TERMINAL DEOXYNUCLEOTIDYL TRANSFERASE GENERATION OF SEQUENCE DIVERSITY ON PLASMIDS

A thesis

Presented to

the Faculty of the College of Science and Technology

Morehead State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

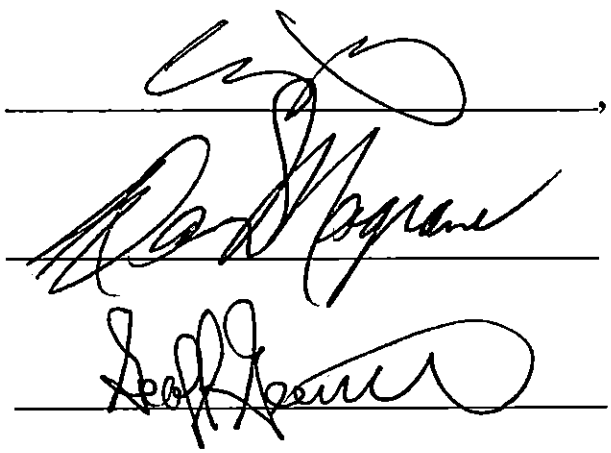
Qingbei Zhang

June, 2002

Accepted by the faculty of the College of Science and Technology, Morehead State University, in partial fulfillment of the requirements for the Master of Science degree.

Director of Thesis

Master's Committee:



, Chair

6/19/2002

Date

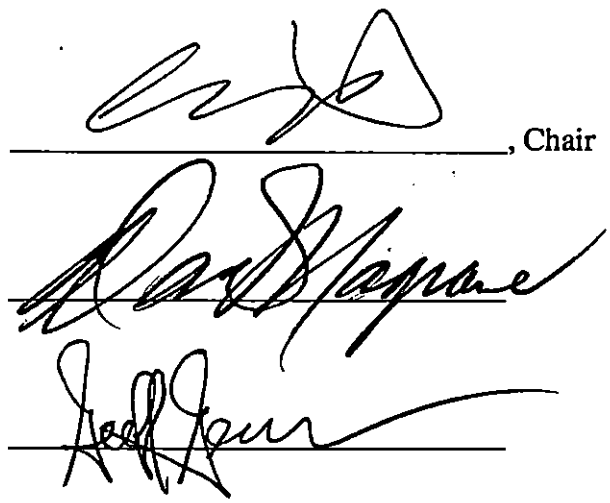
TERMINAL DEOXYNUCLEOTIDYL TRANSFERASE GENERATION OF SEQUENCE DIVERSITY ON PLASMIDS

Qingbei Zhang, M.S.
Morehead State University, 2002

Director of Thesis: 

Peptide libraries provide new technologies to develop drugs and vaccines faster, more cost-effectively, and in greater numbers than ever before. Current peptide libraries cannot encode large, three-dimensional, and folded proteins due to the size restriction placed on them by oligonucleotide synthesis and synthetic peptide chemistry. In order to synthesize large random DNA sequences, terminal deoxynucleotidyl transferase (TdT) was used to generate sequence diversity on the circular, double-stranded plasmid pGEM.. Cutting pGEM with the restriction endonuclease Pst I creates a free 3' terminus. TdT added deoxynucleotides to the 3' terminus in a template-independent manner. Using this procedure, long random protein coding sequences (342-1747 base pairs) were constructed on plasmid DNA. These TdT extensions on plasmid can provide for selection of in vivo functions or can be transcribed and translated *in vitro* to conduct *in vitro* selection by binding with targets.

Accepted by: _____, Chair



Acknowledgements

I would like to express my deepest thanks to Dr. Craig Tuerk for his guidance and assistance with the tremendous amount of work required to complete the research and paper.

I wish to thank my graduate committee, Dr. Geoff Gearner, Dr. David Magrane, and Dr. Craig Tuerk for their guidance in my professional and personal development.

I wish to thank Dr. Ted Pass as my graduate advisor and working supervisor. I had a wonderful year under his direction.

In addition, I express my thanks to the faculty of the Department of Biological and Environmental Sciences. I was very lucky to receive the excellent education during these two years.

My gratitude also goes out to the following graduate students: Sean Thatcher, Mike Kenawell, Zhen Lei, Lisa Hawkins, and Cassandra Garrett. I will remember all the good times during our graduate years.

Finally, I wish to thank my husband Yang and my parents. I appreciate their love and support during my graduate education.

Table of Contents

List of Figures

List of Tables

Tables **Page**

Introduction

Traditional drug discovery has primarily depended on classical processes to screen compounds for novel pharmacological activity. This empirical approach is time-consuming, inefficient, and the drugs are developed against a limited number of *in vivo* targets. In recent years new pharmaceutical compounds have been discovered by new methods that are faster, more cost-effective and in greater numbers than ever before. Peptide libraries provide a convenient means to screen a large number of naturally occurring or synthetic peptides and has attracted great interest from pharmaceutical companies and the biotechnology industry (Fitzgerald, 2000). As a natural modulator of phenotype, proteins play important structural and catalytic roles and are widely used in diagnostic, therapeutic, and industrial applications.

In the post-genome era, the focus has shifted to the study of the proteome, the complete set of proteins produced from a given genome. The proteome is much larger than the genome would predict. From a single gene, transcriptional, translational, and post-translational modifications may give rise to more than one protein. It is estimated that our 35,000 genes would make approximately 5 million proteins.

Protein library technology is a necessary and effective method to recover the information encoding a protein. By using *in vivo* and *in vitro* display systems, the resulting protein is expressed in such a way as to remain linked to the encoding nucleic acid, which then is recovered for amplification and further selection. Biological library technologies have emerged during the past 15 years as powerful tools for basic research and drug discovery and

development. In recent years, these libraries have been shown to be very useful for finding compounds, such as inhibitors, binding agents, enzymes, and antibodies (Cull et al., 1992; Matthews and Wells, 1993; Martens et al, 1995), which satisfy specific requirements. Thus these techniques represent an excellent resource for the discovery of compounds for a variety of biotechnological process needs. An advantage of the biological libraries is the ability of each of the library members to be replicated, carrying with it its own coding sequence. Data derived from the biologically active libraries show that of the libraries reported during the 1992-1997 period, approximately 50% were in fact peptide-based (Dolle, 2000).

Phage display

Phage display describes a selection technique in which a peptide or protein is expressed as a fusion with a coat protein of bacteriophage, resulting in display of fused protein on the exterior surface of the phage virion, while the DNA encoding the fusion resides within the virion (Scott and Smith, 1990). Phage display uses a bacterial virus (the M13-related filamentous phage) to create a physical linkage between a vast library of random peptide sequences to the DNA encoding each sequence, allowing rapid identification of peptide ligands for a variety of target molecules (antibodies, enzymes, and cell-surface receptors, etc.) by an *in vitro* selection process called " biopanning." "Biopanning" (Figure 1) is carried out by incubating a library of phage-displayed peptides with a plate (or bead) coated with the target, washing away the unbound phage, and eluting the specifically-bound phage.

Alternatively the phage can be reacted with the target in solution, followed by affinity

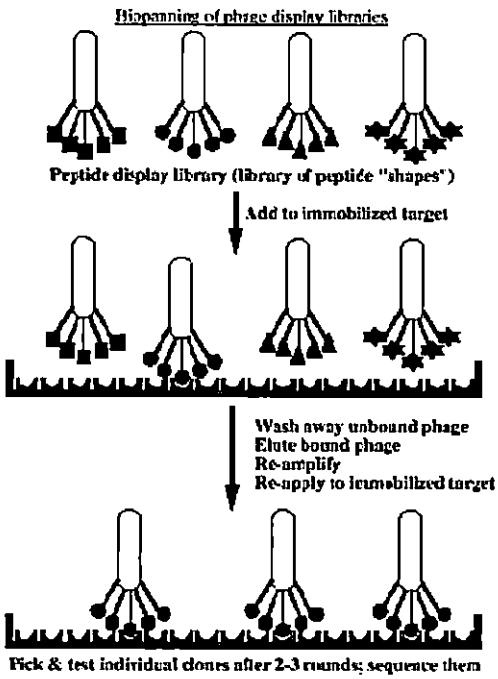


Figure 1. Biopanning of phage display libraries

capture of phage-target complexes onto a plate or bead that specifically binds the target. The eluted phage is then amplified and taken through additional cycles of panning and amplification to successively enrich the pool of phage in favor of the tightest binding sequences. After 3 or 4 rounds, individual clones are characterized by DNA sequencing and ELISA.

In this filamentous phage system, foreign peptide sequence can be expressed on the surface of infectious phage, thereby conferring two significant benefits. First, these phages can be screened in a large number at one time because the phages can be

applied at a very high concentration (10^{13} phages per milliliter). Second, even though the recovered phages encoding the target sequence have very low yield in the first round they could be amplified through additional rounds of infection and selection (Delvin et al, 1990). Random peptide libraries displayed on phage have been used in a number of applications, including epitope mapping, identification of peptide ligands, identification of enzyme substrates or inhibitors, and identification of peptide mimics of non-peptide ligands. However, there are limitations to the technique. First, the reliance of the method on the relatively low efficiency with which DNA can be introduced into *E.coli* cells (the transformation efficiency) typically imposes a ceiling on library size of 10^9 recombinants (FitzGerald, 2000). Second, since displayed proteins larger than 20-30 amino acids have a deleterious effect on the infectivity function of PIII in phage vectors, this vector is only suitable for the display of short peptides (New England Biolab, 2000-2001).

Plasmid display

Plasmid display, or "peptide on plasmids", employs peptides fused to the C terminus of *lac* repressor *Lac*I. The *E.coli lac* repressor is a DNA binding protein that regulates expression of the *lac* operon. The repressor protein physically links the peptides to the plasmids encoding them by binding to the *lac* operator sequence on the plasmid (Mattheakis et al., 1994). The ability of the plasmids to replicate after transformation of *E.coli* allows amplification of the selected population for

subsequent rounds of enrichment (panning). After several rounds of panning and amplification, sequencing of individual plasmid clones reveals the structure of the peptides. This "peptide-*Lac*I-plasmid complexes" screening method has been used successfully to isolate peptide ligands for a number of antibodies and mammalian cell surface receptors (Cull et al., 1992; Gates et al., 1996). These isolated proteins have similar binding affinities with phage-displayed libraries for their targets, 10 to 100 μM (Gates et al., 1996).

SELEX

SELEX (Selective Evolution of Ligands by EXponential enrichment) is a method for rapidly selecting preferred binding sequences from a population of random sequences. The interaction between the bacteriophage T4 DNA polymerase (gp43) and the ribosome binding site of the mRNA that encodes it was studied by Tuerk and Gold (1990). First, they ligated three pieces of synthetic oligonucleotides to create a pool of single-stranded DNA templates, with 36 nucleotides used for the T4 DNA polymerase recognition site and 3' and 5' flanking sequences used for RT-PCR and sequencing. In the case of the gp43 recognition site, the sequence AAUAACUC encoding the wild type hairpin loop was replaced with completely randomized sequences at these eight positions creating a pool of 65,536 species of variant sequences. Second, by *in vitro* transcription, a pool of RNA sequences with variable sequences in the hairpin loop were produced. These variant transcripts were subjected

to rounds of selection for binding to T4 DNA polymerase followed by RT-PCR. The amplified double-stranded DNA products were transcribed in vitro and used as an enriched population in the next round of selection. By this way, they found 2 different sequences from 65,536 species having the highest affinity for gp43. One is the wild-type loop sequence, and one is varied from wild type at four nucleotides in the loop sequence. So, SELEX analyses could be used to determine the optimal binding sequence for any nucleic acid binding protein including both natural and unnatural solutions. The principle of SELEX is depicted below in Figure 2.

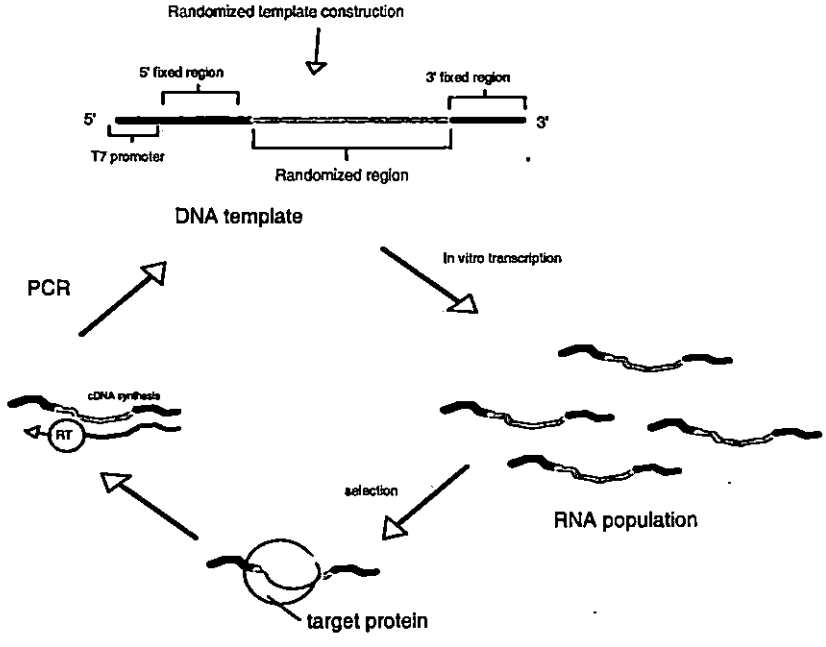


Figure 2. The SELEX procedure

Polysome display

In the polysome display system (see Figure 3), a polysome is a physical linkage between nascent peptides and their encoding mRNAs. Polysome complexes, consisting of messenger RNA (mRNA), ribosome, and encoded protein, which are used for selection, were stabilized by controlling the concentration of magnesium ions and the addition of chloromphenicol (Mattheakis et al., 1994). The bound complexes were dissociated with EDTA, and the encoding mRNAs were copied into cDNA and

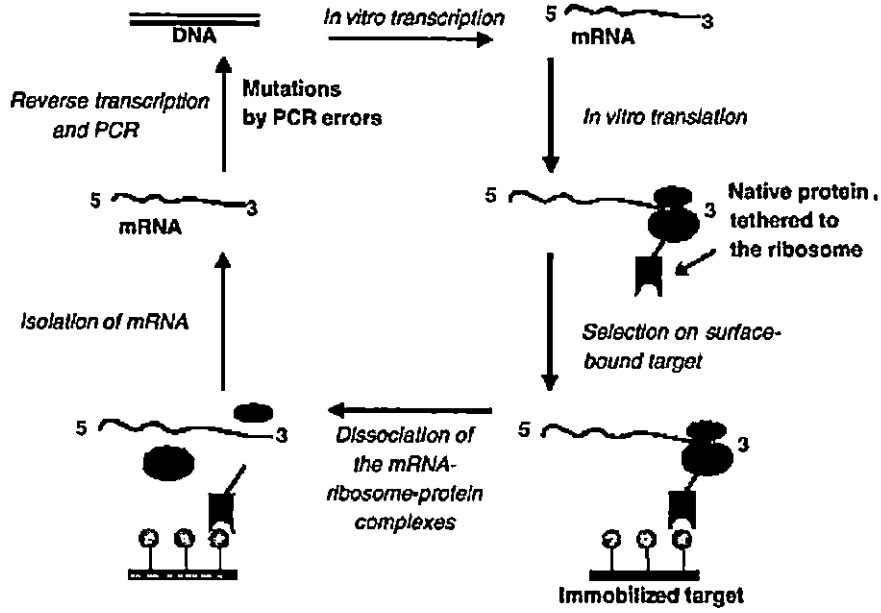


Figure 3. The principle of polysome display.

PCR-amplified to produce DNA templates. Several rounds of *in vitro* synthesis and selection resulted in pools of enriched sequences that were cloned in phagemid vector to determine the specificity of peptide binding by phage ELISA and to sequence the DNA. By this process, Mattheakis and others (1994) created a polysome library that yielded proteins with affinities ranging from 7 to 140 nM. If a high fidelity

proofreading DNA polymerase is used during the PCR amplification steps (Fig. 3), the repertoire of the library employed is virtually maintained. However, a particularly interesting feature of the ribosome display technology is that it can also be used for the directed evolution and affinity maturation that occurs during the selection process, if a low-fidelity DNA polymerase could be used that introduces mutations during amplification (Viguera et al., 2001). A directed evolution can also be achieved when *in vivo* selection technologies are used, for example by transforming into particular mutant strains when using phage display (Irving et al., 1996). However, this entails the risk of simultaneously introducing unwanted and possibly detrimental mutations in the plasmid or the host genome (Hanes et al., 2000).

RNA-peptide fusions

Alternative *in vitro* selection methods such as "RNA-peptide fusion," in which an *in vitro*-synthesized polypeptide is covalently attached to its encoding message, have been demonstrated to work for peptide libraries. Roberts and Szostak (1997) did this by covalently linking puromycin, an antibiotic that mimics an aminoacylated tRNA, to the 3' end of a synthetic mRNA through a DNA linker. A ribosome begins translation of such a template as usual, generating a peptide as it transits the open reading frame (ORF). When the ribosome reaches the end of the open reading frame and hits the DNA linker it stalls, allowing the nearby puromycin to enter the A site of the ribosome and accept the nascent peptide chain. The resulting peptidyl-puromycin

molecule contains a stable amide linkage between the peptide and the O-methyl tyrosine portion of the puromycin. The O-methyl tyrosine is, in turn, linked by a stable amide bond to the 3'-amino group of the modified adenosine portion of puromycin. Thus a synthetic mRNA with puromycin at its 3' end would be expected to generate stable mRNA-peptide fusions.

The ability to synthesize covalent mRNA-peptide fusions by *in vitro* translation provides a different approach to the *in vitro* selection and directed evolution of peptides and proteins. This will allow the unreadable information in the protein portion to be read via the attached mRNA. Current protocols are consistent with the generation of mRNA-peptide fusion libraries consisting of 10^{12}-10^{13} independent members. Roberts and Szostak (1997) have generated a library of 10^{12} fusions, in which the peptide domain contains a random 27-aa sequence, in a 10 ml translation reaction. Improvements in the efficiency of the transfer reaction, coupled with a modest degree of scale-up, could increase the accessible library complexity to as much as 10^{15}.

In general, an *in vivo* display system provides a peptide library with 10^{9}—10^{10} recombinants because of the transformation efficiency and capacity of *E.coli*. By avoiding the bacterial transformation, the *in vitro* peptide library exceeds the diversity of cell-based systems and provides 10^{14}—10^{15} recombinants (Singer et al., 1997). Moreover, *in vitro* expression of libraries provides the opportunity to continuously

introduce variation into the sequence pool by low-fidelity amplification between rounds of selection. Repeated mutagenesis coupled with affinity-selective screening should provide the rapid evolution and identification of peptide ligands of high affinity (Mattheakis et al., 1994)

Alpha complementation

The *lacZ* gene of *E. coli* encodes a protein called β-galactosidase. The functional enzyme is a tetramer of four identical subunits each of which is a single protein product of the *lacZ* gene (see Figure 4). Each of the subunits contains an

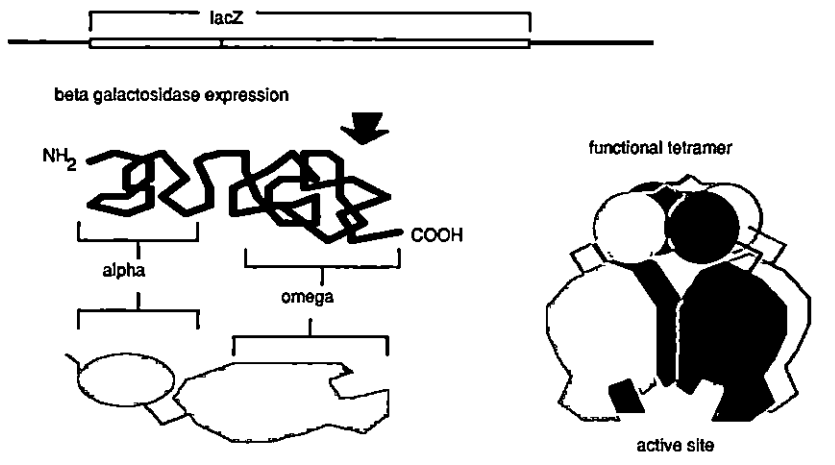


Figure 4. Tetramerization of beta-galactosidase

amino terminal domain called the alpha domain and a carboxyl terminal domain called the omega domain. The alpha domains of the subunits are responsible for bringing together the four subunits to make the functional enzyme; the omega domains form the active site of the enzyme. There are bacterial mutants of the *lacZ* gene that have neatly deleted the amino terminal encoding portion of the gene so that

only the omega fragment is expressed. Even though these expressed omega fragments exist in the cell, they do not combine to form functional β-galactosidase. If alpha domains are expressed from a separate gene found on a plasmid in these cells, these four alpha fragments will both interact with each other and with four separate omega domains to create a functional β-galactosidase. This process is called α-complementation as shown in Figure 5.

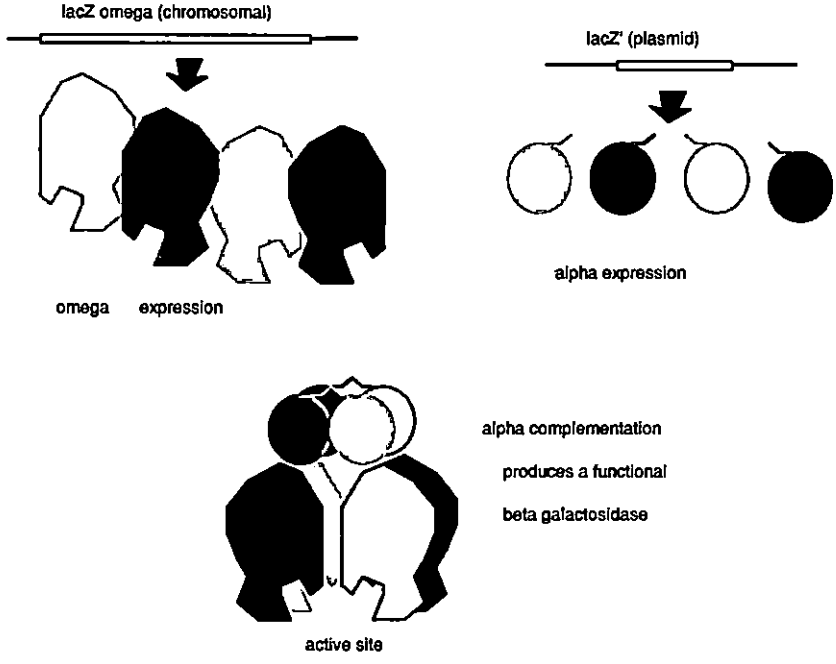


Figure 5. Alpha-complementation of beta-galactosidase.

Beta-galactosidase is normally used to cleave the disaccharide lactose into the monosaccharides galactose and glucose. Beta-galactosidase cleaves the colorless chromogenic substrate X-gal to produce the blue chemical group (Figure 6).

On the other hand, insertion of a fragment of foreign DNA into the polylinker region of the plasmid results in an amino terminal segment that is not capable of alpha complementation. The cells carrying recombinant plasmids form white colonies and provide a visual means of identification.

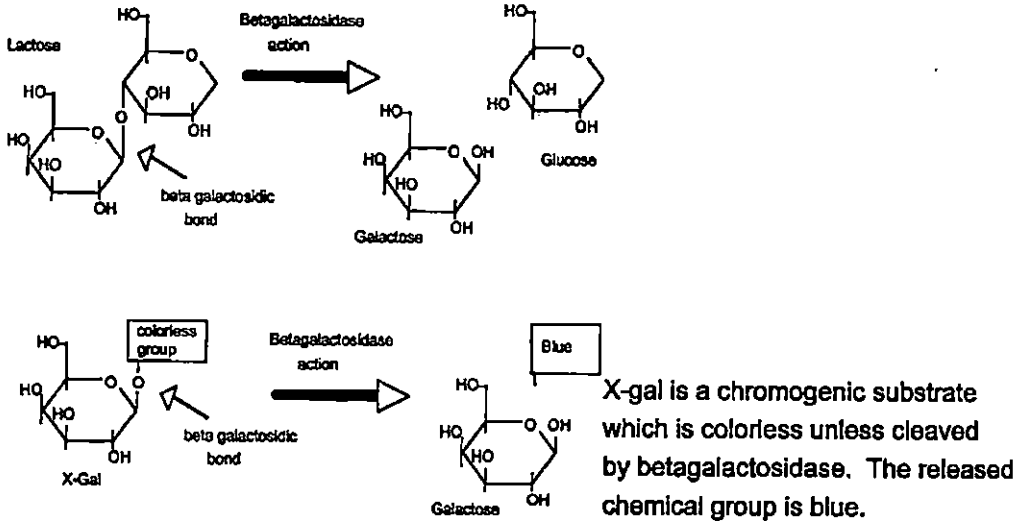


Figure 6. Beta-galactosidase action on lactose and X-gal.

Limitations of peptide libraries

At present, random peptide libraries do not encode large molecular weight, three-dimensional, folded proteins due to the size restriction placed on them by the length limit of oligonucleotide synthesis and synthetic peptide chemistry (Spencer and Tuerk, 1999). This size restriction limits the ability of current peptide libraries to find highly specific binders due to the inconsistent shapes of flexible short peptide chains

(Ladner et al., 1995). However, most biological reactions are carried out by large, well-folded proteins. A folded protein usually has one or several protein domains. For example, peptide-binding PDZ domains usually have an average length of 85-100 amino acids (Li, 2000). But the current biological peptide libraries have been created that code for a random peptide ranging from 6-38 amino acids (Kay et al., 1993), limiting their ability to find highly specific binders.

Previous work using TdT to create random regions

Spencer and Tuerk (1999) used terminal deoxynucleotidyl transferase (TdT), a template-independent DNA polymerase that catalyzes the addition of nucleotides to a 3' terminal hydroxyl group, to generate random nucleotide additions to the 3'-terminus of an oligonucleotide. This random extension reaction generated random ssDNA over 1500 base pairs. The random extension products were converted to DNA duplex and used as templates for *in vitro* transcription and translation. The yielded proteins ranged from 30 kDa to 50 kDa, predicting proteins with 250-415 amino acids. This number of random amino acids far exceeded the current number of random amino acids, 6-38, in peptide libraries. Such a system would allow the display and selection of large random proteins. However, numerous manipulations are required to prepare the 3' ends of these oligo extensions so that they may be utilized during the replicative rounds of such selections, and these manipulations greatly lower the yield

and consequently the diversity. In addition, the oligos used must be in high concentrations and are a significant expense in this system.

Objectives of the study

In order to generate large random DNA sequences, and then large peptide molecules, terminal deoxynucleotide transferase (TdT) was used to produce DNA sequence diversity on the plasmid pGEM. An outline of the experimental design is shown in Figure 7. pGEM has an ampicillin-resistance gene and a polylinker region. When pGEM is digested by the restriction endonuclease *Pst*I, a 3' sticky end is produced. TdT could add random nucleotides to the 3'terminus of linear pGEM in a template-independent manner. In this way, a long random protein coding sequences can be constructed. Ligation and transformation of *E.coli* DH5α with these extended plasmids produced white colonies on X-gal containing culture media. Plasmid was prepared from selected white colonies and analyzed for length of random DNA insertion.

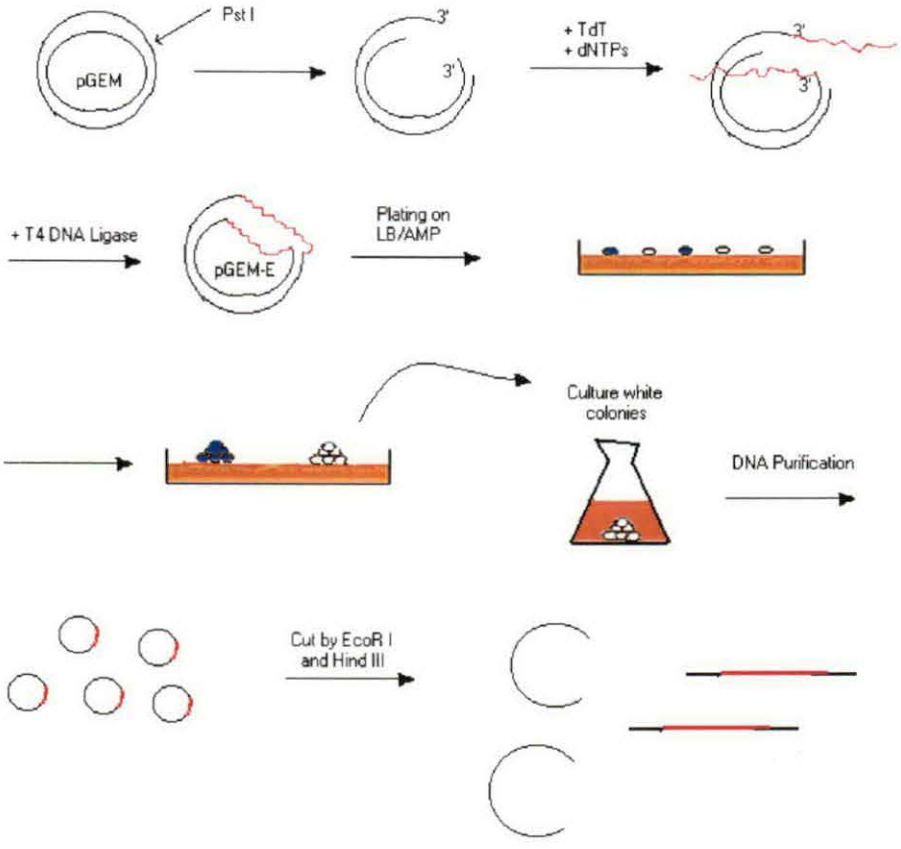
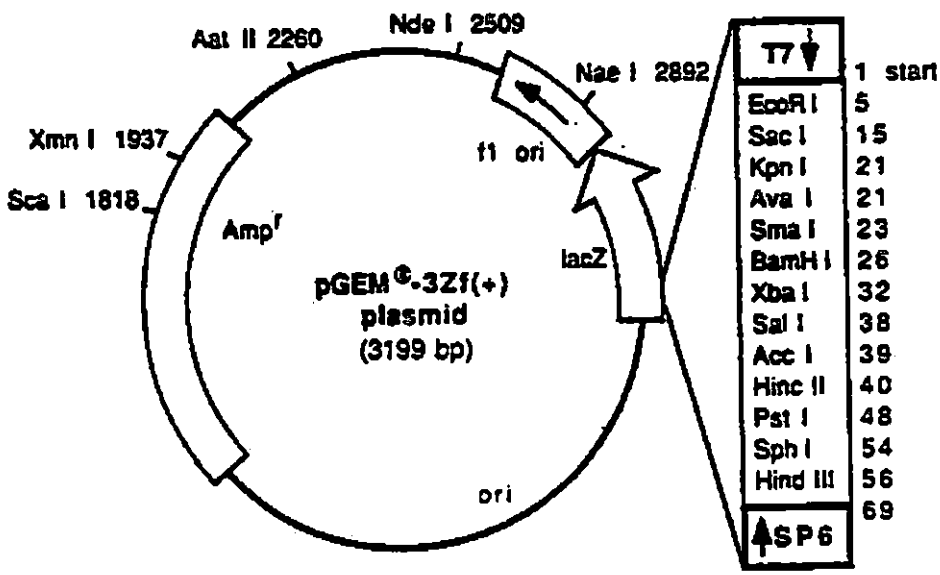


Figure 7. TdT extension on plasmids, ligation and transformation.

Materials and Methods

Materials

Plasmid: pGEM-3Zf(+) (Promega, Madison, Wisconsin) carries a polylinker inserted within the alpha region of the *lacZ* gene. The polylinker region contains a clustering of unique cleavage sites into which DNA fragments generated by endonuclease restriction digestion can be inserted (Figure 8).



Selectable marker: Amp^R Ori: Origin of replication
Multiple coloning site (polylinker): 5—61 *LacZ*: *LacZ* alpha gene

Figure 8. pGEM-3Zf(+) plasmid map.

***E. coli* DH5 alpha:** A mutant strain of *E. coli* that expresses the omega domain of *lacZ*.

TdT: Terminal deoxynucleotide transferase (Promega, Madison, Wisconsin) is a template-independent DNA polymerase found only in the prelymphocytes. In the presence of a divalent cation (Mg^{2+}, Co^{2+}, Mn^{2+}) the enzyme catalyzes the addition of dNTPs to the 3'-hydroxyl termini of DNA. Depending on the reaction conditions, anywhere from three to several thousand bases can be added.

Restriction endonucleases: (New England Biolabs, Beverly, Massachusetts) Restriction endonucleases are enzymes that cleave DNA molecules at specific nucleotide sequences depending on the particular enzyme used. Enzyme recognition sites are usually 4 to 6 base pairs in length. Specific reaction buffer for the particular enzyme is used.

T4 DNA Ligase: (New England Biolabs, Bervely, Massachusetts): Catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in duplex DNA or RNA. This enzyme will join both blunt-end and cohesive-end termini.

Mung Bean Nuclease: (New England Biolabs, Beverly, Massachusetts) A single-strand specific DNA and RNA endonuclease that will degrade single-stranded extensions from the ends of DNA and RNA molecules, leaving blunt, ligatable ends.

X-gal: 5-bromo-4-chloro-3-indolyl-beta-galactoside. Dissolved 100 mg of X-gal in 5 ml of dimethylformamide in a sterile polypropylene tube. Aliquot 1 ml into eppendorf tubes wrapped in foil (to prevent damage by light) and store at -20 °C. It is not necessary to filter sterilize X-gal solutions.

Methods: All preparation of plasmid DNAs, phenol/ether extraction, agarose gel analysis, restriction digests, ligations, polyacrylamide gel electrophoresis and visualization with ethidium bromide, and transformation of competent bacterial cells and plating with X-gal are essentially as treated by the Molecular Cloning Manual (Sambrook 1989).

Terminal extension reaction by TdT

TdT extensions were performed following manufacturer's instructions except the concentrations of dATP, dCTP, dGTP and dTTP were 1mM each and extensions were cultured for 18 hours at 37° C. In some cases homopolymeric tailing with only dGTP or dCTP were carried out as noted.

Results

Effect of *Pst* I and *Eco*R I on plasmid (Figure 9)

In order to verify that complete digestion of pGEM by *Pst*I and *Eco*RI release the expected fragment size, endonucleases *Pst*1 and *Eco*R I were used to cut pGEM. NE Buffer *Eco*R I was suggested to use for this double digest. The first lane is the 100 bp DNA ladder used as molecular weight standards for electrophesis. The second lane showed the DNA cut by both *Eco*R I and *Pst*1, as evidenced by a small band that appeared on the gel. It was about 43 bp in length, so it runs faster than the 100 bp DNA marker, suggesting both endonucleases were cuting well. The third lane showed the pGEM DNA cut only by *Pst*1, so there was no small band showing up. This could be used as a negative control. Because the band was relatively small (43 bp), polyacrylamide gel electrophoresis was used instead of agarose gel allowing high resolution of very small fragments (<0.2 kb).

Large scale digestion of pGEM by *Pst* I (Figure 10)

Pst I (25 μl) was used to incubate with 50 μl pGEM at 37°C overnight in 1 ml. After cutting by *Pst*1, pGEM became linear and ran slower than the uncut pGEM which is supercoiled. *Pst*1 is an enzyme which can cut double stranded DNA and produce a 3' sticky end. This 3' sticky end could be used as the 3' terminus for TdT's random extension to construct the random DNA sequences.

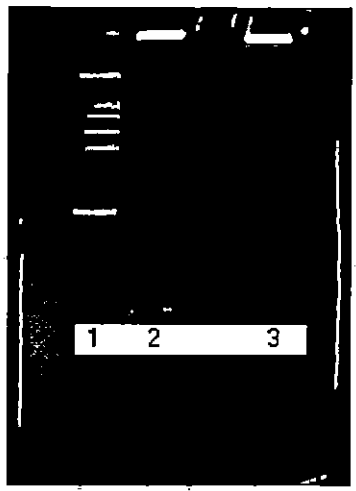


Figure 9. Double restriction digest of pGEM. Lane 1, 100bp ladder. Lane 2, pGEM+ *Pst*I, *Eco*R I. Lane 3, pGEM + *Pst*I.

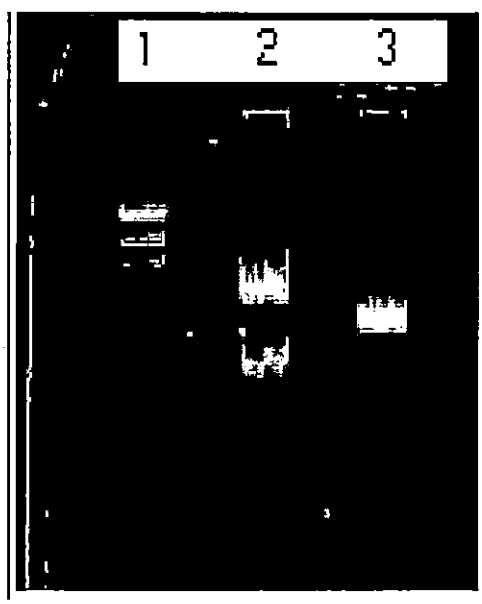


Figure 10. Large-scale digest of pGEM by *Pst*I. Lane 1, Hind III digest phage lambda DNA. Lane 2, uncut pGEM. Lane 3, pGEM cut with *Pst*I

Terminal extension reaction by TdT (Figure 11)

As a positive control, TdT was used to extend the 3' terminus of an oligonucleotide GS (45 bp DNA sequence) and pGEM cut by *Pst*1. This took place in a 10 μl reaction system, including pGEM DNA 1 μl, TdT 1 μl and 10mM dNTPs 1μl. TdT is a DNA polymerase that catalyzes the template-independent addition of deoxyribonucleotides to the 3' hydroxyl terminus of single or double-stranded DNA. After incubation overnight, 1 μl *Eco*R I and 1 μl *Eco*R I buffer were added to the extended and non-extended pGEM, incubate overnight. The 43bp small band should be seen from the non-extension pGEM lane and the small band should disappear in extended pGEM lane. However, these results were not seen (lane 1, 2). On the other hand, TdT did extend the GS DNA sequence since there was no small band shown in the extended GS only the DNA smear image could be seen in lane 3. The extensions on the 43bp fragment of *Eco*RI-*Pst*I cut pGEM may have not been visible because of insufficient amounts of plasmid. The extension on *Pst*I-cut pGEM was repeated by increasing the amount of plasmid in the reaction, followed by *Eco*RI digestion. Results (Figure 12) showed that a small band appeared on the non-extension lane (right lane), but no small band was seen "smear" throughout the lane in the extension lane (left lane). This suggested that all the enzymes (TdT, *Eco*R I, *Pst* I) are functional.

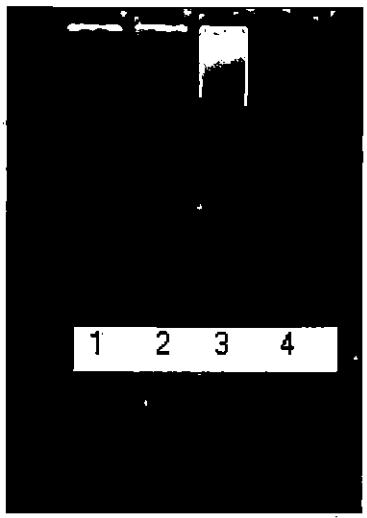


Figure 11. Extension of oligo GS by TdT. Lane 1, *Pst*I cut pGEM. Lane 2, TdT extended pGEM (*Pst*I cut). Lane 3, TdT extended GS. Lane 4, GS oligo (45bp).

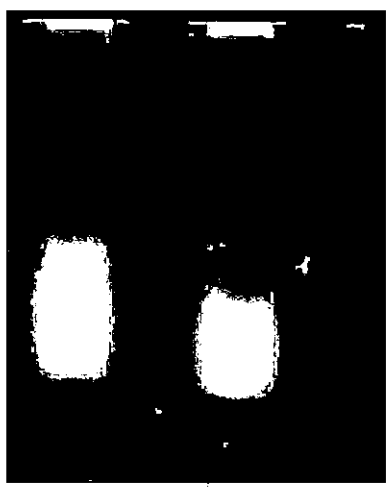


Figure 12. Extension on *Pst*I cut pGEM by TdT. Left lane, *Pst*I cut pGEM with TdT extension, then *Eco*RI digestion. Right lane, *Pst*I cut pGEM, no TdT extension, then *Eco*RI digestion.

Timed extension of pGEM 3' terminus generated by *Pst* I

TdT was used to extend pGEM 3' terminus produced by *Pst* I. TdT and dNTPs were added to incubate with pGEM for 5', 15', 1h, 2h, and overnight. Following the

extension reaction, *Eco*RI was added to each tube and incubated for 6 more hours. In Figure 13, lane 1 (no extension) showed the appearance of small band, which should be cut by both *Pst* I and *Eco*R I with 43 bp in length. From lane 2 to lane 6, no small bands were observed, suggesting the 3' terminus of pGEM was extended by TdT, and this extension reaction could occur within 5 minutes. The extended smears indicative of extended DNA increased in molecular weight (traveled a shorter distance in each gel lane) with increasing time of extension (highest molecular weight products in lane 2, Figure 13). In order to eliminate the possibility that TdT buffer was producing these results, reactions were varied for presence of TdT with a common background buffer showing that the enzyme was required for the extensions visualized by gel electrophoresis (Figure 14). In lane 3 (Figure 14), there was still a faint small band that could be seen after 5 minutes extension, but completely disappeared in lane 1 (1 hour extension). So, it was estimated that the minimum time for TdT extension was between 1 hour and 2 hours, even though some extension could occur within 5 minutes as seen in Figure 13.

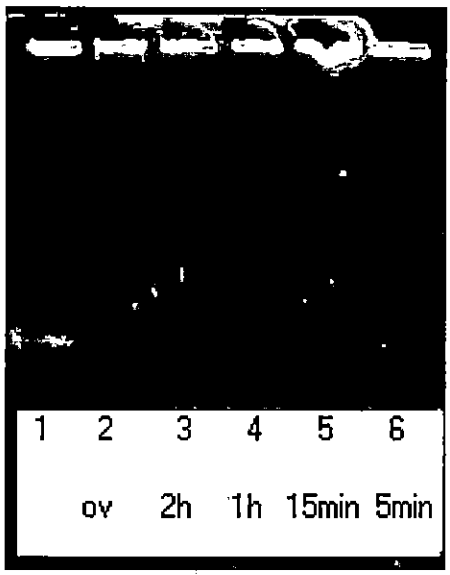


Figure 13. Timed extension on PstI cut pGEM by TdT. Lane 1, pGEM + *Pst*I, *Eco*RI. Lane 2-6, pGEM + *Pst*I, TdT extension overnight, 2 hours, 1 hour, 15 minutes, 5 minutes, respectively, and then cut by *Eco*RI.

White colony numbers of extension products (Table 1)

Various treatments were performed on PstI-cut pGEM as listed in Table 1. "Extension" refers to mixed dNTPs (A,C,G, and T) during TdT extention. The label "poly A" refers to plasmids that were extended with mixed dNTPs as above, then homopolymer tailed with dATP. A similar pattern was repeated for polyG, polyC and polyT. The label "polyC+G" refers to random extension followed by polyC then polyG then a single round of replication by PCR. It was expected there would be annealing of the polyC/G tracts to each other to give a filled in random region but the transformation

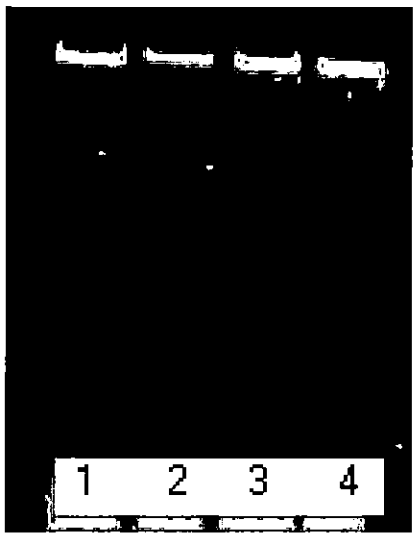


Figure 14. Timed extension on *Pst*I cut pGEM by TdT or buffer control. Lane 1, pGEM + *Pst*I, TdT extension 1 hour, then cut by *Eco*RI. Lane 2, pGEM + *Pst*I, TdT buffer only 1 hour, then cut by *Eco*RI. Lane 3, pGEM + *Pst*I, TdT extension 5 minutes, then cut by *Eco*RI. Lane 4, pGEM + *Pst*I, TdT buffer only 5 minutes, then cut by *Eco*RI.

efficiencies were so low that this could not be determined. The extended pGEM DNA (+/- homopolymeric tail) were ligated by T4 DNA Ligase (room temperature for 2 days), and then transformed to DH 5 alpha competent cells. Cells transformed with ten µl and 100 µl of each DNA were applied to ampicillin plates (50 µl X-gal / plate), which were incubated at 37 °C for 24 hours. The numbers of white versus total transformants are shown in Table 1.

Table 1. White colonies generated by transformation of extended pGEM

	w/o extension	Random extension	poly A	poly C	poly G	poly T	polyC+G
10 µl DNA (W/T)	6394	7/49	0/5	0/1	0/1	0/1	0/0
90 µl DNA (W/T)	57546	65/500	1/143	1/86	8/150	2/100	0/12
Percentage of White colonies	--	13%	0.75%	1.2%	5.3%	2%	0%
Transformation Efficiency (Transformants/µ g DNA)	63940	549	296	174	302	202	24
Ligation efficiency (compared to unextended)		0.9%	0.5%	0.3%	0.5%	0.3%	0.03%

W: # of white colonies; T: # of total colonies

Diversity of extension size generated by TdT (Figure 15 and 16)

The white colonies were selected and put into 1.5 ml LB broth in eppendorf tubes with 50 µg/ml Ampicillin and grown for 48 hours. Purified plasmid DNAs were cut by *Eco*R I and *Hind* III overnight, and run on 1% Agarose gel to check the extension size generated by TdT. The DNA of blue colonies from each plate was used as negative control. Lambda DNA digested by Hind III was used as molecular marker. The following are the abbreviations used in Figures 15 and 16. W: White colony; B: Blue colony; EL: Extension + Ligation; A: +dATP; C: +dCTP; G: +dGTP; T: +dTTP; CG: dCTP + dGTP + replication then ligation.

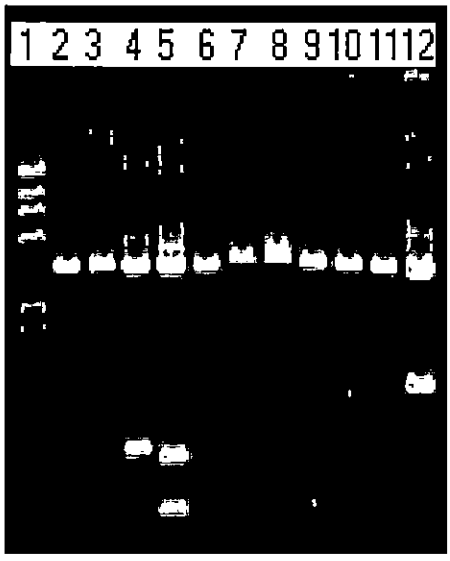


Figure 15. *Hind*III/*Eco*RI digests of transformant DNA. Lane 1, lambda DNA + *Hind*III. Lane 2, CGB2. Lane 3, CGB3. Lane 4, TW1. Lane 5, TW2. Lane 6, TB1. Lane 7, TB2. Lane 8, TB3. Lane 9, ELB1. Lane 10, ELB2. Lane 11, ELB3. Lane 12, ELW1.

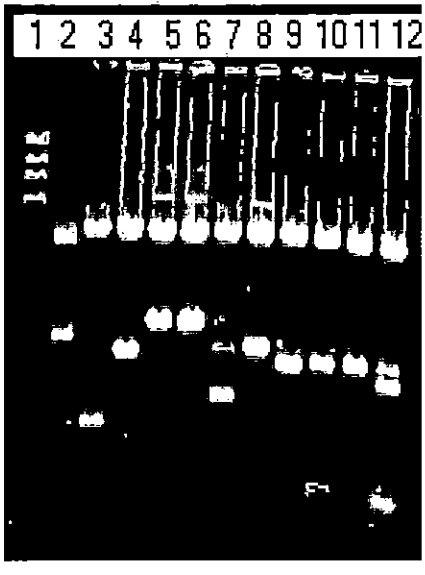


Figure 16. More *Hind*III/*Eco*RI digestions of transformant DNAs. Lane 1, lambda DNA + *Hind*III. Lane 2-12, ELW2-12.

Most of the DNA from white colonies (except those poly G white colonies) showed there is an extension generated by TdT, with the DNA size from 500 bp to 1500 bp. Some of the lanes even showed two bands in those white colonies. Such extension bands were never seen in lanes of blue colony DNA, suggesting no extended pGEM occured in those blue colonies.

Extension size of white colonies (Table 2)

Migration distance of DNA is inversely proportional to the log of the size (MW, bp) in the gel electrophoresis. A molecular size marker (Lambda DNA *Hin*d III digest) was used to generate a standard curve of mobility vs. log bp and the standard curve was used to calculate the size of the small fragments (cut by *Eco*R I and *Hin*d III) isolated from each white colony separated on the gel.

Shown below in Table 2 are the sizes of the random region *Hind*III/*Eco*RI digest fragments determined from white colonies as indicated.

Table 2. Extension size of white colony inserts

Colony name	Insert size (bp)	Colony name	Insert size (bp)
CW1	856	ELW8	1113
AW1	1065	ELW9	938
AW2	669	ELW10	1357
TW1	538	ELW11	993
TW2	836	ELW12	1056
ELW1	1075	ELW13	625
ELW2	1247	ELW14	342
ELW3	437	ELW16	1747
ELW4	993	ELW17	437
ELW5	1561	ELW18	342
ELW6	1476	ELW19	703
ELW7	625	ELW20	836

Restriction digest by *Pst* I on extended pGEM (Figure 17)

In order to investigate if the *Pst* I recognition site is interrupted by TdT extension on pGEM DNA, *Pst* I was used to digest every extended pGEM at 37°C overnight. Only two out of 28 isolate's random regions of the extended pGEM could be excised by *Pst* I (see supercoiled DNA on most lanes), suggesting TdT extension on 3' termini generated by *Pst* I digestion on pGEM interrupted this recognition site. (Note: not all isolates shown.)

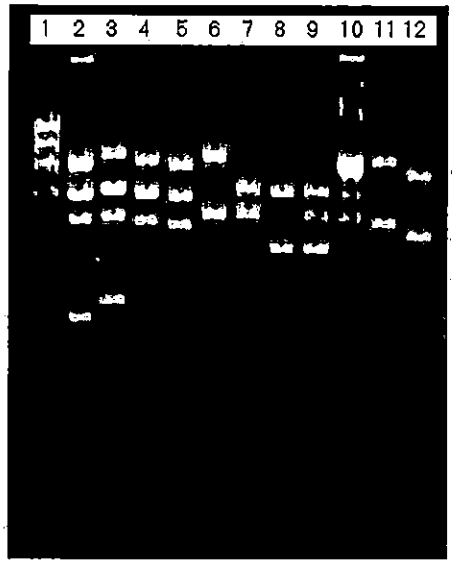


Figure 17. *Pst*I digest of extended random regions. Lane 1, lambda DNA + *Hind*III. Lane 2-12, different white colonies from different extensions. Lane 2, CW1, Lane 3, AW1. Lane 4, AW2. Lane 5 TW1, Lane 6, TW2. Lane 7, GW1. Lane 8, GW2. Lane 9, GW3. Lane 10-12, ELW 1-3, respectively. C, dCTP extension. A, dATP extension. T, dTTP extension. G, dGTP extension. EL, elongation and ligation extension.

Discussion

Libraries of random sequence polypeptides are useful as sources of antibody epitopes, novel ligands, enzyme substrates and potential lead compounds for the development of vaccines and therapeutics. The strength of this technology lies in the large amount of molecular diversity displayed that can be easily obtained and rapidly tested. As a result of screening peptide libraries, novel peptide sequences can be identified, which mimic native protective epitopes, and have the potential for use as drug candidates.

Current peptide libraries such as phage display or polysome display typically code for 6 to 38 random amino acids. They cannot display proteins with large random sequences due to the size restriction placed on them by oligonucleotide synthesis and synthetic peptide chemistry (Spencer and Tuerk, 1999). This size restriction limits the ability of current peptide libraries to find high affinity ligands due to the lack of folded motifs.

Terminal deoxynucleotidyl transferase (TdT) is a template-independent DNA polymerase catalyzing the irreversible terminal addition of deoxyribonucleotides to the 3'-OH termini of DNA. It was used by Damiani (1982) to create random DNA duplexes that were cloned and sequenced. In the presence of Co^{2+} (in 5X buffer), the preference of TdT for purines was avoided and it was allowed for longer extensions by increasing

primer binding. Spencer and Tuerk (1999) designed a new library construction technology based on the idea of polysome display. They used TdT to add over 1500 random nucleotides to the oligonucleotide primer TdT3P to generate long protein coding sequences. These ssDNA extension products were converted to dsDNA duplexes with T7 promoter sequences for transcription. Proteins translated from these coding sequences ranged from 30 kDa to 50 kDa, indicating that proteins range from 250-415 amino acids.

In this study, TdT was used to generate DNA sequence diversity on the PstI cut plasmid pGEM. These random extension products ranging from 342 to 1747 base pairs, correlated well with previous results (500 to 1500 bp extensions), predicting a peptide libraries coding sequence of large protein. In order to achieve individual extended pGEM, they have to be ligated and transformed to competent *E.coli* DH5 alpha cells. Because less than 1% of the bacterial cells will take up the plasmid DNA, transformation efficiency is a key factor limiting the diversity of peptide and protein libraries (Hanes and Pluckthun, 1997).Transformation efficiency is defined as the number of transformants formed per 1 μg plasmid DNA. In our experiment system, we put 1 μg extended and ligated pGEM, and 0.5 μg extended, ligated polymeric pGEM (including polyA, polyC, polyG, and polyT) into 75 μl DH5 alpha cells to perform transformation. The transformation efficiencies are between 174 transformants to 549 transformants/ug DNA (Table 1). Even more striking is the effect of extension on ligatability of plasmids compared to DNA that was *PstI* cut and NOT extended by TdT. The ligation efficiency for the *PstI* cut and religated pGEM was 63940 compared to 549 for extended. This

dramatically illustrates how difficult it is to ligate these extended single-stranded DNAs and suggests that for this system to be useful, some additional modifications to the procedure must be instituted. Without such modifications, the major part of the diversity is lost because of failure to ligate.

In our experiments, T4 DNA Ligase was used to ligate both "sticky" ends and "blunt" ends. After the random extension by TdT on pGEM (cut by *Pst* I) free 3' termini on double strands, a hetero-duplex DNA was produced because of the template-independency of TdT. From the above calculations it is obvious that the substrate presented in our experiment is poorly recognized or processed by DNA ligase.

A simple statistical analysis can provide information about what happened during these ligations as illustrated in Figure 18. The frequency with which one expects to find a nucleotide at a specific position is ¼. If four bases are required for ligation, then the frequency of ligatable ends would be $(¼)^4$ or $3.9X10^{-3}$. If five bases are required the frequency is $1X10^{-3}$. We produced 65 white colonies from 0.9 μg of extended DNA (90 μl in Table 2) and 57000 transformants from 0.9 μg on unextended plasmid DNA. That frequency is $1.14X10^{-3}$ suggesting very strongly that five bases of complementarity are required for ligation in our system.

As for the rest of the heteroduplex, it is difficult to determine how they appeared subsequent to transformation. Probably there are two possibilities: they may just form a bubble on plasmid dsDNA, or they may develop some regions of base pairing. This

complicated mechanism can only be clear after the extended DNA fragment sequencing is done.

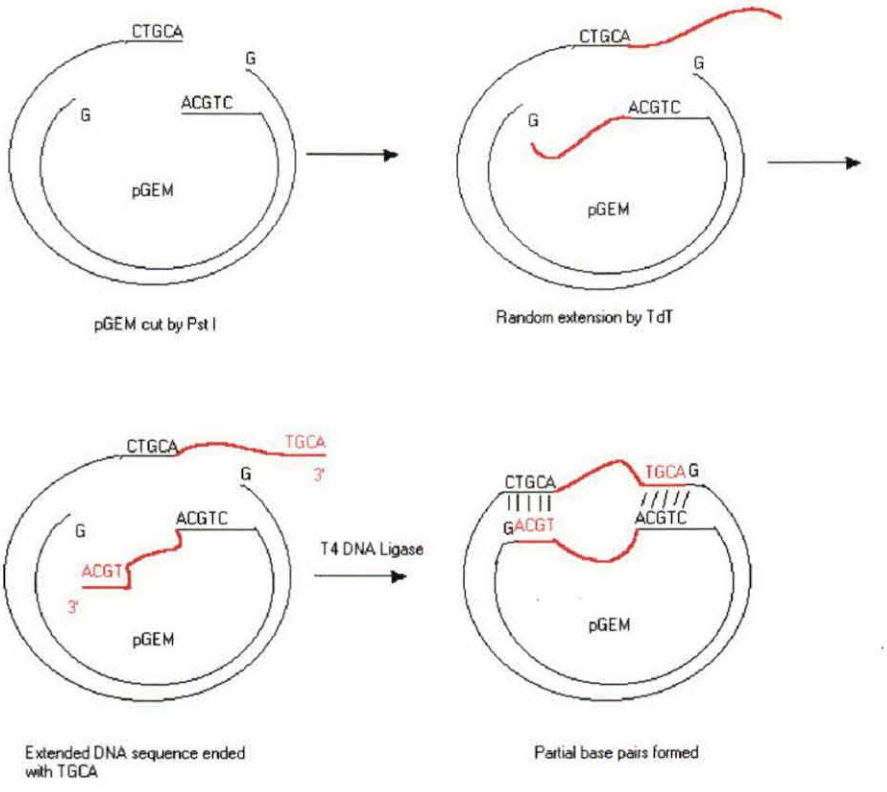


Figure 18. Hypothetical ligation of extended DNAs.

The rest of this analysis assumes that the 5 bases are required for efficient ligation. Because *Pst* I digestion produce a 3' sticky end, the five base sequence **CTGCA** exists at the junction point of the plasmids. The proportion of these ligatable junctions that recreate a *Pst*I site can be estimated by the following logic. In Figure 19, X is the first deoxynucleotide on extension sequence 1 added by TdT, and it may be any of the 4

deoxynucleotides: A, C, G, and T. So, there is 25% chance for G to be added. To regenerate the *Pst*I site the complementary base must be a C (Y, in Figure 19), also added at the frequency of 25% at the end of the extension sequence just prior to the required 3'-ACGT-5' for ligation. The expected frequency of that CG pair which would allow religation and regeneration of a *Pst*I site is therefore the product of the individual frequencies for X=G and Y=C or 1/8. Two clones could be digested at two sites releasing a fragment (Figure 17) and two clones (not shown) had a single *Pst*I site. This is 6/56 *Pst*I sites for ligatable ends or a frequency of 0.107 compared to 0.125 as expected statistically.

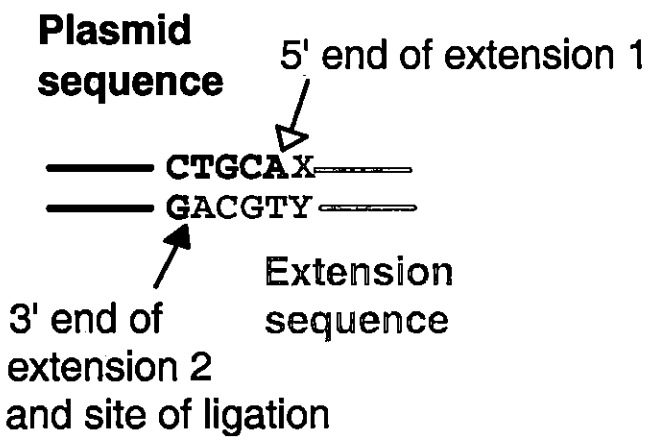


Figure 19. Hypothetical recreation of *Pst*I sites

Conclusions

Large random regions were successfully generated using terminal deoxynucleotidyl transferase on the plasmid pGEM. A small population of these plasmids can be religated and it is statistically probable that 5 bases of complementarity are required for ligation to close the circle. Corroborating this statistical inference is the frequency of *PstI* sites that are regenerated by extensions as a subpopulation of the ligatable plasmids. In order to harvest the diversity formed by TdT in a plasmid system such as this, redesign of the plasmid vectors or more complicated enzymatic manipulation of these extensions will be necessary in future experiments.

Literature Cited

Cull, M.G., J.F. Miller, & P.J. Schatz. 1992. Screening for receptor ligands using large libraries of peptides linked to the C terminus of the lac repressor. *Proceedings of the National Academy of Science* **89**: 1865-1869

Damiani, G., Scovassi, I., Romagnoli, S., Palla, E., Bertazzoni, U., and Sgaramella, V. 1982. Sequence analysis of heteropolymeric DNA synthesized in vitro by the enzyme terminal deoxynucleotidyl transferase and cloned in Esherichia coli. *Nucleic Acid Research* **10**: 6401-6405.

Devlin, J.J., L.C. Panganiban, , P.E. Delvin, . 1990. Random peptide libraries: a source of specific protein binding molecules. *Science* **249**: 404-406.

Dolle, Roland E. 2000. Comprehensive survey of combinatorial library synthesis: 1999 *Journal of Combinatorial Chemistry* **2** (5): 383-433.

FitzGerald, K. 2000. In vitro display technologies-new tools for drug discovery. *Drug Discovery Today*. **5**(6): 253-258.

Gates, C.M., W.P.C. Stemmer, R. Kaptein, and P.J. Schatz. 1996. Affinity selective isolation of ligands from peptide libraries through display on a *lac* repressor "Headpiece Dimer". *Journal of Molecular Biology* **255**: 373-386.

Hanes, J. and A. Plückthun. 1997. In vitro selection and evolution of functional proteins by using ribosome display. *Proceedings of the National Academy of Science* **94**: 4937-4942.

Hanes, J., C. Schaffitzel, A. Knappik, A. Plückthun. 2000. Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. *Nature Biotechnology* **18**: November.

Irving RA, Kortt AA, Hudson PJ. 1996. Affinity maturation of recombinant antibodies using E. coli mutator cells. *Immunotechnology* **2**(2):127-43

Kay, B., Adey, N., He, Y., Manfredi, J., Mataragnon, A., and Fowlkes, D. 1993. An M13 phage library displaying random 38-amino-acid peptides as a souce of novel sequences with affinity to selected targets. *Gene* **128**: 59-65.

Ladner, R. 1995. Constrained peptides as binding entities. *Trends in Biotechnology* **13**: 426-430.

Li, Min. 2000. Applications of display technology in protein analysis. *Nature Biotechnology* **18**: December 1251-1256.

Liu, Y. and E. Haggard-Ljungquist. 1994. Studies of bacteriophage P2 DNA replication: localization of the cleavage site of the A protein. *Nucleic Acids Research* **22**: 5204-5210.

Martens, C., Cwirla, S., Lee, R., Whitehorn, E., Chen, Esther, Bakker, A., Martin, E., Wagstrom, C., Gopalan, P., Smith, C., Tate, E., Koller, K., Schatz, P., Dower, W., and Barrett, R. 1995. peptides which bind to E-selectin and block neutrophil adhesion. *Journal of Biological Chemistry* **270**: 21129-21136.

Mattheakis, L., Bhatt., R., and Dower, W. 1994. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proceedings of the National Academy of Science* **91**: 9022-9026 .

Matthews, D., and Wells, J. 1993. Substrate phage: selection of protease substrates by monovalent phage display. *Science* **260**: 1113-1117.

New England Biolabs 2000-2001 Catalog and technical Reference.

Roberts, R.W. and J.W. Szostak. 1997. RNA-peptide Fusions for the in vitro Selection of Peptides and Proteins. *Proceedings of the National Academy of Science* **94**(23):12297-12302.

Sambrook, J., D.F. Fritsch, and T. Maniatis. 1989. Molecular Cloning: A laboratory manual. 2 nd Edition. Cold Spring Harbor Press, New York.

Scott, J.K. and J.P. Smith. 1990. Searching for peptide ligands with an epitope library *Science* **249**: 386-389.

Singer, B.S., T. Shtatland, D.Brown, and L. Gold. 1997. Libraries for genomic SELEX. *Nucleic Acids Research* **25**(4): 781-786.

Spencer, ML. and C. Tuerk. 1999. Translation of random transcripts generated by TdT: potential use in polysome peptide libraries. *SAAS bulletin: Biochemistry and Biotechnology* **12**: 23-28.

Tuerk, C. and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**: 505-510.

Viguera E, Canceill D, Ehrlich SD 2001. In vitro replication slippage by DNA polymerases from thermophilic organisms. *J Mol Biol* **312**(2):323-33