Development of The MAX Randomisation Technique

David Andrew Nagel

Doctor of Philosophy

ASTON UNIVERSITY
September 2003

ASTON UNIVERSITY

Development Of The MAX Randomisation Technique

David Andrew Nagel
Doctor of Philosophy

September 2003

THESIS SUMMARY

During the last decade the use of randomised gene libraries has had an enormous impact in the field of protein engineering. Such libraries comprise many variations of a single gene in which codon replacements are used to substitute key residues of the encoded protein. The expression of such libraries generates a library of randomised proteins which can subsequently be screened for desired or novel activities.

Randomisation in this fashion has predominantly been achieved by the inclusion of the codons NNN or $NN^{G/}_{C\ or\ T}$, in which N represents any of the four bases A,C,G or T. The use of these codons however, necessitates the cloning of redundant codons at each position of randomisation, in addition to those required to encode the twenty possible amino acid substitutions. As degenerate codons must be included at each position of randomisation, this results in a progressive loss of randomisation efficiency as the number of randomised positions is increased. The ratio of genes to proteins in these libraries rises exponentially with each position of randomisation, creating large gene libraries, which generate protein libraries of limited diversity upon expression.

In addition to these problems of library size, the cloning of redundant codons also results in the generation of protein libraries in which substituted amino acids are unevenly represented. As several of the randomised codons may encode the same amino acid, for example serine which is encoded six time using the codon NNN, an inherent bias may be introduced into the resulting protein library during the randomisation procedure.

The work outlined here describes the development of a novel randomisation technique aimed at eliminating codon redundancy from randomised gene libraries, thus addressing the problems of library size and bias, associated with the cloning of redundant codons. The design, development and implementation of the technique are described along with suggestions for its future development and implementation.

Keywords: Gene randomisation, gene/protein libraries, selectional hybridisation, randomised codon, zinc finger(s).

For My Family

Who taught me how to stand up

For My Family and Friends

Who kept me from falling back down

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Anna Hine for her help, encouragement, support and friendship throughout this study, these thanks are also extended to Dr Andy Sutherland. I would also like to thank Albie Santos and Amersham Biosciences for their support throughout this study.

I would also like express my gratitude to the post-doc and plutonic conversationalists, Dr. Marcus Hughes, Dr. Chad Zhang and Dr Richard Darby for their help, expertise, coffee breaks and friendship. A special thanks to Dr. Marcus Hughes for looking after our friend Max and letting me play with him occasionally. I must also extend a special thanks to Mr. David Palfrey for his continued, help, expertise and valued friendship.

I extend my thanks also to Dr. Peter Lambert for his support, help, advice and limitless knowledge and understanding of absolutely everything. Thanks also to Dr. Peter Hanson for the same reasons and apologies for all the statistics questions. I must also thank Dr. Ann Vernallis for her help and valued advice, and for allowing me to "borrow" many of her reagents. Thanks also to Dr. Jwan Khal for her help and friendship and a special thanks to Dr. Sue Lang for her help, friendship, coffee and conversation. I extend my thanks also to all the other students, who have provided valued help and friendship and who have tolerated my often disturbing presence. My thanks are also extended to the staff of the microbiology department, Mrs. Rita Chohan and Alan Richardson for their help with urgent ordering, Mr Roy McKenzie for his help with my often urgent demands for glassware and especially Kevin Hughes for his immediate help with all my problems and for making me laugh. Thanks also to all at Aston University who made my work there so enjoyable and are too numerous to list by name.

I would also like to thank Mr. R. Dodgson for his valued mathematical help, Mr. And Mrs Hogg for their help and support and friendship, and a special thank you to Mr. L. Palmer from who's valued teaching I will always continue to learn (thanks for getting me here).

I would also like to thank my family and friends for their support, especially Dr. Lucy Harper for her love, valued help in all matters scientific, support in all other matters, for

giving me words and for making it better. Mr. Adam Scott for all his support, boundless knowledge and very valued friendship and David and Sophie Hogg for their support, help and most of all friendship. Finally I must and will, always thank my family who support me in every way there is.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

ATP	Adenosine triphosphate
bp	Base pair(s)
BSA	Bovine serum albumin
cDNA	Complementary DNA
CIP	Calf intestinal alkaline phosphatase
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
DTT	Dithioreitol
EDTA	Ethylenediaminetetraacetic acid
g	Grams
IPTG	Isopropyl β-D-thiogalactoside
K_d	Dissociation constant
L	Litre
LB	Luria broth
M	Moles / Litre
Mins	Minutes
MOPS	4-Morpholinepropanesulfonic acid
NEB	New England Biolabs
OD	Optical Density
PAGE	Polyacrylamide gel electrophoresis
PEG	Polyethyleneglycol
PCR	Polymerase Chain Reaction
PNK	Polynucleotide kinase
RNA	Ribonucleic acid
rpm	Revolutions per minute
sec(s)	Second(s)
TAE	Tris acetate EDTA
TF	Transcription factor
Tm	Theoretical melting temperature
TRIS	Tris(hydroxymethyl)aminomethane
W/V	Weight to volume
V/V	Volume to volume

Chapter 1 Introduction

1.1 Background

During the last decade the construction of randomised gene libraries and the use of combinatorial methodology has had an enormous impact in the field of protein engineering. Initially reported in the late 80's (Reidhaar-Olson and Sauer, 1988; Hermes *et al.*, 1989) randomised gene libraries have little in common with conventional genomic or cDNA libraries. Conventional gene/cDNA libraries are derived from the genome/transcriptome of a single organism. Thus a complete or representative, library would be expected to contain clones which would collectively cover the entire genome/transcriptome of the organism from which that library was derived. The presence of particular genes or proteins within the genome/proteome of the library organism may then be identified by screening processes such as nucleic acid hybridisation or serological techniques.

In contrast randomised gene libraries contain many variations of a single gene. In these libraries mutations are introduced into the coding sequence of a target gene, creating a library of randomised genes. Expression of the library allows the effects of genetic mutation on the translated products of that gene to be studied.

Although novel DNA enzymes (Santoro *et al.*, 2000), antisense RNA oligonucleotides (Patzel and Sczakiel, 2000) and antisense DNA oligonucleotides (Ho *et al.*, 1996) have been generated using randomised libraries, this technique is most commonly employed in the study of proteins. The utility of these libraries in protein study cannot be underestimated. Often described as combinatorial libraries, randomised gene libraries permit the effects of multiple amino acid substitutions within a target protein to be studied at the same time. Combinatorial methods have not only facilitated the study of protein structure/function relationships but have emerged as one of the most powerful tools in the discovery of new biologically and pharmacologically important proteins. Interacting proteins can be identified from large combinatorial libraries, a process termed deconvolution (described in combinatorial chemistry as the identification of the active constituent from within a mixture) thus offering the potential to discover mutants with new or anticipated skills (Avalle *et al.*, 1997).

1.2 Combinatorial Protein libraries and Display Technologies

The selection of proteins with new or improved properties from combinatorial libraries has been described as directed evolution (Mössner and Plückthun, 2001) and indeed the parallel can be drawn between natural evolution and selection from combinatorial libraries. The process of natural selection favours the selection of phenotypes harbouring favourable genotypic mutations, once introduced these mutations are passed on to the resulting offspring via the genome of the selected parent cells.

In combinatorial methodology, the occurrence of genetic mutation within the coding sequence of a target protein is defined experimentally (methods employed in the introduction of mutations are discussed later in this section) creating a library of mutant genes. Selection pressure upon the resulting proteins is also user defined, as the conditions imposed during the screening of the proteins expressed from these libraries, allows proteins of desired phenotypic characteristics to be selected.

Once selected from the library, the identity of the mutations which generated the desired protein must be established. Effective screening of peptides from a combinatorial library therefore requires a connection between the selected protein and the nucleic acid which encoded it. This linkage between genotype and phenotype has been established logically, both by limiting the constituents of each library to certain proteins(Jamieson *et al.*,1996, & Choo and Klug, 1994) and by logically encoding the identity of interacting target sites for DNA binding proteins in the length of the DNA target sites (Desjarlais and Berg, 1994) permitting the selection of proteins from conventionally expressed libraries. However more commonly, a physical linkage between genotype and phenotype is established using display technologies developed to facilitate the efficient selection of proteins from combinatorial libraries.

Efficient display technologies must not only maintain the phenotypic/genotypic link between the mutant protein and its encoding sequence, but also ensure that mutant proteins are functionally displayed in an environment which is conducive to their selection from the library. For example, small peptides may require the physical constraint of a scaffold protein to adopt an active conformation in solution (Koide *et al.*, 1998 & McConnell and Hoess, 1995), the generation of potentially toxic proteins may

preclude the use of display technologies which require transformation of eukaryotic or prokaryotic host cells (Rungpragayphan *et al.*, 2002), whilst the selection of mutant enzymes may dictate that proteins are expressed in a eukaryotic system to ensure the functionality of the enzyme (Jermutus *et al.*, 2001).

The effective selection of proteins from combinatorial libraries has been addressed by the development of a number of different display technologies. Each of these techniques has merit, but also associated disadvantages. Thus the selection conditions for each particular protein must be considered in order to utilise the most effective display technique.

1.2.1 Phage Display

Phage display developed in the mid eighties (Smith, 1985) has become a widely accepted technique for the display of combinatorial libraries. In this technique, peptides encoded by randomised genes are displayed on the surface of filamentous bacteriophage through fusion to the outer coat PIII (Scott and Smith, 1990, Cheng *et al.*, & Danielsen *et al.*, 2001) and PVIII (Felici *et al.*, 1991 & Petrenko *et al.*, 2002) proteins. Phage encoding interacting proteins are selected by several rounds of affinity purification, followed by amplification of the bound phage by re-infection of *E. coli* cells. Increased stringency of selection during the latter selections leads to the identification of tightly interacting proteins which are subsequently identified by sequence analysis of the isolated phage. Phage display is perhaps the most widely adopted technique for the display of combinatorial libraries. Commonly used in the generation of antibodies for use as diagnostic and therapeutic molecules (Mao *et al.*, 1999, Barbas *et al.*, 1992 and reviewed in Rader and Barbas, 1997), phage display libraries have also been used in epitope mapping (Scala *et al.*, 1999), antagonist identification (Röttgen and Collins, 1995 & Chirinos-Rojas *et al.*, 1998), the generation of enzymes with improved or extended functionality (Avalle *et al.*, 1997 & Danielsen *et al.*, 2001), the selection of high affinity peptide hormone variants (Lowman and Wells, 1993) and the identification and design of novel DNA binding proteins (Rebar and Pabo, 1994, Cheng *et al.*, 1996 & Wolfe *et al.*, 1999).

Phage display can typically be used to generate combinatorial libraries containing 10^9 – 10^{10} clones (Noren and Noren, 2001). Selection from phage display libraries is straight forward and relatively fast, although repeated rounds of selection and re-infection may be time-consuming. The technique also relies upon infection of prokaryotic host cells which may prohibit the expression of proteins toxic to these cells and also proteins which require post translational modification by eukaryotic hosts for their activity. Selection of interacting phage relies upon affinity purification which may also preclude the use of this technique in the selection of mutant enzymes. In addition, four to five copies of the PIII coat protein and thousands of copies of the PVIII protein are present on the phage coat (Zwick *et al.*, 1998). Avidity effects caused by the multivalency of displayed proteins may therefore affect the affinity selection of the displayed proteins, although the use of monovalent phage display systems (Röttgen and Collins, 1995, Lowman and Wells, 1993, Collins *et al.*, 2001) overcomes this problem.

1.2.2 *In Vitro* and *In Vivo* Display Technologies

Display technologies have recently been developed to facilitate *in vitro* expression and display of combinatorial libraries. *In vitro* expression eliminates the need for vector transformation into host cells, circumventing the problems associated with the display of potentially toxic proteins. In addition, larger libraries can be displayed using these technologies, with libraries containing 10^{12} – 10^{13} members being reported (Mattheakis *et al.*, 1994, Cho *et al.*, 2000). *In vitro* display can be achieved using several methods including, ribosome display in which the translated peptide remains associated with the corresponding mRNA in the polysome complex (Hanes *et al.*, 1998, He and Taussig, 1997), or using mRNA display in which the addition of the tRNA mimic puromycin to the 3' end of the encoding mRNA, results in the covalent attachment of the mRNA to the translated protein (Roberts and Szostack, 1997, He and Taussig, 1997, Hanes *et al.*, 1998). Proteins selected from mRNA and ribosome display libraries are identified after reverse transcription of the encoding mRNA and sequencing of the generated cDNA products. Applications of *in vitro* display are reviewed in Takahashi *et al.*, (2003). Although a powerful technique, *in vitro* display is usually limited to the display of proteins with relatively short chain lengths (approximately 100 amino acids) as larger proteins form fusion products with somewhat reduced affinity (Takahashi *et al.*, 2003).

In addition, as the protein identity is mRNA encoded, care must be taken to ensure that selection or panning of the library is undertaken in a RNase free environment.

In libraries in which *in vivo* expression of the target protein is required, display technologies such as the peptides on plasmids approach may be used. In this technique, randomised proteins are expressed as fusion proteins with DNA binding proteins such as the *lac* repressor (Cull *et al.*, 1992) or the nuclear factor κB p50 (Speight *et al.*, 2001). The inclusion of target sites for the DNA binding moiety of the fusion protein in the plasmid, allows expressed proteins to bind to the encoding plasmid *in vivo*. As this genotypic-phenotypic linkage is maintained after cell lysis, *in vitro* selection of proteins expressed *in vivo* can still be performed.

Alternatively, microbial cell surface display systems may be used, in which proteins can be displayed on the surface of yeast or bacterial cells by fusing them with a carrier protein or anchoring motif (reviewed in Lee *et al.*, 2003). The advantages of microbial cell surface display in vaccine development and in the development of biocatalysts are easily appreciated. In addition, ligand binding by peptides displayed in this fashion can be easily quantified by flow cytometry by incorporation of fluorescent labels (Wittrup, 2001). Selection from combinatorial libraries may also be performed *in vivo* using hybrid systems and reporter gene assays, to study both protein-DNA and protein-protein interactions, the use of, both yeast (Cheng *et al.*, 1997) and bacterial (Joung *et al.*, 2000) hybrid systems have been reported.

1.3 Combinatorial Protein Libraries and Randomisation

Randomised gene libraries are created by the defined introduction of mutations to the coding sequence of a target gene. The introduction of mutations, or randomisation of the target gene may be accomplished using several techniques. Although many such techniques have been developed, these techniques can be broadly defined as falling into one of the two general formats or "strategies" described below. The most appropriate randomisation strategy is usually defined by the existing knowledge of the protein encoded by the target gene. As with display technologies, employment of the most appropriate randomisation strategy is crucial to the success of combinatorial experimentation.

1.3.1 Random Mutagenesis

Random mutagenesis is usually applied in the study of proteins for which a structure/function relationship model has not been established and is often used to identify the roles of individual amino acids and to help establish such a model. The technique relies upon the introduction of point mutations at random intervals along the sequence of an entire gene. Although mutations are introduced at random locations, the frequency of mutation is experimentally defined. Mutation frequencies that are too high result in the accumulation of multiple mutations in a single gene resulting in a library of genes bearing numerous mutations. The fraction of functional mutants in such a population will be low (Shafikani *et al.*, 1997). Mutation frequencies which are too low may result in a large background of wild type genes or fail to identify mutations which affect the function of the protein being studied.

Random mutagenesis is usually achieved *in vitro,* using error prone PCR (Shafikani *et al.*, 1997, Xu *et al.*, 1999, Doi and Yanagawa, 1999) although *in vivo* random mutagenesis has been reported (Fabret *et al.*, 2000). Error prone PCR relies upon the amplification of target genes in low fidelity PCR reactions. The frequency of mutation is controlled by varying the concentrations of mutagenic components such as, manganese and dITP (Xu *et al.*, 1999) within the reaction, or by using imbalanced concentrations of dNTPs (Shafikani *et al.*, 1997). Although error prone PCR has been

used routinely to identify the roles of amino acids within uncharacterised proteins, it may also be applied in the creation of novel proteins (Doi and Yanagawa, 1999) and in the process of affinity maturation, in which further mutations are introduced to proteins selected from combinatorial libraries in an attempt to further improve their function (Daugherty *et* al 2000 & Mössner and Plückthun, 2001).

1.3.2 Targeted Randomisation

Perhaps the most widely adopted strategy used in the creation of randomised gene libraries is the use of targeted randomisation. Targeted randomisation is used to directly replace specific codons within a target gene. At the extreme end of the spectrum, randomisation of all codons of a target gene generates libraries of completely randomised peptides. Although no pre-determined function or structure can be attributed to these random peptides, the screening of such libraries has been successfully employed in, the identification of antibody ligands (Felici *et al.*, 1991, Cull *et al.*, 1991, Mattheakis *et al.*,1994) including the identification of HIV-specific immunogenic epitopes (Scala *et al.*, 1998), cytokine and calcium binding protein antagonist generation (Chirinos-Rojas *et al.*, 1998 & Pierce *et al.*, 1998) and the identification of DNA binding motifs (Cheng *et al.*, 1996).

More frequently targeted randomisation may be viewed as a type of site directed mutagenesis in which only specific codons within the target gene are replaced. Randomisation in this fashion is particularly applicable in the mutagenesis of well characterised proteins, where the replacement of specific amino acid residues can be predicted to alter the structure or function of the protein. This approach has been applied extensively in protein engineering with remarkable success. Examples demonstrating the diverse applications of targeted randomisation are detailed below.

The randomisation of the complementarity determining region (CDR) of antibody molecules and antibody domains has been used to generate combinatorial libraries for the selection of high affinity antibodies (Barbas III *et al.*, 1992, Reiter *et al.*, 1999, Knappik *et al.*, 2000), for diagnostic, therapeutic and biocatalytic use. As the randomised target gene encodes human antibody domains, the therapeutic use of these

antibodies avoids the human antimurine antibody response associated with monoclonal antibodies generated by conventional hybridoma technology (Mao *et al.*, 1999). The generation of very high affinity antibodies against HIV-1 (Yang *et al.*, 1995) and cytomegalovirus (Pini *et al.*, 1997) demonstrate the potential therapeutic application of such proteins.

The generation of small antibody like proteins has also been achieved by the randomisation of scaffold proteins such as the fibronectin type III domain (Koide *et al.*, 1998), and lipocalins (Beste *et al.*, 1999 & Schlehuber and Skerra, 2002). The generation of these ligand binding domains from small scaffold proteins not only demonstrates the utility of combinatorial randomisation, but has also provided the first framework which permits the specific complexation of small molecules such as metabolic compounds (Beste *et al.*, 1999).

Randomisation of key amino acid residues in target enzymes has resulted in the selection of enzymes with new specificities for industrial uses (Danielsen *et al.*, 2001). Randomisation of enzymes in this fashion has also been used to identify interacting residues of biologically important enzymes such as β-lactamases (Cantu III *et al.*, 1996, Avelle *et al.*, 1997 & Gaytan *et al.*, 2002) glycinamide ribonucleotide transformylase (Warren and Benkovic, 1997) and have been used to generate mutant enzyme cofactors to facilitate electron transfer studies (Robles and Youvan, 1993).

In addition to the identification of key amino acid residues in protein function, targeted randomisation has also been used to probe the determinants of protein function. Limited randomisation, in which substituted residues are replaced with only a subset of amino acids (such as the polar or hydrophobic amino acids) has been applied to the identification of structurally important residues in target proteins (Jeffery and Koshland Jr., 1999), and used to randomise scaffold proteins in order to study and establish general patterns involved in protein folding (Kamtekar *et al.*, 1993, Lahr *et al.*, 1999, Petrenko *et al.*, 2002).

The use of targeted randomisation and combinatorial libraries has also identified a number of other biologically important proteins, including trypsin inhibitors (Röttgen and Collins, 1995), novel cytokine variants, with improved pharmacological properties

(Klein *et al.*, 1999), ligands which promote receptor mediated endocytosis in mammalian cells (Legendre and Fastrez, 2002), RNA binding peptides (Barrick *et al.*, 2001) and acid stable green fluorescent protein (GFP) (Sawano and Miawaki, 2000).

Perhaps the most promising application of targeted randomisation has been the generation of novel sequence specific DNA binding proteins based upon Cys_2 His_2 zinc finger frameworks (for review see Choo and Islan, 2002) suggesting that these techniques may hold the key to designing artificial transcription factors and gene silencing elements at will. As the randomisation of Cys_2 His_2 zinc fingers was used as a model in the present study, the targeted randomisation of these proteins is discussed in more detail in section (1.11.1).

1.4 The Randomisation Process: Technicalities and Problems.

The potential power of randomised gene libraries, evident in their successful application in the examples in section 1.3.3, is derived from the ability of such combinatorial techniques to identify biologically important proteins from large mixtures of diverse protein species, a process termed deconvolution. It may therefore be assumed that an increase in the diversity of expressed proteins may consequently increase the possibility of identifying proteins with desired abilities from within that mixture. Thus any limitation imposed upon library diversity at the protein level, has the potential to significantly affect the subsequent deconvolution of that library.

Library diversity at the protein level is often described in terms of the representation of each potential protein within that library. For example a library which is under-represented will not be physically large enough to contain all the theoretically possible randomised proteins. As such, potentially tightly interacting proteins may not be present in the library population during the deconvolution process. Alternatively a library in which randomised proteins are present in disproportionate concentrations may be described as misrepresentative. This misrepresentation may also interfere with the deconvolution process, for example affinity based selection which is governed by the laws of mass action, may favour the selection of proteins whose representation in the library is greatest, even though their true affinity may be lower than that of less abundant proteins (Choo and Klug, 1994b).

Thus any factors which affect representation at the protein or amino acid level of a randomised library may consequently affect library diversity and as such interfere with the deconvolution process. Representation may be affected by several diverse factors such as the negative selection of toxic proteins in libraries generated by *in vivo* expression (Rottgen and Collins, 1995 & Scott and Smith, 1990) or the persistent reoccurrence of the "wild type" gene within the library population (Warren and Benkovic, 1997). However perhaps the greatest potential problems in the creation of a representative protein library and the greatest limitations upon library diversity occur at the genetic level, during the randomisation process itself and predominately result from the inherent degeneracy of the genetic code.

1.4.1 Randomisation and the Degeneracy of the Genetic Code

The degeneracy of the genetic code does affect libraries generated by random mutagenesis, as a degenerate code is utilised in nature to minimise the harmful effects of random mutations (Eisinger and Trumpower, 1996). As a result of this a significant fraction of mutations introduced during random mutagenesis will encode silent mutations or conservative amino acid changes (Shafikhani *et al.*, 1997). However it is in the creation of gene libraries by targeted mutagenesis that this degeneracy poses the greatest problem.

Targeted randomisation is usually achieved by replacing a section of DNA from the parental gene with synthetic DNA in which specific codons have been randomised. This may be achieved using cassette mutagenesis or PCR amplification using degenerate primers. In both techniques, the identities of the randomised amino acids are encoded synthetically. At each position of randomisation, codons representing all possible amino acid substitutions must be generated. Since the genetic code is degenerate, this leads to problems both with library size and unequal representation of amino acids.

1.4.2 Degeneracy and Library Size

In order to encode all 20 possible amino acid substitutions at a randomised position, a synthetic oligonucleotide may be randomised with the codon NNN, in which N represents any one of the four possible nucleotides A, C, G or T. This would be expected to generate oligonucleotides containing each of the possible 64 codon possibilities at the position of randomisation. As the code is degenerate, 41 of these 64 codons are redundant (which also results in the unequal representation of amino acids) and three of the codons encode termination codons, which will result in the generation of truncated proteins. The use of these oligonucleotides in the complete randomisation of a single amino acid in a target protein, would therefore generate a gene library of 64 individual genes encoding only the 20 possible target proteins, a gene : protein ratio of 3.2 : 1. As 64 codons are required at each position of randomisation, the ratio of genes to proteins increases exponentially as the number of randomised positions is increased.

Complete randomisation of two amino acids for example, requires 4096 genes to encode the 400 possible randomised proteins, a gene : protein ratio of 10.24 : 1. Due to the exponential rise in the ratio of genes : proteins, the practical limits of transformation of a phage display library are reached after the randomisation of only five codons (1.07 X 10^9 genes), however despite the large genetic diversity such a library will only contain 3.2 X 10^6 different proteins.

Several mutagenesis strategies have been developed to try and address this limitation on library diversity imposed by the degeneracy of the genetic code. Complete randomisation of a target amino acid may be achieved using the randomised codon $NN^G/_T$ (Scott and Smith, 1990, Gunneriusson *et al.*, 1999, Petrenko *et al.*, 2002) or $NN^G/_C$ (Reidhaar-Olson and Sauer, 1988, Jamieson *et al.*, 1994, Parikh and Guengerich, 1997). Randomisation with either codon results in the generation of 11 redundant codons and one termination codon, however protein truncation by this termination codon can be avoided by expression in an amber suppressor strain of bacterial cells(Söderlind *et al.*, 1995).

With a total of 32 possible randomised codons the ratio of genes : proteins is reduced to 1.6 : 1 when using the codon $NN^G/_T$ or $NN^G/_C$ to randomise a single amino acid. However as some degeneracy still exists, this ratio still increases exponentially as the number of randomised positions is increased. Using these codons, the practical limits of transformation of a phage display library are reached after the randomisation of six amino acid residues (1.07 X 10^9 genes), although diversity at the protein level is limited to only 6.4 X 10^7 species. Thus the use of degenerate codons to encode randomised amino acids results in the limitation of library diversity being reached at the genetic level long before the limit of protein diversity is approached. This may result in the generation of a library which is physically too small to contain all possible protein variants when a number of positions are randomised and this limited diversity may adversely effect the deconvolution of such a library (Scott and Smith, 1990 & Gunneriusson *et al.*, 1999).

Technical constraints imposed by library size therefore limit the number of positions which can be completely randomised to approximately six codons (Lowman and Wells, 1993). Randomisation techniques have been developed to overcome this limitation,

several of which rely upon the combination of libraries randomised independently of each other to increase library diversity. These techniques usually involve the generation of subset libraries in which a number of codons are randomised independently within each library subset. Selection is performed on these subset libraries to reduce their complexity, and the subset libraries combined using recombination (Collins *et al.*, 2001) or DNA shuffling techniques (Kitamura *et al.*, 2002 & Matsuura *et al.*, 2002) to generate libraries of increased diversity.

Alternatively the process of affinity maturation may be used to overcome the limitation of library size upon the selection of highly interacting proteins. Affinity maturation relies upon the assumption that the accumulation of favourable mutations within a target protein will have additive affects. This may be achieved sequentially by the introduction of further mutations to a randomised protein after its selection from a combinatorial library (Gunneriusson *et al.*, 1999 & Schlehuber *et al.*, 2000), or in a parallel manner with the use of subset libraries followed by the combination of pre-selected protein domains (Lowman and Wells, 1993 & Yang *et al.*, 1995). Although powerful, affinity maturation is still constrained by the limitations imposed on library size, as it is crucial in this process to create an initial library with maximal functional diversity (Schlehuber and Skerra, 2002). However in order to obtain a completely randomised library with maximal functional diversity the degeneracy of the genetic code must be overcome to achieve the ideal genes : proteins ratio of 1 : 1. This would ensure that the only limitation imposed upon diversity at the protein level was the physical size of the generated library.

1.4.3 Degeneracy and amino acid representation,

In addition to imposing constraints upon library size, the degeneracy of the genetic code may also affect the representation of individual amino acids in a combinatorial protein library. As amino acids are encoded disproportionately by the genetic code, an amino acid bias may be reflected in libraries generated by randomisation techniques which necessitate the cloning of redundant codons. Table 1.0 shows the representation of amino acids by codons typically used in the creation of randomised gene libraries.

AMINO ACID	REPRESENTATION WITH NNN	REPRESENTATION WITH $NN^{G}/_{T}$ OR $NN^{G}/_{C}$
ALA	4	2
CYS	2	1
ASP	2	1
GLU	2	1
PHE	2	1
GLY	4	2
HIS	2	1
ILE	3	1
LYS	3	1
LEU	6	3
MET	1	1
ASN	2	1
PRO	4	2
GLN	2	1
ARG	6	3
SER	6	3
THR	4	2
VAL	4	2
TRP	1	1
TYR	2	1

Table 1.0 Representation of amino acids when randomisation is carried out using the codons NNN $NN^{G}/_{T}$ and $NN^{G}/_{C}$.

As the randomisation strategies shown in Table 1.0 necessitate the cloning of redundant codons, amino acids encoded by the most degenerate codons will be over-represented in libraries generated in this fashion. As an example, when the codon NNN is employed in randomisation, the amino acids serine and leucine are represented by six codons, whereas the amino acids tryptophan and methionine are represented only once.

Libraries generated using the codon NNN will therefore contain a 6 : 1 ratio of the most degenerately encoded amino acids : least degenerately encoded, for example the ratio of serine : tryptophan residues, or leucine : methionine residues. The use of the codon $NN^{G}/_{T}$ or $NN^{G}/_{C}$ limits redundancy reducing this ratio to 3 : 1. Although this ratio may appear inconsequential, it only reflects the disparity in the total amount of each of the respective residues contained within the library. As the degeneracy of the code is reflected at each position of randomisation, the effect of this bias on the protein representation of libraries randomised at multiple positions may be profound.
This can be demonstrated by comparing the theoretical distribution of genes containing a degenerately encoded amino acid, such as serine, with the distribution of genes containing an amino acid encoded only once such as tryptophan within a calculated library population (Table 1.1). An explanation of the calculation of this binomial distribution can be found in appendix (A3)

No. of Randomised Positions Encoding the Target Amino Acid	**Theoretical Distribution of Genes (NNN)**		**Theoretical Distribution of Genes ($NN^{G}/_{C}$ or $NN^{G}/_{T}$)**	
	Serine	**Tryptophan**	**Serine**	**Tryptophan**
6 Positions	46656	1	729	1
5 Positions	2706048	378	42282	186
4 positions	65396160	59535	1021815	14415
3 Positions	8.4×10^{8}	5.0×10^{6}	1.3×10^{7}	5.9×10^{5}
2 Positions	6.1×10^{9}	2.3×10^{8}	9.5×10^{7}	1.3×10^{7}
1 Position	2.3×10^{10}	5.9×10^{9}	3.6×10^{8}	1.7×10^{8}
0 Positions	3.8×10^{10}	6.2×10^{10}	5.9×10^{8}	8.8×10^{8}

Table 1.1 An example of the theoretical distribution of genes encoding one of the most degenerately encoded amino acids (serine) and one of the least degenerate amino acids (tryptophan) when six positions are randomised using the codons NNN, $NN^{G}/_{T}$ or $NN^{G}/_{C}$.

As demonstrated in Table 1.1, the cloning of degenerate codons during the randomisation procedure can be expected to have profound results upon the relative protein concentrations of the library. Although rarely discussed in the literature, (with the exception of Reidhaar-Olsen and Sauer, 1998 who discuss the under representation of the amino acids histidine, asparagine and lysine, encoded only once using the codon $NN^{G}/_{C}$, despite prior knowledge that these are acceptable mutations at the randomised positions) such profound differences in relative protein concentrations must be assumed to affect the deconvolution of libraries generated in this fashion. As with the problems associated with library size, the degeneracy of the genetic code must be overcome to establish a 1 : 1 ratio of codons : amino acids in order to remove this inherent bias from randomised libraries.

1.5 Further Problems Associated with Conventionally Randomised Codons

In addition to the problems associated with the inherent degeneracy of the genetic code, the codons NNN, $NN^{G}/_{C}$ and $NN^{G}/_{T}$ are generated synthetically from equimolar mixtures of the four nucleotides A,C,G and T during oligonucleotide synthesis. Thus the generation of a truly randomised codon relies upon the stoichiometric coupling of each of the four possible nucleotides as the synthetic oligonucleotide is extended. Differing coupling efficiencies and molar ratios of the four phosphoramidite precursors may therefore result in further bias being introduced into the library. To ensure an equimolar distribution of the four nucleotides at positions of randomisation a biased mixture of phosphoramidites can be added during the synthesis of randomised oligonucleotides to compensate for the differing coupling efficiencies of each phosphoramidite (Ho *et al.*, 1996). However the appearance of excluded nucleotides at the terminal position of codons in libraries randomised using the codons $NN^{G}/_{C}$ and $NN^{G}/_{T}$ (Rötgen and Collins, 1995, Söderlind *et al.*, 1995, Rungpragayphan *et al.*, 2003) highlights the difficulty in obtaining truly randomised codons during the synthesis of randomised oligonucleotides.

1.6 Current Strategies Used to Address the Problems of Randomisation.

The use of subset libraries and processes such as affinity maturation still predominantly rely upon conventional randomisation techniques to generate libraries of genes containing randomised codons. Recent developments in the creation of randomised gene libraries have targeted the randomisation process itself, as a means of addressing the problems associated with the generation of representative libraries. This is often achieved using "limited randomisation", a process in which only a limited number of substitutions are permitted at positions of randomisation. Limited randomisation can be employed not only to reduce library size, but can also be used to generate "Patterned Libraries" in which substitution is limited to only certain classes of amino acid.

As a means of reducing library size, codon substitution may be limited by excluding those codons which would be expected to generate unfavourable mutations based upon prior structural knowledge of the target protein. This has been achieved by simply limiting the nucleotide possibilities at each base of the randomised codon during synthesis (Greisman and Pabo, 1997, Doi and Yanagawa, 1999, Frenkel *et al.*, 2000 & Barrick *et al.*, 2001). Using this technique the number of codon substitutions is governed by the limitations imposed upon the individual nucleotides during synthesis. The codon $TN^{G}/_{C}$ (Frenkel *et al.*, 2000) for example allows only five amino acid substitutions (Phe, Leu, Ser, Tyr and Trp) at randomised positions. The codon (A/C/G)N(G/C) (Greisman and Pabo, 1997) encodes 16 amino acids excluding only the codons for Cys, phe, tyr and trp. Limiting codons in this fashion provides a simple means of reducing library size, however as codons are limited by the omission of certain nucleotides during synthesis, only codons which are similar in sequence can be excluded from the generated library. In addition the exclusion of certain codons based solely on the assumption that the encoded amino acids will generate unfavourable mutations may unintentionally exclude highly interacting proteins from the generated library. This is highlighted by the appearance of a specified aromatic residue in the consensus sequence of a number of strong binding zinc fingers (Joung *et al.*, 2000), residues which had been excluded from the corresponding phage display libraries (Greisman and Pabo, 1997).

Patterned libraries, in which amino acid substitutions conform to a certain class, are often constructed as a means of studying proteins structure/function relationships by creating libraries which mimic natural protein folds. This may be achieved by limiting amino acid substitutions to alternating patterns of polar and non polar residues (Jeffery *et* al 1999, Kamtekar *et* al 2000, Larsson *et al.*, 2002) or by mimicking codon distribution in known protein folds (Cho *et al.*, 2000). To achieve randomisation in this fashion, the ratios of each individual nucleotide in a randomised codon are calculated to ensure a high probability that the desired codons will be generated during oligonucleotide synthesis. The nucleotide bias in each codon required to minimise the occurrence of termination codons within the randomised gene may also be calculated.

Although this technique is capable of generating libraries which show good correlation with the sequences of known protein folds, the technique is complicated and often requires several rounds of optimisation to achieve the correct nucleotide bias (Jeffery *et al.*, 1999). The main limitation of this approach is that the similarity of the base composition of some codons prevents their exclusion/inclusion in the generated libraries. This limit on flexibility results in the unwanted incorporation of termination codons and other amino acids, or the partial loss of some amino acids from the library (Larsson *et al.*, 2002).

A more direct form of control over the identities of randomised codons can be achieved by employing a "resin splitting" technique during the synthesis of randomised oligonucleotides. This technique involves the redistribution of the resin support containing the randomised oligonucleotide between two or more synthesis columns during the creation of the randomised oligonucleotide. Although complicated and time consuming, the strength of this technique is the ability to control the identities of randomised codons.

Such control has been used to limit the degree of randomisation in target genes, by distributing the resin between columns in which the wild type and randomised codons are individually synthesised (Cormack and Struhl, 1993). The technique also affords the ability to generate single codon replacements at a number of positions within a target gene (Pakula and Simon, 1992, Chatellier *et al.*, 1995). Applied to less limited randomisation strategies, this resin splitting approach has been used to generate

patterned libraries in which all but one of the randomised codons were limited to hydrophilic residues (Lahr *et al.*, 1999).

Combined with the use of dinucleotide phosphoramidites, resin splitting has been applied in the generation of completely randomised libraries (Neuner *et al.*, 1998). The combination of resin splitting prior to the joining of pre-synthesised dimer phosphoramidites was used to generate 20 randomised codons at randomised positions during oligonucleotide synthesis. Similar to the use of trinucleotide phosphoramidites, this technique can be used to exclude termination codons from randomised positions, reduce library size and address the problem of bias within generated libraries. Although dinucleotide phosphoramidites show higher coupling efficiency than trinucleotides (Neuner *et al.*, 1998), the coupling efficiencies of these dinucleotides differs, requiring the empirical determination of the relative concentrations of each dinucleotide, to avoid their overrepresentation within the generated library. Thus the involved nature of the technique suggests it would be difficult to apply to multiple positions of randomisation.

The use of trinucleotide phosphoramidites during the synthesis of randomised oligonucleotides (Virenkäs *et al.*, 1994, Kayusihin *et al.*, 1996, Bruanagel and little, 1997) is currently the only randomisation technique able to afford full control over codon identity at positions of randomisation. Using this technique, randomised codons are synthesised as trinucleotide "building blocks" which can be employed in standard oligonucleotide synthesis to generate randomised oligonucleotides. These trinucleotide building blocks can be used to generate standard libraries in which the ratio of codons : amino acids is reduced to 1 : 1 (Virenkäs *et al.*, 1994, Kayusihin *et al.*, 1996, & Bruanagel and little, 1997) consequently eliminating termination codons and bias from generated libraries. As any combination of trinucleotides can be employed during the synthesis of randomised oligonucleotides, codon identity at positions of randomisation can be completely defined in accordance with any randomisation strategy, facilitating the production of libraries in which certain codons are excluded from randomised positions, to mimic the amino acid distribution of known protein folds (Knappik *et al.*, 2000). In addition biased ratios of trinucleotide phosphoramidites can be employed during oligonucleotide synthesis to control the probable distribution of chosen codons at positions of randomisation (Matsuura *et al.*, 2002).

The use of trinucleotide phosphoramidites produces a powerful randomisation methodology capable of controlling the identity of randomised codons, excluding termination codons and controlling library size and bias by the establishment of a 1 : 1 ratio of codons : amino acids. However the technique is involved and not without associated problems. Low coupling efficiency and varied coupling efficiency of these trinucleotides has been reported (Virenkäs *et al.*, 1994, Kayusihin *et al.*, 1996). In addition trinucleotide phosphoramidites are not yet commercially available necessitating complicated synthesis of these building blocks prior to experimentation.

1.7 The Max Randomisation Technique

1.7.1 Introduction

The MAX randomisation technique was designed to produce, by means of standard oligonucleotide synthesis, a randomisation technique designed for use with all current display technologies, in which each possible amino acid is encoded by only one randomised codon. The establishment of a 1 : 1 ratio of codons : amino acids during the randomisation process effectively eliminates the cloning of redundant codons, facilitating the production of gene libraries in which the genes : proteins ratio is maintained at 1 : 1 irrespective of the number of randomised positions. As each amino acid is encoded by a single codon, the technique should also address the problem of amino acid bias within libraries, as each possible amino acid will be encoded equally in sequences randomised using MAX methodology.

In addition, as the randomisation procedure relies upon a "selection" procedure to encode randomised positions, the technique also enables the creation of "Designer libraries" in which only amino acids defined by the user are contained at each position of randomisation. In the design of the MAX technique this ability to create "designer libraries" was considered an important extension to the utility of randomised gene libraries.

This ability to completely define the amino acids encoded at randomised positions allows the construction of limited libraries in which no amino acids are unintentionally excluded (See Section 1.6). In addition as codon identity at positions of randomisation can be assigned arbitrarily using this technique, limited libraries can be constructed in which groups of unrelated amino acids are included at positions of randomisation (for example where several different favourable substitutions have been identified by structure function relationship studies). In the work following the current study, the ability to assign codons at random is to be used to establish a logical connection between selected zinc finger proteins and the randomised gene which encoded them.

1.7.2 Introduction to the Technique

The MAX randomisation technique employs a "Selectional hybridisation" procedure to ensure that the amino acids contained at each position of randomisation are only encoded by a single codon, with the exclusion of all termination codons. Specific codon assignment is not fundamental to the randomisation process and as such each amino acid can be arbitrarily assigned a single randomised codon. As libraries generated in the current study are intended for use with an *in vivo* expression system, randomised codons were assigned to each amino acid using the most abundant codon for that particular amino acid in the most highly expressed genes of *E. coli* (Nakamura *et al.*, 2000). These codons, termed MAX codons in the following experimentation, are shown in Table (1.2).

Selected Amino Acid	MAX Codon (In *E. Coli*)
ALA (A)	GCG
CYS (C)	TGC
ASP (D)	GAT
GLU (E)	GAA
PHE (F)	TTT
GLY (G)	GGC
HIS (H)	CAT
ILE (I)	ATT
LYS (K)	AAA
LEU (L)	CTG
MET (M)	ATG
ASN (N)	AAC
PRO (P)	CCG
GLN (Q)	CAG
ARG (R)	CGC
SER (S)	AGC
THR (T)	ACC
VAL (V)	GTG
TRP (W)	TGG
TYR (Y)	TAT

Table 1.2. The selected "MAX" codons for each of the twenty amino acids employed in the MAX randomisation technique.

1.7.3 Selectional Hybridisation

The selectional hybridisation process is used to generate randomised synthetic DNA cassettes, which can be employed in the cassette mutagenesis of target genes. The hybridisation process, demonstrated graphically in figure 1.1 relies upon the selection of single stranded oligonucleotides, from pools of oligonucleotides encoding the twenty possible amino acid substitutions.

This is achieved using a single template oligonucleotide (Fig 1.1) which encodes one strand of the excised sequence of the target gene, and is conventionally randomised using the codon NNN at the designated positions of randomisation. Double stranded DNA is created by the hybridisation of selection oligonucleotides to this template strand.

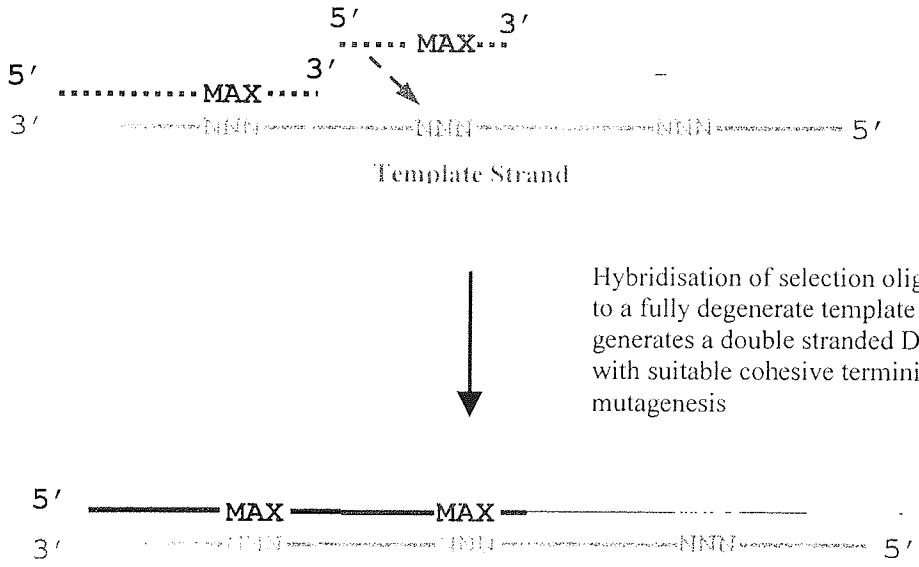


Fig 1.1 Schematic representation demonstrating how MAX randomised double stranded DNA cassettes can be selected from a conventionally randomised template by the hybridisation of individual selection oligonucleotides during the selectional hybridisation procedure. Complementary DNA strands are shown as straight lines, the term MAX is used to represent any one of the 20 MAX codons listed in table 1.2, the template strand is shown in grey and the selection oligonucleotides are shown in blue, red and green.

Since in the current study, the template strand contains three positions of randomisation (α, β and γ) three separate pools of selection oligonucleotides are employed. These pools of selection oligonucleotides each complement a consecutive region of the conserved sequence of the template strand (Fig. 1.2). The remaining three base pair

region of each selection oligonucleotide, which correspond to the randomised positions of the template strand, encodes one of the twenty possible MAX codons listed in table 1.2. To achieve full randomisation, three pools containing equimolar amounts of each of the twenty selection oligonucleotides are added to the hybridisation reaction. The reaction is heated to denature any non specific interaction between the oligonucleotides and then cooled slowly to allow base pairing of the template and selection oligonucleotides. Base pairing between the complementary regions of the template and selection oligonucleotides orientates the selection oligonucleotides, allowing the region encoding the MAX codon to "select" its complementary sequence from the template strand.

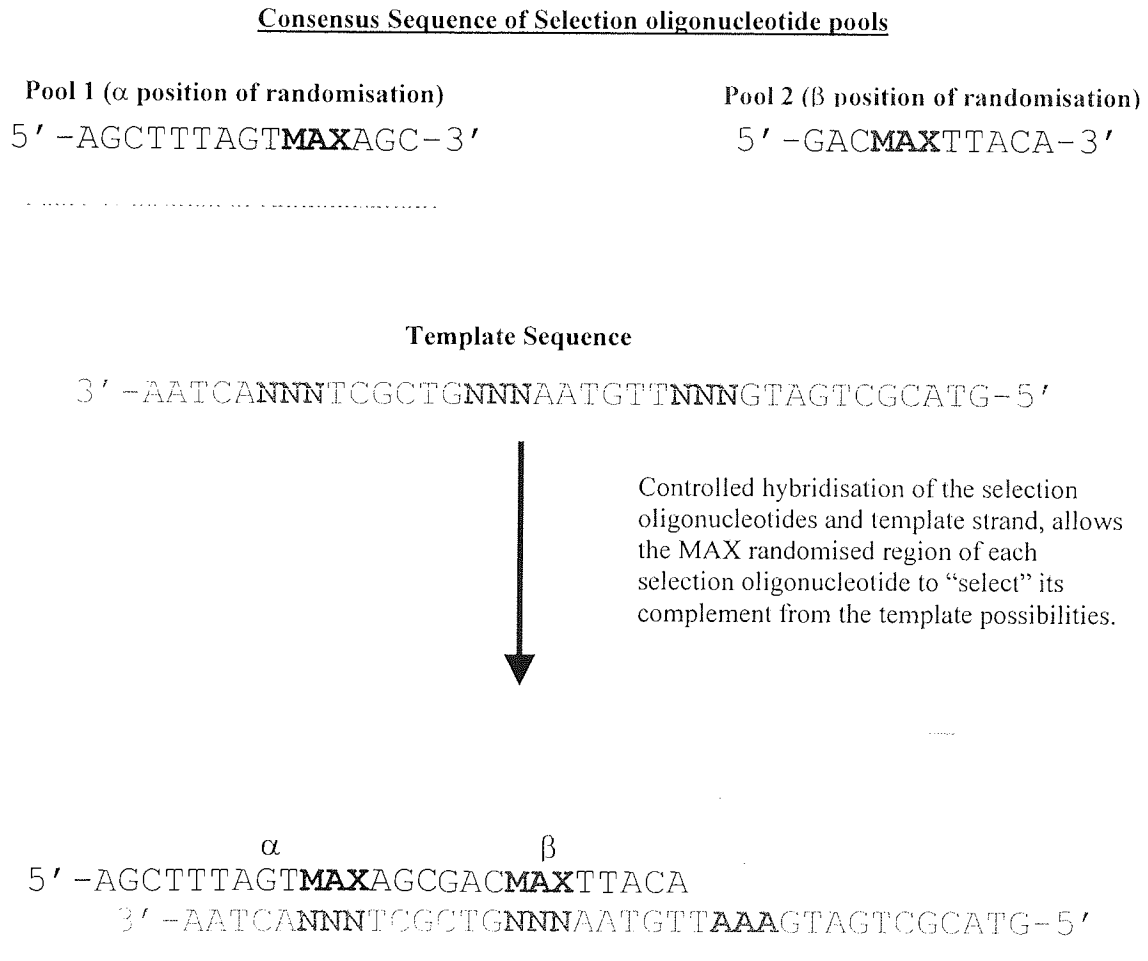


Fig 1.2 Diagram demonstrating the selection of the complementary randomised positions on the template strand by selection oligonucleotides, during selectional hybridisation. The consensus sequences of the equimolar pools of selection oligonucleotides are shown in the diagram. MAX denotes any one of the possible MAX codon sequences listed in table 1.2. The hybridisation of the γ selection oligonucleotide encoding phenylalanine (TTT) (underlined in the figure) effectively selects the complementary AAA codon from the template possibilities.

In this fashion only twenty codon possibilities are selected from the possible 64 NNN codon possibilities of the template strand. Hybridisation of three consecutive selection oligonucleotides generates a double stranded DNA cassette, which can then be employed in the cassette mutagenesis of the target protein.

As the twenty MAX codons are selected from 64 possible codons on the template strand unbound template DNA will be present in the hybridisation reactions. However as demonstrated in Figure 1.1b, the binding of the α and γ selection oligonucleotides to the template DNA generates the restriction termini necessary for the cassette mutagenesis of the target gene. Template DNA which is not hybridised to both α and γ selection oligonucleotides will therefore be lost at the cassette mutagenesis stage of randomisation. Cassettes in which the only the β selection oligonucleotide is not present contain a region of single stranded DNA. In the event that such cassettes maintained sufficient stability to participate in the cassette mutagenesis of the target gene, plasmids encoding the target gene would still contain a region of single stranded DNA and as such should not be established in the host cells upon cloning of the mutagenised gene. In this fashion the selectional hybridisation procedure effectively eliminates redundant codons, this not only reduces library size but should also eliminate any amino acid bias from libraries generated using the MAX randomisation procedure.

In the current study, carried out to develop and assess the MAX randomisation technique, the MAX randomisation strategy is employed in the creation of Cys_2 His_2 zinc finger libraries by the randomisation of three base contacting residues of the zinc finger protein (Section 1.10.4). Complete randomisation of the three base contacting residues generates a protein library of 8000 possible zinc finger proteins (as any of the twenty amino acid possibilities may be present at each position of randomisation). To achieve this with conventional randomisation using the codons NNN or $NN^{G}/_{T}$ would respectively generate libraries of 262144 and 32768 individual genes. In addition approximately 13% of the total number of genes in the library generated using the codon NNN would contain one or more termination codons, this figure is reduced to 9% for libraries generated using the codon $NN^{G}/_{T}$. However establishment of a 1 : 1 genes : proteins using the MAX technology reduces the number of genes to only 8000, with the exclusion of all termination codons.

The utility of the MAX randomisation technique extends further than the establishment of a 1 : 1 ratio of codons to amino acids. As randomised codons are effectively selected by the pools of selection oligonucleotides, the composition of these pools can be varied in accordance with differing randomisation strategies. As an example codons, may be randomised using selection oligonucleotides encoding only hydrophobic or polar residues. Unlike the creation of patterned libraries using conventional techniques (Section 1.6) this may be achieved without the exclusion or misrepresentation of any residue.

As any codon can be included or excluded from the selection pools, randomisation may also be limited to groups of unrelated amino acids (Hughes *et al.*, 2003). Presently such randomisation strategies may only be achieved using trinucleotide phosphoramidites, (problems associated with such specialised phosphoramidite chemistry are discussed in section 1.6). This ability to arbitrarily include any amino acid at each position of randomisation, facilitates the logical linkage of genotype and phenotype within generated libraries, enabling fast and efficient deconvolution strategies. Although not included in the current study, which concentrated upon the development of the MAX randomisation technique itself, a brief description of the proposed library deconvolution process is included below, to provide an example of such a logical linkage and further demonstrate the utility of the technique.

1.8 The Deconvolution strategy

Subsequent to the development of the MAX randomisation technique in this study, the MAX technique is to be employed in the creation of a Cys_2 His_2 zinc finger library. These zinc finger proteins are discussed in detail in section (1.9). Briefly Cys_2-His_2 zinc fingers are small protein moieties capable of sequence specific binding to a single DNA trinucleotide. DNA binding by zinc fingers is predominately mediated by three specific residues within the zinc finger protein (See section 1.10.4), which has encouraged many groups to generate libraries of zinc fingers, in which these residues are randomised, in an attempt to identify zinc finger proteins capable of binding novel DNA target sites. As several such zinc finger libraries have been constructed (Section 1.11.1), such libraries provide an excellent model to assess the utility of the MAX

randomisation technique. In addition, the successful development of the MAX randomisation technique will provide a means of generating a set of zinc finger libraries in which the identities of the randomised residues are logically defined in each library to expedite the subsequent deconvolution of the library.

The adoption of such a deconvolution strategy is enabled by the MAX randomisation process. The only other means of creating such defined libraries is the use of trinucleotide phosphoramidite chemistry (Section 1.6), as the deconvolution process requires multiple libraries, the cost and time required to synthesise each library using trinucleotide phosphoramidite chemistry would be prohibitive.

1.8.1 Library Deconvolution By Positional Fixing

The creation of a zinc finger protein library in which the three base contacting residues have been completely randomised could be easily achieved using the MAX randomisation technique, as the number of genes needed to encode such a library would be reduced to 8000 genes. However the proposed library deconvolution strategy utilises a number of smaller, defined subset libraries to identify interacting zinc fingers from within the library mixture.

Codon assignment in these subset libraries can be "Fixed" as the MAX randomisation technique affords the ability to arbitrarily include any number of randomised codons at each position of randomisation. In the deconvolution strategy libraries are created in which the identity of certain bases within the randomised codons included at a single position are fixed. As an example at the first position of randomisation (α position), twelve subset libraries are created. The first of these libraries contains only randomised codons which have an adenine at the first position of the randomised codon. The second contains only those codons with cytosine at the first position and the third and fourth libraries are limited to those codons beginning with guanine and thymine respectively. Library subsets 5 – 8 are limited at the second position of the randomised codon to the bases A, C, G and T as described above. The library subsets 9 – 12 are limited at the third position of the randomised codon in the same fashion. As these twelve libraries only cover the first position of randomisation the second and third

positions of randomisation (β and γ) are completely randomised using all 20 codon possibilities.

This position fixing is continued to create 12 libraries in which the identities of the randomised codons at the second position of randomisation are fixed as described above with positions 1 and 3 randomised with all 20 amino acids and another 12 libraries in which the third position of randomisation is fixed and positions 1and 2 are completely randomised. In total 36 libraries (12 libraries for each position of randomisation) are required to positionally fix the codon identity of the 3 positions of randomisation.

The three subset libraries are then screened individually against each of the possible 64 DNA triplet target sites, shown in microtitre format in Figure (1.3).

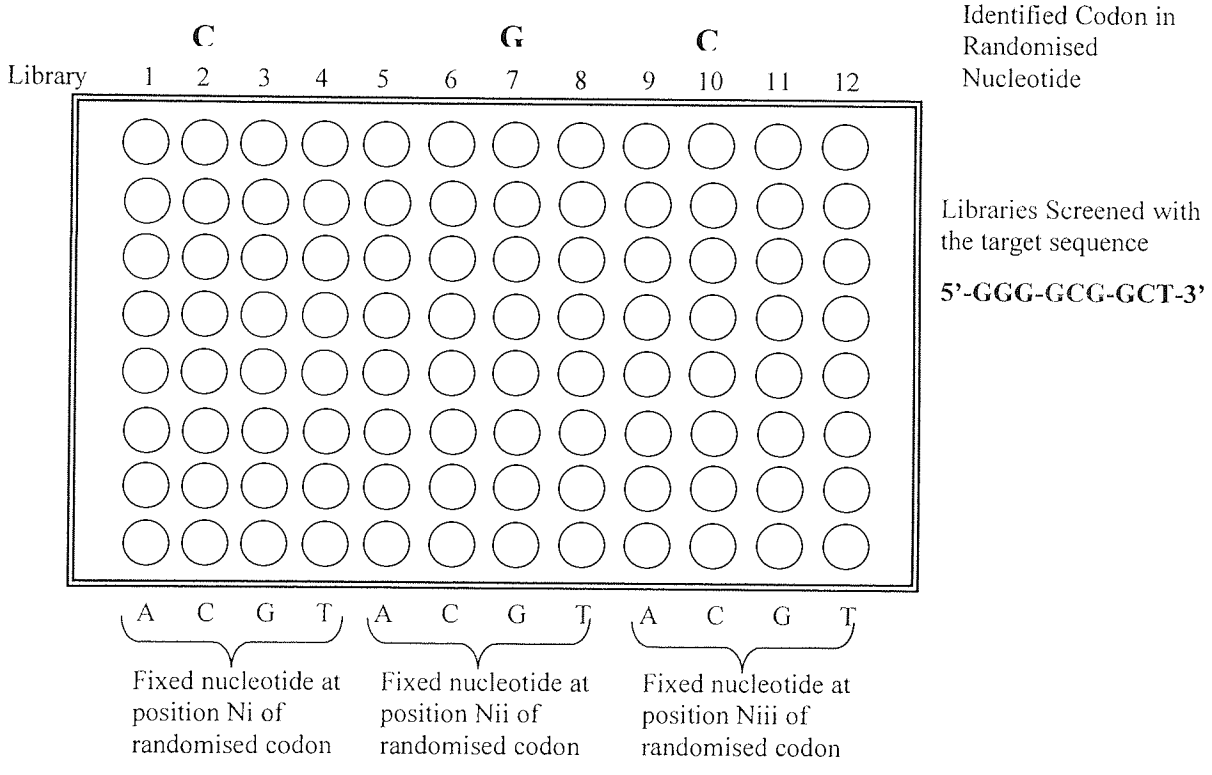


Fig 1.3 Schematic representation of library deconvolution by positional fixing. The plate shown contains libraries 1 – 12 which encompass only the α position of randomisation. In this example the target sequence of the QDR-RER-RHR protein (used in library construction; section 3.1) has been chosen to screen these libraries, from the 64 possible 5'- GGG-NNN-GCT-3' target sequences. The use of this target sequence would be expected to select for arginine at the α position of the middle finger. After washing until only three wells "light up", the identity of the most strongly interacting amino acid can be read directly from the plate. Fixing the identity of the first nucleotide as A, C, G or T in libraries 1 to 4 results in only library 2 "lighting up" as the MAX codon for arginine is CGC. A positive result in library 7 identifies the second base of the randomised codon as guanine, and the third cytosine base of the arginine codon is identified by the positive result in library 10. Thus the amino acid present at the α position can be identified as arginine. Libraries 13 – 24 and 25 – 36 can then be screened with the same target sequence to identify the interacting amino acids at the β and γ positions respectively.

As described in the legend, as the identity of each fixed position is known, the identity of the interacting residue at position α can be directly read from the plate during the assay. Screening the positionally fixed β and γ libraries with the same DNA target site identifies the interacting residues at all three positions of randomisation.

The ability to logically define the identities of the randomised bases removes the need to individually sequence the genes encoding interacting proteins, and also removes the need for multiple selection steps associated with other randomisation techniques such as phage display.

In addition wash steps, used to remove proteins which bind the target site with low affinity can be continued until only three wells "light up". Thus only the highest affinity clones will be identified, unlike conventional assays, in which the number of wash steps must be empirically determined and can result in the identification of a number of interacting clones.

1.9 Zinc Finger Proteins

1.9.1 Background

The name "zinc finger" was originally used to describe repeated protein motifs identified as constituent moieties of the transcription factor IIIA (TFIIIA), isolated from the Southern Clawed Frog *Xenopus laevis* (Miller *et al.*, 1985). These protein motifs were termed Cys_2-His_2 zinc fingers, derived from the coordination of a single zinc ion by two invariant cysteine and histidine residues in each protein motif. As these proteins are perhaps the most intensively studied of all the zinc finger proteins and as the basis of the current study, these zinc fingers are described in greater detail in section 1.10.

The term Zinc finger is now used to describe a group of related protein domains in which the three dimensional structure is stabilised by the binding of one or more zinc ions. In common with the Cys_2-His_2 class, these related zinc finger domains typically function as interaction modules, binding a wide variety of compounds such as nucleic acids, proteins and small molecules (Krishna *et al.*, 2003). This class of proteins continues to expand as new zinc finger domains have been identified in a wide variety of organisms, regulating a remarkable array of biological functions (Laity *et al.*, 2001). Some examples of these proteins are listed below to illustrate the structural and functional diversity of this class of proteins.

In addition to Cys_2-His_2 fingers, the study of various transcription factors and other proteins associated with the regulation of gene expression has identified other members of the zinc finger class. As examples, the GATA transcription factor contains zinc finger domains stabilised by the binding of a zinc ion by four conserved cysteine residues (Fox *et al.*, 1999) and CCHC zinc finger domains characterised as containing five absolutely conserved (three cysteine and two histidine) zinc binding residues have been identified in the myelin transcription factor NZF-1 (Berkovits and Berg, 1999). Zinc fingers identified as constituent moieties of transcription factors also include a number of domains which coordinate two zinc ions such as the GAL-4 DNA binding domain, a binucleate cluster in which two zinc ions are coordinated by six cysteine residues (Kraulis *et al.*, 1992), the DM domain found in transcription factors which regulate sexual differentiation, which coordinates two zinc ions via a CCHC domain

containing three cysteine and one histidine residues and a similar HCCC domain (reviewed in Laity *et al.*, 2001) and the PHD zinc finger domain found in the human Williams-Beuren syndrome transcription factor (WSTF) which coordinates two zinc ions using a Cys_4-$HisCys_3$ motif (Pascual *et al.*, 2000).

The zinc finger class also includes the familiar nuclear hormone receptor binding domains of the estrogen receptor and glucocorticoids receptor (reviewed in Schmiedeskamp and Klevit, 1993). These domains containing two zinc fingers, stabilised by the binding of a zinc ion by a tetrahedral arrangement of cysteine residues, function as the DNA binding domain of the activated receptors, binding palindromic half sites separated by an intervening region of sequence. Similar nuclear receptor binding domains activated by non-steroid hormones have also been identified (reviewed in Rastinejad, 2001). These domains are almost identical in structure and mode of DNA interaction to those of the steroid receptor elements although the DNA consensus sequence bound by these domains differs from that of the steroid receptors.

The identification of zinc finger domains is not limited to their identification within eukaryotic cells. Zinc finger motifs stabilised by the binding of a zinc ion by four conserved cysteine residues (CCCC fingers) have been identified in primase enzymes isolated from bacteriophage (Kusabe *et al.*, 1999 & Tseng *et al.*, 2000), and the CCHC zinc knuckle motif containing three cysteine and one histidine residue has been identified in all retroviral nuclear capsid proteins, with the exception of the spumaviridae (Amarasinghe *et al.*, 2000). The importance of this zinc finger motif in viral replication has led to its use as a target in the synthesis of anti HIV-1 agents (Turpin *et al.*, 1999). Prokaryotic zinc finger motifs have also been identified, including the transcriptional regulator Ros, a Cys_2His_2 zinc finger of the bacteria *Agrobacterium tumefaciens* in which the invariant loop of twelve amino acids in a classical zinc finger is replaced by a smaller nine residue loop (Chou *et al.*, 1998). The solution structure of the *Escherichia coli* chaperone protein DnaJ also identified zinc binding domains, in which zinc ions were coordinated by two cysteine and two glycine residues (Martinez-Yamout *et* al 2000).

The cysteine rich domains of the *Escherichia coli* DnaJ protein, also demonstrate the functional diversity of the zinc finger domains, as this zinc finger domain is implicated in protein binding (Martinez-Yamout *et* al 2000). This function has also been identified in other zinc finger domains such as, the zinc finger domains of members of the friend of GATA (FOG) protein (Fox *et al.*, 1999), the FOG related drosophila protein U-Shaped (Matthews *et al.*, 2000) which modulate the transcriptional activity of GATA-1, the TAZ2 (CH3) domain of the transcriptional adaptor protein CBP (De Guzman *et al.*, 2000) and the LIM domain found in the transcription factors Lin-11, Isl-1 and Mec-3 (Reviewed in Schwabe and Klug, 1994).

A zinc finger domain described as a treble clef finger has been identified in a wide variety of protein structures (reviewed in Grishin, 2001). This single domain highlights the functional diversity of the zinc finger class as its reported functions include protein binding, nucleic acid binding, small ligand binding and enzymatic catalysis (Grishin, 2001).

1.10 Cys_2-His_2 Zinc Finger Domains

1.10.1 Introduction

The first of the zinc finger classes to be identified (Miller *et al.*, 1985), the Cys_2-His_2 zinc finger domain, has been described as perhaps the most versatile nucleic acid binding protein known (Lee and Garfinkel, 2000). These domains continue to be identified in a wide variety of eukaryotic organisms and appear to be the most numerous single protein motif identified within the human genome (Miller and Pabo, 2001).

In addition to their initial discovery in TFIIIA (Miller *et al.*, 1985) these domains have been identified in a number of other transcription factors. These notably include the mouse immediate early gene protein Zif268 (Christy *et al.*, 1988), the human transcription factor Sp1 (Gidoni *et al.*, 1984) and the Wilms tumor-supressor gene product, the transcription factor WT1 (Deuel *et al.*, 1999). Such domains continue to be identified in many proteins involved in the control of gene expression, such as

transcriptional repressors (Cook *et al.*, 1999 & Hemavathy *et al.*, 2000) and proteins which mediate the effect of other proteins involved in gene expression, such as the ZBP-89 repressor which represses activation by the transcription factor Sp1 (Wieczorek *et al.*, 2000), and the RIZ (Abbondanza *et al.*, 1999) and CTIP proteins (Avram *et al.*, 2000) which mediate the effects of nuclear hormone receptor elements.

1.10.2 Function of the Cys_2His_2 Zinc Finger Domains

Novel functions of Cys_2His_2 zinc finger domains have been reported, including a role in zinc sensing, in zinc excess regulated transcription (Bird *et al.*, 2000), endonuclease activity of a single zinc finger, in the absence of zinc (Lima & Crooke, 1999) and a postulated role in cofactor recruitment by the middle finger of the transcription factor AEBP2, (mutagenesis of which abolished the repressor function of AEBP2 but did not alter the ability of the protein to bind DNA; He *et al.*, 1999).

Although these novel functions have been described, the predominant and classical function of the Cys_2His_2 zinc finger motif is the sequence specific recognition of DNA and RNA. The zinc finger domains described above, with the exception of the endonucleolytic zinc finger (generated synthetically, but which corresponds to a single finger of the human ZYF transcription factor), were isolated as constituent parts of eukaryotic transcription factors which contained several zinc finger domains. Although single fingers of these proteins displayed novel functions, the remaining Cys_2His_2 zinc finger motifs were involved in the sequence specific binding of DNA by the transcription factor.

The examples of the identification of Cys_2His_2 domains in transcription factors and nuclear hormone response elements (Section 1.10.1) demonstrates the role of these domains in sequence specific nucleic acid recognition. In addition, the identification of Cys_2His_2 domains in the bovine Adx gene product, involved in the cAMP responsiveness of other genes (Cheng *et al.*, 2000), and the RIZ protein which binds RNA and DNA/RNA hybrids (Yand *et* al., 1999) demonstrate that these domains can play an important part in all aspects of transcriptional regulation.

The importance of these Cys_2His_2 domains in transcriptional regulation is evidenced by the identification of mutations in zinc finger encoding genes in a number of proliferative disease states. Mutations in the human proto-oncogene BCL6 (which encodes a zinc finger transcriptional repressor) have been identified in large cell lymphoma (Ye *et al.*, 1993). Experiments with proteins encoded by the Evi9 gene, which is upregulated in murine myeloid leukaemia's, suggested that the zinc finger domains of the protein may be important in its oncogenic potential (Nakamura *et al.*, 2000). The Wilms' Tumour gene product WT1, provides perhaps the best example of the importance of Cys_2His_2 domains in transcriptional regulation. The WT1 gene product has been shown to arrest macrophage development through its zinc finger domain (Deuel *et al.*, 1999). Mutations of the WT1 gene product are associated with neoplastic diseases such as Wilms' Tumours and myeloid leukaemia's (Deuel *et al.*, 1999) and the developmental diseases Denys-Drash syndrome (Patek *et al.*, 1999) and Fraisier syndrome (Laity *et al.*, 2000).

1.10.3 Structure of Cys_2His_2 Zinc Finger Domains

Proteins of the Cys_2His_2 contain 28 – 30 amino acids and can be described by the consensus sequence (Tyr, Phe)-X-Cys-X_{2-4}Cys-X_3-Phe-X_5-Leu-X_2-His-$X_{3,4}$_His-X_{2-6} where X represents any amino acid (Miller *et al.*, 1985). The tertiary structure of the Cys_2His_2 zinc finger family was initially proposed by Berg (1987) and subsequently confirmed by the structural analysis of isolated zinc finger domains (Lee *et al* 1989) and the construction and analysis of a consensus zinc finger peptide (Krizek *et al.*, 1991) and minimalist zinc finger peptide in which all but seven structurally important residues were replaced by alanine (Michael *et al.*, 1992).

Analysis of these domains showed that much of the amino acid sequence was variable, with only the cysteine, histidine and hydrophobic residues being conserved between different Cys_2His_2 domains. A schematic diagram of a single zinc finger domain is shown in Figure 1.4.

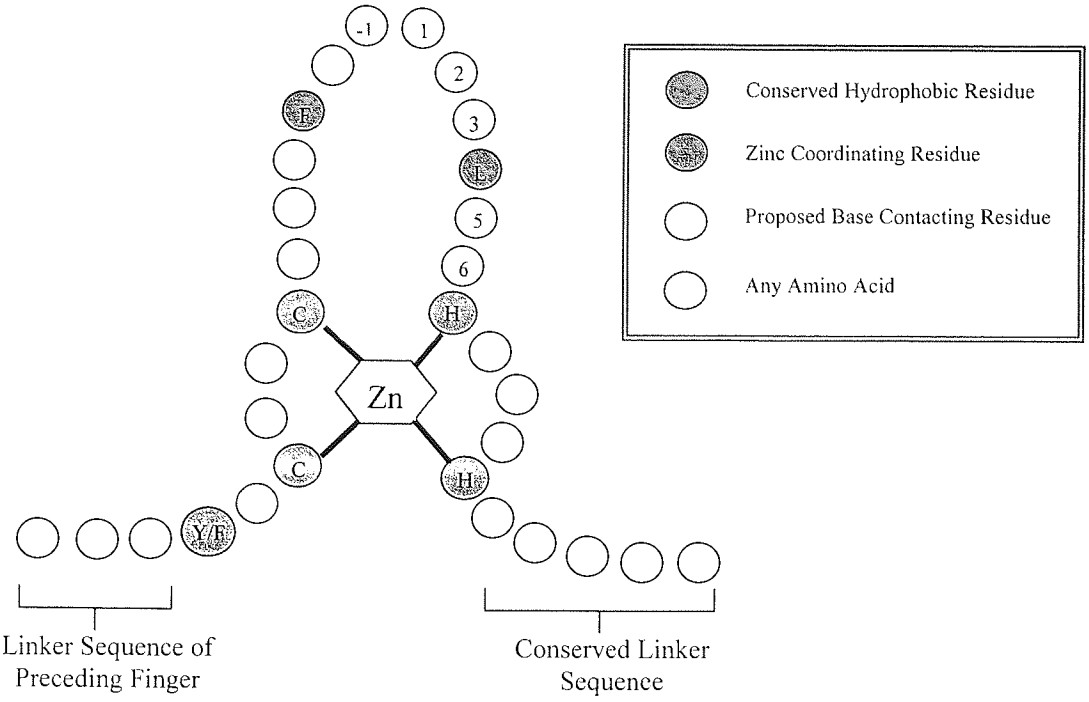


Fig 1.4 Schematic representation of a Cys_2His_2 zinc finger. The identities of conserved residues are represented using single letter code. Residues –1 to +6 of the α helical region of the zinc finger are highlighted in the figure.

The three dimensional structure of a single zinc finger is stabilised by the coordination of a zinc ion and the interaction of the three conserved hydrophobic residues (Berg, 1995). The structure contains two antiparallel β sheets containing the conserved cysteine and hydrophobic residues and an opposing α helix, containing the histidine zinc binding residues and the proposed DNA contacting residues of the domain (Krizek *et al.*, 1991, Micheal *et al.*, 1992). The folding of the domain creates a hydrophobic core containing the three conserved hydrophobic and the zinc binding residues (Krizek *et al.*, 1991).

Approximately 25 of the amino acid residues are folded around the zinc ion to form the finger structure, the remaining residues serve as a linker region connecting consecutive zinc fingers (Rhodes and Klug, 1993). Zinc finger domains usually occur as tandem arrays of multiple zinc fingers, with as many as 37 such domains being identified in a single sequence (Krizek *et al.*, 1991). The linking of single domains in such fashion facilitates the recognition of contiguous regions of DNA. In addition to the conserved residues within the zinc finger domain, a conserved linker sequence connecting zinc fingers has also been identified in approximately 50% of Cys_2His_2 domains (Laity *et al.*, 2000).

1.10.4 Cys_2His_2 Zinc Finger / Nucleic Acid Interaction.

Elucidation of the mechanism of Cys_2His_2 zinc finger / DNA interaction was provided by X-ray crystallographic studies of the mouse transcription factor Zif268 complexed with a DNA oligonucleotide target site. (Pavletich and Pabo, 1991). These studies showed the three-fingered Zif268 protein wound around the DNA target site with the amino terminus of the α helix region of each zinc finger domain inserted into the major groove of the DNA (Fig 1.5).

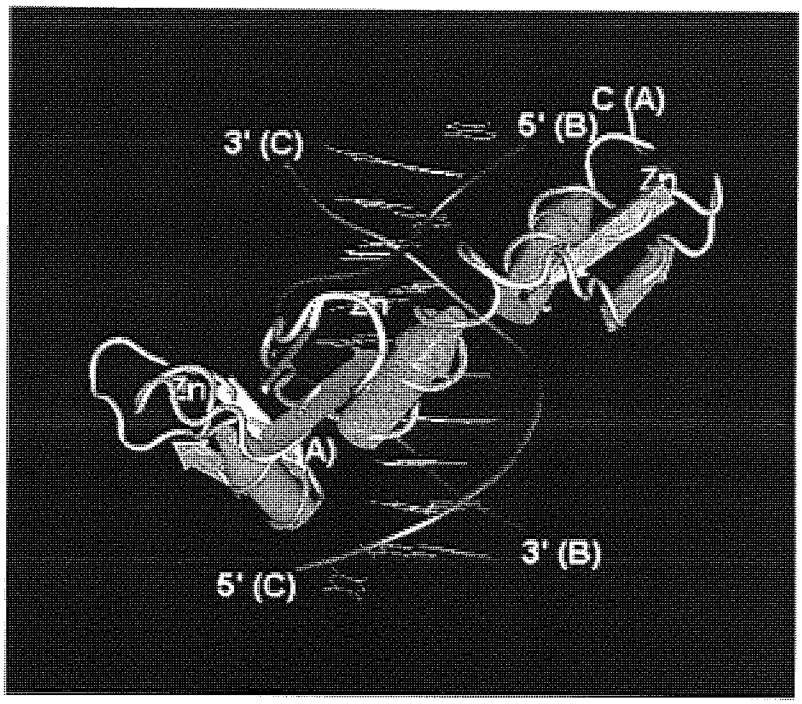


Fig 1.5 Crystal structure of a ZIF268 variant interacting with its DNA target site (at 1.6 Å). The protein backbone and the two nucleotide backbones of the DNA are labelled (A), (B) and (C) respectively. The Carboxyl terminus of the protein is labelled C. β Pleated sheets are denoted by yellow arrows and the α helices are shown as green cylinders. Structure obtained from www.ncbi.nlm.nih.gov/entrez/structure (MMDB 5814) original citation (Elrod-Erikson *et al* 1996) and viewed in schematic format using cn3D viewer version 4.1 (available at www.ncbi.nlm.nih.gov).

Specific base contacts were made to only one strand, known as the primary strand of the DNA target site. Each finger of the three fingered domain contacted three base pairs of the nine base pair sub-site (Fig 1.6), with the specific contacts made by only three amino acids at positions –1, 3 and 6 of the α helix (referred to as residues 13, 16, and 19 by Desjarlais and Berg (1992) of each domain (see fig 1.4).

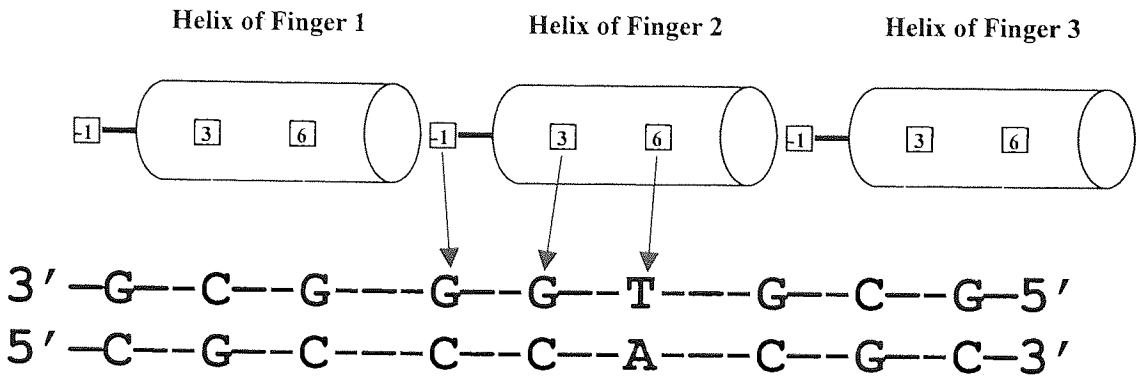


Fig 1.6 Schematic representation of the interactions between Zif268 and its DNA target site as proposed by Pavletich and Pabo (1991). The modular interactions of the base contacting residues (-1, 3 and 6) with the contacted strand of the target site are based upon those reviewed by Berg (1995). The diagram is adapted from a similar representation of zinc finger helical domains in Miller and Pabo (2001). Note as the zinc finger protein is represented in the amino to carboxyl convention, the contacted DNA strand is shown in its antiparallel orientation.

Several non-specific contacts with the phosphate backbone of the DNA were observed in the Zif268 / DNA complex, mediated by residues in both the zinc finger domain itself, including one of the zinc coordinating histidine residues, and residues in the conserved linker sequence of the protein. However sequence-specific contacts to the contacted strand of the target site were only made by three residues in each zinc finger domain, suggesting these residues in each domain may be responsible for DNA recognition. A crystal structure highlighting these residues is shown in Figure 1.7

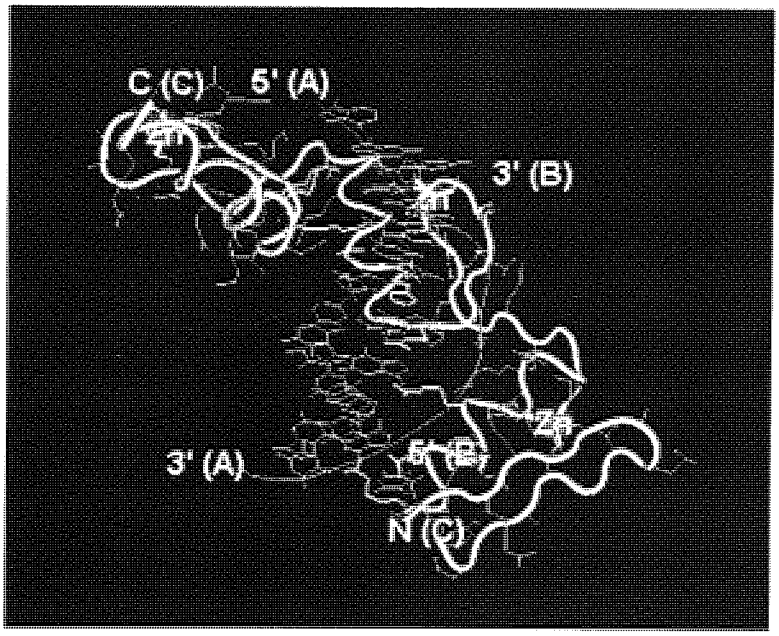


Fig 1.7 Crystal structure of Zif268 interacting with DNA (at 2.1 Å). The amino acid side chains at positions –1, 3 and 6 of each of the α helices are highlighted in yellow. The backbone of the α helices are shown in green with β pleated sheets shown in gold. Structure obtained from www.ncbi.nlm.nih.gov/entrez/structure (MMDB 2496) original citation (Pavletich and Pabo 1991).

Evidence implicating these residues in the sequence specific recognition of nucleotide sequences was also provided by mutagenesis studies (Nardelli *et al.*, 1991) which showed that mutagenesis of two of these residues altered the base discrimination of the zinc finger domains. Further mutagenesis studies performed on Cys_2His_2 domains from the transcription factor Sp1 (Desjarlais and Berg, 1992), demonstrated that the target site of a zinc finger domain could be altered by mutation of two of the base contacting (positions 13 and 16) residues and a further residue (aspartic acid at position 15) which was believed to be important in stabilising the interaction of the arginine at position 13 in the native Sp1 zinc finger.

The ability to define the target site of the mutant zinc finger domains led the researchers to suggest that a code of zinc finger / DNA interactions may exist and that the elucidation of such a code would enable the design of zinc finger proteins with pre-selected DNA specificity (Desjarlais and Berg, 1992a). Applying these design rules, to the base contacting residues of a consensus zinc finger the authors subsequently generated a novel zinc finger protein capable of binding to its preferred sub-site with a dissociation constant of 2nM (Desjarlais and Berg, 1993).

1.10.5 Studies of the Zinc Finger / DNA Recognition Code

Early studies of these domains tentatively suggested that specific binding of DNA may be the result of complex interactions, such as inter-domain effects and DNA contacts mediated by residues other than those at positions –1, 3 and 6 of the α helix. This has been confirmed in subsequent studies of Cys_2His_2 domains which have suggested that simple rules defining binding site preferences can not be applied to all zinc fingers or all target sites. Such studies have provided surprising and sometimes conflicting evidence in the identification of interacting residues within these domains.

As examples, binding studies (Kuwahara *et al.*, 1993) and structural studies (Narayan *et al.*, 1997) of the zinc finger domains from transcription factor Sp1 suggested that only five bases of the nine target site were important for recognition by the three fingered protein. In addition to the fact that only some of the base contacting residues were used in target site recognition, one of the contacted bases was contained in the non contacted

(secondary) strand of the DNA target site. Results of mutagenesis studies on Zif268 (Choo and Klug, 1994b) led the authors to suggest that an aspartic acid residue at position 2 of the third finger made contact with the secondary strand at the cytosine complement of the 5' guanine base of the middle triplet of the target site. Although interactions mediated by an aspartic acid residue at position 2 had been predicted in earlier experimentation, this interaction appeared to be dominant over that of the base contacting residue at position 6 of the preceding finger (Choo and Klug, 1994b), which would be expected to specify the guanine base at this position. These results and further mutagenesis studies suggested that base contact by residues at position 2 of these domains resulted in the recognition of overlapping 4 base pair target sites (Islan *et al.*, 1998) (Fig 1.7).

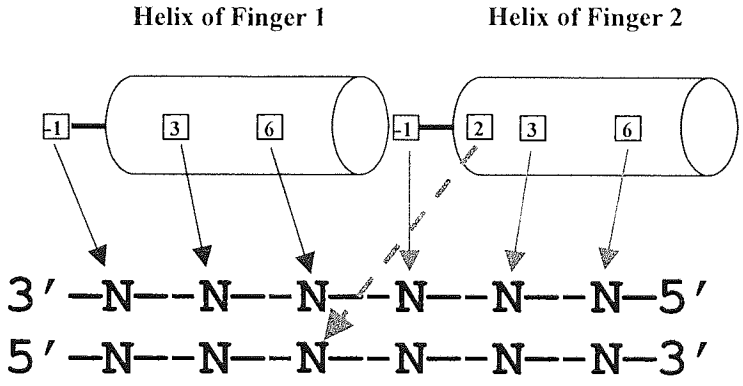


Fig 1.8 Schematic representation of the postulated recognition of overlapping four base pair target sites by Cys_2His_2 zinc fingers (Islan *et al.*, 1998). Diagram adapted from Wolfe *et al.* (1999).

This rule however has also proven to be difficult to apply in the elucidation of a zinc finger / DNA recognition code. Mutants of Zif268 containing aspartic acid at position 2 of finger 1 (which would be predicted to limit sub-site recognition of finger 2 to sites having a guanine at the 5' position) have been shown to also recognise sub-sites containing adenine at this position (Beerli *et* al., 1998). In studies of a Zif268 mutant in which this aspartic acid residue at position 2 of the α helix was replaced with alanine, this substitution produced no dramatic effect on the affinity of the mutant protein for its target site (Miller and Pabo 2001). In addition, a large number of Cys_2His_2 domains have a serine residue at position 2 (approximately 50%, Jacobs, 1992) and mutagenesis

studies of zinc fingers with serine at this site have suggested that this residue is not involved in base recognition (Kim and Berg, 1995).

Further complexities in the "rules" of DNA recognition by zinc finger proteins have also been highlighted. Mutagenesis studies provided evidence that altering the identity of a single base of a zinc finger target site could not simply be achieved by the single substitution of its base contacting residue, but was context-dependent and required additional substitutions at the two remaining base contacting residues (Desjarlais and Berg, 1992b). Subsequent research has also highlighted the context-dependent nature of such substitutions in the base contacting residues of Cys_2His_2 domains (Choo and Klug, 1994b, Elrod-Erikson and Pabo, 1999, Wolfe *et al.*, 2001).

Inter-domain context-dependent effects have also been recognised, as the study of zinc finger / DNA interactions usually employs tandem arrays of multiple zinc fingers (single fingers have generally found to be incapable of site specific interactions; Krizek *et al.*, 1990). These include the previously discussed problem of target site overlap caused by aspartic acid at position 2 of the α helix proposed by Choo and Klug (1994) which has been suggested to limit recognition of the preceding finger to target sites beginning with guanine or thymine (Dreier *et al.*, 2001). In addition, the selection of base contacting residues in a single finger has been shown to be dependent upon the identities of the residues of adjacent fingers in phage display libraries (Wolfe *et al* 1999). Further inter-domain effects have recently been recognised by Lui *et al.* (2002) with the demonstration that different zinc fingers were needed to specify the same DNA triplet while at different positions in a tandem array.

1.11 Experimental Approaches to Generating Zinc Finger Proteins with Novel DNA Target Sites

1.11.1 Zinc finger Randomisation

Despite the apparent complexities of the rules governing zinc finger / DNA interactions, the elucidation of such a code has been the centre of research since the early 1990's. This highlights the enormous potential of these motifs for the creation of "designer" DNA binding proteins. The modular nature of these proteins (independent binding domains joined by a conserved linker sequence) also suggests that independently selected fingers may be joined together to target novel DNA target sites. Examples of the experimental generation of zinc fingers to target novel sequences are shown below to highlight the potential utility of these domains, which have been described as the most promising DNA recognition motifs for the construction of artificial transcription factors (Dreier *et al.*, 2001).

The structure based design of zinc finger proteins has been successfully employed in the generation of proteins with novel DNA binding characteristics (Desjarlais and Berg, 1993, Shi and Berg, 1995) however this approach is somewhat limited by the apparent complexities of the postulated zinc finger / DNA recognition code.

This has led many researchers to create randomised zinc finger libraries in which binding specificities have been altered by randomisation of specific residues within the zinc finger domains. This selection approach to the identification of proteins with novel DNA binding specificities has been adopted successfully in many cases (for review see Segal and Barbas III, 2000 and Choo and Islan, 2000).

Many of these selection experiments focussed upon the randomisation of base-contacting residues of a single finger within a triplet array of zinc finger proteins (Desjarlais and Berg 1994, Jamieson *et* al., 1994, Choo and Klug 1994b, Rebar and Pabo, 1994). The randomisation of only a single finger allowed these proteins to be screened against consensus based target sites, in which the three base pair sub-sites of the conserved fingers act as a docking station allowing the interaction of the randomised finger with different sub-sites to be assessed.

The identification of zinc fingers from within such libraries has often relied upon the use of phage display technology to select randomised proteins on the basis of their affinity for immobilised oligonucleotides containing altered binding sites, with the identity of interacting proteins revealed by sequence analysis of the bound phage (Jamieson *et* al., 1994, Choo and Klug 1994b, Rebar and Pabo, 1994). Zinc finger proteins with novel binding specificities were identified from these libraries after the randomisation of four interacting residues (positions –1, 2, 3 and 6) (Jamieson *et* al., 1994, Rebar and Pabo, 1994). The Choo and Klug library was randomised at seven positions (-1, 1, 2, 3, 5, 6 and 8) with the codon (A/C/G)NN to prevent the inclusion of hydrophobic and cysteine residues, which may perturb the structure of the mutagenised domain. This library was also successful in identifying novel zinc finger proteins, despite the fact that the randomisation of seven residues, resulted in the generated library being 200 times too small to encompass the entire number of finger possibilities (Choo and Klug, 1994b).

Desjarlais and Berg (1994) reversed this conventional form of library screening, by screening a single three zinc finger protein with a library of target sites to identify the optimum sub-site recognised by zinc finger mutants by affinity selection. Target sites with the consensus sequence 5'-GAG-NNN-GAT-3' were generated in which the identity of the NNN sub-site was encoded in the length of the oligonucleotides used to generate the target site. After denaturation of the zinc finger DNA complexes, the identity of the target site/s recognised by each finger was ascertained directly from a sequencing gel after PAGE analysis of the recovered DNA.

Some researchers have adopted randomisation strategies designed to account for possible inter-domain effects upon the target site recognition of the randomised proteins. To address the concern that interactions from a preceding finger may limit the recognition of the 5' base of a target sub-site, two zinc finger domains of triplet zinc finger motifs were randomised simultaneously in libraries constructed by Islan *et al* (1998) and Islan and Choo (2000). In both cases this approach was successful in aiding the discrimination of different bases at the 5' position of the target sub-site, suggesting synergistic interaction between the two randomised fingers. A sequential randomisation strategy was adopted by Greisman and Pabo (1997) and Wolfe *et al.* (1999) to allow context-dependent selection of zinc fingers. Both strategies relied upon the

randomisation of six positions within a single zinc finger of a Zif268 mutant in a sequential manner. After randomisation of the first finger, selections were performed by phage display using DNA containing a randomised three base pair sub-site and the Zif268 sub-sites of fingers 2 and 3. Finger 2 of the selected clones was subsequently randomised and the selection process repeated. This randomisation and selection process was repeated for finger 3 to effectively select each finger contacting a nine base pair target site. As with the libraries described previously, zinc finger proteins which bound novel target sites were identified in both libraries.

Such sequential selection techniques however are experimentally involved, requiring multiple rounds of randomisation and selection. In addition this technique evolved as a means of addressing the perceived limitations imposed by inter-domain effects, upon the selection of randomised single fingers within a zinc finger array. However despite such limitations this approach has resulted in some notable successes. Zinc fingers recognising each of the possible 5'-GNN-3' sub-sites have been identified by the randomisation of a single finger in a tandem array (Beerli *et al.*, 1998). In addition these predefined zinc fingers were then linked together to form a single six finger protein capable of binding an 18 base pair target site, demonstrating the modular nature of these single domains. Fusions of this six finger domain and a repressor domain, were shown experimentally to completely repress the promoter activity of the *erbB-2* promoter, which plays an important role in the development of human malignancies (Beerli *et al.*, 1998).

Zinc finger domains capable of recognising sub-sites containing the consensus sequence 5'-ANN-3' have also been generated by the randomisation of a single finger (Dreier *et al.*, 2001) of Zif268, in which the limitations imposed by potential target site overlap were addressed by the mutagenesis of aspartic acid at position 2 of the preceding helix. Once again the utility of these selected fingers was demonstrated by the construction of six finger proteins fused to regulatory domains and the use of these fusions to regulate *in vivo* transcription of both reporter genes and endogenous human genes.

Unlike the selection of the 5'-GNN-3' sub-sites, zinc fingers capable of recognising all the possible 5'-ANN-3' sub-sites were not identified. This may highlight the importance of the use of correct randomisation strategy in the generation of zinc finger

libraries. Randomisation by Dreier *et al.* (2001) had been carried out using the codon (A/C/G)N(G/C) excluding termination, cysteine and hydrophobic residue encoding codons, a strategy previously employed by Wolfe *et al.* (1999). However in later experimentation this group used resin splitting technology (Section 1.6) to exclude only cysteine and termination codons, in the generation of libraries of Zif268 zinc finger mutants (Joung *et al* 2000). Analysis of these libraries showed that the inclusion of aromatic residues, especially tryptophan, at potential base contacting sites generated mutants capable of binding a 5'-AAA-3' sub-site, despite the presence of aspartic acid at position 2 of the preceding helix, highlighting that these excluded bases may be important in the recognition of 5' A-containing subsites.

1.11.2 Synthetic Zinc Finger Proteins for the Targeting of Biologically Important Nucleotide Sequences

As in the examples in section 1.11.1 (Beerli *et al.*, 1998, Dreier *et al.*, 2001) the potential for the use of Cys_2His_2 zinc fingers in the creation of designer DNA binding proteins, has been highlighted by the use of these domains linked together to form synthetic proteins capable of targeting a number of biologically important target sites (reviewed in Nagaoka and Sugiura, 2000, Beerli and Barbas III, 2002)

As triplet zinc finger proteins recognise only a nine base pair sub-site, the construction of multi-fingered proteins, allowing larger target sites to be recognised, may be important in targeting unique sites within complex genomes, such as the human genome where the recognition of at least 16 base pairs is required to target a unique locus (Liu *et al.*, 1997).

The linking of zinc finger domains has been achieved using only the linker sequence (Thr-Gly-Glu-Lys-Pro) a conserved linker sequence found in many zinc finger proteins. This approach has been used to generate proteins containing nine zinc finger domains capable of binding a 30 base pair target site (Kamiuchi *et al.*, 1998) and six finger proteins, coupled with fusion to activator and repressor domains, capable of regulating the expression of reporter genes within human cells (Liu *et al.*, 1997).

However, studies of triplet zinc finger domains have suggested that the helical periodicity of the zinc fingers does not completely match the helical periodicity of B-DNA, thus as more fingers are added, the helical periodicities of the protein and DNA may be shifted further out of phase (Kim, and Pabo, 1998). This problem may be overcome by the application of structure based design approaches in the construction of multiple zinc finger proteins. The introduction of longer linking sequences containing glycine residues to allow flexibility has been shown to increase both affinity and specificity of polydactyl zinc finger proteins (Kim and Pabo, 1998, Imanishi *et al.*, 2000, Moore *et al.*, 2001), with the six finger peptides generated by Kim and Pabo (1998) demonstrating 6000 times greater affinity for their respective target site than three finger proteins. In addition, six finger peptides generated by Moore *et al.* (2001) by the linking of three two finger proteins showed greater specificity than those generated from two three finger proteins, suggesting that the flexible linker sequence reduced any discrepancy in helical periodicity allowing these fingers to adopt a more active conformation.

Novely, a zinc finger domain has also been used to link triplet zinc finger arrays to create a six finger protein capable of spanning up to ten base pairs of DNA between the recognition subsites of each three finger moiety (Moore *et al.*, 2000). The linking finger designed to make no base contacts within the target DNA, was used to act as a bridge over the minor groove of the DNA. Interestingly separation of the two respective subsites by smaller gaps (1-2 base pairs) did not prevent recognition by the seven finger peptide, suggesting this linking finger could "flip out" from the DNA to enable binding by the linked three finger peptides.

This ability to link pre-selected zinc finger domains to create polydactyl proteins of known specificity, further extends the utility of these versatile domains. Examples in which such proteins have been used to regulate transcriptional activity at endogenous sites within human cells (Kang and Kim, 2000, Zhang *et al.*, 2000) highlights the remarkable potential of these proteins. Recent notable examples of the use of Cys_2His_2 domains, such as, the use of multiple zinc finger transcription factors to scan for gene regulatory elements (Blancafort *et al.*, 2003), and perhaps more importantly the use of such proteins to inhibit the replication of HIV-1 (Reynolds *et al.*, 2003) and to inhibit

gene expression of the herpes simplex virus (Papworth *et al.*, 2003) suggest that these domains may eventually provide a means to address a number of biological problems.

1.12 Aims and Objectives.

The aim of the current study primarily centres upon the development of the MAX randomisation process and the assessment of the ability of the technique to specify a single codon to represent each of the possible twenty amino acids, in the randomised DNA sequence. As the MAX randomisation technique is to be eventually tested in the construction of zinc finger libraries, this development work was carried out within this context.

Thus with reference to the intended use of the randomisation technique, the study initially aims to produce a suitable target gene in which the randomisation process can be assessed. The gene should encode a triplet zinc finger protein, in which the base contacting residues of the middle finger are amenable to randomisation by cassette mutagenesis.

Subsequent work in the study is aimed at the development of the MAX randomisation process itself. This includes the development of the technique itself and the assessment of its success in constructing gene libraries, in which the need to clone redundant codons is removed. In addition, this development work is also intended to identify and address, any problems or potential problems associated with the intended use of the MAX randomisation technique.

Chapter 2 Materials and Methods

All reagents listed in the materials and methods were purchased from Sigma (Poole, UK) unless otherwise stated.

2.1 Media Recipes

(2.1.1) LB Broth

1 % ($^{w}/_{v}$) Bacto tryptone (Oxoid, Basingstoke), 0.5 % ($^{w}/_{v}$) Yeast extract (Oxoid Basingstoke), 0.5 % ($^{w}/_{v}$) NaCl (Fisher Loughborough).

(2.1.2) LB Agar

1 % ($^{w}/_{v}$) Bacto tryptone (Oxoid Basingstoke), 0.5 % ($^{w}/_{v}$) Yeast extract (Oxoid Basingstoke), 0.5 % ($^{w}/_{v}$) NaCl (Fisher Loughborough). 1.5 % ($^{w}/_{v}$) Bactoagar.

(2.1.3) SOB Broth

2 % ($^{w}/_{v}$) Bacto tryptone (Oxoid Basingstoke), 0.5 % ($^{w}/_{v}$) Yeast extract (Oxoid Basingstoke), 1 % NaCl ($^{w}/_{v}$) (Fisher Loughborough), 0.25 % KCl ($^{w}/_{v}$) (Fisher Loughborough), 1 % ($^{w}/_{v}$) $MgCl_2$ and 1 % ($^{w}/_{v}$) $MgSO_4$.

All Media were sterilised by autoclaving for 20 minutes at 121°C.
Selective media were prepared by the addition of ampicillin solution (2.2.15) or kanamycin solution (2.2.16) to cooled media (<50°C) to final concentrations of 50 μg/ml and 30 μg/ml respectively.

2.2 Buffer Recipes

(2.2.1) TAE

1 x TAE (0.04 M Tris-acetate, 0.001 M EDTA) was prepared from a 50 x stock solution (2m Tris-acetate, 0.05 M EDTA pH 8.0).

(2.2.2) Loading Buffer

Loading buffer was prepared using double distilled water, containing 30% ($^{v}/_{v}$) Glycerol, 0.025 % ($^{w}/_{v}$) Xylene cyanol, and/or 0.025 % ($^{w}/_{v}$) Bromophenol blue.

(2.2.3) RFB 1 Buffer

100 mM RbCl, 50 mM $MnCl_2 \cdot 4H_2O$ 30 mM potassium acetate 10 mM $CaCl_2 \cdot 2H_2O$) and 15 % ($^w/_v$) Glycerol.

(2.2.4) RFB 2 Buffer

10 mM MOPS, 10 mM RbCl, 75 mM 10mM $CaCl_2 \cdot 2H_2O$) and 15 % ($^w/_v$) Glycerol.

(2.2.5) Hybridisation Buffer 1

(1 x Buffer) 50 mM Tris-HCl (pH 7.6), 10 mM $MgCl_2$ and 4 % ($^w/_v$) PEG 8000.

(2.2.6) Hybridisation Buffer 2

(1 x Buffer) 40 mM Tris-HCl (pH 7.4) and 10 mM $MgCl_2$.

(2.2.7) β Agarase buffer

(10 x Buffer) 100mM Bis Tris-HCl (pH 6.5) 1mM Na_2EDTA. (NEB Hertfordshire).

(2.2.8) Ligase Buffer

(5 x Buffer) 250 mM Tris-HCl, 50 mM $MgCl_2$, 5 mM ATP, 5mM DTT, 25 % ($^w/_v$) PEG 800 (Gibco UK).

(2.2.9) T4 PNK Buffer

(10 x Buffer) 700 mM Tris-HCl (pH 7.6), 100 mM $MgCl_2$, 50 mM DTT (NEB Hertfordshire).

(2.2.10) CIP Buffer

CIP reactions were carried out in NEB restriction enzyme buffers 2 or 4 dependant upon the preceding digestion reaction.

(2.2.11) PCR Buffer

(10 x Buffer) 160 mM $(NH_4)_2SO_4$, 670 mM Tris-HCl (pH 8.8) and 0.1 % Tween-20 (Bioline London).

(2.2.12) *Pfu* polymerase Buffer

(10 x Buffer) 200 mM Tris-HCl (pH 8.8), 20 mM $MgSO_4$, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 1% ($^{W}/_{V}$) Triton X-100, 1 mg/ml nuclease free BSA (Stratagene UK).

(2.2.13) Restriction Enzyme Buffers

All restriction digest reactions were performed in the appropriate buffer supplied by NEB (Hertfordshire).

NEB Buffer 2 (*Hin*dIII, BsiWI, SpeI, *Eco*RI, *Bsm*I)

(10 x Buffer) 100 mM Tris-HCl, 500 mM NaCl, 10 mM DTT (pH 7.9 at 25°C).

NEB Buffer 3

(10 x Buffer) 500 mM Tris-HCl, 100 mM $MgCl_2$, 10 mM DTT (pH 7.9 at 25°C).

NEB Buffer 4 (*Sma*I, *Sna*BI, *Bpu*1012 I)

(10 x Buffer) 200 mM Tris-acetate, 100 mM Magnesium acetate, 500 mM Potassium acetate, 10 mM DTT, (pH 7.9 at 25°C).

(2.2.14) Ampicillin Solution

Stock solutions of ampicillin were prepared using ampicillin sodium salt and double distilled water. Solutions were filter sterilised using a 0.2 μm syringe filter (Nalgene UK) according to the manufacturer's instructions.

(2.2.15) Kanamycin Solution

Stock solutions (30 mg/ml of kanamycin were prepared using kanamycin sulphate and double distilled water. Solutions were filter sterilised using a 0.2 μm syringe filter (Nalgene UK).

(2.2.16) ATP Solutions

ATP solutions were prepared using 100 mM ATP stock solution (Amersham Pharmacia, Little Chalfont) and sterile double distilled H_2O.

(2.2.17) DTT Solutions

DTT solutions were prepared using molecular biology grade DTT and sterile double distilled H_2O.

(2.2.18) $CaCl_2$

50 mM $CaCl_2$ was prepared using $CaCl_2$ (Analar R) and double distilled water, before sterilisation at 121°C for 20 minutes.

2.3 Cell lines

E. coli DH5α cells (Gibco Paisley) (SupE44 ΔLacU169 (Ø80 Lac Z ΔM15) hsdR17 recA1 end A1 gyrA96 thi-1 rel A1) were employed in all transformations.

2.4 General Techniques

(2.4.1) Preparation and Transformation of Competent Cells ($CaCl_2$ Method)

Overnight cultures of cells were prepared using a single colony of DH5α cells (2.3) to inoculate 10 ml of LB broth (2.1.1). A 1% inoculum of this overnight culture was added to 30 ml of sterile LB media and incubated at 37°C until the cells entered logarithmic growth phase, (approximately 0.4 OD at 550 nm). The cells were pelleted in a cooled centrifuge (4000 rpm 5mins, 4°C) using a Beckman JA 20 rotor and the media discarded. Cells were re-suspended in ice-cold sterile 50mM $CaCl_2$ (2.2.18) at 20 % of the original volume of media and left on ice for 20 minutes. This wash step was repeated, and the cells incubated on ice for a further 20 minutes. Subsequently the cells were pelleted as previously described and the $CaCl_2$ discarded. Cells were re-suspended in ice cold $CaCl_2$ at 4 % of the original volume of media and incubated on ice for a further 30 minutes.

In each transformation 100 µl of competent cells were added to 5 ng of transforming DNA in pre chilled 1.5 ml eppendorf tubes. The contents were mixed by flicking the tubes and placed on ice for 30 minutes. Subsequently the cells were heat-shocked for 1 minute in a 37°C water bath and returned to ice for a further two minutes. After the

addition of 0.9 ml of sterile LB broth (2.1.1) the cells were incubated at 37°C for 45 minutes to allow expression of conferred phenotypic characteristics. Subsequently 200 µl aliquots of the transformed cells were plated on LB media (2.1.2) containing ampicillin at a concentration of 50 µg/ml, and incubated at 37°C.

(2.4.2) Preparation and Transformation of Competent Cells (Rubidium Chloride Method)

Overnight cultures of cells were prepared using a single colony of DH5α cells (2.3) to inoculate 10 ml of SOB broth (2.1.3). A 0.25 % ($^{v}/_{v}$) inoculum of this overnight culture was added to sterile SOB media and incubated at 37°C until the cells entered logarithmic growth phase, (approximately 0.4 OD at 550 nm). Once the cells had reached a logarithmic growth phase the culture was chilled on ice for 30 minutes. Subsequently the cells were pelleted by centrifugation (2500 rpm for 15 mins in a Beckman JA 14 rotor, at 4°C) and the supernatant removed. The cells were re-suspended in ice-cold RFB 1 buffer (2.2.3) at 33 % of the original volume of media and chilled on ice for 60 minutes. The cells were then pelleted by centrifugation (2500 rpm for 15 mins in a Beckman JA 14 rotor, at 4°C) and the supernatant removed. The cells were re-suspended in RFB 2 buffer (2.2.4) at 8 % of the original volume of media and chilled for a further 15 minutes. The prepared cells were aliquotted into chilled 1.5ml eppendorf tubes, flash frozen in liquid nitrogen and stored at – 80°C until required.

Prior to transformation stocks of competent cells were thawed on ice. Once defrosted 100 µl of competent cells were added to 5 ng of transforming DNA in pre chilled 1.5 ml eppendorf tubes. The contents were mixed by flicking the tubes and placed on ice for 30 minutes. Subsequently the cells were heat-shocked for 1 minute in a 37°C water bath and returned to ice for a further two minutes. After the addition of 0.9 ml of sterile LB broth (2.1.1) the cells were incubated at 37°C for 45 minutes to allow expression of conferred phenotypic characteristics (Cells transformed with plasmids encoding kanamycin resistance as a selection marker were incubated for 75 mins at 37°C). 200 µl aliquots of the transformed cells were plated on LB media (2.1.2) containing ampicillin at a concentration of 50 µg/ml or kanamycin at a concentration of 30 µg/ml, and incubated at 37°C.

(2.4.3) Ethanol Precipitation

DNA was precipitated by the addition of a one tenth volume of 3 M tri-sodium acetate (pH 5.5) and a 2 x volume of ice-cold absolute ethanol (Fisher UK). The mixture was placed on ice for 30 minutes. DNA was pelleted by centrifugation (14000 r.p.m. for 20 mins in an Eppendorf microfuge) and the pellet dried by aspiration. DNA pellets were washed with 70 % ethanol and dried by aspiration/evaporation before re-suspension in the appropriate buffer.

(2.4.4) Isopropanol Precipitation

DNA was precipitated by the addition of a one tenth volume of 3 M tri-sodium acetate (pH 5.5) and a 2 x volume of isopropanol (Fisher UK). The mixture was placed at -20°C for 20 minutes and the DNA subsequently pelleted by centrifugation (14000 r.p.m. in an eppendorf microfuge for 20 mins). The isopropanol was removed by aspiration and the pellet washed with 70 % ethanol. The DNA was dried by aspiration/evaporation before re-suspension in the appropriate buffer.

(2.4.5) Phenol Chloroform Extraction of DNA

DNA was extracted by the addition of an equal volume of phenol:chloroform: isoamyl alcohol at a ratio of 25:24:1. The mixture was vortexed and then centrifuged at 14000 r.p.m. in an Eppendorf microfuge for 30 seconds. The upper aqueous phase was removed, and the process repeated until no protein was visible at the interface between the organic / aqueous phases. The recovered aqueous phase was mixed with an equal volume of chloroform:isoamyl alcohol at a ratio of 24:1 to remove any residual phenol. The mixture was vortexed and centrifuged at 14000 r.p.m. for 30 minutes. The upper aqueous phase was removed and the DNA recovered by ethanol precipitation (2.4.3) before re-suspension in the appropriate buffer.

(2.4.6) Purification of Plasmid DNA (Small Scale)

Plasmid DNA was recovered from cells grown in liquid culture using the Wizard™ minipreps system (Promega Madison) in accordance with the manufacturer's instructions.

(2.4.7) Purification of Plasmid DNA (Large Scale)

Plasmid DNA was recovered from cells grown in liquid culture using the Wizard™ maxipreps system (Promega, Madison) in accordance with the manufacturer's instructions.

(2.4.8) BLAST Searching of the *E. coli* Genome

BLAST searches were performed using the BLAST search facility at the NCBI website (www.ncbi.nlm.nih.gov/BLAST). Blast searches were performed for short nearly exact matches and the search limited to the *E. coli* database, selecting no complexity filter.

2.5 Agarose Gel Electrophoresis

(2.5.1) Agarose Gels

Agarose gels (1, 2 and 3 % $^{w}/_{v}$) were prepared using molecular biology grade agarose (Gibco Paisley). Gels were prepared and electrophoresed in 1 x TAE buffer (2.2.1) Electrophoresis was carried out using a 5 volt/cm potential difference for DNA of less than 1 Kb in length and 4 volts/cm potential difference for DNA greater than 1Kb. Gels were stained by the addition of ethidium bromide at a final concentration of 0.5 μg/ml and the DNA visualised using the UVP transilluminator (UVP Products UK, or GeneSnap transilluminator photographic systems, (GeneSnap UK).

(2.5.2) Gel Purification (Agarose Gel)

Agarose gels were prepared using SeaKem™ high purity low melting point agarose gel (Flowgen Rockland). Gels were prepared and electrophoresed in 1 x TAE containing guanosine to a final concentration of 1 mM to protect the DNA from UV damage. Gels were electrophoresed and stained as described in (2.5.1). DNA was visualised using a transilluminator and the desired bands excised from the gel using sterile scalpel blades. DNA was recovered from the gel slice by β agarase digestion (2.8.7) or by the use of a Quiax II kit (Qiagen) in accordance with the manufacturers instructions.

(2.5.3) DNA Quantitation using Agarose Gel Electrophoresis.

DNA was quantitated by comparison to quantitative DNA molecular weight ladders (2.6) using agarose gels prepared as described in (2.5.1). Comparison of relative band intensities was performed using Phoretix 1D Advanced Gel Analysis program (Phoretix International UK).

2.6 DNA Molecular Weight Markers

(2.6.1) Hyperladder IV

The quantitative 100 – 1000 bp, Hyperladder IV DNA ladder was obtained from Bioline (London).

(2.6.2) MassRuler DNA Ladder High Range

The quantitative 1500 – 10 000 bp, MassRuler DNA Ladder was obtained from MBI Fermentas (Distrib. Helena Biosciences Sunderland).

(2.6.3) GeneRuler 100bp DNA Ladder

The quantitative 80 – 1031 bp, GeneRuler DNA ladder was obtained from MBI Fermentas (Distrib. Helena Biosciences Sunderland).

(2.6.4) Sigma PCR Low Ladder

The 100 – 1000 PCR Low Ladder was obtained from Sigma (Poole).

(2.6.5) Promega 100 bp DNA ladder

The 100 – 1000 bp DNA ladder was obtained from Promega (Madison).

(2.6.6) λ *Hin*dIII DNA Markers

λ *Hin*dIII DNA markers were prepared by the digestion (2.8.4) of λ DNA (dam^- dcm^-) (MBI Fermentas, distrib. Helena Biosciences Sunderland) with the restriction enzyme *Hin*dIII.

2.7 Plasmid DNA

(2.7.1) pUC19 DNA

pUC19 DNA was obtained from MBI Fermentas (Distrib. Helena Biosciences Sunderland).

(2.7.2) pGEX-2TK

The expression vector pGEX-2TK was obtained from Amersham Biosciences (Amersham).

(2.7.3) pET-42a

pET-42a DNA was obtained from Novagen (Distrib. Merk Biosciences, Nottingham).

2.8 Enzyme Dependent Techniques

(2.8.1) Calf Intestinal Alkaline Phosphatase (CIP) Reactions.

CIP reactions were performed in 1 x NEB buffer 2 or NEB buffer 4 (2.2.14). DNA was added to these reactions to a maximum concentration of 50 ng/μl. One unit CIP (NEB Hertfordshire) / pmol "DNA ends" was employed to hydrolyse the 5' phosphates from blunt or recessed termini.

(2.8.2) Phosphorylation (T4 Polynucleotide Kinase) Reactions

Phosphorylation reactions were performed in 1 x T4 PNK buffer (NEB Hertfordshire) containing 1mM ATP (Amersham Biosciences, Amersham). Oligonucleotides were kinased at a maximum concentration of 50 pmol/μl using 5 units of T4 PNK (NEB Hertfordshire) per reaction.

(2.8.3) Ligation Reactions.

Ligations were performed in 1 x ligation buffer (2.2.8). Ligations were performed in reaction volumes of 20 μl, containing 100 ng of the appropriate vector, a 3 fold molar excess of insert DNA and 1 Weiss unit ligase (Gibco UK). Reactions were incubated at 14°C for cohesive end ligation and 4°C for blunt end ligation, for 12 - 16 hours.

(2.8.4) Restriction Digest Reactions.

Restriction digest reactions were performed in a 1 x concentration of the appropriate restriction digest buffer (2.2.14). When indicated by the supplier the buffer was supplemented with bovine serum albumin (BSA) at a final concentration of 0.1 μg/μl to prevent star activity of the enzymes. The DNA was digested at concentrations of 50 ng/μl or below. Enzyme concentration was varied in differing reactions in accordance with the suppliers' recommendation. All restriction enzymes were obtained from NEB (Hertfordshire).

(2.8.5) PCR and Colony PCR Reactions.

PCR reactions were carried out in 1 x PCR buffer (2.2.11) containing 100 μM dNTPs, (Amersham Biosciences, Amersham) 1.5mM $MgCl_2$ (Bioline London), the appropriate forward and reverse primers (Appendix) at a final concentration of 0.4 μM and approximately 0.8 units Taq polymerase (Bioline London). PCR reactions were carried out in an PTC-100 thermal cycler (MJ Research Watertown MA USA), employing a standard PCR cycle of 30 cycles at the temperatures listed below unless otherwise stated.

95°C for 30 seconds
55°C for 30 seconds
72°C for 60 seconds.

Amplification of plasmid DNA was carried out using 0.5 ng of plasmid per reaction. Colony screening was carried out by direct amplification of the colonies (colony PCR) as described above with the exception that reactions were incubated at 95°C for 3 minutes prior to cycle 1 to lyse the bacterial cells.

(2.8.6) *Pfu* Polymerase amplification

Amplification of insert DNA with *Pfu* polymerase was carried out in 1 x *Pfu* polymerase buffer (2.2.12) containing 100 μM dNTPs (Amersham Biosciences, Amersham), the appropriate forward and reverse primers at a final concentration of 0.25 μM and 1 ng of template DNA in each reaction. 4 units of cloned *Pfu* Polymerase were added to each reaction immediately before amplification to prevent degradation of the

oligonucleotide primers. 30 cycles of amplification were carried out on an PTC-100 thermal cycler (MJ Research Watertown MA USA) using the cycling conditions listed below.

95°C for 30 seconds
55°C for 30 seconds
72°C for 60 seconds.

(2.8.7) β agarase Digestion of Gel Slices

Excised gel slices were placed in pre-weighed Eppendorf tubes and the volume of the gel slice calculated. This volume was incorporated into a 200 µl reaction containing 1 x β agarase buffer (2.2.7). The reaction was incubated at 65°C for 15 minutes to melt the gel slice and then cooled to 40°C. After cooling, 2 units of β agarase (NEB Hertfordshire) was added to each reaction and the reactions incubated at 40°C for a minimum of 4 hours. Subsequent to incubation the reactions were vortexed and then placed at –20°C for 20 minutes. The reactions were centrifuged at 14 000 r.p.m. for 10 minutes to pellet any insoluble material. The supernatant was removed and transferred to a fresh eppendorf tube whereupon the DNA was recovered from the mixture by isopropanol precipitation (2.4.4).

(2.8.8) Sequencing Reactions

Unless stated otherwise, sequencing reactions were carried out in conjunction with the Birmingham University Functional Genomics Laboratory. Sequencing was performed using the Big Dye 3 chain termination protocol (Functional Genomics Laboratory Birmingham University), in accordance with the supplier's instructions.

(2.8.9) Sequence reactions (Lark Technologies)

Fluorescent chain termination sequencing was performed by Lark technologies. Ethanol precipitated plasmid DNA purified (2.4.6) from liquid culture was supplied to lark technologies as template DNA.

2.9 Oligonucleotide Synthesis and Hybridisation.

(2.9.1) Oligonucleotide Synthesis

All oligonucleotides were obtained HPSF purified from MWG Biotech (Ebersberg Germany). Sequences of oligonucleotide primers not listed in the text or figures are contained in the appendix (A1).

(2.9.2) Hybridisation of Oligonucleotides to Create Insert DNA

Phosphorylated complementary oligonucleotides were added to hybridisation reactions at a final concentration of 1 pmol/µl. Hybridisation was carried out using the DVTCH1 protocol described in (2.9.3).

(2.9.3) Hybridisation of Selection Oligonucleotides (PNK Buffer)

Individual pools of α, β and γ selection oligonucleotides were generated by mixing equimolar amounts of each of the twenty respective selection oligonucleotides. Prior to hybridisation, the β and γ oligonucleotide pools were treated with T4 PNK (2.8.2), the non-phosphorylated α oligonucleotide pools were diluted in 1 x T4 PNK buffer (2.2.9). Aliquots from each pool, corresponding to 5 pmoles of the mixed selection oligonucleotides were added to a hybridisation reaction containing 320 pmoles of the template oligonucleotide (this 64 : 1 molar ratio of template : selection oligonucleotides, was maintained in all hybridisation reactions used to create the MAX randomised cassettes). Hybridisation reactions were carried out using a PTC-100 thermal cycler (MJ Research Watertown MA USA). Reactions was heated at 94°C to denature any non-specific interactions between the oligonucleotides, and then subjected to controlled cooling using the DVTCH1 protocol listed below.

95°C for 2 minutes.
Controlled cooling to 75°C @ -1°C per 2 seconds.
75°C for 30 seconds.
Controlled cooling to 65°C at -1 °C per 20 seconds.
65 °C for 30 seconds
Controlled cooling to 4 °C at -1 °C per 2 minutes

(2.9.4) Hybridisation of Selection Oligonucleotides (Hybridisation Buffers 1 and 2)

Individual pools of α, β and γ selection oligonucleotides were generated by mixing equimolar amounts of each of the twenty respective selection oligonucleotides. Prior to hybridisation, the β and γ oligonucleotide pools were treated with T4 PNK (2.8.2), the non-phosphorylated α oligonucleotide pools were diluted with sterile water. Aliquots from each pool, corresponding to 5 pmoles of the mixed selection oligonucleotides were added to a hybridisation reaction containing 320 pmoles of the template oligonucleotide and 320 pmoles of pre-kinased ENDMAX oligonucleotides (generating a 64 : 1 ratio of template oligonucleotides to selection oligonucleotides and a 1 : 1 ratio of template and ENDMAX oligonucleotides). Hybridisation reactions were carried out in 1 x hybridisation buffer 1 (2.2.5) or 2 (2.2.6) using a PTC-100 thermal cycler (MJ Research Watertown MA USA). Reactions were heated at 94°C to denature any non-specific interactions between the oligonucleotides, and then subjected to controlled cooling using the DVTCH1 protocol listed previously (2.9.3).

(2.9.5) Hybridisation and Pre-ligation of Selection Oligonucleotides

Individual pools of α, β and γ selection oligonucleotides were generated by mixing equimolar amounts of each of the twenty respective selection oligonucleotides. Prior to hybridisation, the β and γ oligonucleotide pools were treated with T4 PNK (2.8.2), the non-phosphorylated α oligonucleotide pools were diluted with sterile water. Aliquots from each pool, corresponding to 5 pmoles of the mixed selection oligonucleotides were added to a hybridisation reaction containing 320 pmoles of the template oligonucleotide and 320 pmoles of pre-kinased ENDMAX oligonucleotides. Hybridisation reactions were carried out in 1x hybridisation buffer 1 (2.2.5) or 2 (2.2.6) using a PTC-100 thermal cycler (MJ Research Watertown MA USA) and a modified version of the DVTCH1 protocol, listed below.

95°C for 2 minutes.
Controlled cooling to 75°C @ -1°C per 2 seconds.
75°C for 30 seconds.
Controlled cooling to 65°C at -1 °C per 20 seconds.
65 °C for 30 seconds
Controlled cooling to 4 °C at -1 °C per 2 minutes
14°C Incubation

Subsequent to hybridisation, reactions were maintained at 14°C. ATP (1 mM final conc. in buffer 1, 0.5 mM final conc. in buffer 2) and DTT (1 mM final conc. in buffer 1, 10 mM final conc. in buffer 2) were added to each reaction at appropriate concentrations. One Weiss unit of ligase was then added to each reaction prior to incubation at 14°C overnight.

Chapter 3 Gene Assembly for Library Construction

3.1 Introduction

The synthesis of randomised gene libraries centres around the mutagenesis of a single parental gene. As a basis for the creation of zinc finger libraries, the zinc finger protein QDR-RER-RHR as described by Desjarlais and Berg (1993) was selected. This previously characterised protein was selected due to its high affinity and specificity for its DNA target site. The protein binds to the site 5'-GGG-GCG-GCT-3' with a dissociation content of 2 nM (Desjarlais and Berg, 1993) and consists of a triplet array of zinc finger motifs each linked by a canonical linker sequence. The name QDR-RER-RHR refers to the respective base contacting residues of the three zinc finger motifs.

Two plasmid constructs were generously provided by D. Palfrey. The plasmid pGEX-ZFH (Palfrey *et al.*, 2002) constructed in the expression vector pGEX-2TK (Amersham Biosciences) encodes the QDR-RER-RHR protein as a C-terminal fusion to glutathione-S-transferase. The plasmid ZFH6 contains a gene encoding the QDR-RER-RHR protein inserted between the *Bam*HI and *Eco*RI sites of pUC19. The amino acid sequence of the QDR-RER-RHR protein and the nucleotide sequence of the ZFH gene are shown in Figs. 3.1a and 3.1b respectively.

As the MAX randomisation technique relies upon cassette mutagenesis, to enable the randomisation of a target gene, the region to be randomised must be flanked on each side by two unique restriction sites. As the zinc finger gene within pGEX-ZFH contained no suitable restriction sites for cassette mutagenesis, it was necessary to introduce suitable restriction sites into the ZFH gene by silent mutagenesis. At the commencement of this project, plasmid pUCD4 was obtained from A. V. Hine. This plasmid contains a fragment of the ZFH gene extending from the start of the *Bsm*I restriction site (base 148) to the *Eco*RI restriction site (base 309) and contains five base changes that introduce *Hin*dIII and *Bsi*WI restriction sites, by silent mutagenesis (Fig 3.2).

H_2N- M E K L R N G S G D P G K K K-

Q H A **C** P E **C** G K S F S Q S S D L Q R **H** Q R T **H** T G E K-
13 16 19

P Y K **C** P E **C** G K S F S R S D E L Q R **H** Q R T **H** T G E K-
13 16 19

P Y K **C** P E **C** G K S F S R S D H L S R **H** Q R T **H** Q N K K-COOH
13 16 19

Predicted binding site 5'- GGG-GCG-GCT-3'

COOH-RHR-RER-RDQ-NH_2

Fig 3.1a The amino acid sequence of the QDR-RER-RHR protein (Desjarlais and Berg 1993), encoded by the ZFH gene. The zinc coordinating residues of each finger are denoted by bold typeface. The putative Base contacting residues are underlined.

1 30

GGA TCC GAG AAA CTT CGT AAT GGT TCG GGC GAC CCA GGA AAG AAG

***Bam*HI**

60 90

AAA CAG CAT GCG TGC CCA GAG TGT GGT AAG AGC TTC AGT CAA TCC

120

TCT GAT CTG CAG CGC CAC CAA CGT ACA CAT ACC GGG GAG AAA CCG

150 180

TAC AAG TGT CCA *GAA TGC G*GG AAG TCC TTT AGT **CGC** AGC GAC **GAA**

***Bsm*I**

210

TTA CAA **CGT** CAT CAG CGC ACT CAC ACC GGG GAA AAG CCA TAT AAA

240 270

TGC CCT GAA TGT GGC AAG TCT TTC AGC CGT AGT GAT CAT CTG TCT

300

CGC CAT CAA CGC ACG CAT CAG AAC AAG AAA TGA *GAA TTC*

***Eco*RI**

Fig 3.1b The nucleotide sequence of the coding strand of the ZFH gene, which encodes the QDR-RER-RHR protein. The DNA encoding the putative base contacting residues of the middle finger are highlighted in bold face. The recognition sequences of the restriction enzymes *Bsm*I, *Bam*HI and *Eco*RI are shown in Italics.

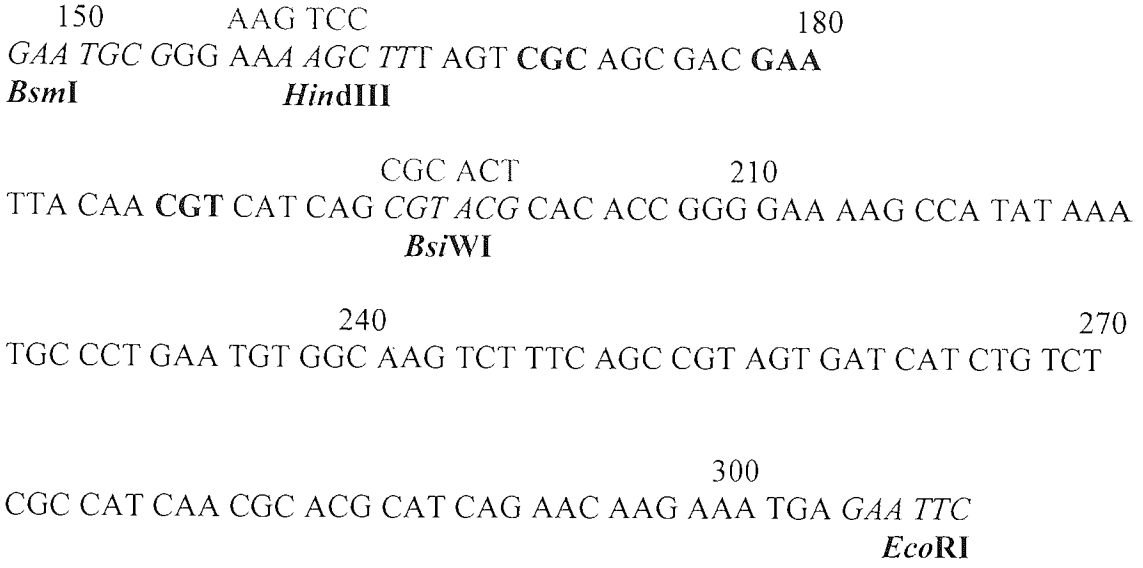


Fig 3.2 The mutagenised fragment of the ZFH gene in clone pUCD4 corresponds to bases 148-309 of the ZFH gene sequence. Silent mutations introduce the *Hin*dIII site, which represents positions 158 – 164 of the complete library zinc finger gene. The numbering sequence from the ZFH gene has been maintained in the figure. The mutagenised fragment substitutes the lysine encoding triplet AAA in place of the lysine codon AAG in the ZFH sequence, the second substitution represents a substitution of the serine codon TCC in position 157 of the ZFH sequence for its alternative AGC triplet, to generate the *Hin*dIII site. The introduction of the *Bsi*WI recognition site was achieved by the substitution of the arginine encoding triplet CGC at position 196 of the ZFH gene for the replacement arginine codon CGT and the substitution of the threonine codon ACT at position 199 of the ZFH gene for the alternative ACG threonine encoding triplet. The codons replaced in the ZFH gene are shown in red face above the pUCD4 sequence. Restriction enzyme sites are shown in italics.

3.2 Construction of the library gene

At the start of this study, the sequence encompassing the *Bsm*I to *Eco*RI region in the pGEX-ZFH gene (Fig 3.1b) had to be replaced with the corresponding region from clone pUCD4. This was initially attempted by the direct insertion of the pUCD4 fragment into the pGEX-ZFH construct.

The pGEX-ZFH plasmid construct was digested (2.8.4) with the enzymes *Bsm*I and *Eco*RI. The digested DNA was treated with calf intestinal alkaline phosphatase (CIP) (2.81) to remove the 5' phosphate groups from the cohesive termini of the plasmid and prevent religation of the parental plasmid, should plasmid DNA digested with only one of the two restriction enzymes persist in subsequent ligation reactions. The digested plasmid was subjected to gel purification (2.5.2). The recovery of the DNA was estimated by visualization on an agarose gel (2.5.1). The concentration of the recovered plasmid (lane 2 Fig 3.3) was estimated to be 20 ng/μl.

The clone pUCD4 was digested (2.8.4) with *Bsm*I and *Eco*RI to remove the mutagenised fragment and the fragment separated using agarose gel electrophoresis (2.4.2). The fragment indicated in Fig 3.4 was excised from the gel and recovered by β agarase digestion of the gel slice (2.8.7). The recovered fragment was employed in a ligation reaction (2.8.3) with the pre-digested plasmid and the ligation reactions used to transform *E. coli* DH5α cells (2.4.1). No colonies were recovered after transforming the DH5α cells with the products of the ligation reaction, and repeated attempts at direct ligation of the purified fragment and transformation also failed to generate any colonies. It was considered that the failure in the subcloning procedure was the result of the poor recovery of the excised D4 fragment after gel purification. As the direct excision of the fragment from the D4 clone generated only small amounts of the fragment DNA in relation to the amount of plasmid digested (Fig 3.4), it was decided to amplify the mutagenised fragment by PCR (2.8.5) prior to digestion (2.8.4) with the restriction enzymes *Bsm*I and *Eco*RI.

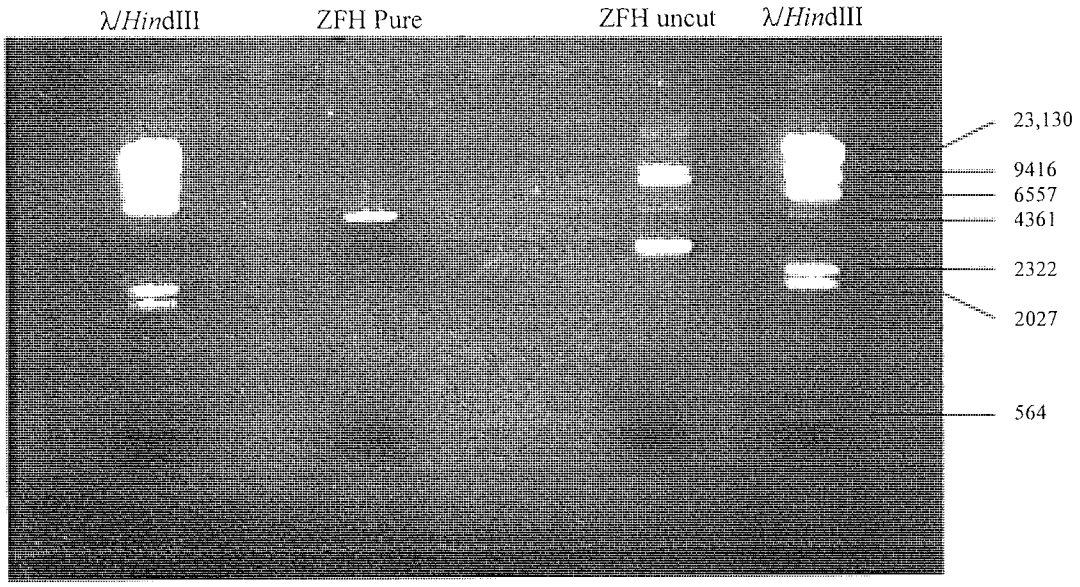


Fig 3.3 Estimation of DNA recovery after gel purification of the pGEX-ZFH construct by visualisation on 1 % agarose gel. Key to figure: λ/*Hin*dIII = 500ng λ *Hin*dIII Ladder; ZFH Pure = 5 μL of the recovered pGEX-ZFH DNA; ZFH Uncut = 850 ng of native pGEX-ZFH DNA.

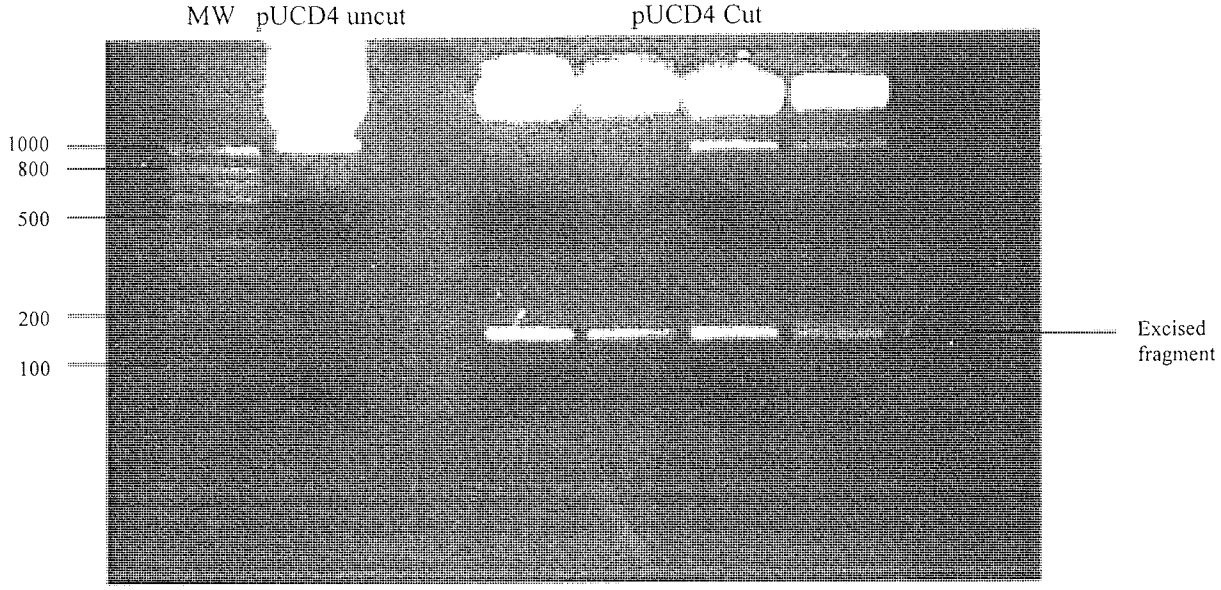


Fig 3.4 Gel purification of the excised pUCD4 mutagenised fragment (2 % Low melting point agarose gel). Key to figure: MW = 500ng 100 bp DNA ladder (Bioline); PUCD4 uncut = 2 μg pUCD4 DNA; pUCD4 cut = 2 μg pUCD4 DNA, digested with *Eco*RI and *Bam*HI. The fragment excised for purification is highlighted in the figure.

The mutagenised fragment was amplified from the pUCD4 construct using the pUC19 forward and reverse PCR primers (Appendix A1) and a standard PCR reaction (2.8.5). The use of a proof reading polymerase enzyme such as *Pfu* DNA polymerase, which posses 3' – 5' exonuclease activity and is able to "correct" missincorperated bases, was considered unnecessary in the amplification of such a small DNA insert. Four PCR reactions were carried out to generate a stock of amplified DNA prior to purification. The products of the PCR reaction were visualized using agarose gel electrophoresis (Fig 3.5a).

The products of the PCR reaction in lane 2 of Fig. 3.5a contained a non specific band of amplified DNA. The reactions represented by lanes 3, 4 and 5 of fig. 3.5a were pooled and the amplicons gel purified (2.5.2). DNA was recovered from the gel slices by β agarase digestion (2.8.7). Recovery of the DNA was estimated by visualization on an agarose gel (2.5.1). The concentration of the recovered DNA (pUCD4 Pure, Fig. 3.5b) was estimated to be 80 ng/μl. The recovered products were subjected to restriction enzyme digestion with the enzymes *Bsm*I and *Eco*RI (2.8.4) and ligated (2.8.3) into the pre-digested pGEX-ZFH construct. The ligation reactions were subsequently used to transform *E. coli* DH5α cells (2.4.1)

The transformation of the *E. coli* cells generated no recoverable clones. Subsequent attempts to subclone the pUCD4 fragment directly into the pGEX-ZFH gene also failed.

As the direct subcloning of the pUCD4 fragment directly into the pGEX-ZFH gene was unsuccessful, the mutagenesis of the ZFH gene was carried out in the pUC19 based ZFH6 construct. The use of a high copy number cloning vector during the subcloning procedure, was expected to ameliorate the difficulties experienced when subcloning the mutagenised fragment directly into the ZFH gene contained in the pGEX-2TK expression vector.

A

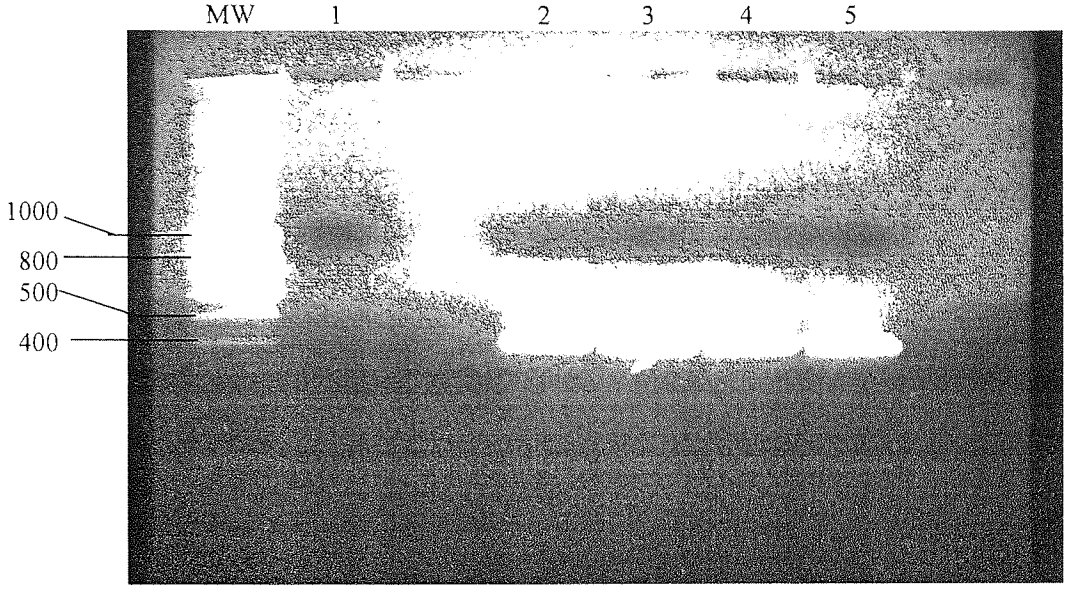


Fig 3.5 (A) Agarose gel analysis of the amplified pUCD4 PCR products. Key to figure: MW = 500 ng 100 bp ladder (Bioline); 1 = 5 % Negative control reaction; 2 – 5 = 5 % pUCD4 PCR reaction.

B

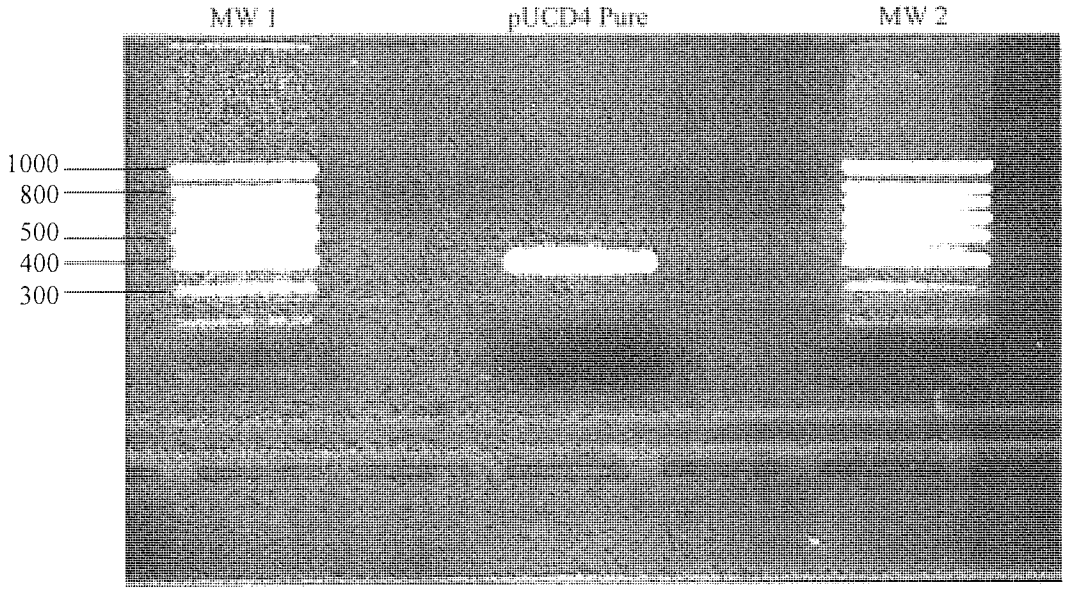


Fig 3.5 (B) Estimation of the recovery of the pUCD4 fragment after gel purification (2 % agarose gel). Key to figure: MW 1 = 500 ng 100 bp DNA Ladder (Bioline); MW 2 = 1 μg 100 bp DNA Ladder (Bioline); PUCD4 Pure = 5 μL purified pUCD4 amplification products.

The ZFH6 plasmid was digested with the enzymes *Bsm*I and *Eco*RI (2.8.4) and the digested plasmid treated with CIP (2.8.1) to prevent religation of the parental vector. The phosphatased vector was gel purified (2.5.2) and the DNA recovered by β agarase digestion of the gel slice (2.8.7). The recovery of the DNA was estimated by visualisation on an agarose gel (2.5.1). The mutagenised pUCD4 fragment was excised from the vector by restriction digest with the enzymes *Bsm*I and *Eco*RI (2.8.4). The fragment was gel purified (2.5.2) and ligated into the pre-digested ZFH6 vector (2.8.3). The ligation reactions were employed in the transformation (2.4.1) of *E. coli* DH5α cells. Transformation results are shown in Table 3.1

TRANSFORMING DNA	COLONIES RECOVERED
No Plasmid	0
Self Ligation	2
Insert Ligation	8

Table 3.1 Transformation results obtained when subcloning the mutagenised pUCD4 fragment into the ZFH6 construct.

The recovered colonies were screened by PCR (2.8.5), using pUC19 forward and reverse primers (Appendix A1). Two colonies produced amplicons corresponding to the expected 465 bp. These products were subsequently digested with the restriction enzymes *Hin*dIII and *Bsi*WI (2.8.4) to verify the presence of the introduced recognition sites for these enzymes. The digested products were analysed using agarose gel electrophoresis (2.5.1) (Fig 3.6). The digestion of the amplified products confirmed the presence of the *Hin*dIII and *Bsi*WI restriction sites in the ZFH gene. The two recovered clones named ZFHD4 and ZFHD8 were subjected to sequence analysis (2.8.8), which confirmed the correct sequence of the mutagenised gene in both of the recovered clones. Plasmid DNA was recovered from clone ZFHD4 using large scale plasmid preparation (2.4.7).

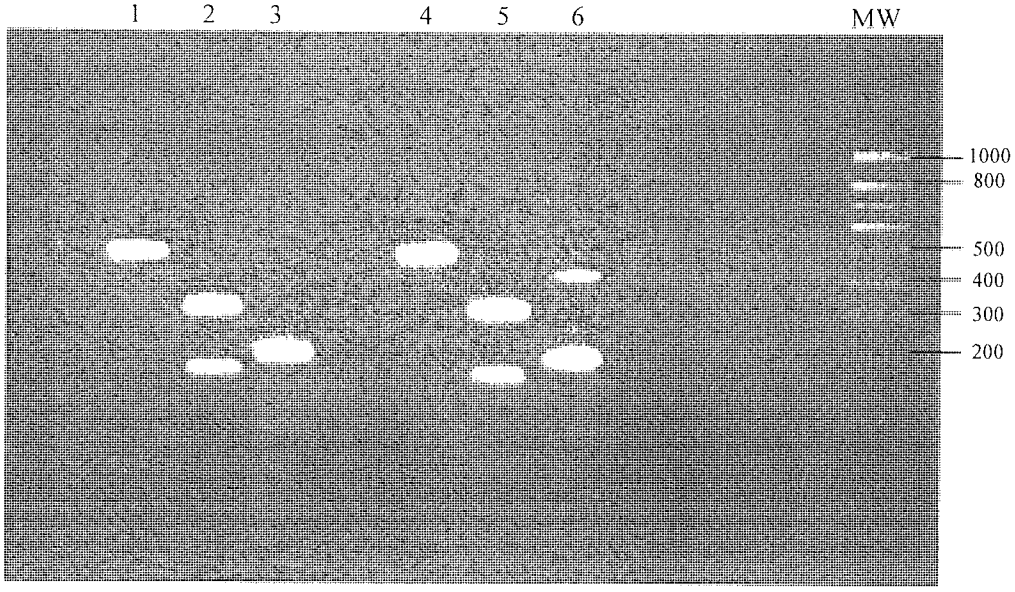


Fig 3.6 Restriction digest analysis of PCR products amplified from clones recovered after the subcloning of the mutagenised pUCD4 fragment into the ZFH6 construct (2 % agarose gel). The predicted size of the undigested PCR product is calculated to be 465bp, corresponding to the observed sizes in lanes 1 and 4. The predicted size of products after digestion with *Bsi*WI can be calculated as 277 and 188bp, lanes 2 and 5. The predicted size of the product formed after digestion with *Hind*III can be calculated as 240 and 225bp lanes 3 and 6. The observed product sizes in these lanes concurs somewhat with the predicted sizes of the digested products, the slight discrepancy in observed and predicted sizes was expected to be due to the differing ionic strengths of the restriction digest buffer or the presence of single stranded DNA at the restriction termini of the digested products. Key to figure: 1) 15 µl ZFHD4 PCR product; 2) 15 µl ZFHD4 PCR product digested with *Bsi*WI; 3) 15 µl ZFHD4 PCR product digested with *Hind*III; 4) 15 µl ZFHD8 PCR product; 5) 15 µl ZFHD8 PCR product digested with *Bsi*WI; 6) 15 µl ZFHD8 PCR product digested with *Hind*III; MW = 500 ng 100 bp DNA ladder (Bioline).

The expression vector pGEX-2TK was digested (2.8.4) with the enzymes *Bam*HI and *Eco*RI and treated with CIP (2.8.1) to prevent religation of the parental vector. The mutagenised ZFH gene was excised from the pUC19 vector by digestion (2.8.8) with the enzymes *Bam*HI and *Eco*RI. The fragment was gel purified (2.5.2) and ligated (2.8.3) into the pre-digested pGEX-2TK expression vector. The ligation reactions were used to transform (2.4.1) *E. coli* DH5α cells.

Three colonies recovered from the transformation were screened by PCR (2.8.5) using the pGEX 5' and pGEX 3' primers (Appendix A1). Three colonies produced amplicons corresponding to the expected 461 bp product. The PCR products amplified from these clones were digested with the enzymes *Hin*dIII and *Bsi*WI (2.8.4) and analysed using agarose gel electrophoresis (2.5.1). Each of the three clones were digested with both *Hin*dIII and *Bsi*WI. The three clones were subject to sequence analysis (2.8.9), one of the clones, named pGEX-ZFHM6 was identified as containing the correct sequence corresponding to the mutagenised ZFH gene. Sequence data obtained from the pGEX-ZFHM6 construct is contained in Figure 3.7.

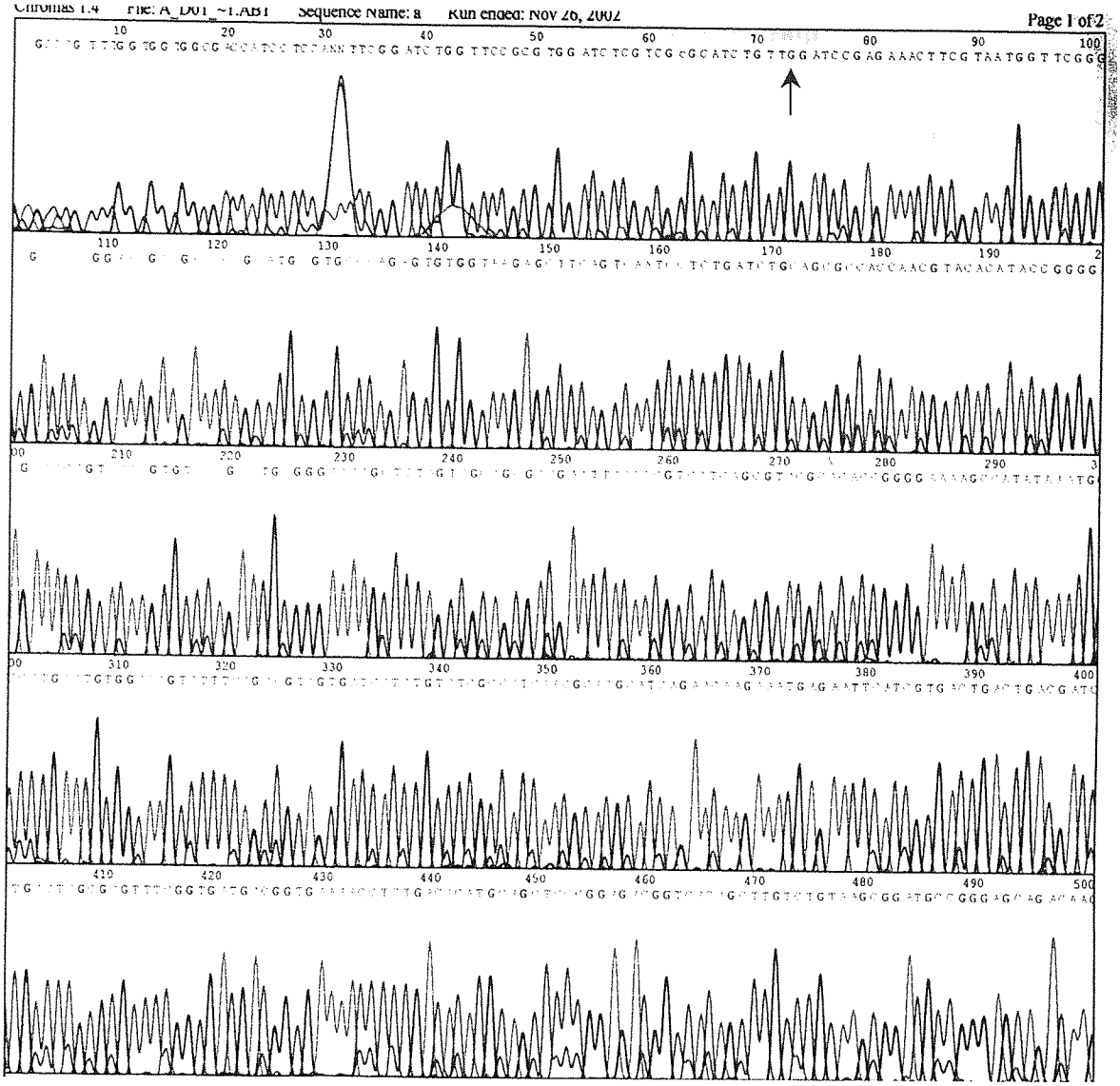


Fig 3.7 Sequence analysis of the pGEX-ZFHM6 gene. Base 73 highlighted in the sequence represents the start of the ZFHM6 gene. The predicted sequences of all genes used in the study are contained in appendix A2.

3.3 Optimisation of the ZFHM6 Gene for Library Construction

The MAX technique is used in the creation of randomised DNA cassettes, which are subsequently employed in the cassette mutagenesis of a target gene to encode a protein library randomised in a controlled fashion at specific positions. Prior to work carried out in the creation and analysis of the MAX randomised cassettes, factors which may affect the representation of any randomised library, irrespective of the randomisation method employed in their mutagenesis, were addressed.

The term representation, when applied to randomised gene or protein libraries, is used to describe the number of separate species contained within the library in comparison to the theoretical numbers of possible clones in a randomised library. Thus, in a fully representative gene library equivalent numbers of clones of each species will be present. The representation of gene libraries is influenced by diverse factors, however these factors generate non-representative libraries in only two ways: under representation and misrepresentation.

Under representation refers to the generation of a library in which the number of recovered clones or phage plaques in a phage display library is insufficient to account for all of the possible randomised species which are generated in the randomisation procedure. For example in the cloning of a gene library randomised at three codon positions by the use of the codon 5'-NNN-3' the transformation must yield a minimum of 262144 clones in order that each possible randomised species be represented once within that library.

Misrepresentation refers to a library in which a disproportionate number of library species are generated from the same genetic sequence. A library which may be numerically representative may still be misrepresented at the genetic level, as the prevalence of certain library species may cause other library constituents to be underrepresented. Misrepresentation within gene libraries may be the result of several diverse factors (Section 1.4), the effect of one of these factors, the synthesis of an abundant population of the parental plasmid during library creation, was addressed by the modification of the pGEX-ZFHM6 construct.

3.3.1 Prevention of self ligation of the pGEX-ZFHM6 construct.

The selectional hybridisation technique used to create the MAX randomised cassettes provides a powerful tool with which to address many of the problems which prevent the creation of representative gene libraries. The technique does however preclude the use of CIP to prevent the religation of the pGEX-ZFHM6 construct during the cassette mutagenesis step, as the addition of phosphate groups to the mutagenic cassette may lead to the formation of concatamers during the selectional hybridisation procedure (Fig 3.8)

As demonstrated in Fig 3.8, the palindromic nature of the restriction termini of the cassette, permits the hybridisation of the template oligonucleotide to other template strands in an antiparallel orientation. The α MAX oligonucleotides may also hybridise to copies of themselves in the same fashion. The addition of phosphates to these termini would permit the covalent attachment of these hybridised strands by the action of ligase. Consequently a cloning strategy antipodal to that of standard cassette mutagenesis must be adopted, in which the mutagenic cassette possesses no 5' phosphates and the phosphate groups on the restriction termini of the cohesive termini of the pGEX-ZFHM6 construct are maintained after it has been prepared for cassette mutagenesis.

Self-ligations are often recovered in routine subcloning experiments, despite the treatment of digested vector DNA with CIP to prevent religation of the parental plasmid. Exclusion of this step would therefore be expected to significantly increase the number of self-ligations recovered after cassette mutagenesis during library construction. This raised concerns regarding library representation, as the pGEX-ZFHM6 construct encodes the QDR-RER-RHR zinc finger protein designed by Desjarlais and Berg (1993). Prior work has demonstrated that this protein possesses a high affinity for its DNA target site and reduced affinity for permutations of this site (Desjarlais and Berg, 1993). The generation of clones encoding the QDR-RER-RHR zinc finger protein by self-ligation of the parental plasmid may result in the over expression of this protein within the zinc finger libraries. If the QDR-RER-RHR zinc finger protein is present in much higher quantities than other library constituents, the signal generated by the interaction of this protein with DNA target sites, other than its

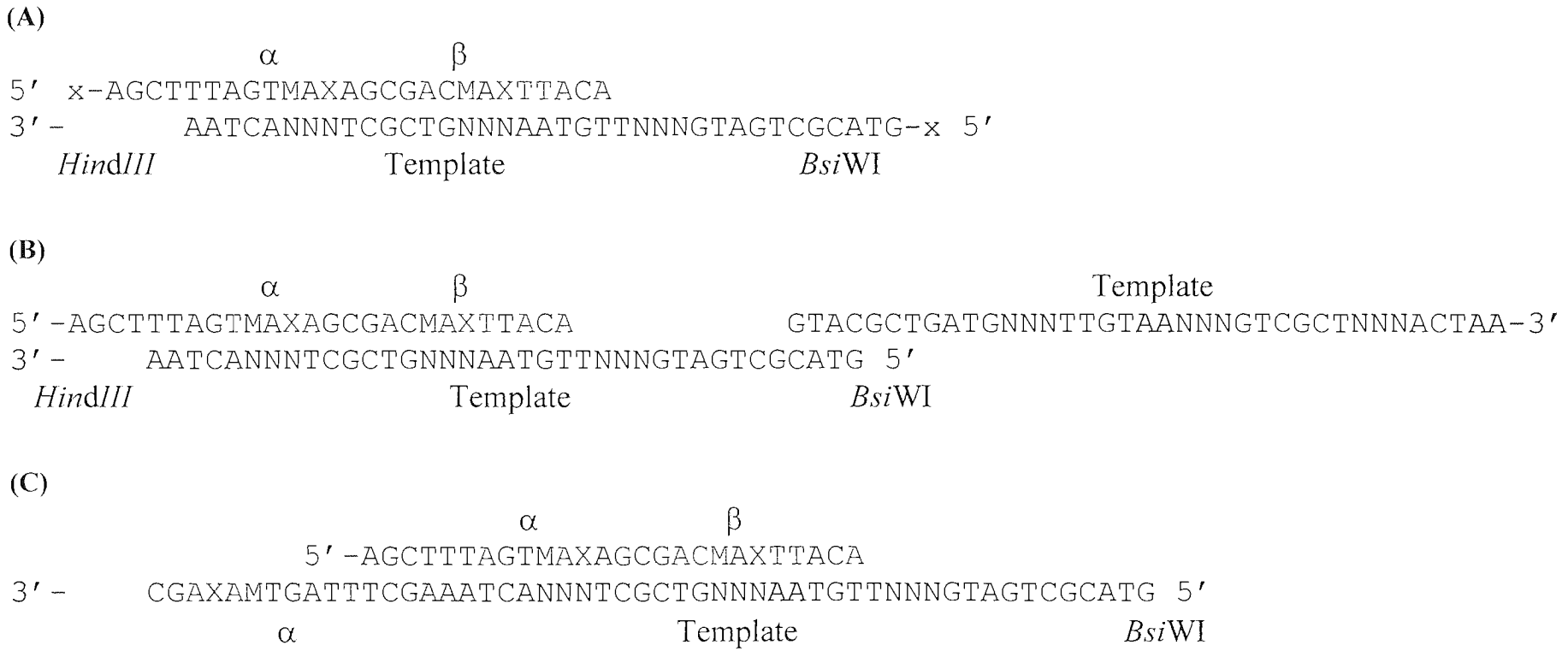


Fig 3.8 Mechanisms by which concatamer formation could arise upon the addition of a 5' phosphate group to the template oligonucleotide or α max selection oligonucleotides. A) Standard Max cassette bearing *Hin*dIII and *Bsi*WI cohesive termini with no terminal 5' phosphates (denoted by –X): B) Binding of the *Bsi*WI recognition sequence by the template strand, the template strand binding to the palindromic *Bsi*WI site of another template strand in the opposite orientation. This positions the 5' end of the template adjacent to the 3' termini of the γ selection oligonucleotide. The addition of a 5' phosphate to the template strand would permit the covalent bonding of these two strands by ligase: C) Hybridisation of an α MAX oligonucleotide to the palindromic *Hin*dIII site of another α MAX oligonucleotide. Inclusion of a 5' phosphate on the α MAX oligonucleotide would allow the bonding of the α MAX oligonucleotide to the template strand. The covalent bonding of the α MAX selection oligonucleotide to the template strand of a cassette, and the bonding of the template strand to the γ MAX selection oligonucleotide by ligase results in the generation of regions of single stranded DNA to which further hybridisation of selection or template oligonucleotides can occur.

specific target site, as a result of mass action, may mask the signal of other high affinity zinc finger proteins which are present in the library at equal concentrations. This problem was addressed by a reappraisal of the subcloning strategy used in the cassette mutagenesis, leading to a redesign of the pGEX-ZFHM6 gene.

As the selectional hybridisation technique precludes the use of CIP to remove the phosphate groups from the digested M6 construct, other means of preventing the re-ligation of the parental plasmid during the subcloning of the MAX randomised cassettes were considered. The use of gel purification to isolate the M6 plasmid construct from the 37 bp fragment, excised by digestion with the restriction enzymes *Hin*dIII and *Bsi*WI, was quickly discounted. The preclusion of re-ligation of the parental plasmid using this technique relies upon the complete digestion of the plasmid population by both HindIII and *Bsi*WI. Thus pGEX-ZFHM6 plasmid constructs which had been digested by only one of the enzymes would still posses the capacity for self ligation. Due to the small size of the excised DNA (37 bp) and the relatively large size of the pGEX-ZFHM6 construct (5262 bp) plasmids digested with only one of the restriction enzymes would be indistinguishable from plasmids cut with both enzymes upon analysis by agarose gel electrophoresis.

The subcloning strategy employed in the cassette mutagenesis involves the removal of the 37 bp fragment encoding a region of the middle finger of the QDR-RER-RHR zinc finger triplet, from the ZFHM6 gene. The gene is reconstituted by the insertion of a 37 bp cassette randomised at the three base contacting positions using the MAX technology (Section 1.7). As libraries may only be generated after the removal of the 37bp fragment an alternative sequence was redesigned to eliminate the potential problem of library misrepresentation caused by the religation of the parental plasmid.

The initial step in the design of this sequence involved the introduction of a further restriction enzyme recognition sequence between these two sites. In the design of the gene it was considered that the insertion of a further recognition sequence, unique within the gene/plasmid construct, would allow the gene to be "pre-cut" prior to digestion with the enzymes *Hin*dIII and *Bsi*WI. The rational behind this design is shown schematically in Fig 3.9.

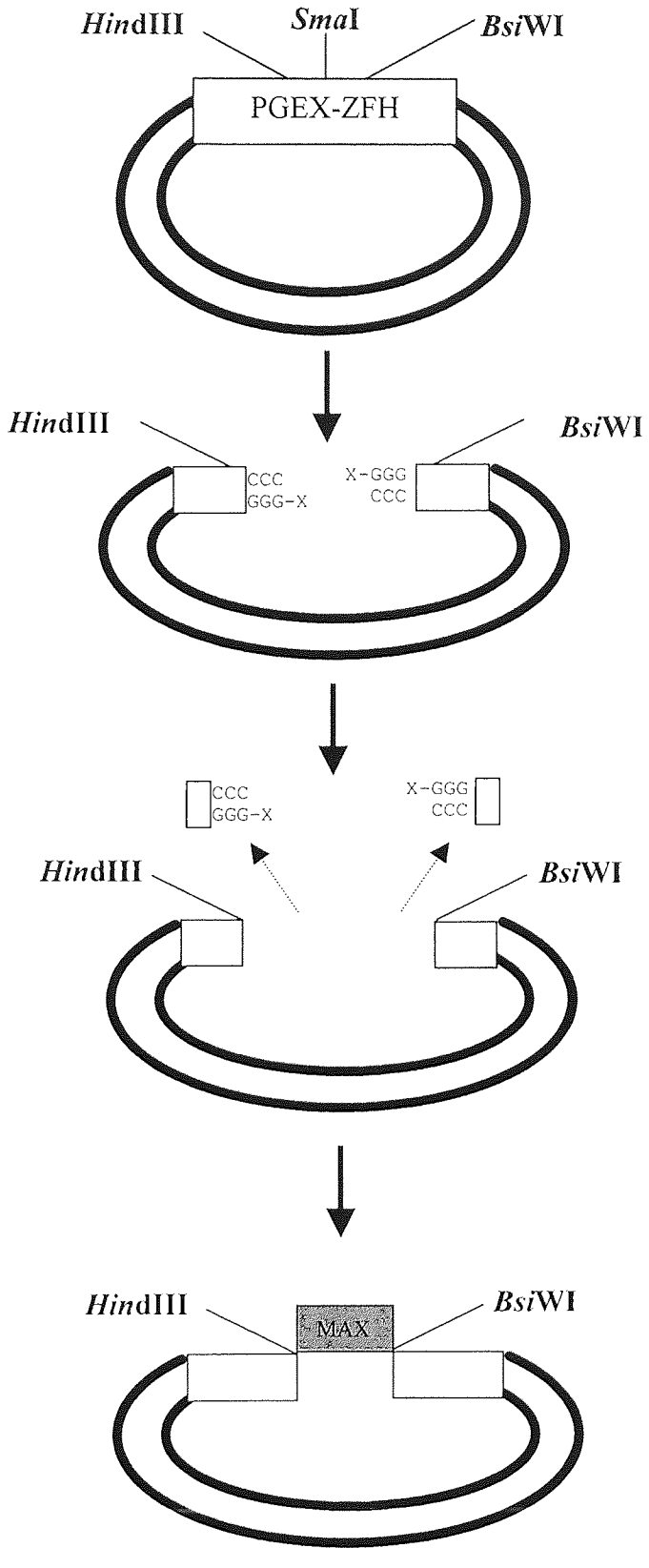


Fig 3.9. The subcloning strategy permitted by the redesign of the pGEX-ZFH6 gene. Pre-digestion of the plasmid with the enzyme *Sma*I permits the use of CIP to prevent religation of the parental plasmid. Subsequent digestion with the enzymes *Hin*dIII and *Bsi*WI maintains the 5' phosphate groups on the cohesive termini of these sites, facilitating the subcloning of a non phosphorylated MAX cassette.

As demonstrated in Fig 3.9, the pre-digestion of the pGEX-ZFHM6 construct by the restriction enzyme *Sma*I permits the use of CIP to remove the phosphate groups from the 5' termini of the digested plasmid. In addition, pre-digestion of the supercoiled plasmid with a single restriction enzyme allows the efficiency of the digestion reaction to be assessed by agarose gel electrophoresis, prior to CIP treatment. The plasmid can subsequently be digested with the restriction enzymes *Hin*dIII and *Bsi*WI, which removes the inserted sequence as two small fragments. Incomplete digestion of the plasmid by *Hin*dIII or *Bsi*WI results in a pGEX-ZFHM6 plasmid bearing a *Hin*dIII or *Bsi*WI cleavage site at one terminus and a phosphatased *Sma*I cleavage site at the other terminus, preventing religation of the parental plasmid.

The recognition sequence for the enzyme *Sma*I was included in the redesign of the inserted sequence after sequence analysis showed this site to be unique within the pGEX-ZFHM6 gene/plasmid construct. The use of the recognition sequence for *Sma*I was considered desirable as DNA cleavage by this enzyme generates blunt ended DNA fragments. The ligation of blunt ended DNA fragments is less efficient than the ligation of DNA bearing cohesive termini as the hybridisation of the cohesive termini bring the 5' phosphate and the 3' hydroxyl group into close proximity. Ligation of blunt ends is routinely carried out at low temperature (4°C), whereas the ligation reaction of the cassette mutagenesis reaction is carried out at higher temperatures (>14°C). This minimises the possibility of self-ligation of pGEX-ZFHM6 plasmids which still posses *Sma*I termini with 5' phosphate groups attached, although the formation of such species is expected to be rare.

3.3.2 Preclusion of the generation of the QDR-RER-RHR zinc finger protein by Regeneration of the parental plasmid

The design of the inserted sequence also considered the possibility that colonies may still be recovered as a result of transformation with the pGEX-ZFHM6 plasmid. In the design of the insert sequence to replace the sequence between the *Hin*dIII and *Bsi*WI sites, the size of the insert DNA was limited to 20 bp. The introduction of a smaller insert was also expected to facilitate PCR identification of colonies recovered as a result of self ligation of the parental plasmid. Limiting the size of the insert to exactly 20 base pairs was carried out to shift the reading frame of the ZFHM6 gene downstream of the *Hin*dIII site. As the *Hin*dIII site precedes the base contacting residues of the second finger of the zinc finger triplet, the shift in the reading frame of the gene after this point perturbs the amino acid sequence of the QDR-RER-RHR protein to such an extent that only the first finger of the protein is functionally encoded. In libraries constructed in the frameshifted gene, colonies recovered as a result of re-ligation of the parental plasmid encode a frameshifted zinc finger protein, as the affinity of a single zinc finger for its target site is low (Krizek *et al.*, 1990), these proteins are not expected to affect the deconvolution of the zinc finger libraries.

3.3.3 Construction of the frameshifted zinc finger gene.

The designed sequence was synthesised as two complementary 20 bp oligonucleotides. The two oligonucleotides, termed Ins 1 and Ins 1R, when hybridised together form the Ins 1 insert sequence which contains the recognition sequence for the restriction enzyme *Sma*I, and possesses *Hin*dIII and *Bsi*WI cohesive termini for directional ligation into the ZFHM6 gene. The sequences of the Ins 1 and Ins 1R oligonucleotides and the hybridised Ins 1 insert DNA are shown in Fig 3.10.

INS I

5′ AGC TTC GTT CCC GGG ATG AC 3′

INS IR

5′ GTA CGT CAT CCC GGG AAC GA 3′

INS I (insert)

5′ ***AGC*** ***T***TC GTT **CCC GGG** ATG AC 3′

3′ AG CAA **GGG CCC** TAC TG***C*** ***ATG*** 5′

***Hind*III** ***Sma*I** ***Bsi*WI**

Fig 3.10 The nucleotide sequences of the two oligonucleotide INS I and INS IR. The sequences are also shown aligned as the hybridised INS 1 insert, designed to replace the 37 bp sequence between the *Hin*dIII and *Bsi*WI site of the ZFH gene. The *Hin*dIII and *Bsi*WI cohesive termini of the hybridised insert are shown in italics. The *Sma*I recognition site is underlined in bold face.

The pGEX-ZFHM6 construct was digested with the restriction enzymes *Hin*dIII and *Bsi*WI (2.8.4). The digested plasmid was treated with CIP (2.8.4) to remove the terminal 5' phosphate groups from the cohesive termini of the plasmid. The plasmid was purified by gel purification (2.5.2), to remove the CIP enzyme and to ensure the removal of the 37 bp excised fragment. The DNA was recovered from the gel by β agarose digestion (2.8.7) of the gel slices and the purified DNA quantitated by estimation on an agarose gel (2.5.1).

The Ins 1 and Ins 1R oligonucleotides were treated with PNK (2.8.2) and then hybridised together in a standard hybridisation reaction (2.9.2), to create the Ins 1 cassette. A ligation reaction (2.8.3) was prepared to subclone the Ins 1 cassette into the pre-digested pGEX-ZFHM6 plasmid. The ligation reaction was subsequently used in the transformation (2.4.1) of *E. coli* DH5α cells. Transformation results are shown in Table 3.2.

TRANSFORMING DNA	COLONIES RECOVERED
No Plasmid	0
Self Ligation	5
Insert Ligation	10

Table 3.2 Transformation results obtained in the subcloning of the Ins 1 cassette into the pGEX-ZFHM6 gene.

All recovered clones were screened using PCR (2.8.5). Three of the recovered clones gave a strong positive signal in the initial PCR screen (data not shown). These three clones were subjected to a further PCR screen (2.8.5) and the products of the PCR reaction digested (2.8.4) with the enzyme *Sma*I, to identify the presence of the inserted recognition sequence for this enzyme. The digestion of the amplified products was assessed using agarose gel electrophoresis (2.5.1), results are shown in Fig 3.11.

The agarose gel analysis of the digested products (Fig 3.11) showed that the products amplified from each of the recovered clones were digested with the enzyme *Sma*I. Control reactions amplified from the pGEX-ZFHM6 construct remained uncut, indicating the digestion of products amplified from the recovered clones must be

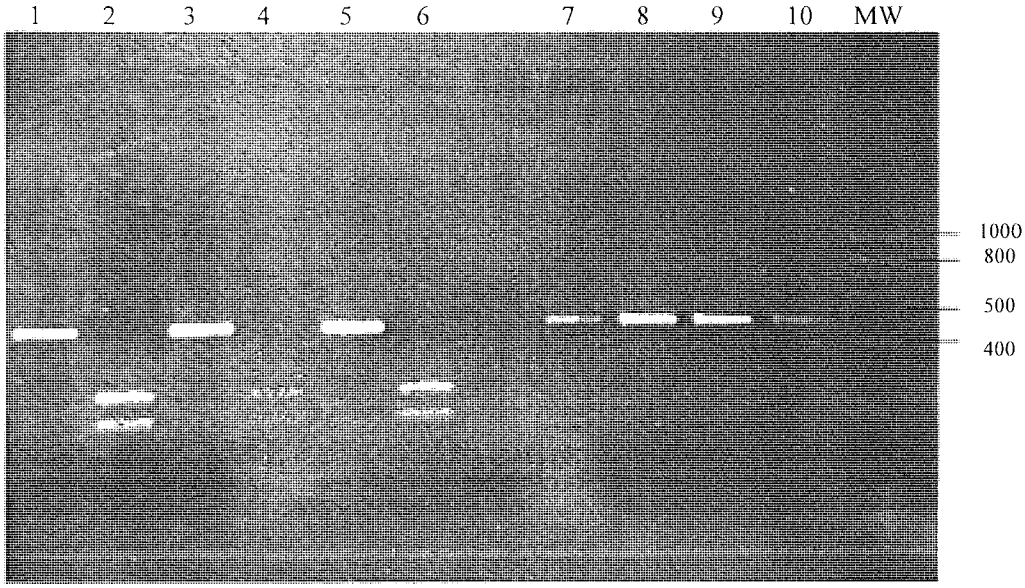


Fig 3.11 Restriction digest analysis of the PCR products amplified when screening clones recovered in the subcloning of the INS 1 insert sequence into the pGEX-ZFHM6 construct. Key to figure: Lanes 1, 3 and 5) 15µL of undigested PCR product corresponding to the predicted 448bp amplicon expected when amplifying the pGEX-ZFMA3 plasmid; Lanes 2, 4 and 6) 15µL PCR product digested with *Sma*I, the expected product sizes after digestion can be calculated as 252 and 196 bp, the products appearing to correspond to these sizes although accurate discrimination is difficult in the gel photo due to the low intensity of the 100 – 300bp bands of the molecular weight marker; Lanes 7 and 9) 15µL PCR products amplified from the parental pGEX-ZFHM6 construct, included as control reactions and corresponding to the 465bp amplicon expected after amplification of this construct; Lanes 8 and 10) 15µl of the products amplified from the parental pGEX-ZFHM6 construct digested with *Sma*I, corresponding to the 465bp amplicon, as no *Sma*I site is present in the parental construct; MW = 500 ng 100bp DNA ladder (Bioline).

the result of the insertion of the recognition sequence for *Sma*I. The three clones were subjected to sequence analysis (2.8.8). The sequence data (Fig 3.12) identified one of the clones, termed ZFMA3, as containing the zinc finger gene with the Ins 1 sequence correctly inserted between the *Hin*dIII and *Bsi*WI recognition sequences of the original zinc finger gene. DNA from the clone pGEX-ZFMA3 was subsequently recovered using large scale plasmid preparation (2.4.7).

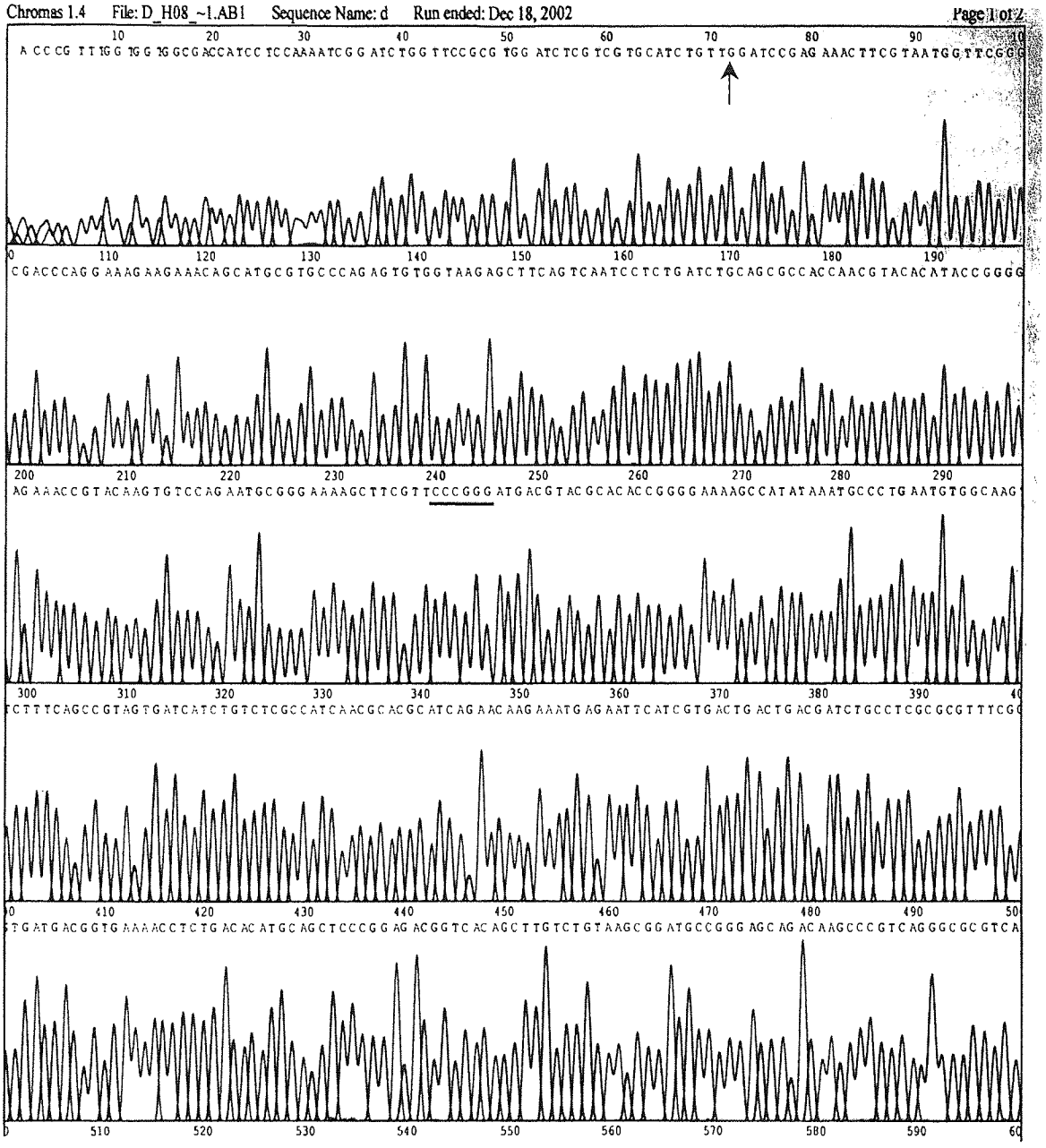


Fig 3.12 Sequence analysis of the pGEX-ZFMA3 construct. Base number 72 highlighted in the sequence represents the start of the pGEX-ZFMA3 gene. The introduced *SmaI* recognition site is underlined. The predicted sequences of all genes used in the experimentation are contained in appendix A2.

Chapter 4 Library Construction

In the previous chapter, the plasmid pGEX- ZFMA3 was constructed successfully (3.3.2) to provide a zinc finger gene amenable to randomisation by cassette mutagenesis. Preliminary library construction was then carried out creating MAX oligonucleotide cassettes containing the full complement of 20 MAX codons at each position. It was anticipated that complete (rather than partial) randomisation would provide the greatest number of clones for subsequent analysis and highlight any potential areas of the technique requiring optimisation.

4.1 Initial MAX Randomisation Strategy

The initial MAX strategy involved hybridisation of a template strand and three pools of selection oligonucleotides, each complementary to a consecutive region of the template (Fig. 4.1). Once hybridised together, three selection oligonucleotides and a single template strand generate a 37bp cassette with *Hin*dIII and *Bsi*WI overhangs for directional ligation into the pGEX-ZFHMA3 construct.

The template strand was randomised with the codon NNN at the positions encoding the base contacting residues of the middle finger of the QDR-RER-RHR protein. The selection oligonucleotides were created by dividing the complementary sequence to the template strand into three groups of oligonucleotides. Each group consisted of 20 oligonucleotides that have a complementary sequence to the template, except for the MAX position which consists of one codon, representing the most abundant codon for that particular amino acid in the most highly expressed genes of *E. coli* (Nakamura *et al.*, 2000). The conserved regions of the selection oligonucleotide hybridise to their complementary sequence on the template strand, allowing the MAX region of the selection oligonucleotide to select the corresponding complementary sequence from the template possibilities. Sequences of these library oligonucleotides are listed in Fig. 4.2.

1) Pool of alpha selection oligonucleotides with the consensus sequence
5′-AGCTTTAGT**MAX**AGC-3′

2) Pool of beta selection oligonucleotides with the consensus sequence
5′-GAC**MAX**TTACA-3′

3) Pool of gamma selection oligonucleotides with the consensus sequence

4) The template strand which consists of the sequence
3′-AATCA**NNN**TCGCTG**NNN**AATGTT**NNN**GTAGTCGCATG-5′

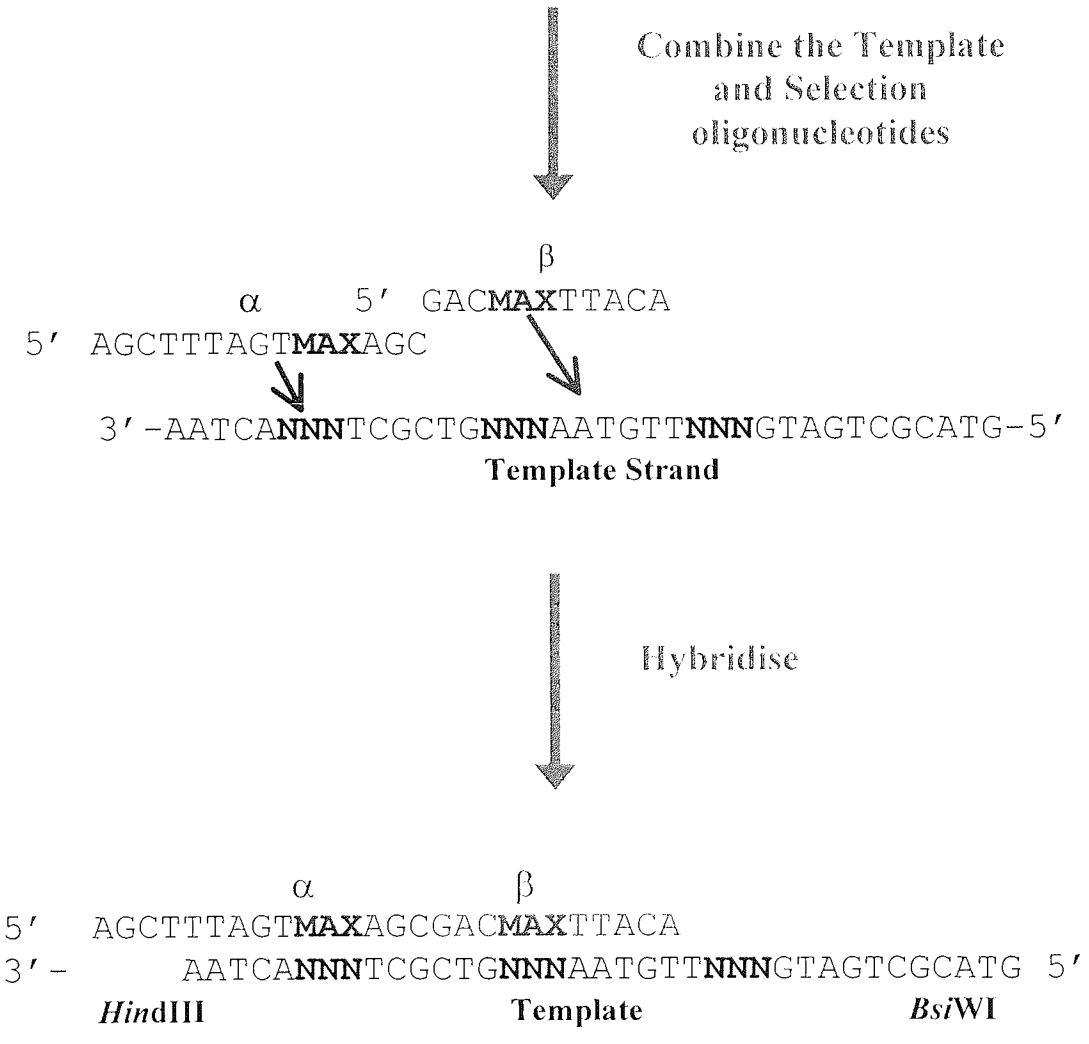


Fig 4.1 Generation of the MAX cassette by hybridisation of the alpha, beta and gamma selection oligonucleotides to the template strand. The additional 4 bp, 5' sequences of the template and alpha selection oligonucleotides generate the required *Hin*dIII and *Bsi*WI cohesive termini for directional ligation of the cassette into the pGEX-ZFMA3 vector. Randomised positions, denoted as MAX, are shown in bold face.

Selected Amino Acid	MAX Codon *(E. Coli)*	Sequence of Selection Oligonucleotide at Each Position 5'-3'		
		ALPHA (α)	BETA (β)	GAMMA (γ)
ALA (A)	GCG	AGCTTTAGTGCGAGC	GACGCGTTACA	AGCGCATCAGC
CYS (C)	TGC	AGCTTTAGTTGCAGC	GACTGCTTACA	ATGCCATCAGC
ASP (D)	GAT	AGCTTTAGTGATAGC	GACGATTTACA	AGATCATCAGC
GLU (E)	GAA	AGCTTTAGTGAAAGC	GACGAATTACA	AGAACATCAGC
PHE (F)	TTT	AGCTTTAGTTTTAGC	GACTTTTTACA	ATTTCATCAGC
GLY (G)	GGC	AGCTTTAGTGGCAGC	GACGGCTTACA	AGGCCATCAGC
HIS (H)	CAT	AGCTTTAGTCATAGC	GACCATTTACA	ACATCATCAGC
ILE (I)	ATT	AGCTTTAGTATTAGC	GACATTTTACA	AATTCATCAGC
LYS (K)	AAA	AGCTTTAGTAAAAGC	GACAAATTACA	AAAACATCAGC
LEU (L)	CTG	AGCTTTAGTCTGAGC	GACCTGTTACA	ACTGCATCAGC
MET (M)	ATG	AGCTTTAGTATGAGC	GACATGTTACA	AATGCATCAGC
ASN (N)	AAC	AGCTTTAGTAACAGC	GACAACTTACA	AAACCATCAGC
PRO (P)	CCG	AGCTTTAGTCCGAGC	GACCCGTTACA	ACCGCATCAGC
GLN (Q)	CAG	AGCTTTAGTCAGAGC	GACCAGTTACA	ACAGCATCAGC
ARG (R)	CGC	AGCTTTAGTCGCAGC	GACCGCTTACA	ACGCCATCAGC
SER (S)	AGC	AGCTTTAGTAGCAGC	GACAGCTTACA	AAGCCATCAGC
THR (T)	ACC	AGCTTTAGTACCAGC	GACACCTTACA	AACCCATCAGC
VAL (V)	GTG	AGCTTTAGTGTGAGC	GACGTGTTACA	AGTGCATCAGC
TRP (W)	TGG	AGCTTTAGTTGGAGC	GACTGGTTACA	ATGGCATCAGC
TYR (Y)	TAT	AGCTTTAGTTATAGC	GACTATTTACA	ATATCATCAGC
Template Strand		5'-GTACGCTGATGNNNTTGTAANNNGTCGCTNNNACTAA-3'		

Fig 4.2 Sequences of the original library oligonucleotides, used in the generation of the MAX randomised cassettes.

Each of the selection oligonucleotides complements 11bp of the template strand. In addition, the alpha oligonucleotides (which are 15bp in length) contain a 4bp sequence at the 5' end, which generates a *Hin*dIII overhang when hybridised correctly to the template strand (Fig 4.1). The MAX randomisation position of the alpha, beta and gamma oligonucleotides correspond to the codons encoding the base contacting residue at positions 13, 16 and 19 of the middle finger of the QDR-RER-RHR protein respectively (Desjarlais and Berg, 1993).

4.2 The Role of Temperature and Molarities in MAX Methodology

The hybridisation of each specific selection oligonucleotide to the template strand is temperature dependent, occurring at temperatures around the theoretical melting temperature (Tm) of the product of this interaction. The theoretical melting temperatures of the products created by the annealing of the α, β and γ oligonucleotides differ due to the difference in sequence between the complementary regions of these oligonucleotides. The theoretical annealing temperature of products formed from each of the 20 oligonucleotides in the α, β and γ pools of oligonucleotides differs due to the differing sequence of the MAX codon at each selection position. This difference in theoretical melting temperatures of the hybridisation of each selection oligonucleotide to its template complement was exploited during the hybridisation procedure. The selection oligonucleotides were added to the hybridisation reaction in equimolar concentrations with regard to the molar concentration of their respective complementary sequence in the template (Fig. 4.3).

The hybridisation reactions were then heated to 94°C to denature any secondary structure within the oligonucleotide mix and subjected to controlled cooling on a thermal cycler using the DVTCH hybridisation protocol (2.9.3). The Tm of the interaction between a specific selection oligonucleotide and the template strand is greatest when the two strands are fully complementary. However, at lower temperatures, hybridisations containing mismatched base pairs may be stabilised. It was anticipated that the use of an equimolar mixture of selection and template oligonucleotides would result in sequestration of the selection oligonucleotide with the highest theoretical melting temperature before the reaction fell to temperatures at which

NNN =															
AAA	ACA	AGA	ATA	CAA	CCA	CGA	CTA	GAA	GCA	GGA	GTA	TAA	TCA	TGA	TTA
AAC	ACC	AGC	ATC	CAC	CCC	CGC	CTC	GAC	GCC	GGC	GTC	TAC	TCC	TGC	TTC
AAG	ACG	AGG	ATG	CAG	CCG	CGG	CTG	GAG	GCG	GGG	GTG	TAG	TCG	TGG	TTG
AAT	ACT	AGT	ATT	CAT	CCT	CGT	CTT	GAT	GCT	GGT	GTT	TAT	TCT	TGT	**TTT**

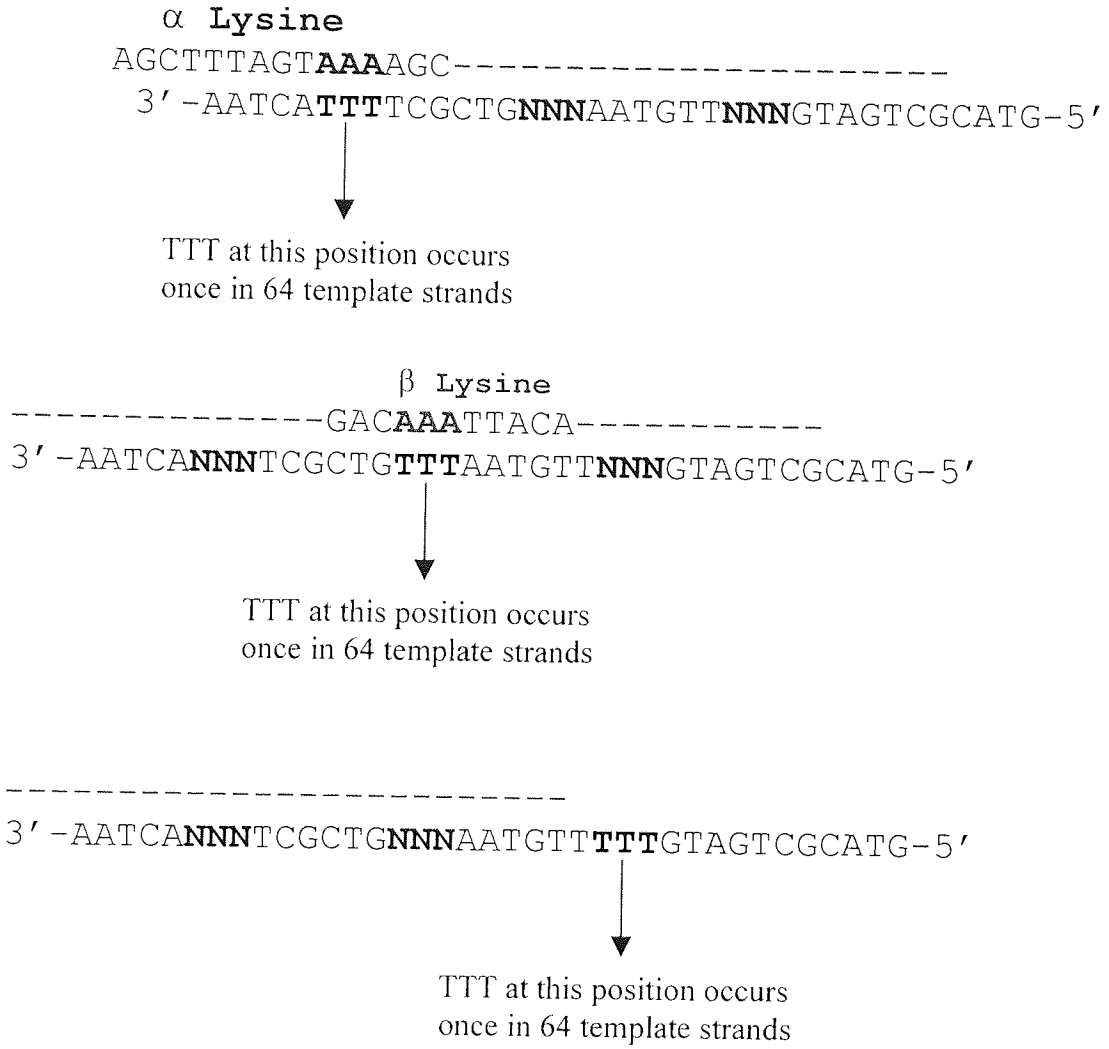


Fig 4.3 Calculation of the ratio of each selection oligonucleotide to corresponding complementary sequences in the template oligonucleotide. The ratio can be calculated as 1 : 64. This calculation is based upon the assumption that each selection oligonucleotide functions as an independent moiety in the hybridisation process. Each selection oligonucleotide is expected to associate with and hybridise to its complementary sequence, at temperatures approaching the Tm of that specific selection oligo/template product. The conserved complementary regions of the α, β and γ oligonucleotides promote association of the MAX randomised region with the NNN randomised region of the template. Using the α selection oligonucleotides as an example, each specific α oligonucleotide is selecting essentially from the sequence NNN after the α complementary region, which can be expected to occur once in 64 template molecules. The β and γ selection oligos also selecting their respective complementary templates from these same template strands. Only the selection oligonucleotides which encodes lysine at the MAX position of randomisation from the pools of α, β and γ selection oligonucleotides are represented in the figure. The MAX lysine codon (AAA) of each of the selection oligonucleotides is highlighted in bold face. The complementary codon TTT is highlighted in bold face on the template strand and in red text in the table of codon possibilities. The dashed regions represent potential hybridisation sites to which further selection oligonucleotides can anneal.

mismatched base pair interactions would be stabilised. As the cooling cycle progressed, each of the selection oligonucleotides were expected to hybridise to their respective complementary sequence, effectively removing them from the reaction and preventing mismatched annealing to non-complementary target sites. The reaction was then cooled to and maintained at 4°C to stabilise the electrostatic interactions between the hybridised oligos. Subsequent reactions were prepared on ice to prevent the hybridised cassettes returning to room temperature.

4.3 Initial Library Synthesis

Prior to hybridisation, the 5' ends of the β and γ selection oligonucleotides were phosphorylated by treatment with PNK (2.8.2), to permit ligation of the selection oligonucleotides. The α selection oligonucleotides and template oligonucleotides were not phosphorylated, to prevent the possibility of concatamer formation (Section 3.3.1, Fig 3.8). Randomised cassettes were generated in a hybridisation reaction (2.9.3) containing the full complement of α, β and γ selection oligonucleotides in equimolar amounts. A ligation reaction (2.8.3) was prepared to ligate the double stranded MAX cassettes into the pGEX-ZFMA3 vector, which had been previously pre-digested with the enzyme *Sma*I and treated with CIP (2.8.1) before digestion with the enzymes *Hin*dIII and *Bsi*WI (2.8.4). The ligation reactions containing the MAX randomised cassettes were performed as a standard ligation reaction (2.8.3) with the exception that only the expected number of double stranded, complete MAX cassettes were included in the 3:1 molar ratio of insert : vector DNA. The calculation was based upon the assumption that each selection oligonucleotide was bound to its respective complementary sequence in the template strands. In this assumption, the 8000 possible MAX cassettes have been selected from 262144 template possibilities, a ratio of 1:37.77. This ratio was then used to adjust the concentration of hybridised DNA to the correct 3:1 molar ratio. Ligations were prepared on ice before incubation at 14°C for 16-18 hours. The ligation reaction was then employed in the transformation of *E. coli* DH5α cells (2.4.1) and plated on LB media (2.1.1). Plasmid DNA was recovered from 20 colonies and subjected to sequence analysis (Lark Technologies 2.8.9).

Sequencing results showing the sequence through the 37bp mutagenised site are listed in Figure 4.4. The full sequences obtained from Lark technologies are contained in Appendix A. An initial survey of these results showed that only two clones (1 and 2) contained MAX codons at all three randomisation positions with the rest of the sequence being correct. A further three clones (4, 5 and 10) possessed MAX codons at all positions of randomisation but showed one or more point mutations in the conserved sequences. All these mutations were A – G transitions occurring directly after the MAX codon in the α section of the cassette. Mutations in cloned sequences are routinely observed when cloning synthetic DNA, possibly as result of cloning unmethylated DNA that is subsequently recognised by *E. coli* repair proteins. As the DH5α hosts are restriction deficient, this phenomena is often presented as recombinational events or insertions or deletions in the synthetic region of the cloned gene (see Discussion). The fact that this point mutation occurred in the same place in several clones and succeeded a guanosine base in all but one of the clones, suggested that it may be the result of the repair of mismatched MAX codons by the host, as the mutation occurs directly after the randomised region. This mutation occurs in the complementary region of the α selection oligonucleotides, which do not contain a guanosine at this position, suggesting that this point mutation resulted from repair of the inserted DNA by the host cell.

Three of the sequenced clones contained inserted bases within the randomised 37bp sequence. The insertion of a single base shifts the remaining sequence out of frame resulting in these clones being unable to code for functional zinc fingers. As the preliminary work was carried out to assess the MAX randomisation at a genetic level, these clones were included in the subsequent analysis.

Five clones (6,7,9,16 and 19) contained mixed codons. The term mixed codon was applied to a randomised codon containing one or more unidentified base/s which could not be accurately identified from analysis of the chromatogram. Analysis of the chromatograms of those sequences which contained an N in the MAX position, showed in some cases two discernable peaks of similar sizes. It was postulated that these peaks may be the result of a mixed population of the same plasmid, created when a mismatch in the randomised positions of the oligonucleotides cassettes was stabilised and ligated into the pGEX- ZFMA3 construct. As plasmid DNA is replicated in a circular fashion

1) met-trp-asn
GGGAAAAGCTTTAGTATGAGCGACTGGTTACAAAACCATCAGCGTACGCACACCGGGGA
AAA

2) ser-arg-ala
GGGAAAAGCTTTAGTAGCAGCGACCGCTTACAAGCGCATCAGCGTACGCACACCGGGGA
AAA

3) phe-asn-leu
GGGAAAAGCTTTAGTTTTAGCGACAATTTACCAACTGCATCAGCGTACGCACACCGGGG
AAAA

4) leu-gln-glu
GGGAAAAGCTTTAGTCTGGGCGACCAGTTACAAGAACATCAGCGTACGCACACCGGGGA
AAA

5) trp-ala-trp
GGGAAAAGCTTTAGTTGGGGCGACGCGTTGCAATGGCATCAGCGTACGCACACCGGGGA
AAA

6) ala-arg/trp-cag
GGGAAAAGCTTTAGTGCCAGCGACNGGTTACAACAGCATCAGCGTACGCACACCGGGGA
AAAG

7) ser/asn-trp-glu
GGGAAAAGCTTTAGTANCGGCGACTGGTTACAAGAACATCAGCGTACGCACACCGGGGA
AAA

8) gln-asp-lys
GGGAAAAGCTTTAGTCAAAGCGACGACTTACAAAAACATCAGCGTACGCACACCGGGGA
AAA

9) phe-xxx-leu/val
NGGGAAAAgCtTTAgTTTTAGCGaCNNTTTaCCAANTGCATCAGCGTACGCACACCGGG
GAAAA

10) leu-gln-glu
GGGAAAAGCTTTAGTCTGGGCGACCAGTTACAAGAACATCAGCGTACGCACACCGGGGA
AAA

11) gln-asp-lys
GGGAAAAGCTTTAGTCAAAGCGACGACTTACAAAAACATCAGCGTACGCACACCGGGGA
AAA

12) val-arg-gln
GGGAAAAGCTTTAGTATGAGCGACCGTTTACAACAGCATCAGCGTACGCACACCGGGGA
AAA

13) arg-cys-gly
GGGAAAAGCTTTAGTAGAAGCCGACTGCTTACAAGGCCAT---CGTACGCACACCGGG
GAAAA

14) phe-asp-glu
GGGAAAAGCTTTAGTTTTAGCGACGACTTACCAAGAGCCATCAGCGTACGCACACCGGG
GAAAA

15) his-thr-pro
GGGAAAAGCTTTAGTCACAGCGACACACTTACAACCTCATCAGCGTACGCACACCGGGG
AAAA

16) pro-asp/ile-his
GGGAAAAGCTTTAGTCCCAGCGACANTTTACAACATCATCAGCGTACGCACACCGGGGA
AAA

17) ser-trp-his
GGGAAAAGCTTTAGTAGTAGCGACTGGTTACAACATCATCAGCGTACGCACACCGGGGA
AAA

18) cys-gly-asn
GGGAAAaGCTTTAGTTGCaGCGACGGGTTACAAAACCATCAGCGTACGCACACCGGGGA
AAA

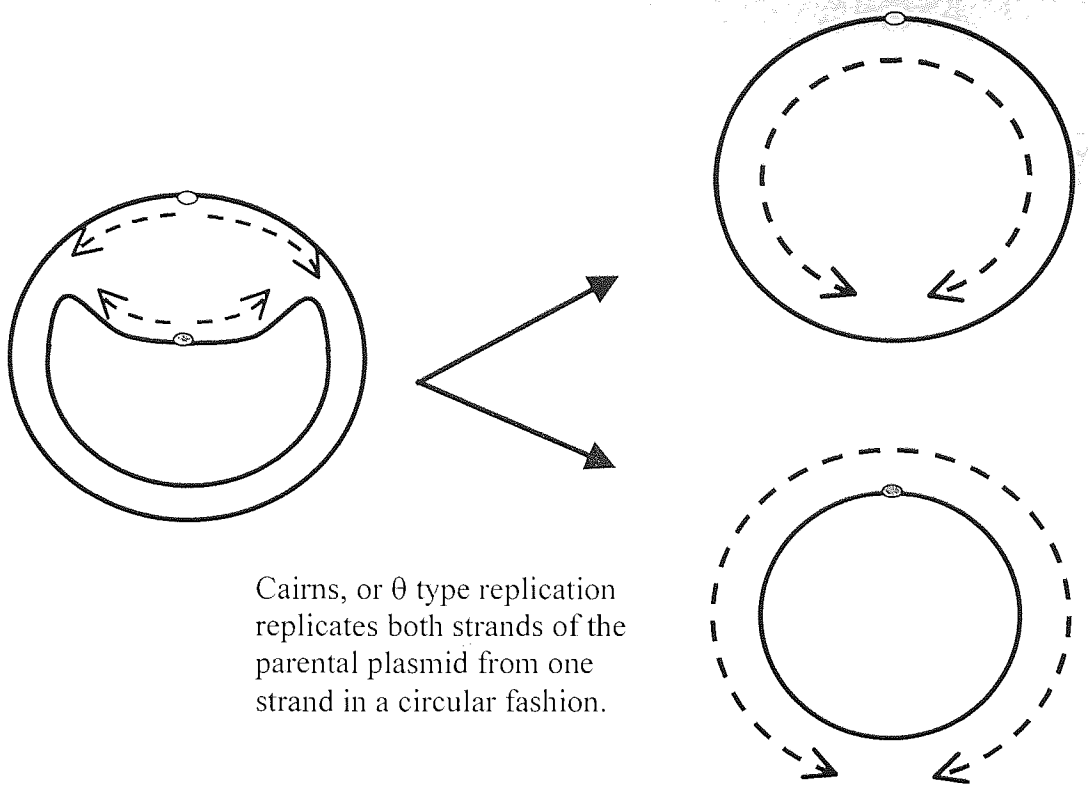
19) arg-arg/leu-gly
GGGAAAAGCTTTAGTCGGAGCGACCNCTTACAAGGCCATCAGCGTACGCACACCGGGGA
AAA

Fig 4.4 Sequence alignments of the randomised region of the ZFH gene, obtained from the 19 sequences recovered after the MAX randomisation of the pGEX-ZFMA3 vector. The amino acids encoded at the randomised positions of each clone are also shown. Key to figure: Lower case letters = Bases represented by N, subsequently discriminated by analysis of the chromatogram; Purple text = MAX codon sequence; Red Text = Non MAX codon sequence; All bases in blue text represent sequence abnormalities. Inserted bases are denoted by the inserted base highlighted in blue text. Deletions are denoted by a blue dash (-). Randomised codons highlighted in blue which contain the letter N, represent sequences in which the base represented by N could not be discriminated from a mixture of bases at that position. The possible amino acids encoded by randomised positions containing only one N are shown in the text.

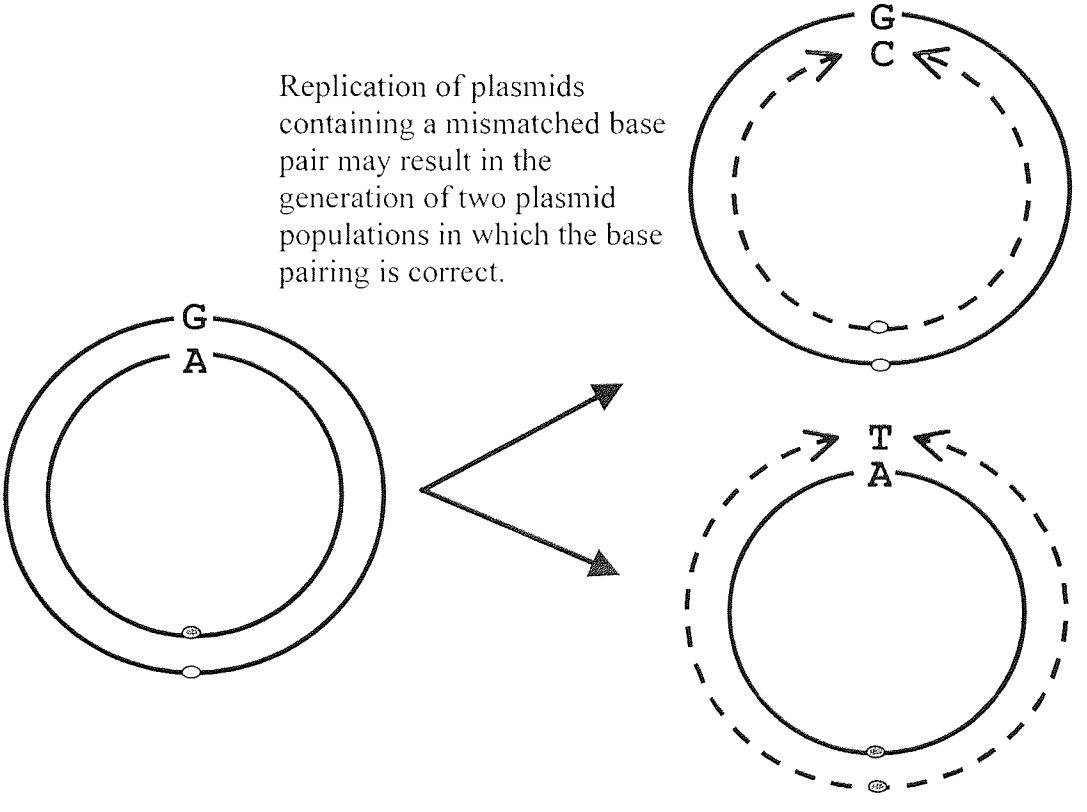
replication of a plasmid bearing a mismatch without any prior repair by the host may give rise to stable populations of two different plasmids (Fig. 4.5).

Finally, 12 clones contained non-MAX codons at one or more positions of randomisation. The generation of non-MAX codons in the clones was postulated to result from repair of mismatched inserted DNA by the host, or from DNA polymerase I activity, which may occur at a "nicked" substrate after incomplete ligation of the selection oligonucleotides. Repair of the DNA based upon the sequence of the template strand (NNN) may result in the generation of non-MAX codons.

The results were used to plot the graph in Figure 4.6. The γ position of randomisation demonstrated the highest number of MAX codons with only two non-MAX codons and 1 mixed codon, indicating the hybridisation of the γ selection oligonucleotide was occurring specifically. In contrast, only 10 MAX codons were observed at the α position of randomisation, one mixed codon was present at this position and the remaining 8 codons consisted of varied non-MAX codons. The number of non MAX codons at the α position in conjunction with the 3 A – G point mutations directly after the α MAX position in three clones suggested that the α selection oligonucleotide was capable of stabilising mismatched base pairing which then was undergoing repair by the *E. coli.* host. The β position of randomisation showed only 8 MAX codons, 7 non max codons and 4 mixed codons again suggesting that the β selection oligonucleotide was stabilising mismatches during hybridisation.



Cairns, or θ type replication replicates both strands of the parental plasmid from one strand in a circular fashion.



Replication of plasmids containing a mismatched base pair may result in the generation of two plasmid populations in which the base pairing is correct.

Fig 4.5 Schematic representation of the replication of a plasmid containing mismatched base pairs without prior repair by the host, postulated as the mechanism by which mixed population of codons were generated in the library. The same origin of replication of the plasmid is highlighted in yellow or green in the figure, to identify the strand of the parental plasmid from which the daughter plasmids were transcribed.

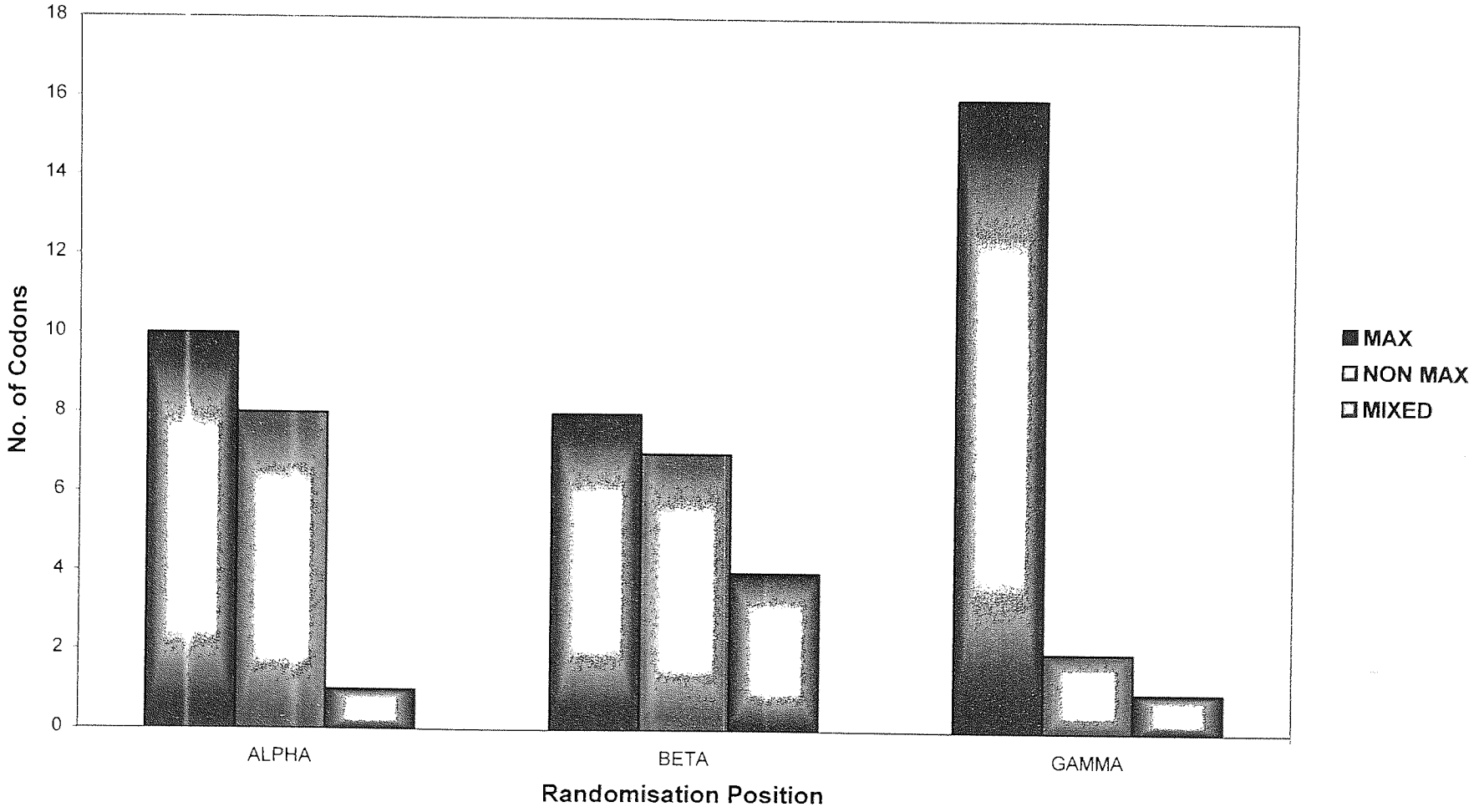


Fig 4.6 Graph demonstrating the identities of the codons present at each position of randomisation, in the 19 sequenced clones obtained from the cassette mutagenesis of the pGEX-ZFMA3 plasmid in the initial attempts at library construction.

4.4 Redesign of the selection oligonucleotides

The results of the initial library synthesis clearly demonstrated that the gamma selection oligos were functioning more effectively than the beta and alpha oligos, which prompted a re-examination of the selection oligonucleotide design.

When considered in terms of the number of base pair interactions made with the template strand, the design had split the three selection positions into equal length oligonucleotides. Figure 4.7 shows the complementary sequences of the selection oligonucleotides with the randomised areas denoted as MAX.

The most obvious difference between the α, β and γ oligonucleotides was the location of the sequence in which the randomised position occurred. The MAX position in the γ oligo occurs near to the 5' end, with only one adenosine flanking the 5' end of the MAX codon. In the α and β oligos, the shortest flanking sequence to the randomised area is 3bp at the 3' and 5' ends respectively. Both of these flanking sequences contain 2 G or C bases, which are capable of making three hydrogen bonding interactions with their complementary base on the template. The presence of these flanking sequences led to the assumption that the bonding of these sequences to their complementary sequence on the template strand may be stabilising base mismatches between the MAX region and the NNN regions of the template, analogous to the way in which mismatched base pairs are stabilised by flanking sequences in oligonucleotide directed mutagenesis. The sequence results from the first library construction, particularly in respect to the results obtained in the γ position of hybridisation suggested that removal of any flanking sequence succeeding the MAX randomisation position of the selection oligonucleotides would improve the specificity of their hybridisation to the template strand.

The redesign of the selection oligonucleotides moved the MAX position of randomisation to the end of each sequence. As the original selection oligonucleotides contained complementary sequences both upstream and downstream of the MAX positions it was apparent that the design would have to include an additional oligonucleotide encoding the remaining part of the complementary sequence.

Fig 4.7 The consensus sequences of the selection oligonucleotides. The randomised regions, denoted as MAX, are highlighted in bold face. The region of the α selection oligonucleotide, which is not complementary to the template strand, used to generate the *Hin*dIII cohesive termini is shown in italics.

Figure 4.8 demonstrates the two possible ways in which the MAX position of randomisation can be moved to the end of the three selection oligonucleotides. Both designs necessitate the inclusion of selection oligonucleotides of unequal length.

Placing the MAX randomisation position at the 3' end of the selection oligonucleotide generates a 12 base α MAX selection oligonucleotide which makes 8 base pairs with the template strand and 9 base β and γ oligonucleotides. Additionally a 7 base pair oligonucleotide termed ENDMAX encodes the remaining complementary sequence. The addition or deletion of a single base pair to one of the selection oligonucleotides in each design was not expected to unduly effect its hybridisation specificity, but simply alter the theoretical melting temperature at which hybridisation occurs.

In contrast, placing the MAX position at the 5' end of the selection oligonucleotides (Fig. 4.8b) generates a 9 base ENDMAX sequence which makes only 5 base pairs with the template strand, 9 bp α and β selection oligonucleotides and a 10 bp γ oligonucleotide. This would not preclude the generation of MAX cassettes as the temperature of the hybridisation reaction falls below the expected Tm of this interaction. However results of the initial library construction had raised some concern that the *E. coli* host may have "filled in" randomised positions from the "nicks" in the phosphodiester backbone of the inserted DNA. The design strategy in Fig. 4.8b places the ENDMAX oligonucleotide at the 5' end of the mutagenic cassette upstream of the α β and γ oligonucleotides. DNA polymerase I activity initiated from the 3' end of the ENDMAX oligonucleotide, could result in the generation of non MAX codons at all positions of randomisation, as the remaining selection oligonucleotides may be removed by the 5'- 3'exonuclease activity of the host polymerase enzyme.

The strategy shown in Fig. 4.8a was therefore adopted in the redesign of the selection oligonucleotides, placing the MAX randomisation position at the 3' end of the oligonucleotides. This design strategy may also result in the "filling in" of cloned cassettes, however the α selection oligonucleotide must be present to generate the *Hin*dIII cohesive termini of the cassette. As the α selection oligonucleotide must be present in cloned cassettes, employing this strategy provided a point of reference with which to compare the occurrence of non MAX codons at the β and γ positions.

(a)

(b)

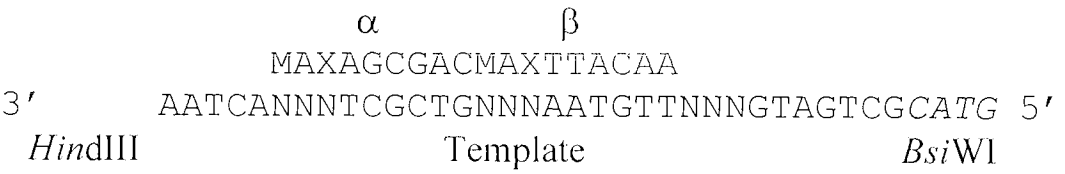


Fig 4.8 Schematic representation demonstrating the two possible ways in which the MAX randomisation position can be moved to the end of the respective oligonucleotides by the inclusion of a fourth oligonucleotide. **a**) The MAX randomisation position situated at the 3' end of each selection oligonucleotide. **b**) The MAX randomisation position situated at the 5' end of the selection oligonucleotides. The additional oligonucleotide (ENDMAX) required to generate the *Bsi*WI and *Hin*dIII cohesive termini of the cassette in each design, is shown in orange. The cohesive termini of the cassette are italicised

The sequences of the new MAX oligonucleotides designed in accordance with the strategy shown in Fig 4.8a are listed in Fig 4.9. The template sequence was maintained from the previous design.

4.5 Library Synthesis with Redesigned Oligonucleotides

In addition to the redesign of the oligonucleotides the conditions in which the selection oligonucleotides were hybridised to the template strand were also reassessed. The synthesis (2.9.1) of the redesigned selection oligonucleotides at higher synthesis yields than those obtained previously, provided the opportunity to further manipulate the conditions of the hybridisation reaction.

Hybridisation buffers were tested which would promote association of the selection oligonucleotides with the template strand and would not inhibit the subsequent ligation of the generated MAX randomised cassettes into the pGEX-ZFMA3 vector. The hybridisation buffers were based upon those used with T4 DNA ligase. Ligation reactions often rely upon the hybridisation of cohesive termini and this is reflected in the ionic environment of the buffers used in this reaction. In addition macromolecules such as PEG and BSA are often included to promote the association of these regions of single stranded DNA.

4.6 Pre-Ligation of Randomised DNA Cassettes

The use of a hybridisation buffer compatible with T4 DNA ligase also enables the "pre-ligation" of the selection oligonucleotides after hybridisation to the template strand (Fig 4.10). Sealing the single stranded nicks between each selection oligonucleotide after hybridisation generates a more stable double stranded DNA cassette. As the Tm of the pre-ligated cassette is increased, the interaction between the template and selection oligonucleotides is less likely to be perturbed during downstream processes. In addition, mutagenesis of the pGEX-ZFMA3 vector with a pre-ligated MAX cassette, generates a plasmid bearing only two nicks in the phosphodiester backbone of the double stranded DNA, as would be generated by conventional cassette mutagenesis, which would not be expected to promote DNA repair in the host.

Selected Amino Acid	MAX Codon *(E. coli)*	Sequence Of Selection Oligonucleotide At Each Position 5'-3'			
		ALPHA (α)	BETA (β)	GAMMA (γ)	ENDMAX
ALA (A)	GCG	AGCTTTAGTGCG	AGCGACGCG	TTACAAGCG	CATCAGC
CYS (C)	TGC	AGCTTTAGTTGC	AGCGACTGC	TTACAATGC	CATCAGC
ASP (D)	GAT	AGCTTTAGTGAT	AGCGACGAT	TTACAAGAT	CATCAGC
GLU (E)	GAA	AGCTTTAGTGAA	AGCGACGAA	TTACAAGAA	CATCAGC
PHE (F)	TTT	AGCTTTAGTTTT	AGCGACTTT	TTACAATTT	CATCAGC
GLY (G)	GGC	AGCTTTAGTGGC	AGCGACGGC	TTACAAGGC	CATCAGC
HIS (H)	CAT	AGCTTTAGTCAT	AGCGACCAT	TTACAACAT	CATCAGC
ILE (I)	ATT	AGCTTTAGTATT	AGCGACATT	TTACAAATT	CATCAGC
LYS (K)	AAA	AGCTTTAGTAAA	AGCGACAAA	TTACAAAAA	CATCAGC
LEU (L)	CTG	AGCTTTAGTCTG	AGCGACCTG	TTACAACTG	CATCAGC
MET (M)	ATG	AGCTTTAGTATG	AGCGACATG	TTACAAATG	CATCAGC
ASN (N)	AAC	AGCTTTAGTAAC	AGCGACAAC	TTACAAAAC	CATCAGC
PRO (P)	CCG	AGCTTTAGTCCG	AGCGACCCG	TTACAACCG	CATCAGC
GLN (Q)	CAG	AGCTTTAGTCAG	AGCGACCAG	TTACAACAG	CATCAGC
ARG (R)	CGC	AGCTTTAGTCGC	AGCGACCGC	TTACAACGC	CATCAGC
SER (S)	AGC	AGCTTTAGTAGC	AGCGACAGC	TTACAAAGC	CATCAGC
THR (T)	ACC	AGCTTTAGTACC	AGCGACACC	TTACAAACC	CATCAGC
VAL (V)	GTG	AGCTTTAGTGTG	AGCGACGTG	TTACAAGTG	CATCAGC
TRP (W)	TGG	AGCTTTAGTTGG	AGCGACTGG	TTACAATGG	CATCAGC
TYR (Y)	TAT	AGCTTTAGTTAT	AGCGACTAT	TTACAATAT	CATCAGC
Template Strand		5'-GTACGCTGATGNNNTTGTAANNNGTCGCTNNNACTAA-3'			

Fig 4.9 Sequences of the redesigned oligonucleotides used in the generation of the MAX randomised cassettes.

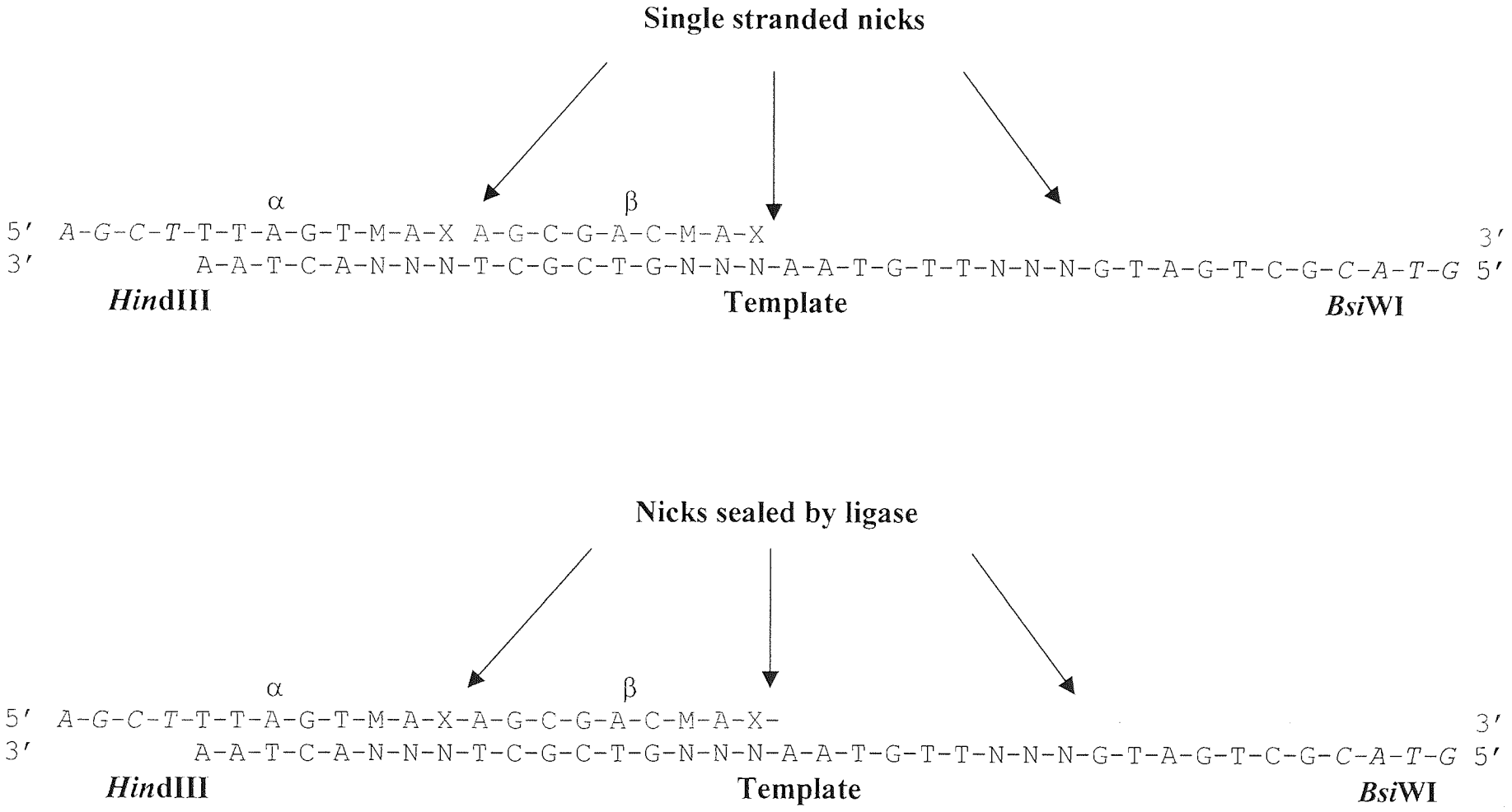


Fig 4.10 Schematic representation of the "Pre-ligation" of the MAX randomised cassette. Ligase is used to seal the nicks in the phosphodiester backbone of the hybridised selection oligonucleotides, generating a contiguous strand of DNA.

Hybridisation buffers 1 (2.2.5) and 2 (2.2.6) were based upon conventional ligase buffers. Hybridisation buffer 1 contained PEG 800 to promote macromolecular crowding of the oligonucleotide DNA. The heat labile components (ATP and DTT) were omitted from both buffers. To test the buffers each buffer was used in the religation of λ DNA, which had been pre-digested with *Hin*dIII. Stocks of hybridisation buffers 1 and 2 were heated to mimic a hybridisation reaction. After cooling, ATP and DTT were added at appropriate concentrations (2.9.5) and the digested λ DNA was re-ligated in the presence of each. The reactions were incubated at 14°C, and frozen at -20 °C after the time intervals listed in Figure 4.11 and 4.12. Prior to analysis by agarose gel electrophoresis, the samples were incubated at 65°C to denature the ligase. Results are shown in Figures 4.11 and 4.12. The figures show that ligation of the digested fragments has occurred in both buffers within 10 minutes of the addition of ligase.

4.7 Analysis of Pre-ligated and Conventionally Hybridised Cassettes

Hybridisation reactions (2.9.5) were prepared using each of the two buffers, each reaction containing the full complement of α, β, γ and ENDMAX selection oligonucleotides in equimolar amounts to their complementary sequences on the template strand. The β, γ and ENDMAX oligonucleotides were phosphorylated with PNK (2.8.2) prior to hybridisation. Replicates of each reaction were then aliquotted into two tubes (A and B) and hybridised (2.9.5). Subsequent to hybridisation ATP (2.2.17), DTT (2.2.18) and one Weiss unit of ligase were added to one of the replicates of each of the hybridisation reactions (A) and the reactions incubated at 14°C overnight.

Ligation reactions (2.8.3) were prepared to ligate the double stranded MAX cassettes generated in the pre-ligated and conventional hybridisation reactions into the pGEX-ZFMA3 vector, which had been previously pre-digested with the enzyme *Sma*I (2.8.4) and treated with CIP (2.8.1) prior to digestion with the enzymes *Hin*dIII and *Bsi*WI (2.8.4). These ligation reactions were then employed in the transformation (2.4.2) of *E. coli* DH5α cells. The numbers of colonies recovered from these transformations were counted to assess the affect of pre-ligation of the MAX cassette upon clone recovery. The results are shown graphically in Figure 4.13.

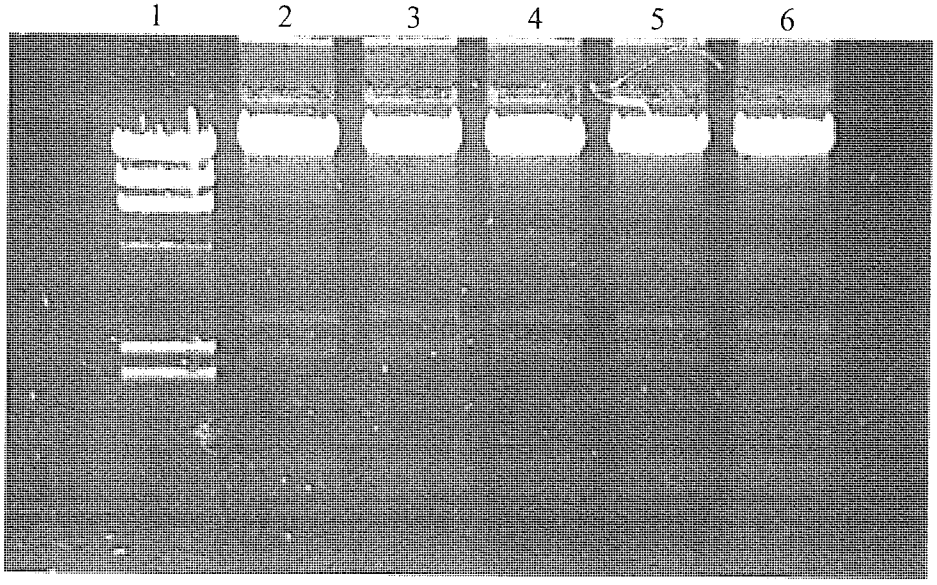


Fig 4.11 Ligation of *Hin*dIII fragments of λ DNA in hybridisation buffer 1, visualised on a 1 % agarose gel. The total volume of each ligation reaction, corresponding to 500 ng *Hin*dIII digested λ DNA and 1 Weiss unit T4 DNA ligase, was loaded in each well. Key to figure: 1) Time 0 (no ligase); 2) 10 mins after the addition of ligase; 3) 20 mins after the addition of ligase; 4) 30 mins after the addition of ligase; 5) 60 mins after the addition of ligase; 6) 90 mins after the addition of ligase.

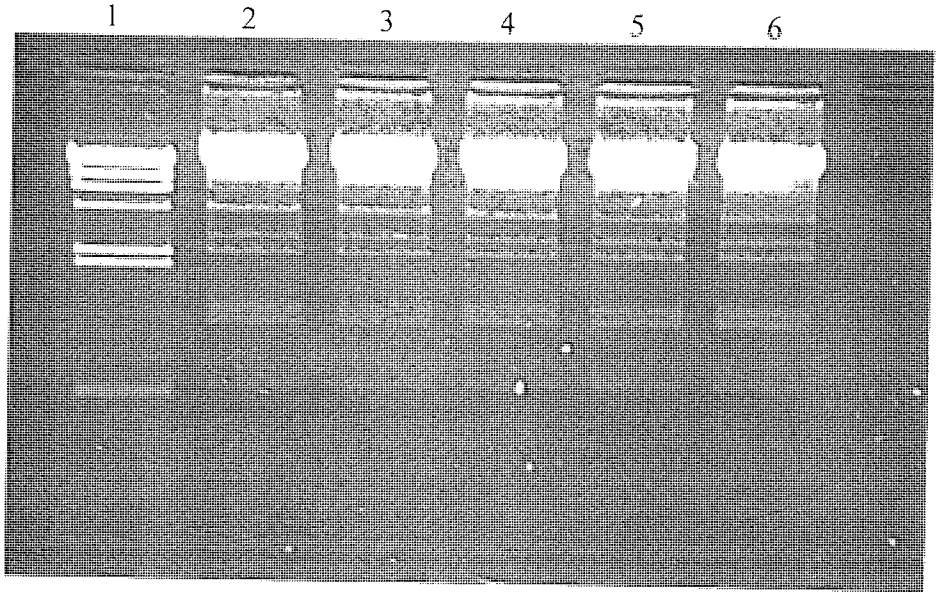


Fig 4.12 Ligation of *Hin*dIII fragments of λ DNA in hybridisation buffer 2, visualised on a 1 % agarose gel. The total volume of each ligation reaction corresponding to 500 ng *Hin*dIII digested λ DNA and 1 Weiss unit T4 DNA ligase, was loaded in each well. Key to figure: 1) Time 0 (no ligase); 2) 10 mins after the addition of ligase; 3) 20 mins after the addition of ligase; 4) 30 mins after the addition of ligase; 5) 60 mins after the addition of ligase; 6) 90 mins after the addition of ligase.

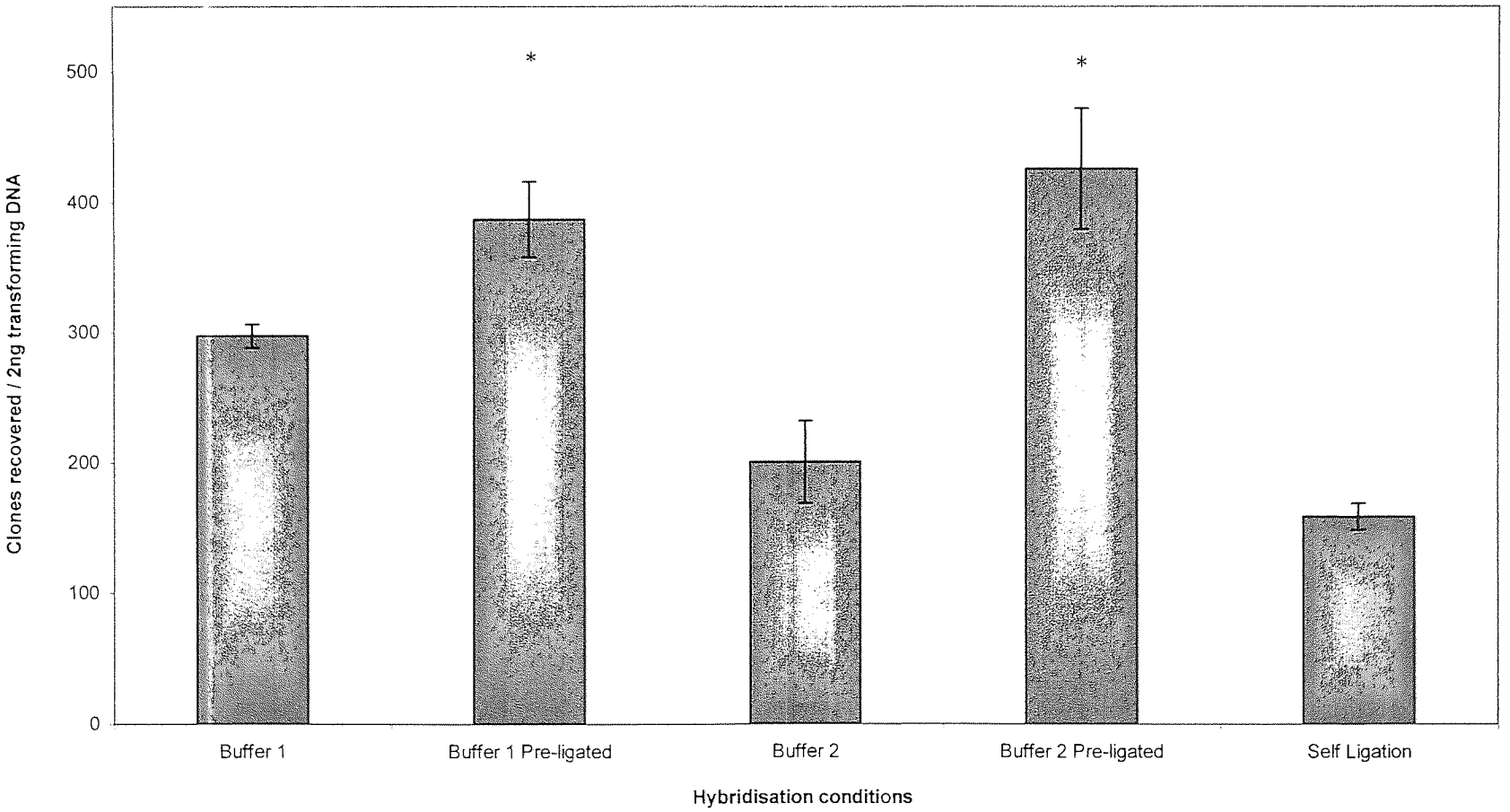


Fig 4.13 Graph showing clone recovery after Pre-ligated and conventionally hybridised MAX cassettes were employed in the cassettes mutagenesis of the pGEX-ZFMA3 construct. Results are means and S.E.M of 8 replicates. ($P = 0.0051$ for difference in recovery between hybridised and pre-ligated inserts in Buffer 1. $P = 0.0054$ for difference in recovery between hybridised and pre-ligated inserts in Buffer 2, using a paired t test). Clones recovered after self ligation of the pGEX-ZFMA3 vector are included in the chart for comparison.

In both buffers, pre-ligation of the selection oligonucleotides increased the numbers of recovered clones (P = 0.0051 Buffer 1, P = 0.0054 Buffer 2, paired t test, GraphPad version 3.02) suggesting that the pre-ligated DNA cassettes may be more stable than conventionally hybridised cassette. The average number of clones recovered after pre-ligation of the cassette was similar in both buffers. The difference in recovery when using pre-ligated and conventionally hybridised cassettes generated using hybridisation buffer 1 was smaller than the difference in recovery when cassettes were hybridised in buffer 2. This suggested that cassettes generated using hybridisation buffer 1 were stabilised prior to ligation, possibly as a result of macromolecular crowding promoting interaction between the selection oligonucleotides and the template strand.

4.8 Sequence Analysis of Pre-ligated and Conventionally Hybridised Cassettes

Clones generated from cassette mutagenesis using pre-ligated and non ligated cassettes were sequenced. As clone recovery from pre-ligated and conventionally hybridised cassettes was similar in buffer 1, samples of these clones were sequenced to asses the affect of pre-ligation on the inserted DNA sequence after cloning. Sequence alignments of the inserted DNA are shown in Figure 4.14.

Four of the sequence reactions prepared using the conventionally hybridised cassettes (Fig. 4.14a) failed to generate sequence data. One clone was shown to result from the religation of the parental pGEX-ZFMA3 vector. Two clones contained the expected *Hin*d III recognition sequence with the downstream sequence completely rearranged, possibly as a result of recombination of the inserted DNA. Of the remaining clones only one clone possessed MAX codons at each position of randomisation (Fig. 4.14a, No. 9) this sequence however also contained an inserted base. Four of the remaining clones contained inserted and deleted bases and the two clones which did not contain insertions or deletions, contained one or more none MAX codons at positions of randomisation. The sequences of the clones generated with pre-ligated cassettes (Fig. 4.14b) also contained two clones which failed to generate any sequence data. Three of the sequenced clones contained inserted bases within the sequence and three of the clones contained the correct sequence, with no insertions or deletion, but possessed non-MAX codons at one or more positions of randomisation.

1 **Self Ligation**

GGGAAAAGCTTCGTTCCCGGGATGACGTACGCACACCGGGGAAAA

2 **No sequence data generated**

3 **No sequence data generated**

4 **No sequence data generated**

5 **No sequence data generated**

6 **tyr-cys-phe**

GGGAAAAGCTTTAGTTATAGCGACTGTTTACAATTCCATCAGCGTACGCA
CACCGGGGAAAA

7 **lys-xxx-xxx**

GGGAAAAGCTTTAGTAAAAGCGACGAGTTTACAACCTCCATCAGCGTACG
CACACCGGGGAAAA

8 **Rearranged sequence**

GGGAAAAGCTTAGTGAACTGACGCGGGCAGACCCGCGTTTGGTAATATCC
TGCAAC

9 **cys-lys-thr**

GGGAAAAGCTTTAGTTGCAGCGACAAATTACCAAACCCATCAGCGTACGC
ACACCGGGGAAAAA

10 **Rearranged sequence**

GGGAAAAGCTTAATCCGATACCAACAAGAATTAATAACTGAAAAAAACGA
TCGCCTTGCA

11 **arg-asp-thr**

GGGAAAAGCTTTAGTAGGAGCGACGATTTACAAACCCATCAGCGTACGCA
CACCGGGGAAAA

12 **pro-xxx-xxx**

GGGAAAAGCTTTAGTCCGAGCGACGAT------CATCAGCGTACG
CACACCGGGGAAAA

Fig 4.14a Sequence data obtained from clones recovered after the cassettes mutagenesis of pGEX-ZFMA3 with conventionally hybridised MAX cassettes generated in hybridisation buffer 1. The amino acids encoded at the randomised positions of each clone are also shown. Key to figure: Purple text = MAX codon sequence: Red text = Non MAX codon sequence. All bases in blue text represent sequence abnormalities. Inserted bases are denoted by the inserted base highlighted in blue text. Deletions are denoted by a blue dash (-).

13 No sequence data generated
14 No sequence data generated
15 No sequence data generated

16 leu-asp-xxx
GGGAAAAGCTTTAGTCTCAGCGACGATTTACAACGAACATCAGCGTACGC
ACACCGGGGAAAA

17 asn-lys-thr
GGGAAAAGCTTTAGTAACAGCGACAAATTACAAACGCATCAGCGTACGCA
CACCGGGGAAAA

18 pro-xxx-xxx
GGGAAAAGCTTTAGTCCGAGCGACCAGTTTACAACCAACATCAGCGTACG
CACACCGGGGAAAAT

19 val-val-gln
GGGAAAAGCTTTAGTGTGAGCGACGTGTTACAACAGCATCAGCGTACGCA
CACCGGGGAAAA

20 ile-asp-tyr
GGGAAAAGCTTTAGTATTAGCGACGATTTACAATATCATCAGCGTACGCA
CACCGGGGAAAA

21 leu-asp-cys
GGGAAAAGCTTTAGTCTGAGCGACGACTTACAATGCCATCAGCGTACGCA
CACCGGGGAAAA

22 thr-ala-xxx
GGGAAAAGCTTTAGTACCAGCGACGCATTAGCCATCGCCATCAGCGTACG
CACACCGGGGAAAA

23 val-val-gln
GGGAAAAGCTTTAGTGTGAGCGACGTGTTACAACAGCATCAGCGTACGCA
CACCGGGGAAAAG

24 met-stop-arg
GGGAAAAGCTTTAGTATGAGCGACTAGTTACAAAGGCATCAGCGTACGCA
CACCGGGGAAAA

Fig 4.14b Sequence data obtained from clones recovered after the cassettes mutagenesis of pGEX-ZFMA3 with pre-ligated MAX cassettes generated in hybridisation buffer 1. Amino acids encoded at the randomised positions of each clone are also shown. Key to figure: Purple text = MAX codon sequence: Red text = Non MAX codon sequence. All bases in blue text represent sequence abnormalities. Inserted bases are denoted by the inserted base highlighted in blue text. Deletions are denoted by a blue dash (-).

Three of the sequenced clones contained MAX codons at all positions of randomisation with no insertions or deletions. Two of the clones with MAX codons at all positions of randomisation, contained the same randomised bases (sequences 19 and 23 Fig. 4.14b). As the hybridised cassettes were generated with all 20 selection oligonucleotides, the cassettes encoded all 8000 possible zinc finger proteins. It was considered unlikely that two identical clones would occur at random in such a small sample of sequenced clones, suggesting that amplification of the VAL-VAL-GLN clone may have occurred prior to the plating of the *E. coli* cells on selective media. No clones recovered from the mutagenesis with the pre-ligated cassette showed sequence rearrangements.

The results suggested that the pre-ligation of the MAX cassette resulted in better clone recovery, and that the sequence of the inserted DNA was not adversely affected, the technique was therefore adopted in the subsequent hybridisation of the MAX cassettes.

4.9 Library Synthesis with Pre-ligated DNA cassettes

A series of hybridisations was carried out creating MAX oligonucleotide cassettes containing the full complement of 20 MAX codons at each position. Prior to hybridisation the β, γ and ENDMAX oligonucleotides were treated with PNK (2.8.2) to phosphorylate the 5' end of each oligonucleotide. Hybridisation reactions (2.9.5) were prepared in both hybridisation buffer 1 and hybridisation buffer 2, and hybridised (2.9.5). ATP (2.2.17), DTT (2.2.18) and one Weiss unit of ligase (2.8.3) were then added to the hybridisation reactions before incubation at 14°C overnight.

Ligation reactions (2.8.3) were prepared to subclone the pre-ligated cassettes into the pGEX-ZFMA3 vector which had been previously digested the enzyme *Sma*I (2.8.4) and treated with CIP (2.8.1) before digestion (2.8.4) with the enzymes *Hin*dIII and *Bsi*WI. The ligation reactions were incubated at 14°C for 14 – 16 hours and then used to transform (2.4.2) *E. coli* DH5α cells.

Plasmid DNA was isolated (2.4.6) from the recovered colonies and sequenced. Sequence alignments of the inserted sequences are shown in Figures 4.15a and 4.15b.

Sequence Alignments Hybridisation Buffer 1

1) Rearranged sequence
GGGAAAAGCTTTCATTTCAGCCAGCGAGTGACCCAACGCGACGTGGCCCAGTTCATGGC

2) phe-pro-ala
GGGAAAAGCTTTAGTTTTAGCGACCCGTTACAAGCGCATCAGCGTACGCACACCGGGG

3) cys-phe-leu
GGGAAAAGCTTTAgTTGCAGCGACTTTTTACAACTGCATCAGCGTACgCACACCGGGG

4) met-pro-xxx
GGGAAAAGNTTTAGTATGAGCGAcCCTTTACAATGNCATCANCGNANGCACACCGGGG

5) asn-ala-asn
GGGAAAAGCTTTAGTAATAgCGACGCGTTACAAAaACATNANCGTACGNNCACCGGG

6) Rearranged
GGGAAAAGCTTTAGTACCNGACGACCATAAACAAATGCAATCNGCGTACGCACACCGGG

7) ala-asp-ser
GGGAAAAGCTTTAGTGCTAGCGACGATTTACAAAGCCATCAGCGTACGCACACCGGGG

8) cys-phe-met
GGGAAAAGCTTTAGTTGCAGCGACTTTTTACAAATGCATCAGCGTACGCACACCGGGG

9) ala-gly-arg
GGGAAAAGCTTTAGTGCGAGCGAACGGCTTACAACGCCATCAGCGTACGCACACCGGGG

10) xxx-xxx-arg
GGGAAAANATTTAGTNAGAGCGACGTATTTACAACGCCATCAGCGTACGNACTTCCCGG

11) glu-gln-lys
GGGAAAAGCTTTAGTGAAAGCGACCAGTTACAAAAACATCAGCGTACGCACACCGGGG

12) arg-phe-phe
GGGAAAAGCTTTAGTCGCAGCGACTTTTTACAATTTCATCAGCGTACGCACACCGGGG

13) pro-his-xxx
GGGAAAAGCTTTAGTCCGAGCGACCACTTACAATA--TCAGCGTACGCACACCGGGG

14) pro-his-xxx
GGGNAAAAGCTTTAGTCCGAGCGACCATTTACAATA-ATCAGCGTACGCACACCGGGG

15) gly-lys-his
GGGAAAAGCTTTAGTGGCAGCGNACAAATTACAACATCATCAGCGTACGCACACCGGGG

16) Rearranged
GGGAAAAGCTTTAGTAACAGTCCATTAACACCCAACTCATCAGCGTACGCACACCGGGG

17) **met-thr-asn**
GGGAAAAAGCTTTAGTATGAGCGACACTTTACAAAACCATCANCGTACGCACANCGGGG

18) **asn-tyr-xxx**
GGGAaAAGcttTAGtAACAGCGCACTATTTACaACCGNANCATCAGCGGACGCANACCG

19) **lys-stop-trp**
GGGAAAAGcTTTAGTAAAAGCGACTAGTTAA-TGGCATCAGCGTACGCACACCGGGGC

20) **pro-pro-his**
GGGAAaAGcTTTAGTCCGAGCGACCCGTTACAAACACATCAGCGTACGCACACCGGGGA

21) **val-pro-lys**
GGGAAAAGCTTTAGTGTGAGCG-CCCATTACAAAAACATCAGCGTACGCACACCGGGG

22) **lys-val-gly**
GGGAAAAGCTTTAGTAAAAGCGACGTATTACAAGGGCATCAAGGCGTACGCACACCGGG

23) **asn-phe-arg**
GGGAaAAgcttTAGTAACAGCGACTTTTTACAACGCCATCAGCGTACGCACANCGGGG

24) **thr-his-xxx**
GGGAaAAGCTtTAGTACCAGCGACCATTTACNAAAGACATCGNNCGCTACGCNCTCCGG

25) **ser-arg-his**
GGGAAAAGCTTTAGTAGCAGCGACAGGTTACAACACCATCAGCGTACGCACACCGGGG

26) **Rearranged**
GGGAAAAgCTTTAGCATGGAGCGACGAATTACNATCGNATCACCGTACGCACNCCGGGG

27) **trp-lys-xxx**
GGGAAAAGCTTTAGTTGGAGCGACAAATTACAACCTCCATCAGCGTACGCACACCGGGG

28) **lys-xxx-gly**
GGGAAAAGcTTTAgTAAAAgCgACNNTATTACAAGGGCATNNAAGGNGCTANGCACAAN

29) **ile-arg-phe**
GGGAAAAGCTTTAGTATTAGCGACAGGTTACAATTTCCCATCAGCGTACGCACACCGGG

30) **xxx-tyr-ser**
GGGAaAAGcTTTAGTTC-AGCGACTATTTACAAAGCCATCAGCGTACGCACACCGGGG

31) **cys-asp-phe**
GGGAAAAgcTTTAGTTGCAGCGACGATTTACAATTTCATCAGCGTACGCACACCGGGG

32) **ser-lys-lys**
GGGAAAAGctttAGtAGCAGcGACAAATTACAAAAACATcAGcGTANGNACACCGGGG

33) **his-ala-ser**
GGGAAAAGCTTTAGTCATAGCGACGCGTTACAAAGCCATCAGCGTACGCACACCGGGG

34) phe-lys-lys
GGGAAAAGCTTTAGTTTTAGCGACAAATTACAAAAACATCAGCGTACGCACACCGGGG

35) met-ser-xxx
GGGAAAAGCTTTAGTATGAGCGACAGCTTACAAAGACCATCAGCGTACGCACACCGGGG

36) gln-xxx-arg
GGGAAaAgCTTTAGTCAGAGCGACTTCTTTAGCCAACGCCATCAGCGTACGCACACCGG

37) tyr-lys-trp
GGGAAAAGCTTTAGTTATAGCGACAAATTACAATGGCATCAGCGTACGCACACCGGGG

38) phe-cys-cys
GGGAAaAGCTTTAGTTTTAGCGACTGCTTACAATGCCATCAGCGTACGCACACCGGGG

39) ile-xxx-gln
GGGNANAAAATTTAGtATTAGCGACGG-TTACAACAGCATCAGCGTACGCACACCGGGG

40) lys-ile-met
GGGAAAAGCTTTAGTAAAAGCGACATTTTAACAAAAGCATCAGCGTACGCACACCGGGG

41) val-xxx-asn
GGGAAAAGCTTTAGTGTTAGCGACCAGTTTACAAAACCATCAGCGTACGCACACCGGGG

42) asn-arg-gln
GGGAAAAgCTTTAGTAACAGCGACCGCTTACAACAGCATCAGCGTACGCACACCGGGG

43) ile-lys-leu
GGGAAAAGCTTTAGTATCAGCGACAAATTACAACTTCATCAGCGTACGCACACCGGGG

44) xxx-ala-ala
GGGAAAAGCTTTAGTGT-AGCGACCCGTTACAAGCGCATCAGCGTACGCACACCGGGG

45) val-arg-xxx
GGGAAAAGCTTTAGTGTGAGCGACAGATTACAATTGCCATCAGCGTACGCACACCGGGG

46) glu-lys-trp
GGGAAAAGCTTTAGTGAAAGCGACAAATTACAATGGCATCAGCGTACGCACACCGGGG

47) tyr-lys-ser
GGGAAAAGCTTTAGTTATAGCGACAAATTACAAAGCCATCAGCGTACGCACACCGGGG

48) lys-asn-leu
GGGAAAAGCTTTAGTAAAAGCGACAACTTACAACTGCATCAGCGTACGCACACCGGGG

49) asp-glu-cys
GGGAAAAGCTTTAGTGATAGCGACGAATTACAATGGCATCAGCGTACGCACACCGGGG

50) cys-ser-ala Rearranged
GGGAAAAGCTTTAGTTGCAGCGACTCATTACAAGCGCATCCAGCCGT---CAGAAAAGA

51) stop-xxx-asp
GGGAAAAGCTTTAGTTAGAGCGACAA-TTACAAGATCATCAGCGTACGCACACCGGGG

52) ala-cys-thr
GGGAAAAGCTTTAGTGCGAGCGACTGCTTACCAAACCCNCTCCGCNCACACACCCCGG

53) met-xxx-xxx
GGGAAAAGCTTTAGTATGAGCGACGA---TTACAAAGCCCATCAGCGTACGCACACCG

54) trp-asn-xxx
GGGAAAAGCTTTAGTTGGAGCGACAATTTACAAAAACACATCAGCGTACGCACACCGGG

55) leu-thr-ala
GGGAAAAGCTTTAGTCTGAGCGACACGTTACCAAGCTCCATCAGCGTACGCACACCGGG

56) xxx-trp-ala
GGGAAAAGCTTTAGTNATAGCGACTGGTTACAAGCGCATCAGCGTACGCACACCGGGG

57) met-phe-phe
GGGAAAAGCTTTAGTATGAGCGACTTTTTACAATTTCATCAGCGTACGCACACCGGGG

58) ser-val-val
GGGAAAAGCTTTAGTAGCAGCGACGTATTACAAGTTCATCAGCGTACGCACACCGGGG

59) lys---
CGGGAAAAGCTTTAGTAAA----------CATCAGCGTACGCACACCGGGG

60) gly-trp-leu
GGGAAAAGCTTTAGTGGCAGCGACTGGTTACAACTGCATCAGCGTACGCACACCGGGG

61) gly-xxx-leu
GGGAAAAGCTTTAGTGGGAGCGACCGCCTTACAATTACATCAGCGTACGCACACCGGGG

62) asn-asp-val
GGGAAAAGCTTTAGTAACAGCGACGATTTACAAGTGCATCAGCGTACGCACACCGGGG

63) ser-asn-acc
GGGAAAAGCTTTAGTTCCAGCGACAACTTACAAACCGCATCAGCGTACGCACACCGGGG

64) asp-ile-his
GGGAAAAGCTTTAGTGATAGCGACATTTTACAACATCATCAGCGTACGCACACCGGGG

65) ser-xxx-tyr
GGGAAAAGCTTTAGTAGCAGCGACAG-TTACAATATCATCAGCGTACGCACACCGGGG

66) his-met-gly
GGGAAAAGCTTTAGTCATAGCGACATGTTAACAAGGCCATCAGCGTACGCACACCGGGG

67) pro-arg-met
GGGAAAAGCTTTAGTCCGAGCGACCGCTTACAAATGCATCAGCGTACGCACACCGGGG

68) **pro-asp-xxx**

GGGAAAAGCTTTAGTCCGAGCGAACCGATTTACCAAAAATCCATCAGCGTACGCACAC

69) **ala-ile-ala**

GGGAAAAGCTTTAGTGCAAGCGACATTTTACAAGCGCATCAGCGTACGCACACCGGGG

70) **xxx-ala-glu**

GGGAAAAGCTTTAGTAT-AGCGACGCGTTACAAGAACATCAGCGTACGCACACCGGGG

71) **met-gln-val**

GGGAAAAGCTTTAGTATGAGCGACCAGTTACAAGTGCATCAGCGTACGCACACCGGGG

72) **Self ligation**

GGGAAAAGCTTCGTTCCCGGGATGACGTACGCACACCGGGGAAAA

73) **gly-met-xxx**

GGGAAAAGCTTTAGTGGCAGCGACATGTT-CCCAACGCGCATCAGCGTACGCACACCGG

Fig 4.15a Sequence data obtained from clones recovered after the cassettes mutagenesis of pGEX-ZFMA3 with pre-ligated MAX cassettes generated in hybridisation buffer 1. The amino acids encoded at the randomised positions of each clone are also shown. Key to figure: Purple text = MAX codon sequence: Red text = Non MAX codon sequence. All bases in blue text represent sequence abnormalities. Inserted bases are denoted by the inserted base highlighted in blue text. Deletions are denoted by a blue dash (-). Point mutations are highlighted in bold text. Rearranged sequences are noted in the figure. Randomised positions in which the identity of the encoded amino acid could not be accurately deduced from the sequence data due to the insertion/deletion of bases or the inclusion of an N represented base are represented by a blue xxx. Nucleotides shown in lower case letters represent bases identified from N classified bases after analysis of the chromatogram.

Sequence Alignments Hybridisation Buffer 2

1) ile-trp-trp
CGGGAAAAGCTTTAGTATTAGCGACTGGTTACAATGGCATCAGCGTACGCACACCGGGG

2) asn-ile-arg
CGGGAAAAGCTTTAGTAACAGCGACATTTTACAAAGGCATCAGCGTACGCACNCCGGGG

3) thr-arg-xxx
CGGGAAAAGCTTTAGTACCAGCGACCGTTTAAAACGCCCATCAGCGTACGCACACCGGG

4) val-gln-leu
GGGAAAAGCTTTAGTGTGAGCGACCAGTTA--ATTACATCAGCGTACGCACACCGGGG

5) leu-his-ala
GGGAAAAGCTTTAGTCTGAGCGACCATTTACAAGCACATCAGCGTACGCACACCGGGG

6) ala-xxx-pro
GGGAAAAGCTTTAGTGCCAGCGACCTGTTTACAACCCCATCAGCGTACGCACACCGGGG

7) Rearranged
CGGGAAAAGCTTTGCTCACCGCATAATCCGTCGCAATAATCNCAATATGGCGCAACCTG

8) Rearranged
CGGGAAAAGCTTCTNGNACAANATCGGGTAACATNNCNNGNACGGNGACATAGCGGGTA

9) xxx-xxx-gly
GGGAAAAGCTTTAGTNTGAGCGACCTCCTTACAAGGCCATCCNNGTGCNCNCNCCGGGG

10) phe-ile-gln
GGGAAAAGCTTTAGTTTTAGCGACATCTTACAACAGCATCAGCGTACGCACACCGGGG

11) gln-ile-leu
GGGAAAAGCTTTAGTCAGAGCGACATATTACAACTACATCAGCGTACGCACACCGGGG

12) pro-lys-trp
GGGAAAAGCTTTAGTCCGAGCGACAAATTACAATGGCATCAGCGTACGCACACCGGGG

13) gln-met-phe
GGGAAAAGCTTTAgTCAGAGCGACATGTTACAATTTCATCAGCGTACGCACACCGGGG

14) Rearranged
GGGAAAAGCTTGTGGCAGGAGCTGGCAGACATCACCGATAAAACGCAGCTTGAATGGC

15) self ligation
GGGAAAAGCTTCGTT**CCCGGG**ATGACGTACGCACACCGGGGAAAA

16) asn-STOP-met
GGGAAAAGCTTTAGTAACAGCAACTAGTTACAAATGCATCAGCGTACGCACACCGGGG

17) **rearranged**
GGGGAAAAGCCATATAAATGCCCTTCNATGTGGCAAGTCTTTCAGCCGTAGTGATCAT

18) **pro-lys-phe**
GGGAAAAGCTTTAGTCCAAGCGACAAATTACAATTTCATCAGCGTACGCACACCGGGG

19) **ser-asp-xxx**
GGGAAAAGCTTTAGTAGCAGCGACGACTTACAAATTTCATCAGCGTACGCACACCGGGG

20) **lys-xxx-leu**
GGGAAAAGCTTTAGTAAAAGCGACCGCCTTACAATTGCATCAGCGTACGCACACCGGGG

21) **cys-xxx-gln**
GGGAAAAGCTTTAGTTGCAGCGACGG-TTACAACAGCATCAGCGTACGCACACCGGGG

22) **leu-ala-ala**
GGGAAAAGCTTTAGTCTGAGCGACGCGTTACAAGCGCATCAGCGTACGCACACCGGGG

23) **Self ligation**
GGGAAAAGCTTCGTT**CCCGGG**ATGACGTACGCACACCGGGGAAAA

24) **ala-lys-xxx**
GGGAAAAGCTTTAGTGCCAGCGACAAATTACAACGCCCATCAGCGTACGCACACCGGGG

25) **glu-xxx-ile**
GGGAAAAGCTTTAGTGAAAGCGAGAA-TTACAAATTCATCAGCGTACGCACACCGGGG

26) **cys-tyr-thr**
GGGAAAAGCTTTAGTTGCAGCGACTATTTACAAACCCATCAGCGTACGCACACCGGGG

27) **his-glu-ala**
GGGAAAAGCTTTAGTCACAGCGACGAATTACAAGCGCATCAGCGTACGCACACCGGGG

28) **ser-asn-phe**
GGGAAAAGCTTTAGTAGTAGCGACAATTTACAATTTCATCAGCGTACGCACACCGGGG

29) **ala-arg-met**
GGGAAAAGCTTTAGTGCGAGCGACCGCTTACAAATGCATCAGCGTACGCACACCGGGG

30) **ile-met-xxx**
GGGAAAAGCTTTAGTATTAGCGACATGTTACAAGACCCATCAGCGTACGCACACCGGGG

31) **met-xxx-gln**
GGGAAAAGCTTTAGTATGAGCGACT--TTACAACAGCATCAGCGTACGCACACCGGGG

32) **lys-leu-tyr**
GGGAAAAGCTTTAGTAAAAGCGACTTATTACAATATCATCAGCGTACGCACACCGGGG

33) **gly-gln-ile**
GGGAAAAGCTTTAGTGGGAGCGACCAGTTACAAATTCATCAGCGTACGCACACCGGGG

34) **val-lys-arg**
GGGAAAAGCTTTAGTGTGAGCGACAAATTACAACGCCATCAGCGTACGCACACCGGGG

35) **ile-met-lys**
GGGAAAAGCTTTAGTATTAGCGACATGTTACAAAAACATCAGCGTACGCACACCGGGG

36) **gln-asp-ala**
GGGAAAAGCTTTAGTCAGAGCGACGATTTACAAGCGCATCAGCGTACGCACACCGGGG

37) **xxx-thr-cys**
GGGAAAAGCTTTAGT-CGAGCGACACCTTA-AATGCCATCAGCGTACGCACACCGGGG

38) **arg-xxx-pro**
GGGAAAAGCTTTAGTCGCAGCGACGTATTTACAACCGCATCAGCGTACGCACACCGGGG

39) **asp-xxx-xxx**
GGGAAAAGCTTTAGTGATAGCGACTANTTACAAANCCATCAGCGTACGCACACCGGGG

40) **gly-xxx-xxx**
GGGAAAAGCTTTAGTGGGAGCGACGCCaTTACAATCGCCATCAGCGTACGCACACCGGG

41) **thr-xxx-pro**
GGGAAAAGCTTTAGTACCAGCGACGCACTTACAACCGCATCAGCGTACGCACACCGGGG

42) **his-thr-xxx**
GGGAAAAGCTTTAGTCATAGCGACACCTTACA-TATCATCAGCGTACGCACACCGGGG

43) **tyr-lys-ser**
GGGAAAAGCTTTAGTTATAGCGACAAATTACAAAGCCATCAGCGTACGCACACCGGGG

44) **thr-xxx-lys**
GGGAAAAGCTTTAGTACCAGCGACACCCTTACAAAAGCATCAGCGTACGCACACCGGGG

45) **ser-xxx-asn**
GGGAaAAgCTTTAGTAGCAGCCGACCAGATTACAAAACCATCAGCGTACGCACACCGGG

46) **val-xxx-Rearranged**
GGGAAAAGcTTTAGTGTTAGCGACNGGTTACCAA*AAC*CTNNNCGANCNNCCNCCGGGN

47) **xxx-xxx-xxx**
GGGAAaAGcATTTAGTTGGCAGCGACCANTTACAAGC-CATCAGCGTACGCACACCGG

48) **asp-gly-leu**
GGGAAaAGCTTTAGTGATAGCGACGGCTTACAACTGCATCAGCGTACGCACACCGGGG

49) **ile-lys-xxx**
GGGAAAAGCTTTAGTATTAGCGACAAATTACAAAAACCATCAGCGTACGCACACCGGGG

50) **tyr-val-xxx**
GGGAAAAgcTTTAGTTATAGCGACGTGTTACACGGGCATCAGCGTACGCACACCGGGG

51) gln-trp-val

GGGAAaAgcTTTAGTCAGAGCGACTGGTTACAAGTGCATCAGCGTACGCACACCGGGG

52) asp-gly-val

GGGAAAAGCTTTAGTGATAGCGACGGATTACCAAGTACATCAGCGTACGCACACCGGGG

Fig 4.15b Sequence data obtained from clones recovered after the cassettes mutagenesis of pGEX-ZFMA3 with pre-ligated MAX cassettes generated in hybridisation buffer 2. The amino acids encoded at the randomised positions of each clone are also shown. Key to figure: Purple text = MAX codon sequence: Red text = Non MAX codon sequence. All bases in blue text represent sequence abnormalities. Inserted bases are denoted by the inserted base highlighted in blue text. Deletions are denoted by a blue dash (-). Point mutations are highlighted in bold text. Rearranged sequences are noted in the figure. Randomised positions in which the identity of the encoded amino acid could not be accurately deduced from the sequence data are represented by a blue xxx. Nucleotides shown in lower case letters represent bases identified from N classified bases after analysis of the chromatogram.

An initial survey of the sequence results is contained in Table 4.1 showing the numbers of clones containing all MAX codons at positions of randomisation, non MAX codons at these positions and the number of clones containing frameshift mutations within the sequence.

Identity of Sequenced clones	**Hybridisation Buffer 1**		**Hybridisation Buffer 2**	
	No. of sequences	% of total sequences	No. of sequences	% of total sequences
Correct Sequence Containing MAX codons at all randomised Positions	24	33.3	12	24
Correct Sequence containing non MAX codons at randomised positions	8	11.1	9	18
Sequences Containing Frameshift Mutations	38	52.8	26	52
Sequences Containing Point Mutations	0	0	2	4
Sequences Containing Mixed Codons	2	2.8	1	1.9
Self Ligations	1	See Legend	2	See Legend
Total No. of sequences	73	100	52	100

Table 4.1 Summary of results obtained when sequencing clones recovered after the MAX randomisation of the pGEX-ZFMA3 plasmid, using MAX cassettes generated in different hybridisation buffers. Clones resulting from the religation of the parental plasmid were omitted from the percentage of total sequences calculation, the religation of parental plasmids is common in cloning reactions and it was expected that the properties of the inserted cassette would not influence the occurrence of self ligated clones.

The initial survey of the sequence results highlighted that a large number (approximately 50 %) of the recovered sequences contained frameshift mutations, such as inserted and deleted bases or sequence rearrangements. The incidence of frameshift mutations was almost identical in the sequences obtained from the two differing hybridisation buffers. Although frameshift mutations are observed when cloning synthetic DNA, the incidence of these mutations in the generated libraries was higher than that obtained when cloning standard oligonucleotide cassettes, which suggested that these sequences were being selected for in the cloning procedure. The reasons for this selection are difficult to ascertain. Initially it was assumed that these sequences may have resulted indirectly from mismatched base pairing between the selection oligonucleotides and the template strand. If this assumption was correct the similar

incidence of these mutations observed when the cassettes were hybridised in the different buffers suggested that the stringency of hybridisation was similar in both buffers. This assumption is difficult to reconcile with the comparisons of the numbers of MAX codons present at each position of randomisation when the cassettes were hybridised in the different buffers. Sequence results showed that cassettes hybridised in buffer 1 contained the greatest number of intact sequences in which MAX codons were present at all positions of randomisation. In addition the number of intact sequences containing non-MAX codons at randomised positions was lower when the cassettes were hybridised in buffer 1, which suggested that the stringency of hybridisation of the selection oligonucleotides may be increased when using hybridisation buffer 1.

4.10 Analysis of Clones

The identities of the randomised positions of the recovered clones were analysed in the context of the generation of libraries. This raised the question of whether clones containing frameshifts should be included in codon analyses. These clones would not encode functional zinc finger proteins and so it could be argued that they should be excluded from any subsequent analysis. Conversely, as the current study reflects the recovery of DNA sequences and the optimisation of the MAX technique, it may be argued that codons in frameshifted clones should be included in the analysis. Both approaches were examined.

4. 10.1 Analysis of Clones *Excluding* Those Containing Frameshift Mutations

The inclusion of MAX codons at each position of randomisation was examined to assess the affect of the redesign of the selection oligonucleotides. The graphs in Figures 4.16a and 4.16b illustrate the numbers of correct MAX codons at each position of randomisation in the intact sequences, generated by the hybridisation of cassettes in hybridisation buffer 1 and buffer 2 respectively. The graphs demonstrate that the incorporation of MAX codons (approximately 88% and 83% in buffers 1 & 2 respectively) is similar at all positions of randomisation in sequences that contain no frameshift mutations. Hence the disparity between the positions of randomisation seen

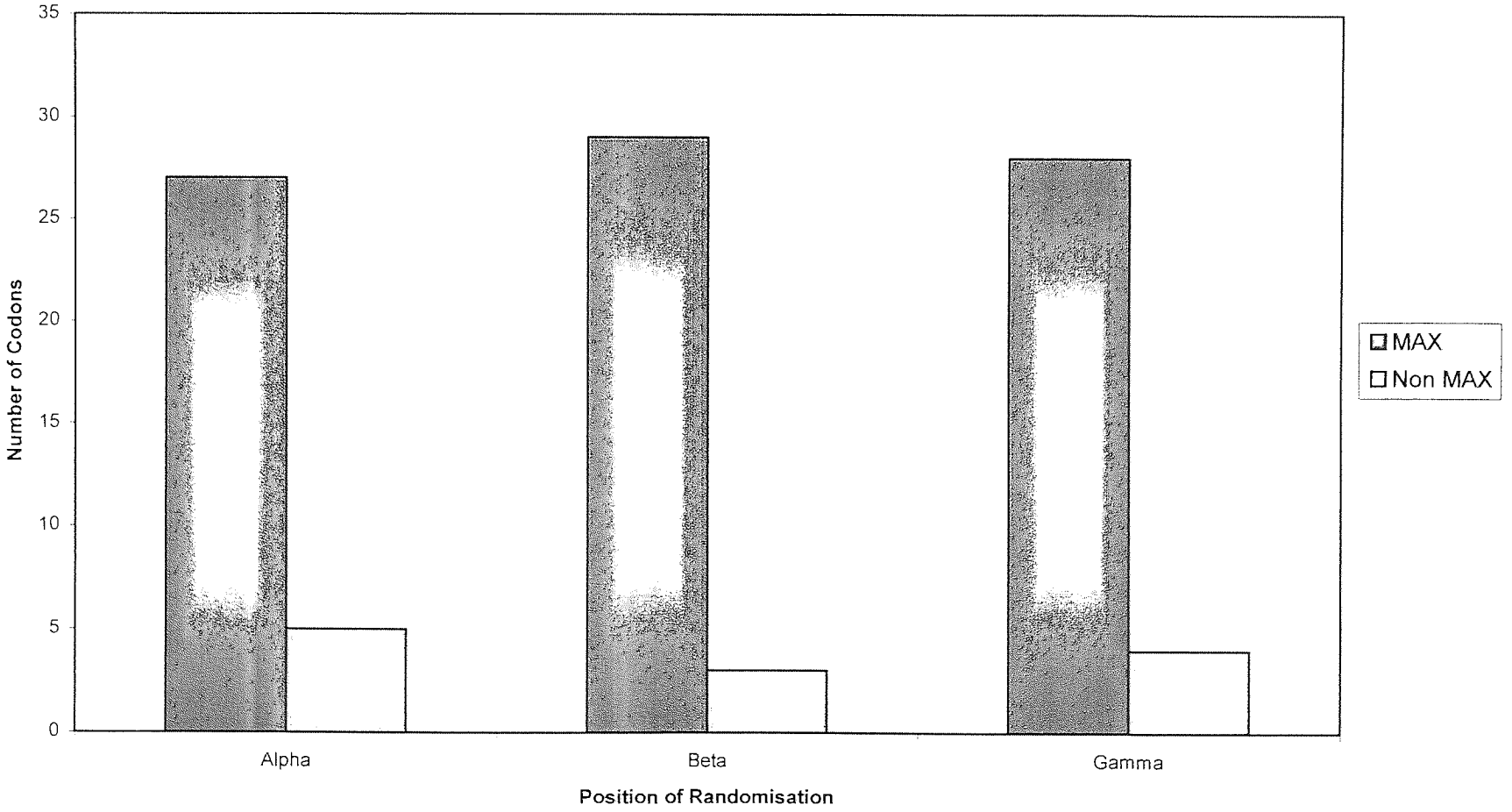


Fig 4.16a Graph demonstrating the identities of the codons present at the randomised positions, in the intact DNA sequences recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX cassettes generated in hybridisation buffer 1.

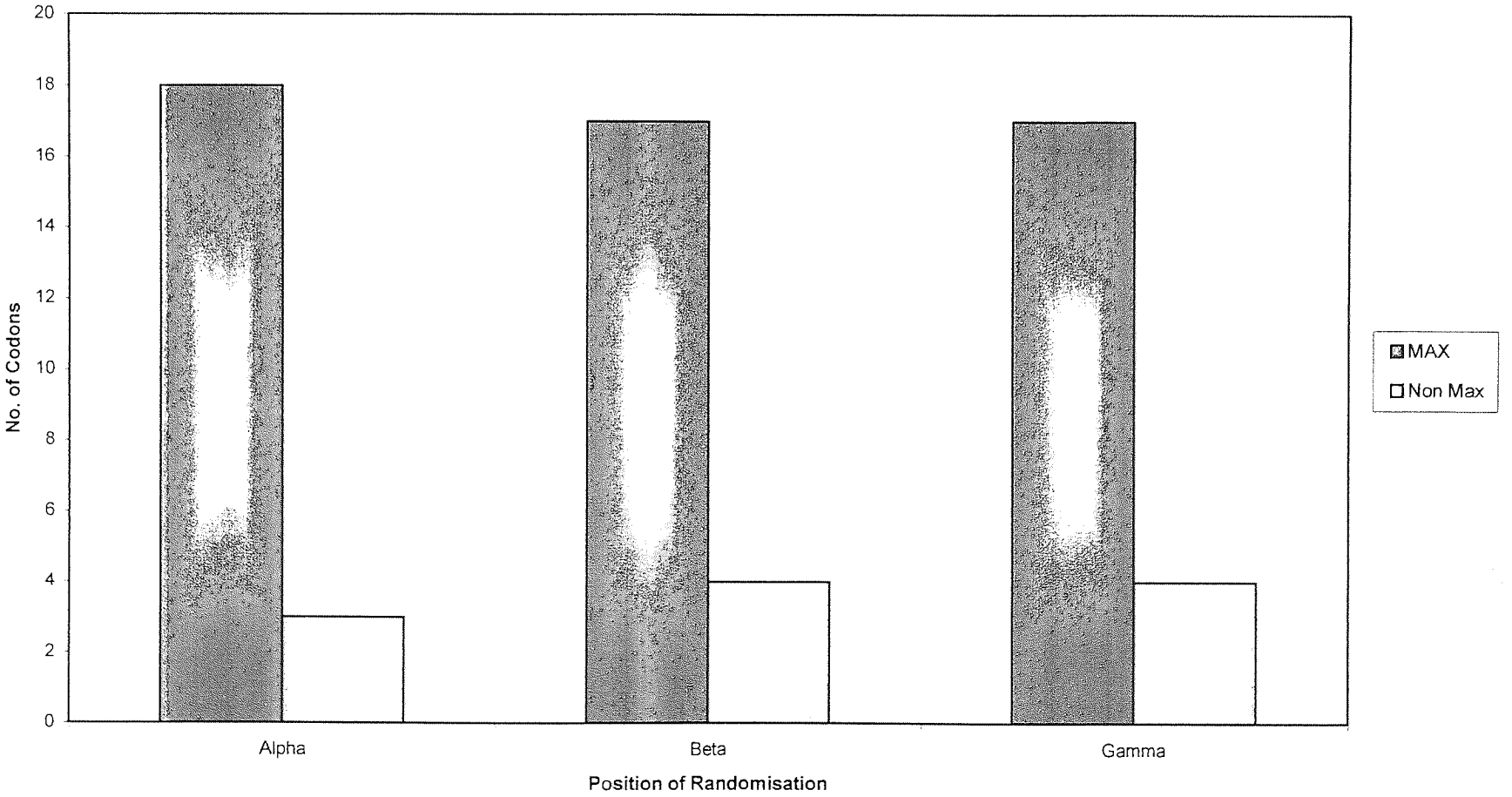


Fig 4.16b Graph demonstrating the identities of the codons present at the randomised positions, in the intact DNA sequences recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX cassettes generated in hybridisation buffer 2

previously (section 4.3), where the incorporation of MAX codons was seen predominantly at position gamma, is no longer evident. It was concluded that the relocation of the MAX codon to the end of the selection oligonucleotides had successfully reduced the levels of non-MAX incorporation at all three locations. Furthermore, the slightly lower recovery of cassettes containing non-MAX codons in buffer 1, suggested that this buffer may be preferable for use in future experiments.

Ideally, MAX randomisation should result in equal representation of each encoded amino acid. Overall representation was therefore assessed (Fig. 4.17a and 4.17b). These graphs show that the distribution of the randomised codons is reasonable in both buffers when the small sample size is taken into account.

Comparison of the representation in both buffers showed that the MAX codon encoding lysine (AAA) was predominant in both hybridisation buffers, which suggested that the lysine selection oligonucleotide may have been favourably selected within the hybridisation reaction. The representation of the MAX phenylalanine codon, which had predominated in buffer 1, fell roughly within the expected values when cassettes were hybridised in buffer 2. In comparison the MAX glutamine codon, which had predominated in buffer 2, fell within the expected values when cassettes were hybridised in buffer 1, suggesting that the predominance of these codons may have been the result of experimental variation, rather than an active selection in the hybridisation procedure.

The representation of MAX codons at each individual randomised position was then examined to assess whether the predominance of certain codons was the result of the predominance of that codon at a single randomised position (Fig. 4.18a and 4.18b). Comparison of the representation at each position in both buffers, highlighted that the overall predominance of the MAX lysine codon resulted from the predominance of this codon at the β position of randomisation in both libraries. Similarly the predominance of the glutamine codon in the buffer 2 hybridised libraries appeared to result from the predominance of this codon at the alpha position. There was little difference seen between results generated in buffer 1 and buffer 2.

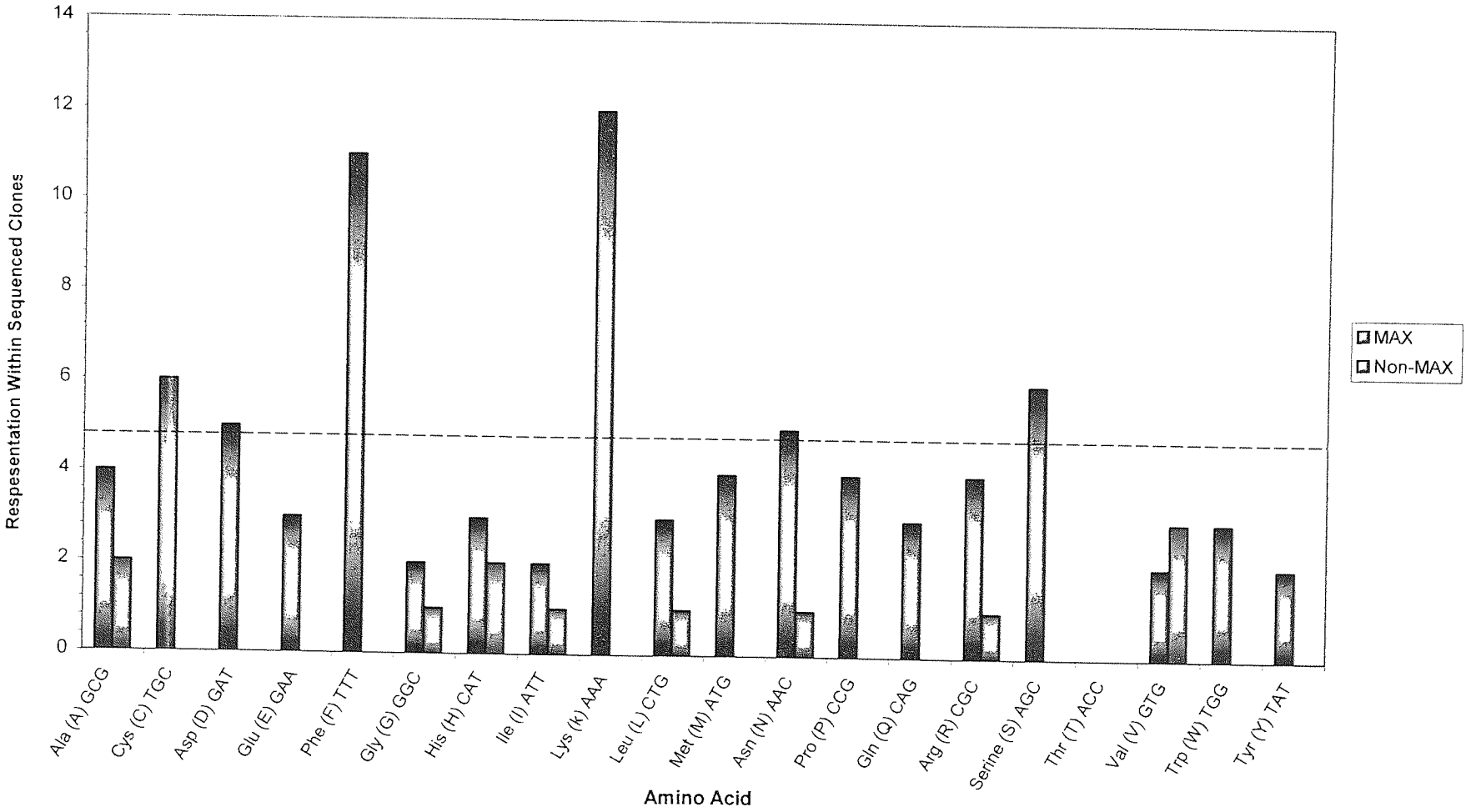


Fig 4.17a Graph demonstrating amino acid representation at all randomised positions within the sequenced clones, recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 1. Data excludes clones with frameshift mutations. The blue line represents the theoretical ideal distribution of MAX codons in the correct sequences, based upon 100 % efficiency of the technique and the recovery of no non-MAX codons.

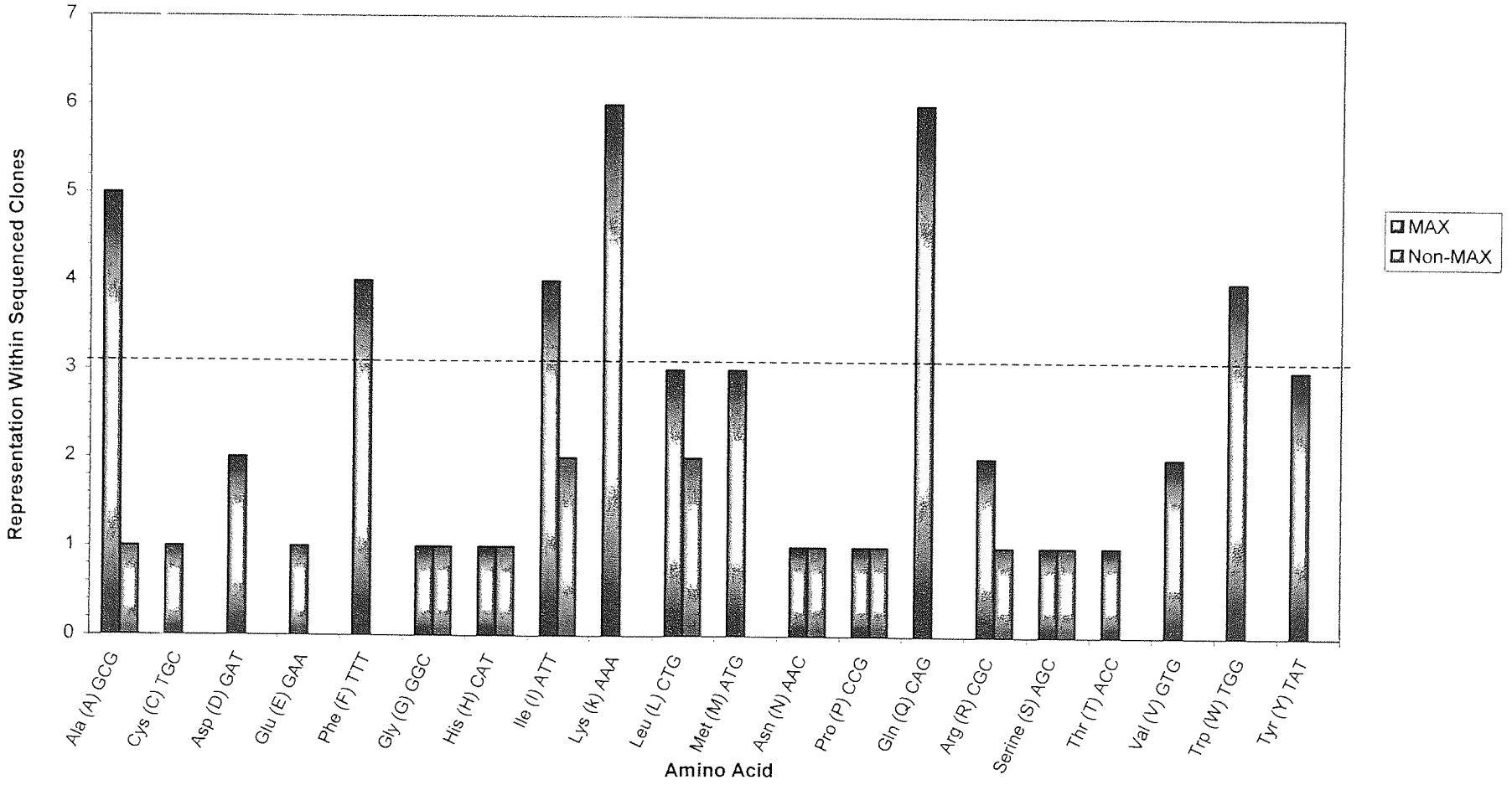


Fig 4.17b Graph demonstrating amino acid representation at all randomised positions, within the sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 2. The blue line represents the ideal distribution of MAX codons in the correct sequences, based upon 100% efficiency and the generation of no non-MAX codons. Data excludes frameshift mutations.

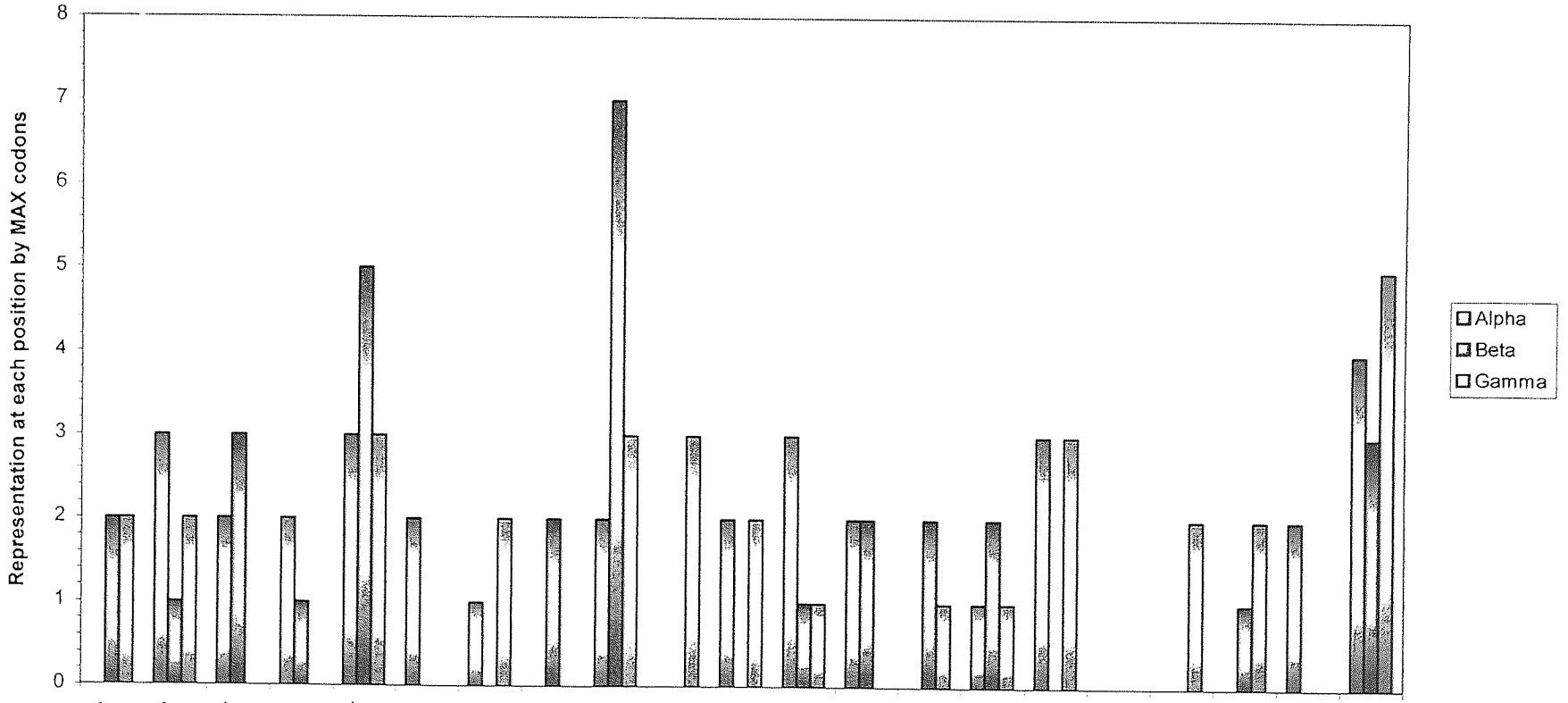


Fig 4.18a Graph demonstrating the amino acid representation by MAX codons at each randomised position, within sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 1. Data excludes sequences containing frameshift mutations. The total numbers of non-MAX codons recovered at each position are included for comparison.

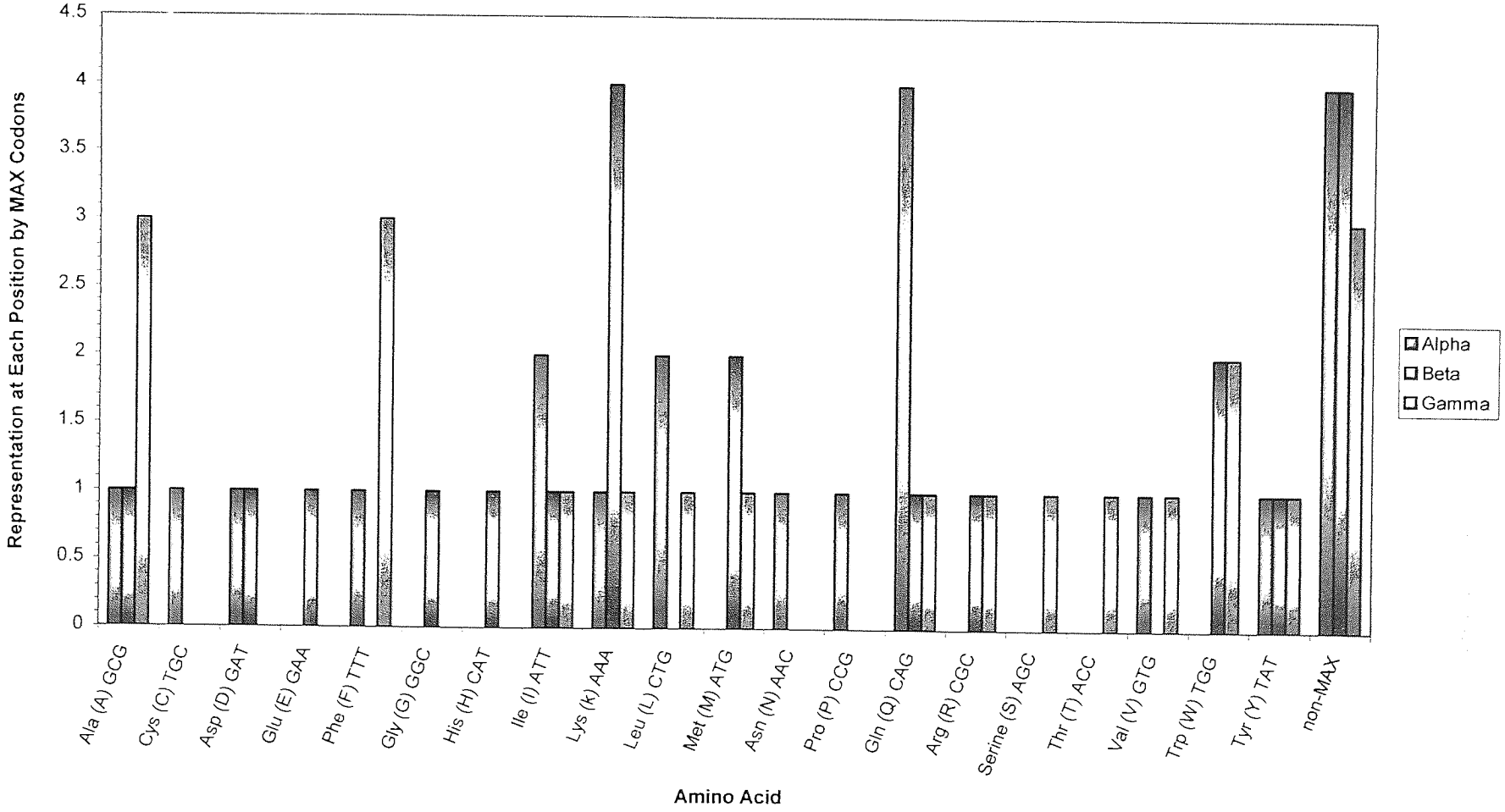


Fig 4.18b Graph demonstrating the amino acid representation by MAX codons at each randomised position, within the sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 2. Data excludes sequences containing frameshift mutations. The total numbers of non-MAX codons recovered at each position are included for comparison.

4.10.2 Analysis of Clones *Including* Frameshift Mutations

As in the previous analysis, the inclusion of MAX codons at each position of randomisation was examined and the data was then used to examine the overall incorporation of MAX codons, overall encoded amino acid representation and finally, amino acid representation at each of the three randomised positions (Figs. 4.19, 4.20 and 4.21 respectively).

Figures 4.19a and 4.19b demonstrate that when including frameshifted clones, the alpha position of randomisation contained the greatest number of correct MAX codons, whereas incorporation of MAX codons at positions beta and gamma was comparable and consistently lower than at position alpha. This was expected to reflect the incorporation of unidentifiable codons (those in which the frameshift mutation occurred actually *within* the codon itself) at the beta and gamma positions, rather than any discrepancy between the hybridisation of the selection oligonucleotides. The numbers of unidentified codons at each position within the libraries were calculated as 7α, 13β and 16γ (buffer 1) and 5α, 15β and 13γ (buffer 2). The lower incorporation of MAX codons at the beta and gamma positions appears therefore to result from the incidence of frameshift mutations within the inserted sequence, as the majority of frameshifts occur after the alpha codon. The low incidence of frameshift mutation within the alpha position of randomisation is expected to reflect the requisite nature of the alpha selection oligonucleotide, which generates the *Hin*dIII cohesive terminus essential for cloning.

Interestingly, inclusion of the frameshift data in the overall representation of encoded amino acids (Figure 4.20a and 4.20 b) makes very little difference in the pattern of amino acid representation. Again, in either buffer, there is a reasonable distribution of encoded amino acids.

When individual positions of randomisation are considered (Figure 4.21a and 4.21b), encoded amino acids representation is still reasonable in both buffers, although there is more codon omission in buffer 2.

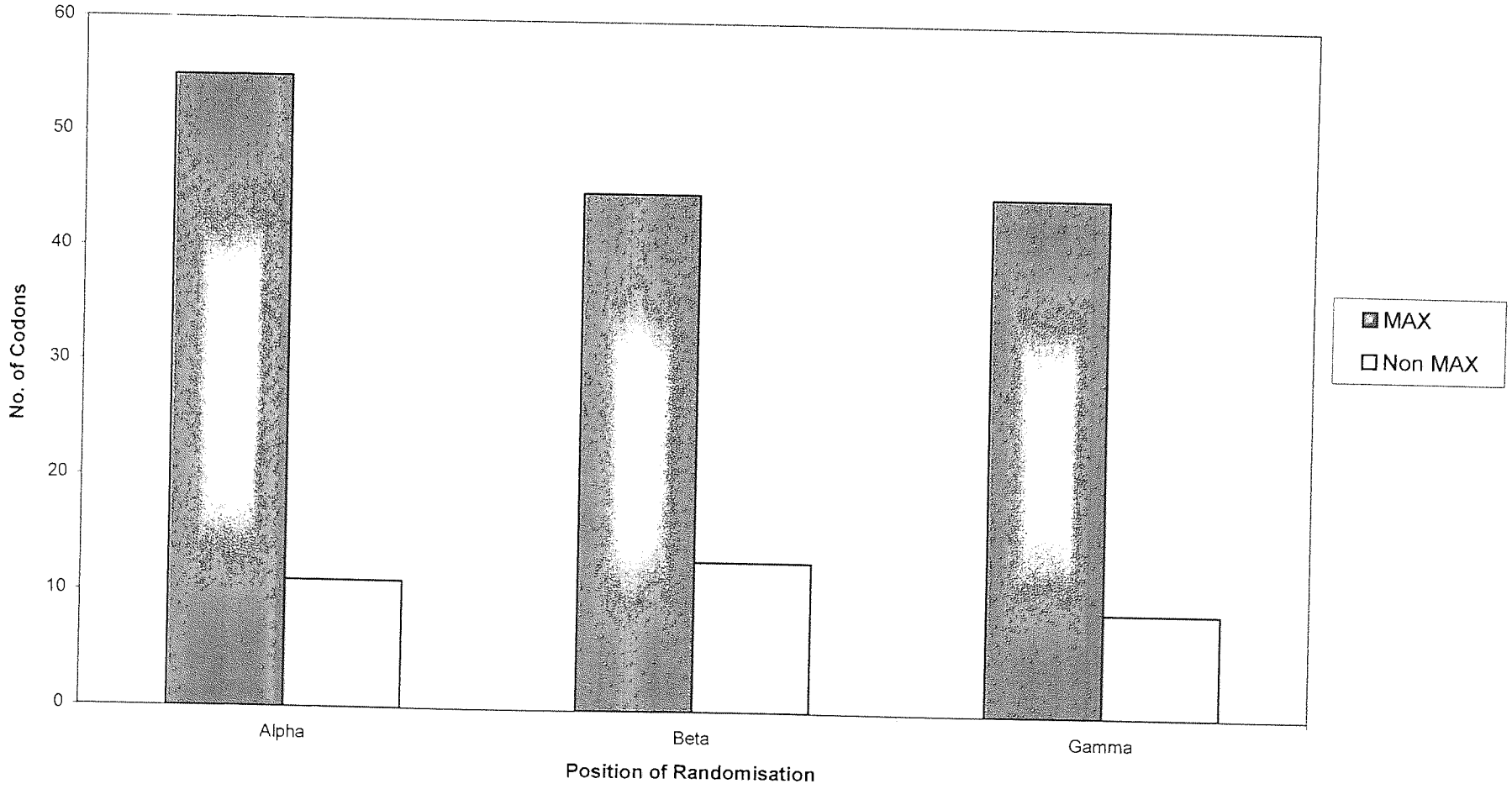


Fig 4.19a Graph demonstrating the identities of all identified codons at the positions of randomisation, in the sequences recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX cassettes generated in hybridisation buffer 1. Data includes frameshift mutations.

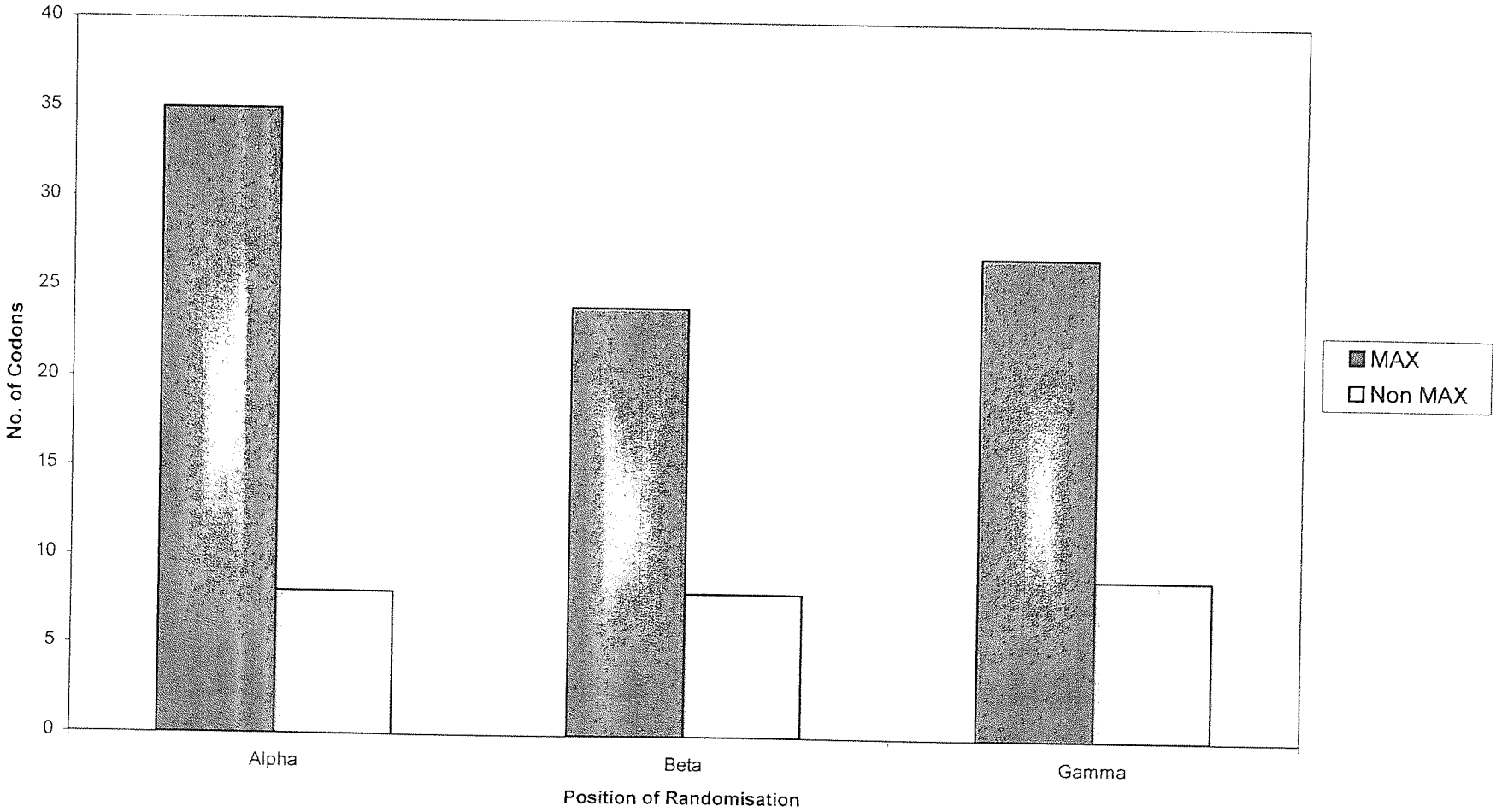


Fig 4.19b Graph demonstrating the identities of all identified codons at the positions of randomisation, in the sequences recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX cassettes generated in hybridisation buffer 2. Data includes frameshift mutations.

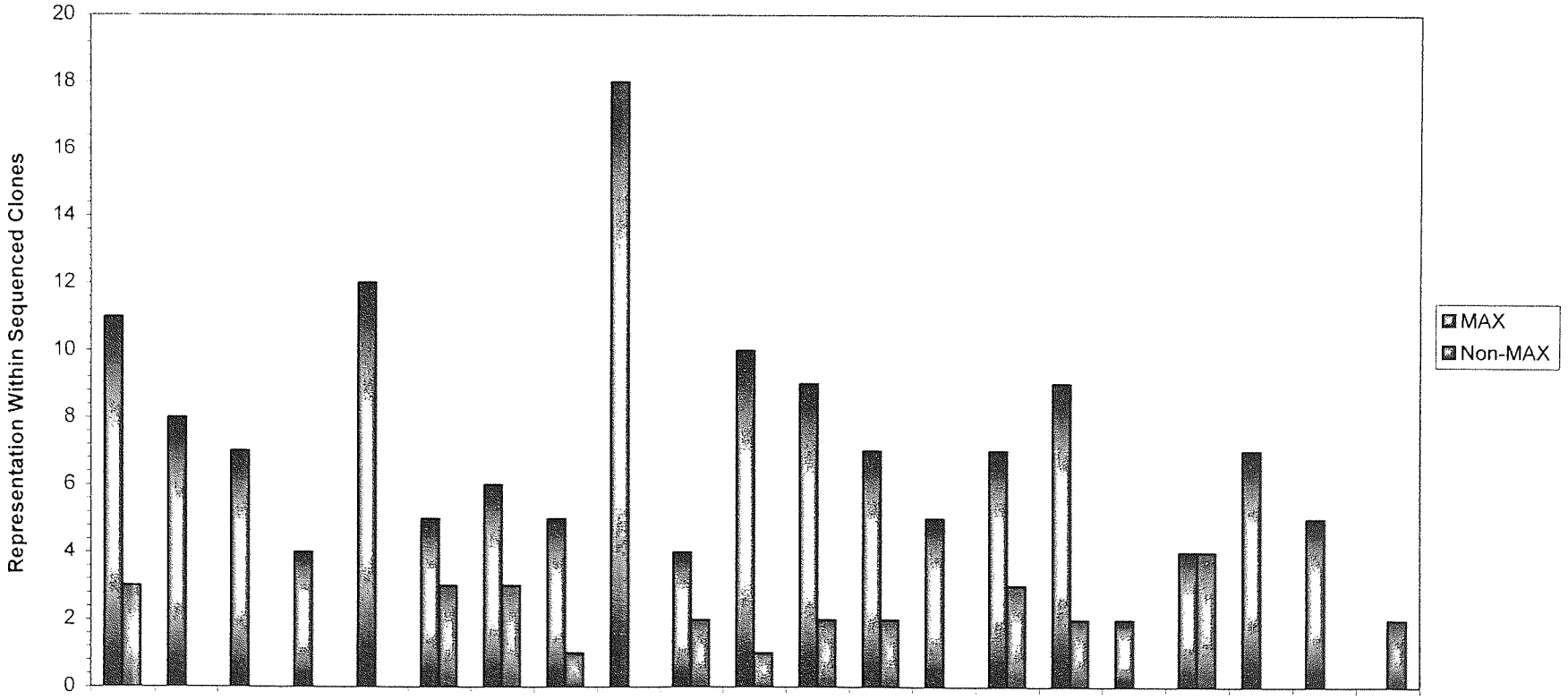


Fig 4.20a Graph demonstrating the amino acid representation at all randomised positions, within the sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 1. Data includes frameshift mutations.

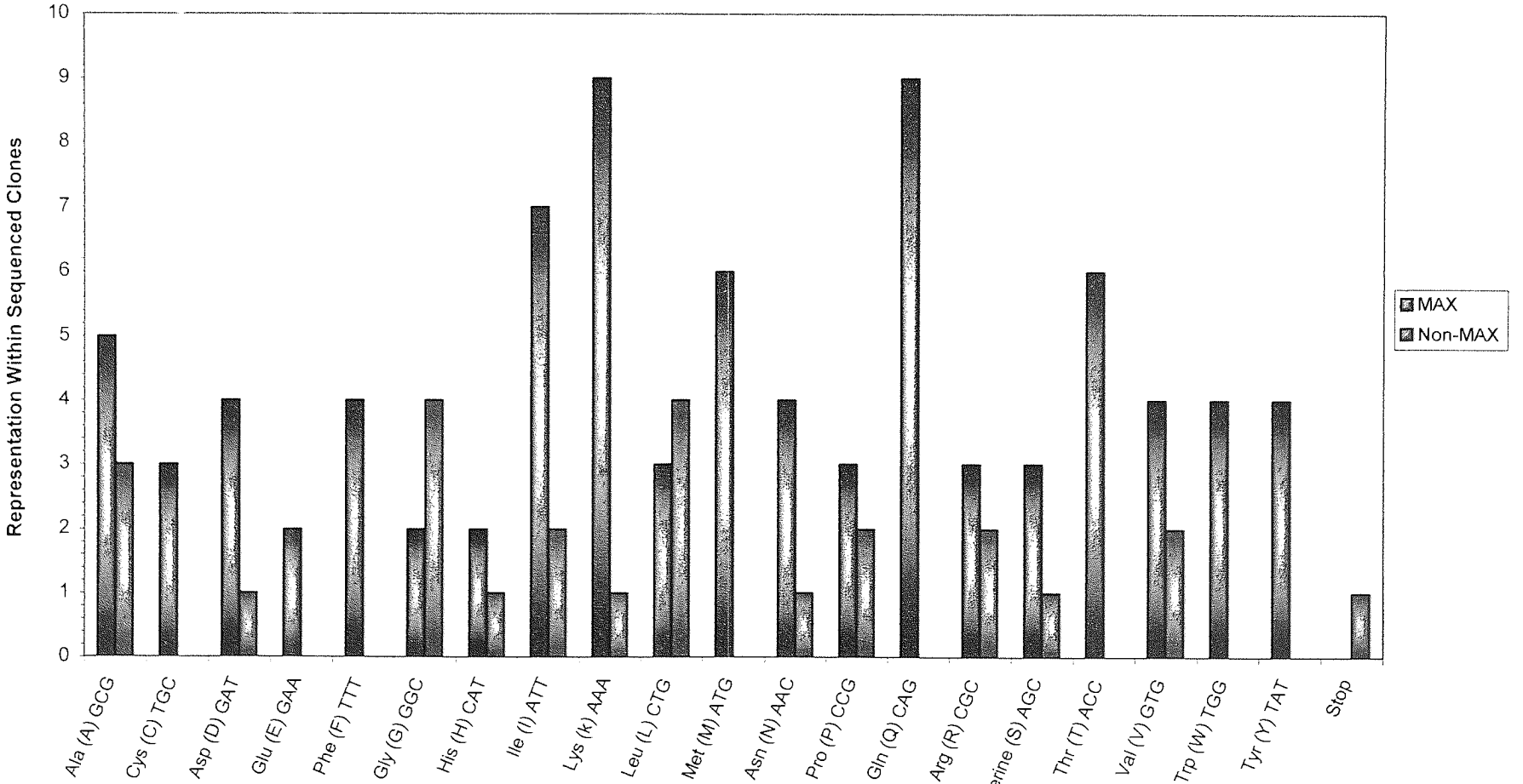


Fig 4.20b Graph demonstrating amino acid representation at all randomised positions within the sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 2. Data includes frameshift mutations

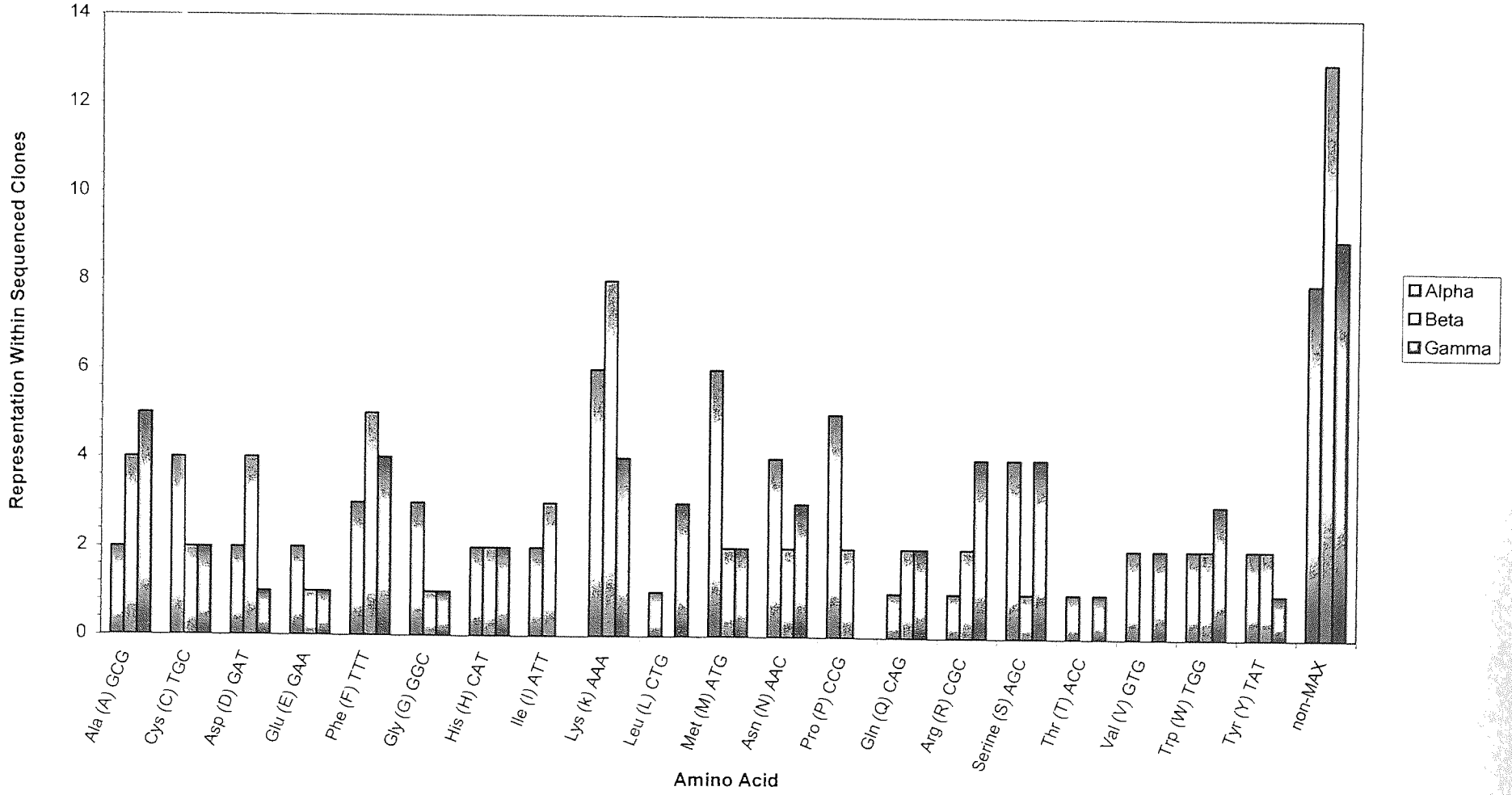


Fig 4.21a Graph demonstrating the amino acid representation by MAX codons at each randomised position, within the sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 1. Data includes frameshift mutations.

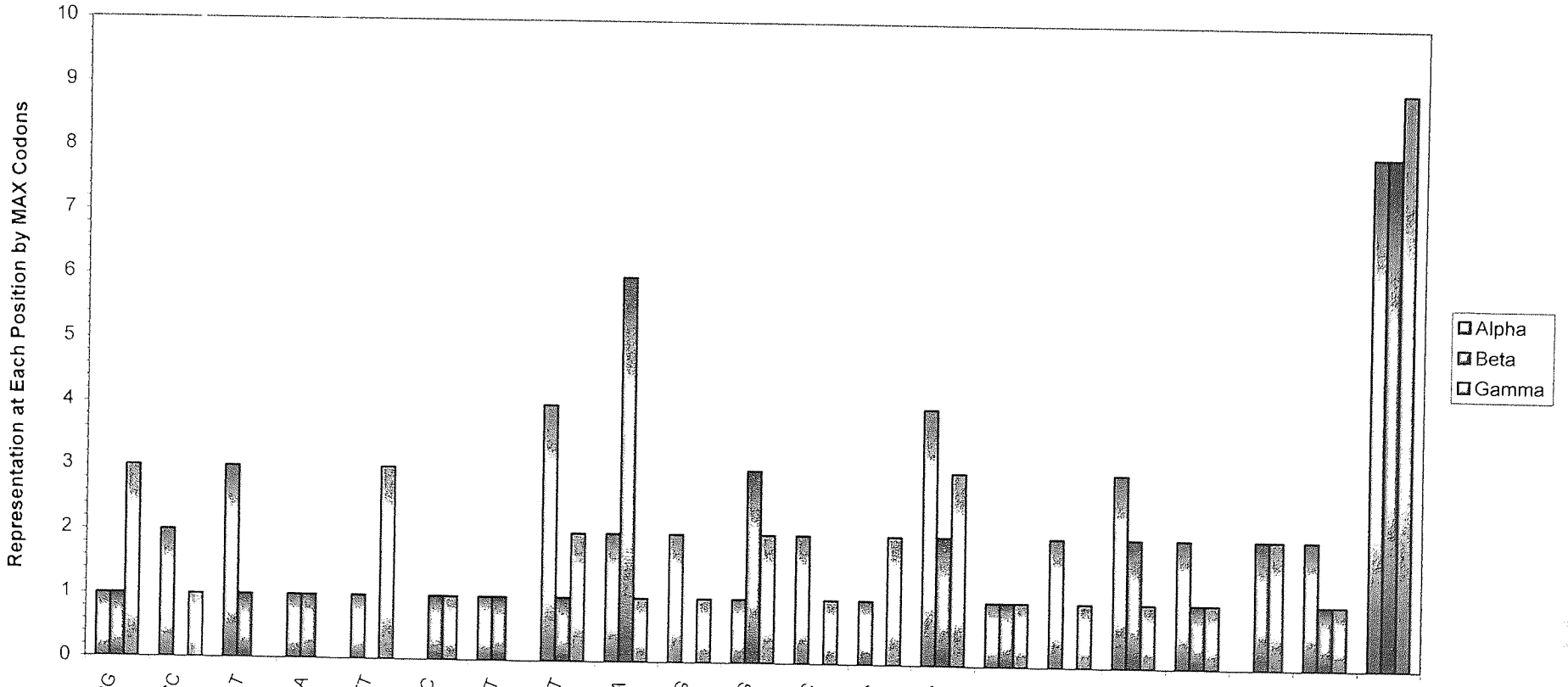


Fig 4.21b Graph demonstrating the amino acid representation by MAX codons at each randomised position, within the sequenced clones recovered from the cassette mutagenesis of the pGEX-ZFMA3 plasmid with MAX randomised cassettes generated in hybridisation buffer 2. Data includes frameshift mutations

4.11 Conclusions

Analysis of libraries constructed with the redesigned oligonucleotides suggested that the relocation of the MAX codon to the 3' ends of the oligonucleotides had successfully reduced the occurrence of non-MAX codons to similar levels in all locations.

Comparison of the libraries generated using the different hybridisation buffers showed that the use of buffer 1 resulted in a slight decrease in the number of non-MAX codons present at the randomised positions. Throughout the experimentation the use of buffer 1 had also resulted in more consistent clone recovery than buffer 2 (data not shown). Buffer 1 was therefore selected for future use. There was little apparent difference in encoded amino acid representation in libraries generated in buffers 1 and 2.

Given the relatively small sample sizes, it was encouraging to note that the majority of amino acids were encoded at all three positions within the sequenced genes. This suggested there were no significant biases in the overall randomisation procedure.

The recovery of a large percentage of clones containing frameshift mutations after randomisation, was unexpected. Inclusion of these clones in the sequencing data made subsequent analysis of libraries difficult, as certain beta and gamma codons could not be identified within the sequenced clones. Initially the inclusion of the frameshifted data in the analysis of codon inclusion at each randomisation position, was considered important, in order to compare the hybridisation of the redesigned selection oligonucleotides. However as the majority of frameshift mutations occurred after the alpha codon, inclusion of the frameshifted data resulted in an uneven distribution of the numbers of MAX codons at each randomised position. Analysis of this skewed data incorrectly suggested that selection oligonucleotides were performing differently, suggesting that this data should be excluded from such analyses.

Inclusion of the frameshifted sequences in the analysis of the encoded amino acid representation within libraries, did not appear to significantly alter the distribution of codons. However as the graphs were plotted from an equal number of randomised positions, inclusion of the frameshifted data did prevent the inclusion of a calculated ideal distribution of codons in the analysis.

The recovery of a large percentage of clones which possessed frameshift mutations after randomisation, clearly needed addressing. This problem might be inherent to the MAX randomisation technique itself (for example, *via* the repair of the synthetic DNA sequences after transformation; see Chapter 6) or else may be the result of selection of these sequences within the library, perhaps as a result of toxicity of the encoded proteins. If the encoded zinc finger protein was subject to leaky expression and was indeed toxic to the *E. coli* cells, then it might be predicted that selection for frameshifted clones would be encountered.

Chapter 5 Examination of Selection Pressure Within the Zinc finger Libraries.

5.1 Introduction

The factors which may affect representation within gene libraries have been discussed previously (Introduction and Section 3.3). The synthesis of randomised gene libraries is generally allied to the expression of these libraries to create a library of randomised proteins. Protein production in these libraries, is usually under the control of an inducible promoter such as those derived from the *lac* promoter sequences found in *E. coli*. As such, expression of the inserted sequence is regulated by *lac* repressor proteins. When the host cells are grown in the absence of lactose, transcription of the library genes is repressed. Once a large population of clones is established in culture, transcription of the gene and the consequent production of the randomised proteins can be induced by the addition of the lactose analogue IPTG.

When constructing gene libraries in *E. coli*, the use of inducible promoter sequences which are recognised by the host's translational machinery may result in some basal level expression of the cloned genes, prior to the induction of expression. This basal level of expression may lead to the negative selection of clones which produce proteins toxic to the host cell.

Initial library construction (Sections 4.3 and 4.6) was carried out in the ZFMA3 gene, which is inserted downstream of a *tac* promoter, in the expression vector pGEX-2TK. The *tac* promoter is derived from the *E. coli lac* promoter sequence and as such may result in some basal level expression of the inserted zinc finger protein gene. If zinc finger proteins present in the cell as a result of basal level expression prove toxic to the host, then clones encoding these proteins may be negatively selected from within the generated libraries

The potential toxicity of the zinc finger proteins generated by the libraries is difficult to ascertain. The libraries are expected to generate zinc finger proteins with novel DNA binding properties within the nominal sequence 5'-GGGNNNGCT-3'. Thus the probability of the binding site for a single zinc finger protein occurring in a random

DNA sequence can be calculated as occurring in every 262144 base pairs and thus is likely to occur within the *E. coli* genome of 4,600,000 bp (Blattner *et al.*, 1997).

The possible toxicity of finger proteins, that possess high affinity for a target site within the host cell chromosome, has profound implications upon the construction of the zinc finger libraries. Plasmids encoding toxic proteins are difficult to maintain in culture. This problem is exacerbated in liquid culture, where at sufficiently high cell densities the ampicillin in the media is completely destroyed. As this point is reached, cells which contain no plasmid are able to grow (Studier and Moffat, 1985). If strongly interacting zinc fingers have the potential to exert toxic effects upon the host *E. coli* cells, then clones containing genes encoding such zinc fingers may be lost from the culture.

In the synthesis of gene libraries, genes encoding products of differing affinities or functions are created from a single gene/plasmid construct. As antibiotic resistance is conferred by the plasmid, selection pressure will be exerted upon those clones which encode proteins that exert deleterious effects upon the host cells, even in the presence of antibiotic selection.

In the example of a fully randomised zinc finger library, randomisation of a single gene would be expected to result in the generation of a mixture of plasmids, capable of encoding zinc finger proteins with differing affinities for different target sites. Included in this mixture may be zinc fingers which bind, with high affinity, target sites which are present in the host cell genome, zinc fingers which bind, with high affinity, target sites which do not appear in the host cell genome and zinc fingers which exhibit low affinity for their respective target sites. In addition plasmids encoding non functional zinc fingers as a result of frameshift mutations in the encoding sequence, or as a result of self-ligation of the parental plasmid may also be present. The binding of target sites within the host cell genome by high affinity zinc finger proteins, (present in the cell as a result of basal level expression) may exert toxic effects upon the host cell, if the binding of such sites interferes with the transcription of important genes. Thus these plasmids will fail to thrive in liquid media in the presence of cells containing other plasmids from the library mixture, as the basal level transcription of non functional and low affinity zinc fingers, along with zinc fingers which bind, with high affinity, target

sites that occur within unimportant regions of the host cell, may not affect the growth and division of the host cell.

Cells containing plasmids which encode products which can affect the growth of the host are difficult to maintain in culture if the plasmid is transcribed by the host. In these cases cells which possess an expression defect are selected for in culture (Spher *et al.*, 2000). In the generation of the fully randomised zinc finger libraries (Section 4.6) a large number of the recovered sequences contained frameshift mutations. This raised the possibility that the recovery of such a large number of these sequences may have been due to the negative selection of plasmids encoding functional zinc fingers, due to the potential toxicity of these proteins to the host cell.

5.2 Assessment of Recovery of Clones Encoding High affinity and Frameshifted Zinc Finger Proteins in a Simple Model Library

To examine the potential selection pressures placed upon plasmids encoding both a high affinity zinc finger protein, and a non-functional (frameshifted) zinc finger, a series of experiments was carried out. The pGEX-ZFHM6 plasmid which encodes the QDR-RER-RHR zinc finger protein (Section 3.1) was used to represent a plasmid encoding a high affinity zinc finger protein and the frameshifted pGEX-ZFMA3 plasmid (Section 3.3.2) was used to represent a plasmid encoding a non-functional zinc finger.

Initially the plasmids encoding the high affinity and frameshifted zinc fingers were used to transform (2.4.1) aliquots of *E. coli* DH5α cells taken from the same cell preparation. The transformed cells were plated on LB media (2.1.2) containing ampicillin and the recovered colonies counted. The graph in Figure 5.1 shows the average number of colonies recovered. The graph demonstrates that transformation with the plasmid encoding the frameshifted zinc finger plasmid, results in greater clone recovery than transformation with a plasmid encoding a high affinity zinc finger. In addition, colonies recovered after transformation with the high affinity zinc finger plasmid were smaller in size than those recovered after transformation with the frameshifted plasmid.

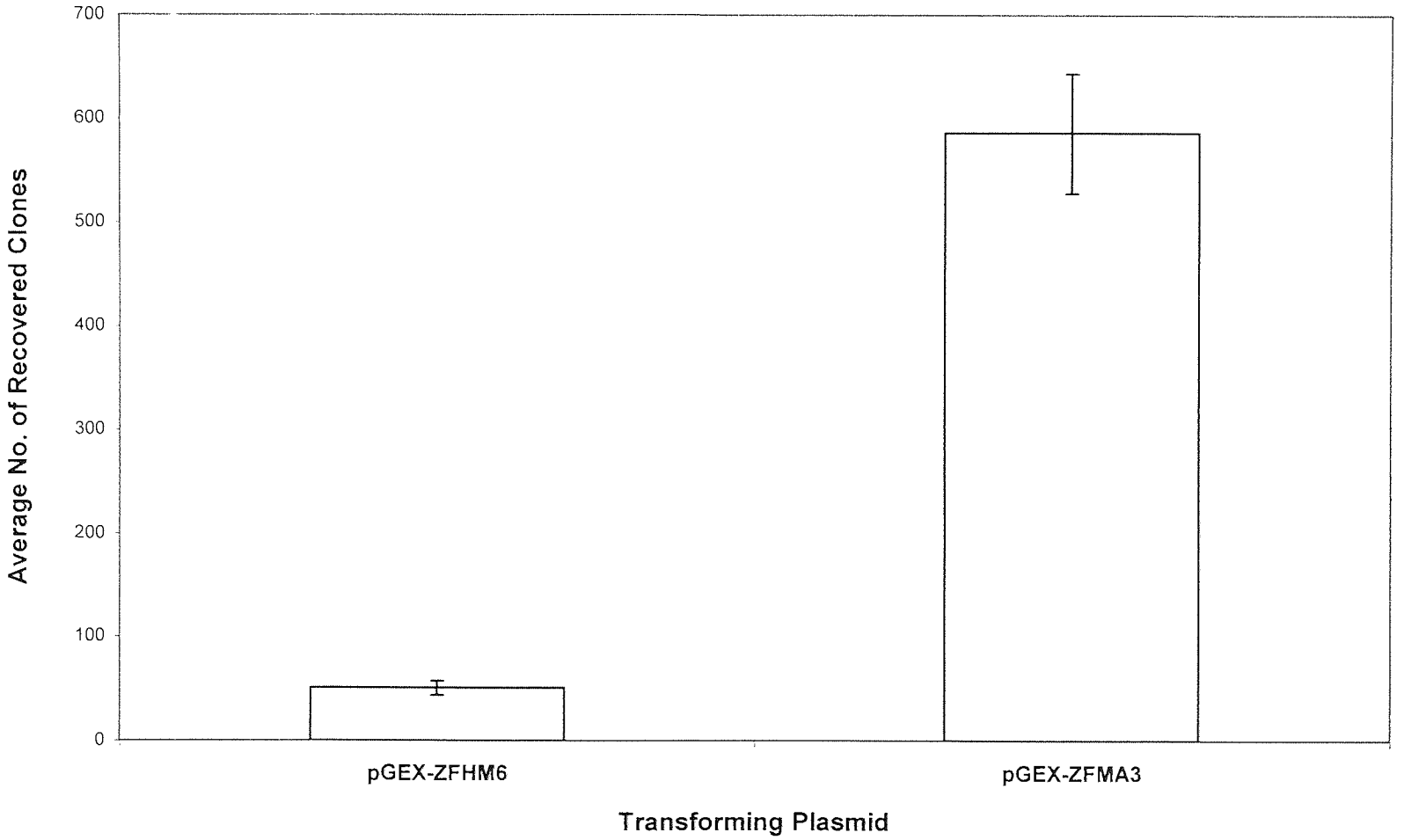


Fig 5.1 Graph demonstrating the average number of colonies recovered, after the transformation of *E. coli* DH5α cells with plasmids encoding a high affinity zinc finger protein (pGEX-ZFHM6) and a non functional zinc finger protein (pGEX-ZFMA3). Results are means and S.E.M of 8 replicates ($P < 0.05$ for difference in means by paired t test and Wilcoxon signed rank test).

The discrepancy in clone recovery after transformation with the two plasmids, which was greater than one order of magnitude, and the small size of the pGEX-ZFHM6 colonies suggested that plasmids encoding high affinity zinc fingers may be negatively selected within a library population.

To better mimic a randomisation experiment, the high affinity pGEX-ZFHM6 and frameshifted pGEX-ZFMA3 plasmids were then used to transform (2.4.2) a single aliquot of *E. coli* DH5α cells. The transformed cells were plated on LB media (2.1.2) containing ampicillin. The recovered colonies were differentiated on the basis of size. To ensure that this discrimination was accurate, samples of large (pGEX-ZFMA3) and small (pGEX-ZFHM6) colonies were screened using a standard PCR reaction (2.8.5) using the pGEX forward and reverse primers (Appendix A1). The resulting PCR products were digested (2.8.4) with the enzyme *Sma*I and visualised using agarose gel electrophoresis (2.5.1) to identify the recognition site for this enzyme, which is unique to the pGEX-ZFMA3 construct. The gel in Figure 5.2 demonstrates that with the exception of one failed PCR reaction, visual identification of the large and small colonies was reliable in the tested samples. The graph in Figure 5.3 shows the average number of large and small colonies recovered. As in the initial experiment, the number of colonies recovered as a result of transformation with the frameshifted plasmid, was greater than that recovered as a result of transformation with the plasmid encoding the high affinity zinc finger, by almost an order of magnitude. These results again suggested that high affinity zinc finger proteins may be negatively selected within the library population.

The results of the initial comparison of these plasmids had suggested that the plasmid encoding the high affinity zinc finger was exerting detrimental effects upon host cells. A second series of experiments was therefore carried out to ensure the difference in growth and/or division of clones was attributable to the zinc finger gene encoded by the plasmid.

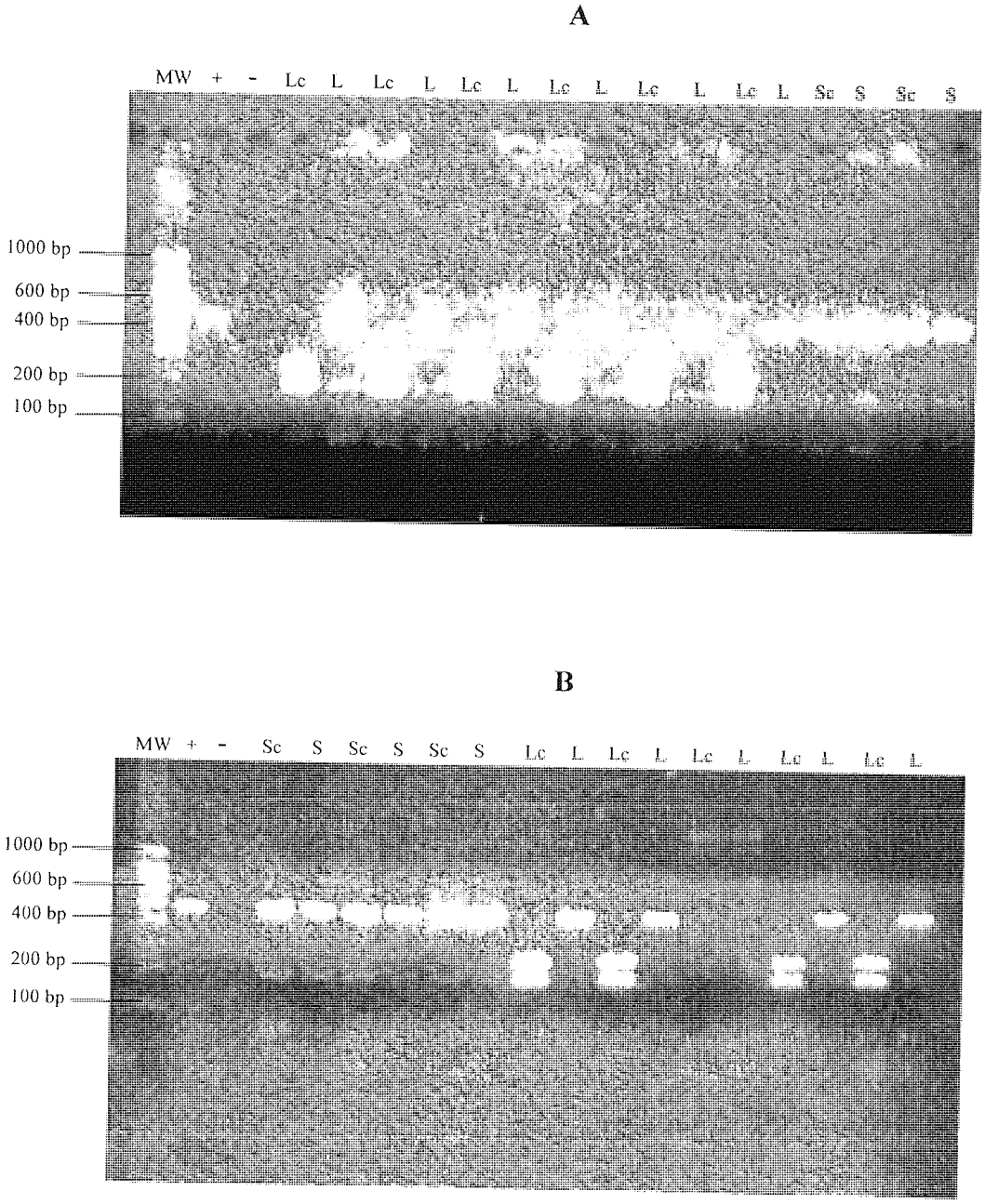


Fig 5.2 Analysis of *Sma*I digested PCR products, amplified from the large and small colonies recovered after the transformation of *E. coli* with the pGEX-ZFMA3 and pGEX-ZFHM6 plasmids (2 % agarose gel). Amplicons generated from amplification of The pGEX-ZFMA3 plasmid were expected to be 448bp in size, those generated by amplification of the pGEX-ZFHM6 plasmid were expected to be 465bp. To ensure accurate discrimination of the two amplicons, all products generated were digested with the enzyme *Sma*I. Products amplified from the pGEX-ZFHM6 plasmid are not digested with this enzyme. Products amplified from the pGEX-ZFMA3 plasmid are digested once with this enzyme producing products of 252 and 196bp. Key to figure: ; S = PCR products amplified from small colonies recovered after transformation; Sc = PCR products amplified from small colonies and digested with *Sma*I; L = PCR products amplified from large colonies recovered after transformation; Lc = PCR products amplified from large colonies and digested with *Sma*I; + = PCR products amplified from 0.5 ng of the pGEX-ZFHM6 plasmid as a positive control; MW = 500 ng 100 bp ladder (Bioline); - = Negative control.

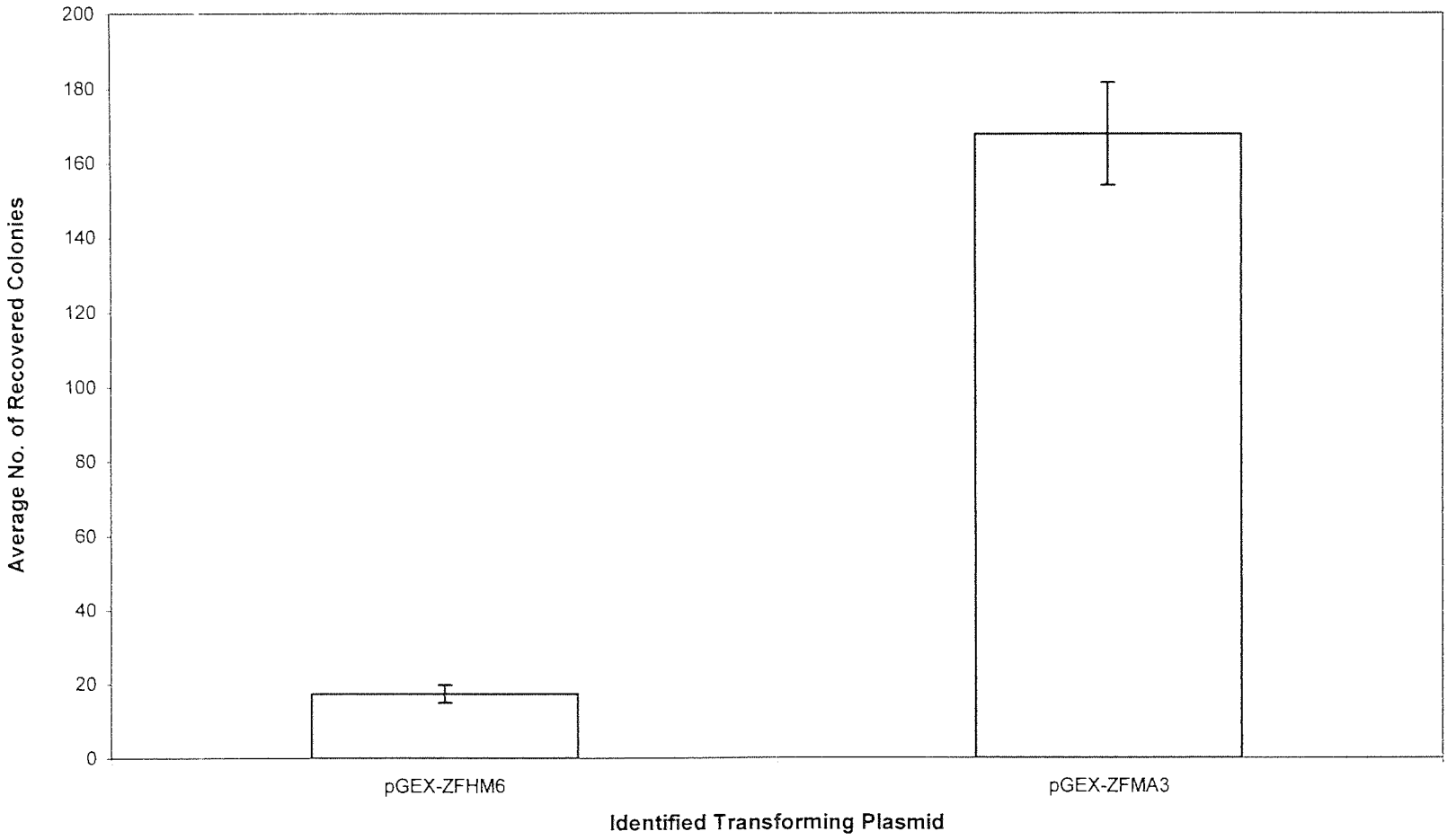


Fig 5.3 A graph demonstrating the average number of colonies identified as containing plasmids encoding a high affinity zinc finger protein (pGEX-ZFHM6) and a non functional zinc finger protein (pGEX-ZFMA3) after the transformation of *E. coli* DH5α cells with equal amounts of both plasmids. Results are means and S.E.M of 6 replicates ($p < 0.05$ for difference in means by paired t test and Wilcoxon signed rank test).

5.3 Confirmation of Toxicity Effects in the Recovery of Recombinant Clones

The initial comparison of the transformation efficiency of the plasmids encoding the high affinity and frameshifted zinc fingers, compared the transformation of two different plasmid constructs. This raised the possibility that inherent properties of the transforming plasmid, or the different plasmid preparations may have influenced the transformation efficiency of the differing constructs.

A second series of experiments was carried out, based upon the generation of plasmids encoding high affinity and frameshifted zinc fingers by cassette mutagenesis of a single plasmid. Generation of the plasmids in this fashion ensured that the differential recovery of clones from the high affinity and frameshifted plasmids, resulted directly from the encoded gene. In addition to the generation of individual plasmids, simple model library systems containing the two plasmids were constructed using both solid and liquid media to examine any potential selection which may occur within generated libraries.

The pGEX-ZFMA3 plasmid (section 3.3.2) used in the experiments is a frameshifted version of the pGEX-ZFHM6 plasmid with a truncated sequence inserted between the *Hin*dIII and *Bsi*WI sites of the zinc finger gene. The plasmid was digested with *Sma*I (2.8.4) and treated with CIP (2.8.1) before digestion (2.8.4) with the enzymes *Hin*dIII and *Bsi*WI.

Two mutagenic cassettes were synthesised (2.9.1). The two oligonucleotides M6 forward and M6 reverse when hybridised (2.9.2) together form a 37 bp cassette which (when subcloned into the pre-digested pGEX-ZFMA3 plasmid) generates the high affinity pGEX-ZFHM6 plasmid. The two oligonucleotides INS1 and INS1R when hybridised (2.9.2) together form a 20 bp cassette which regenerates the frameshifted pGEX-ZFMA3 plasmid when subcloned into the pre-digested pGEX-ZFMA3 plasmid. The cassettes were termed the High affinity (HAF) insert and the frameshift (FS) insert (Fig 5.4).

INS I

5′-AGCTTCGTTCCCGGGATGAC 3′

INS IR

5′ GTACGTCATCCCGGGAACGA 3′

Frameshift (FS) insert

5′-*AGCT*TCGTT**CCCGGG**ATGAC-3′
3′-AGCAA**GGGCCC**TACTG*CATG*-5′
***Hind*III** ***Sma*I** ***Bsi*WI**

M6 Forward

5′-AGCTTTAGTCGCAGCGACGAATAACAACGTCATCAGC-3′

M6 Reverse

5′-GTACGCTGATGACGTTGTAATTCGTCGCTGCGACTAA-3′

High affinity (HAF) Insert

5′-*AGCT*TTAGTCGCAGCGACGAATAACAACGTCATCAGC-3′
3′-AATCAGCGTCGCTGCTTAATGTTGCAGCAGTCG*CATG*-5′
***Hind*III** ***Bsi*WI**

Fig 5.4 Nucleotide sequences of the INS I and INS IR oligonucleotides and the M6 forward and M6 reverse oligonucleotides. The sequences are also show aligned as the hybridised frameshift insert and high affinity insert, designed to replace the 37 bp sequence between the *Hin*dIII and *Bsi*WI site of the ZFHM6 gene. The *Hin*dIII and *Bsi*WI cohesive termini of the hybridised insert are shown in italics. The *Sma*I recognition site is underlined in bold face.

5.3.1 Generation of Individual Plasmids Encoding High Affinity and Frameshifted Zinc Fingers by Cassette Mutagenesis

The two inserts were ligated (2.8.3) into the pGEX-ZFMA3 plasmid as described above and the ligation reactions used to transform (2.4.2) *E. coli* DH5α cells.

Interestingly both large and small colonies were recovered after transformation with the plasmid containing the high affinity insert (Fig. 5.5a). Since the clones containing the high affinity insert had been generated by the cassette mutagenesis of the pGEX-ZFMA3 plasmid, it was assumed that the large colonies may have resulted from the self ligation of the parental pGEX-MA3 plasmid. In contrast, transformation with the plasmid containing the frameshifted insert generated only large colonies (Fig 5.5b). The average numbers of colonies recovered in these experiments are shown in Figure 5.6, which again demonstrates that transformation with a plasmid encoding a non-functional zinc finger protein generates approximately one order of magnitude more colonies than transformation with a plasmid encoding a high affinity zinc finger protein.

Collectively these results suggest that negative selection pressure is placed upon cells transformed with plasmids encoding the QDR-RER-RHR protein, suggesting that plasmids containing frameshift mutations may be actively selected in library construction.

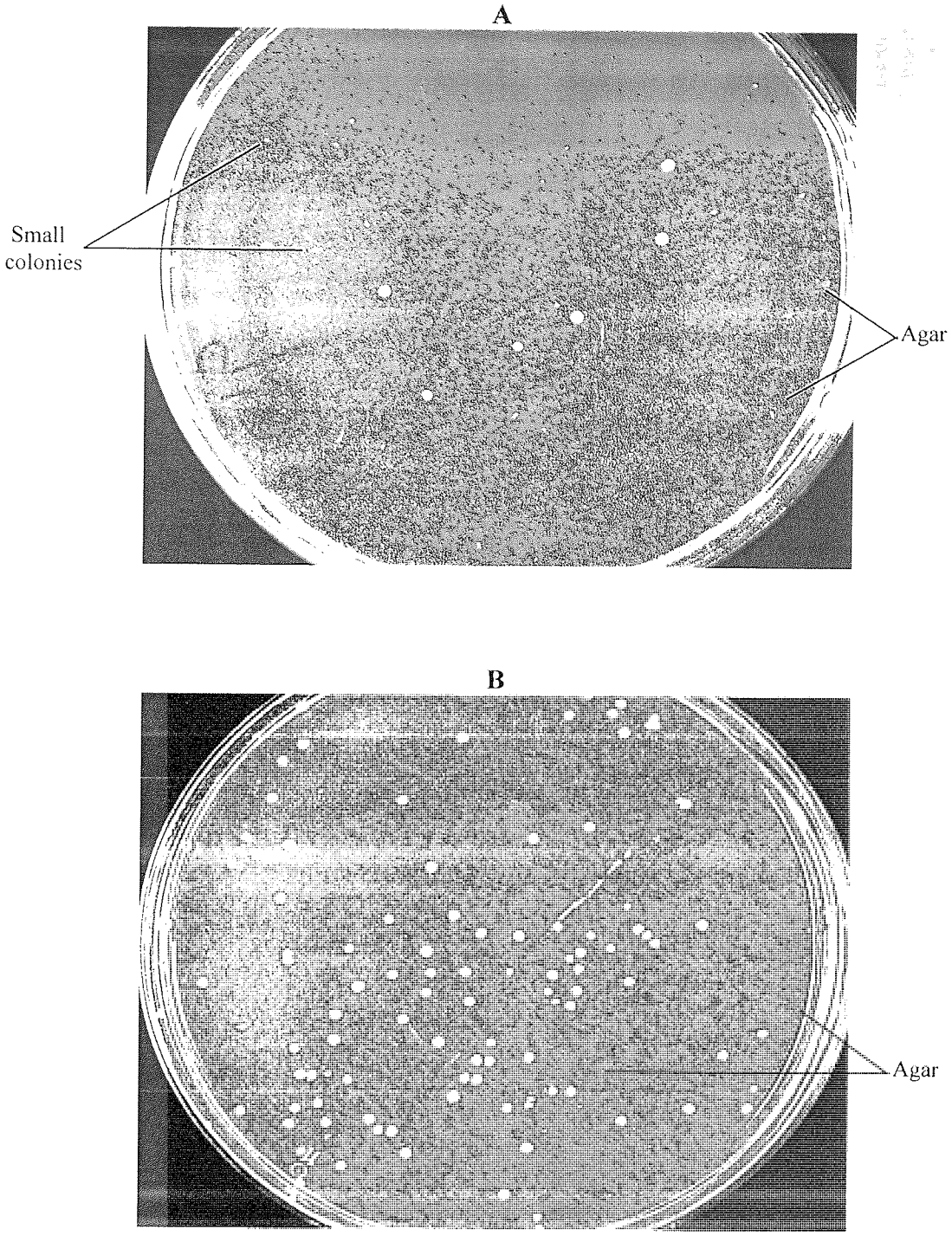


Fig 5.5 **A)** Photograph demonstrating the different sized colonies recovered after transformation of *E. coli* cells with pGEX-ZFMA3 plasmids containing the high affinity insert. **B)** Photograph demonstrating that similar sized colonies are recovered after transformation with pGEX-ZFMA3 plasmids containing the frameshifted insert. Small colonies are highlighted in the figure. Small pieces of agar scuffed from the surface of the plate, which resemble small colonies in the photograph but which were readily discernable on the plates are also highlighted.

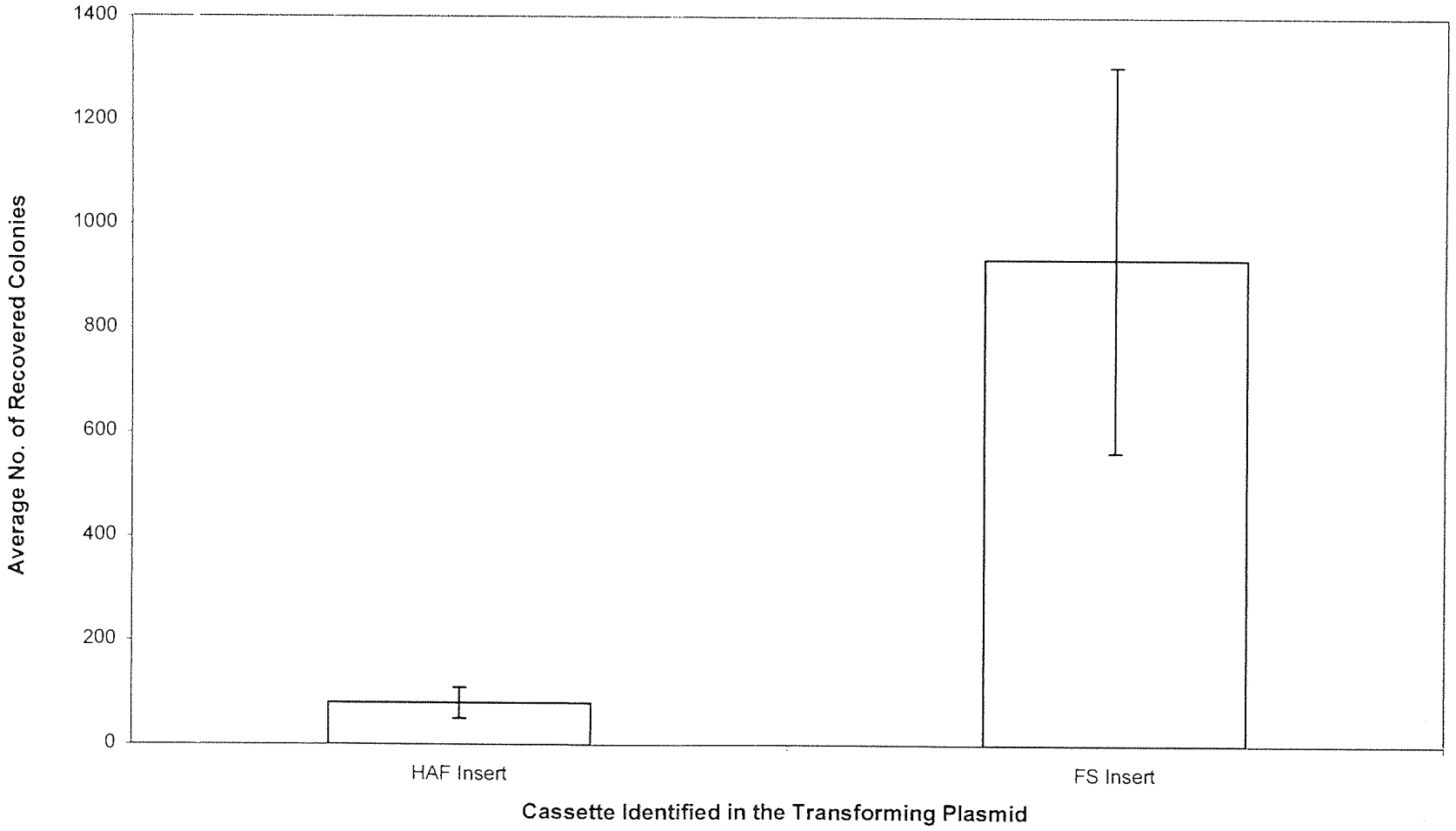


Fig 5.6 Graph demonstrating the average number of colonies recovered after the cassette mutagenesis of the pGEX-ZFMA3 plasmid with a cassette which generates a high affinity zinc finger gene (HAF insert) and a cassette which generates a non functional frameshifted zinc finger gene (FS Insert). Results are means and S.E.M of 8 replicates ($P < 0.05$ for difference in recovery by paired t test and Wilcoxon signed rank test). As the number of colonies recovered as a result of religation of the parental plasmid could not be ascertained in the experiments using the frameshifted insert, all recovered colonies were included in the analysis

5.3.2 Generation of a Simple Model Library Using Solid Media

The initial experimentation used to asses the selection of plasmids encoding high affinity and frameshifted zinc finger proteins, had relied upon the generation of individual plasmids by cassette mutagenesis. In the creation of libraries a mixture of cassettes, which generate plasmids of differing affinities, are cloned into a pre-digested plasmid in a single reaction, which could possibly increase selection pressure upon toxic clones.

To test this assumption reactions were performed in which equimolar amounts of the high affinity and frameshifted insert were pre-mixed and ligated (2.8.3) into the pre-digested pGEX-ZFMA3 plasmid. The ligation reactions were subsequently used to transform *E. coli* DH5α cells (2.4.2) as previously described.

Large and small colonies were recovered after transformation (Fig 5.7). It was assumed that the transforming plasmid in the small colonies contained the high affinity insert, and would encode a functional zinc finger. Transforming plasmids in the large colonies were expected to contain the frameshifted insert.

To re-verify this assumption, samples of large (pGEX-ZFMA3) and small (pGEX-ZFHM6) colonies were screened using a standard PCR reaction (2.8.5) using the pGEX forward and reverse primers (Appendix A1). The resulting PCR products were digested (2.8.4) with *Sma*I and visualised using agarose gel electrophoresis (2.5.1). The results are shown in Figure 5.8, which demonstrates that products amplified from the large colonies were digested with *Sma*I. Large colonies recovered in the transformation were classified as containing the frameshifted insert, and small colonies counted as containing the high affinity insert. Total numbers of large and small colonies (Fig. 5.9) demonstrate that after transformation of *E. coli* cells with plasmids containing an equimolar mix of the HAF and FS inserts, plasmids encoding the frameshifted zinc finger predominated in the recovered colonies. Plasmids which were expected to contain the high affinity insert DNA represented only 16 % of the total number of recovered transformants. In addition the total number of recovered transformants was approximately four times lower than that obtained when transforming cells with plasmids containing only the frameshifted insert.

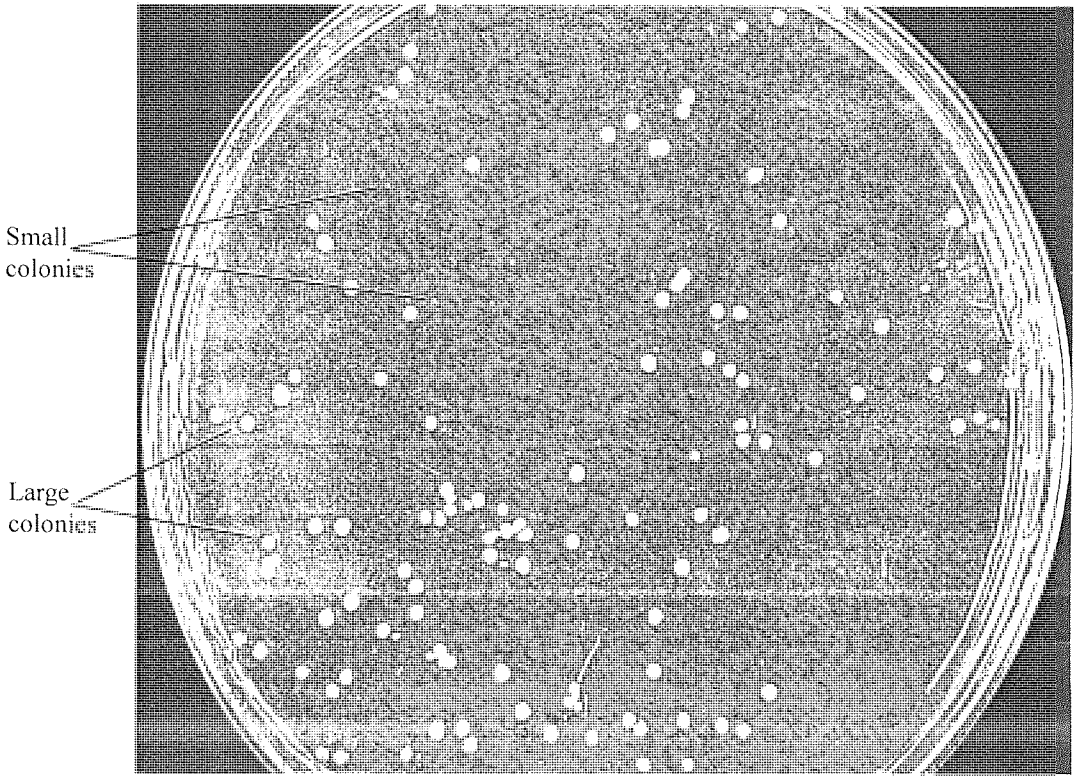


Fig 5.7 Photograph demonstrating the size difference in colonies, recovered after the transformation of *E. coli* with the pGEX-ZFMA3 plasmid containing an equimolar mix of the high affinity (HAF) and frameshifted (FS) insert.

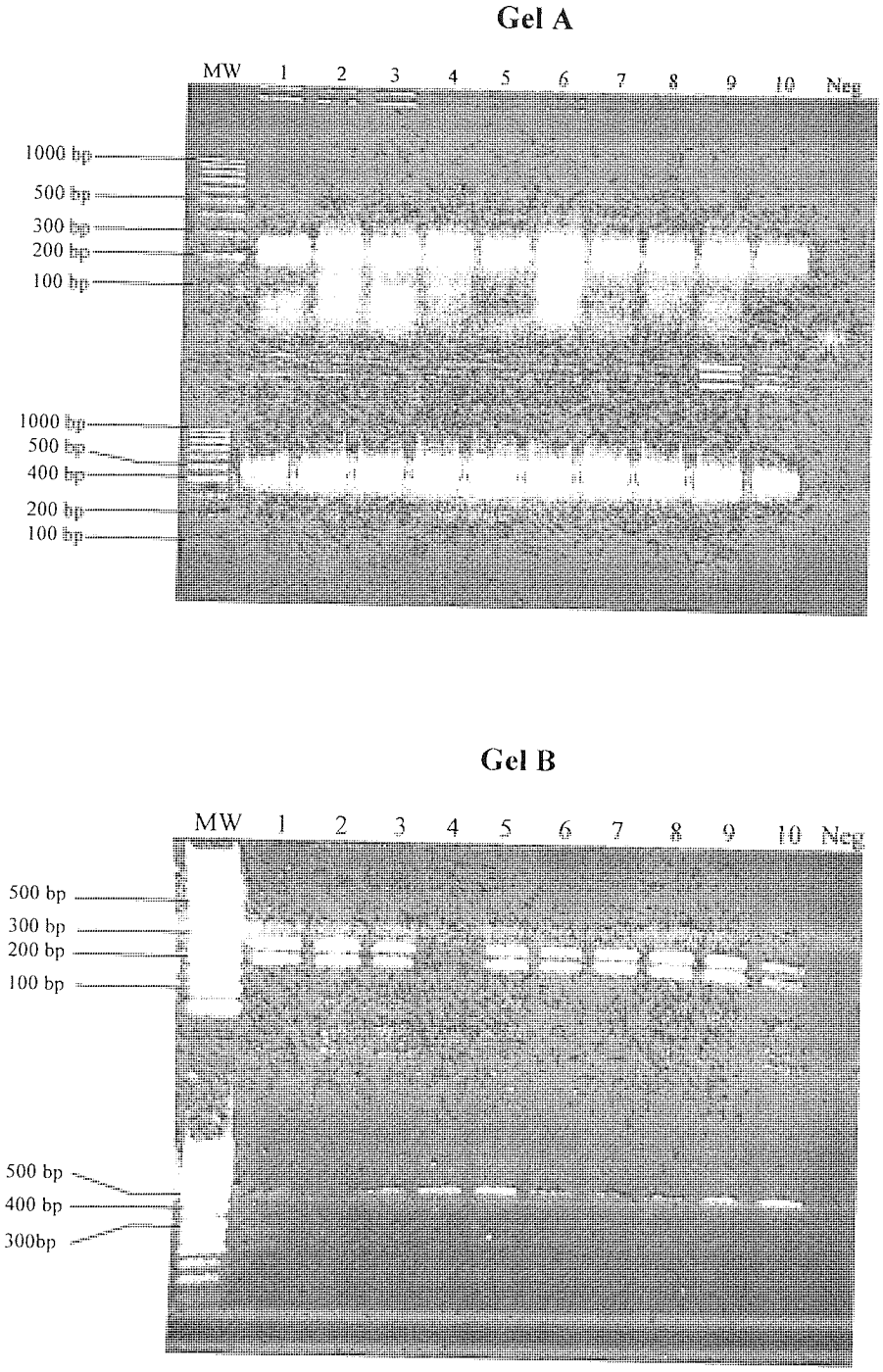


Fig 5.8 Analysis of *SmaI* digested PCR products, amplified from the large and small colonies obtained in the cassette mutagenesis of the pGEX-ZFMA3 plasmid using the high affinity and frameshifted DNA inserts (2 % agarose gel). Key to figure: MW = Sigma PCR low ladder in gel A and MBI Ferments GeneRuler 50bp ladder in gel B; Lanes 1 – 10 on the top row of both gels represent PCR products amplified from large colonies (corresponding to the predicted 252 and 198bp products of the *SmaI* digestion of PCR products amplified from a regenerated pGEX-ZFMA3 plasmid); Lanes 1 – 10 on the bottom row of each gel represent PCR products amplified from small colonies (corresponding 465bp PCR products amplified from the regenerated pGEX-ZFHM6 plasmid which does not contain a *SmaI* recognition site); Neg = Negative control.

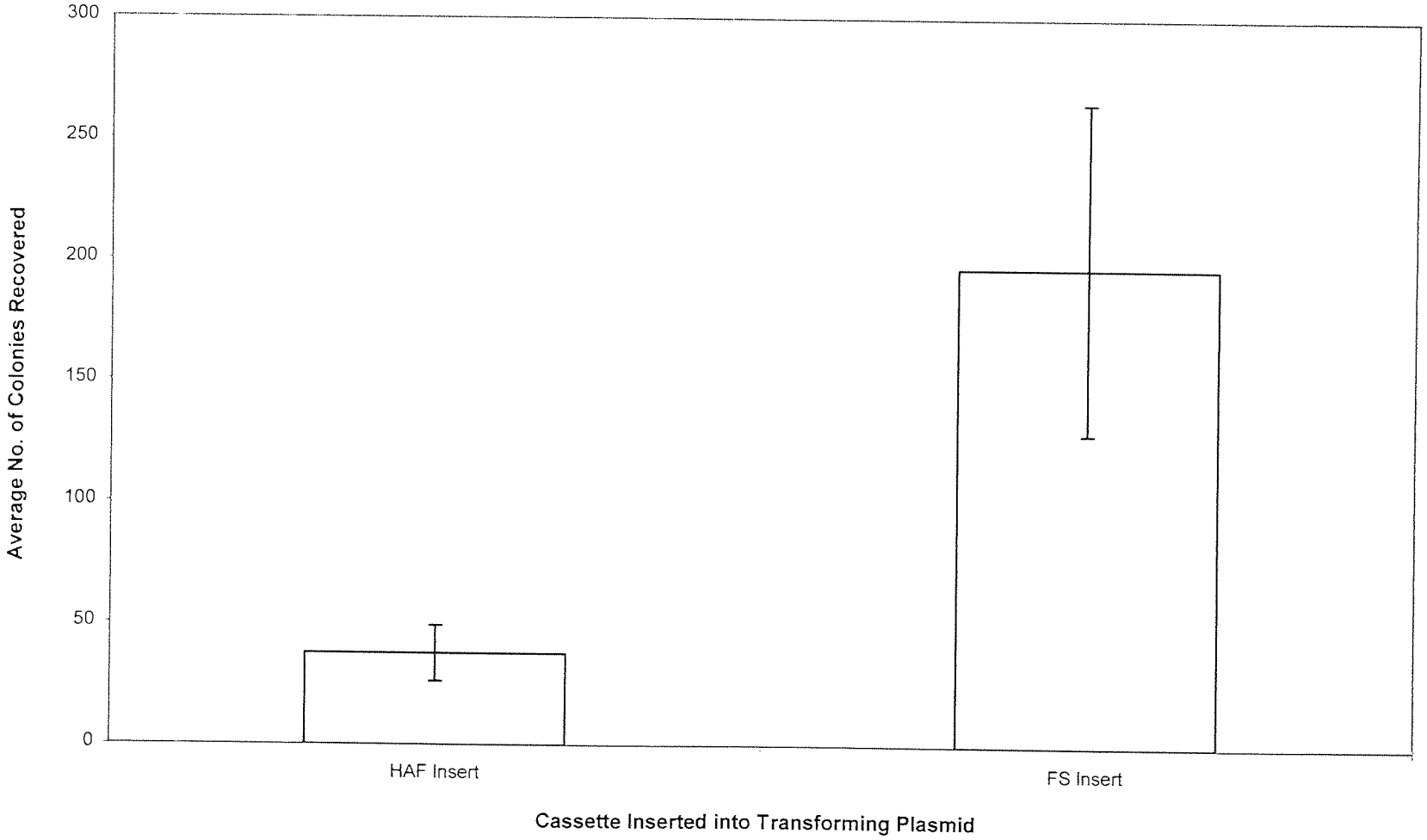


Fig 5.9 Graph demonstrating the average number of colonies recovered after the cassette mutagenesis of the pGEX-ZFMA3 plasmid with equimolar amounts of a cassette which generates a high affinity zinc finger gene (HAF insert) and a cassette which generates a frameshifted zinc finger gene (FS insert). Results are means and S.E.M. of 5 replicates ($P < 0.05$ for difference in means by paired t test and Wilcoxon signed rank test).

The results suggested that negative selection of the plasmids encoding the high affinity zinc finger was occurring when plasmids containing both inserts were transformed into *E. coli* cells and the cells plated directly onto selective media.

5.3.3 Generation of a Simple Model Library Using Liquid Media

Protein production from gene libraries requires cells to be grown in liquid media, which would be expected to place greater selection pressure upon plasmids encoding proteins toxic to the host cell. Experiments detailed in section 5.3.2 were repeated except that cells were cultured overnight in 30ml of liquid media containing ampicillin (2.1.1) prior to plating. After incubation, the cultures were diluted in series and plated onto LB media containing ampicillin (2.1.2).

The serial dilution of the cultured cells did not generate colonies which were easily differentiated by their size. Plates which had been inoculated with cells diluted at a ratio of 1: 10^8 contained between 5 and 30 colonies per plate. All colonies from these plates were screened by PCR (2.8.5) and the resulting products digested (2.8.4) with the enzyme *Sma*I. The digested products were visualised using agarose gel electrophoresis (2.5.1) to identify the insert contained in the transforming plasmid (Fig. 5.10).

The PCR results revealed that 50 of the screened clones contained the frameshifted insert, whilst five of the tested clones contained the high affinity insert. Clones recovered as a result of transformation with the plasmid encoding the functional zinc finger accounted for only 9 % of the total recovery. This result corresponded with those obtained in the previous experiments, suggesting that negative selection pressure placed upon these clones, was resulting in the selection of clones containing frameshift mutations. The negative selection of the high affinity zinc finger genes implied that basal level expression of this gene may interfere with the growth and/or division of the host cell.

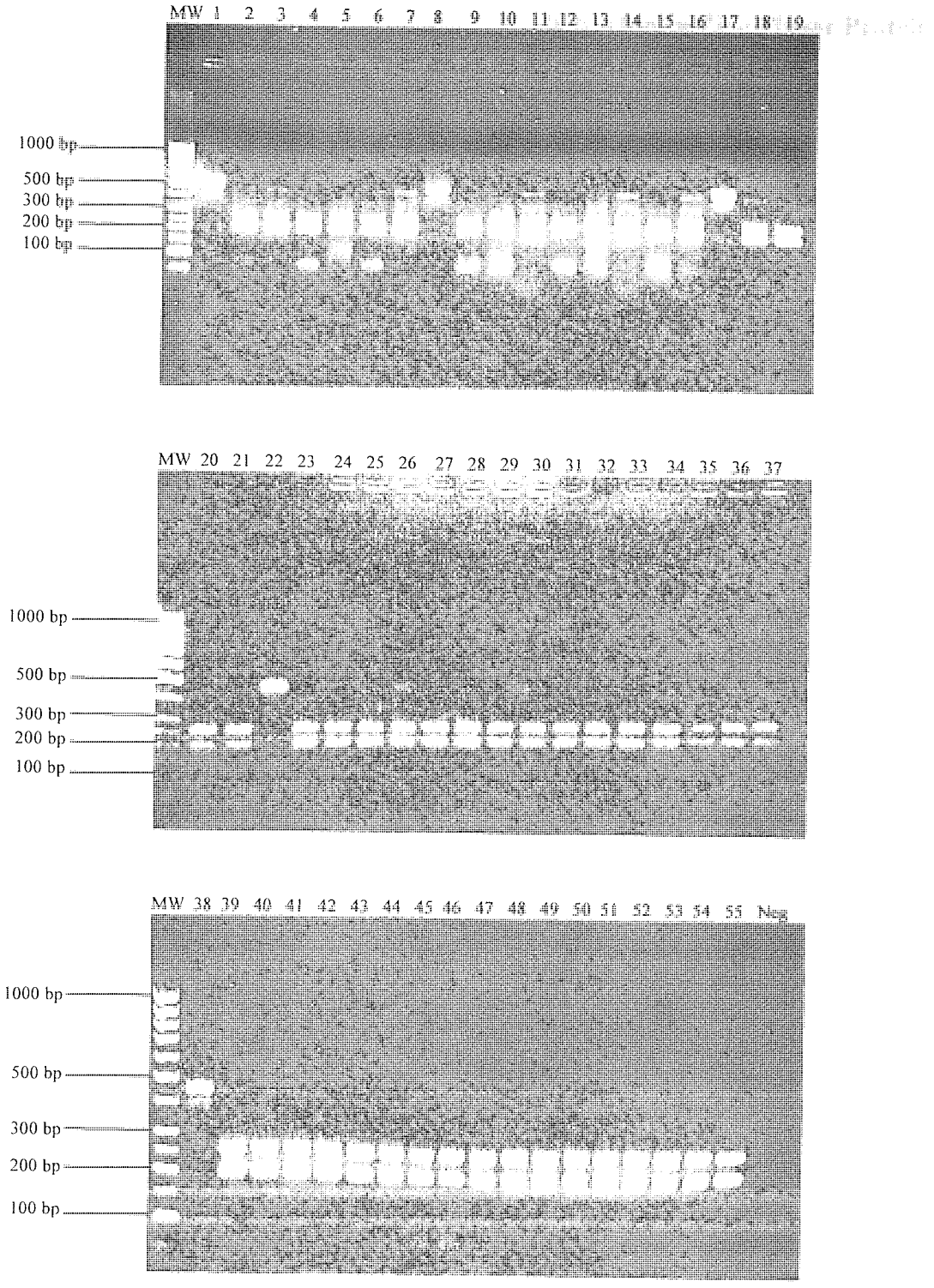


Fig. 5.10 Analysis of *SmaI* digested PCR products amplified from clones recovered after *E. coli* DH5α cells were transformed with the pGEX-ZFMA3 plasmid containing high affinity and frameshifted DNA inserts and the transformed cells grown in liquid media (2 % agarose gel). Key to figure: MW = GeneRuler 50 bp ladder (MBI Fermentas); Neg = Negative control; 1 – 55 = sample clones. Clones containing the frameshifted insert generate digested products of 252 and 196 bp. Clones containing the high affinity insert generate a single undigested product of 465 bp.

5.4 Search for Putative Binding Sites of the High Affinity Zinc Finger Protein Within the *E. coli* Genome

A BLAST search (2.4.8) of the *E. coli* genome was performed to identify the presence of the target site for the QDR-RER-RHR zinc finger protein encoded by the high affinity zinc finger gene. The target site (5'-GGG-GCG-GCT-3') was identified within the open reading frame of 14 genes within the *E. coli* genome. The identified genes and the proposed functions of the products of these genes are listed in Table 5.1.

Gene	NCBI Accession No. (*E Coli*)	Position In Genome	Gene Product	Function
AidB	AE000490	10132 - 10140	Putative acyl coenzyme a dehydrogenease	Energy metabolism
NrfE	AE000481	722 - 730	Formate dependant nitrite reductase	Anaerobic respiration
Spf	AE000461	12579 - 12571	Spot 42 RNA	Inhibition of DNA synthesis
NarG	AE000221	2002 - 2010	Nitrase reductase 1 subunit	Anaerobic respiration
MltE	AE000217	10455 - 10463	Murein transglycosylase E	Peptidoglycan synthesis
YbfD	AE000174	4557 - 4565	Putative DNA ligase	Not classified
YbfL	AE000174	3290 - 3298	Putative receptor	Not classified
YfiF	AE000344	7634 - 7626	Unknown	Not classified
YdjZ	AE000270	2922 - 2930	Unknown	Not classified
YdcC	AE000243	1947 - 1955	Putative receptor	Not classified
YhhS	AE000423	2521 - 2513	Putative transport protein	Not classified
RtcR	AE000418	6749 - 6741	Putative regulator	Not classified
YggA	AE000375	9123 - 9115	Unknown	Not classified
YfiM	AE000345	1723 - 1715	Unknown	Not classified
YhhI	AE000424	5847 - 5855	Unknown	Not classified

Table 5.1 Identification of the 9 bp target site of the QDR-RER-RHR zinc finger protein encoded by the high affinity zinc finger gene ZFHM6, within the genome of *E. coli*.

The presence of the target site in genes involved in metabolic functions, peptidioglycan synthesis and genes involved in the regulation of DNA synthesis, suggest possible mechanisms by which growth of transformed cells may be inhibited if these sites are bound by zinc finger proteins present in the cell as a result of basal level transcription.

The presupposition that the negative selection of clones containing the pGEX-ZFM6 plasmid was due to the toxic effects on the host of the QDR-RER-RHR protein, present in the cell as a result of basal level expression, has implications upon the generation of libraries in the pGEX-ZFMA3 construct. This negative selection pressure may be exerted upon other zinc finger proteins generated within the libraries. As a result of the negative selection of these clones, plasmids which encode non functional zinc fingers, low affinity zinc fingers, and zinc fingers which do not recognise target sites within the host, or recognise sites within non essential genes, may predominate within the library population. This presupposition corresponds with the data obtained in these experiments and the sequence results, obtained when sequencing the libraries constructed with the redesigned oligonucleotides, which highlighted that approximately 50 % of the sequenced clones frameshift mutations. To ascertain if this selection was a result of basal level transcription of zinc finger proteins and to minimise the potential for selection pressure to exist within the generated libraries, the library gene (ZFMA3) was subcloned into a different expression vector.

5.5 Subcloning of the Library Gene (ZFMA3) into a T7-Based Expression Vector

The basal expression of plasmid-encoded proteins results from the interaction of the host RNA polymerase with plasmid promoter sequences not regulated by repressor proteins within the host cell. The problem of maintaining plasmids which express proteins that are toxic to the *E. coli* host, can be overcome by cloning the gene of interest under the control of a promoter sequence for T7 RNA polymerase. T7 RNA polymerase and *E. coli* RNA polymerase recognise different promoter sequences (Dunn & Studier, 1983). As T7 RNA polymerase, derived from the T7 bacteriophage, does not occur naturally within the host cell, transcription of the cloned gene is prevented, as this promoter sequence is not recognised by the host's transcriptional machinery.

Induction of protein expression from the cloned gene can subsequently be achieved by utilising host cells carrying the T7 RNA polymerase gene under the control of a *lac*UV5 promoter and inducing with IPTG (Studier and Moffatt, 1986). Alternatively more stringent control of expression can be achieved by cloning the plasmid in an *E. coli* host lacking T7 RNA polymerase. Expression is then induced by infection with phage particles which encode T7 RNA polymerase. In this way a T7-based expression system is ideal for maintaining and expressing genes which encode products toxic to *E. coli* (Studier & Moffat, 1985). In addition rifampicin can be added to the culture after the induction of protein expression. The addition of rifampicin inhibits *E. coli* RNA polymerase and directs RNA synthesis from the T7 promoter alone, preventing growth and protein production from cells which have lost plasmid (Studier & Moffat, 1985).

In the context of library production, the use of a T7-based expression system would be expected to prevent any selection pressure placed upon clones encoding highly interactive zinc finger proteins, due to the basal level expression of these proteins.

The vector pET-42a (2.7.3) was chosen as a suitable T7-based expression vector for the cloning of the ZFMA3 gene under the control of a T7 promoter, as proteins expressed from this vector possess GST fusion tags. In addition, the recognition site for the enzyme *Bsi*WI which would be used in the preparation of the vector for library construction (Section 3.1, Fig 3.10) does not occur within the plasmid. The recognition

site for *Hin*dIII, which occurs once in pET-42a, is easily removed by cloning into the multicloning site of the vector. A vector map of pET-42a is contained in Fig 5.11.

Two potential problems were highlighted from examination of the pET-42a vector sequence. Firstly the vector possesses a *Sma*I site which is used in the preparation of the gene for library construction. This site is difficult to remove from the pET-42a vector as it occurs within the kanamycin resistance gene, thus the removal of this recognition site from the plasmid could only be achieved by silent mutation. In addition mutagenesis of this region is difficult as a functioning kanamycin gene is essential to clone recovery.

Secondly the vector also encodes two His.Tag sequences upstream and downstream of the multicloning site. The His.Tag sequences encode a series of histidine amino acids, which can be employed in the purification of the protein using the affinity of these residues for nickel ions. The zinc finger motif is stabilised by the binding of the divalent cation, zinc. The close proximity of these poly histidine sequences raised concern regarding the potential destabilisation of expressed zinc finger proteins, due to the ability of these residues to bind divalent cations, which could possibly interfere with zinc binding by the conserved cystine and histidine residues of the zinc finger proteins. A subcloning strategy was designed to remove the His.Tag sequences of the vector upon insertion of the zinc finger gene into the pET-42a vector. The design also permitted the pre-digestion of the inserted gene in preparation of the plasmid for library construction, despite the recognition site for the enzyme *Sma*I occurring within the pET-42a vector.

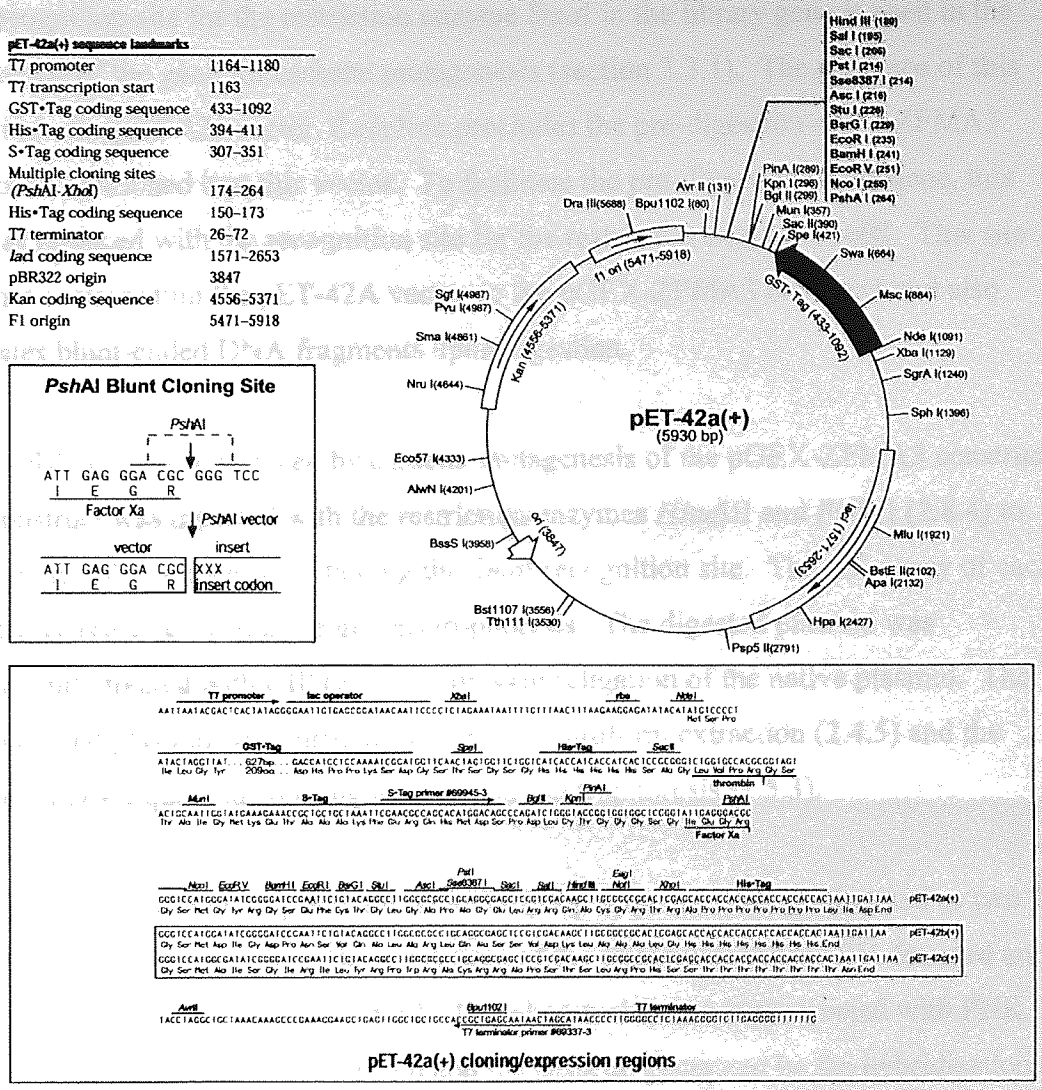


Fig 5.11 Map of the pET-42a vector. (Vector information downloaded from WWW.Novagen.com)

5.5.1 Replacement of the *Sma*I Recognition Site Within the ZFMA3 Gene

The recognition site for the restriction enzyme *Sma*I in the library gene is used in the preparation of the gene for cassette mutagenesis (section 3.3.1). The presence of this site within the pET-42a vector, therefore precludes the pre-digestion of the ZFMA3 gene once subcloned into this vector. To facilitate the pre-digestion of the gene, this site was replaced with the recognition site for the restriction enzyme *Sna*BI. This site does not occur within the pET-42A vector or the pGEX-ZFMA3 construct and also generates blunt-ended DNA fragments upon digestion.

The *Sna*BI site was introduced by cassette mutagenesis of the pGEX-ZFMA3 construct. The construct was digested with the restriction enzymes *Hind*III and *B*siWI (2.8.4) to remove the 20bp sequence containing the *Sma*I recognition site. The efficiency of each digest was assessed by agarose gel electrophoresis. The digested plasmid was subsequently treated with CIP (2.8.1) to prevent religation of the native plasmid. The phosphatased plasmid was purified by phenol:chloroform extraction (2.4.5) and the recovered DNA quantitated using agarose gel electrophoresis (2.5.3).

The insert, DN1, used to replace the removed 20bp sequence was synthesised as two 20 mer oligonucleotides (2.9.1). The sequences of the oligonucleotides DN1 forward and DN1 reverse are shown in Figure 5.12. The oligonucleotides were treated with PNK (2.8.2) to phosphorylate the 5' termini and the cassette generated by the hybridisation (2.9.2) of equimolar amounts of the phosphorylated oligonucleotides. The hybridisation of the two oligonucleotides creates a 20 bp cassette (Fig 5.12) bearing complementary overhangs for *Hin*dIII and *Bsi*WI for directional ligation into the pGEX-MA3 plasmid. The DN1 cassette was ligated (2.8.3) into the pre-digested pGEX-MA3 plasmid and the resulting ligation used to transform (2.4.2) *E. coli* DH5α cells. The numbers of recovered colonies are shown in Table 5.2.

TRANSFORMING DNA	COLONIES RECOVERED
No Plasmid	0
Self Ligation MA3	2
Insert Ligation	364

Table 5.2 Colonies recovered after the cassette mutagenesis of the pGEX-ZFMA3 plasmid, using the DN1 mutagenic cassette.

DN1 forward

5′-AGCTTCGTGTACGTACTGAC-3′

DN1 reverse

5′-GTACGTCAGTACGTACACGA-3′

DN1 insert

5′-*AGCT*TCGTG**TACGTA**CTGAC-3
3′-AGCAC**ATGCAT**GACTG*CATG*-5′
***Hind* III** ***Sna*BI** ***Bsi* W I**

Fig 5.12 Sequences of the DN1 forward and DN1 reverse oligonucleotides. The sequences are shown in the 5'-3' direction and hybridised together to form the DN1 mutagenic cassette. The *Hin*dIII and *Bsi*WI cohesive termini of the cassette are italicised, the *Sna*BI recognition site is underlined in boldface.

Several of the recovered clones were screened by PCR (2.8.5). The amplified products were digested (2.8.4) with *Sna*BI and the digested products analysed by agarose gel electrophoresis (2.5.1) to verify the presence of the introduced site. (Fig 5.13).

Figure 5.13 shows that only one of the screened clones did not contain the recognition site for *Sna*BI. Plasmid DNA was recovered on a small scale (2.4.6) from the colonies represented by the PCR products in samples 1 - 8, and the recovered plasmid DNA subjected to sequence analysis (2.8.8). The clone represented by the PCR products in sample 8 was identified as containing the correct sequence. The sequence of this plasmid, which was termed pGEX-ZFDN1, is contained in Figure 5.14.

5.5.2 Subcloning the Library Gene into the pET-42a Expression Vector

The synthesis of pGEX-ZFDN1 created a gene which could be inserted behind the T7 promoter of the pET-42a vector. The subsequent stages in the subcloning strategy were designed to remove the His.Tag encoding sequences from the pET-42a vector by the insertion of the ZFDN1 gene.

The His.Tag sequences of the pET-42 vector were removed by digestion with the restriction enzymes *Spe*I and *Bpu*1102 I. The removal of this sequence also removes the sequences which encode the thrombin cleavage site, used to facilitate separation of expressed zinc finger proteins from the GST fusion tag, the termination codon, which is contained within the multicloning site and part of the T7 termination primer sequence used in the commercial sequencing of the vector. In addition the distance between the stop codon and the T7 termination sequence is reduced by the removal of the multicloning site. The DNA cassette encoding ZFDN1 gene was therefore designed to encode the removed thrombin cleavage site, the T7 termination primer sequence and to re-establish the distance between the internal stop codon of the zinc finger gene and the T7 terminator sequence (to account for the possibility that this distance may be important in the termination of transcription by T7 RNA polymerase). The DNA cassette was also designed to encode a protein kinase site after the GST sequence of the pET-42a vector to permit the radioactive labelling of expressed proteins.

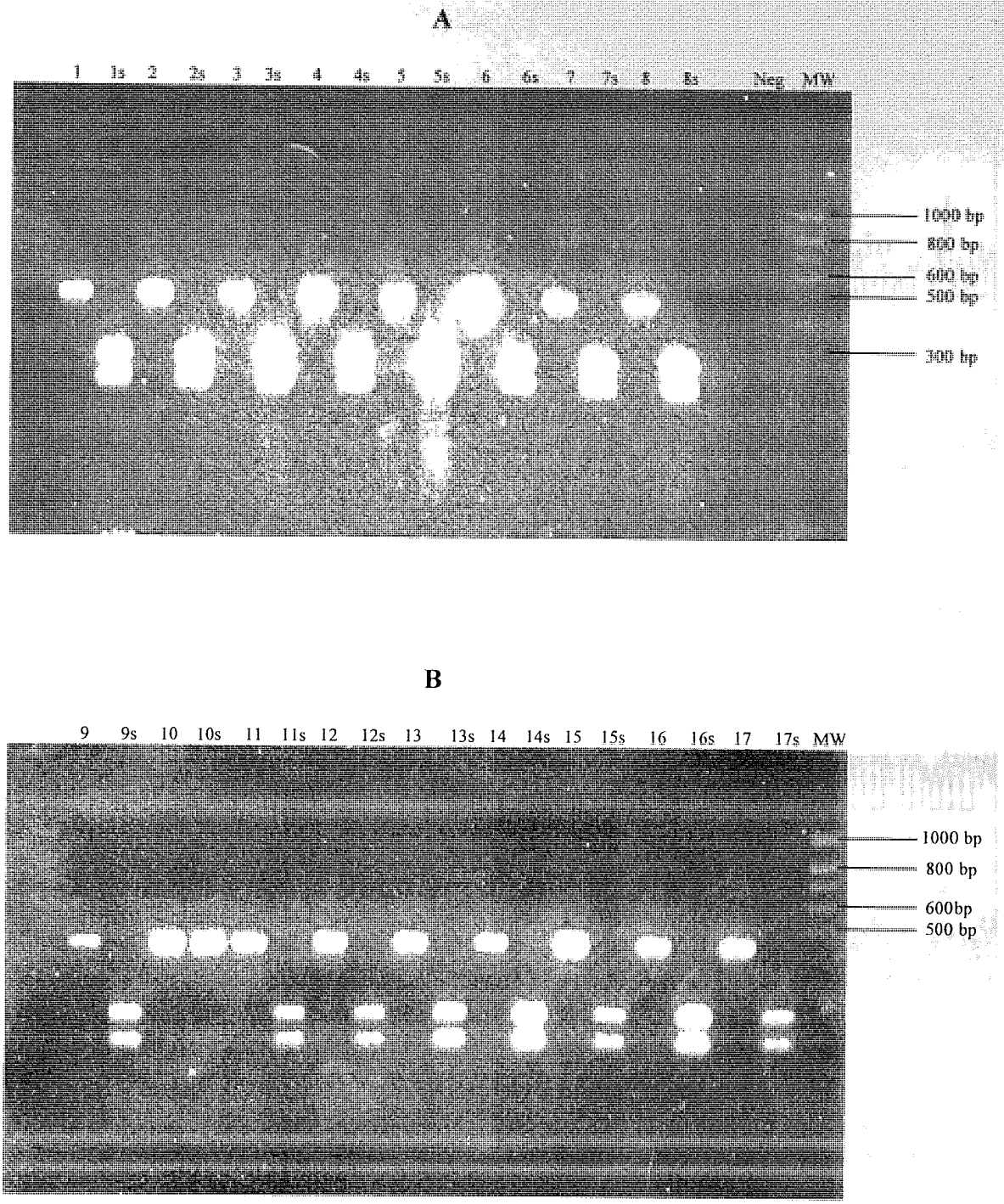


Fig 5.13 Agarose gel analysis of *Sna*BI digested PCR products, amplified from clones recovered after the cassette mutagenesis of the pGEX-ZFMA3 plasmid with the DN1 DNA insert (2 % agarose gel). Key to figure: MW = 100 bp ladder (Bioline); 1 – 17 = PCR products amplified from samples 1 – 17; 1s – 17s = PCR products amplified from samples 1 – 17 after digestion with *Sna*BI. The expected PCR product generated from the plasmid after successful insertion of the DN1 insert is 448bp. Digestion of this product at the introduced *Sna*BI site would be expected to generate two products of 252 and 196bp.

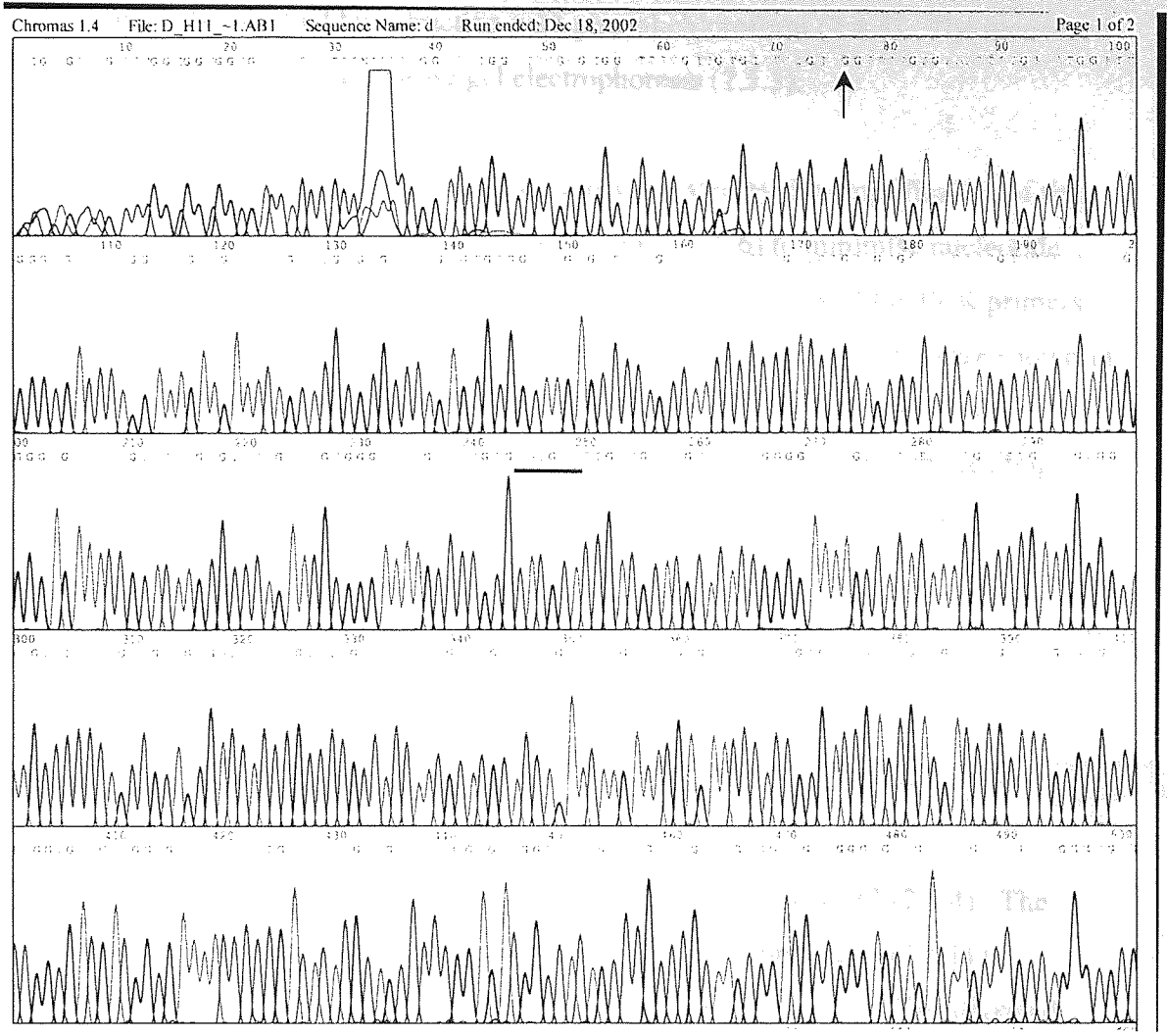


Fig 5.14 Sequence analysis of the pGEX-ZFDN1 construct. Base number 76 highlighted in the figure represents the start of the DN1 gene. The introduced *Sna*B I site is underlined. The predicted sequences of all the genes used in the experimentation are contained in appendix A2.

The pET-42a vector was digested with the enzymes *Spe*I and *Bpu*1102 I (2.8.4) and the efficiency of each digestion analysed by agarose gel electrophoresis (2.5.2). The digested vector was treated with CIP (2.8.1) to prevent religation of the native plasmid and subsequently purified by extraction with phenol:chloroform (2.4.5). The recovered DNA was quantitated using agarose gel electrophoresis (2.5.3).

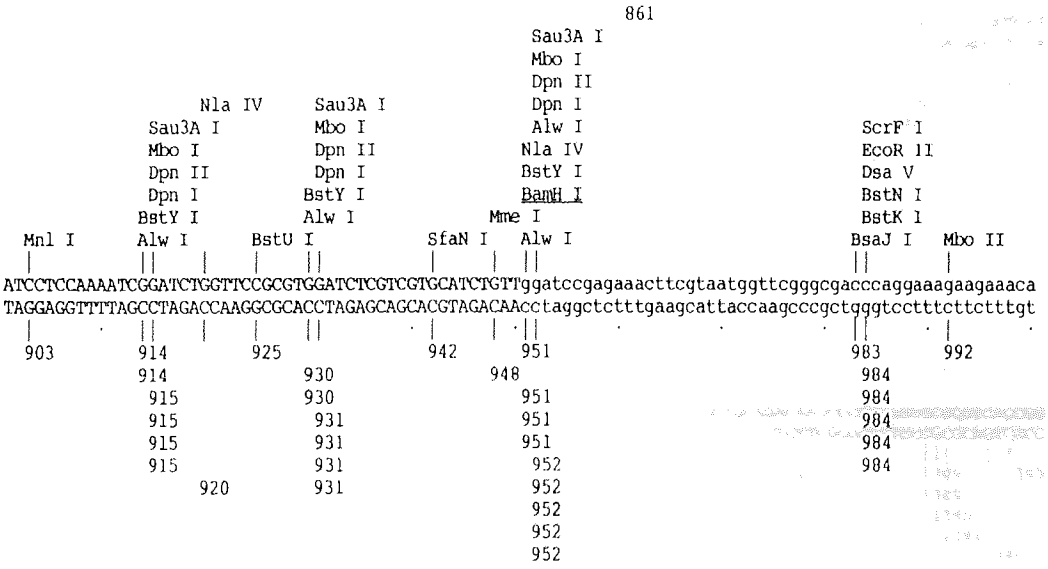
The insert DNA encoding the ZFDN1 gene was generated by the amplification of the pGEX-ZFDN1 construct using *Pfu* DNA polymerase (2.8.6) to minimise nucleotide misincorporation during amplification. The nucleotide sequences of the PCR primers termed ZFRET forward and ZFRET reverse employed in the amplification are shown in Figure 5.15 A and B. The predicted sequence of the vector after insertion of the expected insert was translated using the DNA Strider package to ensure the reading frame of the zinc finger pET-42a vector construct was correct (Appendix A2).

The amplified DNA was analysed using agarose gel electrophoresis (2.5.2). Analysis of the product showed a single band, corresponding to the expected 390 bp amplicon. The product of the PCR reaction was gel purified (2.5.2) and the DNA recovered by β agarase digestion of the gel slice (2.8.7).

The purified DNA was digested with the enzymes *Spe*I and *Bpu*1102 I (2.8.4). The products of the digestion were analysed by agarose gel electrophoresis (2.5.1) in comparison to an undigested sample of the amplified insert (Fig 5.16). The digested insert DNA obtained from the amplification of the pGEX-ZFDN1 construct was gel purified (2.5.2) and the DNA recovered by β agarase digestion of the gel slice (2.8.7). The recovered DNA was ligated (2.8.3) into the pre-digested pET-42a vector and the ligation reaction used in the transformation (2.4.2) of *E. coli* DH5α cells.

Sequence of the ZRFET Forward primer

5'-TT*ACTAGT*CTGGTTCCGCGTGGATC-3'

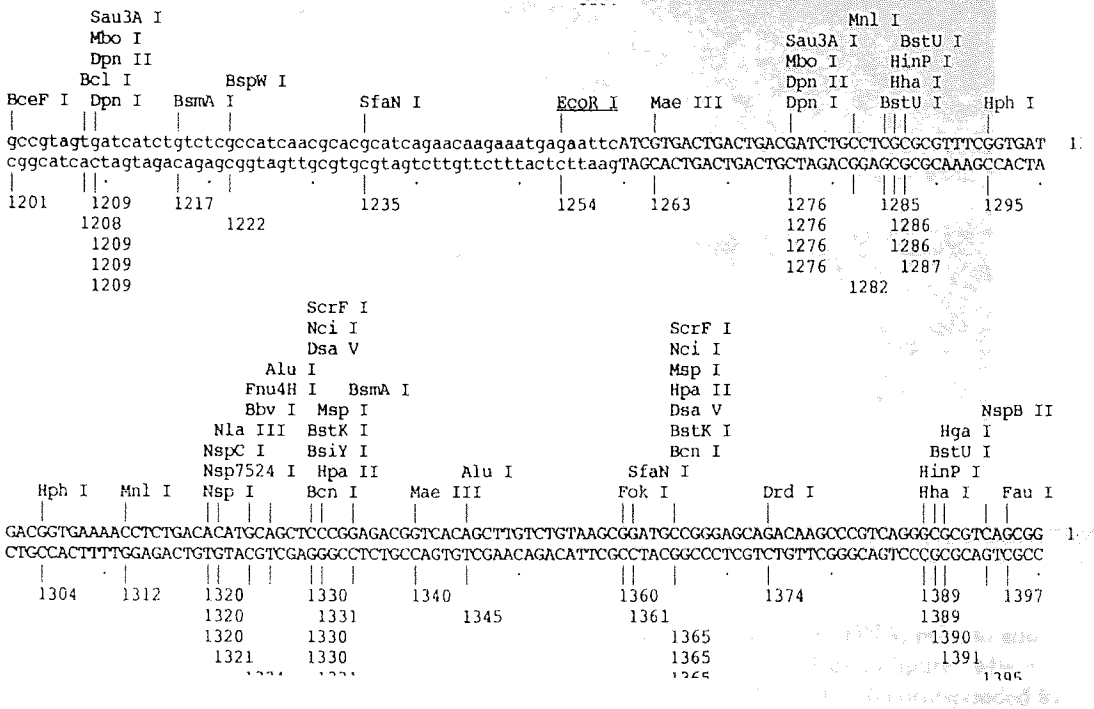


Sequence of the pGEX-ZFHM6 construct (positions 900 – 1000 bp)

Fig 5.15a Sequence of the ZFRET forward primer used to generate the correct *Spe* I recognition site in the amplification of the ZFDN1 zinc finger gene, before directional ligation into the pET42a vector. The sequence of the pGEX-ZFHM6 plasmid, on which the primer design was based, including the thrombin cleavage site (position 918 – 936) and the kinase site (position 936 – 950) is also shown. The use of the ZFHM6 sequence in the design ensures that library genes constructed in the pET42a vector will be maintained in the correct reading frame for protein expression. Key to figure: The underlined sequence of the ZFRET Forward primer is congruent to the sequence of the pGEX-ZFHM6 construct between positions 918 – 934 and is designed to amplify these two sites in addition to the ZFDN1 gene. The remaining sequence of the primer is non complementary to the pGEX-ZFHM6 construct and encodes a recognition site for the enzyme *Spe* I, highlighted in italics, with two additional flanking bases to aid the binding of the enzyme.

Sequence of the ZFRET Reverse primer

5'-TT*GCTCAGC***GGT**CGTCATCACCGAAACGCG-3'



Sequence of the pGEX-ZFHM6 construct (positions 1200 – 1400 bp)

Fig 5.15b Sequence of the ZFRET Reverse primer used to generate the *Bpu*1102 I recognition site in the amplification of the ZFDN1 zinc finger gene, before directional ligation into the pET42a vector. The ZFRET reverse primer is complementary to positions 1287 – 1304 of the non coding strand of the pGEX-ZFHM6 construct. During the amplification of the pGEX-ZFDN1, this primer amplifies an additional 50 bp downstream of the succeeding base to the stop codon of the ZFDN1 gene (position 1254 in the pGEX-ZFHM6 construct). This additional 50 bp sequence which is used to re-establish the distance between the stop codon and T7 terminator sequence present in the pET-42a vector before the removal of the His.Tag encoding sequences. The use of the pGEX-ZFHM6 sequence in the design of the primers ensured that the distance between the stop codon and T7 terminator would be maintained upon the generation of library genes constructed in the pET42a vector. Key to figure: The complementary sequence of the primer is underlined. The remaining sequence of the ZFRET Reverse primer comprises of the recognition sequence of the restriction enzyme *Bpu*1102 I highlighted in italics, with an additional two bases flanking this sequence to facilitate the recognition and binding of this site by the enzyme. The ZFRET Reverse primer also contains a 3 bp sequence, highlighted in bold face, designed to re-establish the T7 Terminator primer site when the amplified DNA is inserted into the pET-42a plasmid

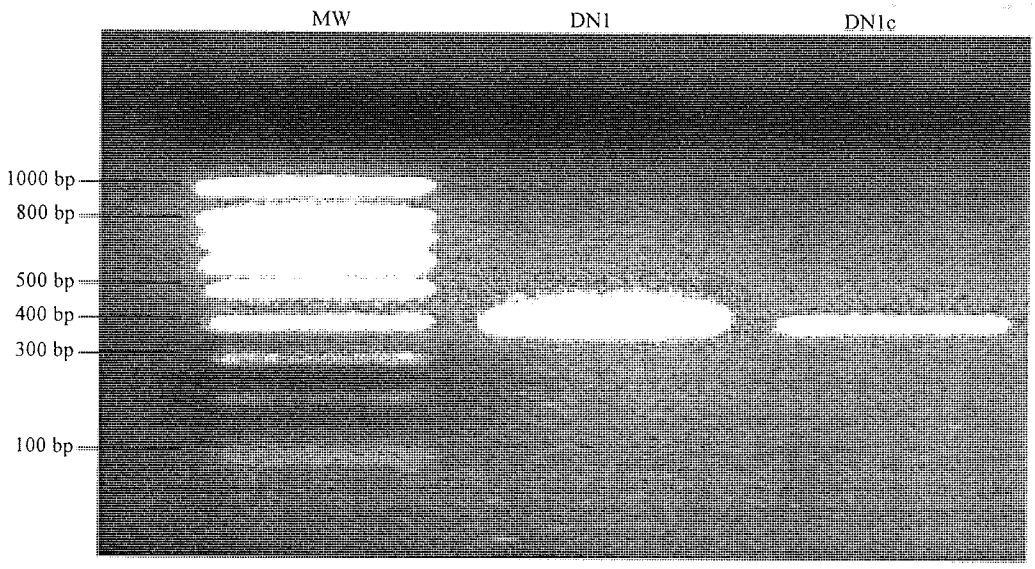


Fig 5.16 Agarose gel analysis of the amplified ZFDN1 and ZFHM6 insert DNA, prior to and after digestion with the enzymes *Spe*I and *Bpu*1102 I (3 % agarose gel). Key to figure: Mw = 500 ng, 100 bp ladder (Bioline); DN1 = Amplified ZFDN1 insert DNA, which corresponded to the expected 390bp amplified product; DN1c = Amplified ZFDN1 insert DNA after digestion with *Spe*I and *Bpu*1102 I which would be expected to generate a product of 383bp.

The results obtained when transforming the pET42a vector containing the amplified insert into *E. coli* DH5α cells are shown in Table 5.3.

TRANSFORMING DNA	COLONIES RECOVERED
No Plasmid	0
PET-42a Vector	13

Table 5.3 Colonies recovered after transformation of *E. coli* DH5α cells with the amplified zinc finger gene, after direct subcloning into the pET-42a expression vector.

The 13 clones were screened by PCR (2.8.5) using the ZFRET forward and reverse primers. Analysis of the PCR products suggested that eight of the clones contained the ZFDN1 insert. Plasmid DNA was extracted from these eight clones using small scale plasmid recovery (2.4.6). The recovered plasmids were subjected to a restriction enzyme digestion (2.8.4) with the enzyme *Sna*BI to identify the *Sna*BI recognition site encoded in the DN1 insert DNA. The presence of this site was identified by agarose gel electrophoresis (2.5.1) of the digestion reactions (Fig 5.17).

Analysis of the agarose gel showed that each of the recovered clones contained the recognition site for the restriction enzyme *Sna*BI, suggesting the presence of the DN1 insert. Plasmids obtained from three of these clones were selected for sequencing (2.8.9). Sequence analysis showed that each of the three clones contained the ZFDN1 zinc finger gene including the thrombin and kinase sites amplified from the pGEX2-TK plasmid and that the inserted DNA was maintained in the correct reading frame of the pET-42a plasmid. A single clone represented by the sequence data in Figure 5.18 was amplified in LB broth (2.1.1) and the plasmid recovered using large scale plasmid recovery (2.4.7) The plasmid was named pET-ZFDN1.

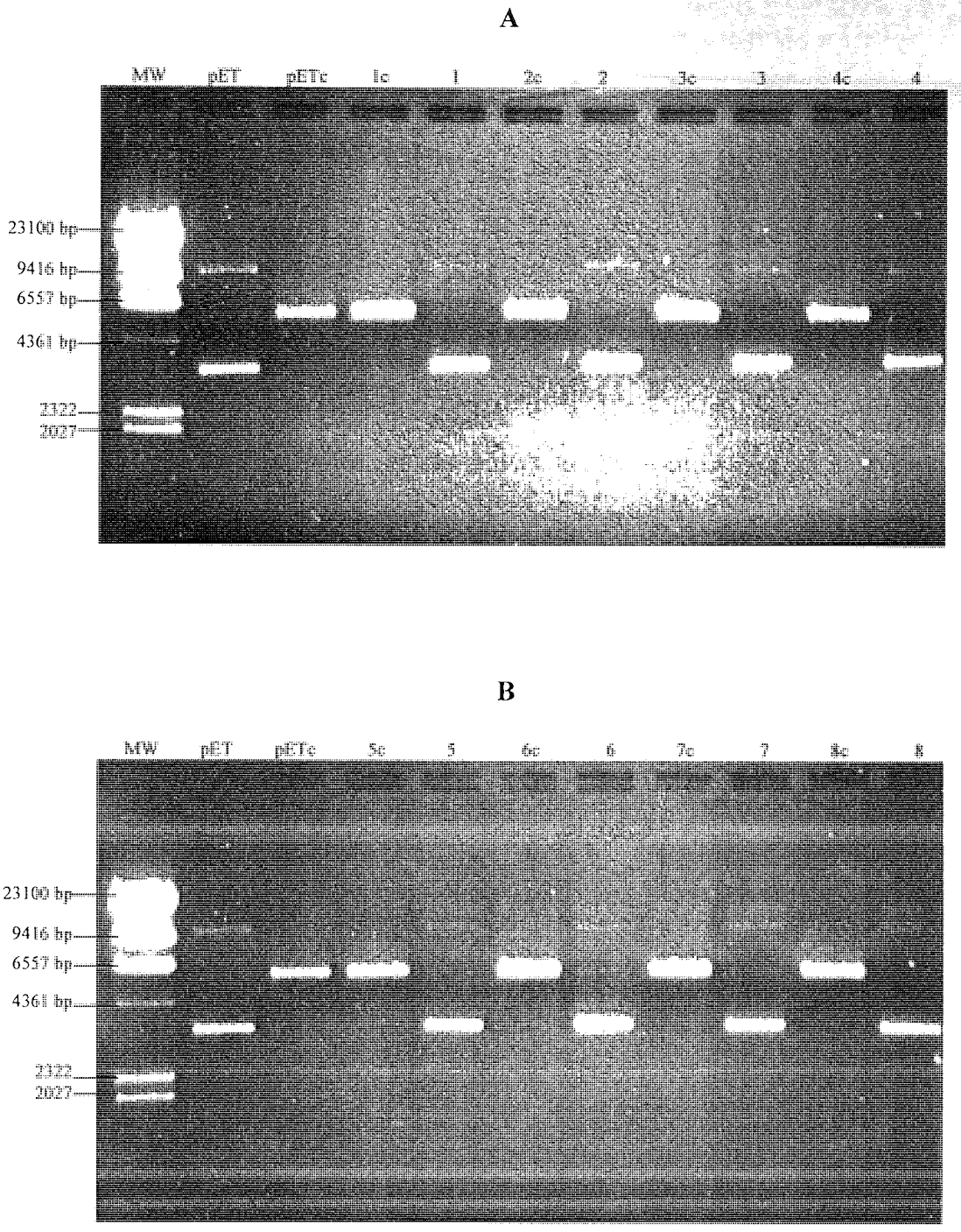


Fig 5.17 Restriction digest analysis of plasmid DNA recovered after the subcloning of the ZFDN1 insert into the pET-42a vector (1 % agarose gel). Key to figure: MW = 500 ng *Hin*dIII digested λ DNA; pET = 500 ng pET-42a; pETc = 500 ng pET-42a linearised by digestion with the restriction enzyme *Nde*I; 1c – 8c = Plasmid DNA recovered from clones 1 – 8 digested with *Sna*BI; 1 – 8 = Native plasmid DNA recovered from clones 1 – 8.

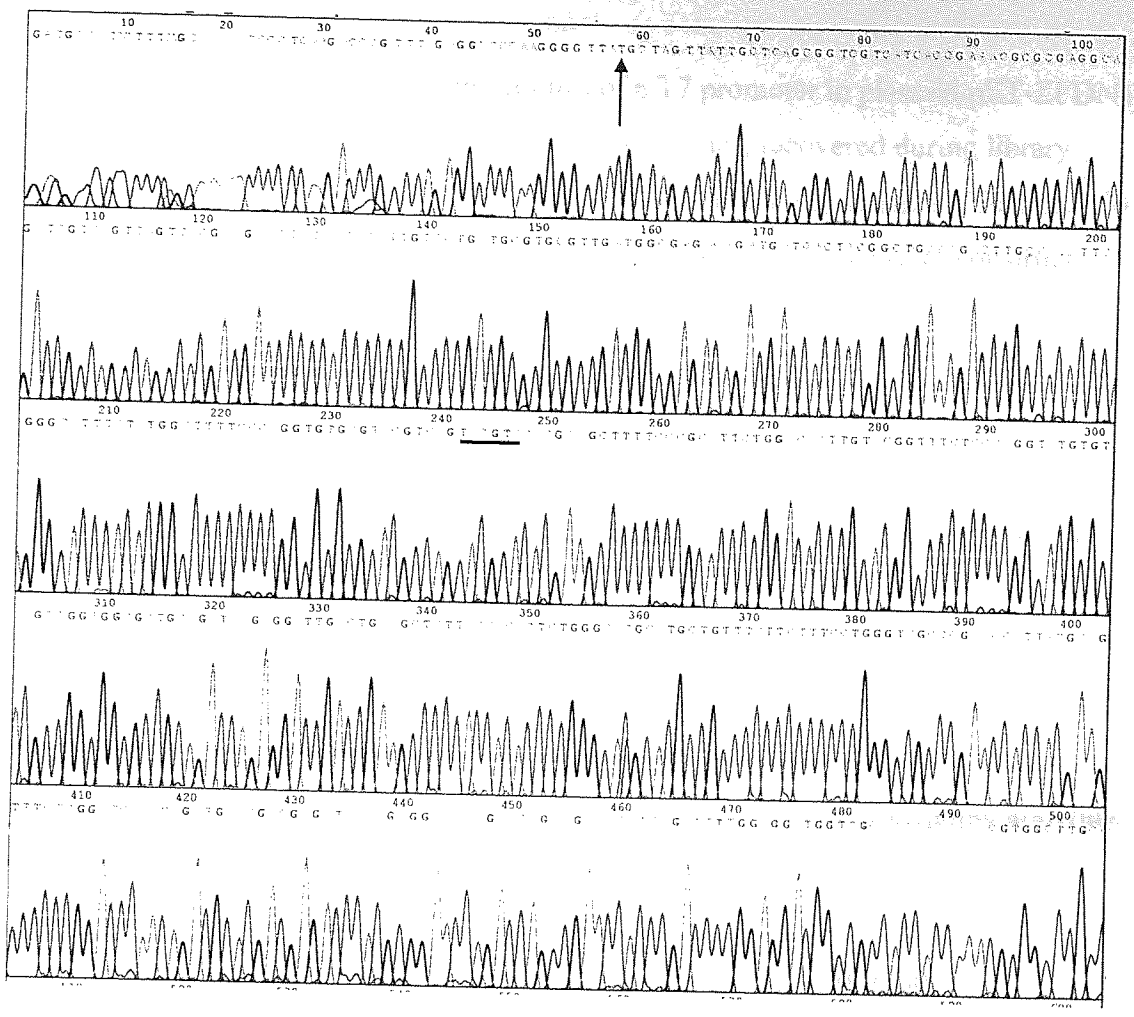


Fig 5.18 Sequence analysis of the pET-ZFDN1 plasmid. Base number 58 highlighted in the sequence represents the T7 terminator primer, followed by the sequence of the pET-ZFDN1 gene. The introduced *Sna*B I site is underlined. Note the plasmid sequence is shown in the reverse orientation as this sequence was used to confirm the regeneration of the T7 terminator primer sequence

5.6 Assessment of the Recovery of Clones Encoding High affinity and Frameshifted Zinc Finger Proteins in the T7-Based Expression Vector

Placing the zinc finger gene under the control of a T7 promoter in plasmid pET-ZFDN1 was expected to reduce the number of frameshifted genes, recovered during library construction. Prior to library construction, the model library experiments, (Section 5.3) were therefore repeated, to assess the effect of the T7-based pET-ZFDN1 construct upon clone recovery.

The pET-ZFDN1 construct was pre-digested (2.8.3) with *Sna*BI and treated with CIP (2.8.1) before digestion with the enzymes *Hind*III and *Bsi*WI (2.8.4). The M6 forward and M6 reverse and the INS1 and INS1R oligonucleotides were hybridised (2.9.2) to form the high affinity and frameshifted inserts used in the previous experimentation. Although the frameshifted insert (FS) generates a *Sma*I recognition site when inserted into the pre-digested pET-ZFDN1 plasmid, the PCR primers used to identify the high affinity and frameshifted clones do not amplify the *Sma*I recognition site of the pET-42a plasmid. Thus this insert could still be identified in the pET-ZFDN1 plasmid, allowing direct comparison with the previous results.

5.6.1 Generation of Individual Plasmids Encoding High Affinity and Frameshifted Zinc Fingers by Cassette Mutagenesis

Ligation reactions (2.8.3) were performed to ligate the high affinity and frameshifted insert into the pre-digested pET-ZFDN1 plasmid. The ligation reactions were then employed in the transformation (2.4.2) of *E. coli* DH5α cells. The transformed cells were plated on LB media (2.1.2) containing kanamycin.

Analysis of the recovered clones showed that colonies formed as a result of transformation with the plasmid containing the high affinity insert were of similar size to those containing the frameshifted insert (Fig 5.19). The recovered colonies were counted and the data used to plot the graph in Figure 5.20 which shows the average number of colonies recovered after the transformation of *E. coli* with plasmids containing each insert.

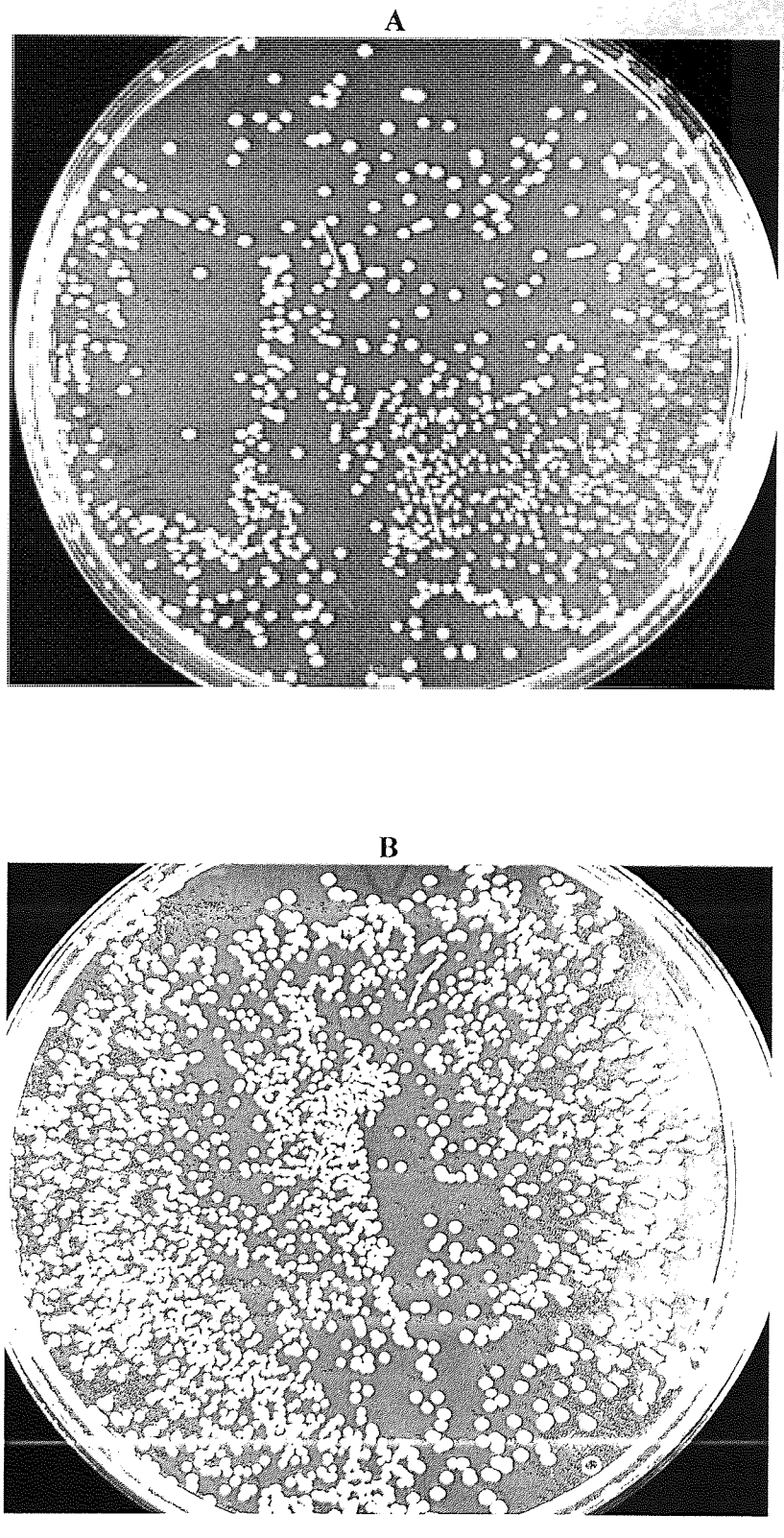


Fig 5.19 The similar sized colonies recovered after cassette mutagenesis of the ZFDN1 gene in the T7-based pET-ZFDN1 construct, using the HAF insert which generates a high affinity zinc finger protein (**A**) and the FS insert which generates a frameshifted zinc finger protein (**B**).

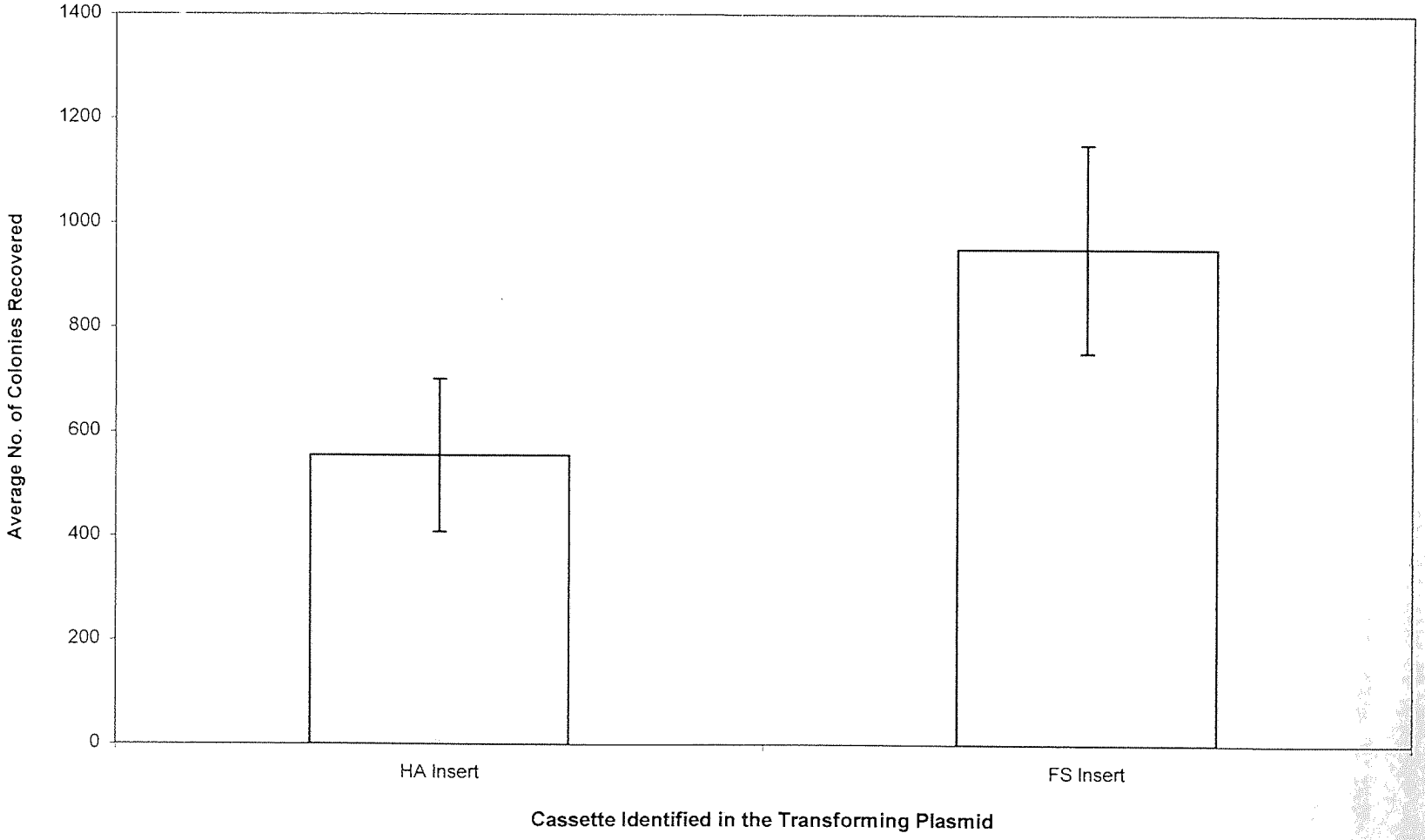


Fig 5.20 Graph demonstrating the average number of recovered transformants after the cassette mutagenesis of the pET-ZFDN1 plasmid with a cassette which generates a high affinity zinc finger gene (HA insert) and a cassette which generates a frameshifted zinc finger gene (FS insert). Results are means and S.E.M. of 8 replicates ($P < 0.05$ for difference in means by paired t test and Wilcoxon signed rank test).

The graph demonstrates that significantly fewer colonies were recovered after transformation with plasmids containing the high affinity insert, although the discrepancy in colony recovery between cells transformed with plasmids containing the high affinity and frameshifted insert, was much smaller than when the same experiment was performed using the pGEX-ZFMA3 construct (section 5.3.1). The similar size of the colonies recovered after transformation of the cells with both plasmids, suggested that the growth of cells was not affected by the presence of the high affinity gene under the control of the T7 promoter, to the extent seen previously. However a degree of selection was still apparent as evidenced by the number of colonies recovered.

5.6.2 Generation of a Simple Model Library Using Solid Media

Ligation reactions (2.8.3) were performed in which equimolar amounts of the high affinity and frameshifted insert were mixed and this mixture ligated into the pre-digested pET-ZFDN1 vector. The ligation reactions were used in the transformation (2.4.2) of *E. coli* DH5α cells and the transformed cells plated on LB media (2.1.2) containing kanamycin.

As the colonies recovered were of equal size (Fig 5.21), the colonies were screened by PCR (2.8.5) to identify the insert contained within each clone. The amplified products were digested (2.8.4) with the enzyme *Sma*I and visualised using agarose gel electrophoresis (2.5.1) to identify colonies containing the frameshifted insert (Fig 5.22).

Figure 5.22 shows that the majority of the PCR products correspond to either the undigested 564 bp product amplified from plasmids containing the HAF insert, or the digested 268 / 279 bp products amplified from plasmids containing the FS insert, which appear as a large single band when resolved on a 2 % agarose gel (Fig 5.22 lane 2). One of the PCR reactions (sample 7, Fig 5.22) failed to generate any product and another reaction (sample 20, Fig 5.22) generated a large non specific product, greater than 1500 bp, in both cases this was expected to result from the addition of excess template during the colony PCR reaction. Two reactions (samples 18 and 40 Fig 5.22)

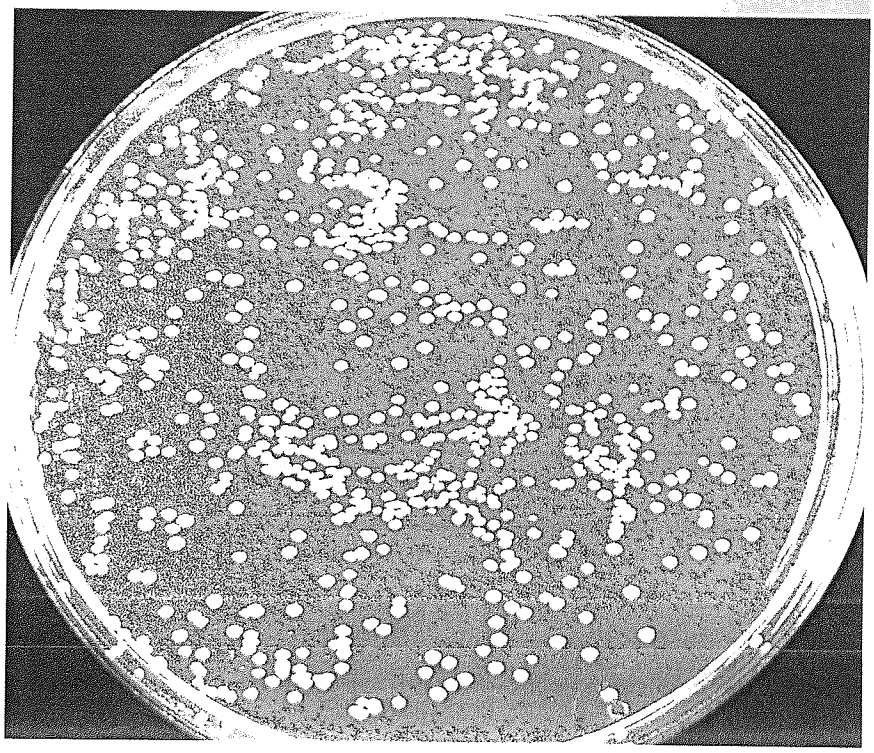


Fig 5.21 Similar sized colonies recovered after the transformation of *E. coli* DH5α cells with the pET-ZFDN1 plasmid containing both the high affinity (HAF) and frameshifted (FS) inserts.

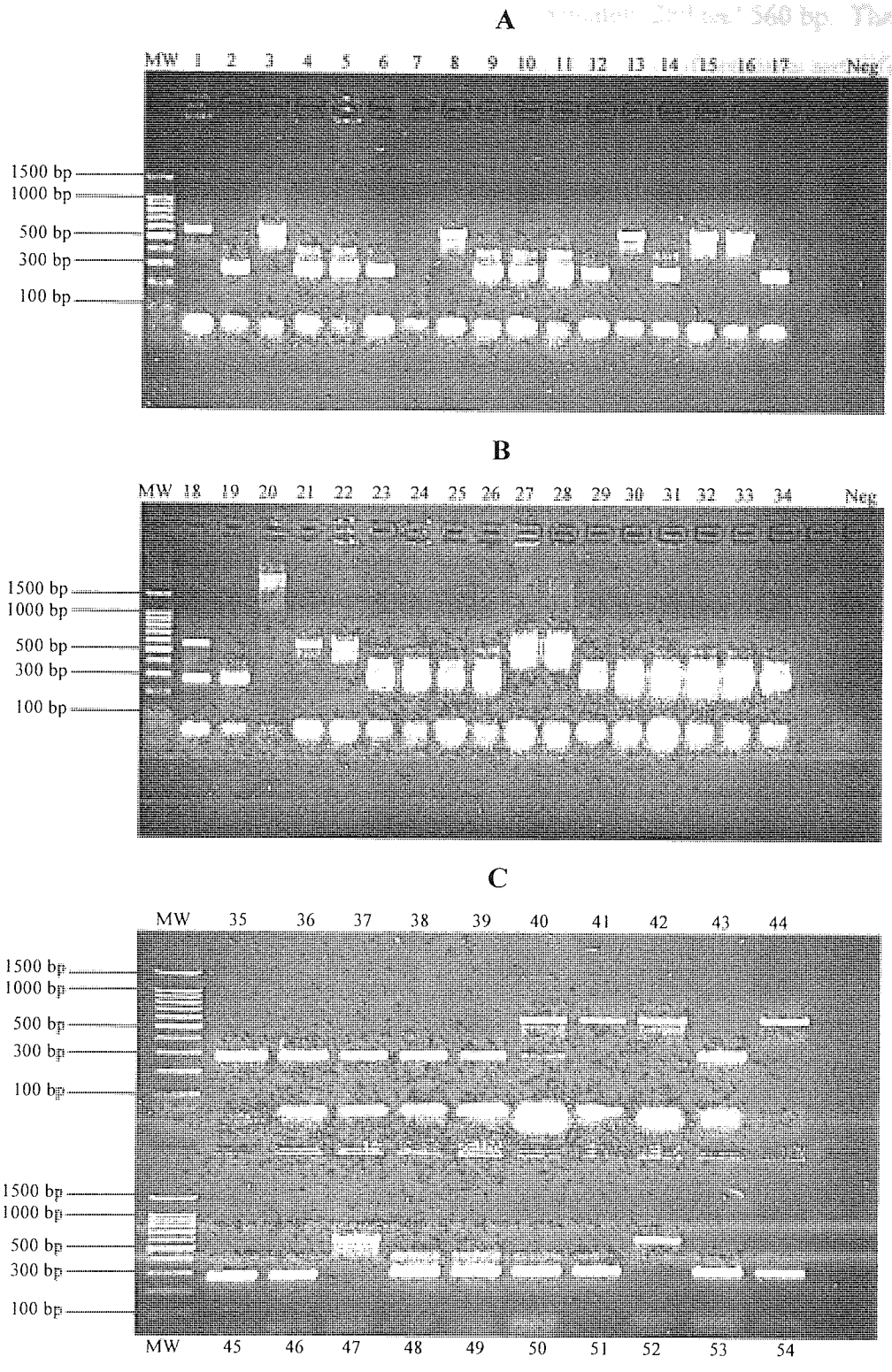


Fig 5.22 Analysis of *Sna*BI digested PCR products, amplified from clones recovered after the transformation of *E. coli* DH5α cells with the pET-ZFDN1 plasmid containing an equimolar mixture of the high affinity and frameshifted inserts (2 % agarose gel). The amplification of plasmids containing the HAF insert would be expected to generate a product of 564 bp. Amplification of plasmids containing the Fs insert and subsequent digestion of the products with *Sma*I would be expected to generate two products of 268 and 279 bp, which was expected to appear as a single band when resolved on a 2 % agarose gel. Key to figure: MW = 100 bp ladder (Promega, Madison); 1 – 55 = Samples 1 – 55; Neg = Negative control.

resulted in the generation of two bands of approximately 280 and 560 bp. The presence of two bands, may have indicated the incomplete digestion of products amplified from plasmids containing the frameshifted insert, resulting in a banding pattern similar to that produced in sample 18 in which the bands demonstrate similar fluorescence. The smaller band in sample 40 shows less intense fluorescence than the larger band, suggesting that it may be the result of product overspill from the previous lane. The possibility that two overlapping colonies may have been amplified in the PCR reaction could also not be discounted. Colonies which generated two bands were discounted in the calculation of the numbers of colonies identified as containing each insert (Fig 5.23).

The figure shows that colonies which contained the high affinity insert accounted for 30 % of the identified clones. This figure is similar to the discrepancy in the recovery observed when generating the high affinity and frameshifted plasmids individually using the pET-ZFDN1 vector (Section 5.6.1). As the recovery of clones in a system where no selection pressure was present would be expected to result in a recovery rate of approximately 50 % for clones containing each insert, the results suggested that a degree of negative selection may still be present when the high affinity zinc finger gene is under the control of a T7 promoter.

5.6.3 Generation of a Simple Model Library Using Liquid Media

The experiments carried out in section 5.6.2 were repeated using the pET-ZFDN1 plasmid, except that cells were cultured overnight in 30ml of liquid media containing kanamycin (2.1.1) prior to plating. After incubation, the cultures were diluted in series and plated onto LB media (2.1.2) containing kanamycin to assess the recovery of clones after the growth of the model library in liquid media.

Colonies recovered from the serially diluted culture were of equal size, all colonies from plates in which the serial dilution of the cells resulted in the generation of less than 30 colonies were screened by PCR (2.8.5) using the pET close forward and pET close reverse primers (Appendix A1). The amplified products were digested (2.8.4) with the enzyme *Sma*I and visualised using agarose gel electrophoresis (2.5.1) to identify the

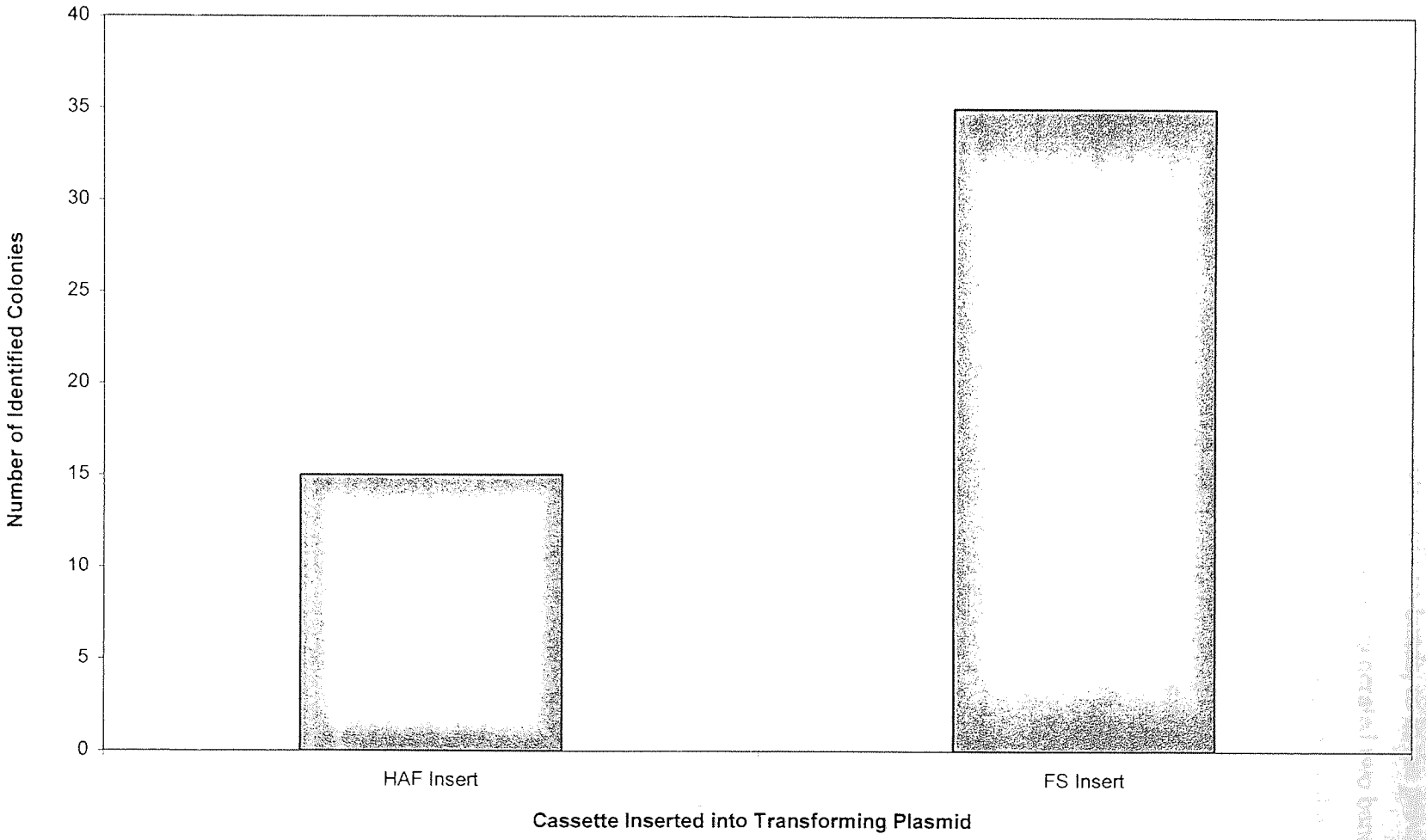
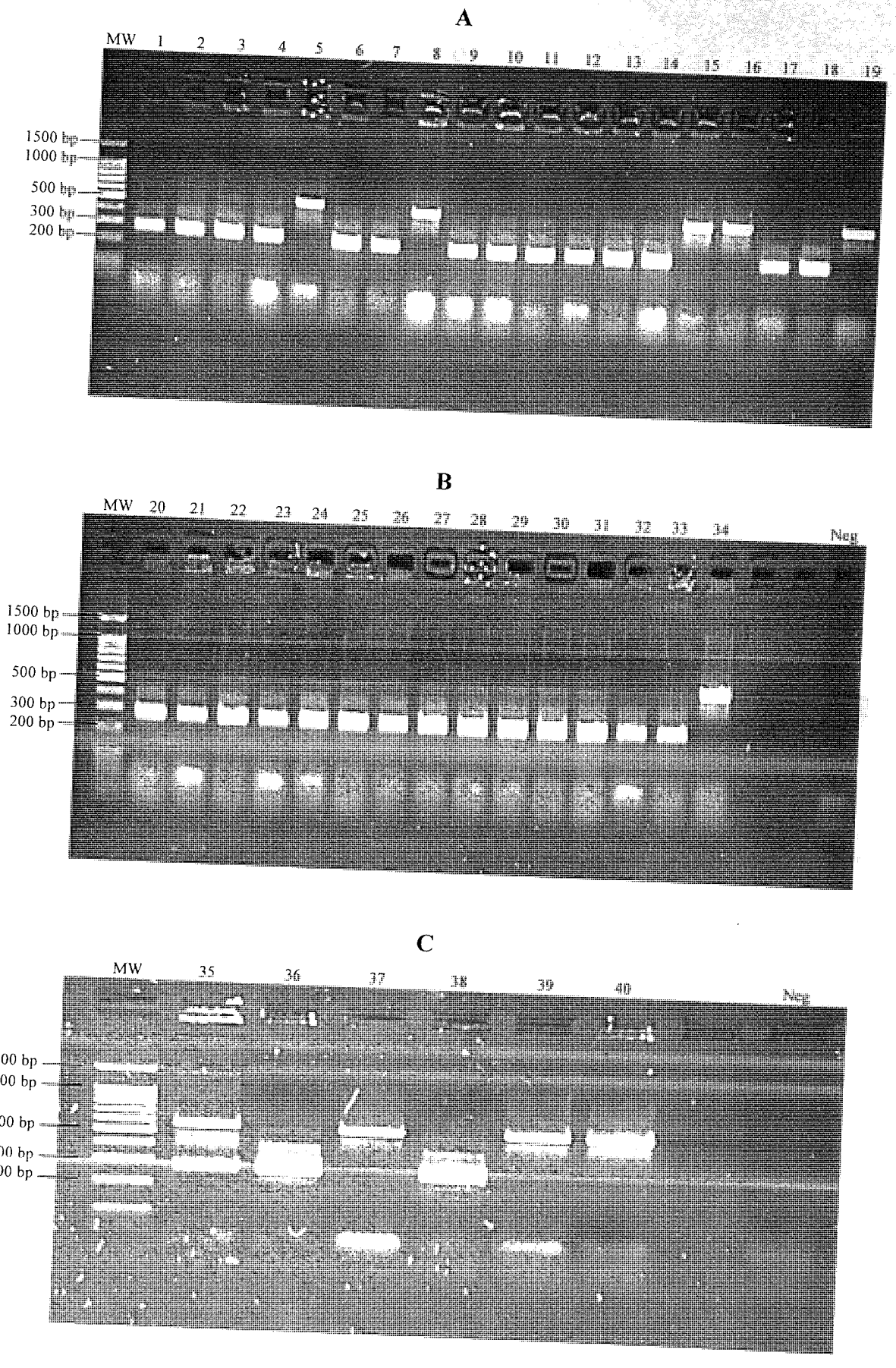


Fig 5.23 Graph demonstrating the number of recovered clones, identified as containing the high affinity (HAF) and frameshifted insert (FS) after the PCR analysis of clones recovered after the transformation of *E. coli* DH5α cells with the pET-ZFDN1 plasmid containing equimolar amounts of each insert.

frameshifted insert (Fig 5.24). Analysis of the recovered products showed that products were amplified in all PCR reactions. In four of the samples (35, 50, 56 and 52) digestion of the products resulted in the generation of two bands, as in the previous experiment (Section 5.6.2). Discounting the colonies which generated two bands, the numbers of colonies identified as containing each insert were calculated and used to plot the graph in Figure 5.25. The figure shows that colonies recovered containing the high affinity insert represent only 25 % of the total number of recovered colonies, again suggesting that some degree of selection still exists despite the use of a T7 based expression system.

The use of a T7 based expression vector was expected to remove any selection pressure within the libraries, caused by basal level expression of zinc finger proteins, which may prove toxic to the host cell. The results of the model library experiments suggested however, that negative selection of plasmids encoding the high affinity zinc finger protein, still occurred under the control of a T7 promoter. It was postulated that other factors may result in the selection of the plasmid encoding the frameshifted zinc finger, such as, increased stability or favourable conformation of the smaller frameshifted insert. However the lowest recovery of the plasmid encoding the high affinity zinc finger was obtained after the growth of transformed cells in liquid culture. As growth of cells in liquid culture will place the greatest selection pressure upon cells containing plasmids encoding toxic proteins (when these proteins are transcribed by the host) this suggested that the negative selection of these plasmids was the result of basal level expression of the protein.


A
MW 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
1500 bp
1000 bp
500 bp
300 bp
200 bp
B
MW 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 Neg
1500 bp
1000 bp
500 bp
300 bp
200 bp
C
MW 35 36 37 38 39 40 Neg
1500 bp
1000 bp
500 bp
300 bp
200 bp

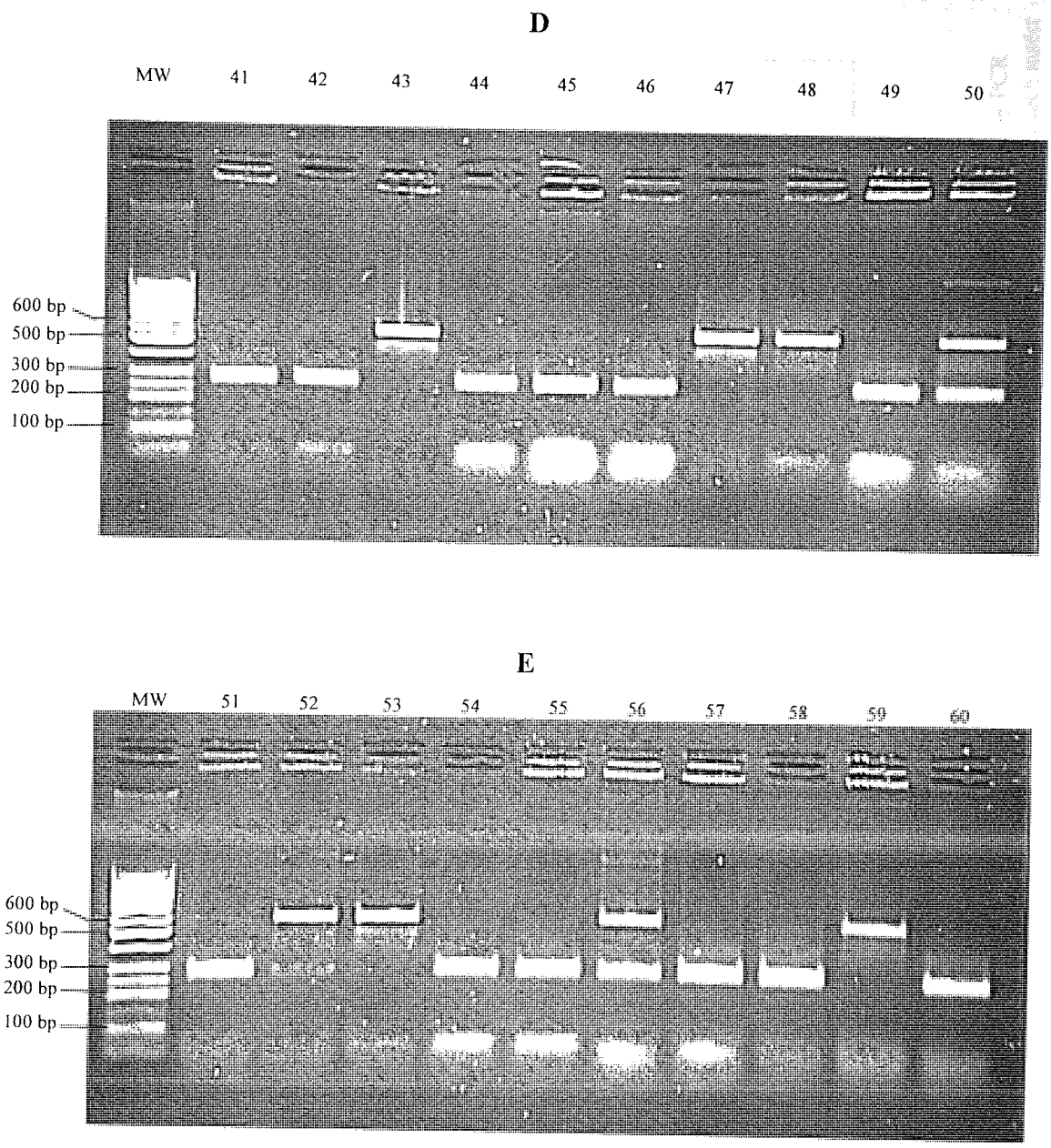


Fig 5.24 Analysis of *Sma*I digested PCR products, amplified from clones recovered after the transformation of *E. coli* DH5α cells with the pET-ZFDN1 plasmid, containing equimolar amounts of the high affinity and frameshifted inserts and the transformed cells grown in liquid media (2 % agarose gel). Amplification of plasmids containing the HAF insert would be expected to produce a single amplicon of 564 bp containing no *Sma*I recognition site. Amplification of plasmids containing the FS insert and subsequent digestion of the product by *Sma*I would be expected to generate two products of 268 and 279 bp, which may appear as a single band when resolved on a 2 % agarose gel. Key to figure: MW = 100 bp ladder (Promega) in gels A – C and GeneRuler 50 bp ladder (MBI Fermentas) in gels D and E; 1 – 60 = Samples 1 – 60; Neg = Negative control.

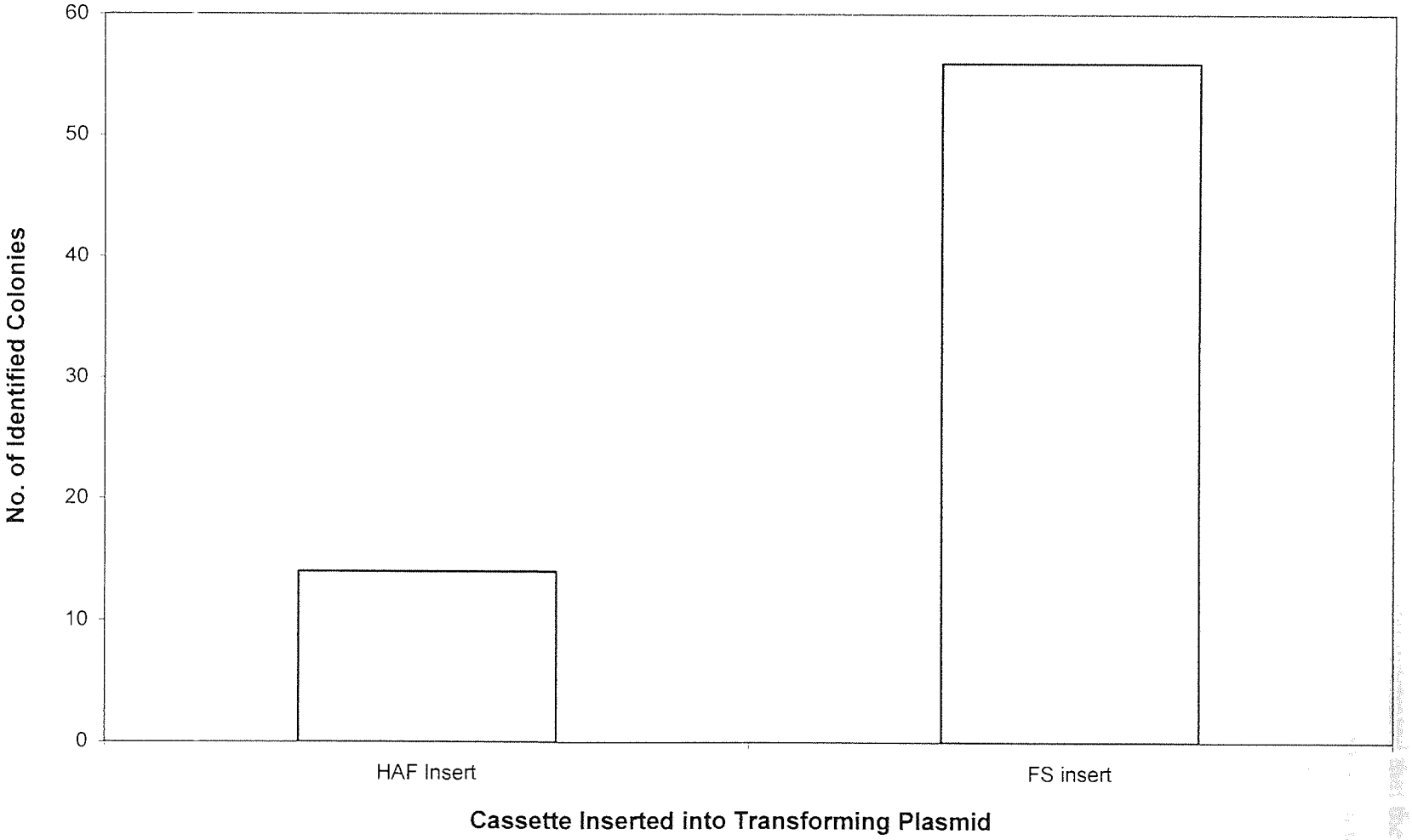


Fig 5.25 Graph demonstrating the number of recovered clones containing the high affinity (HAF) and frameshifted insert (FS), identified by PCR analysis of clones recovered after the transformation of *E. coli* DH5α cells with the pET-ZFDN1 plasmid containing equimolar amounts of each insert and growth of the transformed cells in liquid media.

5.7 Library Construction in the T7-Based Expression Vector

The model library experiments (Sections 5.3 and 5.6) had demonstrated that the selection of plasmids encoding the frameshifted zinc finger was reduced when using a T7-based expression vector. If the recovery of library clones containing frameshift mutations was due to the negative selection of clones encoding high affinity zinc finger proteins, the construction of a MAX library in the T7-based expression vector would be expected to result in the lower recovery of clones containing these mutations. To test this assumption, and to further assess the MAX technique, fully randomised MAX zinc finger libraries were constructed using the pETZFDN1 construct.

The pETZFDN1 plasmid was digested (2.8.4) with *Sna*BI and treated with CIP (2.8.1) prior to digestion with the enzymes *Hin*dIII and *Bsi*WI (2.8.4).

Hybridisations (2.9.4) containing the full complement of 20 MAX codons at each position of randomisation, were carried out in Hybridisation buffer 1 (2.2.5). Prior to hybridisation the β, γ and ENDMAX oligonucleotides were treated with PNK (2.8.2) to phosphorylate the 5' end of each oligonucleotide. The reactions were hybridised (2.9.3) and ATP (2.2.17), DTT (2.2.18) and one Weiss unit of ligase were then added to the hybridisation reactions before incubation at 14°C overnight.

Ligation reactions (2.8.3) were carried out to ligate the pre-ligated MAX cassettes into the pre-digested pET-ZFDN1 vector. These ligation reactions were used in the transformation (2.4.2) of *E. coli* DH5α cells and the transformed cells grown on LB media (2.1.2) containing kanamycin. Plasmid DNA was isolated (2.4.6) from the recovered colonies and sequenced. Sequence alignments of the inserted sequences are shown in Figure 5.26.

Sequence Alignments of Libraries Generated in pET-ZFDN1

1 leu-leu-lys
GGGAAAAGCTTTAGTCTCAGCGACCTGTTACAAAAACATCAGCGTACGCACACCGGGG

2 ile-pro-glu
GGGAAAAGCTTTAGTATTAGCGACCCGTTACAAGAACATCAGCGTACGCACACCGGGG

3 ser-gly-phe
GGGAAAAGCTTTAGTAGCAGCGACGGCTTACAATTTCATCAGCGTACGCACACCGGGG

4 asp-xxx-pro
GGGAAAAGCTTTAGTGATAGCGACATTCTTACAACCCCATCAGCGTACGCACACCGGGG

5 Rearranged
GGGAAAAGCTTTAGTNCGAGCGANNNGTTCACAAACCC-TNCNCNNNCNCCCCCCNGGG

6 met-arg-asn
GGGAAAAGCTTTAGTATGAGCGACAGGTTACAAAACCATCAGCGTACGCACACCGGGG

7 tyr-val-val
GGGAAAAGCTTTAGTTATAGCGACGTGTTACAAGTGCATCAGCGTACGCACACCGGGG

8 glu-met-phe
GGGAAAAGCTTTAGTGAAAGCGACATGTTACAATTTCATCAGCGTACGCACACCGGGG

9 tyr-arg-stop
GGGAAAAGCTTTAGTTATAGCGACCGCTTACCAATAGCATCAGCGTACGCACACCGGGG

10 glu-gly-arg
GGGAAAAGCTTTAGTGAAAGCGACGGTTTACAACGCCATCAGCGTACGCACACCGGGG

11 val-xxx-xxx
GGGAAAAGCTTTAGTGTGAGCGACCGTCTTACAAAGGCCATCAGCGTACGCACACCGGG

12 xxx-xxx-gly Mixed
GGGAAAAGCTTTAGTNNGAGCGACTNGTTACAAGGGCATCAGCGTACGCACACCGGGG

13 ile-ala-phe
GGGAAAAGCTTTAGTATTAGCGACGCGTTACAATTTCATCAGCGTACGCACACCGGGGA

14 his-xxx-xxx
GGGAAAAGCTTTAGTCATAGCGAACCTTATTACAACAGCCATCAGCGTACGCACACCGA

15 asp-thr-xxx
GGGAAAAGCTT-AGTGATAGCGACACTTTACAAGCGCCATCAGCGTACGCACACCGGG

16 pro-thr-ile
GGGAAAAGCTTTAGTCCGAGCGACACCTTACAAATTCATCAGCGTACGCACACCGGGG

17 arg-xxx-thr
GGGAAAAGCTTTAGTAGGAGCGACCAGTTTACAAACACATCAGCGTACGCACACCGGGG

18 thr-glu-phe
GGGAAAAGCTTTAGTACCAGCGACGAATTACAATTTCATCAGCGTACGCACACCGGGG

19 tyr-phe-xxx
GGGAAAAGCTTTAGTTATAGCGACTTTTTACAATAATCATCAGCGTACGCACACCGGGG

20 asn-xxx-his
GGGAAAAGCTTTAGTAACAGCGACCACATTACAACACCATCAGCGTACGCACACCGGGG

21 ile-xxx-leu
GGGAAAAGCTTTAGTATTAGCGACAA-TTACAATTACATCAGCGTACGCACACCGGGG

22 thr-phe-gln
GGGAAAAGCTTTAGTACCAGCGACTTTTTACAACAGCATCAGCGTACGCACACCGGGG

23 thr-thr-xxx
GGGAAAAGCTTTAGTACCAGCGACACATTACAAAGGCCATCAGCGTACGCACACCGGGG

24 tyr-xxx-xxx
GGGAAAAGCTTTAGTTATAGCGACCCAATTAACCCAACCCACATCAGCGTACGCACACC

25 gly-val-ser
GGGAAAAGCTTTAGTGGCAGCGAACGTGTTACAAAGCCATCAGCGTACGCACACCGGGG

26 ala-glu-xxx
GGGAAAAGCTTTAGTGCGAGCGACGAGTTACAACCAGCCATCAGCGTACGCACACCGGG

27 xxx-arg-xxx
GGGAAAAGCTTTAGTGT-AGCGACAGGTTACAAATGCCATCAGCGTACGCACACCGGGG

28 arg-xxx-met
GGGAAAAGCTTTAGTAGGAGCGACCCTTTTTACAAATACATCAGCGTACGCACACCGGG

29 arg-asn-pro
GGGAAAAGCTTTAGTAGGAGCGACAATTTACCAACCACATCAGCGTACGCACACCGGGG

30 asp-ile-val
GGGAAAAGCTTTAGTGATAGCGACATTTTACAAGTGCATCAGCGTACGCACACCGGGG

31 phe-xxx-phe Mixed
GGGAAAAGCTTTAGTTTTAGCGACGNTTTACAATTTCATCAGCGTACGCACACCGGGG

32 gln-pro-pro
GGGAAAGCTTTAGTCAAAGCGACCCGTTACAACCGCATCAGCGTACGCACACCGTGGG

33 his-cys-his
GGGAAAAGCTTTAGTCATAGCGACTGCTTACAACATCATCAGCGTACGCACACCGGGG

34 lys-xxx-met
GGGAAAAGCTTTAGTAAAAGCGACTAATTTACAAATGCATCAGCGTACGCACACCGGGG

35 ser-ser-asn
GGGAAAAGCTTTAGTAGCAGCGACAGCTTACAAAACCATCAGCGTACGCACACCGGGG

36 leu-pro-thr
GGGAAAAGCTTTAGTCTGAGCGACCCATTACCAAACTCATCCAGCGTACGCACACCGGG

37 his-pro-tyr
GGGAAAAGCTTTAGTCACAGCGACCCGTTACAATATCATCAGCGTACGCACACCGGGG

38 lys-gln-val
GGGAAAAGCTTTAGTAAAAGCGACCAGTTACAAGTGCATCAGCGTACGCACACCGGGG

39 arg-glu-lys
GGGAAAAGCTTTAGTAGGAGCGACGAATTACAAAAACATCAGCGTACGCACACCGGGG

40 arg-gly-arg
GGGAAAAGCTTTAGTCGCAGCGACGGCTTACAACGCCATCAGCGTACGCACACCGGGG

41 trp-val-trp
GGGAAAAGCTTTAGTTGGAGCGACGTTTTACAATGGCATCAGCGTACGCACACCGGGG

42 lys-xxx-met
GGGAAAAGCTTTAGTAAAAGCGACCCCCTTACCAAATGCATCAGCGTACGCACACCGGG

43 lys-xxx-xxx
GGGAAAAGCTTTAGTAAAAGCGACGACTTTACAATTACCATCAGCGTACGCACACCGGG

44 thr-asn-arg
GGGAAAAGCTTTAGTACCAGCGACAACTTACAACGTCATCAGCGTACGCACACCGGGG

45 thr-xxx-val
CGGGAAAAGCTTTAGTACCAGCGACCTAGTTACAAGTTCATCAGCGTACGCACACCGGG

46 ile-pro-cys
GGGAAAAGCTTTAGTATTAGCGACCCCTTACCAATGCCATCAGCGTACGCACACCGGGG

47 pro-pro-cys
GGGAAAAGCTTTAGTCCGAGCGACCCGTTACAATGCCATCAGCGTACGCACACCGGGG

48 glu-stop-val
GGGAAAAGCTTTAGTGAAAGCGACTAGTTACAAGTCCATCAGCGTACGCACACCGGGG

49 leu-ser-ile
GGGAAAAGCTTTAGTCTGAGCGACAGCTTACAAATTCATCAGCGTACGCACACCGGGG

50 met-xxx-ser
GGGAAAAGCTTTAGTATGAGCGACCCACTTACAAAGTCATCAGCGTACGCACACCGGGG

51 his-xxx-xxx
GGGAAAAGCTTTAGTCATAGCGACCCATGTTACAACCTCCATCCCAGCGTACGCACACC

52 his-glu-gly
GGGAAAAGCTTTAGTCATAGCGACGAATTACCAAGGGCATCAGCGTACGCACACCGGGG

53 asn-arg-trp
GGGAAAAGCTTTAGTAATAGCGACAGGTTACAATGGCATCAGCGTACGCACACCGGGG

54 pro-phe-met
GGGAAAAGCTTTAGTCCGAGCGACTTTTTACAAATGCATCAGCGTACGCACACCGGGG

55 phe-tyr-trp
GGGAAAAGCTTTAGTTTTAGCGACTACTTACAATGGCATCAGCGTACGCACACCGGGG

56 glu-xxx-gly
GGGAAAAGCTTTAGTGAAAGCGACTCAATTACAAGGACATCCAGCGTACGCACACCGGG

57 met-cys-trp
GGGAAAAGCTTTAGTATGAGCGACTGCTTACAATGGCATCAGCGTACGCACACCGGGG

58 xxx-met-gly
GGGAAAAGCTTT--TCCGAGCGACATGTTACAAGGACATCAGCGTACGCACACCGGGG

59 arg-his-asn
GGGAAAAGCTTTAGTAGGAGCGACCATTTACAAAACCATCAGCGTACGCACACCGGGG

F8 Self Ligation
GGGAAAAGCTTCGTGTACGTACTGACGTACGCACACCGGGGAAAA

61 val-ile-thr
GGGAAAAGCTTTAGTGTGAGCGACATTTTACAAACCCATCAGCGTACGCACACCGGGG

62 asn-tyr-ser
GGGAAAAGCTTTAGTAACAGCGACTATTTACAAAGCCATCNGCGTACGCAACCGGGGG

63 asn-asn-thr
GGGAAAAGCTTTAGTAACAGCGACAACTTACAAACGCATCAGCGTACGCACACCGGGG

64 asp-xxx-ile
GGGAAAAGCTTTAGTGATAGCGACCCGTTTACAAATACATCAGCGTACGCACACCGGGG

65 his-asp-xxx
CAGGGAAAAGCTTTAGTCATAGCGACGACTTACAATCGGCATCAGCGTACGCACACCGG

66 ile-ser-arg
GGGAAAAGCTTTAGTATTAGCGACAGTTTACAACGCCATCAGCGTACGCACACCGGGG

67 gln-xxx-xxx Mixed
GGGAAAAGCTTTAGTCAGAGCGACNCGTTACAAACNCATCAGCGTACGCACACCGGGG

68 his-xxx-glu
GGGAAAAGCTTTAGTCATAGCGACGGCATTACAAGAACATCAGCGTACGCACACCGGGG

69 pro-xxx-tyr
GGGAAAAGCTTTAGTCCGAGCGACCTCATTACCAATACCATCAGCGTACGCACACCGGG

70 asp-gln-asp
GGGAAAAGCTTTAGTGATAGCGACCAGTTACAAGACCATCAGCGTACGCACACCGGGGA

71 asn-phe-xxx
GGGAAAAGCTTTAGTAACAGCGACTTTTTACAACCCTCATCAGCGTACGCACACCGGGG

G8 SELF LIGATION
GGGAAAAGCTTCGTGTACGTACTGACGTACGCACACCGGGGAAAA

G9 SELF LIGATION
GGGAAAAGCTTCGTGTACGTACTGACGTACGCACACCGGGGAAAA

74 xxx-gln-xxx
GGGAAAAGCTTTA-GGCAGCGACCAATTACA-TTTCATCAGCGTACGCACACCGGGG

75 thr-thr-glu
GGGAAAAGCTTTAGTACCAGCGACACGTTACAAGAACATCAGCGTACGCACACCGGGG

76 pro-thr-xxx
GGGAAAAGCTTTAGTCCGAGCGACACGTTACAAACTCCATCAGCGTACGCACACCGGGG

77 asp-lys-phe
GGGAAAAGCTTTAGTGACAGCGACAAGTTACAATTCCATCAGCGTACGCACACCGGGG

78 cys-pro-thr
GGGAAAAGCTTTAGTTGCAGCGACCCGTTACAAACGCATCAGCGTACGCACACCGGGG

Fig 5.26 Sequence data obtained from clones recovered after the cassettes mutagenesis of pET-ZFDN1 with pre-ligated MAX cassettes generated in hybridisation buffer 1. The amino acids encoded at the randomised positions of each clone are also shown. Key to figure: Purple text = MAX codon sequence: Red text = Non MAX codon sequence. All bases in blue text represent sequence abnormalities. Inserted bases are denoted by the inserted base highlighted in blue text. Deletions are denoted by a blue dash (-). Point mutations are highlighted in bold text. Rearranged sequences and mixed sequences are noted in the figure. Randomised positions in which the identity of the encoded amino acid could not be accurately deduced from the sequence data due to the insertion/deletion of bases or the inclusion of an N represented base are represented by a blue xxx.

An initial survey of the sequence results is contained in Table 5.4, showing the numbers of clones containing all MAX codons at positions of randomisation, non MAX codons at these positions and the number of clones containing frameshift mutations within the sequence.

Identity of Sequenced Clones	No. of Sequences	% of Total Sequences
Correct Sequence Containing MAX codons at all randomised Positions	19	25.3
Correct Sequence containing non MAX codons at one or more randomised positions	18	24
Sequences Containing Frameshift Mutations	35	46.7
Sequences Containing Point Mutations	0	0
Sequences Containing Mixed Codons	3	4
Self Ligations	3	See Legend
Total No. of sequences	78	100

Table 5.4 Summary of results obtained when sequencing clones recovered after the MAX randomisation of the pET-ZFDN1 plasmid. Clones resulting from the religation of the parental plasmid were omitted from the percentage of total sequences calculation, the religation of parental plasmids is common in cloning reactions and it was expected that the properties of the inserted cassette would not influence the occurrence of self ligated clones.

The initial survey of the results showed that only 25.3 % of the recovered clones contained DNA inserts with no frameshift mutations and MAX codons present at all positions of randomisation. Clones which contained no frameshift mutations but possessed non-MAX codons at one or more positions of randomisation represented 24 % of the total number of sequenced clones. In comparison to the results obtained in the MAX mutagenesis of the pGEX-ZFMA3 plasmid (which employed the same hybridisation conditions) this represented an 8 % decrease in the number of clones which contained MAX codons at all positions of randomisation and a 12.9 % increase in the number of clones containing non MAX codons, suggesting a decrease in the specificity of the hybridisation reaction used to generate the MAX cassettes. The results also showed that clones containing frameshift mutations still predominated within the generated library accounting for 46.7 % of the total number of sequenced clones.

Frameshifted clones accounted for 52.8 % of the total number of sequenced clones when the libraries were generated the pGEX-ZFMA3 plasmid (Section 4.9). The reduction in the number of frameshifted clones recovered when the libraries were cloned in the T7-based expression vector, did not reflect the decrease in the selection of frameshifted clones observed in the model library experiments (Section 5.6). This suggested that the recovery of these clones did not result from the negative selection of high affinity zinc finger proteins as a result of basal level expression.

5.7.1 Analysis of Clones

The sequences of the clones, excluding those containing frameshift mutations, were examined, to assess the inclusion of MAX codons at each randomisation position. The graph in Figure 5.27 illustrates the numbers of correct MAX codons at each position of randomisation in the recovered sequences. As in the libraries constructed in the pGEX-ZFMA3 plasmid (Section 4.10) the inclusion of MAX codons was similar at each position of randomisation (approximately 80.2%). In comparison to the library constructed previously using hybridisation buffer 1, the inclusion of MAX codons at each position was slightly decreased, suggesting a decrease in the efficiency of the hybridisation reaction, used to generate the MAX cassettes.

The amino acid representation of the generated library was calculated at all positions of randomisation (Fig 5.28). The amino acid representation within the library was reasonable, with all amino acids represented at least once on the generated library. The results showed that none of the amino acids predominated to any great extent, when representation at all randomised positions was considered, although the alanine encoding MAX codon was only represented once within the sequenced clones. The amino acid representation at each position of randomisation was calculated and is shown in Figure 5.29. The graph illustrates that again the representation of the library was reasonable when considering the small sample size. The predominance of certain amino acids at different randomisation positions, such as the inclusion of a larger number of proline codons in the beta position and a larger number of tryptophan and phenylalanine codons at the gamma position, was expected to reflect the variance of the sampled population of clones rather than an active selection of these amino acids at these positions.

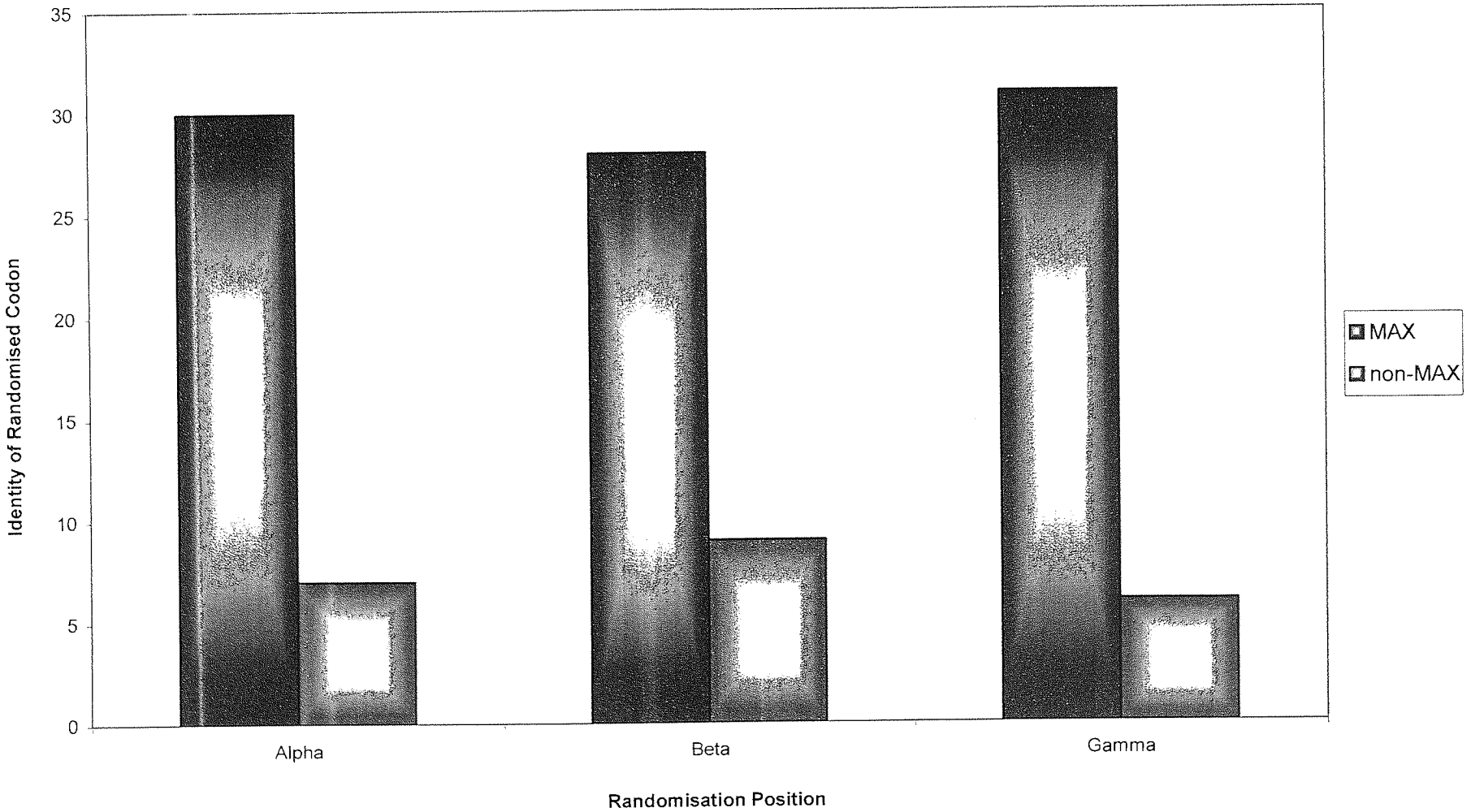


Fig 5.27 Graph demonstrating the identities of the codons present at the randomised positions in the intact DNA sequences recovered from the cassette mutagenesis of the pET-ZFDN1 plasmid with MAX cassettes generated in hybridisation buffer 1.

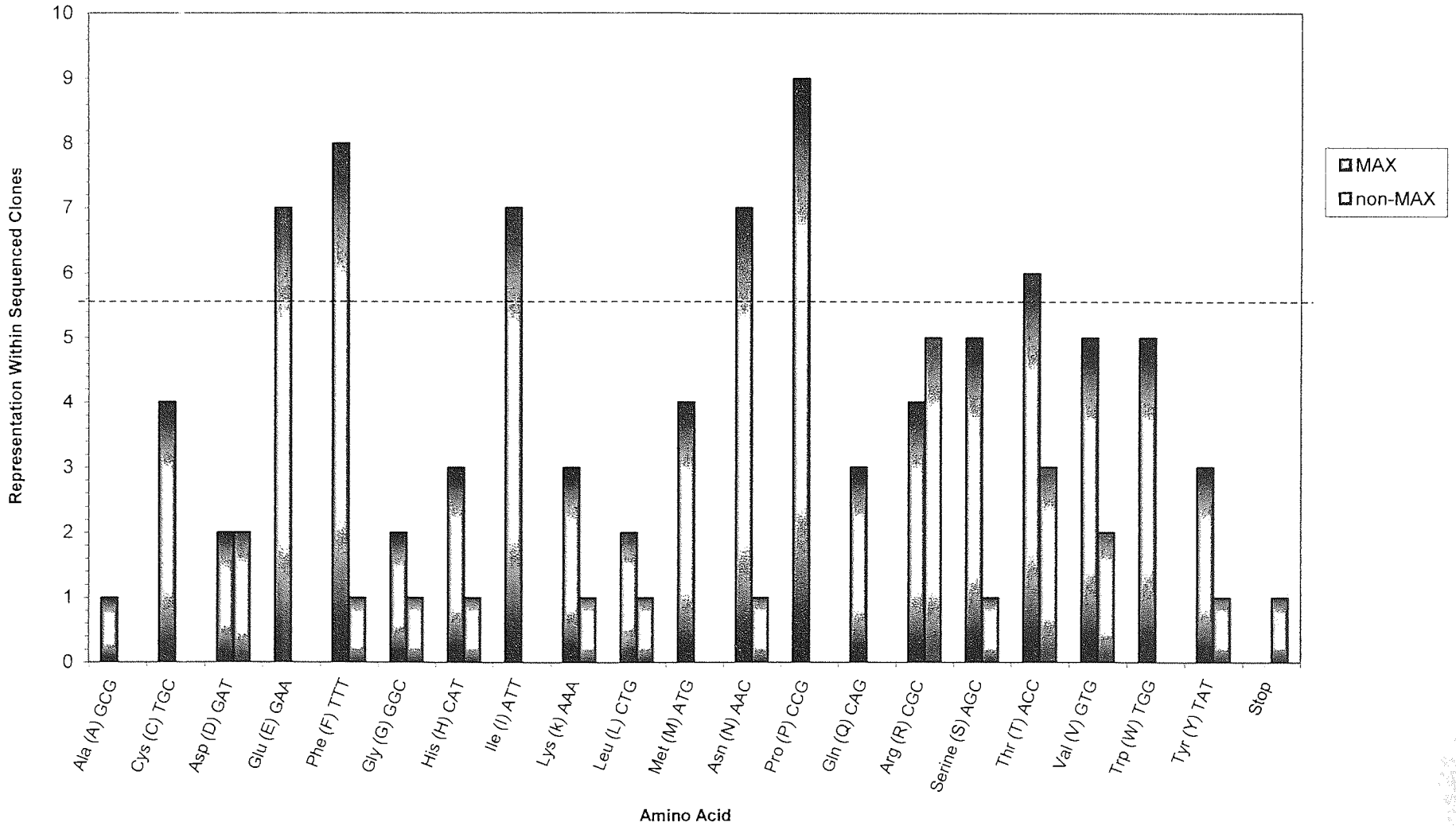


Fig 5.28 Graph demonstrating the amino acid representation at all randomised positions within the sequenced clones, recovered from the cassette mutagenesis of the pET-ZFDN1 plasmid with MAX randomised cassettes generated in hybridisation buffer 1. The blue line represents the theoretical ideal distribution of MAX codons in the correct sequences, based upon 100 % efficiency of the technique and the recovery of no non-MAX codons.

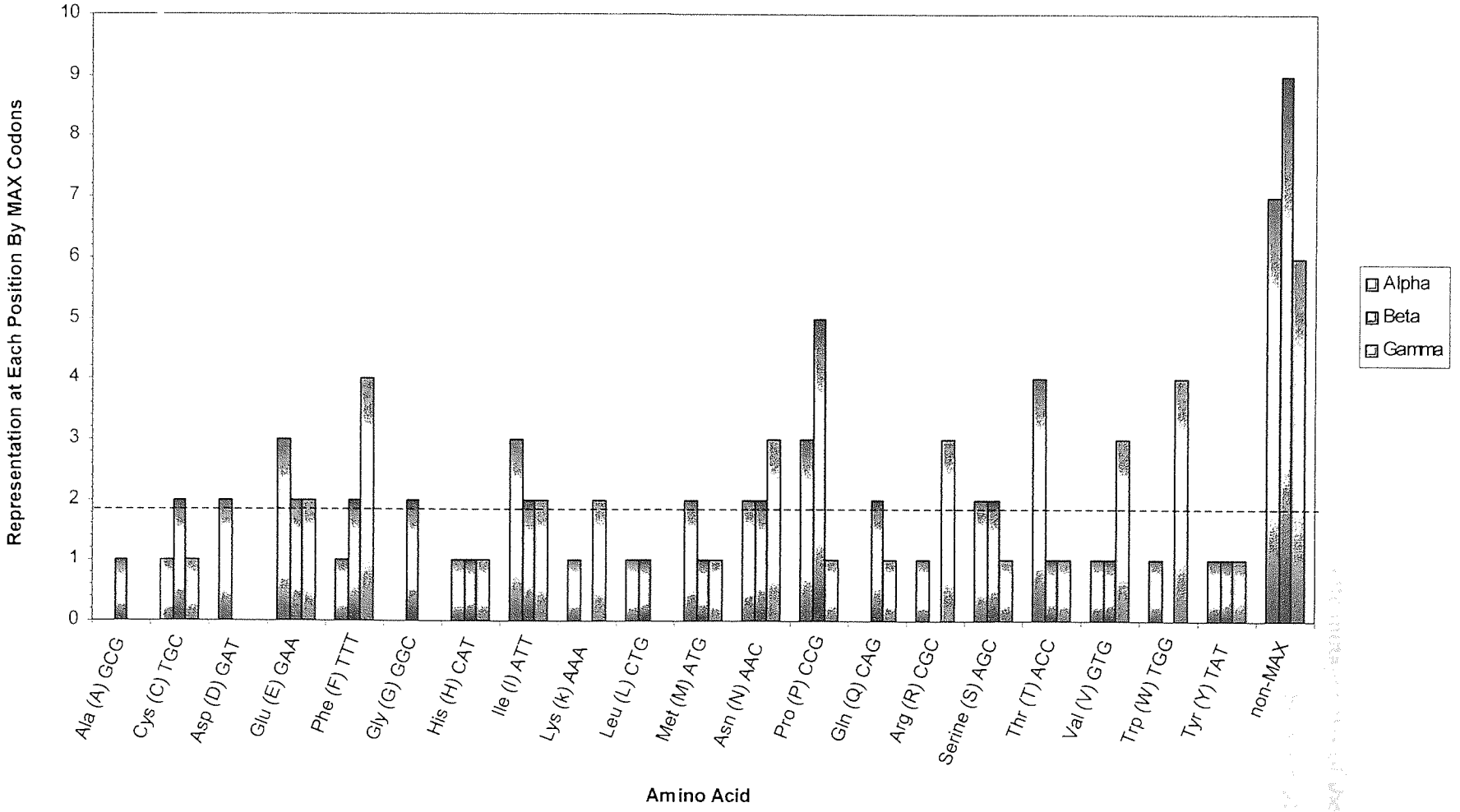


Fig 5.29 Graph demonstrating the amino acid representation by MAX codons at each randomised position, in the sequences of clones recovered from the cassette mutagenesis of the pET-ZFDN1 plasmid with MAX randomised cassettes generated in hybridisation buffer 1. Representation at each position by non-MAX codons is included for comparison. The blue line represents the theoretical ideal distribution of MAX codons in the correct sequences based upon 100 % efficiency of the technique and the recovery of no non-MAX codons.

5.8 Conclusions

The generation of the simple model libraries by the cassette mutagenesis of the pGEX-ZFMA3 vector (Section 5.3) highlighted that negative selection of the plasmid encoding the high affinity zinc finger was occurring when this gene was placed under the control of a promoter sequence which could be recognised by the hosts transcriptional machinery. This not only raised concerns regarding the possible negative selection of high affinity zinc fingers within libraries generated using this vector, but also suggested a mechanism by which frameshifted clones may be selected for within the libraries. The subcloning of the zinc finger gene into the pET-42a plasmid, under the control of a T7 promoter, was expected to prevent the negative selection of plasmids encoding high affinity zinc fingers, resulting from the basal level expression of these proteins. However, repetition of the model library experiments using the pET-ZFDN1 plasmid (Section 5.6) highlighted that negative selection of plasmids encoding the high affinity zinc finger protein, although reduced, was not completely abolished by the use of a T7 based expression vector. This was expected to be the result of a residual basal level expression of these plasmids, as the greatest degree of selection occurred when the model libraries were grown in liquid media. Mechanisms by which this basal level expression may have occurred are discussed in Chapter 6.

As the use of the T7 based expression vector reduced the selection of frameshifted clones in the model library experiments, then it would be expected that a similar reduction in the recovery of such clones would be evident within libraries generated in this vector, assuming that the recovery of these clones was simply the result of negative selection of high affinity clones from within the generated libraries. The analysis of the library results showed only a slight decrease in the recovery of frameshifted clones, suggesting initially that the high incidence of frameshifted clones within the library, was not the result of a simple selection process but was perhaps inherent in the current MAX randomisation technique. Although the use of a T7 based expression system did not significantly reduce the number of frameshifted clones recovered in the generated libraries, the model library experiments, suggested that negative selection of plasmids encoding high affinity zinc fingers does occur within the libraries, and that this problem can be addressed by controlling the basal level expression of these proteins. The model library experiments highlighted the importance of the use of a T7 based expression system in the construction

of the MAX randomised zinc finger libraries, and suggested that perhaps even greater stringency over the control of expression of the inserted gene was required.

Analysis of the library sequences generated in the pET-ZFDN1 vector (Section 5.71) highlighted the inclusion of a greater number of non-MAX codons at each position than that observed in previous library construction (Section 4.6), suggesting a decreased stringency of the hybridisation of the selection oligonucleotides. Despite this, the representation of amino acids within the generated library was reasonable, with no omission or striking predominance of any codon, suggesting that the MAX methodology could select each of the twenty amino acids at the randomised positions.

The recovery of a large number of frameshifted clones despite the use of a T7 based expression system, suggested that the generation of these clones was resulting from an inherent property of the MAX randomisation technique, and that the use of T7 based expression system in which expression is more stringently controlled, may not significantly reduce the recovery of these clones. Mechanisms by which these clones may be generated, and possible means by which the recovery of these clones may be prevented are discussed in Chapter 6.

Chapter 6 Discussion of Results

6.1 Introduction

The initial aim of this study was to address the problems associated with the creation of representative randomised gene libraries, by the development and implementation of the MAX randomisation technique. As a model the MAX randomisation technique was used to create randomised gene libraries encoding Cys_2-His_2 zinc finger proteins. In addition, other factors which may affect the representation of randomised gene libraries were also addressed. As the development of the technique was accomplished in discrete stages, each stage of development is discussed individually with suggestions for and notation of, the further development of the MAX randomisation procedure.

6.2 Gene Assembly for Library Construction

The mutagenesis of the pGEX-ZFH plasmid to create the pGEX-ZFHM6 plasmid was successfully completed. This provided a gene in which the libraries could be constructed. The pGEX-ZFHM6 plasmid encodes the QDR-RER-RHR triplet zinc finger protein (Desjarlais and Berg, 1993), with the restriction sites necessary to allow cassette mutagenesis of the base contacting residues of the middle finger of the protein.

Over-representation of the QDR-RER-RHR protein within the libraries, which may result from the religation of the native pGEX-ZFHM6 plasmid in the cassette mutagenesis stage of library construction, was addressed by the redesign of the pGEX-ZFHM6 plasmid to create the pGEX-ZFMA3 plasmid. This plasmid was designed to reduce the number of self ligated plasmids recovered after the cassette mutagenesis stage of library construction, by enabling the pre-digestion and phosphatase treatment of the plasmid, prior to digestion with the restriction enzymes employed in the cassette mutagenesis (described in section 3.3.1). In addition, and perhaps more importantly the pGEX-ZFMA3 gene encodes a frameshifted version of the ZFHM6 gene. As a functional zinc finger protein can therefore only be encoded after the successful cassette mutagenesis of the pGEX-ZFMA3 plasmid, parental plasmids recovered in the libraries

as a result of self ligation, would not be expected to affect the subsequent deconvolution of these libraries.

6.3 Control of Selection Pressure Within the Generated Libraries

The creation of simple model libraries, highlighted that the recovery of clones containing plasmids encoding the QDR-RER-RHR protein was almost an order of magnitude lower than the recovery of plasmids encoding a non functional, frameshifted zinc finger protein (Section 5.32 and 5.33). This suggested that clones containing the high affinity zinc finger plasmid were being negatively selected from within the library population, presumably as a result of basal level expression of the zinc finger protein. Potential binding sites for the QDR-RER-RHR protein were identified within the open reading frame of several metabolic and regulatory genes within the *E. coli* genome, which suggested possible mechanisms by which the basal level expression of this zinc finger protein may exert toxic effects upon the host cell.

The possible toxicity of these zinc finger proteins to the host cells has profound implications on the construction of the randomised zinc finger libraries. If the binding of target sites by high affinity zinc finger proteins proves toxic to the host cell, then plasmids encoding these high affinity zinc finger proteins may be lost from the library population, leading to the under-representation of these plasmids within the library. The negative selection of highly interacting zinc fingers from within the library population and the overrepresentation of low affinity clones, would be expected to adversely effect the screening process, as key clones will have been lost from the library.

The potential loss of plasmids encoding potential high affinity zinc finger proteins within the libraries was addressed by the use of a T7 based expression vector (pET-42a), to prevent basal level expression of the encoded proteins. The creation of model libraries by the cassette mutagenesis of the pET-ZFDN1 plasmid (section 5.6) resulted in an increased recovery of plasmids encoding the high affinity zinc finger, suggesting that the negative selection of these clones in the pGEX-ZFMA3 vector had resulted from the toxicity of basally expressed proteins. However the negative selection of these

plasmids was not completely abolished by the use of the pET-42a vector. The worst recovery of high affinity plasmids was obtained when model libraries were grown in liquid media (Section 5.6.3), suggesting that the use of the T7 based expression system did not completely prevent basal level expression of the zinc finger protein. This basal level expression was unexpected, since no T7 RNA polymerase was present in the host DH5α strain used in the experimentation. It was postulated that the basal level transcription of the encoded protein, might have resulted from the "read-through" of termination signals by transcription initiated elsewhere in the plasmid. Basal level transcription in the absence of T7 RNA polymerase has been reported (Studier & Moffat, 1985, Davanloo *et al.*, 1984) with the read-through of termination signals, or initiation at relatively weak promoters within the plasmid, postulated as mechanisms by which this expression may occur.

The results of the model library experiments suggested that the construction of the zinc finger libraries in the pET-ZFDN1 vector, may still result in the loss of plasmids encoding high affinity zinc finger proteins and that basal level expression of these proteins must be further reduced in order to construct a fully representative library.

Development work using T7 based expression systems for the construction of the MAX zinc finger libraries has been continued in the laboratory, following the completion of the experimental work in this study (Dr. Z. Zhang, pers. comm.). The basal level expression of the encoded proteins has been further reduced by the use of the dual replicon pETcoco expression system (Novagen). The pETcoco plasmid can be maintained in a single copy number state within transformed cells (prior to the induction of protein expression) to minimise basal level expression of encoded proteins. Prior to the induction of protein expression, the plasmid copy number can be amplified by induction of the (oriV) medium copy number origin of replication (Novagen Technical information, WWW.Novagen.Com).

6.4 Development of the MAX Randomisation Technique.

Initial randomisation was carried out with MAX randomised cassettes generated by the hybridisation of three selection oligonucleotides and a template oligonucleotide (Section 4.3). Sequence results obtained from this library showed that the inclusion of MAX codons was greatest at the gamma position of randomisation suggesting that the design of the selection oligonucleotides was critical to the inclusion of MAX codons at positions of randomisation.

The selection oligonucleotides were redesigned, relocating the MAX randomisation position to the 3' end of the oligonucleotide, to prevent any flanking sequence stabilising mismatched base pairing between the randomised position and the template strand. Library construction with the redesigned oligonucleotides (Sections 4.6 and 5.7) showed that the discrepancy in the inclusion of MAX codons at the individual positions of randomisation had been successfully addressed by this redesign.

Sequence analysis of the libraries constructed with the redesigned oligonucleotides was also encouraging with regard to the amino acid representation of the generated libraries. In most cases each individual amino acid was encoded at least once when all positions of randomisation were taken into account. Although some codons predominated in each generated library, the identity of these codons varied between the libraries, suggesting that this predominance was the result of experimental variance as opposed to any direct selection of these amino acids.

Sequence analysis of intact clones recovered after libraries were generated in the pGEX-ZFMA3 and pET-ZFDN1 plasmids, showed clones which included non-MAX codons at one or more positions of randomisation, accounted for 11.1 % and 24 % of the total number of clones recovered in each respective library. The generation of non-MAX codons at the randomised positions was expected to result from the repair of the inserted DNA by the host cell, as no complementary double stranded DNA bearing non-MAX codons can be generated during the hybridisation of the mutagenic cassettes.

The generation of non-MAX codons by repair of the inserted DNA by the host cell can be expected if the repair is based upon the sequence of the template strand. Mismatch

repair in *E. coli* utilises the methylation of DNA to ensure the mismatched base is correctly replaced. The repair of mismatched base pairs in unmethylated DNA has been shown to occur on either strand (Modrich, 1989, Dohet *et al.*, 1984, Lyons & Schendel, 1984). As the inserted MAX cassette is unmethylated, then the correction of mismatched bases, prior to replication and methylation, could be dictated by the base present on the template strand. The generation of non-MAX codons may also have resulted from the nick translation of the inserted DNA, by DNA polymerase I present in the host cell. Polymerase activity initiated in the breaks in the phosphodiester backbone between un-ligated selection oligonucleotides, could result in the generation of non-MAX codons, as the conventionally randomised template strand would act as template DNA in this reaction.

Sequence analysis of the libraries generated with the redesigned oligonucleotides highlighted that approximately 50 % of the recovered clones contained frameshift mutations within the inserted sequence. As the recovery of such clones was not diminished by the use of a T7 based expression vector, this suggested that the selection of such clones was inherent within the randomisation technique. It was postulated that the generation of these clones may be due to the incomplete or incorrect repair of the inserted DNA by the host cell. Frameshift mutations are routinely observed when cloning synthetic DNA and accounted for approximately 5 % of recovered clones when the zinc finger libraries were generated using conventionally randomised DNA cassettes (Dr. M. Hughes, pers. comm.). The increased recovery of these clones when libraries are generated using the MAX randomised cassettes would therefore suggest that certain properties of the randomised cassettes may stimulate DNA repair by the host.

The stimulus for such a postulated increase is difficult to ascertain as DNA repair in *E. coli* may involve a variety of discrete and complex systems. Mismatch repair systems in *E. coli* are utilised by the host cell to eliminate errors, such as base pair mismatches, from newly synthesised strands of DNA (Jiricny, 2000). It was expected that the presence of DNA cassettes containing mismatched base pairs may stimulate mismatch repair by the host cell. However although the stimulation of one or more of the pathways involved in the repair of mismatched bases, may have resulted in the generation of non-MAX codons as described earlier, it appears unlikely that mismatch

repair would result in the generation of frameshifted clones, as inserted and deleted bases are also substrates for this repair pathway (Yang, 2000).

Frameshift mutations resulting from DNA repair would therefore suggest the involvement of a lower fidelity repair pathway, such as those initiated in the SOS response of *E. coli* to high levels of DNA damage. The presence of single stranded DNA molecules within *E. coli* has been demonstrated to induce the SOS response (Higashitani *et al.*,1992), which suggested that the presence of single stranded template DNA may have initiated this response in the transformed cells. However binding of single stranded DNA by the *RecA* gene product of *E. coli* is implicated in the induction of the SOS response (Sassanfar & Roberts, 1990, Bhattacharya & Beck, 2002). The *E.* coli DH5α cells used in the construction of the libraries are *RecA*1 deletion mutants, suggesting that the induction of SOS repair by single stranded DNA was unlikely.

The inclusion within *E. coli* cells of multiple copies of msDNA retro-elements, which contain mismatched base pairs, has been linked to an increase in specific frameshift mutations within these cells (Maas *et al.*, 1994 and Mao *et al.*, 1996). The inclusion of retro-elements in which the mismatched base pairs were removed however, did not result in an increase in mutation. This led to the postulation that the mismatched base pairs in the msDNA may sequester a cellular mismatch repair system, which subsequently results in an increase of frameshift mutations as the cells utilise lower fidelity repair pathways (Mao *et al.*, 1996). It seems unlikely however that a similar response would be elicited by plasmids containing MAX cassettes, even with cassettes containing a number of mismatched base pairs, as these plasmids would be expected to be present at only single copy levels after transformation. Replication of these plasmids without any repair would be expected to generate discrete populations of fully complementary plasmids (See Fig. 4.5, section 4.3), unlike the replication of msDNA in which mismatched base pairs are maintained by the defined secondary structure of these moieties.

DNA polymerases including those of *E. coli* have been implicated in the introduction of frameshift mutations in transcribed DNA, in both *in vivo* and *in vitro* experimentation. (Harris *et al.*, 1997, Pham *et al.*, 1998, Bebenek & Kunkel, 1990, Kunkel, 1990). Perhaps then the most plausible explanation for the generation of both non-MAX

codons and frameshift mutations involves the nick translation of the inserted DNA by polymerase enzymes within the host, without subsequent repair by the host cell.

Models which explain the insertion or deletion of bases by DNA polymerases, centre around the concept of slippage of the two strands of DNA being transcribed. Two theories have been described to account for this template slippage. The first involves the slippage of the DNA strands by mis-priming at iterated bases (Streisinger *et al.*, 1966), the second theory suggests slippage results from misincorporation of bases by the polymerase, followed by a subsequent realignment of the priming strand of DNA (Pham *et al.*, 1998). Both theories rely upon the misalignment of the priming strand and template DNA, followed by extension of the misprimed DNA, by polymerase enzymes, to stabilise the inserted or deleted base. The diagrams in figure 6.1 demonstrate how DNA slippage could account for the insertion or deletion of bases from MAX cassettes in which selection oligonucleotides were not hybridised correctly to the template strand.

(A)

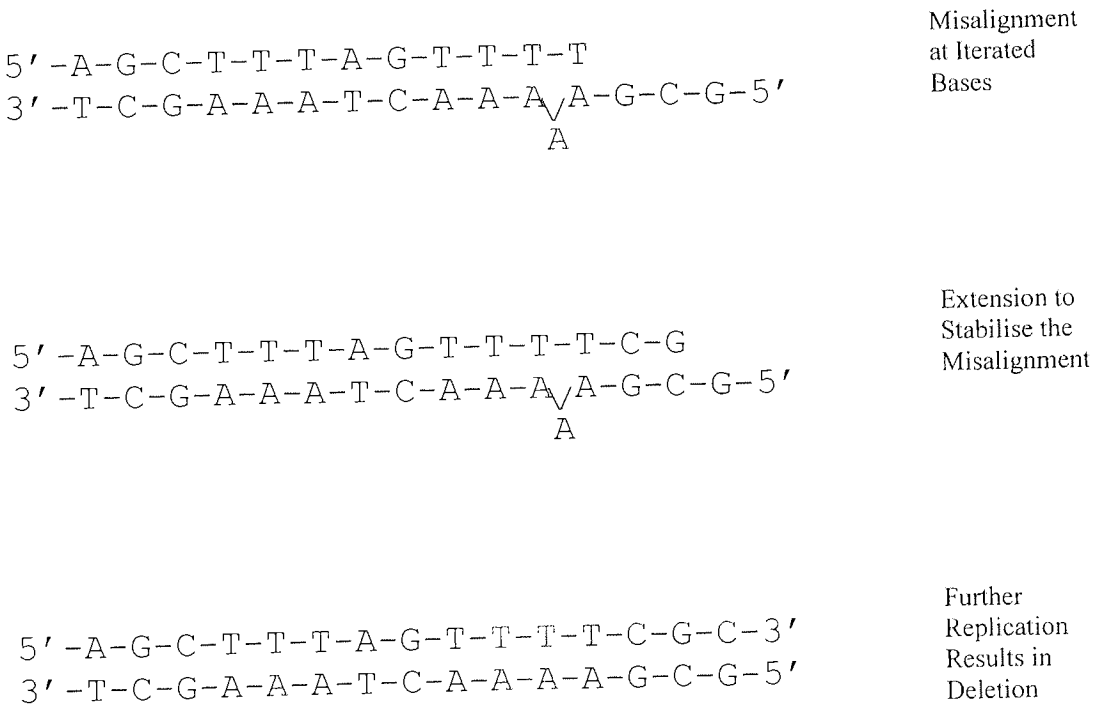


Fig 6.1 **(A)** Misalignment of the α selection oligonucleotide at repeated bases within the randomised position, leads to the dislocation of a thymine base in the template strand. Polymerisation from the correctly base paired 3' primer terminus, stabilises the unpaired base. Further replication of the DNA causes the deletion of the unpaired base.

(B)

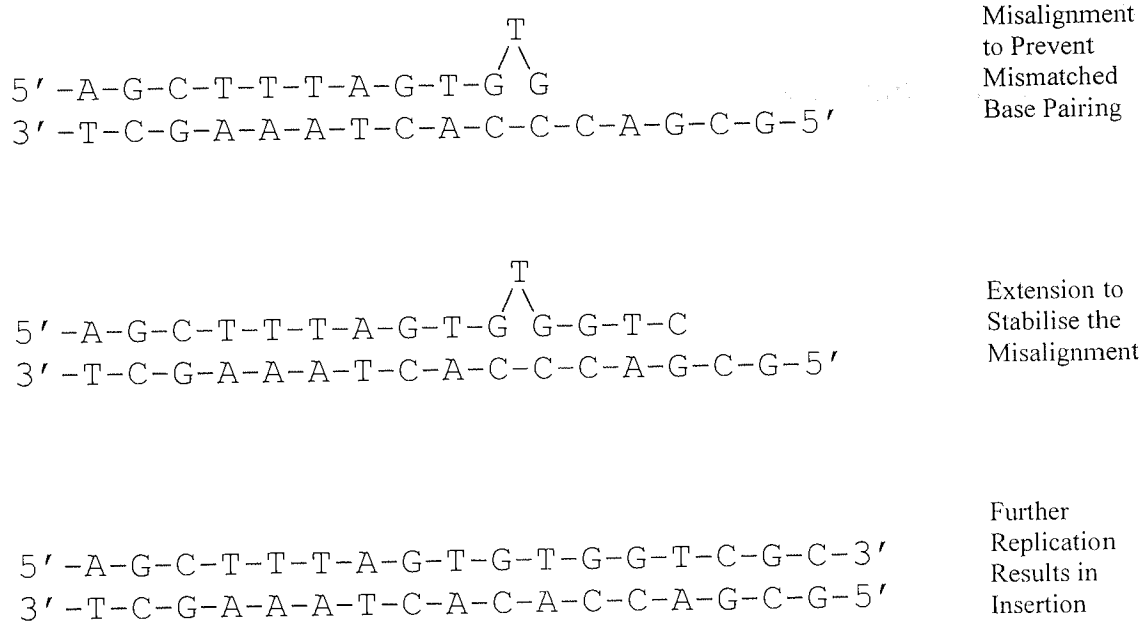


Fig 6.1 (B) Misalignment of the α selection oligonucleotide by dislocation of a thymine base facilitates the correct G:C base pairing of the 3' terminus of the selection oligonucleotide. Polymerisation from the correctly base paired 3' primer terminus, stabilises the unpaired base. Further replication of the DNA results in an insertion to complement the unpaired base. Key to figures: Template strand and polymerised DNA are shown in black text; α selection oligonucleotide is shown in blue text; Position of randomisation shown in red text

The diagrams in figure 6.1 demonstrate only the insertion and deletion of a single base pair. However template slippage may result in the insertion or deletion of a number of base pairs, dependant upon the alignment of the priming strand of DNA. As can be seen in the figure, the insertion or deletion of bases is dependant on the position of the unpaired nucleotide.

The predominant frameshift mutation occurring within the randomised sequences involved the insertion of additional bases. This would be expected if the DNA slippage model was correct, as the unpaired base(s) would be dislocated from the selection oligonucleotides. This dislocation may not only be more easily achieved in these short oligonucleotides, but also results in the generation of an unpaired nick between the mispaired strand and the succeeding selection oligonucleotide. This nick would not be

sealed by the subsequent ligation of the cassette and would provide an exposed 3' terminus for the initiation of DNA polymerase I activity.

Analysis of the sequence at which frameshift mutations occurred is difficult as the majority of mutations occurred within the randomised regions of the inserted DNA. However analysis of those mutations which occurred in the conserved sequence of the inserted DNA, shows that approximately 70 % of these mutations occur within a sequence of the γ oligonucleotide which contains two iterated bases (TTACAA).

The mispriming of DNA polymerases within the host cell appears to be the most plausible explanation for the generation of frameshift mutations within the MAX libraries. As nicks in double stranded DNA have been shown to be the only substrate required to initiate DNA polymerase I activity (Kelly *et al.*, 1970), it is likely that this would be due to the activity of DNA polymerase I, initiated at unsealed nicks within the selection oligonucleotides. As strand slippage may also be initiated during disassociation and reassociation of the enzyme (Kunkel & Bebeneck, 2000) the presence of a number of single stranded nicks in the inserted DNA may also have contributed to the generation of frameshifted inserts. It is interesting to note that when the MAX cassette was generated with only three selection oligonucleotides (Section 4.3), only 16 % of the recovered clones contained frameshift mutations.

The generation of clones containing frameshift mutations and non-MAX codons therefore appears to result from cloning mutagenic cassettes containing misaligned or mismatched selection oligonucleotides. This led to the idea of using PCR to amplify only the top strand of the cassette, which is comprised of only the selection oligonucleotides. This work was developed by Dr Marcus Hughes after the completion of the experimental work in this study. The PCR amplification of the MAX cassettes is shown schematically in Figure 6.2.

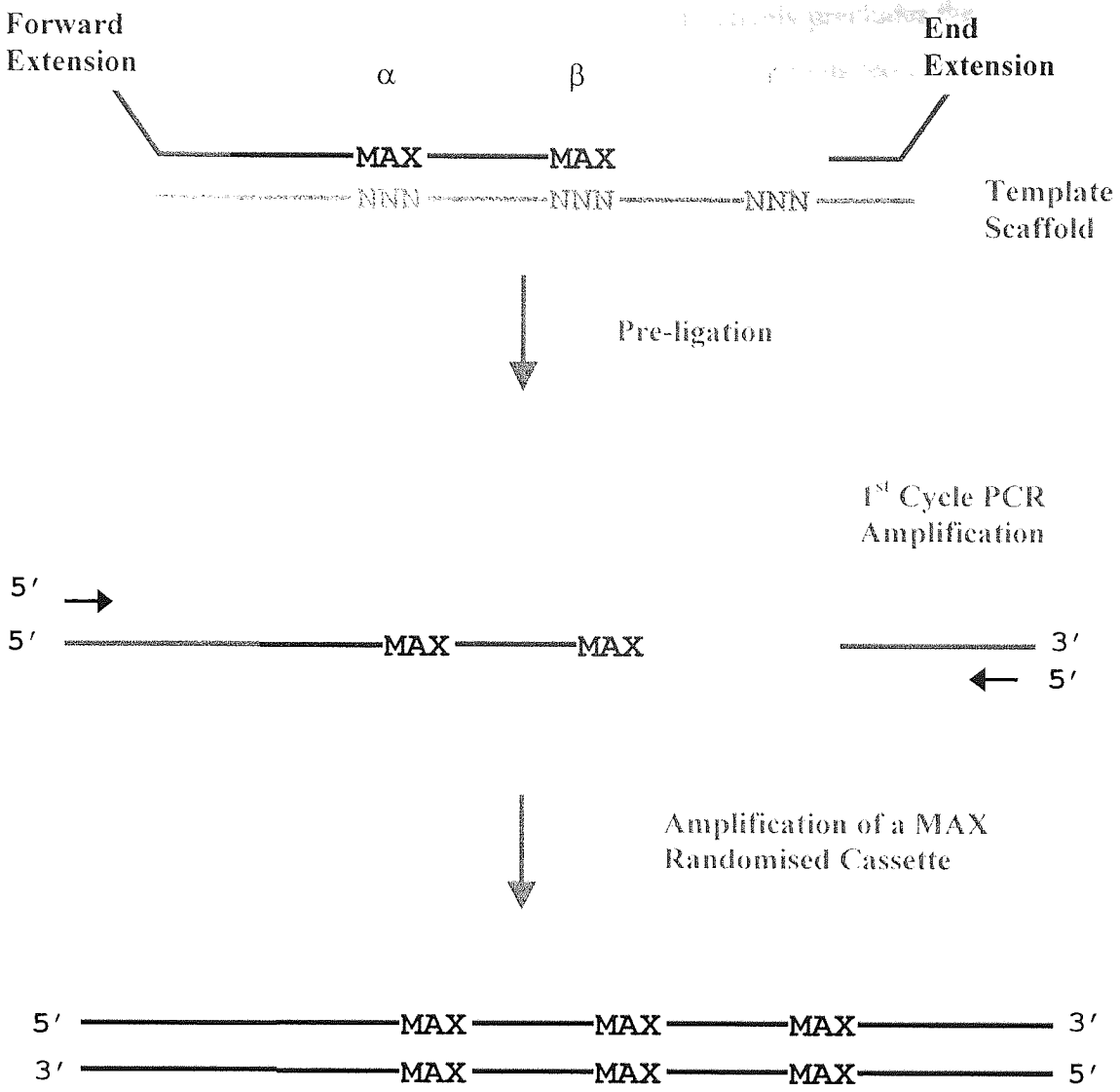


Fig 6.2. Schematic representation of the generation of MAX randomised cassettes by PCR. In the PCR generation of the MAX cassettes, selection oligonucleotides are hybridised to a DNA template scaffold. Two further oligonucleotides (Forward Extension & End Extension), which contain short regions complementary to the scaffold strand are also hybridised in the reaction. These oligonucleotides are ligated together to form a contiguous single strand of DNA. The double stranded cassette is employed in a PCR reaction, using primers directed at the non complementary regions of the Forward Extension & End Extension oligonucleotides. As the primers used in the PCR bear no complementarity to the template scaffold, amplification of the scaffold strand is prohibited.

As only the top strand of the DNA can be amplified in the first cycle of the reaction, amplification of the conventionally randomised template strand is prohibited, standard double stranded DNA cassettes are generated which contain MAX codons at the positions of randomisation. Cassettes must contain each selection oligonucleotide and

the two additional oligonucleotides, ligated together in a single strand, to generate the two primer binding sites of the PCR reaction. This effectively precludes the amplification of cassettes containing misaligned selection oligonucleotides.

Library generation using the PCR generated MAX cassettes has reduced the generation of both frameshifted clones and clones which contain non-MAX codons to approximately 3 % of the total number of recovered clones (Dr. M. Hughes, pers. comm.). In addition the amplification of the cassettes by PCR facilitates the creation of much smaller libraries, as the need to clone large amounts of template DNA is removed. This further increases the capability of the technique to "fix" positions of randomisation to a defined subset of amino acids, to such a degree that single clones have been generated using MAX randomisation.

The development of a new randomisation procedure throughout this study, highlighted the difficulties associated with the generation of fully representative randomised gene libraries. In addition to the theoretical difficulties inherent in the creation of representative gene libraries when using a degenerate genetic code, further problems associated with library generation were highlighted in this study. These problems fall into two categories, those associated with the MAX technique itself such as the generation of frameshifted clones, and those associated with the construction of all randomised gene libraries such as the negative selection of toxic clones. These problems have been addressed by the development of the MAX technique and continued work carried out on the basis of the results and techniques developed throughout this study.

The continued development of the MAX randomisation technique has enabled the creation of small, representative and controlled randomised gene libraries. Currently the technique is being employed to create small positionally fixed libraries of zinc finger proteins, in an attempt to identify zinc fingers which bind with high affinity, each of the possible 64, three base pair zinc finger target sites.

The utility of a randomisation technique with which multiple codons can be substituted with those encoding any directed subset of amino acids cannot be underestimated. The MAX randomisation technique could provide a powerful tool in the elucidation of protein structure/function relationships and in the discovery/generation of proteins with

desired or novel activity. The possible applications of the MAX technique in future studies are myriad and diverse, as the technique can be applied in the randomisation of most proteins.

Future development of the MAX randomisation technique could centre around the extension of the technique to encompass the randomisation of contiguous DNA codons. In the current study the MAX randomisation technique relies upon a complementary flanking sequence to facilitate the hybridisation of the selection oligonucleotides. This in theory limits randomisation to two adjacent codons contained at the 5' and 3' ends of two adjoining oligonucleotides. Future development of the technique, could attempt to include small oligonucleotides of 3 – 4 bp in the hybridisation reaction, to facilitate the mutagenesis of contiguous codons, enabling virtually any randomisation strategy to be performed using this technique.

References

Abbondanza, C., Medici, N., Nigro, V., Rossi, V., Gallo, L., Piluso, G., Belsito, A., Roscigno, A., Bontempo, P., Puca, A. A., Molinari, A. M., Moncharmont, B. & Puca, G. A. (1999). "The Retinoblastoma-interacting Zinc Finger Protein RIZ is a Downstream Effector of Estrogen Action". *Proc. Nat. Acad. Sci. USA.* **98**: 3130 – 3135.

Amarasinghe, G. K., De Guzman, R. N., Turner, R. B., Chancellor, K. J., Wu, Z. R. & Summers M. F. (2000). "NMR Structure of the HIV-1 Nucleocapsid Protein Bound to Stem-Loop SL2 of the Ψ-RNA Packaging Signal. Implications for Genome Recognition". *J. Mol. Biol* **301**: 491 – 511.

Avalle, B., Vanwetswinkle, S. & Fastrez, J. (1997). "*In vitro* Selection for Catlytic Turnover from a Library of β-Lactamase and Penicillin-binding Mutants". *Bioorganic and Medicinal Chemistry Letters* **7**: 479 – 484.

Avram, D., Fields, A., Pretty, K., Nevrivy, D. J., Ishmael, J. E., & Leid, M. (2000). "Isolation of a Novel Family of C_2H_2 Zinc Finger Proteins Implicated in Transcriptional Repression Mediated by Chicken Ovalbumin Upstream Promoter Transcription Factor (COUP-TF) Orphan Nuclear Receptors". *Journal of Biological Chemistry* **275**: 10315 – 10322.

Barbas III, C. F., Bain, J. D., Hoekstra, D. M. & Lerner, R. A. (1992). "Semisynthetic Combinatorial antibody Libraries: A Chemical Solution to the Diversity Problem". *Proc. Nat. Acad. Sci. USA.* **89**: 4457 – 4461.

Barrick, J. E., Takahashi, T. T., Ren, J., Xia, T. & Roberts R. W. (2001). "Large Libraries Reveal Diverse Solutions to a RNA Recognition Problem". *Proc. Nat. Acad. Sci. USA.* **98**: 12374 – 12378.

Bebeneck, K. and Kunkel, T. A. (1990). "Frameshift Errors Initiated by Nucleotide Misincorporation". *Proc. Nat. Acad. Sci. USA.* **87**: 4946 – 4950.

Bebeneck, K. and Kunkel, T. A. (2000). "DNA Replication Fidelity". *Annu. Rev. Biochem.* **69**: 497 – 529.

Beerli, R. R., Segal, D. J., Dreier, B. & Barbas III, C. F. (1998). "Toward Controlling Gene Expression at Will: Specific Regulation of the *erbB-2/HER-2* Promoter by Using Polydactyl Zinc Finger Proteins Constructed From Modular Building Blocks". *Proc. Nat. Acad. Sci. USA.* **95**: 14628 – 14633.

Beerli, R. R. & Barbas III, C. F. (2002). Engineering Polydactyl Zinc Finger Transcription Factors". *Nature Biotechnology* **20**: 135 – 141.

Berg, J. M. (1987). "Proposed Structure for Three Zinc Binding Domains From Transcription Factor IIIA End Related Proteins". *Proc. Nat. Acad. Sci. USA.* **85**: 99 – 102.

Berg, J. M. (1995). "Zinc Finger Domains: From Prediction to Design". *Acc. Chem. Res.* **28**: 14 – 19.

Berkovits, H. J. & Berg, J. M. (1999). "Metal and DNA Properties of a Two Domain Fragment of Neural Zinc Finger Factor 1, A CCHC Zinc Binding Protein". *Biochemistry* **38**: 16826 – 16830.

Beste, G., Schmidt, F. S., Stibora, T. & Skerra, A. (1999). "Small Antibody-like Proteins with Prescribed Ligand Specificities Derived from the Lipocalin Fold". *Proc. Nat. Acad. Sci. USA.* **96**: 1898 – 1903.

Bhattacharya, R. & Beck, D. J. (2002). "Survival and SOS Induction in Cisplatin-treated *Escherichia coli* Deficient in Pol II, RecBCD and RecFOR Functions". *DNA Repair* **1**: 955 – 966.

Bird, A.J., Zhao, H. Luo, H., Jensen, L. T., Srinivasan, C., Evans-Galea, M., Winge, D. R. & Eide, D. J. (2000). A Dual Role for Zinc Fingers in Both DNA Binding and Zinc Sensing by the Zap1 Transcriptional Activator". *EMBO Journal* **19**: 3704 – 3713.

Blancafort, P., Magnenat, L. & Barbas III, C. F. (2003). "Scanning the Human Genome With Combinatorial Transcription Factor Libraries" *Nature Biotechnology* **21**: 269 – 274.

Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F.,Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeded, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). "The complete Genome sequence of *Escherichia coli* K-12". *Science.* **277**: 1453-1473.

Braunagel, M. & Little, M. (1997). "Construction of a Semisynthetic Antibody Library Using Trinucleotide Oligos". *Nucl. Acids Res.* **25**: 4690 – 4691.

Cantu III, C., Hnang, W. & Palzkill, T. (1996). "Selection and Characterisation of Amino Acid Substitutions at Residues 237-240 of TEM-1 β-Lactamase with Altered Substrate Specificity for Aztreonam and Ceftazidime". *Journal of Biological Chemistry* **271**: 22538 – 22545.

Chantellier,J., Mazza, A., Brousseau, R. & Vernet, T. (1995). "Codon-Based Combinatorial Alanine Scanning Site-Directed Mutagenesis: Design, Implementation and Polymerase Chain Reaction Screening". *Analytical Biochemistry* **229**: 282 – 290.

Cheng, P-Y., Kagawa, N., Takahashi, Y. & Waterman, M. R. (1999). "Three Zinc Finger Nuclear Proteins, Sp1, Sp3 and a ZBP-89 Homologue, Bind to the Cyclic Adenosine Monophosphate-Responsive Sequence of the Bovine Adrenodoxin Gene and Regulate Transcription". *Biochemistry* **39**: 4347 – 4357.

Cheng, X., Kay, B. K. & Juliano R.I. (1996). "Identification of a Biologically Significant DNA-binding Peptide Motif by the use of a Random Phage display Library". *Gene* **171**: 1 – 8.

Cheng, X., Boyer, J. L. & Juliano, R. L. (1997). "Selection of Peptides that Functionally Replace a Zinc Finger in the Sp1 Transcription Factor by using a Yeast Combinatorial Library". *Proc. Nat. Acad. Sci. USA.* **94**: 14120 – 14125.

Chirinos-Rojas, C. L., Steward, M. W. & Partidos, C. D. (1998). "A peptidomimetic Antagonist of TNF-α-Mediated Cytotoxicity Identified from a Phage-displayed Random Peptide Library". *J. Immunology* **161**: 5621 – 5626.

Cho, G., Keefe, A. D., Liu, R., Wilson, D. S. & Szostack, J. W. (2000). "Constructing High Complexity Synthetic Libraries of Long ORFs Using *In Vitro* Selection". *J. Mol. Biol.* **297**: 309 – 319.

Choo, Y. & Islan, M. (2000). "Advances in Zinc Finger Engineering". *Current Opinion In Structural Biology* **10**: 411 – 416.

Choo, Y. & Klug, A. (1994a). "Selection of DNA Binding Sites for Zinc Fingers Using Rationally Randomised DNA Reveals Coded Interactions". *Proc. Nat. Acad. Sci. USA.* **91**: 11168 – 11172.

Choo, Y. & Klug, A. (1994b). "Towards a Code for the Interactions of Zinc Fingers with DNA: Selection of Randomised Fingers Displayed on Phage". *Proc. Nat. Acad. Sci. USA.* **91**: 11163 – 11167.

Choo, Y. & Islan, M. (2000). "Advances in Zinc Finger Engineering". *Current Opinion in Structural Biology* **10**: 411 – 416.

Chou, A. Y., Archdeacon, J. & Kado, C. I. (1998). "*Agrobacterium* Transcriptional Regulator Ros is a Prokaryotic Zinc Finger Protein that Regulates the Plant Oncogene *ipt*". *Proc. Nat. Acad. Sci. USA.* **95**: 5293 – 5298.

Christy, B. A., Lau, L. F. & Nathans D. (1988). "A Gene Activated in the Mouse 3T3 Cells by Serum Growth Factors Encodes a Protein with "Zinc Finger" Sequences". *Proc. Nat. Acad. Sci. USA.* **85**: 7857 – 7861.

Collins, J., Horn, N., Wadenbäck, J. & Szardenings, M. (2001). "Cosmix-plexing a Novel Recombinational Approach for Evolutionary Selection from Combinatorial Libraries". *Reviews in Molecular Biotechnology* **74**: 317 – 338.

Cook, T., Gebelein, B., Belal, M., Mesa, K. & Urritia, R. (1999). "Three Conserved Transcriptional Repressor Domains are a Defining Feature of the TIEG Subfamily of Sp1-like Zinc Finger Proteins". *Journal of Biological Chemistry* **274**: 29500 – 29504.

Cormack, B. P. & Struhl, K. (1993). "Rational Codon Randomisation: Defining a TATA-Binding Protein Surface Required for RNA Polymerase III Transcription". *Science* **262**: 244 – 248.

Cull, M. G., Miller, J. F. & Schatz, P. J. (1992). "Screening for Receptor Ligands Using Large Libraries of Peptides Linked to the C Terminus of the *lac* Repressor". *Proc. Nat. Acad. Sci. USA.* **89**: 1865 – 1869.

Danielsen, S., Eklund, M., Deussen, H-J., Gräslund, T., Nygren, P-A. & Borchert, T. V. (2001). "*In Vitro* Selection of Enzymatically Active Lipase Variants from Phage Libraries Using a Mechanism Based Inhibitor". *Gene* **272**: 267 – 274.

Daugherty, P. S., Chen, G., Iverson, B. L. & Georgiou, G. (2000). "Quantitative Analysis of the Effect of the Mutation Frequency on the Affinity Maturation of Single Chain Fv Antibodies". *Proc. Nat. Acad. Sci. USA.* **97**: 2029 – 2034.

Davanloo, P., Rosenberg, A. H., Dunn, J. J. and Studier, W. F. (1984). "Cloning and Expression of the Gene for Bacteriophage T7 RNA Polymerase". *Proc. Nat. Acad. Sci. USA.* **81**: 2035 – 2039.

De Guzman, R. N., Liu, H. Y., Martinez-Yamout, M., Dyson, J. H. & Wright, P. E. (2000). Solution Structure of the TAZ2 (CH3) Domain of the Transcriptional Adaptor Protein CBP". *J. Mol. Biol.* **303**: 243 – 253.

Desjarlais, J. R. & Berg, J. M. (1992a). "Towards Rules Relating Zinc Finger Protein Sequences and DNA Binding Site Preferences". *Proc. Nat. Acad. Sci. USA.* **89**: 7345 – 7349.

Desjarlais, J. R. & Berg, J. M. (1992b). "Redesigning the DNA binding Specificity of a Zinc Finger Protein: A Data Base-Guided Approach". *PROTEINS: Structure, Function and Genetics* **12**: 101 – 104.

Desjarlais, J. R. & Berg, J. M. (1993). "Use of a Zinc Finger Consensus Framework and Specificity Rules to Design Specific DNA Binding proteins". *Proc. Nat. Acad. Sci. USA.* **90**: 2256 – 2260.

Desjarlais, J. R. & Berg, J. M. (1994). "Length-encoded Multiplex Binding Site Determination: Application to Zinc Finger Proteins". *Proc. Nat. Acad. Sci. USA.* **91**: 11099 – 11103.

Deuel, T. F., Guan, L. & Wang Z. (1999). "Wilms' Tumor Gene Product WT1 Arrests Macrophage Differentiation Through Its Zinc Finger Domain". *Biochemical and Biophysical Research Communications.* **254**: 192 – 196.

Dohet, C., Wagner, R. & Radman, M. (1984). "Repair of Defined Single Base-pair Mismatches in *Escherichia coli*". *Proc. Nat. Acad. Sci. USA.* **82**: 503 – 505.

Dohet, C., Wagner, R. & Radman, M. (1986). "Methyl-directed Repair of Frameshift Mutations in Hetroduplex DNA". *Proc. Nat. Acad. Sci. USA.* **83**: 3395 – 3397.

Doi, N. & Yangawa, H. (1999). "Design of Generic Biosensors Based on Green Fluorescent Proteins with Allosteric Sites by Directed Evolution". *FEBS Letters* **453**: 305 – 307.

Doi, N. & Yangawa, H. (1999). "STABLE: Protein-DNA Fusion System for Screening of Combinatorial Protein Libraries *in Vitro*". *FEBS Letters* **457**: 227 – 230.

Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D. & Barbas III, C. F. (2001). Development of Zinc Finger Domains for Recognition of the 5'-ANN-3' Family of DNA Sequences and Their use in the Construction of Artificial Transcription Factors". *Journal of Biological Chemistry* **274**: 29466 – 29478.

Džidić, S. & Petranović, M. (2003). "Mismatch Repair in the antimutator *Escherichia coli mud*". *Mutation Research.* **522**: 27 – 32.

Eisinger D. P. & Trumpower, B. L. (1996). "Long-Inverse PCR to Generate Regional Peptide Libraries by Codon Mutagenesis". *Biotechniques* **22**: 250 – 254.

Elrod-Erickson, M., Benson, T. E. & Pabo, C. O. (1996). "Zif268 Protein-DNA Complex Refined at 1.6 Å: A Model System for Understanding Zinc Finger-DNA interactions". *Structure* 4: 1171.

Elrod-Erickson, M. & Pabo, C. O. (1999). "Binding Studies with Mutants of Zif268". *Journal of Biological Chemistry* **274**: 19281 – 19285.

Felici, F., Castagnoli, L., Musacchio, A. Japelli, R. & Cesareni, G. (1991). "Selection of Antibody Ligands from a large Library of Oligopeptides Expressed on a Multivalent Exposition Vector". *J. Mol. Biol* **222**: 301 – 310.

Fox, A. H., Liew, C., Holmes, M., Kowalski, K., Mackay, J. & Crossley, M. (1999). "Transcriptional Cofactors of the FOG Family Interact with GATA Proteins by Means of Multiple Zinc Fingers". *EMBO Journal* **18**: 2812 – 2822.

Frenkel, D., Solomon, B. and Benhar, I. (2000). "Modulation of Alzheimer's β-amyloid Neurotoxicity by Site-directed Single Chain Antibody". *Journal of Neuroimmunology* **106**: 23 – 31.

Gaytan, P., Osuna, J. & Soberón, X. (2002). "Novel Ceftazidime-resistance – Lactamase Generated by a Codon-Based Mutagenesis Method and Selection". *Nucl. Acids Res.* **30**: e84 1 –13.

Greisman, H. A. and Pabo, C. O. (1997). "A General Strategy for Selecting High-Affinity Zinc Finger Proteins for Diverse DNA Target sites". *Science* **275**:657 – 661.

Grishin, N. V. (2001). "Treble Clef Finger – A Functionally Diverse Zinc-Binding Structural Motif". *Nucl. Acids Res.* **29**: 1703 – 1714.

Gunneriusson, E., Nord, K., Uhlén, N. & Nygren P-Å. (1999). "Affinity Maturation of a *Taq* DNA Polymerase Specific Antibody by Helix Shuffling". *Protein Engineering* **12**: 873 – 878.

Hanes, J. Jermutas, L., Webber-Bornhauser, S., Bosshard, H. R. & Plückthun, A. (1998). "Ribosome Display Effectively Selects and Evolves High-affinity Antibodies *in vitro* from Immune Libraries". *Proc. Nat. Acad. Sci. USA.* **95**: 14130 – 14135.

Harris, R. S., Bull, H. J. and Rosenberg, S. M. (1997). "A Direct Role for DNA Polymerase in Adaptive Reversion of a Frameshift Reversion Mutation in *Escherichia coli*". *Mutation Research* **375**: 19 – 24.

He, G-P., Kim, S. & Ro, H-S. (1999). "Cloning and Characterization of a Novel Zinc Finger Transcriptional Repressor". *Journal of Biological Chemistry* **274**: 14678 – 14684.

He, M. & Taussig, M. J. (1997). "Antibody-Ribosome-mRNA (ARM) Complexes as Efficient Selection Particles for *in vitro* Display and Evolution of Antibody combining sites". *Nucl. Acids Res.* **25**: 5132 – 5134.

Hemavathy, K., Shovon, I., Ashraf, Y. & Tony I. P. (2000). "Snail/Slug Family of Repressors: Slowly Going into the Fast Lane of Development and Cancer". *Gene* **257**: 1 – 12.

Hermes, J. D., Parekh, S. m., Blacklow, S. C., Koster, H. & Knowles, J. R. (1989). "A Reliable Method for Random Mutagenesis: The Generation of Mutant Libraries using Spiked Oligodeoxyribonucleotide Primers". *Gene* **84**: 143 – 151.

Higashitani, N., Higashitani, A., Roth, A. & Horiuchi, K. (1992). "SOS Induction in *Escherichia coli* by Infection with Mutant Filamentous Phage that are Defective in Initiation of Complementary Strand Synthesis". *Journal of Bacteriology* **174**: 1612 – 1618.

Ho, S. P., Britton, D. H. O., Stone, B. A., Behrens, D. L., Leffet, L. M., Hobbs, F. W., Miller, J. A. & Trainor, G. L. (1996). "Potent Antisense Oligonucleotides to the Human Multidrug Resistance-1 mRNA are Rationally Selected by Mapping RNA-accessible Sites with Oligonucleotide Libraries". *Nucl. Acids Res.* **24**: 1901 – 1907.

Hughes, M.D., Nagel, D.A., Santos, A.F., Sutherland, A.J. & Hine, A.V. (2003). "Removing the redundancy from randomised gene libraries". *J. Mol. Biol.* **331**: 967 - 972.

Huse, W. D., Sastry, L., Iverson, S. A., Kang, A. S., Alting-mees, M., Burton, D. R., Benkovic, S. J. & Lerner, R. A. (1989). "Generation of a Large Combinatorial Library of the Immunological Repertoire in Phage Lambda". *Science* **246**: 1275 – 1281.

Imanshi, M., Hori, Y., Nagaoka, M. & Sugiura, Y. (2000). "DNA Bending Finger: Artificial Design of 6-Zinc Finger Peptides with Polyglycine Linker and Induction of DNA Bending". *Biochemistry* **39**: 4383 – 4390.

Islan, M., Klug, A. & Choo, Y. (1998). "Comprehensive DNA Recognition Through Concerted Interaction from Adjacent Zinc Fingers". *Biochemistry* **37**: 12026 – 12033.

Islan, M. & Choo, Y. (2000). "Engineered Zinc Finger Proteins That Respond to DNA Modification by *Hae*III and *Hha*I Methyltransferase Enzymes". *J. Mol. Biol.* **295**: 471 – 477.

Jacobs, G. H. (1992). "Determination of the Base Recognition Positions of Zinc Fingers From Sequence Analysis". *EMBO Journal* **11**: 4507 – 4517.

Jamieson, A. C., Kim, S-H. & Wells, J. A. (1994). "*In Vitro* Selection of Zinc Fingers with Altered DNA-Binding Specificity". *Biochemistry* **33**: 5689 – 5695.

Jamieson, A. C., Wang, H. & Kim, S-H. (1996). "A Zinc Finger Directory for High-affinity DNA Recognition". *Proc. Nat. Acad. Sci. USA.* **93**: 12834 – 12839.

Jeffery, C. J. & Koshland, Jr. D.E. (1999). "The *Escherichia coli* Aspartate Receptor: Sequence Specificity of a Transmembrane Helix Studied by Hydrophobic-biased Random Mutagenesis". *Protein Engineering* **12**: 863 – 871.

Jermutus, L., Tessier, M., Pasamontes, L., van Loon, A. P. G. M. and Lehmann, M. (2001). "Structure-based Chimeric Enzymes as an Alternative to Directed Enzyme Evolution: Phytase as a Test Case". *Journal of Biotechnology* **85**: 15 – 21.

Jiricny, J. (2000). "Mismatch Repair: The Praying Hands of Fidelity". *Current Biology* **10**: R788 – R790.

Joung, J. K., Ramm, E. I. & Pabo, C. O. (2000). "A Bacterial Two-Hybrid Selection System for Studying Protein-DNA and Protein-Protein Interactions". *Proc. Nat. Acad. Sci. USA.* **97**: 7382 – 7387.

Kamiuchi, T., Emiko, A., Miki, I., Kaji. T., Nagaoka, M. & Sugiura, Y. (1998). "Artificial Nine Finger Peptide With 30 Base Pair Binding Sites". *Biochemistry* **37**: 13827 - 13834

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). "Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids". *Science* **262**: 1680 – 1685.

Kang, J. S. & Kim, J-S. (2000). "Zinc Finger Proteins as Designer Transcription Factors". *Journal of Biological Chemistry* **275**: 8742 – 8748.

Kayushin, A. L., Korosteleva, A. I., Miroshnikov, W., Kosch, W., Zubov, D. & Piel N. (1996). "A Convenient Approach to the Synthesis of Trinucleotide Phosphoramidites Synthons for the Generation of Oligonucleotide/peptide Libraries". *Nucl. Acids Res.* **24**: 3748 - 3755.

Kelly, R. B., Cozarelli, N. R., Deutscher, M. P., Lehman, I. R. & Kornberg, A. (1970). "Enzymatic Synthesis of DNA. XXXII. Replication of Duplex DNA by Polymerase at a Single Strand Break". *Journal of Biological Chemistry* **245**: 35 – 49.

Kim, C. A. & Berg, J. M. (1995). "Serine at Position 2 in the DNA Recognition Helix of a Cys_2-His_2 Zinc Finger Peptide is Not, in General, Responsible for Base Recognition". *J. Mol. Biol.* **252**: 1 – 5.

Kim, J-S. & Pabo, C. O. (1998). "Getting a Handhold on DNA: Design of Polydactyl Zinc Finger Proteins with Femtomolar Dissociation Constants". *Proc. Nat. Acad. Sci. USA.* **95**: 2812 – 2817.

Kitamura, K., Kinoshita, Y., Narasaki, S., Nemoto, N., Husimi, Y. & Nishigaki, K. (2002). "Construction of Block-shuffled Libraries of DNA for Evolutionary Protein Engineering: Y-ligation-based Block Shuffling". *Protein Engineering* **15**: 843 – 853.

Klein, B.K., Olins, P. O., Bauer, C., Caparon, M. H., Easton, A. M., Bradford S. R., Abrams, M. A., Klover J. A., Paik, K., Thomas, J. W., Hood, W.F., Shieh, J-J., Polazzi, J. O., Donnelly, A. M., Zeng, D. L., Welply, J. K. & McKearn, J. P. (1999). "Use of Combinatorial Mutagenesis to Select for Multiply Substituted Human Interleukin-3 Variants with Improved Pharmacologic Properties". *Experimental Haematology* **27**: 1746 – 1756.

Koide, A., Bailey, C. W., Huang, X. & Koide, S. (1998). "The Fibronectin Type III Domain as a Scaffold for Novel Binding Proteins". *J. Mol. Biol.* **284**: 1141 – 1151.

Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G., Hoess, A., Wölle, J., Plückthun, A. & Virnekäs, B. (2000). "Fully Synthetic Combinatorial Antibody Libraries (HuCAL) Based on Modular Consensus Frameworks and CDRs Randomised with Trinucleotides". *J. Mol. Biol.* **296**: 57 – 86.

Kraulis, P. J., Raine A. R. C., Gadhavi, P. L., & Laue E. D. (1992). "Structure of the Zinc Containing Domain of GAL4". *Nature* **356**: 448 – 450.

Krishna, S. S., Madjumdar, I. & Grishin, N. V. (2003). "Structural Classification of Zinc Fingers". *Nucl. Acids Res.* **31**: 532 – 550.

Krizek, B. A., Ammann, B. T., Kilfoil, V. J., Merkle, D. L. & Berg, J. M. (1991). "A consensus Zinc Finger Peptide: Design, high affinity Metal Binding, a pH dependant structure and a His to Cys sequence variant". *Journ. American Chem. Soc.***113**: 4518-4523.

Kunkel, T.A. (1990). "Misalignment-Mediated DNA Synthesis Errors". *Biochemistry* **29**: 8003 – 8011.

Kunkel, T.A. & Bebeneck, K. (2000). "DNA Replication Fidelity". *Annu. Rev. Biochem.* **69**: 497 – 529.

Kusabe, T., Hine, A. V., Hyberts, S. G. & Richardson, C. C. (1999). "The Cys_4 Zinc Finger of Bacteriophage T7 Primase in Sequence-Specific Single-Stranded DNA Recognition". *Biochemistry* **96**: 4295 – 4300.

Kuwahara, J., Yonezawa, A., Futamura, M. & Sugiura, Y. (1993). "Binding of Transcription Factor Sp1 to GC Box DNA Revealed by Footprinting Analysis: Different Contact of Three Zinc Fingers and Sequence Recognition Mode". *Biochemistry* **32**: 5994 – 6001.

Lahr, S. J., Broadwater, A., Carter, C. W., Collier, M. L., Hensley, L., Waldner, J. C., Pielak, G. J. & Edgell, M. H. (1999). "Patterned Library Analysis: A Method for the Quantitative Assessment of Hypotheses Concerning the Determinants of Protein Structure". *Proc. Nat. Acad. Sci. USA.* **96**: 14860 – 14865.

Laity, J.H., Dyson, H. J. & Wright, P. E. (2000). "Molecular Basis for Modulation of Biological Function by Alternate Splicing of the Wilms' Tumour Suppressor Protein". *Proc. Nat. Acad. Sci. USA.* **97**: 11932 – 11935.

Laity, J.H., Lee, B. M. & Wright, P. E. (2001). "Zinc Finger Proteins: New Insights into Structural and Functional Diversity". *Current Opinion in Structural Biology.* **11**: 39 – 46.

Larsson, O., Thormeyer, D., Asinger, A., Wihlén, B., Wahlestedt, C. and Liang, Z. (2002). "Quantitative Codon Optimisation of DNA Libraries Encoding Sub-random Peptides: Design and Characterisation of a Novel Library Encoding Transmembrane Domain Peptides". *Nucl. Acids Res.* **30**: e133 pages 1 – 8.

Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A. & Wright, P. E. (1989). "Three Dimensional Structure of a Single Zinc Finger DNA Binding Domain". *Science* **245**: 635 – 637.

Lee, S. & Garfinkel, M. D. (2000). "Characterisation of *Drosphila* OVO Protein DNA Binding Specificity Using Random DNA Oligomer Selection Suggests Zinc Finger Degeneration". *Nucl. Acids Res.* **28**: 826 – 834.

Lee, S. Y., Choi, J. H. & Xu, Z. (2003). "Microbial Cell Surface Display". *TRENDS in Biotechnology* **21**: 45 – 52.

Legendre, D. & Fastrez, J. (2002). "Construction and Exploitation in Model Experiments of Functional Selection of a Landscape Library Expressed from a Phagemid". *Gene* **290**: 203 – 215.

Lima, W. F. & Crooke, S. T. (1999). "Highly Efficient endonucleolytic Cleavage of RNA by a Cys_2His_2 Zinc-Finger Peptide". *Proc. Nat. Acad. Sci. USA.* **96**: 10010 – 10015.

Liu, Q., Segal, D. J., Ghiara, J. B. & Barbas III, C. F. (1997). "Design of Polydactyl Zinc-Finger Proteins for Unique Addressing Within Complex Genomes". *Proc. Nat. Acad. Sci. USA.* **94**: 5525 – 5530.

Lowman, H. B. & Wells, J. A. (1993). "Affinity Maturation of Human Growth Hormone by Monovalent Phage Display". *J. Mol. Biol.* **234**: 564 – 578.

Lui, Q., Xia, Z. Q. & Case, C. C. (2002) Validated Zinc Finger Protein Designs for All 16 GNN Triplet Targets". *Journal of Biological Chemistry* **277**: 3850 – 3856.

Lyons, S. M., & Schendel, P. F. (1984). "Kinetics of Methylation in *Escherichia coli* K-12". *Journal of Bacteriology* **159**: 421 – 423.

Maas, W. K., Wang, C., Lima, T., Zubay, G. & Lim, D. (1994). "Multicopy Single-stranded DNAs with Mismatched Base Pairs are Mutagenic in *Escherichia coli*". *Mol. Microbiology*. **14**: 437 – 441.

Mansour, C. A., Doiron, K. M. J. & Cupples, C. G. (2001). "Characterisation of Functional Interactions Among the *Escherichia coli* Mismatch Repair Proteins Using a Bacterial Two-Hybrid System". *Mutation Research*. **485**: 331 – 338.

Mao, J., Inouye, S. & Inouye, M. (1996). "Enhancement of Frameshift Mutation by the Overproduction of msDNA in *Escherichia coli*". *FEMS Microbiology Letters* **144**: 109 – 115.

Mao, S., Goa, C., Lo, C-H. L., Wirsching, P., Wong, C-H. & Janda, K. D. (1999). "Phage-display Library selection of High-affinity Single-chain Antibodies to Tumor-associated Carbohydrate Antigens Sialyl Lewis^x and Lewis^x". *Proc. Nat. Acad. Sci. USA*. **96**: 6593 – 6598.

Martinez-Yamout, M., Legge, G. B., Zhang, O., Wright, P. E. & Dyson, H. J. (2000). "Solution Structure of the Cystine-Rich Domain of the *Escherichia coli* Chaperone Protein DnaJ". *J. Mol. Biol* **300**: 805 – 818.

Matsuura, T., Ernst, A. & Plückthun, A. (2002). "Construction and Characterisation of Protein Libraries Composed of Secondary Structure Modules". *Protein Science* **11**: 2631 – 2643.

Mattheakis, L. C., Bhatt, R. R. & Dower W. J. (1994). "An *in vitro* Polysome Display System for Identifying Ligands from Very Large Peptide Libraries". *Proc. Nat. Acad. Sci. USA*. **91**: 9022 – 9026.

Matthews, J. M., Kowalski, K., Liew, C. K., Sharpe, B. K., Fox, A. H., Cossley, M. & Mackay, J. P. (2000). "A Class of Zinc Fingers Involved in Protein-Protein Interactions: Biophysical Characterization of CCHC Fingers from Fog and U-Shaped" *Eur. J. Biochem.* **267**: 1030 – 1038.

McConnell, S. J. & Hoess, R. H. (1995). "Tendamistat as a Scaffold for Conformationally Constrained Phage Peptide Libraries". *J. Mol. Biol.* **250**: 460 – 470.

Michael, S. P., Kiddoh, V. I., Schmide, M. H., Amann, B. T. & Berg, J. M. (1992). "Metal Binding and Folding Properties of a Minimalist Cys_2His_2 Zinc Finger Peptide". *Proc. Nat. Acad. Sci. USA*. **91**: 4796 – 4800.

Mierendorf, R., Yeager, K. & Novoy, R. (1994) *InNovations* 11. No. 1

Miller, J., McLachlan, A. D. & Klug A. (1985). "Repetetive Zinc Binding Domains in the Protein Transcription Factor IIIA from *Xenopus*". *EMBO Journal*. **4**: 1609 – 1614.

Miller, J. C. & Pabo, C. O. (2001). "Rearrangement of Side Chains in a Zif268 Mutant Highlights the Complexities of Zinc Finger-DNA Recognition". *J. Mol. Biol.* **313**: 309 – 315.

Modrich, P. (1989). "Methyl Directed DNA Mismatch correction". *Journal of Biological Chemistry*. **264**: 6597 – 6600.

Moore, M., Klug, A. & Choo, Y. (2001). "Improved DNA Binding Specificity from Polyzinc Finger Peptides by Using Strings of Two-Finger Units". *Proc. Nat. Acad. Sci. USA*. **98**: 1437 – 1441.

Moore, M., Choo, Y. & Klug, A. (2001). "Design of Polyzinc Finger Peptides with Structured Linkers". *Proc. Nat. Acad. Sci. USA*. **98**: 1432 – 1436.

Mössner, E. & Plückthun, A. (2001). "Directed Evolution with Fast and Efficient Selection Technologies" *CHIMA*. **55**: 324 – 328.

Nagaoka, M. & Sugiura, Y. (2000). "Artificial Zinc Finger Peptides: Creation, DNA recognition and Gene Regulation". *Journal of Inorganic Biochemistry* **82**: 57 – 63.

Nagaoka, M., Nomura, W. Shiraishi, Y. & Sugiura, Y. (2001). "Significant Effect of Linker Sequence on DNA Recognition by Multi-Zinc Finger Protein". *Biochemical and Biophysical Research Communications* **282**: 1001 – 1007.

Nakamura, T., Yamazaki, Y., Saiki, Y., Moriyama, M., Largaespada, D. A., Jenkins, N. A. & Copeland, N. G. (2000). "*Evi9* Encodes a Novel Zinc Finger Protein That Physically Interacts with BCL6, a Known Human B-Cell Proto-Oncogene Product". *Molecular and Cellular Biology* **20**: 3178 – 3186.

Nakmura, Y., Gojobori, T. & Ikemura, T. (2000). "Codon Usage Tabulated from the International DNA Sequencing databases: Status for the year 2000". *Nucl. Acids Res.* **28**: 292.

Narayan, V. A., Kriwaki, R. W. & Caradonna, J. P. (1997). "Structures of Zinc Finger Domains from Transcription Factor Sp1". *Journal of Biological Chemistry*. **272**: 7801 – 7809.

Nardelli, J., Gibson, T. J., Vesque, C. & Charnay, P. (1991). "Base Sequence Discrimination by Zinc Finger Domains". *Nature* 349: 175 –178.

Nemoto, N., Miyamoto-sato, E., Husimi, Y. & Yanagawa, H. (1997). "*In vitro* Virus: Bonding of mRNA bearing puromycin at the 3'-Termininal End to C-terminal End of its encoded Protein on the Ribosome *In Vitro*". *FEBS Letters* **414**: 405 – 408.

Neuner, P., Cortese, R. & Monaci, P. (1998). "Codon Based Mutagenesis Using Dimer-Phosphoramidites". *Nucl. Acids Res.* **26**: 1223 – 1227.

Noren, K. A. & Noren, C. L. (2001). "Construction of High Complexity Combinatorial Phage Display Peptide Libraries". *Methods* **23**: 169 – 178.

Pakula, A. A. & Simon, M. I. (1992). "Determination of Transmembrane Protein Structure by Disulphide Cross-Linking: The *Escherichia Coli* Tar Receptor". *Proc. Nat. Acad. Sci. USA*. **89**: 4144 – 4148.

Papworth, M., Moore, M., Islan, M., Minczuk. M., Choo, Y. & Klug, A. (2003). "Inhibition of Herpes Simplex Virus 1 Gene Expression by Designer Zinc Finger Transcription Factors". *Proc. Nat. Acad. Sci. USA.* **100**: 1621 – 1626.

Parikh, A. & Guengerich F. P. (1997). "Random Mutagenesis by Whole Plasmid PCR Amplification". *Biotechniques* **24**: 428 – 431.

Pascual, J., Martinez-Yamout, M., Dyson, H. J. & Wright, P. E. (2000). "Structure of the PHD Zinc Finger Domain from Human Williams-Beuren Syndrome Transcription Factor". *J. Mol. Biol.* **304**: 723 – 729.

Patek, C. E., Little, M. H., Fleming, S., Miles, C., Charieu, J. P., Clarke, A. R., Miyagawa, K., Christie, S., Doig, J., Harrison, D. J., Porteous, D. J., Brookes, A. J., Hooper, M. L. & Hastie, N. D. (1999). "A Zinc Finger Truncation of Murine WT1 Results in the Characteristic Urogenital Abnormalities of Denys-Drash Syndrome". *Proc. Nat. Acad. Sci. USA.* **96**: 2931 – 2936.

Patzel, V. & Sczakiel, G. (2000). "*In Vitro* Selection Supports the View of Kinetic control of Antisense RNA-mediated Inhibition of Gene Expression in Mammalian Cells". *Nucl. Acids Res.* **28**: 2462 – 2466.

Pavletich, N. P. & Pabo, C. O. (1991). "Zinc Finger DNA Recognition: Crystal Structure of a Zif268 DNA Complex at 2.1 Å". *Science* **252**: 809 – 817.

Petrenko, V. A., Smith, G. P., Mazooji, M. M. & Quinn, T. (2002). "α-Helically Constrained Phage Display Library". *Protein Engineering* **15**: 943 – 950.

Pham, P. T., Olson, M. W., McHenrey, C. S. and Schaaper, R. M. (1998). "The Base Substitution and Frameshift Fidelity of *Escherichia coli* DNA Polymerase III Holoenzyme *in Vitro*". *Journal of Biological Chemistry.* **273**: 23575 – 23584.

Pierce, H. H., Schachat, F., Brandt, P. W., Lombardo, C. R. & Kay, B.K. (1998). "Identification of Troponin C Antagonists from a Phage-displayed Random Peptide Library". *Journal of Biological Chemistry.* **273**: 23448 – 23453.

Pini, A., Spreafico, A., Botti, R., Neri, D. & Neri, P. (1997). "Hierarchical Affinity Maturation of a Phage Library Derived Antibody for the Selective Removal of Cytomegalovirus from Plasma". *Journal of Immunological Methods* **206**: 171 – 182.

Rader, C. & Barbas III, C. F. (1997). "Phage display of Combinatorial Antibody Libraries". *Current Opinion in Biotechnology* **8**: 503 – 508.

Rastinejad, F. (2001). "Retinoid X receptor and its Partners in the Nuclear Receptor Family". *Current Opinion In Structural Biology* **11**: 33 – 38.

Rebar, E. J. & Pabo C.O. (1994). "Zinc Finger Phage: Affinity Selection of Fingers With New DNA Binding Specificities". *Science* **263**: 671 – 673.

Reidhaar-Olson, J. F. & Sauer, R. T. (1988). "Combinatorial Cassette Mutagenesis as a Probe of the Informational Content of Protein Sequences". *Science* **241**: 53 – 57.

Reiter, Y., Schuck, P., Boyd, L. F. & Plaksin, D. (1999). "An Antibody Single-domain Phage Display Library of a Native Heavy Chain Variable Region: Isolation of Functional Single-domain VH Molecules with a Unique Interface" *J. Mol. Biol.* **290**: 685 – 698.

Reynolds, L., Ullman, C., Moore, M., Islan, M., West, M. J., Clapham, P., Klug, A., and Choo, Y. (2003). "Repression of the HIV-1 5' LTR Promoter and Inhibition of HIV-1 Replication by Using Engineered Zinc-Finger Transcription Factors". *Proc. Nat. Acad. Sci. USA.* **100**: 1615 – 1620.

Rhodes, D. & Klug, A. (1993). "Zinc Fingers – They Play a Key Part in Regulating the Activity of Genes in Many Species, from Yeast to Humans. Fewer Than 10 Years ago no one Knew they Existed". *Scientific American* **Feb**: 56 – 65.

Roberts, R. W. & Szostak, J. W. (1997). "RNA-Peptide Fusions for the *in vitro* Selection of Peptides and Proteins". *Proc. Nat. Acad. Sci. USA.* **94**: 12297 – 12302.

Robles, S. J. & Youvan, D. C. (1993). "Hydropathy and Molar Volume Constraints on Combinatorial Mutants of the Photosynthetic Reaction Center". *J. Mol. Biol.* **232**: 242 – 252.

Röttgen, P. & Collins, J. (1995). "A Human Pancreatic Secretory Trypsin Inhibitor Presenting a Hypervariable Highly Constrained Epitope Via Monovalent Phagemid Display". *Gene* **164**: 243 – 250.

Rungpragayphan, S., Kawarasaki, Y., Imaeda, T., Kohda, K., Nakano, H. & Yamane, T. (2002). "High-throughput, Cloning-independent Protein Library Construction by Combining Single-molecule DNA Amplification with *in Vitro* Expression". *J. Mol. Biol.* **318**: 395 – 405.

Rungpragayphan, S., Nakano, H. & Yamane, T. (2003). "PCR-Linked *in vitro* Expression: A Novel System for High-Throughput Construction and Screening of Protein Libraries". *FEBS Letters* **540**: 147 – 150.

Santoro, S. W., Joyce, G. F., Kandasamy, S., Gramaticova, S. & Barbas, III, C.f. (2000). "RNA cleavage by a DNA Enzyme with Extended Chemical Functionality". *J. Am. Chem. Soc* **122**: 2433 – 2439.

Sassanfar, M. & Roberts, J. W. (1990). "Nature of the SOS Inducing signal in *Escherichia coli* the Involvement of DNA Replication". *J. Mol. Biol.* **212**: 79 – 96.

Sawano, A. & Miyawaki, A. (2000). "Directed Evolution of Green Fluorescent Protein by a New Versatile PCR Strategy for Site Directed and Semi-random Mutagenesis". *Nucl. Acids Res.* **28**: e78 1 - VII

Scala, G., Che, X., Liu, W., Telles, J. T., Cohen, O. J., Vaccarezza, M., Igarashi, T. & Fauci, A. S. (1999). "Selection of HIV-specific Immunogenic Epitopes by Screening Random Peptide Libraries with HIV-1 Positive Sera". *J. Immunology* **162**: 6155 – 6161.

Schlehuber, S., Beste, G. & Skerra, A. (2000). "A Novel Type of Receptor Protein, Based on the Lipocalin Scaffold, with Specificity for Digoxigenin". *J. Mol. Biol.* **297**: 1105 – 1120.

Schlehuber, S. & Skerra, A. (2002). "Tuning Ligand Affinity, Specificity, and Folding Stability of an Engineered Lipocalin Variant – A So Called "Anticalin" – Using a Molecular Random Approach". *Biophysical Chemistry* **96**: 213 – 228.

Schmeideskamp, M. & Klevit, R. E. (1993). "Zinc Finger Diversity". *Current Opinion in Structural Biology*. **4**: 28 – 35.

Schwabe, J. W. R. & Klug, A. (1994). "Zinc Mining for Protein Domains". *Structural Biology*. **1**: 345 – 349.

Scott, J. K. & Smith, G. P. (1990). "Searching for Peptide Ligands with an Epitope Library". *Science* **249**: 386 – 390.

Segal, D. J. & Barbas III, C. F. (2000). "Design of Novel Sequence-Specific DNA-Binding Proteins". *Current Opinion in Chemical Biology* **4**: 34 – 39.

Shafikani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. (1997). "Generation of Large Libraries of Random Mutants in *Bacillus subtilis* by PCR Based Plasmid Multimerization". *Biotechniques* **23**: 304 – 310.

Shi, Y. & Berg, J. M. (1995). "A Direct Comparison of the Properties of Natural and Designed Zinc Finger Proteins". *Chemistry and Biology*. **2**:83 – 89.

Smith, G. P. (1985) "Filamentous Fusion Phage: Novel Expression Vectors that Display Cloned Antigens on the Surface of the Virion". *Science* **228**: 1315 – 1317.

Söderlind, E., Vergeles, M. & Borrebaeck, C. A. K. (1995). "Domain Libraries: Synthetic Diversity for *de novo* Design of Antibody V-Regions". *Gene* **160**: 269 – 272.

Spehr, V., Frahm, D. & Meyer, T. F. (2000). "Improvement of the T7 Expression System by the use of T7 Lysozyme". *Gene* **257**: 259 – 267.

Speight, R. E., Hart, D. J., Sutherland J. D., & Blackburn, J. M. (2001). "A New Plasmid Display Technology for the *in vitro* Selection of Functional Phenotype-Genotype Linked Proteins". *Chemistry and Biology* **8**: 951 – 965.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi,, E. & Inouye, M. (1966) *Cold Spring Harbour Symp. Quant. Biol.* **31**: 77 – 84. (Quoted in Kunkel 1990).

Studier, W. F. & Moffat, B. A. (1986). "Use of Bacteriophage T7 to Direct Selective High Level Expression of Cloned Genes". *J. Mol. Biol.* **189**: 113 – 130.

Takahashi, T.T., Austin, R. J. & Roberts R. W. (2003). "mRNA Display: Ligand Discovery, Interaction Analysis and Beyond". *Trends in Biochemical Sciences* **28**: 159 – 165.

Tseng, T. Y., Frick, D. N. & Richardson, C. C. (2000). "Characterisation of a Novel DNA Primase from the *Salmonella typhimurium* Bacteriophage SP6". *Biochemistry* **39**: 1643 – 1654.

Turpin, J. A., Song, Y., Inman, J. K., Huang, M., Wallquist, A., Maynard, A., Covell, D. G., Rice, W. G. & Appella, E. (1999). "Synthesis and Biological Properties of Novel Pyridinioalkanoyl Thiolesters (PATE) as Anti-HIV-1 Agents That Target the Viral Nucleocapsid Protein Zinc Fingers". *J. Med. Chem.* **42**: 67 – 68.

Virnekäs, B., Ge, L., Plückthun, A., Schneider, K. C., Wellnhofer, G. & Moroney, S. E. (1994). "Trinucleotide Phosphoramidites: Ideal Reagents for the Synthesis of Mixed Oligonucleotides for Random Mutagenesis". *Nucl. Acids Res.* **23**: 5600 – 5607.

Warren, M. S. & Benkovic, S. J. (1997). Combinatorial Manipulation of Three Key Active Site Residues in Glycinamide Ribonucleotide Transformylase". *Protein Engineering* **10**: 63 – 68.

Wieczorek, E., Lin, Z., Perkins, E. B., Law, D. J., Merchant, J. L. &Zehner, Z. E. (2000). "The Zinc finger Repressor ZBP-89 Binds to the Silencer Element of the Human Vimentin Gene and Complexes with the Transcriptional Activator SP1". *Journal of Biological Chemistry* **275**: 12879 – 12888.

Wittrup, K. D. (2001). "Protein Engineering by Cell Surface Display". *Current Opinion in Biotechnology* **12**: 395 – 399.

Wolfe, S. A., Greisman, H. A., Ramm, E. I. & Pabo, C. O. (1999). "Analysis of Zinc Fingers optimized *Via* Phage Display: Evaluating the Utility of a Recognition Code". *J. Mol. Biol.* **285**: 1917 – 1934.

Wolfe, S. A., Grant, R. A., Elrod-Erickson, M. & Pabo, C. O. (2001). "Beyond the "Recognition Code": Structures of Two Cys_2His_2 Zinc Finger/TATA Box Complexes". *Structure* **9**: 717 – 723.

Xu, H., Petersen, E. I., Petersen, S. B. & El-Gewely, M. R. (1999). "Random Mutagenesis Libraries: Optimization and Simplification by PCR". *Biotechniques* **27**: 1102 – 1108.

Yang, M., May, W. S. & Ito, T. (1999). "Jaz Requires the Double-Stranded RNA-Binding Zinc Finger Motifs for Nuclear Localization". *Journal of Biological Chemistry* **274**: 27399 – 27406.

Yang, W-P., Green, K., Pinz-Sweeney, S., Briones, A. T., Burton, D. R. & Barbas III, C. F. (1995). "CDR Walking Mutagenesis for the Affinity Maturation of a Potent Human Anti-HIV-1 Antibody into the Picomolar Range. *J. Mol. Biol.* **254**: 392 – 403.

Yang, W. (2000). "Structure and Function of Mismatch Repair Proteins". *Mutation research* **460**: 245 – 256.

Ye, B. H., Lista, F., Lo-Coco, F., Knowles, D. M., Offit, K., Changanti, R. S. K. & Dalla-Favera, R. (1993). "Alterations of a Zinc Finger Encoding Gene, BCL6, in Diffuse Large-Cell Lymphoma". *Science* **262**: 747 – 750.

Zhang, L., Spratt, S. K., Liu, Q., Johnstone, B., Qi, H., Raschke, E. E., Jamieson, A. C., Rebar, E. J., Wolffe, A. P. & Case, C. C. (2000). "Synthetic Zinc Finger Transcription Factor Action At an Endogenous Chromosomal Site". *Journal of Biological Chemistry* **275**: 33850 – 33860.

Zwick, M.B., Bonnycastle, L. L. C., Noren, K. A., Venturini, S., Leong, E., Barbas, C. F., Noren, C. J. & Scott, J. K. (1998). "The Maltose Scaffold Protein for Monovalent Display of Peptides Derived From Phage Libraries". *Analytical Biochemistry* **264**: 87 – 97.

Appendix

A1) Oligonucleotide Sequences.

pGEX 5'	5'-GGGCTGGCAAGCCACGTTTGGTG-3'
pGEX 3'	5'-CCGGGAGCTGCATGTGTCAGAGG-3'

The pGEX 5' and pGEX 3' primers were employed in the PCR amplification of all genes contained within the pGEX-2TK plasmid. Both primers have a theoretical melting temperature of 67.8°C. When employed in PCR reactions the temperature during the annealing cycle of the PCR was maintained at 62°C.

T7 Promoter	5'-TAATACGACTCACTATAGGG-3'
T7 Terminator	5'-GCTAGTTATTGCTCAGCGG-3'

The T7 promoter and T7 terminator primers were used in the PCR amplification of all genes contained within the pET-42a plasmid. The T7 Promoter primer has a theoretical melting temperature of 53.2°C and the T7 Terminator primer has a theoretical melting temperature of 56.7°C. Annealing temperatures in PCR reactions performed using these primers were maintained at 55°C.

T7 Close Forward	5'-AGCATGGCCTTTGCAGGG-3'
T7 Close Reverse	5'-CGTCCCATTCGCCAATCC-3'

The T7 Close forward and reverse primers were used in the sequencing of all genes contained within the pET-42a plasmid.

pGEX Close Forward	5′-TAGCATGGCCTTTGCAGG-3′
pGEX Close Reverse	5′-GTGTCAGAGGTTTTCACC-3′

The pGEX close forward and reverse primers were used in the sequencing of all genes contained within the pGEX-2TK plasmid.

Appendix

A2) Sequences of the ZFHM6, ZFMA3, ZFDN1 and Genes

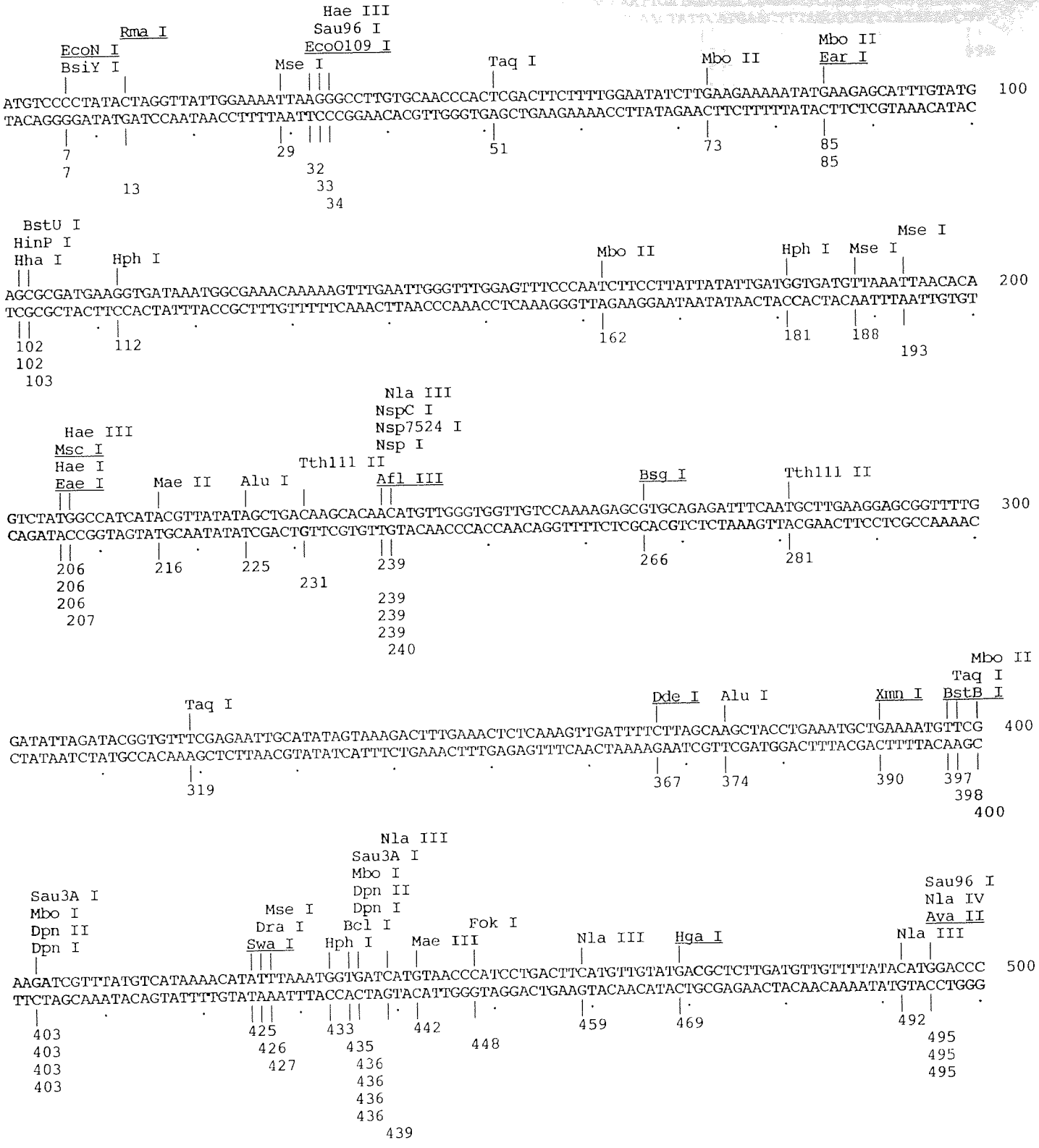
The genes are shown as fusions with the GST sequence of the pGEX-2TK plasmid (ZFHM6, ZFMA3, ZFDN1) and pET-42a plasmid (ZFDN1). The sequences are shown both as restriction maps of the double stranded DNA sequence and also translated into their respective amino acid sequences.

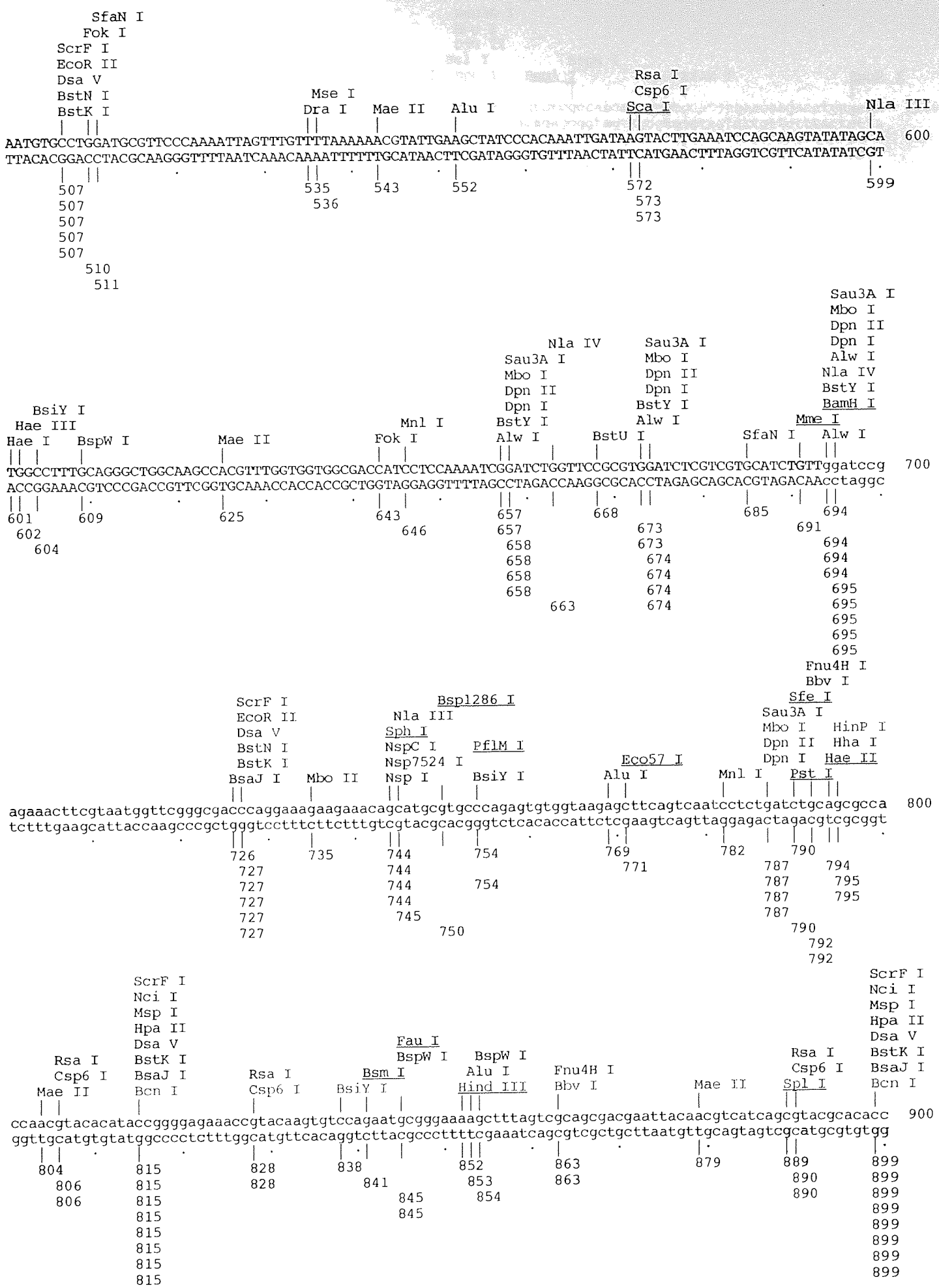
ZFHM6 Sequence

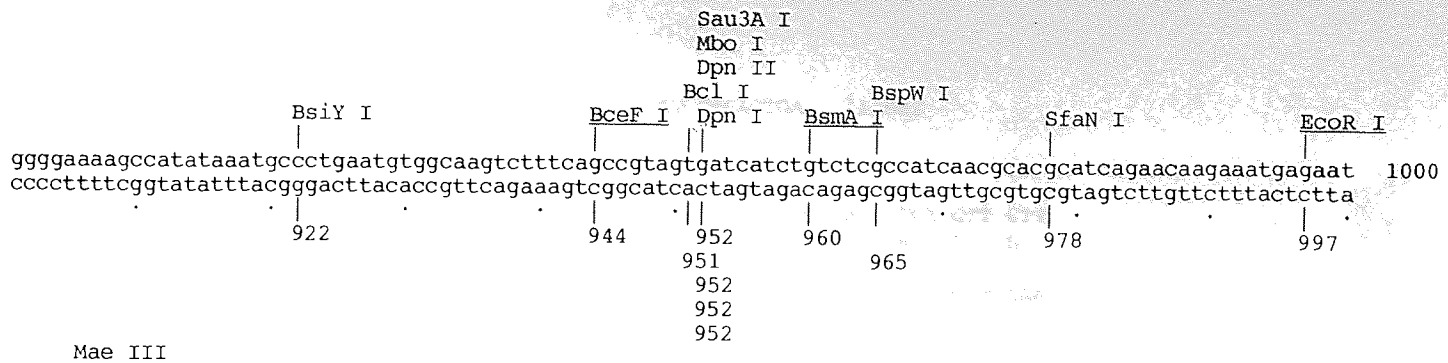
Libgene - GST fusion -> Full Restriction Map

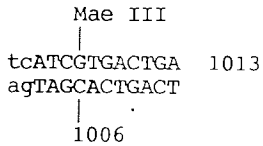
DNA sequence 1013 b.p. ATGTCCCCTATA ... cATCGTGACTGA linear

Positions of Restriction Endonucleases sites (unique sites underlined)




SfaN I
Fok I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Mse I
Dra I
Mae II
Alu I
Rsa I
Csp6 I
Sca I
Nla III
AATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTTAAAAAACGTATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATAGCA 600
TTACACGGACCTACGCAAGGGTTTTAATCAAACAAAATTTTTTGCATAACTTCGATAGGGTGTTTAACTATTCATGAACTTTAGGTCGTTCATATATCGT
507 535 543 552 572 599
507 536 573
507 573
507
507
507
510
511
Nla IV
Sau3A I
Mbo I
Dpn II
Dpn I
BstY I
Alw I
Sau3A I
Mbo I
Dpn II
Dpn I
BstY I
Alw I
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
Nla IV
BstY I
BamH I
BsiY I
Hae III
Hae I
BspW I
Mae II
Mnl I
Fok I
BstU I
SfaN I
Mme I
Alw I
TGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAATCGGATCTGGTTCCGCGTGGATCTCGTCGTGCATCTGTTggatccg 700
ACCGGAAACGTCCCGACCGTTCGGTGCAAACCACCACCGCTGGTAGGAGGTTTTAGCCTAGACCAAGGCGCACCTAGAGCAGCACGTAGACAAcctaggc
601 609 625 643 646 657 657 658 658 658 658 663 668 673 673 674 674 674 674 685 691 694 694 694 694 695 695 695 695 695
602
604
Fnu4H I
Bbv I
Sfe I
Sau3A I
Mbo I
Dpn II
Dpn I
HinP I
Hha I
Hae II
Pst I
Eco57 I
Alu I
Mnl I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Bsp1286 I
Nla III
Sph I
NspC I
Nsp7524 I
PflM I
BsaJ I
Mbo II
Nsp I
BsiY I
agaaacttcgtaatggttcgggcgacccaggaaagaagaaacagcatgcgtgcccagagtgtggtaagagcttcagtcaatcctctgatctgcagcgcca 800
tctttgaagcattaccaagcccgctgggtcctttcttctttgtcgtacgcacgggtctcacaccattctcgaagtcagttaggagactagacgtcgcggt
726 735 744 754 769 771 782 787 787 787 787 790 790 792 792 794 795 795
727 744 754
727 744
727 744
727 745
727 750
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
BsaJ I
Bcn I
Rsa I
Csp6 I
Mae II
Rsa I
Csp6 I
Fau I
BspW I
Bsm I
BsiY I
BspW I
Alu I
Hind III
Fnu4H I
Bbv I
Mae II
Rsa I
Csp6 I
Spl I
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
BsaJ I
Bcn I
ccaacgtacacatacccggggagaaaccgtacaagtgtccagaatgcgggaaaagctttagtcgcagcgacgaattacaacgtcatcagcgtacgcacacc 900
ggttgcatgtgtatggcccctctttggcatgttcacaggtcttacgccctttttcgaaatcagcgtcgctgcttaatgttgcagtagtcgcatgcgtgtgg
804 815 828 838 841 845 845 852 853 854 863 863 879 889 890 890 899 899 899 899 899 899 899 899 899
806 815 828
806 815
815
815
815
815
815





Restriction Endonucleases site usage

Aat II	-	Bsm I	1	Eco57 I	1	Mcr I	-	Rsr II	-
Acc I	-	BsmA I	1	EcoN I	1	Mlu I	-	Sac I	-
Afl II	-	Bsp120 I	-	Eco0109 I	1	Mme I	1	Sac II	-
Afl III	1	Bsp1286 I	1	EcoR I	1	Mnl I	2	Sal I	-
Age I	-	BspE I	-	EcoR II	2	Msc I	1	Sap I	-
Aha II	-	BspH I	-	EcoR V	-	Mse I	5	Sau3A I	7
Alu I	5	BspM I	-	Ehe I	-	Msp I	2	Sau96 I	2
Alw I	4	BspW I	4	Esp I	-	Nae I	-	Sca I	1
AlwN I	-	Bsr I	-	Fau I	1	Nar I	-	ScrF I	4
Apa I	-	BssH II	-	Fse I	-	Nci I	2	SfaN I	3
ApaL I	-	BstB I	1	Fnu4H I	2	Nco I	-	Sfe I	1
Ase I	-	BstE II	-	Fok I	3	Nde I	-	Sfi I	-
Asp718	-	BstK I	4	Fsp I	-	Nhe I	-	SgrA I	-
Ava I	-	BstN I	2	Gdi II	-	Nla III	6	Sma I	-
Ava II	1	BstU I	2	Gsu I	-	Nla IV	3	Sna I	-
Avr II	-	BstX I	-	Hae I	2	Not I	-	SnaB I	-
BamH I	1	BstY I	3	Hae II	1	Nru I	-	Spe I	-
Ban I	-	Bsu36 I	-	Hae III	3	Nsi I	-	Sph I	1
Ban II	-	Cfr10 I	-	Hga I	1	Nsp I	2	Spl I	1
Bbe I	-	Cla I	-	HgiA I	-	Nsp7524 I	2	Sse8337 I	-
Bbs I	-	Csp6 I	4	Hha I	2	NspB II	-	Ssp I	-
Bbv I	2	Dde I	1	Hinc II	-	NspC I	2	Stu I	-
BceF I	1	Dpn I	7	Hind III	1	Pac I	-	Sty I	-
Bcl I	2	Dpn II	7	Hinf I	-	PaeR7 I	-	Swa I	1
Bcn I	2	Dra I	2	HinP I	2	PflM I	1	Taq I	3
Bgl I	-	Dra III	-	Hpa I	-	Ple I	-	Tfi I	-
Bgl II	-	Drd I	-	Hpa II	2	Pml I	-	Tth111 I	-
Bsa I	-	Dsa I	-	Hph I	3	PpuM I	-	Tth111 II	2
BsaA I	-	Dsa V	4	Kas I	-	PshA I	-	Xba I	-
BsaB I	-	Eae I	1	Kpn I	-	Pst I	1	Xca I	-
BsaJ I	3	Eag I	-	Mae II	5	Pvu I	-	Xcm I	-
Bsg I	1	Ear I	1	Mae III	2	Pvu II	-	Xho I	-
BsiE I	-	Ecl136 I	-	Mbo I	7	Rma I	1	Xma I	-
BsiY I	5	Eco47 III	-	Mbo II	5	Rsa I	4	Xmn I	1

Enzyme	Site		Use	Site position (Fragment length) Fragment order		
Afl III	a/crygt		1	1(238) 2	239(775) 1	
Ava II	g/gwcc		1	1(494) 2	495(519) 1	
BamH I	g/gatcc		1	1(693) 1	694(320) 2	
BceF I	acggc	12/13	1	1(943) 1	944(70) 2	
Bsg I	gtgcag	16/14	1	1(265) 2	266(748) 1	
Bsm I	gaatgc	1/-1	1	1(840) 1	841(173) 2	
BsmA I	gtctc	1/5	1	1(959) 1	960(54) 2	
Bsp1286 I	gdgch/c		1	1(749) 1	750(264) 2	
BstB I	tt/cgaa		1	1(396) 2	397(617) 1	
Dde I	c/tnag		1	1(366) 2	367(647) 1	
Eae I	y/ggccr		1	1(205) 2	206(808) 1	
Ear I	ctcttc	1/4	1	1(84) 2	85(929) 1	
Eco57 I	ctgaag	16/14	1	1(770) 1	771(243) 2	
EcoN I	cctnn/nnnagg		1	1(6) 2	7(1007) 1	
Eco0109 I	rg/gnccy		1	1(31) 2	32(982) 1	
EcoR I	g/aattc		1	1(996) 1	997(17) 2	
Fau I	cccgc	4/6	1	1(844) 1	845(169) 2	
Hae II	rgcgc/y		1	1(793) 1	794(220) 2	
Hga I	gacgc	5/10	1	1(468) 2	469(545) 1	
Hind III	a/agctt		1	1(851) 1	852(162) 2	

Libgene - GST fusion -> 1-phase Translation

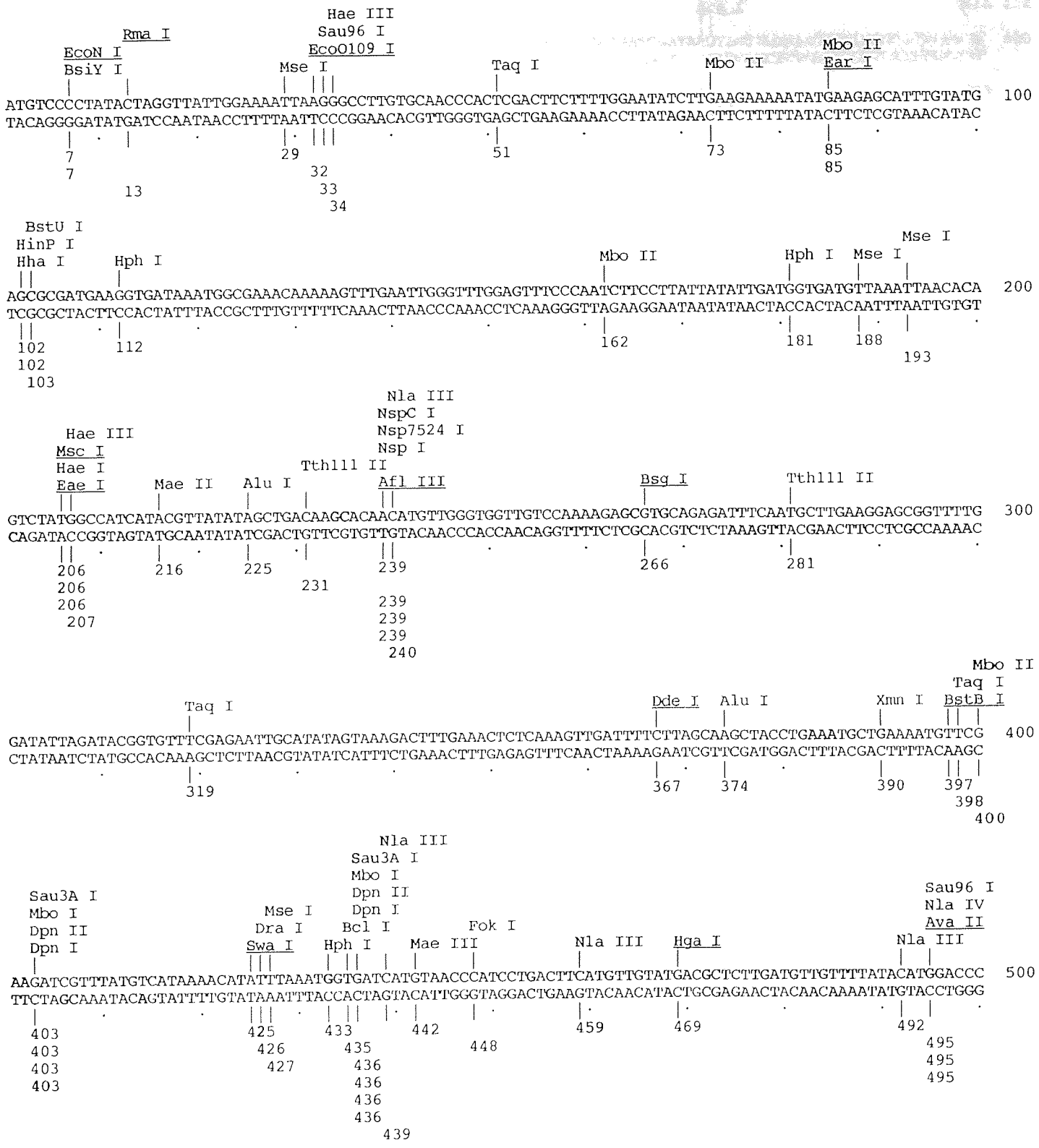
DNA sequence 1013 b.p. ATGTCCCCTATA ... cATCGTGACTGA linear

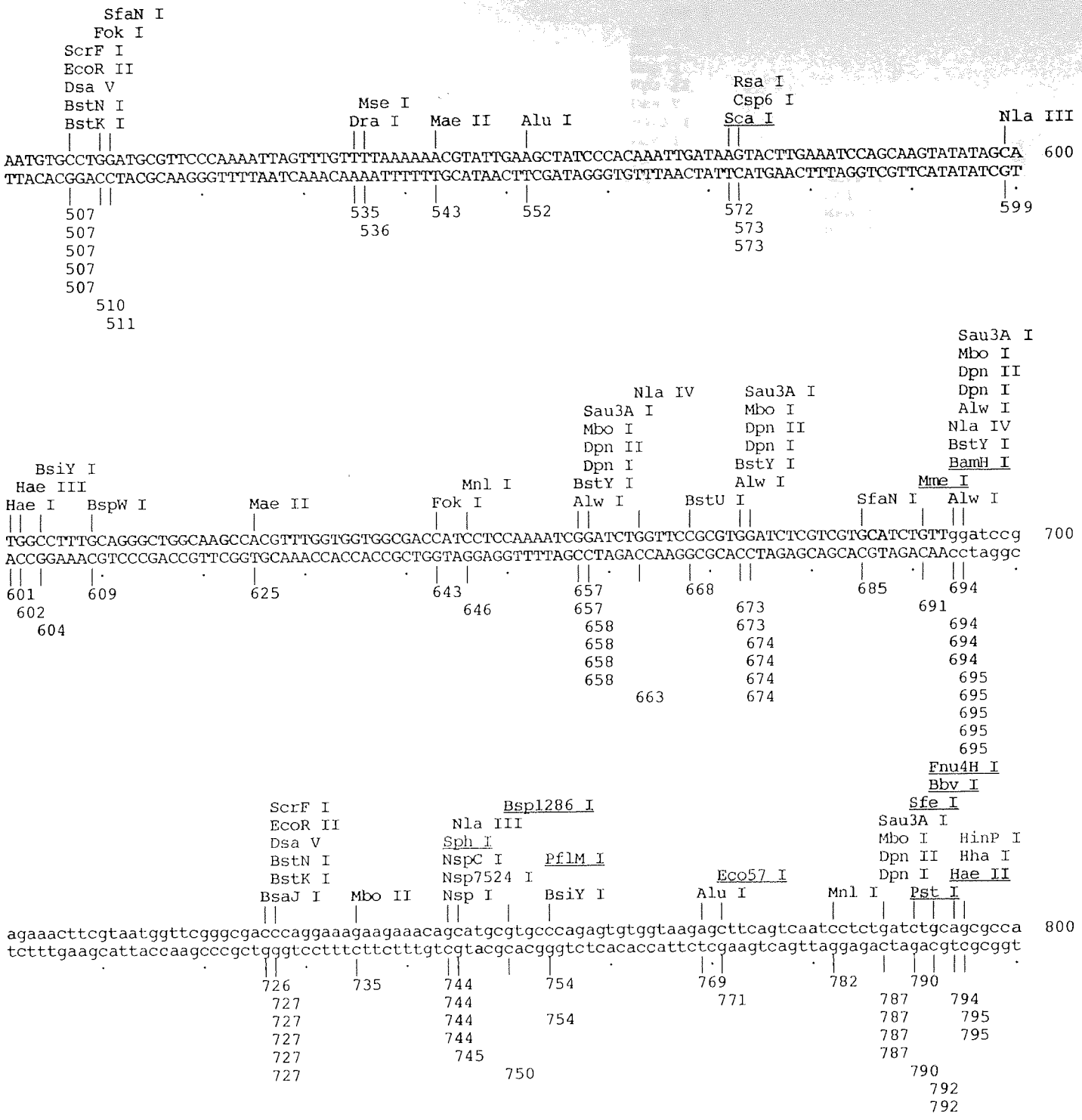
```
1/1                                             31/11
ATG TCC CCT ATA CTA GGT TAT TGG AAA ATT AAG GGC CTT GTG CAA CCC ACT CGA CTT CTT
M   S   P   I   L   G   Y   W   K   I   K   G   L   V   Q   P   T   R   L   L
61/21                                           91/31
TTG GAA TAT CTT GAA GAA AAA TAT GAA GAG CAT TTG TAT GAG CGC GAT GAA GGT GAT AAA
L   E   Y   L   E   E   K   Y   E   E   H   L   Y   E   R   D   E   G   D   K
121/41                                          151/51
TGG CGA AAC AAA AAG TTT GAA TTG GGT TTG GAG TTT CCC AAT CTT CCT TAT TAT ATT GAT
W   R   N   K   K   F   E   L   G   L   E   F   P   N   L   P   Y   Y   I   D
181/61                                          211/71
GGT GAT GTT AAA TTA ACA CAG TCT ATG GCC ATC ATA CGT TAT ATA GCT GAC AAG CAC AAC
G   D   V   K   L   T   Q   S   M   A   I   I   R   Y   I   A   D   K   H   N
241/81                                          271/91
ATG TTG GGT GGT TGT CCA AAA GAG CGT GCA GAG ATT TCA ATG CTT GAA GGA GCG GTT TTG
M   L   G   G   C   P   K   E   R   A   E   I   S   M   L   E   G   A   V   L
301/101                                         331/111
GAT ATT AGA TAC GGT GTT TCG AGA ATT GCA TAT AGT AAA GAC TTT GAA ACT CTC AAA GTT
D   I   R   Y   G   V   S   R   I   A   Y   S   K   D   F   E   T   L   K   V
361/121                                         391/131
GAT TTT CTT AGC AAG CTA CCT GAA ATG CTG AAA ATG TTC GAA GAT CGT TTA TGT CAT AAA
D   F   L   S   K   L   P   E   M   L   K   M   F   E   D   R   L   C   H   K
421/141                                         451/151
ACA TAT TTA AAT GGT GAT CAT GTA ACC CAT CCT GAC TTC ATG TTG TAT GAC GCT CTT GAT
T   Y   L   N   G   D   H   V   T   H   P   D   F   M   L   Y   D   A   L   D
481/161                                         511/171
GTT GTT TTA TAC ATG GAC CCA ATG TGC CTG GAT GCG TTC CCA AAA TTA GTT TGT TTT AAA
V   V   L   Y   M   D   P   M   C   L   D   A   F   P   K   L   V   C   F   K
541/181                                         571/191
AAA CGT ATT GAA GCT ATC CCA CAA ATT GAT AAG TAC TTG AAA TCC AGC AAG TAT ATA GCA
K   R   I   E   A   I   P   Q   I   D   K   Y   L   K   S   S   K   Y   I   A
601/201                                         631/211
TGG CCT TTG CAG GGC TGG CAA GCC ACG TTT GGT GGT GGC GAC CAT CCT CCA AAA TCG GAT
W   P   L   Q   G   W   Q   A   T   F   G   G   G   D   H   P   P   K   S   D
661/221                                         691/231
CTG GTT CCG CGT GGA TCT CGT CGT GCA TCT GTT gga tcc gag aaa ctt cgt aat ggt tcg
L   V   P   R   G   S   R   R   A   S   V   G   S   E   K   L   R   N   G   S
721/241                                         751/251
ggc gac cca gga aag aag aaa cag cat gcg tgc cca gag tgt ggt aag agc ttc agt caa
G   D   P   G   K   K   K   Q   H   A   C   P   E   C   G   K   S   F   S   Q
781/261                                         811/271
tcc tct gat ctg cag cgc cac caa cgt aca cat acc ggg gag aaa ccg tac aag tgt cca
S   S   D   L   Q   R   H   Q   R   T   H   T   G   E   K   P   Y   K   C   P
841/281                                         871/291
gaa tgc ggg aaa agc ttt agt cgc agc gac gaa tta caa cgt cat cag cgt acg cac acc
E   C   G   K   S   F   S   R   S   D   E   L   Q   R   H   Q   R   T   H   T
901/301                                         931/311
ggg gaa aag cca tat aaa tgc cct gaa tgt ggc aag tct ttc agc cgt agt gat cat ctg
G   E   K   P   Y   K   C   P   E   C   G   K   S   F   S   R   S   D   H   L
961/321                                         991/331
tct cgc cat caa cgc acg cat cag aac aag aaa tga gaa ttc ATC GTG ACT GA
S   R   H   Q   R   T   H   Q   N   K   K   *   E   F   I   V   T
```

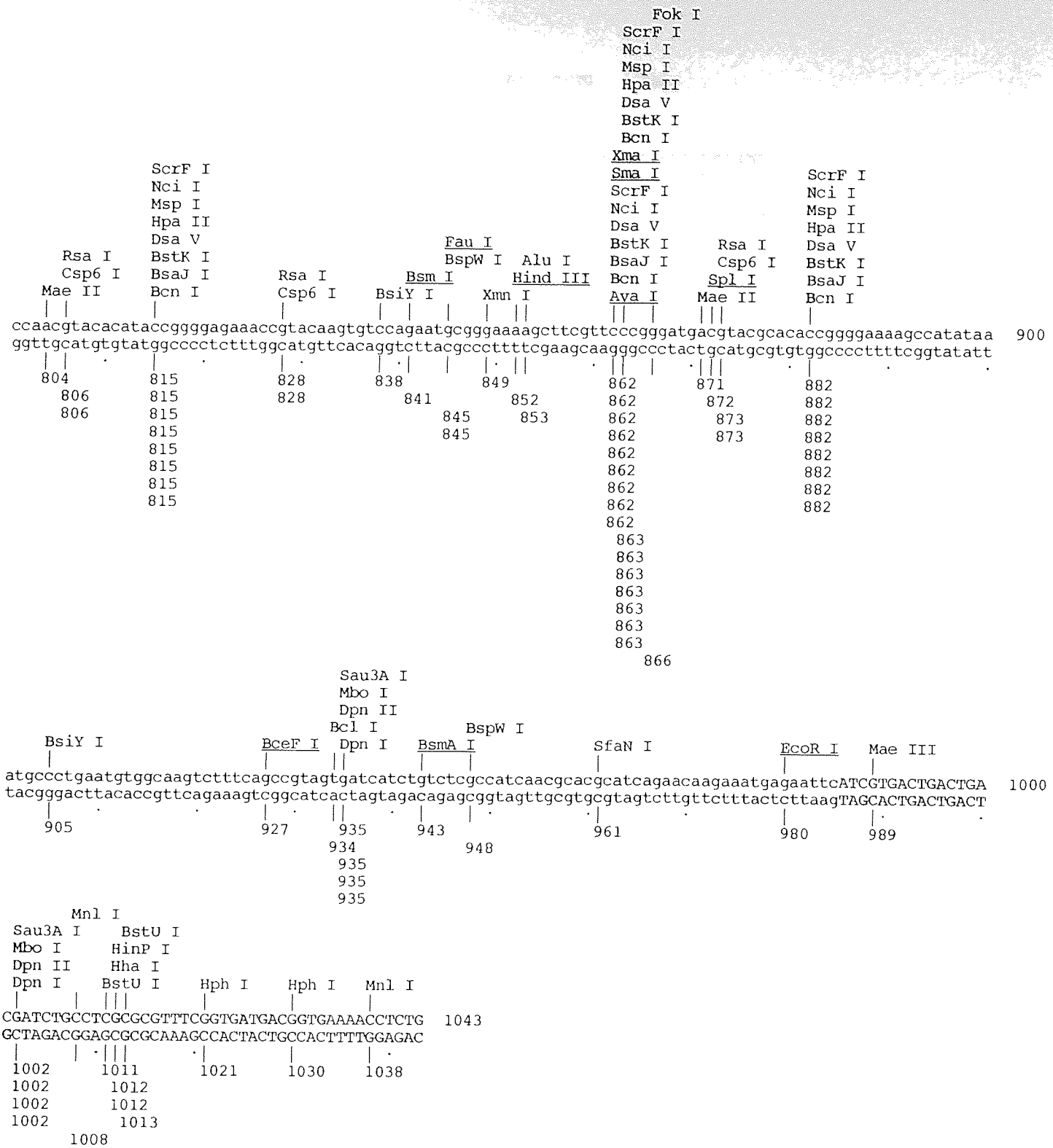
ZFMA3 Sequence

ma3gst -> Restriction Map

DNA sequence 1043 b.p. ATGTCCCCTATA ... TGAAAACCTCTG linear




SfaN I
Fok I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
Mse I
Dra I
Mae II
Alu I
Rsa I
Csp6 I
Sca I
Nla III
AATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTTAAAAAACGTATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATAGCA 600
TTACACGGACCTACGCAAGGGTTTTAATCAAACAAAATTTTTTGCATAACTTCGATAGGGTGTTTAACTATTCATGAACTTTAGGTCGTTCATATATCGT
507
507
507
507
507
510
511
535
536
543
552
572
573
573
599
Sau3A I
Mbo I
Dpn II
Dpn I
Alw I
Nla IV
BstY I
BamH I
Mme I
BsiY I
Hae III
Hae I
BspW I
Mae II
Mnl I
Fok I
BstU I
SfaN I
TGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAATCGGATCTGGTTCCGCGTGGATCTCGTCGTGCATCTGTTggatccg 700
ACCGGAAACGTCCCGACCGTTCGGTGCAAACCACCACCGCTGGTAGGAGGTTTTAGCCTAGACCAAGGCGCACCTAGAGCAGCACGTAGACAAcctaggc
601
602
604
609
625
643
646
657
657
658
658
658
658
663
668
673
673
674
674
674
674
685
691
694
694
694
694
695
695
695
695
695
Fnu4H I
Bbv I
Sfe I
Sau3A I
Mbo I
Dpn II
Dpn I
HinP I
Hha I
Hae II
Pst I
ScrF I
EcoR II
Dsa V
BstN I
BstK I
BsaJ I
Mbo II
Bsp1286 I
Nla III
Sph I
NspC I
Nsp7524 I
Nsp I
PflM I
BsiY I
Eco57 I
Alu I
Mnl I
agaaacttcgtaatggttcgggcgacccaggaaagaagaaacagcatgcgtgcccagagtgtggtaagagcttcagtcaatcctctgatctgcagcgcca 800
tctttgaagcattaccaagcccgctgggtcctttcttctttgtcgtacgcacgggtctcacaccattctcgaagtcagttaggagactagacgtcgcggt
726
727
727
727
727
727
735
744
744
744
744
745
750
754
754
769
771
782
787
787
787
787
790
790
792
792
794
795
795


Fok I
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
Bcn I
Xma I
Sma I
ScrF I
Nci I
Dsa V
BstK I
BsaJ I
Bcn I
Ava I
Rsa I
Csp6 I
Spl I
Mae II
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
BsaJ I
Bcn I
ScrF I
Nci I
Msp I
Hpa II
Dsa V
BstK I
BsaJ I
Bcn I
Fau I
BspW I
Alu I
Bsm I
Hind III
Xmn I
Rsa I
Csp6 I
Mae II
BsiY I
ccaacgtacacataccggggagaaaccgtacaagtgtccagaatgcgggaaaagcttcgttcccgggatgacgtacgcacaccggggaaaagccatataa 900
ggttgcatgtgtatggcccctctttggcatgttcacaggtcttacgccctttcgaagcaagggccctactgcatgcgtgtggcccctttcggtatatt
804
806
806
815
815
815
815
815
815
815
828
828
838
841
845
845
849
852
853
862
862
862
862
862
862
862
862
862
862
863
863
863
863
863
863
863
863
866
871
872
873
873
882
882
882
882
882
882
882
882
Sau3A I
Mbo I
Dpn II
Bcl I
BspW I
BsiY I
BceF I
Dpn I
BsmA I
SfaN I
EcoR I
Mae III
atgccctgaatgtggcaagtctttcagccgtagtgatcatctgtctcgccatcaacgcacgcatcagaacaagaaatgagaattcATCGTACTGACTGACTGA 1000
tacgggacttacaccgttcagaaagtcggcatcactagtagacagagcggtagttgcgtgcgtagtcttgttctttactcttaagTAGCACTGACTGACT
905
927
934
935
935
935
935
943
948
961
980
989
Mnl I
Sau3A I
Mbo I
Dpn II
Dpn I
BstU I
HinP I
Hha I
BstU I
Hph I
Hph I
Mnl I
CGATCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTG 1043
GCTAGACGGAGCGCGCAAAGCCACTACTGCCACTTTTGGAGAC
1002
1002
1002
1002
1008
1011
1012
1012
1013
1021
1030
1038

ma3gst -> 1-phase Translation

DNA sequence 1043 b.p. ATGTCCCCTATA ... TGAAAACCTCTG linear

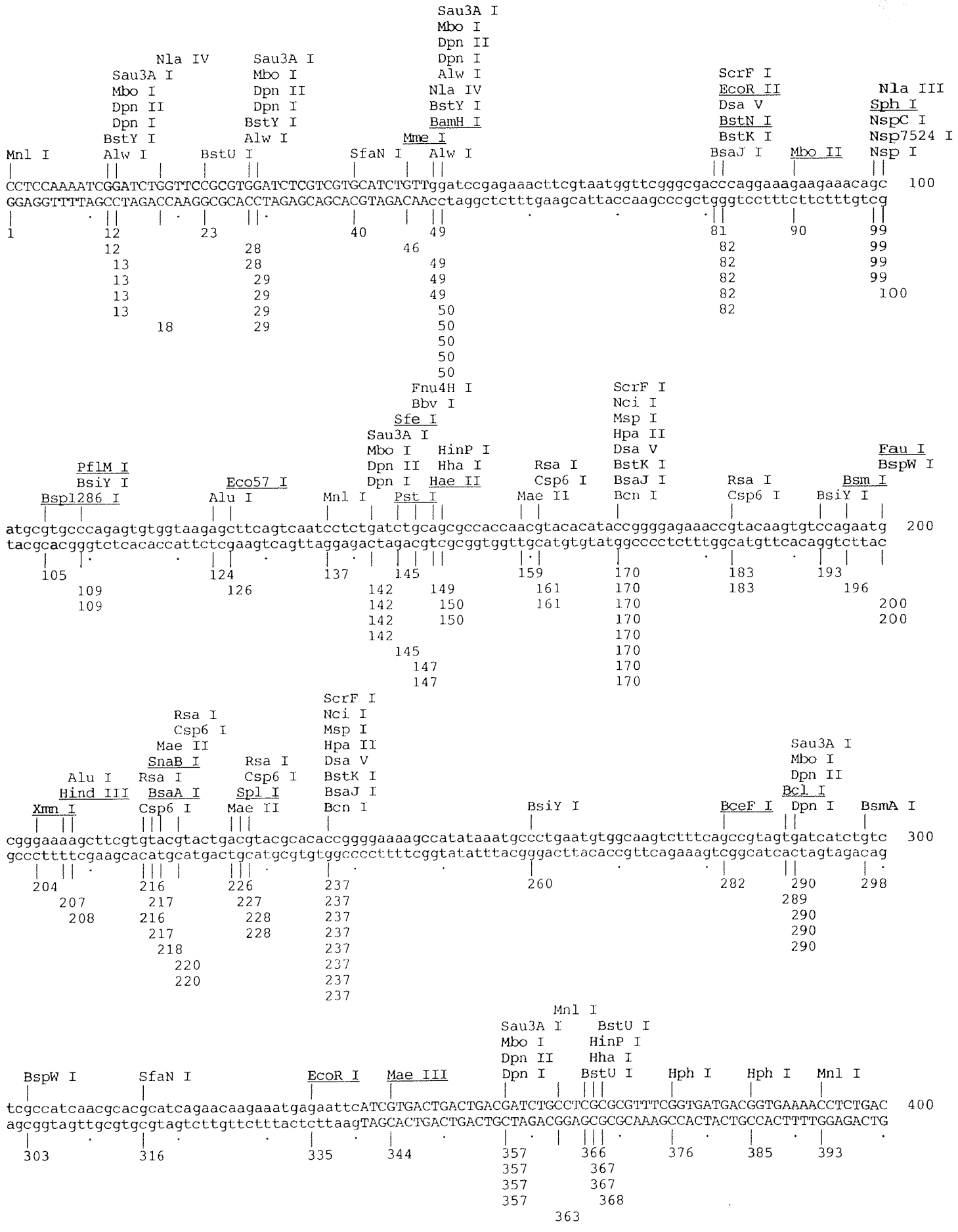
```
1/1                                             31/11
ATG TCC CCT ATA CTA GGT TAT TGG AAA ATT AAG GGC CTT GTG CAA CCC ACT CGA CTT CTT
M   S   P   I   L   G   Y   W   K   I   K   G   L   V   Q   P   T   R   L   L
61/21                                           91/31
TTG GAA TAT CTT GAA GAA AAA TAT GAA GAG CAT TTG TAT GAG CGC GAT GAA GGT GAT AAA
L   E   Y   L   E   E   K   Y   E   E   H   L   Y   E   R   D   E   G   D   K
121/41                                          151/51
TGG CGA AAC AAA AAG TTT GAA TTG GGT TTG GAG TTT CCC AAT CTT CCT TAT TAT ATT GAT
W   R   N   K   K   F   E   L   G   L   E   F   P   N   L   P   Y   Y   I   D
181/61                                          211/71
GGT GAT GTT AAA TTA ACA CAG TCT ATG GCC ATC ATA CGT TAT ATA GCT GAC AAG CAC AAC
G   D   V   K   L   T   Q   S   M   A   I   I   R   Y   I   A   D   K   H   N
241/81                                          271/91
ATG TTG GGT GGT TGT CCA AAA GAG CGT GCA GAG ATT TCA ATG CTT GAA GGA GCG GTT TTG
M   L   G   G   C   P   K   E   R   A   E   I   S   M   L   E   G   A   V   L
301/101                                         331/111
GAT ATT AGA TAC GGT GTT TCG AGA ATT GCA TAT AGT AAA GAC TTT GAA ACT CTC AAA GTT
D   I   R   Y   G   V   S   R   I   A   Y   S   K   D   F   E   T   L   K   V
361/121                                         391/131
GAT TTT CTT AGC AAG CTA CCT GAA ATG CTG AAA ATG TTC GAA GAT CGT TTA TGT CAT AAA
D   F   L   S   K   L   P   E   M   L   K   M   F   E   D   R   L   C   H   K
421/141                                         451/151
ACA TAT TTA AAT GGT GAT CAT GTA ACC CAT CCT GAC TTC ATG TTG TAT GAC GCT CTT GAT
T   Y   L   N   G   D   H   V   T   H   P   D   F   M   L   Y   D   A   L   D
481/161                                         511/171
GTT GTT TTA TAC ATG GAC CCA ATG TGC CTG GAT GCG TTC CCA AAA TTA GTT TGT TTT AAA
V   V   L   Y   M   D   P   M   C   L   D   A   F   P   K   L   V   C   F   K
541/181                                         571/191
AAA CGT ATT GAA GCT ATC CCA CAA ATT GAT AAG TAC TTG AAA TCC AGC AAG TAT ATA GCA
K   R   I   E   A   I   P   Q   I   D   K   Y   L   K   S   S   K   Y   I   A
601/201                                         631/211
TGG CCT TTG CAG GGC TGG CAA GCC ACG TTT GGT GGT GGC GAC CAT CCT CCA AAA TCG GAT
W   P   L   Q   G   W   Q   A   T   F   G   G   G   D   H   P   P   K   S   D
661/221                                         691/231
CTG GTT CCG CGT GGA TCT CGT CGT GCA TCT GTT gga tcc gag aaa ctt cgt aat ggt tcg
L   V   P   R   G   S   R   R   A   S   V   G   S   E   K   L   R   N   G   S
721/241                                         751/251
ggc gac cca gga aag aag aaa cag cat gcg tgc cca gag tgt ggt aag agc ttc agt caa
G   D   P   G   K   K   K   Q   H   A   C   P   E   C   G   K   S   F   S   Q
781/261                                         811/271
tcc tct gat ctg cag cgc cac caa cgt aca cat acc ggg gag aaa ccg tac aag tgt cca
S   S   D   L   Q   R   H   Q   R   T   H   T   G   E   K   P   Y   K   C   P
841/281                                         871/291
gaa tgc ggg aaa agc ttc gtt ccc ggg atg acg tac gca cac cgg gga aaa gcc ata taa
E   C   G   K   S   F   V   P   G   M   T   Y   A   H   R   G   K   A   I   *
901/301                                         931/311
atg ccc tga atg tgg caa gtc ttt cag ccg tag tga tca tct gtc tcg cca tca acg cac
M   P   *   M   W   Q   V   F   Q   P   *   *   S   S   V   S   P   S   T   H
961/321                                         991/331
gca tca gaa caa gaa atg aga att cAT CGT GAC TGA CTG ACG ATC TGC CTC GCG CGT TTC
A   S   E   Q   E   M   R   I   H   R   D   *   L   T   I   C   L   A   R   F
1021/341
GGT GAT GAC GGT GAA AAC CTC TG
G   D   D   G   E   N   L
```

ZFDN1 Sequence

dn1gsttrans -> Full Restriction Map

DNA sequence 421 b.p. CCTCCAAAATCG ... TCCCGGAGACGG linear

Positions of Restriction Endonucleases sites (unique sites underlined)



```
          ScrF I
          Nci I
          Dsa V
       Alu I
     Fnu4H I   BsmA I
     Bbv I  Msp I
  Nla III  BstK I
 NspC I    BsiY I
 Nsp7524 I  Hpa II
 Nsp I     Bcn I
 ||  | |   ||  |
 ACATGCAGCTCCCGGAGACGG  421
 TGTACGTCGAGGGCCTCTGCC
 ||  | |  .||  |    .
 401       411
 401        412
 401       411
  402      411
     405    412
     405       415
       407
          411
          411
          411
```

Restriction Endonucleases site usage

Aat II	-	Bsm I	1	Eco57 I	1	Mcr I	-	Rsr II	-
Acc I	-	BsmA I	2	EcoN I	-	Mlu I	-	Sac I	-
Afl II	-	Bsp120 I	-	Eco0109 I	-	Mme I	1	Sac II	-
Afl III	-	Bsp1286 I	1	EcoR I	1	Mnl I	4	Sal I	-
Age I	-	BspE I	-	EcoR II	1	Msc I	-	Sap I	-
Aha II	-	BspH I	-	EcoR V	-	Mse I	-	Sau3A I	6
Alu I	3	BspM I	-	Ehe I	-	Msp I	3	Sau96 I	-
Alw I	4	BspW I	2	Esp I	-	Nae I	-	Sca I	-
AlwN I	-	Bsr I	-	Fau I	1	Nar I	-	ScrF I	4
Apa I	-	BssH II	-	Fse I	-	Nci I	3	SfaN I	2
ApaL I	-	BstB I	-	Fnu4H I	2	Nco I	-	Sfe I	1
Ase I	-	BstE II	-	Fok I	-	Nde I	-	Sfi I	-
Asp718	-	BstK I	4	Fsp I	-	Nhe I	-	SgrA I	-
Ava I	-	BstN I	1	Gdi II	-	Nla III	2	Sma I	-
Ava II	-	BstU I	3	Gsu I	-	Nla IV	2	Sna I	-
Avr II	-	BstX I	-	Hae I	-	Not I	-	SnaB I	1
BamH I	1	BstY I	3	Hae II	1	Nru I	-	Spe I	-
Ban I	-	Bsu36 I	-	Hae III	-	Nsi I	-	Sph I	1
Ban II	-	Cfr10 I	-	Hga I	-	Nsp I	2	Spl I	1
Bbe I	-	Cla I	-	HgiA I	-	Nsp7524 I	2	Sse8337 I	-
Bbs I	-	Csp6 I	5	Hha I	2	NspB II	-	Ssp I	-
Bbv I	2	Dde I	-	Hinc II	-	NspC I	2	Stu I	-
BceF I	1	Dpn I	6	Hind III	1	Pac I	-	Sty I	-
Bcl I	1	Dpn II	6	Hinf I	-	PaeR7 I	-	Swa I	-
Bcn I	3	Dra I	-	HinP I	2	PflM I	1	Taq I	-
Bgl I	-	Dra III	-	Hpa I	-	Ple I	-	Tfi I	-
Bgl II	-	Drd I	-	Hpa II	3	Pml I	-	Tth111 I	-
Bsa I	-	Dsa I	-	Hph I	2	PpuM I	-	Tth111 II	-
BsaA I	1	Dsa V	4	Kas I	-	PshA I	-	Xba I	-
BsaB I	-	Eae I	-	Kpn I	-	Pst I	1	Xca I	-
BsaJ I	3	Eag I	-	Mae II	3	Pvu I	-	Xcm I	-
Bsg I	-	Ear I	-	Mae III	1	Pvu II	-	Xho I	-
BsiE I	-	Ecl136 I	-	Mbo I	6	Rma I	-	Xma I	-
BsiY I	4	Eco47 III	-	Mbo II	1	Rsa I	5	Xmn I	1

Enzyme	Site	Use	Site position (Fragment length) Fragment order		
BamH I	g/gatcc	1	1(48) 2	49(373) 1	
BceF I	acggc 12/13	1	1(281) 1	282(140) 2	
Bcl I	t/gatca	1	1(288) 1	289(133) 2	
BsaA I	yac/gtr	1	1(216) 1	217(205) 2	
Bsm I	gaatgc 1/-1	1	1(195) 2	196(226) 1	
Bsp1286 I	gdgch/c	1	1(104) 2	105(317) 1	
BstN I	cc/wgg	1	1(81) 2	82(340) 1	
Eco57 I	ctgaag 16/14	1	1(125) 2	126(296) 1	
EcoR I	g/aattc	1	1(334) 1	335(87) 2	
EcoR II	/ccwgg	1	1(81) 2	82(340) 1	
Fau I	cccgc 4/6	1	1(199) 2	200(222) 1	
Hae II	rgcgc/y	1	1(148) 2	149(273) 1	
Hind III	a/agctt	1	1(206) 2	207(215) 1	
Mae III	/gtnac	1	1(343) 1	344(78) 2	
Mbo II	gaaga 8/7	1	1(89) 2	90(332) 1	
Mme I	tccrac 20/18	1	1(45) 2	46(376) 1	

dn1gsttrans -> 1-phase Translation

DNA sequence 421 b.p. CCTCCAAAATCG ... TCCCGGAGACGG linear

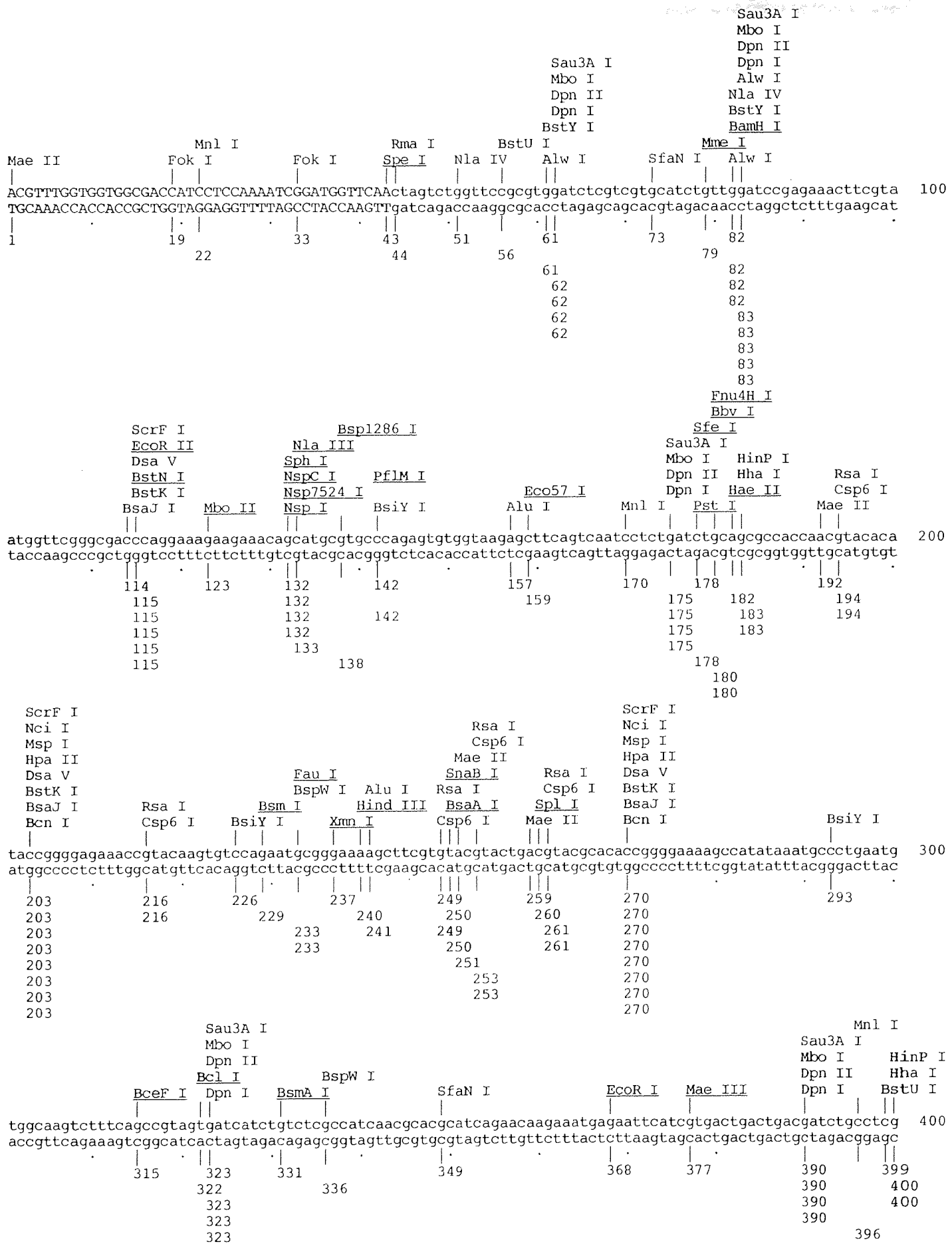
```
1/1                                             31/11
CCT CCA AAA TCG GAT CTG GTT CCG CGT GGA TCT CGT CGT GCA TCT GTT gga tcc gag aaa
P   P   K   S   D   L   V   P   R   G   S   R   R   A   S   V   G   S   E   K
61/21                                           91/31
ctt cgt aat ggt tcg ggc gac cca gga aag aag aaa cag cat gcg tgc cca gag tgt ggt
L   R   N   G   S   G   D   P   G   K   K   K   Q   H   A   C   P   E   C   G
121/41                                          151/51
aag agc ttc agt caa tcc tct gat ctg cag cgc cac caa cgt aca cat acc ggg gag aaa
K   S   F   S   Q   S   S   D   L   Q   R   H   Q   R   T   H   T   G   E   K
181/61                                          211/71
ccg tac aag tgt cca gaa tgc ggg aaa agc ttc gtg tac gta ctg acg tac gca cac cgg
P   Y   K   C   P   E   C   G   K   S   F   V   Y   V   L   T   Y   A   H   R
241/81                                          271/91
gga aaa gcc ata taa atg ccc tga atg tgg caa gtc ttt cag ccg tag tga tca tct gtc
G   K   A   I   *   M   P   *   M   W   Q   V   F   Q   P   *   *   S   S   V
301/101                                         331/111
tcg cca tca acg cac gca tca gaa caa gaa atg aga att cAT CGT GAC TGA CTG ACG ATC
S   P   S   T   H   A   S   E   Q   E   M   R   I   H   R   D   *   L   T   I
361/121                                         391/131
TGC CTC GCG CGT TTC GGT GAT GAC GGT GAA AAC CTC TGA CAC ATG CAG CTC CCG GAG ACG
C   L   A   R   F   G   D   D   G   E   N   L   *   H   M   Q   L   P   E   T

G
```

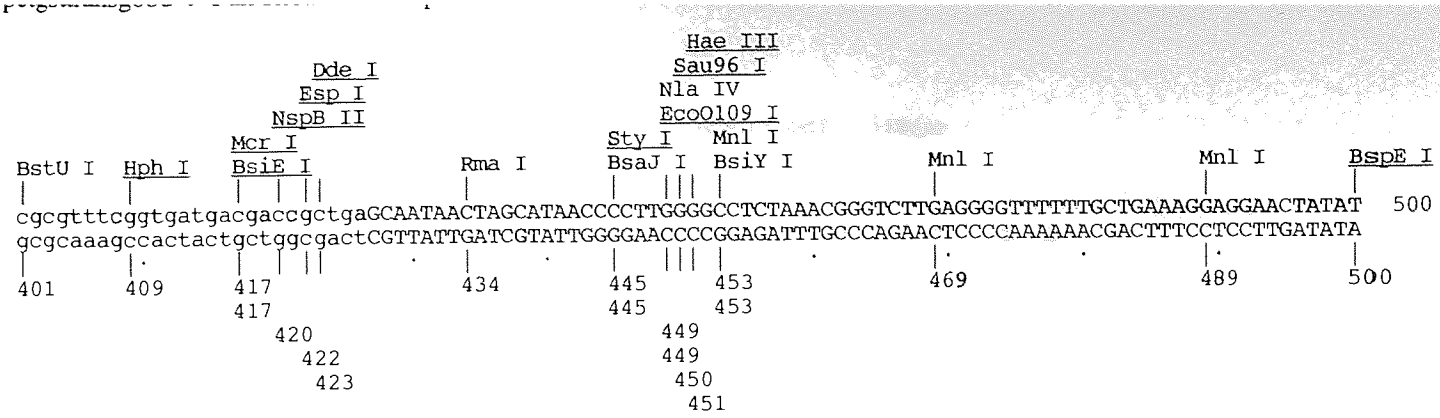
ZFDN1 Sequence (pET-42a)

petgsttransgood -> Full Restriction Map

DNA sequence 506 b.p. ACGTTTGGTGGT ... CTATATCCGGAT linear

Positions of Restriction Endonucleases sites (unique sites underlined)





Msp I
Hpa II
|
CCGGAT 506
GGCCTA
|
501
501

Restriction Endonucleases site usage

Aat II	-	Bsm I	1	Eco57 I	1	Mcr I	1	Rsr II	-
Acc I	-	BsmA I	1	EcoN I	-	Mlu I	-	Sac I	-
Afl II	-	Bsp120 I	-	Eco0109 I	1	Mme I	1	Sac II	-
Afl III	-	Bsp1286 I	1	EcoR I	1	Mnl I	6	Sal I	-
Age I	-	BspE I	1	EcoR II	1	Msc I	-	Sap I	-
Aha II	-	BspH I	-	EcoR V	-	Mse I	-	Sau3A I	5
Alu I	2	BspM I	-	Ehe I	-	Msp I	3	Sau96 I	1
Alw I	3	BspW I	2	Esp I	1	Nae I	-	Sca I	-
AlwN I	-	Bsr I	-	Fau I	1	Nar I	-	ScrF I	3
Apa I	-	BssH II	-	Fse I	-	Nci I	2	SfaN I	2
ApaL I	-	BstB I	-	Fnu4H I	1	Nco I	-	Sfe I	1
Ase I	-	BstE II	-	Fok I	2	Nde I	-	Sfi I	-
Asp718	-	BstK I	3	Fsp I	-	Nhe I	-	SgrA I	-
Ava I	-	BstN I	1	Gdi II	-	Nla III	1	Sma I	-
Ava II	-	BstU I	3	Gsu I	-	Nla IV	3	Sna I	-
Avr II	-	BstX I	-	Hae I	-	Not I	-	SnaB I	1
BamH I	1	BstY I	2	Hae II	1	Nru I	-	Spe I	1
Ban I	-	Bsu36 I	-	Hae III	1	Nsi I	-	Sph I	1
Ban II	-	Cfr10 I	-	Hga I	-	Nsp I	1	Spl I	1
Bbe I	-	Cla I	-	HgiA I	-	Nsp7524 I	1	Sse8337 I	-
Bbs I	-	Csp6 I	5	Hha I	2	NspB II	1	Ssp I	-
Bbv I	1	Dde I	1	Hinc II	-	NspC I	1	Stu I	-
BceF I	1	Dpn I	5	Hind III	1	Pac I	-	Sty I	1
Bcl I	1	Dpn II	5	Hinf I	-	PaeR7 I	-	Swa I	-
Bcn I	2	Dra I	-	HinP I	2	PflM I	1	Taq I	-
Bgl I	-	Dra III	-	Hpa I	-	Ple I	-	Tfi I	-
Bgl II	-	Drd I	-	Hpa II	3	Pml I	-	Tth111 I	-
Bsa I	-	Dsa I	-	Hph I	1	PpuM I	-	Tth111 II	-
BsaA I	1	Dsa V	3	Kas I	-	PshA I	-	Xba I	-
BsaB I	-	Eae I	-	Kpn I	-	Pst I	1	Xca I	-
BsaJ I	4	Eag I	-	Mae II	4	Pvu I	-	Xcm I	-
Bsg I	-	Ear I	-	Mae III	1	Pvu II	-	Xho I	-
BsiE I	1	Ecl136 I	-	Mbo I	5	Rma I	2	Xma I	-
BsiY I	4	Eco47 III	-	Mbo II	1	Rsa I	5	Xmn I	1

Enzyme	Site		Use	Site position (Fragment length) Fragment order
BamH I	g/gatcc		1	1(81) 2 82(425) 1
Bbv I	gcagc	8/12	1	1(179) 2 180(327) 1
BceF I	acggc	12/13	1	1(314) 1 315(192) 2
Bcl I	t/gatca		1	1(321) 1 322(185) 2
BsaA I	yac/gtr		1	1(249) 2 250(257) 1
BsiE I	cgry/cg		1	1(416) 1 417(90) 2
Bsm I	gaatgc	1/-1	1	1(228) 2 229(278) 1
BsmA I	gtctc	1/5	1	1(330) 1 331(176) 2
Bspl286 I	gdgch/c		1	1(137) 2 138(369) 1
BspE I	t/ccgga		1	1(499) 1 500(7) 2
BstN I	cc/wgg		1	1(114) 2 115(392) 1
Dde I	c/tnag		1	1(422) 1 423(84) 2
Eco57 I	ctgaag	16/14	1	1(158) 2 159(348) 1
Eco0109 I	rg/gnccy		1	1(448) 1 449(58) 2
EcoR I	g/aattc		1	1(367) 1 368(139) 2
EcoR II	/ccwgg		1	1(114) 2 115(392) 1

petgsttransgood -> 1-phase Translation

DNA sequence 506 b.p. ACGTTTGGTGGT ... CTATATCCGGAT linear

```
1/1                                             31/11
ACG TTT GGT GGT GGC GAC CAT CCT CCA AAA TCG GAT GGT TCA Act agt ctg gtt ccg cgt
T   F   G   G   G   D   H   P   P   K   S   D   G   S   T   S   L   V   P   R
61/21                                           91/31
gga tct cgt cgt gca tct gtt gga tcc gag aaa ctt cgt aat ggt tcg ggc gac cca gga
G   S   R   R   A   S   V   G   S   E   K   L   R   N   G   S   G   D   P   G
121/41                                          151/51
aag aag aaa cag cat gcg tgc cca gag tgt ggt aag agc ttc agt caa tcc tct gat ctg
K   K   K   Q   H   A   C   P   E   C   G   K   S   F   S   Q   S   S   D   L
181/61                                          211/71
cag cgc cac caa cgt aca cat acc ggg gag aaa ccg tac aag tgt cca gaa tgc ggg aaa
Q   R   H   Q   R   T   H   T   G   E   K   P   Y   K   C   P   E   C   G   K
241/81                                          271/91
agc ttc gtg tac gta ctg acg tac gca cac cgg gga aaa gcc ata taa atg ccc tga atg
S   F   V   Y   V   L   T   Y   A   H   R   G   K   A   I   *   M   P   *   M
301/101                                         331/111
tgg caa gtc ttt cag ccg tag tga tca tct gtc tcg cca tca acg cac gca tca gaa caa
W   Q   V   F   Q   P   *   *   S   S   V   S   P   S   T   H   A   S   E   Q
361/121                                         391/131
gaa atg aga att cat cgt gac tga ctg acg atc tgc ctc gcg cgt ttc ggt gat gac gac
E   M   R   I   H   R   D   *   L   T   I   C   L   A   R   F   G   D   D   D
421/141                                         451/151
cgc tga GCA ATA ACT AGC ATA ACC CCT TGG GGC CTC TAA ACG GGT CTT GAG GGG TTT TTT
R   *   A   I   T   S   I   T   P   W   G   L   *   T   G   L   E   G   F   F
481/161
GCT GAA AGG AGG AAC TAT ATC CGG AT
A   E   R   R   N   Y   I   R
```

Appendix

A3) Calculation of the theoretical distribution of codons within a library using Binomial distributions.

The theoretical distribution of a single codon at multiple positions of randomisation within a library can be calculated from the probability of that codon occurring at any particular randomised codon. This probability can be expected to follow a binomial distribution as there are only two possible outcomes (codon present / codon not present) at each position of randomisation. These outcomes can be considered as success (S) or Failure (F) respectively. As binomial distributions are usually applied to a set number of experiments to give (n), in this case the number of randomised positions is applied to give (n). Also needed in the calculation are the probability of a single success (p) and a single failure (q). In conjunction with the binomial coefficient (${}_nC_x$) which can be derived from Pascal's triangle (shown below), the probability of x number of successes (Px) i.e. x number of correct codons, can be calculated from the following equation $P(x) = {}_nC_x \cdot p^x \cdot q^{n-x}$ where x = the number of successes or 0, 1, 2,..., n.

For the benefit of the non mathematically inclined (i.e. myself). This is provided as a stepwise worked example below, showing how this equation may be used to calculate the number of serine codons present in a library randomised at 6 positions using the codon NNN which has 64 possible sequences.

Calculating the probability of six serine codons

Step 1) At a library randomised at **six** positions the binomial coefficients for each number of positions follow those in row **six** of Pascal's triangle, see below

			1				row 0
		1		1			row 1
	1		2		1		row 2
1		3		3		1	row 3
	1		4		6		4		1	row 4
1		5		10		10		5		1	row 5
	1		6		15		20		15		6		1	row 6
1		7		21		35		35		21		7		1	row 7

The starting point is the number 1 of row six which is ignored in counting across the triangle.

Step 2) Find the binomial coefficient by counting across the triangle for the number of successes you wish to calculate (i.e. all six codons randomised with serine would be six successes) you would count across six numbers. As the first position is ignored this takes you across the triangle to a binomial coefficient of 1. (This means there is only one way of generating six serine codons at six positions.

Step 3) The probability of generating six serine codons can be found by multiplying the probability of generating one serine codon 6 in 64 by itself six times $(^{6}/_{64})^{6}$ which gives $^{46656}/_{6..9 \times 10}{}^{10}$ or 46656 genes with serine codons at six positions in 6.9×10^{10} genes (the total number of possible genes in the library)

Calculating the Probability of 4 serine codons

Step1) Work along row six (six randomised positions) four steps, the binomial coefficient for four successes. (This should be 15, as there are 15 ways to generate 4 serine codons when six positions are randomised.

Step 2). The probability of generating a single serine codon remains 6 in 64. Thus the probability of generating 4 serine codons becomes $(^{6}/_{64})^{4}$. However as two of the randomised codons **must** be any other codon bar serine the probability of this must also be taken into account. The probability of a single codon not being serine (or failure q) is 58 in 64 thus the probability of 2 codons not being serine becomes $(^{58}/_{64})^{2}$. As serine must be present at 4 positions **and** a non serine codon at two positions must also be present, these probabilities must be multiplied as stated in the and / or rule of probability.

Thus the equation becomes $15 \times (^{6}/_{64})^{4} \times (^{58}/_{64})^{2}$

The equation $(^{6}/_{64})^{4} \times (^{58}/_{64})^{2}$ gives us the probability of four codons being serine and two codons being non serine which is 4359744 in 6.9×10^{10}. However the binomial coefficient states that there are 15 ways for this to be achieved so this must be

multiplied by 15 giving 65396160 genes with serine codons at 4 positions in the library population (6.9×10^{10} genes).

Calculating the binomial distribution of 3 Tryptophan codons

Step 1) Work three steps into row six of Pascal's triangle to obtain the binomial coefficient (20).

Step 2) calculate the probability of a single tryptophan codon ($^1/_{64}$) thus the probability of three tryptophan codons becomes$(^1/_{64})^3$. Calculate the probability of a single non tryptophan codon ($^{63}/_{64}$). Thus the probability of three may be calculated from $(^{63}/_{64})^3$.

The probability of three tryptophan codons being generated in the library may then be calculated from $20 \times (^1/_{64})^3 \times (^{63}/_{64})^3$ which should generate a probability of 5.0×10^6 genes with tryptophan codons at 3 randomised positions.

Calculations based upon libraries randomised at 5 positions are carried out in the same fashion but will use row 5 of Pascal's triangle to generate the binomial coefficient. Libraries randomised at 7 codons will use row 7.