J. Bioactive and Compatible Polymers
(Submitted)

Genetic Synthesis of Periodic Protein Materials

M.J. Fournier[1,3,#], H.S. creel[2], K.P. McGrath[2,+], M.T. Krejchi[2],
E.D.T. Atkins[4], T.L. Mason[1,3] and D.A. Tirrell[2,3]

[1]Department of Biochemistry, [2]Department of Polymer Science and Engineering, [3]Program in Molecular and Cellular Biology

Lederle Graduate Research Center

University of Massachusetts, Amherst, MA 01003

413-545-2732

+,* Permanent address: +Department of the Army,
Natick Research Development and
Engineering Center
Natick, MA 01760

*H.H. Wills Physics Laboratory,
University of Bristol,
Bristol BS8 1TL, UK

corresponding author:

Running Title: Recombinant protein materials

REVIEW

1. INTRODUCTION

Genetic engineering offers a novel approach to the development of advanced polymeric materials, in particular protein-based materials. Biological synthesis provides levels of control of polymer chain architecture that cannot yet be attained by current methods of chemical synthesis. In addition to employing naturally occurring genetic templates artificial genes can be designed to encode completely new materials with customized properties. In the present paper we: 1) review the concepts and technology of creating protein-based materials by genetic engineering, 2) discuss the merits of producing crystalline lamellar proteins by this approach, and 3) review progress made by our group in generating such materials by genetic strategies. Full descriptions appear elsewhere about the parameters to be considered in designing artificial protein genes of this type, the effectiveness of different gene construction and expression strategies utilized by us thus far and, the specific properties of the various materials derived from these efforts (1,2).

Progress made by other groups involved in developing periodic proteins by molecular biological strategies are described in refs. 3-8. The latter studies include genetic engineering of artificial silk-like proteins (3,4), poly-aspartylphenylalanine (5), an α/β barrel domain (octarellin; 6), the collagen tripeptide GlyProPro (7) and human tropoelastin (8). Advances with the silk-like proteins (SLP) have been particularly impressive. In addition to producing multi-gram quantities of pure SLP homopolymers, this group has successfully generated block copolymers of SLP interspersed with core peptides of mammalian elastin and the human fibronectin cell attachment element. While publications are still lacking it appears that a number of groups are striving to create genetically engineered variants of the repetitive bioadhesive proteins produced by mussels and barnacles (9).

2. BACKGROUND

2.1 A Role For Genetic Engineering In Materials Science

The merits of protein-based materials have been well established for a few natural proteins, in particular the silks, elastin, collagen and marine bioadhesives. With the advent of genetic engineering technology it is now possible to produce these and other proteins with materials potential from both natural and artificial genes. In addition to providing an alternative means for producing important natural proteins, recombinant DNA technology offers the exciting potential of creating completely new proteins with wide ranging properties. Synthetic proteins can be designed from first principles drawn from the disciplines of protein biochemistry and materials science. At least initially artificial proteins would logically feature structural motifs that occur in natural proteins, including a-helices, β-pleated sheets and β-turns. Using chemically synthesized genetic material it should be possible to engineer entirely new classes of protein-based materials in which natural protein elements are joined together in creative and novel ways. The properties of these materials could be expanded yet further by subsequent chemical or biochemical processing.

Key benefits of protein-based materials include biocompatibility and biodegradability. While such materials can be produced by chemical or biological strategies (or a combination of both) biological synthesis offers superior control over the strictly chemical methodologies presently available. Current chemical methods provide less precise control of molecular architecture and also suffer from size and yield limitations. On the other hand proteins made biologically will be monodisperse in size and sequence - because of the presence of defined start and stop signals in the genetic template and the extraordinary specificity and accuracy of the protein biosynthetic machinery. In addition, biologically-derived products will be pure stereochemically as only L-amino acids are utilized.

2.2 Genetic Svnthesis of Protein Materials

A generic approach to developing recombinant protein materials is depicted in Figure 1. Concepts from materials science and protein structure are used to design a product with specific properties. An appropriate genetic template consisting of ribonucleic acid (RNA) is then designed, from the known genetic code words (codons) for the amino acid sequence desired. The sequence of nucleotides in the messenger RNA template, i.e., mRNA, dictates the DNA sequence of the artificial gene. Double-stranded DNA encoding the desired protein is chemically synthesized and installed in an appropriate DNA vector molecule. It is the role of the vector to ensure that the synthetic coding segment is stably maintained and expressed in the recombinant host organism. Vector DNA contains information for self-replication, i.e., DNA synthesis, and the new coding sequence is flanked by DNA signals for production of mRNA (transcription) and decoding of the mRNA to yield the desired protein (translation). The host cell currently favored for expression of recombinant proteins is the bacterium *Escherichia coli.* A superior base of molecular genetic knowledge exists for E. *coli* and growth and processing technologies are well established for recombinant products expressed by this organism.

In addition to the actual protein sequence decisions about the design of a synthetic protein gene must also include: 1) consideration of host cell preference for specific codons (there are from one to six codons per amino acid); 2) potential untoward effects on expression (e.g., intramolecular folding of mRNA which could impair translation and composition of an early mRNA region known to influence activity); and, 3) the strategy for cloning.

Thus, the final cloning strategy is based on: 1) how the synthetic DNA will be joined enzymatically to the cloning and expression vectors; 2) analysis of the cloned DNA by sequencing - to verify correct construction; 3) possible re-engineering to create related variant products; 4) the expression strategy selected; and 5) the scheme for purifying the product protein. Expression is best assured by fusing the artificial coding sequence to an upstream gene segment that specifies a portion of a natural protein produced at high efficiency by E. *coli.* In these cases the recombinant product will be a bi- or tripartite fusion protein with a foreign peptide moiety at the amino end and possibly another at the carboxy terminus. The fusion segments can provide stability against damage by host cell proteases. These 'tails' can also be exploited in specific purification strategies, for example, as ligands in affinity purification. Design considerations also take into account the eventual goal of removing fusion peptides, by chemical or enzymatic cleavage. Proven approaches include methionine-specific hydrolysis with CNBr and cleavage with proteolytic enzymes, both at sites engineered into the recombinant protein sequence.

2.3 The Case For Periodic Proteins

Our initial efforts in this area have focused on exploring the potential for producing recombinant proteins that form crystalline lamellar materials of defined thickness and surface function (Fig. 2). These proteins feature repeating stem-turn elements predicted to form antiparallel @-pleated sheets. The sheets, in turn, are expected to stack to create growth in the third dimension. The stem elements correspond to repeating alanine and glycine dyads (AlaGly or AG) known to form @-strands in both natural protein (e.g. silk fibroin) and poly(AG) produced chemically (4,5). The @-strands are interrupted at regular intervals with amino acids anticipated to encourage formation of reverse turns (@-turns). The width of the resulting @-sheet and thickness of the corresponding lamellar slab is determined by the length of the β-stem. The turn residues are predicted to populate the surface of the lamellae and thereby define surface functionality. Among the 20 natural amino acids there is a broad range of side groups that can be featured at the surface, including: alkyl, aryl, -OH, -COOH, $-NH_2$, and -SH moieties. One of the major aims of our program is to assess the potential for extending the range of synthetic potential through biological incorporation of unnatural amino acids, some of which are already known to be utilized by the E. coli protein synthetic apparatus.

Our protein engineering strategy is being evaluated with artificial genes that encode sequential peptides of the general sequence:

1 $-[(GlyAla)_3L_1L_2]_n-$

where L_1, L_2 are turn residues.

3. ENGINEERING OF REPETITIVE PROTEINS

3.1 Strategy For Constructing Recombinant Genes

The feasibilty of our approach was first tested with DNA encoding the protein repeats

2 $-(GlyAla)_3GlyProGlu-$

3 $-(GlyAla)_4GlyProGlu-$

Proline is known to disrupt chain folding and both proline and glutamic acid were expected to be excluded from the β-sheet. Additionally, the presence of glutamate in the turn sequence would functionalize the hypothetical lamellar surface.

The scheme used to assemble a synthetic DNA fragment encoding peptide repeat 3 is shown in Figures 3 and 4. Stepwise, the strategy involved: 1) chemical synthesis of double stranded DNA encoding two repeats of 3 ; 2) in vitro ligation of this DNA into a plasmid cloning vector; 3) introduction of the recombinant vector-insert DNA into E. coli, by transformation; 4) recovery of the recombinant plasmid; 5) sequencing of the DNA insert; 6) in vitro multimerization of the synthetic DNA fragment, to create larger coding sequences; and, 7) cloning of size-selected DNA multimers into an expression vector.

DNA synthesis was carried out with an automated DNA synthesizer. The synthetic DNA included three different types of DNA restriction endonuclease recognition/cleavage sites - two at the very ends for cloning (*Eco*RI and *Bam*HI) and two flanking restriction sequences for subsequent multimerization (*Ban*I). The artificial DNA was inserted into a standard cloning vector, pUC18 (Fig. 3). Insert DNA from one clone was determined to have the correct sequence and this fragment was excised and self-ligated to create a family of multimers (Fig. 5). The population of multimers, ranging upwards of 20 DNA monomers was then cloned into a *Ban*I site of a small, high copy vector p937.51. Coding sequences inserted into this vector are endowed with flanking codons for methionine, making possible eventual cleavage of heterologous leader and trailer peptides with cyanogen bromide. Insert DNA was then cloned into the expression vector pET3-b (at the *Bam*HI site; Fig. 5). One clone encoding 14 DNA repeats of the undecapeptide sequence 3) was selected for additional characterization.

Inserts in the pET3-b vector are expressed by RNA and protein synthesis signals from the E. coli bacteriophage T7, in particular from signals of a coat protein gene, i.e., gene 10. The T7 virus elements include a very strong promoter signal for transcriptional initiation, a transcriptional stop signal, and start and stop signals for translation of the gene 10 coat protein; the translation signals include; 1) a ribosome binding site (Shine-Dalgarno sequence), 2) an important 'leader' segment that encompasses the ribosome recognition segment and early mRNA region and, 3) the translational start codon AUG, specific for methionine. The fusion specifies a hybrid protein with 11 amino acids of the gene 10 protein at the amino terminal end and 19 amino acids at the carboxy terminus.

The chromosome of the host bacterium includes a gene for T7 phage RNA polymerase, an enzyme that has robust activity and is specific for T7 phage promoters; E. coli promoters are ignored. Synthesis of T7 polymerase is triggered by addition of a chemical inducer (isopropyl-β-thiogalactoside) which inactivates a negative controlling protein, i.e., a repressor. T7 polymerase is produced and DNA encoding the synthetic protein is transcribed. The resulting mRNA, in turn, is decoded by the bacterial translation machinery.

3.2 Stabilitv of the Artificial Gene and Protein Expression

Expression plasmids encoding 14 repeats of peptide 3 showed no sign of being genetically unstable. Over the course of many cell culturings recombinant vector was maintained in the host cells and the insert DNA did not undergo any apparent rearrangement. Similar stability has been observed for a plasmid containing 54 repeats of peptide 2. Using in vivo radiolabeling recombinant proteins 2 and 3 were observed to accumulate at relatively high levels following induction of expression; data for repetitive peptide 3 are shown in Fig. 6. No material corresponding to putative protein 3 was detected in cells lacking recombinant plasmid or for cells harboring vector without the artificial DNA insert. Comparisons of protein patterns for both induced and non-induced cells indicated that the protein 3 product accumulated to levels corresponding to several percent of total cell protein. The recombinant protein appears to occur as a single electrophoretic band suggesting that it is resistant to damage by cellular proteolytic enzymes. The apparent size of the protein is 40 kDa, larger than the 17.2 kDa deduced from the coding sequence. Although the basis of the anomalously slow migration is not known this behavior has been observed for other proteins featuring related sequence elements, including both natural and recombinant silk fibroins (J. Cappello, pers. commun.). Similar expression results have been obtained for the peptide 2 product and this protein also exhibited anomalously slow electrophoretic mobility (McGrath et al, submitted).

A gene encoding another peptide repeat 4 has also been successfully assembled and expressed, by the strategy outlined above for the product 3 protein. This construct features the same β-strand sequence, but a turn of two amino acids rather than three. The basic repeat is

$$4 \qquad -(AlaGly)_3GluGly-$$

Product corresponding to 36 repeats of 4 has been produced with the pET3-b system. This protein also accumulates at levels similar to that observed for the product 3 variant and, like proteins 2 and 3, also migrates with a lower electrophoretic mobility than predicted.

4. PROPERTIES OF THE SEQUENTIAL PROTEINS

Repetitive polymers 2, 3 and 4 have been partially characterized. The structures of the recombinant proteins have been verified by: 1) amino acid compositional analysis (2,3,4); 2) N-terminal protein sequencing - through 58 residues (2); 3) matrix-assisted laser desorption mass spectrometry (2,3); 4) ^{1}H and ^{13}C NMR spectrometry (2,3,4); 5) cyanogen bromide cleavage (2,3,4); and combustion analysis (2,3,4). Taken together, the results from these analyses demonstrate that the biosynthetic strategy is sound and effective. While the correct proteins were produced there was no evidence that proteins 2 or 3 are able to assume the desired @-sheet or ordered lamellar structure of interest. Characterization by: 1) x-ray scattering, 2) Fourier transform infrared spectrometry and 3) differential scanning calorimetry yielded data indicating that both the primary fusion proteins and CNBr cleavage products form amorphous glasses at room temperature.

Preliminary x-ray diffraction patterns have been obtained for protein 4 precipitated from formic acid. While the diffraction analyses are still in progress the early data support the conclusion that sequence 4 polymers form β-sheets and that these sheets associate to form a membrane-like multi-lamellar solid. Taken together with computer-assisted modeling these results argue that linkers of three amino acids (or alternatively, repeating units comprising odd numbers of amino acids) do not allow adjoining β-strand stems to align in complete register. The results with repetitive peptide 4 indicate that this requirement can be satisfied by at least some two-amino acid turns.

5. CONCLUSIONS

We consider that the biological feasibility of our genetic strategy has been demonstrated. Artificial genes encoding repetitive peptide sequences have been synthesized, assembled and incorporated into host bacterial cells.

The particular genes constructed thus far are genetically stable and have been expressed accurately and with good efficiency by the T7-phage based expression vector adopted. Within the limits of the results available recombinant proteins containing sequences 2,3 and 4 appear to be resistant to degradation by host cell proteolytic enzymes. Finally, first data from x-ray diffraction analysis of a sequence 4 polymer have shown that repetitive proteins of the type featured here can form highly ordered materials. With the biological feasibility of our approach established, at least for the cases described, emphasis can now be shifted to specific issues of materials design and application.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Polymers and Genetics Programs of the National Science Foundation(DMR 8914359) and by the Materials Research Laboratory at the University of Massachusetts.

REFERENCES

1. K.P. McGrath, D.A. Tirrell, M. Kawai, T.L. Mason and M.J. Fournier, ***Biotechnol. Prog.*** **6**:188-192 (1990).

2. H.S. Creel, M.J. Fournier, T.L. Mason and D.A. Tirrell, ***Macromolecules.*** **24**:1213-1214 (1991).

3. J. Cappello, J. Crissman, M. Dorman, M. Mikolajczak, G. Textor, M. Marquet, and F. Ferrari, ***Biotechnol. Prog.*** **6**:198-202 (1990).

4. J. Cappello, J. Crissman, M. Dorman, M. Mikolajcak, G. Textor, M. Marquet and F. Ferrari, ***Materials Res. Soc. Symp. Proc.*** **174**:267-276 (1990).

5. M.T.Doel, M. Eaton, E.A. Cook, H. Lewis, T. Patel and N.H. Carey, ***Nucleic Acids Res.*** **8**:4575-4592 (1980).

6. K. Goraj, A. Renard, and J.A. Martial, ***Protein Engineering*** **3**:259-266 (1990).

7. I. Goldberg, A.J. Salerno, T. Patterson and J.I. Williams, ***Gene*** **80**:305-314 (1989).

8. Z. Indik, W.R. Abrams, U. Kucich, C.W. Gibson, R.P. Mecham and J. Rosenbloom, ***Arch. Biochem. Biophys.*** **280**:80-86 (1990).

9. S.C. Stinson, ***Chemical & Engineering News*** July 16: 26-32 (1990).

FIGURE LEGENDS

Figure 1. Production of novel protein-based materials by genetic engineering.

Figure 2. Structure of repetitive polypeptides designed to form crystalline lamellae of pre-determined thickness and surface function. Single polypeptide chains consist of repeating units of β-strands and linkers that form reverse turns. Chain folding creates anti-parallel β-pleated sheets, which in turn, self-associate to generate lamellar crystals. Lamellar thickness is determined by the length of the @-stems while surface function is speciifed by the amino acids in the @-turn elements.

Figure 3. Cloning of a synthetic DNA encoding two repeats of peptide sequence 3. The nucleotide sequence of the artificial DNA monomer is shown replete with restriction sites for cloning and subsequent ligation to form high molecular weight coding units. The DNA monomer was synthesized chemically and incorporated into the E. coli cloning vector pUC18 at the asymmetric *Eco*RI and *Bam*HI restriction sites. Bacteria containing recombinant plasmids were identified by blue-white color screening of colonies on medium containing an indicator dye - based on insertional inactivation of the gene for β-galactosidase(white). DNA inserts were examined by restriction enzyme analysis and DNA sequencing. Monomer DNA was prepared by digestion of hybrid plasmids with restriction enzyme BanI. DNA digests were fractionated by electrophoresis in gels of agarose or polyacrylamide.

Figure 4. Strategy for construction and expression of a synthetic gene specifying repetitive polypeptide 3. DNA multimers were produced by in vitro ligation of pUC18-derived BanI inserts and cloning into a second vector, p937.51 with adjoining sites for *Bam*HI. Inserts from this latter vector were cloned into the *Bam*HI site of expression vector pET3-b. In the final construct the artificial DNA segment was fused with the T7 phage gene 10 (ϕ10) protein coding sequence and upstream signals for ϕ10 transcription and translation. Recombinant vector pET3-x was introduced into E. coli strain BL21(DE3)pLysS by transformation. Expression of the target protein is achieved by induction of a chromosomal copy of the T7 RNA polymerase gene - by addition of isopropyl-β-D-thiogalactoside (IPTG); IPTG inactivates a repressor protein which prevents expression of the polymerase gene. Recombinant protein is produced in mid-logarithmic phase cultures and isolated from clarified cell lysates by differential precipitation with acetic acid and then ethanol (1,2).

Figure 5. Electrophoretic pattern of DNA multimers formed in vitro. BanI monomer fragments isolated from recombinant pUC18 plasmids were ligated with T4 phage DNA ligase and fractionated by polyacrylamide gel electrophoresis. Multimers containing upwards of 20 DNA repeats were observed. The repeat size is shown at the right. Numbers at the left correspond to DNA size standards measured in base pairs.

Figure 6. Electrophoretic analysis of cloned DNA 3 multimers. The DNA fragments shown are from different recombinant p937.51 clones. The number of peptide 3 repeats encoded by the cloned DNA fragments is shown at the top. The right lane contains DNA size markers; sizes are given at the right edge in base pairs.

Figure 7. Pattern of protein production by transformants containing recombinant pET3-b vectors encoding 54 repeats of peptide 2 and 14 repeats of peptide 3. Proteins were labeled in *vivo* with ^{3}H-glycine and fractionated on a gel of SDS-polyacrylamide. Lanes 1-4 are negative controls corresponding to cells lacking the expression plamid or artificial DNA insert; lanes 5 and 6 (pET3-27) contains protein from cells with an expression plasmid encoding 54-repeats of peptide 2 (from 27 dimeric DNA units; McGrath *et* al, submitted); lanes 7-11 (pET4-14) are protein samples from transformants containing an artificial coding unit specifying 14 repeats of peptide 3. Times of sampling after induction by IPTG addition are indicated at the top (in min.). The pattern shown is a fluorogram prepared by soaking the gel in fluor and exposure to x-ray film. Reproduced with permission from Ref. 2, ***Macromolecules.***

Fig 1

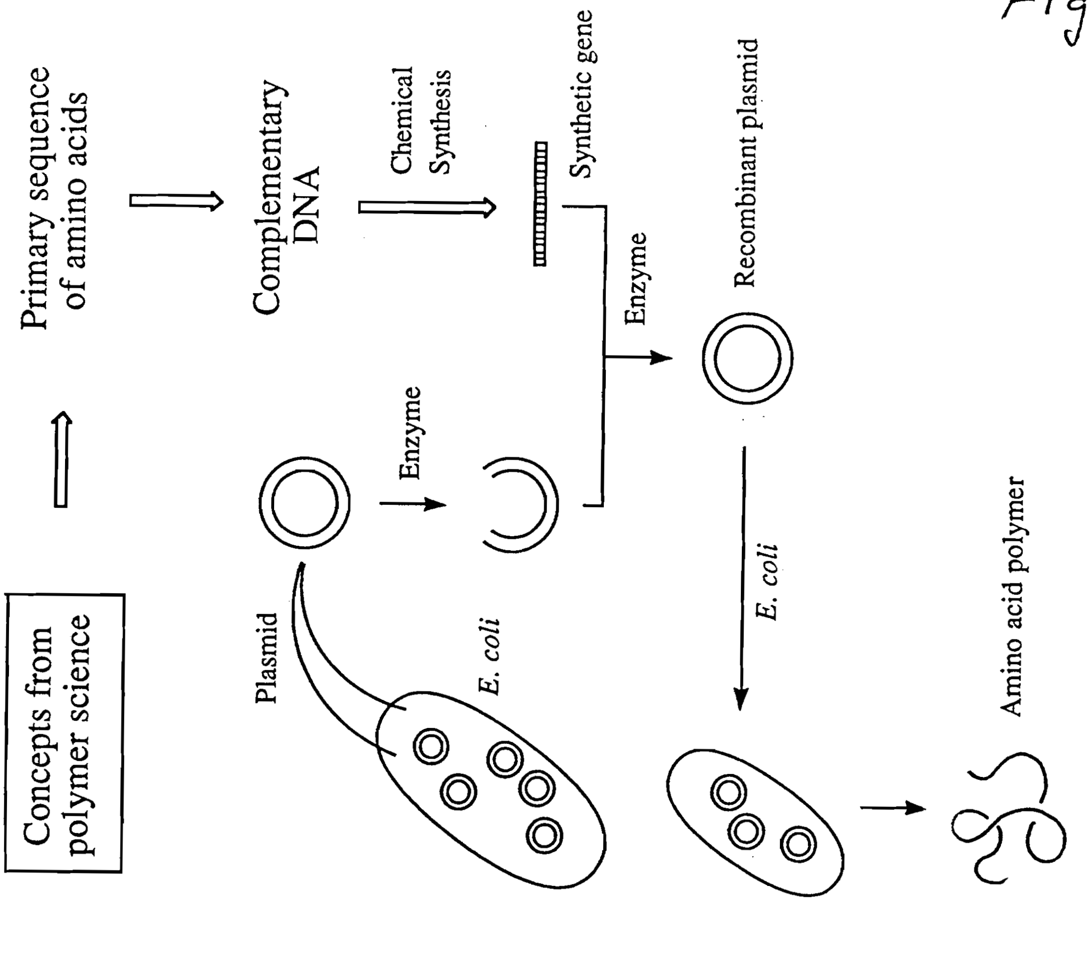

Concepts from polymer science
Primary sequence of amino acids
Complementary DNA
Chemical Synthesis
Synthetic gene
Enzyme
Recombinant plasmid
Plasmid
Enzyme
E. coli
E. coli
Amino acid polymer

Fig 2

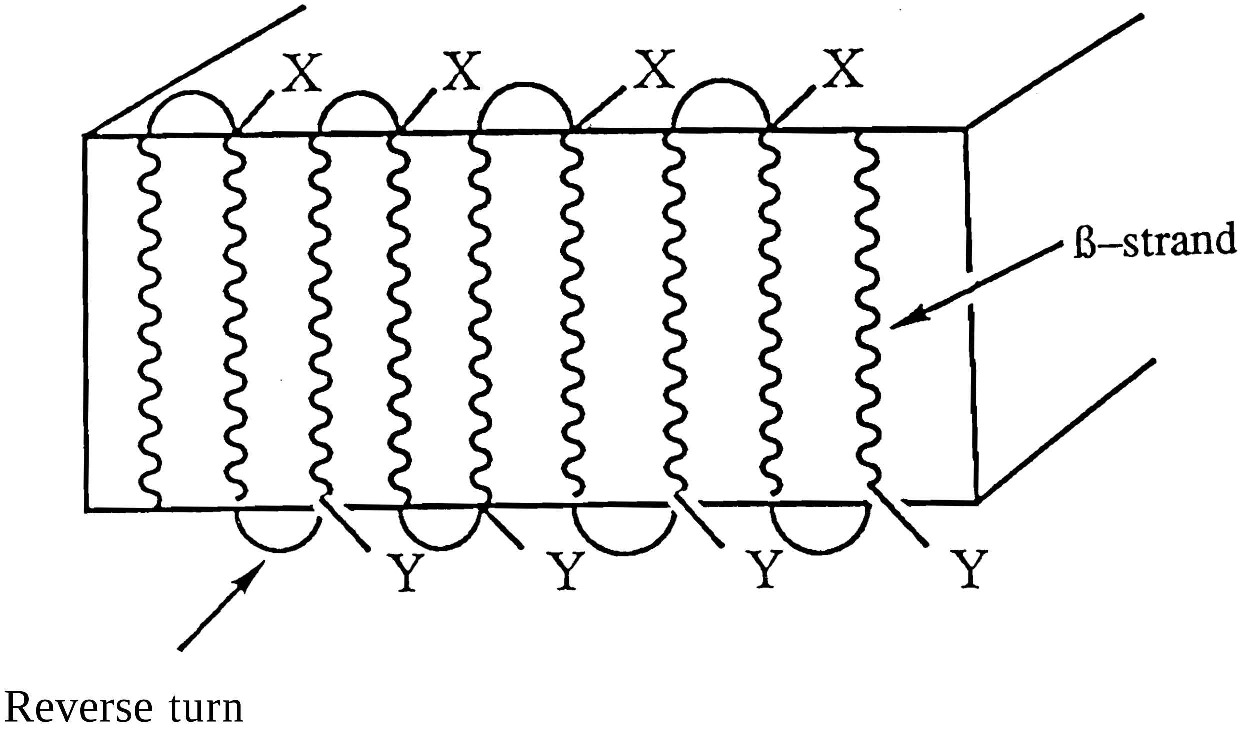

X
X
X
X
ß–strand
Y
Y
Y
Y
Reverse turn

Fig 3

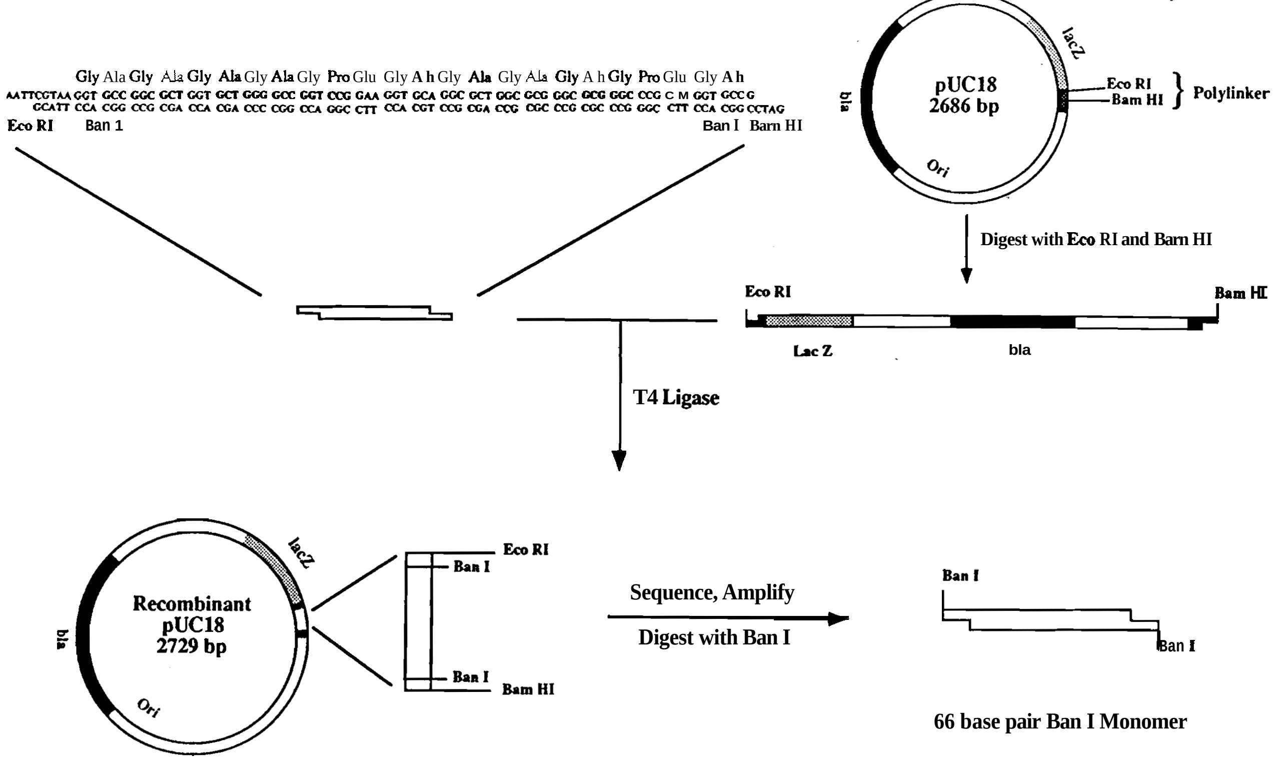

Gly Ala Gly Ala Gly Ala Gly Ala Gly Pro Glu Gly A h Gly Ala Gly Ala Gly A h Gly Pro Glu Gly A h
AATTCGTAA GGT GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT GCA GGC GCT GGC GCG GGC GCG GGC CCG C M GGT GCC G
GCATT CCA CGG CCG CGA CCA CGA CCC CGG CCA GGC CTT CCA CGT CCG CGA CCG CGC CCG CGC CCG GGC CTT CCA CGG CCTAG
Eco RI
Ban 1
Ban I
Barn HI
pUC18
2686 bp
lacZ
bla
Ori
Eco RI
Bam HI
Polylinker
Digest with Eco RI and Barn HI
Eco RI
Bam HI
Lac Z
bla
T4 Ligase
Recombinant
pUC18
2729 bp
lacZ
bla
Ori
Eco RI
Ban I
Ban I
Bam HI
Sequence, Amplify
Digest with Ban I
Ban I
Ban I
66 base pair Ban I Monomer

Fig 4

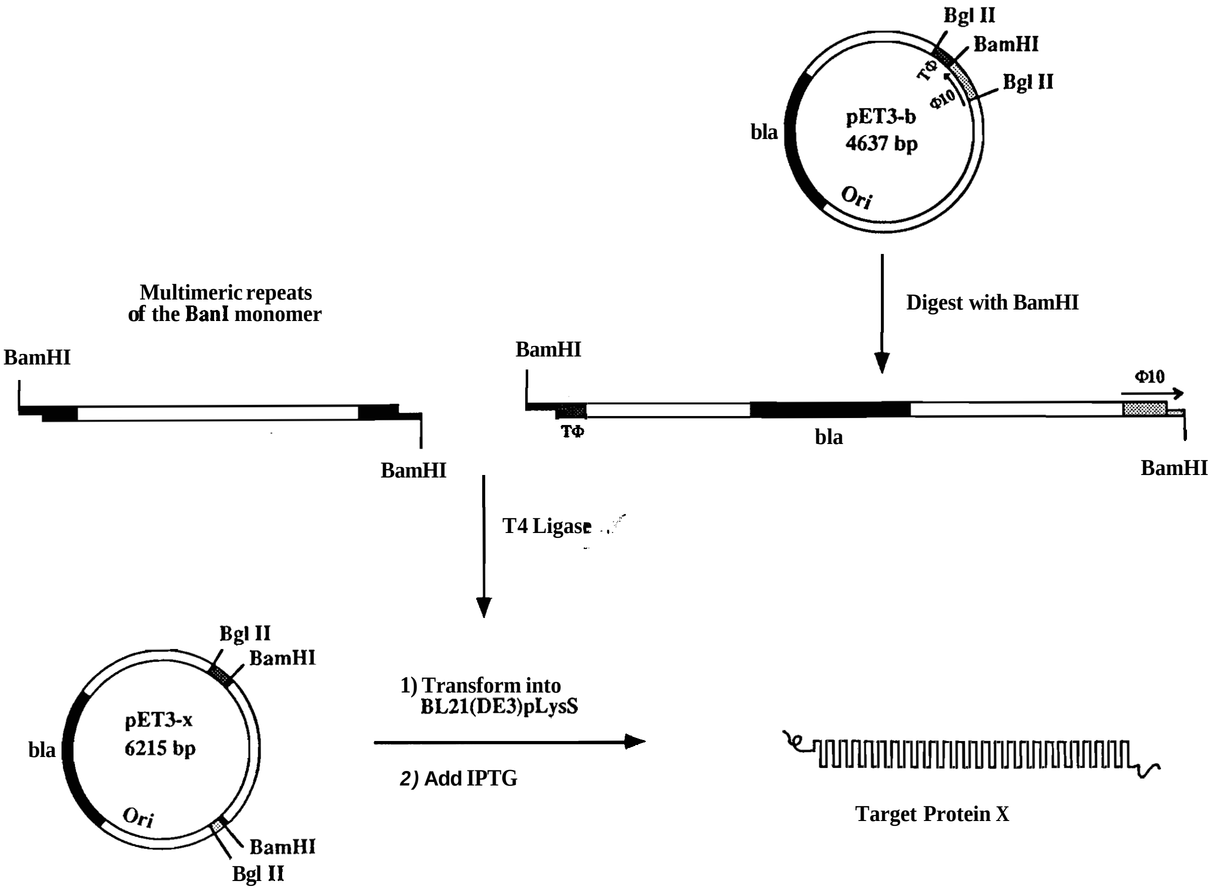

Bgl II
BamHI
Bgl II
TΦ
Φ10
bla
pET3-b
4637 bp
Ori
Digest with BamHI
Multimeric repeats
of the BanI monomer
BamHI
BamHI
BamHI
Φ10
TΦ
bla
BamHI
T4 Ligase
Bgl II
BamHI
pET3-x
6215 bp
bla
Ori
BamHI
Bgl II
1) Transform into
BL21(DE3)pLysS
2) Add IPTG
Target Protein X

Fig 5

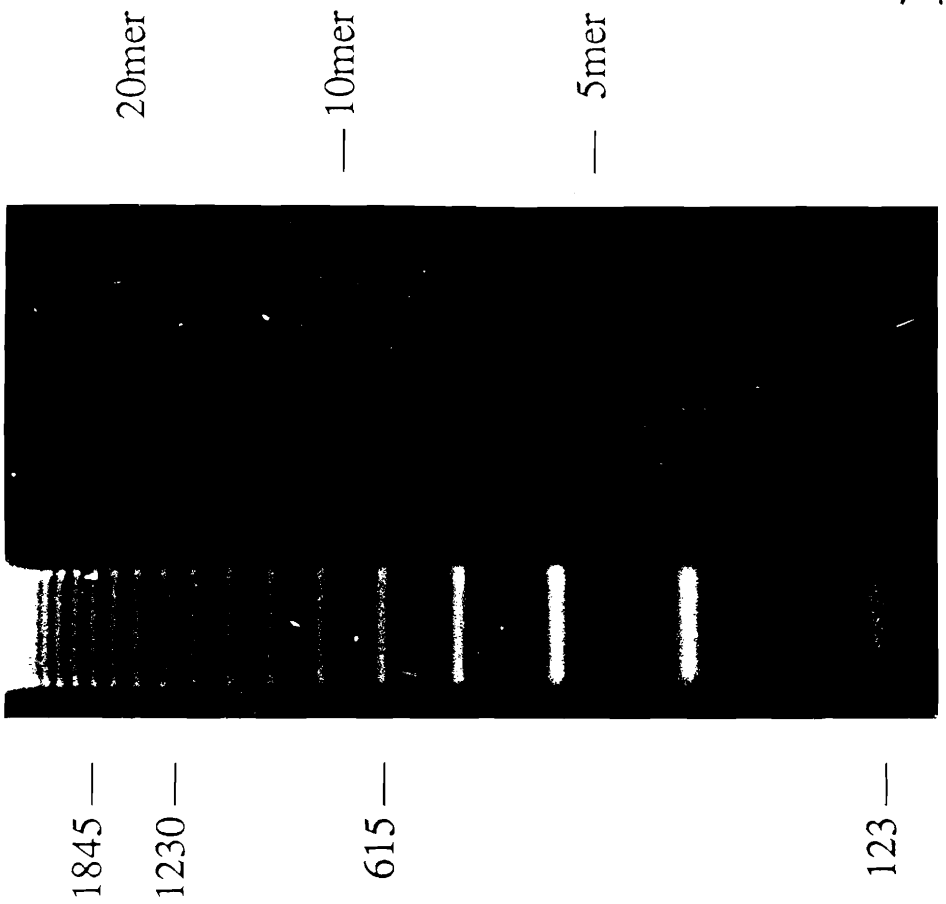


Fig 6

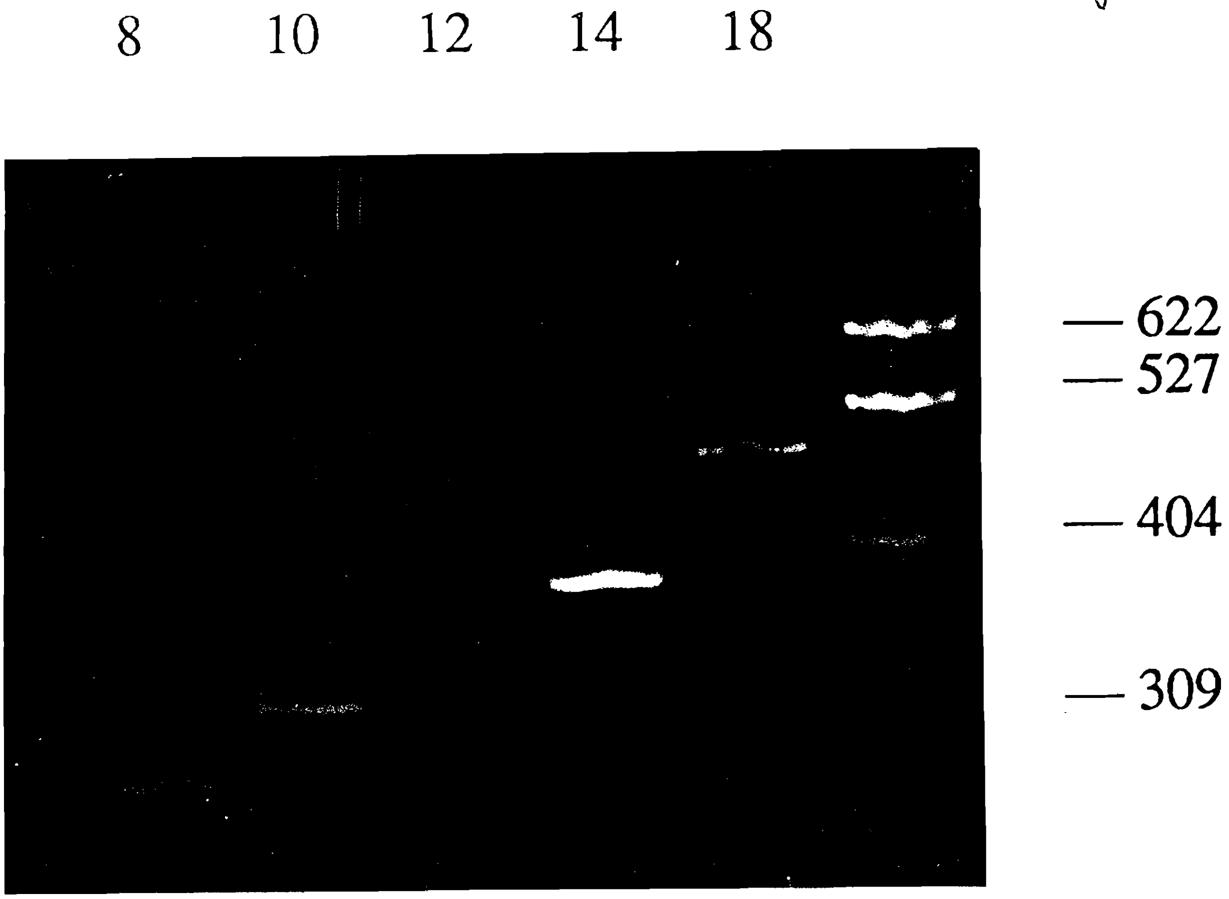
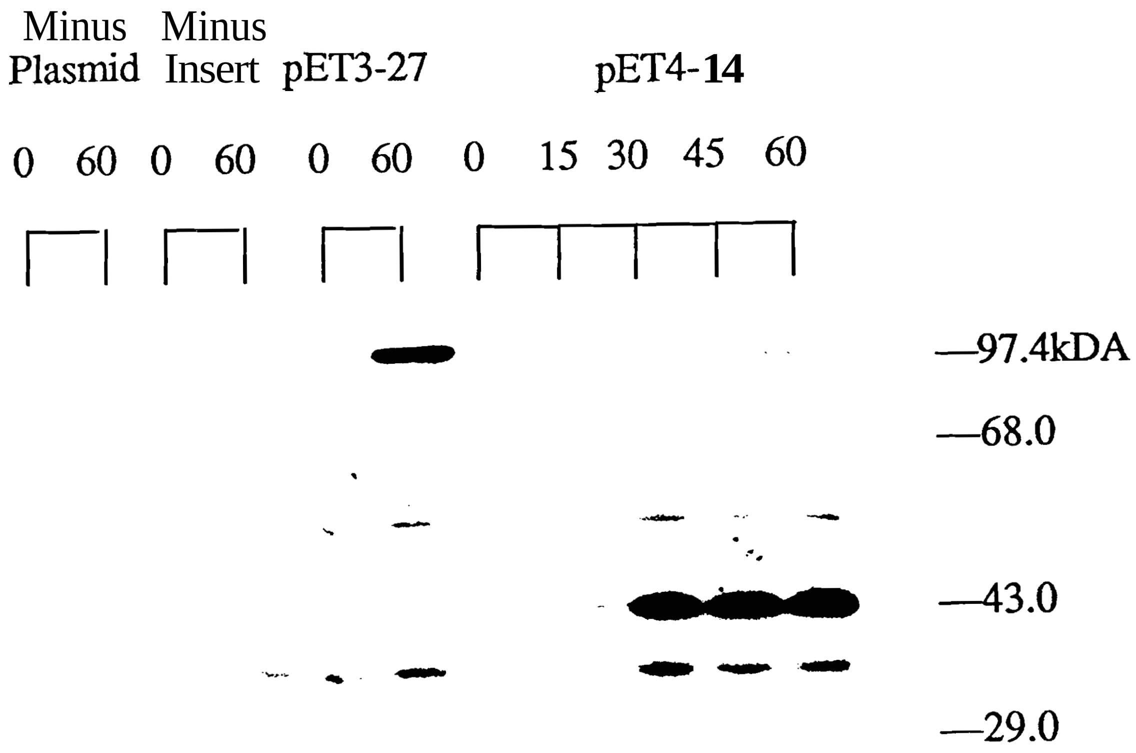


Fig 7


Minus Plasmid
Minus Insert
pET3-27
pET4-14
0
60
0
60
0
60
0
15
30
45
60
—97.4kDA
—68.0
—43.0
—29.0