

2 Chapter II: Construction of the synthetic *gusA*^{Ssp} gene and its variants

2.1 Introduction

2.1.1 Strategies for synthetic gene construction

From a technical viewpoint, synthetic gene construction simply means building up DNA molecules of desirable length from units of chemically synthesized oligonucleotides. Some small genes (coding for tRNAs or hormones, for example) were constructed as early as in the 70's (Khorana 1971; Caruthers et al. 1972; Kleppe et al. 1976; Itakura et al. 1977; Crea et al. 1978).

Although all construction strategies are based on the complementary nature of DNA to form longer "contigs", various methods of construction have been devised. In the original strategy introduced by Khorana, DNA in the double-stranded form is carefully divided into short single-stranded segments (oligonucleotides) with suitable overlaps in the complementary strands. All segments are then chemically synthesized. Appropriate segments are 5'-phosphorylated by polynucleotide kinase. All segments are annealed in aqueous solution, to form a complete duplex with single strand nicks, which are subsequently sealed off by polynucleotide ligase (Khorana 1979). An alternative strategy by Rossi et al. (1982) uses only partially overlapping oligonucleotides to form incomplete duplex with gaps on both strands. These gaps are then "filled in" by DNA polymerase I (Rossi et al. 1982; Rink et al. 1984).

Many variations/improvements of the two basic schemes mentioned above are readily devisable. For example, direct transformation of non-ligated duplex into *E. coli* for nick repair (Narang et al. 1986), transformation of partially single stranded DNA into *E. coli* for *in vivo* gap repair (Adams et al. 1988), use of long and chemically phosphorylated oligonucleotides (Wosnick et al. 1987), use of long oligonucleotides with short inverted

repeat at their 3'-end to serve as primer for synthesis of the second strand by DNA polymerase I (Uhlmann 1988; Georges et al. 1990), sequential ligation of overlapping oligonucleotides directly into cloning vector (Hayden & Mandecki 1988; Ivanov et al. 1990), ligation at high temperature (to avoid mispairing and formation of secondary structures) using thermostable *Pfu*-ligase (Fuhrmann et al. 1999). PCR techniques have also been used extensively in various steps of the construction process (Dillon & Rosen 1990; Ciccarelli et al. 1991; Ye et al. 1992; Sandhu et al. 1992; Au et al. 1998; Fuhrmann et al. 1999; Te'o et al. 2000; Chalmers & Curnow 2001). For example, the strategy used by Fuhrmann et al. (1999) is to create an incomplete duplex, with one complete strand joined by "bridging" oligonucleotides on the opposite strand, followed by ligation of the complete strand. PCR is then used to obtain the complete DNA molecule (Fuhrmann et al. 1999).

Whichever strategy one might use, the characteristics of oligonucleotides are vital to the successful construction. Oligonucleotides should be free of secondary structures, as these structures would prevent correct annealing to form the expected product. The length of oligonucleotides is also an important consideration. In early construction work, short oligonucleotides of about 15-40 mers were often used, whereas oligonucleotides longer than 80-mers were routinely used in later constructions. Long oligonucleotides allow more rapid annealing, forming more stable duplex that could sustain further processing without the need for ligation. Furthermore, the number of oligonucleotides for a specific construction is also reduced. However, long oligonucleotides have their own disadvantages. Due to coupling efficiency of the oligonucleotide synthesis reaction, the longer the oligonucleotide, the lower the synthesis yield, and more importantly, the greater the chance of accumulated sequence errors (Hecker & Rill 1998). Therefore, although synthesis of oligonucleotides in excess of 150 bases is not unusual, oligonucleotides in the range of 80-100 bases have been most routinely used for gene construction. Purification by PAGE to isolate only correct-sized product is often

recommended. 5'-phosphorylation of oligonucleotides (necessary for ligation by DNA polymerase I) can be done enzymatically with polynucleotide kinase, or chemically during oligonucleotide synthesis, although the latter is now more often used due to its convenience and higher efficiency.

2.1.2 Construction strategy for the *gusA*^{Ssp} gene

A gene construction strategy was developed to take advantage of various improvements mentioned above, especially those from Wosnick et al. (1987), which were shown to be suitable for rapid construction of large synthetic genes. The main features of the construction strategy are:

- Division of the gene into subfragments of about 500 bases each, for easy assembly, and sequence verification.
- Use of long oligonucleotides of roughly 80 bases, all without 5'-phosphorylation, because ligation of single strand nicks is not necessary.
- Oligonucleotides to form a complete duplex, so that enzymatic "fill-in" is not necessary.
- Direct cloning of duplex containing single strand breaks into *E. coli* for *in vivo* nick repair, so that 5'-phosphorylation of oligonucleotides, and subsequent *in vitro* enzymatic ligation step is not necessary.

2.2 Materials and Methods

2.2.1 Oligonucleotides

A total of 50 oligonucleotides used for gene construction were chemically synthesized by GenSet Corp. (CA, USA). The oligonucleotides range from 35-mer to 100-mer, with the majority of 80-mers. The scale of the synthesis was 0.2 μmol for each oligonucleotide. All oligonucleotides were desalted and resuspended in deionized water. The sequences of all these oligonucleotides are presented in appendix 1.

All other oligonucleotides, used as primers for PCR, mutagenesis and sequencing, were chemically synthesized by various commercial sources using standard methods. The sequences of all these oligonucleotides are listed in appendix 2.

2.2.2 Enzymes, chemicals, plasmids and bacterial strains

All enzymes, buffers and chemicals were from New England Biolabs and Sigma, unless otherwise stated. pLITMUS39 was purchased from New England Biolabs. pLADF48 was a derivative of pTTQ18 (Stark 1987). *E. coli* K-12 strains DH5 α and KW1, a GUS-operon-deleted strain (Wilson et al. 1995), were used in all cloning and expression procedures as appropriate.

2.2.3 General methods

Restriction digestion, cloning, PCR, sequencing, and other common DNA manipulation techniques were standard, following the supplier's recommendations.

2.2.4 Gene construction

2.2.4.1 Purification of oligonucleotides

A total of 50 oligonucleotides required for the gene construction (12, 15, 13, 10 oligonucleotides for fragment A, B, C, and D, respectively) were purified using denaturing polyacrylamide gel electrophoresis (8% polyacrylamide gel containing 7M

urea) on a Protean II xi Cell apparatus (Biorad Laboratories, CA, USA). The gel was stained with toluidine blue, and destained with water until DNA bands were visible. Oligonucleotides of expected size were isolated from gel slides by elution with TE at 37°C overnight, ethanol precipitated, and resuspended in a small amount of water.

2.2.4.2 Annealing of overlapping oligonucleotides and isolation of correctly annealed fragments

For each fragment, equimolar amounts of purified oligonucleotides were mixed, heated to 95°C for 5 minutes, and cooled down to room temperature for annealing. The mixture was run on agarose gel, and annealed products of expected sizes (457, 555, 489, and 374 bp respectively for fragment A, B, C, and D) were isolated and used directly for cloning.

2.2.4.3 Cloning of annealed fragment into pLITMUS 39 and joining of fragment to form the final gene

All annealed fragments have restriction site overhangs ready to clone into pLITMUS39. Fragment A was cloned to Sall/BsrGI, fragment B to BsrGI/BspEI, fragment C to BspEI/Mlul, and D to Mlul/ApaI restriction sites. Each fragment was sequenced to confirm its correctness prior to joining together. Correct clones of fragment B and D were found among sequenced clones. Shuffling among partially correct clones were necessary to recover correct fragment A (using BsmFI/BstBI sites) and C (using AatII/Asel sites – note: Asel site is in the pLITMUS39 backbone). The compatible restriction sites allow the fragments to be sequentially joined by subcloning to form the complete gene in pLITMUS39.

2.2.4.4 Enzymatic 5'-phosphorylation of oligonucleotides and ligation of annealed products

Purified oligonucleotides, except the two at the most 5' ends of each fragment, were phosphorylated individually with T4 polynucleotide kinase, then ethanol precipitated

and resuspended in a small amount of water. All oligonucleotides of each fragment were then mixed, heated and cooled down to allow annealing. The annealed products were ligated with either T4 DNA ligase (at 16⁰C) or Taq DNA ligase (10 cycles of 60⁰C/5min, 70⁰C/30sec).

2.2.4.5 PCR of ligated products for cloning

Each ligated fragment was PCR-amplified with primers that generate additional 5' EcoRI site and 3' BamHI site. This enables cloning of fragments to other plasmids, should the pLITMUS39 have a recombination problem with its two T7 promoters, leading to the deletion of the cloned fragment. The primers used were AT and AB for fragment A; BT and BB for fragment B; CT and CB for fragment C; DT and DB for fragment D (appendix 2). PCR products were digested with appropriate restriction enzymes before cloning to pLITMUS39.

2.2.4.6 Generation of different 5'-ends of the gene by PCR

Various 5' end variations of the synthetic gene were created by PCR using fragment A as template. The primers used were AI, AII, AIII, and AB (appendix 2). The PCR products were cloned into Sall/BsrGI sites of the originally constructed *gusA*^{Ssp} variant, replacing the original fragment A.

2.2.5 Site-directed mutagenesis

♣ Using uracil-containing, single stranded phagemid template method:

For reversing four non-silent changes in the coding sequence of the assembled *gusA*^{Ssp} gene, site-directed mutagenesis was carried out according to the protocols by Trower et al. (1996). Basically, the mutagenesis scheme involves the generation of a single stranded uracil-containing phagemid template by infecting a uracil N-glycosylase defective *ung*⁻ *E. coli* strain with M13. The single stranded template was subsequently used for fill-in and ligase reactions with mutagenesis primers, T4 DNA polymerase, and T4 DNA ligase. Recovery of mutagenized plasmids was done by transforming the DNA

mix to wild-type *E. coli* (*ung*⁺), in which the uracil-containing template cannot replicate. For detailed protocols, see Trower (1996). The primers used for these mutagenesis reactions were AV, DY, AT, and HL. Primers for PCR-screening of correctly mutated DNA were YDB and TAB (appendix 2).

♣ **Using “Quik-change” mutagenesis:**

Site-directed mutagenesis for randomizing the potential N-glycosylation site (N118) and the cysteine residue (C499) in GUS^{Ssp} was carried out with the “quik-change” mutagenesis kit from Stratagene following the manufacturer’s recommended protocols. The primers used for N118X mutagenesis were BGUS-Asn-T and BGUS-Asn-B. Another pair of primers for specific mutagenesis of N118Q was BGUS-N118Q-T and BGUS-N118Q-B. These primers also eliminated a BstBI restriction site to support quick screening of mutants. The primers used for C499X mutagenesis were BGUS-C499X-T and BGUS-C499X-B. These primers also eliminated an MluI restriction site to support quick screening of mutants (appendix 2). The template was pTANE95.1.

2.2.6 Quantitative GUS assay on permeabilized *E. coli* cells

Quantitative GUS assay was done on permeabilized *E. coli* cells following the protocols described in (Wilson et al. 1992).

2.2.7 Hexahistidine-tagged protein purification

See details in chapter III (section 3.2.4).

2.3 Results and Discussions

2.3.1 A codon-optimized version of the *gusA*^{Ssp} gene constructed

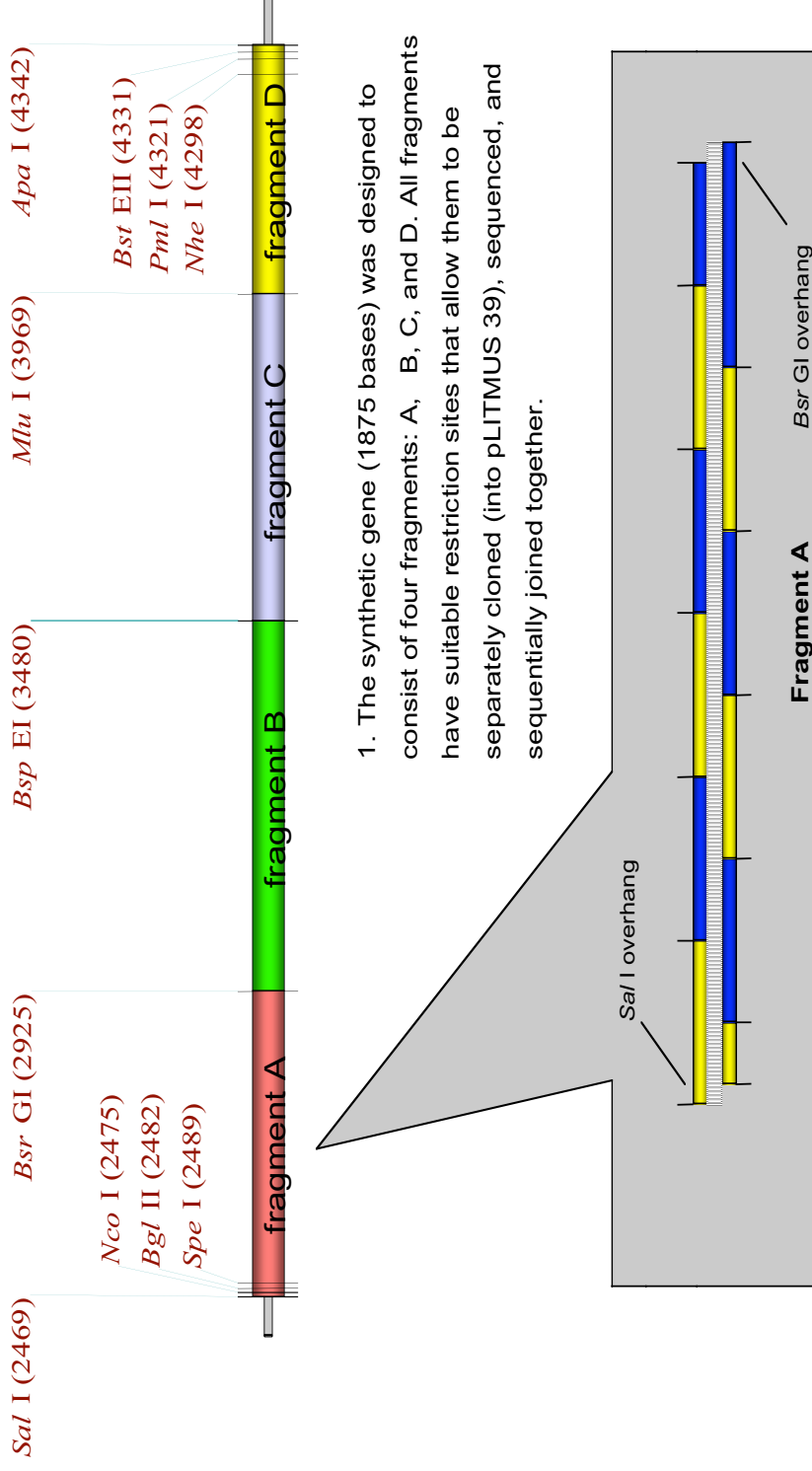
The gene was designed for optimized expression in *E. coli* and plants using the “ecohigh” codon usage from GCG Wisconsin Package (Genetics Computer Group, WI, USA). As a result, the A+T content was decreased from 65.8% to 44.9%. In addition, the polyadenylation signals (AATAAA), splice sequences (ATTTA AGGT), and restriction sites found in polylinkers, were eliminated. Other changes were also made to reduce potential secondary structures. Useful restriction sites were introduced to facilitate construction and downstream handling and manipulation of the gene. The codon usage of the native and synthetic *gusA*^{Ssp} gene is summarized in table 2.1.

The gene (1875 bp in length) was constructed from four fragments of about 500 bases each. Each fragment is formed by 10 to 15 oligonucleotides of roughly 80 bases. These oligonucleotides overlap (by about 40 bases) with their corresponding complementary oligonucleotides, and therefore can anneal to form complete double-stranded fragments. Each fragment was cloned into pLITMUS39, and sequenced separately to verify its correctness, then sequentially joined together to form a complete gene. A schematic description of the design and construction of the *gusA*^{Ssp} gene is presented in figure 2.1.

| Amino acid | Codon | No. in native gene | No. in synthetic gene | Amino acid | Codon | No. in native gene | No. in synthetic gene |
|------------|-------|--------------------|-----------------------|--------------|-------|--------------------|-----------------------|
| Ala | GCA | 14 | 7 | Leu | CUG | 6 | 19 |
| | GCC | 2 | 10 | | CUU | 3 | 1 |
| | GCG | 8 | 15 | | UUA | 16 | 0 |
| | GCU | 9 | 1 | | UUG | 6 | 2 |
| Arg | CGA | 4 | 0 | Lys | AAA | 28 | 13 |
| | CGC | 6 | 12 | | AAG | 6 | 21 |
| | CGG | 3 | 3 | Met | AUG | 11 | 11 |
| | CGU | 10 | 13 | | Phe | UUC | 8 |
| | AGA | 5 | 10 | UUU | | 28 | 10 |
| | AGG | 0 | 0 | Pro | CCA | 13 | 9 |
| Asn | AAC | 15 | 29 | | CCC | 1 | 0 |
| | AAU | 23 | 9 | | CCG | 3 | 16 |
| Asp | GAC | 5 | 23 | | CCU | 9 | 1 |
| | GAU | 30 | 12 | Ser | UCA | 6 | 0 |
| Cys | UGC | 0 | 1 | | UCC | 3 | 0 |
| | UGU | 1 | 0 | | UCG | 2 | 3 |
| Gln | CAA | 5 | 5 | | UCU | 5 | 4 |
| | CAG | 6 | 6 | | AGC | 2 | 11 |
| Glu | GAA | 41 | 22 | | AGU | 3 | 3 |
| | GAG | 14 | 33 | Thr | ACA | 14 | 0 |
| Gly | GGA | 10 | 9 | | ACC | 13 | 23 |
| | GGC | 11 | 29 | | ACG | 8 | 9 |
| | GGG | 8 | 3 | | ACU | 5 | 8 |
| | GGU | 17 | 5 | Trp | UGG | 11 | 11 |
| His | CAC | 5 | 11 | | Tyr | UAC | 7 |
| | CAU | 11 | 5 | UAU | | 21 | 10 |
| Ile | AUA | 7 | 0 | Val | GUA | 19 | 0 |
| | AUC | 8 | 21 | | GUC | 7 | 27 |
| | AUU | 15 | 9 | | GUG | 4 | 32 |
| Leu | CUA | 3 | 0 | | GUU | 33 | 4 |
| | CUC | 5 | 17 | <i>Total</i> | | 602 | 602 |

Table 2.1. Codon usage of the native and synthetic *gusA^{Ssp}* gene (*)

(*) All sequence errors (section 2.3.3) corrected.



2. Each fragment is formed by 10 - 15 oligonucleotides of roughly 80 bases each. These oligonucleotides overlap (by about 40 bases) with their corresponding complementary oligonucleotides, and therefore can anneal to form complete fragments.

Figure 2.1. Schematic representation of the synthetic *gusA^{SSP}* gene and its construction strategy.

Using this construction strategy, fragment A, C, and D were successfully assembled and cloned. Annealing of oligonucleotides of fragment B, however, failed to generate product of expected size (figure 2.2.B). It was suspected that the correct product was present in small quantity, therefore PCR was used in an attempt to recover it. To enable this, appropriate oligonucleotides were kinased before annealing. The annealed products, containing single-stranded nicks, were then ligated for use as PCR templates. PCR amplification, however, also failed to recover fragment B, although it successfully recovered all other 3 fragments (figure 2.2.C).

The results suggested that there was at least one problematic oligonucleotide that did not anneal properly, and therefore prevented the correct formation of fragment B. Because all fragment B's oligonucleotides appeared normal in size and concentration (figure 2.2.E), pair-wise annealing analysis was needed to identify those problematic oligonucleotides. It was clear that oligonucleotides B4 and B11 were unable to anneal with their corresponding oligonucleotides (figure 2.2.F). After the B4 and B11 oligonucleotides were resynthesized, fragment B was successfully assembled (figure 2.2.D) and cloned using the original strategy used for other three fragments. The conclusion drawn was that quality control in the initial synthesis by GenSet had been inadequate to ensure the requested design sequence was actually prepared.

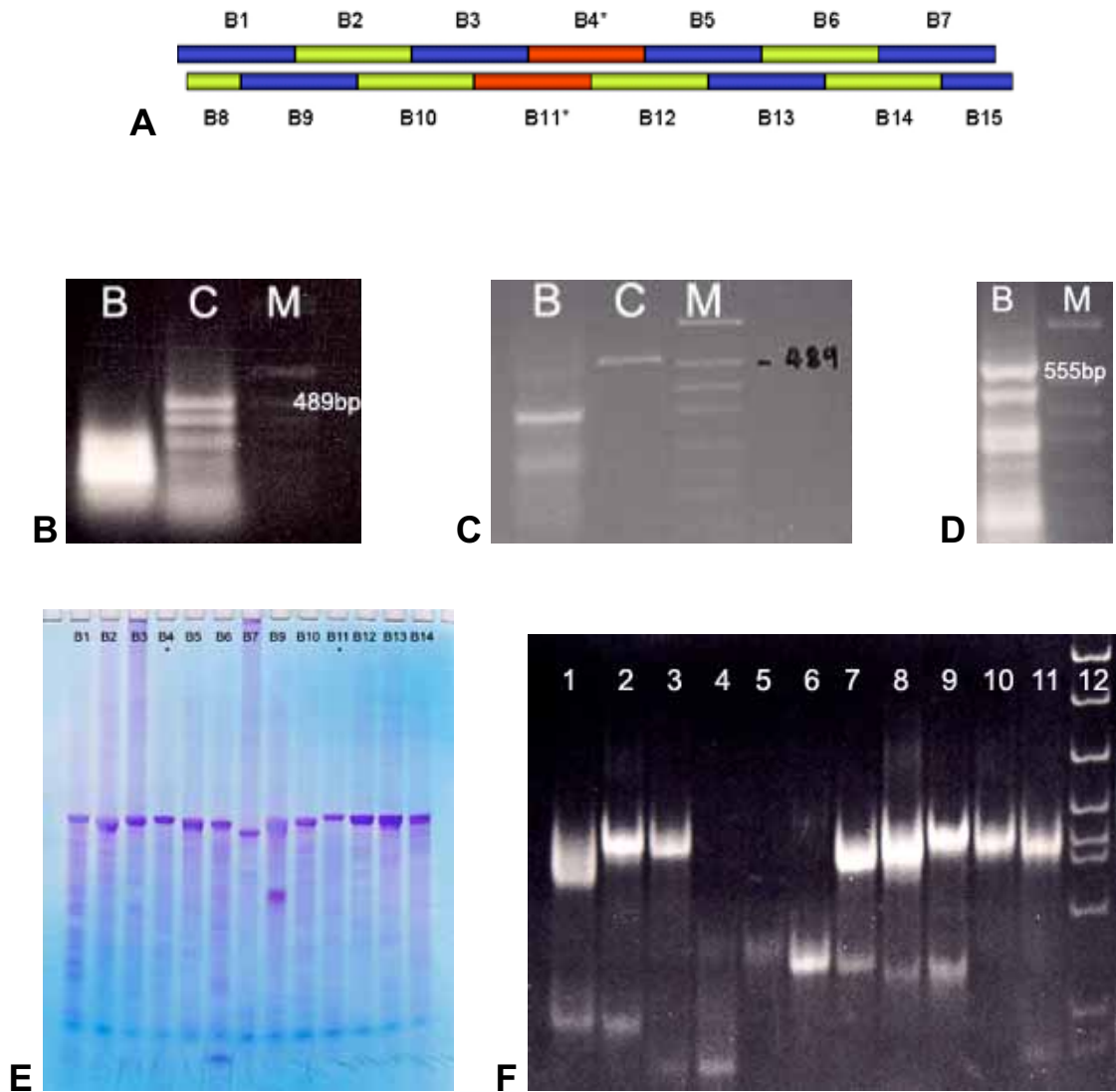


Figure 2.2. Identification of problematic oligonucleotides that prevented the correct formation of fragment B.

A. Schematic representation of fragment B with names and positions of all oligonucleotides.

B. Annealed product of fragment B and C separated on agarose gel. DNA band of expected size was visible for fragment C (489 bp), but not for fragment B (555 bp).

C. PCR amplification using ligated fragment B and C as templates. DNA band of expected size was visible for fragment C, but not for fragment B.

D. Annealed product of fragment B after problematic oligonucleotides B4 and B11 were resynthesized. DNA band of expected size (555 bp) was recovered.

E. Oligonucleotides of fragment B run on denaturing acrylamide gel and stained with toluidine blue. Although oligonucleotides B4 and B11 appeared normal in size and concentration as seen in this gel, they failed to anneal with other oligonucleotides, as judged by pair-wise annealing analysis.

F. Pair-wise annealing analysis of fragment B's oligonucleotides. Lane 1: B9/B2, lane 2: B2/B10, lane 3: B10/B3, lane 4: B3/B11, lane 5: B11/B4, lane 6: B4/B12, lane 7: B12/B5, lane 8: B5/B13, lane 9: B13/B6, lane 10: B6/B14, lane 11: B14/B7. It was clear that annealing was not possible in lane No. 4, 5, and 6. B4 and B11 were therefore identified as the cause for the unsuccessful annealing of fragment B.

Cloned fragments were sequenced to confirm their correctness before joining together to form the complete gene. Sequencing revealed that most cloned fragments contained randomly distributed mutations. In total, I have sequenced more than 100 clones to identify mutation-free clones. Correct clones for fragments B and D were found among those sequenced. However, for fragment A and C, shuffling among partially correct clones were necessary to recover fully correct clones.

A mutation analysis with 58 fully sequenced clones, representing a total of 27,547 base pairs of relevant sequence, is presented in table 2.2. Nearly 50 % of all mutations were single base deletions. Deletions of more than one base (consecutive) were much less common. The number of transition and transversion mutations was similar. The overall mutation rate was 10.2 mutations per 1000 base pairs.

The above results were rather expected, as it has been known that deletion mutations are the most common sequence errors in chemically synthesized oligonucleotides (Temsamani et al. 1995; Hecker & Rill 1998). The very high mutation rate of one error per 100 base pairs found in this study also agreed with the mutation rate reported by Hecker & Rill (1998). However, it should be mentioned that these authors only examined a total of 1000 bps. The data here was based on 27,547 bps, therefore is more statistically reliable.

All the correct fragments were then joined together by sub-cloning using compatible restriction sites. The complete sequence of synthetic gene is available in GenBank (acc. nr. AF354047).

2.3.2 Lessons learned from the construction

The construction strategy proved to be straightforward, with many advantages as outlined in section 2.1.2. However, actual construction work took longer than planned, due to the two technical difficulties. First, there were some problematic oligonucleotides that failed to anneal properly. Resynthesis of these oligonucleotides was necessary to

the successful annealing of a whole fragment to be used for cloning. Secondly, the very high error rate in cloned fragments made it difficult to identify flawless clones. Intensive sequencing and shuffling of correct parts from different clones was necessary to obtain correct fragments. Both problems were found to be associated with the quality of the oligonucleotides used for construction, and therefore, highlight the need to use highest quality oligonucleotides for similar work.

2.3.3 Non-silent changes in the coding sequence reversed using site-directed mutagenesis

The synthetic *gusA*^{Ssp} gene originally assembled did not produce GUS activity when introduced to KW1, a GUS negative *E. coli* strain. At that time, this prompted us with many possible reasons, two of which could be quickly verifiable: incorrect amino acid sequence based on which the gene was designed, and impaired expression/activity of GUS^{Ssp} due to the addition of amino acids at the N- and C-termini. The results of the former are presented below, whereas those of the latter are given in section 2.3.4.

I have completely re-sequenced the native *gusA*^{Ssp} gene, and found five incorrect nucleotides (table 2.3). These were errors of previous sequencing work (A. Kilian, personal communication). Four of them resulted in non-silent amino acid changes. The corresponding errors on the synthetic construct were corrected using site-directed mutagenesis, leading to the restoration of GUS activity. All these sequence errors, on the native and synthetic genes, have been corrected in the GenBank database (GenBank accession number AF354044 and AF354047, respectively).

| Fragment | No. of clones analysed | No. of mutations | | | | | | Total |
|--------------|------------------------|------------------|--------------|-------------|-------------|-------------|----------|------------|
| | | Substitution | | Deletion | | | | |
| | | transition | transversion | single base | double base | triple base | higher | |
| A | 24 | 23 | 21 | 70 | 8 | 6 | 2 | 154 |
| B | 8 | 0 | 0 | 16 | 0 | 0 | 2 | 26 |
| C | 21 | 6 | 4 | 45 | 9 | 2 | 0 | 87 |
| D | 5 | 2 | 3 | 2 | 0 | 1 | 1 | 14 |
| Total | 58 | 31 | 28 | 133 | 17 | 9 | 5 | 281 |

Table 2.2.A. Number of mutations found in fragment A, B, C and D clones. The majorities were single base deletions (nearly 50% of all mutations). The number of transition and transversion mutations was similar.

| Fragment | Total length analysed (bp) | Mutation rates | | | | | | Overall |
|----------------|----------------------------|----------------|--------------|-------------|-------------|-------------|-------------|--------------|
| | | Substitution | | Deletion | | | | |
| | | transition | transversion | single base | double base | triple base | higher | |
| A | 10,968 | 2.10 | 1.91 | 6.38 | 0.73 | 0.55 | 0.18 | 14.04 |
| B | 4,440 | 0.00 | 0.00 | 3.60 | 0.00 | 0.00 | 0.45 | 5.86 |
| C | 10,269 | 0.58 | 0.39 | 4.38 | 0.88 | 0.19 | 0.00 | 8.47 |
| D | 1,870 | 1.07 | 1.60 | 1.07 | 0.00 | 0.53 | 0.53 | 7.49 |
| Overall | 27,547 | 1.13 | 1.02 | 4.83 | 0.62 | 0.33 | 0.18 | 10.20 |

Table 2.2.B. Mutation rates per 1000 base pairs. The average mutation rate, calculated over 27,547 bps of sequence, was 10.2 mutations per 1000 bps. Because of such high mutation rate, finding mutation-free clones was difficult, and in some cases involved shuffling among partially correct clones (see text). The high mutation rate agreed well with a mutation analysis of chemically synthesized oligonucleotides by Hecker & Rill, 1998.

| Native sequence | | Synthetic sequence | | Amino acid change |
|-----------------|-----------|--------------------|-----------|-------------------|
| Error | Corrected | Error | Corrected | |
| GCA | GTA | GCC | GTC | A128V |
| CAC | CTC | CAT | CTA | H141L |
| GAT | TAT | GAT | TAT | D204Y |
| CCT | CCC | unchanged (CCA) | | unchanged (P405) |
| GCA | ACA | GCC | ACC | A560T |

Table 2.3. Non-silent changes in the coding sequence of the synthetic *gusA*^{Ssp} gene corrected using site-directed mutagenesis.

2.3.4 Several 5' end variations of the synthetic gene created

The synthetic gene was originally constructed as a “modular” cassette. This cassette allows direct cloning of the gene into many existing yeast and plant expression vectors, and convenient incorporation of other elements, such as introns and signal peptides, into the gene later on. As a result, the protein has additional amino acids at both N- and C-termini (figure 2.3).

The “modular” cassette was designed under the assumption that such addition of amino acids would not severely compromise GUS^{Ssp} activity, as has been proven for GUS^{Eco}. However, since no GUS activity was detected when the synthetic *gusA*^{Ssp} gene was first introduced to *E. coli*, such assumption needed immediate verification. Several 5' end variations of the synthetic gene were constructed for testing. The original construct (designated as A0) has all the additional N-terminal amino acids of the “modular” cassette. In contrast, A1 construct has a native N-terminus without any amino acid addition/change. It also has the native Shine-Dalgarno sequence of the native *gusA*^{Ssp} gene, to ensure proper transcription. All and AIII are based on A1 with minor changes as the results of the restriction sites included (figure 2.3).

5'-ends of various synthetic *gusA^{Ssp}* gene variants:

```

                NcoI      BglIII      SpeI
                ~~~~~~ ~~~~~~ ~~~~~~
A0:  AGTGGGTCGACCC ATG GTA GAT CTG ACT AGT CTG...
                Met Val Asp Leu Thr Ser Leu

```

```

AI:  AGGAGTGCTATC ATG CTG...
                Met Leu

```

```

                NcoI
                ~~~~~~
AII: AGGAGTGCTACC ATG GTG...
                Met Val

```

```

                NcoI      BglIII
                ~~~~~~ ~~~~~~
AIII: AGGAGTGCTACC ATG GTA GAT CTG...
                Met Val Asp Leu

```

3'-end of all synthetic *gusA^{Ssp}* gene variants:

```

                NheI                                PmlI      BstEII
                ~~~~~~ ~~~~~~ ~~~~~~
...AAC GCT AGC CAT CAC CAT CAC CAT CAC GTG TGA ATTGGTGACC
    Asn Ala Ser His His His His His His Val ***

```

Figure 2.3. 5' and 3' ends of synthetic *gusA^{Ssp}* gene variants. A0 is the originally assembled gene in full “modular” cassette to support further engineering. For example, a signal peptide can be inserted between NcoI and BglIII site. An intron can be inserted between BglIII and SpeI. As a result, five additional amino acids are added between the first and second amino acids of the native protein. AI has the completely native N-terminus. Upstream of its ATG is the Shine-Dalgarno sequence (ribosome binding site) native to the *Staphylococcus gusA^{Ssp}* gene, to ensure proper transcription. Obviously, this construct lacks restriction sites for cloning into other vectors. AII and AIII are slightly modified from AI to accommodate some restriction sites. AII has an NcoI site. As a result, it has the second amino acid changed from Leu to Val, and the second base upstream of the ATG is changed from C to T. AIII has NcoI and BglIII sites. As a result, it has modification upstream of the ATG like AII, and two amino acids added between the first and second amino acids of the native protein. All variations have an identical 3' end, which includes a hexa-histidine tag for affinity purification purpose.

Construction of these 5' end variants was done in parallel with resequencing of the native *gusA*^{Ssp} gene. The mutations found (table 2.3) were corrected and then incorporated into all 5' end variations by subcloning. Corrected variants were subsequently introduced to KW1 cells, and their specific activities in crude cell extracts determined (figure 2.4.A). Only the AI construct gave a high level of GUS activity. All other three constructs produced only 1-2% of total GUS activity obtained with AI construct.

The reduced levels of total GUS enzyme activity could be the result of reduced specific activities of these variants, or reduced steady-state quantities of them. The latter is the consequence of the whole cascade of transcription, translation efficiency, protein degradation etc.

To determine the amounts of GUS^{Ssp} variants being produced, equal amounts of cell extract expressing GUS^{Ssp} variants were loaded on a Ni-NTA column, and the hexahistidine-tagged proteins isolated and visualized with SDS-PAGE (figure 2.4.B). It was clear that except AI, all other variants were present in very small amounts, proportional to their reduced levels of GUS activities.

Therefore, it was concluded that the reduced levels of total enzyme activity were the consequence of reduced amounts of total GUS^{Ssp} variants. This finding, although preliminary, is important for the assessment of the N-terminal tolerance of GUS^{Ssp} fusions. The cause for the reduced amounts of GUS^{Ssp} variants remains to be identified. The change of the second amino acid from Leu (AI) to Val (A0, AII, AIII), should not target the protein for rapid degradation by N-end rule pathway, because Val is a stabilizing amino acid according to the N-end rule (Varshavsky 1992). Determination of the mRNA levels, and translation efficiency due to changes around the ATG start codon will be necessary for further understanding of the expression of these variants in *E. coli*.

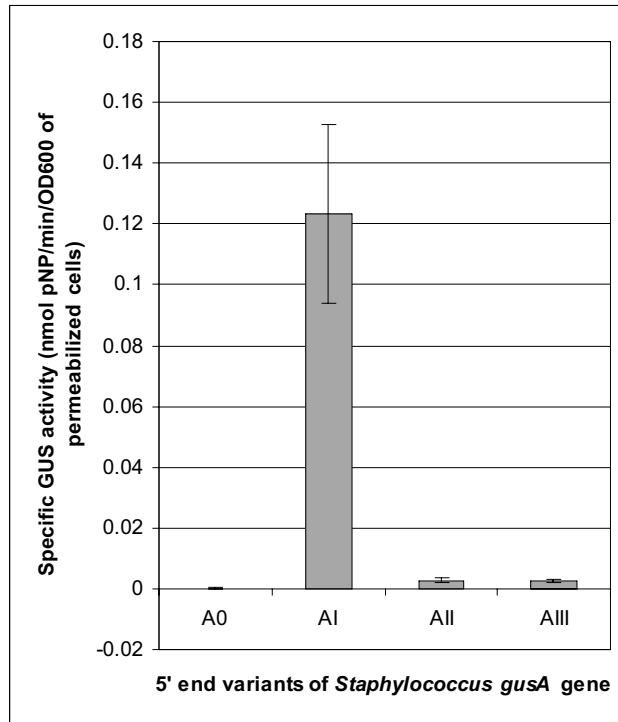


Figure 2.4 A. Specific GUS activities of different 5' end variations of the synthetic *gusA*^{Ssp} gene, measured in permeabilized KW1 cells. Only AI produced high specific activity, others had specific activities just about 1-2% compared to AI.

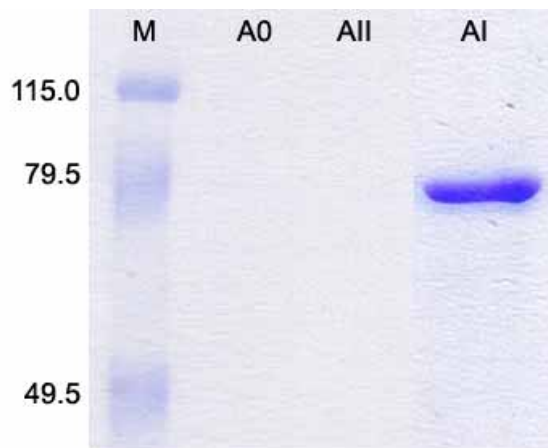


Figure 2.4 B. Hexa-histidine tagged GUS^{Ssp} variants isolated from equal amounts of extract of cells expressing these variants. Except AI, other variants were present in very small amounts, proportional to their reduced levels of GUS activities seen in figure 2.4 A.

Expression of GUS^{Ssp} in KW1 lead to colonies which were irregular in shape, size, and level of blue color on X-glcA containing plates (figure 2.5). Small colonies were often darker blue. Segregation within a colony was clear, resulting in dark and light blue patches. In liquid culture with KW1 host, maintaining a strong selection pressure was necessary to obtain consistent levels of GUS activity. It appeared that expression of the new enzyme was deleterious to the *E. coli* cells, leading to segregation.

Since expression of the *gusA*^{Ssp} gene in *E. coli* was not considered the principal topic of my research, many questions were left for future studies. For example, as discussed previously, we do not know the exact cause of the significantly reduced level of total GUS^{Ssp} seen in various 5' end variants. Also, possible effects of the C-terminal hexahistidine tag on activities of GUS^{Ssp} have not been assessed, although various results have indicated such effects would not be profound.

It should be recalled that all 5' end variants were not originally designed for a comprehensive study on GUS^{Ssp} expression in *E. coli*. Such studies will involve future design of proper expression cassettes in inducible systems. A *recA*⁻ version of KW1 is also needed to minimize recombination issues due to possible unfavorable conditions caused by the expression of GUS^{Ssp}. Such bacterial strain (JEMA99.9) is now available (J. Mayer, unpublished).

A1 and A0 variants were used in all subsequent work described in the thesis. The A1 construct was used for GUS^{Ssp} production in *E. coli* and subsequent purification for antibody production (chapter III) and biochemical characterizations (P. Wenzl and T. Nguyen, unpublished). The A0 construct was used for various clonings into yeast and plant vectors (chapter IV and V), as well as additional engineering like the addition of an intron into the gene (chapter VI).

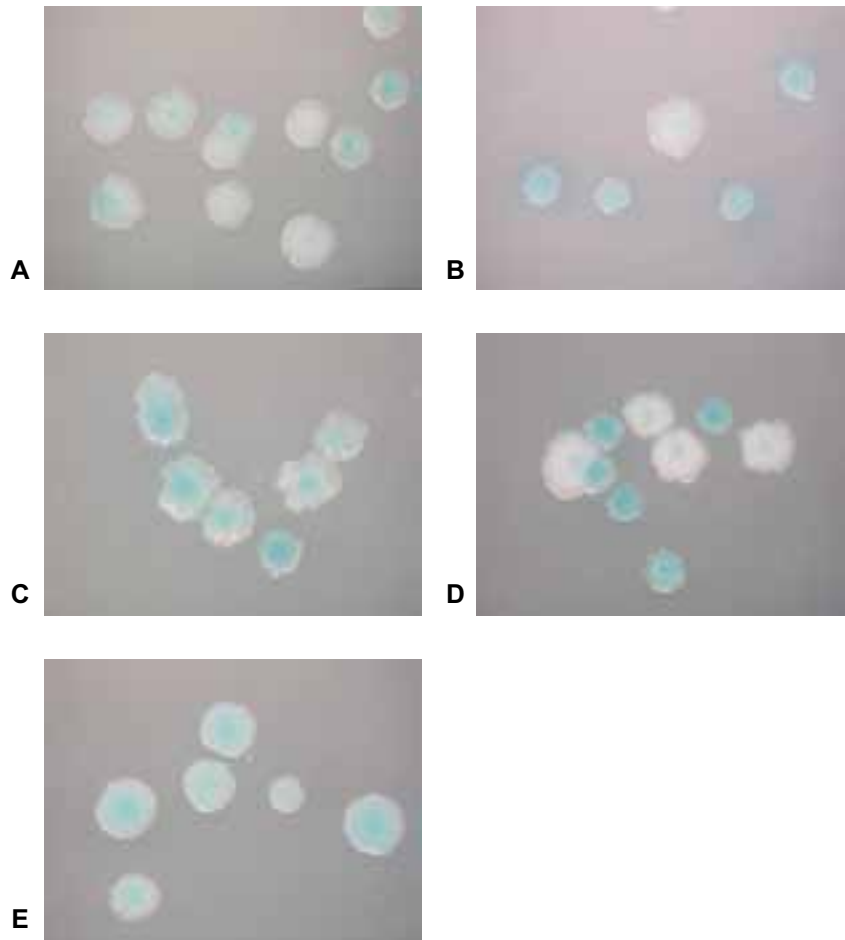


Figure 2.5. Colony phenotypes of KW1 cells expressing GUS^{Ssp} on X-glcA-containing media. A, B, C, D: A0, A1, AII, and AIII versions of GUS^{Ssp} in pLITMUS 39, respectively. E: GUS^{Eco} in pLADF48. Note that colonies expressing GUS^{Ssp} vary in size, shape and color intensity (A, B, C, and D). In contrast, colonies expressing GUS^{Eco} were much more uniform (E).

2.3.5 Cysteine-free, N-glycosylation-free, and cysteine-and-N-glycosylation-free GUS^{Ssp} variants created with reasonable GUS activity retained

GUS^{Ssp} contains a potential N-glycosylation site (N118) and a cysteine residue (C499). These two elements could interfere with GUS^{Ssp} activity and/or secretion when it is targeted into the ER (section 1.5). Site-directed mutagenesis was used to “randomize” the two positions (figure 2.6).

With the N118 codon, the first two bases were randomized, whereas the last base was changed to G, so that no non-silent N118 mutant should be introduced. Theoretically, this strategy could generate a total of 13 different amino acids at the N118 position. The mutagenesis primers were designed so that mutants would have the BstBI restriction site eliminated. A total of 60 mutagenized clones with various color intensities from blue to white on X-glcA plates were selected. Among them, 21 clones were found to be mutants after quick screening for the absence of BstBI restriction site. This translated into an overall efficiency of 35%. Sequencing of these 21 clones revealed that 8 different amino acids at the N118 position were obtained (out of the theoretical 13). However, some mutant clones (in fact, all white clones) also contained extra single base deletions leading to a frame-shift.

Five true N118 mutants were obtained: Ser, Arg, Leu, Pro, Met. In theory, Gln would be the best replacement for Asn. However, we could not obtain this mutation among the mutagenized clones, therefore, a direct mutagenesis scheme to change Asn to Gln was carried out, and the desired mutation obtained.

Remaining GUS activities (measured in crude KW1 cell extract) of these candidates were compared to that of the wild-type (figure 2.6). As expected, Gln mutant retained more than 60% of the original activity.

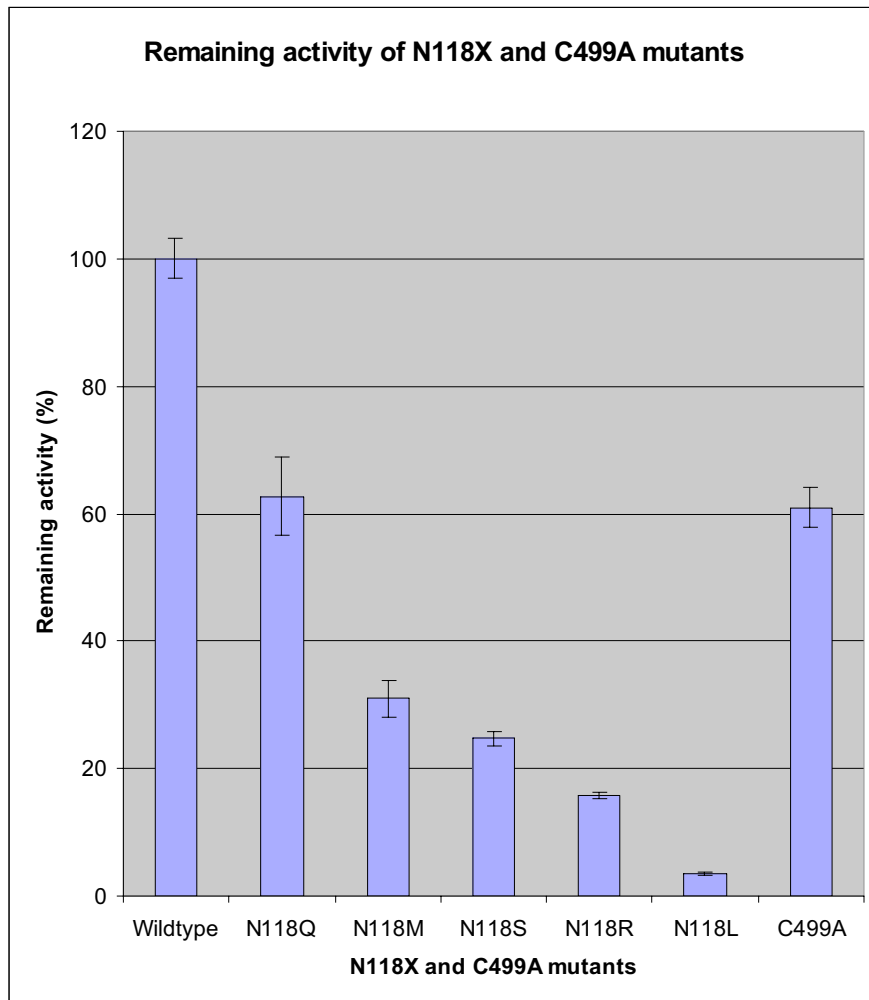


Figure 2.6. Remaining GUS activities of selected GUS^{Ssp} N118X and C499A mutants. The mutants N118Q and C499A were most promising, both retaining more than 60 % of the original activity.

Similar strategy was planned for the C499 codon: the first two bases were randomized, whereas the last base was changed to G, so that no non-silent C499 mutant should be introduced. Theoretically, this strategy could generate a total of 14 different amino acids at the C499 position. The mutagenesis primers were designed so that mutants would have the MluI restriction site eliminated. A total of 51 mutagenized clones with various color intensities from blue to white on X-glcA plates were selected. Among them, 28 clones were found to be mutants after quick screening for the absence of MluI restriction site. This translated into an overall efficiency of 55%. After sequencing confirmation, four true C499X mutants were obtained: Ala, Leu, Met, Gln. Among them, C499A was the most promising as it produced stronger blue colonies compared to others. It retained more than 60% of activity compared to the wild-type (figure 2.6).

The N118Q and C499A mutants were used in further studies in yeast and plants (chapter IV and V).