range of peptides by means of small-molecule inhibitors or defined substitution experiments, e.g. alanine scanning) are required, and various HPLC based protocols have been proven to deliver comprehensive results within complex mixtures of potential substrates. Schellenberger *et. al.* have devised a method for the accurate testing of extended P' positions based on the blocking of the P1-2 positions with a preferred substrate and then using the P' peptide to trigger an acyl transfer reaction, i.e. reversing the peptide hydrolysis reaction and using nucleophilic P' peptides along with an acyl transfer ester, yielding data comparable with conventional hydrolysis (Schellenberger *et al.*, 1993b; Schellenberger, Turck, and Rutter, 1994). The resultant peptide products are monitored at continuous intervals by HPLC and fluorescence detection. Reverse-phase HPLC has a proved a particularly effective method is studying both hydrolysis rates and points of cleavage with a semi-randomised competitive peptide library (Antal *et al.*, 2001). Optimisation of buffers ensures well-spaced retention times of the various peptide cleavage products and UV quantification of peaks previously analysed by conventional HPLC provided a simple and rapid method of establishing enzyme preference. The accuracy and reproducibility of this method make it an effective qualitative specificity test for uncharacterised or mutant proteases, with the crucial advantage that potential

multiple reaction types may be followed within a mixture (a feat possible with adsorption or fluorescence assays only through the use of diverse-wavelength leaving groups), but the time-lag of analysis and the possibility that the complex data acquired may need secondary rounds of analysis could prove problematic for a true high-throughput screen. A further issue is that of the need to purify the sample prior to the digestion reaction, in order to prevent contamination, and thus degradation, of the column. Providing that the target strain had been engineered to express soluble protein, then, in a direct-measurement spectroscopic assay, the peptide mix could be added directly to the cellular lysate and measurements started immediately. For a chromatographic assay, the soluble protein must first be separated from the remainder of the cellular soluble fraction, a simple matter if appreciable levels of secreted material have been obtained, but considerable more complicated if the target protein was retained in either the periplasm or cytoplasm. Release of periplasmic proteins could be effected by the use of lysozyme buffers (as discussed in Section 5.2.5), essentially creating a suspension of periplasts, which may then be pelletted by gentle centrifugation. The use of specialist 96-format filter plates (with the filtrate being retained) could also prove useful for such methodologies. If the target protein was contained predominantly in soluble form within the cytoplasm, then high-throughput methods available for purification are reduced to antibody-mediated affinity binding, either onto nitrocellulose membrane or through specially coated wells. An effective methodology would involve the tagging the C-terminus of r-trypsin with a 6His motif, followed by a suitable protease digestion site, and purification via a nickel-algenate resin coated well-plate. Although such methods have been proven to be viable, and with the added advantage of working with now purified protein, they are usually coupled with a simple activity screen and, in a screen involving a complex, partially randomised peptide library, the added complexity could prove a deterrent.

In terms of direct measurements, discarding the use of radioisotopes which require specialised handling, synthesis and disposal and are used only when exceptional degrees of sensitivity are needed, chromogenic and fluorescent substrates allow for simple, rapid quantitative analysis of a single reaction type. Chromogenic substrate types are discussed in depth in Chapter 2, and have the advantage of being able to be employed in both solution-based screening and also in semi-quantitative initial agar-based selection procedures, providing of course that the target enzyme is accessible and it is possible to tie its activity either to cell viability or simply to generate a halo. Fluorescence provides a considerably more sensitive assay method and probably constitutes

the majority of modern bioassays. Such assays are similar to chromogenic methods in that the fluorophore, typically 7-hydroxycoumarin (umbelliferone) or 7-amino 4-methylcoumarin (AMC), is attached to the C-terminus of the P1 amino acid by an ester or amide bond, and the release of the fluorescent group can be monitored continuously be measuring at its signature wavelengths. Free AMC exhibits the characteristics of $\lambda_{ex}$ 345nm, $\lambda_{em}$ 445 and allows for sensitivity of detection approximately two orders of magnitude higher than assays based on classic chromogenic leaving groups, such as nitrophenol or nitroanilide (John, 1992). A variety of intramolecular quenching methodologies, where a quenching agent is incorporated within the substrate and the fluorophore comprises part of the scissile bond or within a distinct portion of the substrate binding region of the enzyme, have been developed (Auld, 1999), and allow for the collection of detailed data on both the release of product and the formation of the ES complex. This can provide an excellent method for the accurate and rapid determination of the $_{i}K_{M}$ & $k_{cat}$ values, and thus the competitive substrate preferences of an enzyme.

### 11.6 Use of an alkanesulfonamide linker

The requirement of a C-terminal chromophore on the tripeptides of the proposed SCL presented a problem as standard protocols for SPPS allowed for only N-terminal modification of the growing peptide. "Safety-catch" linkers, linker groups composed of relatively stable bonds that can be cleaved only by a specific 2-step activation-release procedure, were developed for use with a number of alternative synthesis strategies. Such synthesis strategies included methods which require only increasingly mild or altered chemical conditions, frequently designed so as to minimise the incidence of damaged side-chains during mid to long polypeptide synthesis (Backes and Ellman, 1999; Novabiochem, 2000). An important additional feature is that while the linkage is stable during the extension of the chain, 'activation' of the linker can then provide a reactive, labile bond susceptible to nucleophilic attack and subsequent displacement by an appropriate group. Numerous strategies for alternative 'non-safety-catch' displaceable linkers do exist, with possible the most relevant to this work being the use of a modified CMtr linker (Furlong *et al.*, 2002), which has already proven to be suitable for the synthesis of pNA tagged peptides, but presents stability issues during synthesis. The alkanesulfonamide "safety-catch" linker was originally developed by Kenner *et. al.* (Kenner *et al.*, 1976), and activation by diazomethane treatment rendered the bond

receptive to aminolysis, hydrazinolysis and saponification, resulting in C-terminal primary amides, hydrazones and carboxylic acids respectively. Modified versions of the linker were devised with refined functionality, such as improved loading efficiency and, particularly, enhanced reactivity following activation. The current version of sulfamylbutyryl resins, in which activation is accomplished by *N*-alkylation of the *N*-acylsulfonamide linker with a haloacetonitrile, results in a bond reactive enough to be displaced by a nucleophilic amine (Backes and Ellman, 1999; Novabiochem, 2000). The cleavage of the linker region thus releases an amide product into solution, a structurally correct peptide, and this strategy allows for the attachment of a chromogenic (or fluorogenic) leaving group to the peptide through the nucleophilic addition of an appropriately derivatised amino acid at the P1 position (Backes *et al.*, 2000; Backes and Ellman, 1999) (see Figure 11-4).

The current generation of sulfamybutyryl resins (Novabiochem, cat. No. 01-64-0152) are constructed so that, prior to activation, they form a stable linker bond, resistant to extremes of pH or strongly nucleophilic reagents, thus making them compatible with a wide range of reaction conditions. Activation is achieved by addition of either diazomethane, iodoacetonitrile or $TMSN_2$ (Backes *et al.*, 2000; Backes and Ellman, 1999).

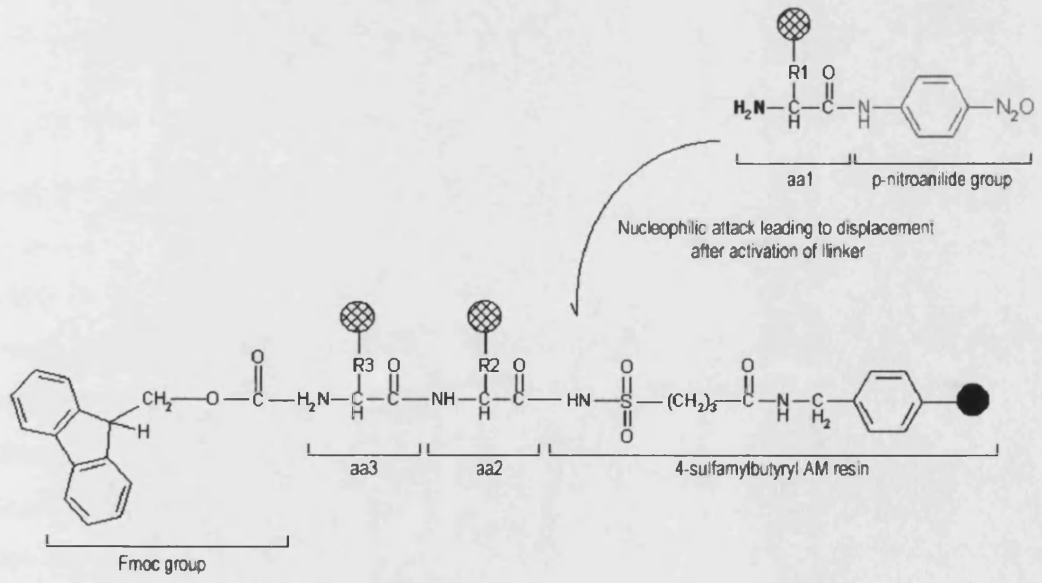



**Figure 11-4: SPPS using linker resin**

Construction of a tripeptide with a C-terminal pNA group by means of standard SPPS and subsequent displacement of sulfonamide group with a nucleophilic amine.

## 11.7 Competitive screening strategy

### 11.7.1 Pooling of library elements

Based on the methodologies described in sections 11.4 & 11.6, a strategy was developed for the synthesis and competitive screening of enzyme mutants against a substantial tripeptide array, capable of determining the P1, P2 & P3 specificities of a library of trypsin variants. The exact nature of the library was obviously dependant upon the 'question' that the screen was designed to ask, and equally by the limitations imposed by the demands of a high-throughput process. This implies that each enzyme mutant should initially be tested as quickly as possible, i.e. by assaying in the minimum number of library wells, and, likewise, the various tripeptide permutations forming the library must be pooled to the highest degree possible, while still allowing for meaningful results. 'Pooling' would be the process of combining a certain number of discrete library tests into a single microplate-well test, but must be done in such a manner that a positive result yields useful information. If the test is qualitative in nature, i.e. it seeks to examine variations in reaction rate of the enzyme on a selection of substrates, then the result from a pooled well would be useless except to indicate that at least one (but equally possible, all) presented substrates were acceptable to the enzyme. An example of this would be the classic form of the PS-SCL (Figure 11-3), where, except possibly for the P1 position, all substrates should be screened separately to fully tease out enzyme specificity. A quantitative test, were reaction kinetics are, at least initially, irrelevant and a positive signal would be considered a rare occurrence, a high degree of pooling, and thus a particularly rapid screening procedure, would be possible. An example of this would be a situation in which one was screening for unnatural substrate specificities, i.e. in the case of trypsin, substrates that did not contain either arginine or lysine at the P1 position, and thus would result in no signal from the wild-type enzyme. Peptides with an OXX sequence (O being any amino acid except arg & lys, X possibly being fully randomised – bearing in mind that only cleavage at the C-terminal of 'O' would release the chromophore) could be pooled in a controlled manner to almost any degree allowing for rapid processing of enzyme variants. A positive result in a single well would require subsequent screening on 'sub-pools' of substrates from the original mix, eventually refining the screen down to individual peptides. Pooling, when appropriate, thus allows for rapid initial screening of large numbers of enzyme variants, with the requirement of secondary and possibly tertiary rounds of screening on a greatly reduced number of mutants.

### 11.7.2 Analysis of multiple substrate kinetics in regards to assessing enzyme specificity

The consumption of a substrate can be modelled through second-order kinetics (Equation 11.1) where [E] is the concentration of unbound (free) enzyme, $[S_i]$ is the concentration of the *i*th substrate form (assuming the presence of multiple substrate analogues) and $k_i$ is the rate constant ($k_{cat}/K_M$) (Schellenberger *et al.*, 1993a).

$$\frac{d[S_i]}{dt} = -k_i[E][S_i] \qquad i = 1,...,n$$

**Equation 11.1**

Consumption of competing substrates by second-order kinetics.

As discussed previously, the specificity of an enzyme can only be determined in context with competing substrates, and so it is useful to derive an expression which allows for the comparison of reaction rates between substrate types. The kinetics of a pair of substrates competing for a single enzyme active site can be described through a variation of the Michaelis-Menten mechanism (Figure 11-5). The figure shows the free enzyme ([E]) associating with two forms of substrate ($S_A$ & $S_B$) and passing through the acyl-intermediate complex (EAc) characteristic of serine proteases prior to the regeneration of the enzyme.

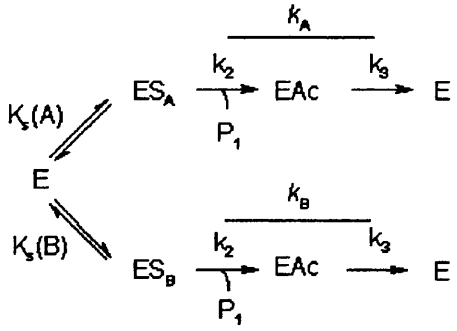



**Figure 11-5: Competition of two substrates for active site of enzyme.**

Adapted from (Fersht, 1999).

In order to relate the kinetic constants of the action of the enzyme on both substrates, it is necessary to adjust the standard Michaelis-Menten equation into a form by which the two ratios may be directly compared. A variation of the standard form, treating $K_M$ as an apparent dissociation constant, is shown in Equation 11.2.

$$v = \frac{k_{cat}}{K_M}[\mathrm{E}][\mathrm{S}]$$

**Equation 11.2**

Relating the concentration of free enzyme [E] and substrate [S] to the kinetic constants. Assumes conditions of low substrate concentrations where enzyme is largely unbound, i.e. $[E] \approx [E]_o$ (total enzyme concentration).

If two substrates, $[S_A]$ & $[S_B]$, then compete for the active site of the enzyme, the equations may be written as derivatives (Equation 11.3), and then combined as shown in Equation 11.4. The important points of this equation are that it allows enzyme specificity to be described in terms of the $k_{cat}/K_M$ ratios, thus avoiding discussion of intermediate states or non-productive binding, and, crucially, it does not require measurement of the concentration of free enzyme at any point, [E], which can be difficult to establish (Antal *et al.*, 2001; Schellenberger, Siegel, and Rutter, 1993a).

$$-\frac{d[\mathrm{A}]}{dt} = v_{\mathrm{A}} = \left(\frac{k_{cat}}{K_M}\right)_{\mathrm{A}}[\mathrm{E}][\mathrm{S_A}]$$

$$-\frac{d[\mathrm{B}]}{dt} = v_{\mathrm{B}} = \left(\frac{k_{cat}}{K_M}\right)_{\mathrm{B}}[\mathrm{E}][\mathrm{S_B}]$$

**Equation 11.3**

Describing the separate rates of consumption of two forms of substrate, $S_A$ & $S_B$.

$$\frac{v_A}{v_B} = \frac{(k_{cat}/K_M)_A[S_A]}{(k_{cat}/K_M)_B[S_B]}$$

**Equation 11.4**

Combined form of Equation 11.3, describing the competition of two substrates, $S_A$ & $S_B$, for molecules of free enzyme.

A simple rearrangement of Equation 11.4 to form Equation 11.5 allows for accurate quantification of the discrimination of the enzyme between substrates. If the kinetic values of the wild-type enzyme can be determined for a substrates then this can be used as a reference point, and the ratios obtained from the screening of variants could be used to construct a meaningful specificity profile.

$$\frac{v_A/[S_A]}{v_B/[S_B]} = \frac{(k_{cat}/K_M)_A}{(k_{cat}/K_M)_B}$$

**Equation 11.5**

Describing the 'specificity ratio' between two competing substrates, $S_A$ & $S_B$.

### 11.8 Library designs

One of the goals of this work was to develop a series of screens by which potential alterations in mutant enzyme specificity could be assessed, i.e. to investigate the plasticity and thus suitability of a serine protease as a target for directed evolution specificity engineering. Aside from this no directive was given, and thus there was considerable flexibility in how this was to be approached, both in terms of whether to assign substrate targets to evolve towards and how best to structure a peptide library to make use of the automated systems available. Although these libraries were never synthesised due to time constraints, two diverse setups are described below, each designed to act as proof-of-concept for a different method of evolving an enzyme for altered substrate specificity.

### 11.8.1 Screening for unnatural substrate specificities

An unnatural substrate in this case would be defined as one that lacked arginine or lysine in the P1 position. The narrow slot presented by the substrate binding pocket at this position (please see Section 1.4.5) with its basal, positively charged asparagine residue, demands a long, negatively charged side-chain on the P1 residue if the reaction is to be sterically favourable. Acceptance of an alternatively structured, bulky side-chain at this position would require a substantial alteration in the conformation of the pocket, and, similarly, the substitution for a small uncharged, relatively innocuous residue (e.g. alanine) has been shown to dramatically reduce the rate of proteolysis by decreasing the strength of the initial enzyme-substrate complex. A question arises as to whether bar arginine and lysine from the P2 & P3 positions as, although cleavage at the C-terminal of either position would not result in the release of the chromophore, it could be argued that the presence of such unproductive cleavage products in each well could somehow compromise the effectiveness of the assay.

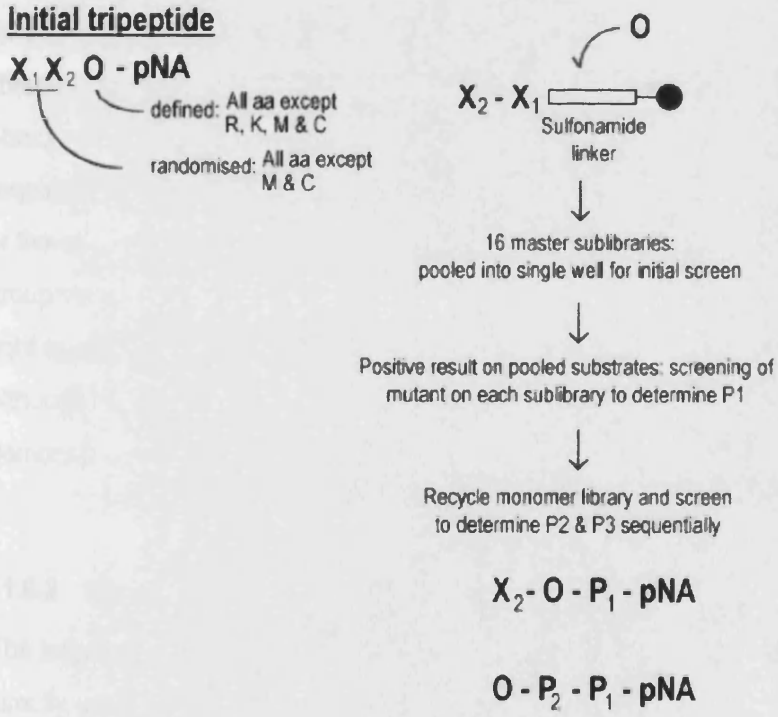



**Figure 11-6: Screening for unnatural specificities**

Schematic of library design for the screening of unnatural substrate specificities in enzyme mutant variants. '$P_1 - P_2$' refer to the 'protease-binding' position on the substrate, 'X' denotes a

randomised position, 'O' refers to a position carrying a defined amino acid. Design is a variant of the standard SCL format (Figure 10.2) and is based on the exclusion of the enzyme's preferred amino acids (Arg & Lys in the case of trypsin) from the P1 position. Cysteine and methionine were excluded from the library because of the problems they present in the batch synthesis of small peptides (see section 11.8.1).

The synthesis and screening of the library would proceed as illustrated in Figure 11-6. A pre-prepared 4-sulfamybutyryl AM resin would undergo two rounds of coupling to an isokinetic mixture of activated amino acids (see Figure 10.4), before displacement of the linker by a defined residue with a C-terminal nitroanilide group. The resultant 16 discrete sublibraries would thus represent every amino acid in the P1 position except for arginine & lysine (natural P1 determinants), and cysteine & methioine (specialist synthesis protocols required, please see below).

As no background signal beyond gradual substrate auto-hydrolysis would be expected from such libraries, and the initial screen would be purely qualitative in nature, i.e. merely testing for the release of any chromophore molecules, a high degree of pooling would be possible. All libraries could be pooled into a single well, allowing 96 enzyme variants to be screened per standard microplate. The principle constraint on the degree of pooling possible would be the required concentration of each tripeptide permutation within the mix, i.e. should the new specificity of the mutant be selective for a single P1-P2-P3 combination, could the release of the nitroanilide group from such a small fraction of the pooled mix be detected? Evidence suggests that such a tight specificity would be highly unlikely to emerge from a single round of mutagenesis and that, although the P1 acts as the prime specificity determinant, the upstream positions usually demonstrate a degree of flexibility.

### 11.8.2 Screening for enhanced specificity towards a targeted substrate

The targeted mutation of enzyme specificity towards a defined sequence represents the most directly useful applications of this technology, allowing for the development of precisely tailored tools for bioprocessing reactions. A library based on the enhancement of specificity to a specific tripeptide would not follow the pattern of positional scanning laid out in Figure 10.2& 10.3, but rather would focus on a competitive assay based on the target sequence and a series of close,

non-nitroanilide tagged analogues. Such an assay would be qualitative in nature, as its success would depend on its ability to distinguish variants that hydrolysed the target sequence, and thus released the chromophore, at a rate significantly greater than the wild-type enzyme. Parallel reactions with two varieties of peptide mixes (mixes ① & ② in Figure 11-7) would complementary data on the kinetic rates in a strictly and loosely competitive environment respectively.

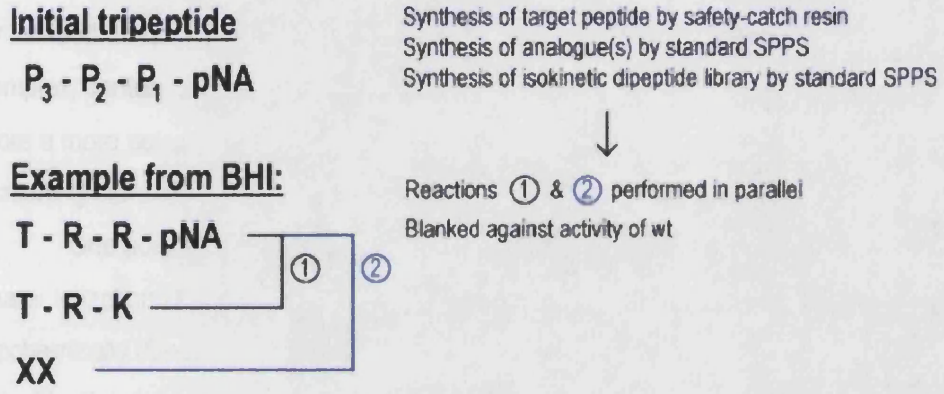



**Figure 11-7: Competitive screening**

Schematic of competitive assay designed to test for enhanced specificity towards a targeted sequence. '$P_1 - P_3$' refer to 'protease-binding' positions on substrate, 'X' refers to a randomised position, all other letters denote the single-letter amino acid abbreviations. 'BHI' stands for biosynthetic human insulin and the sequence used (TRR↓) corresponds to an optimal cleavage point on the molecule (please see Section 1.1.4).

The methods presented here provide a practical and rapid series of approaches for the screening of enzyme substrate specificity in a high-throughput format. The use of a context independent library (PS-SCL) versus a progressive, adaptive library (SCL) is dependant upon a number of factors, most notably the time required for library synthesis, through-put capacity and the 'depth' of the data acquired. Clearly, the time required for the continual resynthesis of a progressive SCL library, moulding the next step of the library based on the results from the previous and reusing existing monomer pools, makes the process unwieldy and impractical for the processing any large number of clones. The variant suggested in Figure 11-6, where unnatural substrate specificities are screened for and thus the number of expected 'hits' would be greatly reduced, has the advantage

over a PS-SCL scheme in that unique structural interactions between neighbouring amino acids may be examined. The complexity of side-chain interactions within the substrate-protease binding region has been the topic of intense study, with an example being the sterically favourable interaction between the prime specificity determinant, Asp189, at the base of the substrate-binding pocket of the protease and Ser190. Replacement of Ser190 with any other amino acid results in a significant decrease in the stability of the ES complex, an effect speculated to occur because of the loss of an additional hydrogen bond to the P1 Arg or Lys residue (Sichler *et al.*, 2002). Further work by this author, examining binding site interactions as a binary matrix reinforced the concept of complex, context dependant relationships, and thus the more work intensive SCL format may yield more a more accurate picture of an enzyme mutants' substrate selectivity, providing that the initial screening 'hurdle' is set high enough to limit the number of clones requiring processing.

One point of note was that none of these methods made allowances for the physical characterization of the putative positional scanning library. Libraries would be characterised biochemically (functionally) through the course of the work, but rest on the assumption that the use of isokinetic mixtures in the synthesis steps (see Table 10.1) has resulted in equimolar representation of each possible peptide combination. These rates were determined empirically and extensive validation has shown them to be robust (Ostresh *et al.*, 1994), but they may be influenced by unusual protocol chemistries or, conceivably, by operator inexperience. As the interpretation of the results from the biological evaluation of a substrate library is dependent upon the synthetic fidelity of the library, it has been argued that some method of initial structural validation should be considered a prerequisite to screening (Backes *et al.*, 2000; Furlong *et al.*, 2002). MALDI-TOF mass spectrometry has been shown to provide a rapid method for the validation of a proposed equimolar peptide mix and involves analysing the spectral data for the distribution of certain distinctive amino acid motifs (selection of a molecular weight 'ladder' of tripeptide motifs, but with the clearest resolution being obtained from those at the high & low extremes of the mixture), and comparing the data to a computationally simulated MALDI-TOF spectrum for a tripeptide sublibrary following an even (regular) distribution (Furlong *et al.*, 2002). Initial analysis of a partially randomised peptide library in this manner would allow for confidence in results, and also aid in the design of individual, robust synthesis protocols.

## 11.9 Library Complexity & oversampling

Mathematical descriptions of error-prone PCR and gene shuffling have been discussed elsewhere (please see Section 1.4) both in terms of their mechanisms and the expected levels of variety that each could introduce into a gene. What is presented here is a simple mechanism establishing library size and also levels of sampling redundancy, i.e. how many clones must be assayed from a gene library before there exists a 95% probability that each possible variant has been sampled at least once.

In an ideal system, where nucleotide mutations occurred without either positional bias (either through degeneracy at any particular position, so-called 'mutational hot-spots' or through untreated secondary structure 'sequestering' certain regions of sequence) or dNTP substitution bias (as can occur when performing PCR under conditions of unequal [dNTP]), nucleotide, and thus amino acid mutations would accumulate as described in Equation 11.6. The resultant library sizes from this factorial expression and dependant mainly on the numbers of simultaneous mutations induced in the target region, are illustrated in Table 11.2.

No. amino acid variants = $19^{M}[N!/(N-M)!M!]$

No. nucleotide variants = $4^{M}[N!/(N-M)!M!]$

**Equation 11.6**

Expressions describing peptide or nucleic acid library sizes, determined by length of sequence / target region exposed to mutagenesis (*N*), and number of mutations introduced simultaneously (*M*).

| | | Sequence length (*N*) | | | |
|---|---|---|---|---|---|
| | | **100nt** | **5aa** | **200aa** | **233aa** |
| No. amino acids / nucleotides changed simultaneously (*M*) | **1** | 400 | 95 | 3800 | 4427 |
| | **2** | 79200 | 3610 | 7183900 | ? |
| | **3** | 10348800 | 68590 | $9.01*10^{9}$ | ? |
| | **4** | $1.00*10^{9}$ | 651605 | $8.43*10^{12}$ | ? |
| | **5** | $7.71*10^{10}$ | 2476099 | $6.28*10^{15}$ | ? |

**Table 11.2: Illustrating absolute protein library size by target sequence length and the number of amino acid substitutions induced simultaneously**

Numbers derived through Equation 11.6. Model assumes that nucleotide mutations occurred randomly on the original gene, with no positional degeneracy or dNTP bias. Frameshift mutations (resulting in truncated proteins) are ignored. The figures shaded dark grey indicate where the number of clones is low enough to allow for complete screening of the library by the high-throughput screening protocol proposed in this work. Figures shaded light grey indicate library sizes which would become practical if a method of nutritional selection could be implemented. Values for $N = 233$ (length of r-trypsin) could not be calculated beyond $M = 1$ because of size of factorial statements required. Table (values for $N = 200$) adapted from (Kuchner and Arnold, 1997)

When screening a library of mutants, the target is of course it to assay all possible variants against the particular substrate in order to exploit the full power of the library, and thus bring to light any potentially beneficial mutations which may have been generated. As all forms of mutation, whether epPCR or by mutator strain, are performed as a controlled mixture, it follows that the resultant gene variants exist as a mix, containing (providing excess quantities of the gene were used) multiple copies of all possible variants. After transformation and plating, the colonies comprising the protease library will represent duplications, triplications, etc of each variant and thus, in order to be sure of screening each unique variant at least once, a degree of oversampling is required above the theoretical size of the library. In any sampling of a mutant library the majority of clones would be represented only once (variants identified by sequencing of target region of gene from purified plasmid obtained after protein expression in individual well-cultures), with a certain percentage being identified two or three times, and this degree of library redundancy can be used to calculate approximately: a) the total complexity of the library (already known as a theoretical value, see Table 11.2, but a useful additional guide as to the validity the model); and b) the degree of oversampling required if all unique variants are to be screened. The method relies on the frequency distribution of a library sampling fitting a poisson function, the form of which is shown in Equation 11.7.

$$P(Y) = (e^{-\mu} \mu^{Y}) / Y!$$

**Equation 11.7**

Formula for the poisson distribution: Expected relative frequency of a rare event, *P(Y)*, where *Y* = the number of occurrences of a particular type in a sample, $\mu$ = the population mean, i.e. average number of occurrences per sampling, and *e* is a constant, 2.71828 (the base of natural logarithms).

The poisson distribution is a form of the binomial, optimised for the representation of extremely large data sets, where the probability of a particular event occurring is relatively small (typically $p < 1.0$). The validity of this extreme form of binomial distribution is dependant upon the results (distribution of sequences within the clone library and the absence of bias in mutagenesis) being purely random, i.e. sample variance / mean ≈ 1 (Heath 1998). Bias, in either sampling or the initial mutagenesis, may be indicated by the occurrence of contagious dispersion, i.e. variance / mean > 1, where data can appear 'clumped', possibly because of biological factors restricting the range of the library.

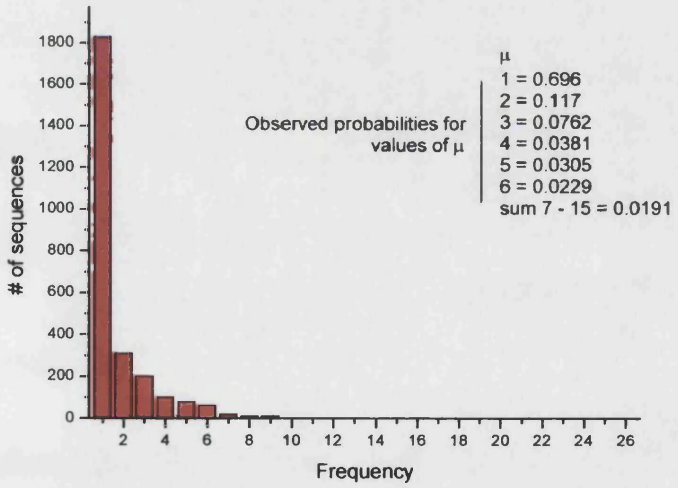



**Figure 11-8: Representation of the distribution of unique sequences in a sampling from a statistically random library**

This frequency histogram illustrates the number of independent clones sequenced following screening, results show that, from a sampling of 4631 clones, 2632 non-overlapping (unique) sequences were identified. 'μ' refers to the frequency with which unique sequences were identified in terms of the poisson distribution (Equation 11.7). Representational data set adapted from (Lin et al 1999) and unpublished data from the Arnold group.

# 12 Appendix B

## 12.1 Suppliers

Lorne Laboratories, Ltd. (UK distributor for Worthington Biochemicals)

7, Tavistock Estate, Ruscombe Business Park, Ruscombe Lane, Twyford, Reading, Berkshire RG10 9NJ. Phone 118 9342400, fax 118 9342788, e-mail lorne.labs@virgin.net

Merck Biosciences, Ltd. (also UK distributor for Novagen)

Boulevard Industrial Park, Padge Road, Beeston, Nottingham NG9 2JR. Phone 0800 622935, fax 0115 9430951

National Diagnostics (UK), Ltd.

Unit 4 Fleet Business Park, Itlings Lane, Hessle, East Riding of Yorkshire HU139LX. Phone 1482 646020, fax 1482 646013, e-mail customer@nduk.clara.ne

Promega UK, Ltd.

Delta House, Chilworth Research Centre, Southampton SO16 7NS. Phone 0800 378994.

Sigma –Aldrich Company, Ltd.

Dorset, England. Phone: 44 1202 733114, fax: 44 1202 715460, E-mail: ukcustsv@eurnotes.sial.com,

Techno-Path, Ltd. (closest distributor for Stratagene)

Rosse Center, Holland Rd. National Technological Park, Plassey, Limerick, Ireland. Phone 353 61 335844, fax 353 61 203 034, e-mail pn@technopath.iol.ie.