Effect of mRNA secondary structure in the regulation of gene expression: Unfolding of stable loop causes the expression of Taq polymerase in *E. coli*

Mandayam Alasingachar Mukund, Tanmay Bannerjee, Indira Ghosh and Santanu Datta*

Astra Research Centre India, 18th Cross, Malleswaram, Bangalore 560 003, India

The presence of secondary structure in the 5′ region of a mRNA modulates the translational efficiency in *E. coli*. A computer program 'ANARCI' which enables us to study the propensity of any single stranded nucleotide sequence to form stable secondary structures has been developed. A unique feature of this program is a mutation module which allows simulation of a gradual unfolding of a secondary structure by changing the nucleotide sequence in a manner such that the corresponding free energy is reduced while the coding amino acid sequence remains unchanged. We have analysed the 5′ region of a number of *E. coli* mRNAs. The results indicated that highly expressed genes had unfolded 5′ region in their mRNA whereas some of the poorly expressed genes had stable fold with $\Delta G < -5.5$ kcal/mole. In order to determine the effect of stable fold on the heterologous expression of genes in *E. coli* we took the test case of Taq polymerase. The wild-type gene of Taq polymerase was poorly expressed in *E. coli* and had a stable fold in 5′ coding region including the initiation codon AUG with a $\Delta G = -9.8$ kcal/mole. This stable fold presumably interferes with the interaction of the initiation AUG codon, with the f-Met-tRNA, thus arresting translation. Using ANARCI we could predict point mutations in the fold region (without changing the amino acid sequence) which would break the predicted folded structure. The gene carrying the designed point mutations was constructed and tested for expression of Taq polymerase. A 24-fold increase in the level of expression of the enzyme was obtained.

TRANSLATIONAL control in *E. coli* is complex and multidimensional. Initiation of translation is usually the rate limiting step. The current paradigm[1] of translation initiation is the binding of 30S subunit of the ribosome (to which the three initiation factors IF1,IF2,IF3 are bound), to the mRNA and t-RNA$_f^{Met}$ in a random order to form a ternary complex. Following a rate limiting conformational change, the 'preinitiaion' complex binds to 50S subunit with concomitant expulsion of IF1 and IF3 and the formation of 70S initiation complex. The last step is the hydrolysis of GTP to GDP and the release of IF2. Peptide elongation starts after the codon anti-codon interaction at the second codon. Since the specificity of the latter interactions is brought about by hydrogen bonding, it is imperative that the appropriate regions of the mRNA is single stranded and complementary. That a non-optimal S–D sequence or strong secondary structure in the 5′ region of the transcript down-regulates the translational efficiency, is well documented[2,3]. Studies have shown that when the putative secondary structure at the translation–initiation site of a gene is broken by point mutations there is a concomitant increase in the level of expression[4,5]. However, some ambiguity still remained. Studies described in the literature indicate that mutations which presumably break the secondary structure, also change the nucleotide sequence in the 5′ region and/or the amino acid sequence in the coding region. It is known that the N-terminal sequence of a protein has a significant role to play in protein turnover[6,7]. Also, some genes like *CspA* show elevated level of expression at low temperature due to increase in chemical half-life[8]. Thus it remains unclear whether the elevated level of expression is solely due to the breakage of the secondary structure or due to a positive input from an altered nucleotide sequence in the 5′ untranslated region and/or an altered amino acid sequence of the protein. In order to understand the basis for the difference in the level of translation we reasoned along the following lines.

If the secondary structure in the 5′ end of a transcript is a major factor in lowering the expression of a protein, then *E. coli* genes with high levels of expression should not have structured 5′ end and vice versa. We analysed the 5′ region of several *E. coli* genes (with documented levels of expression) with the computer program 'ANARCI'. If the secondary structure in the 5′ end of the mRNA is a major cause of translational blockage, then the only unambiguous way to break the secondary structure would be by altering the 5′ coding sequence using synonymous codons. This would keep the 5′ non-coding sequence and the amino acid sequence coded by the mRNA unaltered. We have tested this hypothesis and describe in this paper the use of this principle in the expression of Taq polymerase in *E. coli*.

To study the effect of secondary structures of mRNA on expression, we chose only those *E. coli* mRNAs whose expression levels are well-documented. Two independent data sets are described in the literature[9,10]. Thanraj and Pandit[10] have catalogued genes according to experimentally measured expression levels. They have classified genes in three distinct categories: high level expression (> 9000 molecules/cell), moderate level expression (between 1400 and 3000 molecules/cell) and low level expression (< 100 molecules/cell). In another data set, Medigue *et al.*[9] have catalogued genes based on their codon usage. It was seen that highly and lowly expressed genes clustered separately in a 61-dimension codon space. A number of genes were chosen from these clusters. Gene sequences were mainly retrieved from two databases: (1) GenBank

*For correspondence. (e-mail: santanu.datta@astra.in.astra.com)

(rel. 89) and (2) EMBL (rel. 32). While selecting the genes for this study we have used the following criteria: (i) Only those genes were taken whose coding DNA sequences were completely known; and (ii) First coding regions were taken in the case of multiple gene operons.

While selecting the required length of window for studying the folding of mRNA relating to the regulation of expression, our analysis was focussed around the translational initiation site. In the present study, we chose a window length of 75 nucleotides. The boundaries of the window in context of a whole gene was fixed by keeping the first A of the conserved S–D sequence AGGA, at the 51st position of the window. In the case of genes lacking a consensus AGGA sequence, the first base of the initiation codon, was placed at the 51st position of the window. The primary reason for selecting this type of window, was to analyse the secondary structure of mRNA in the upstream region rather than in the coding region.

ANARCI, written in Turbo C, is compatible with EGA or VGA monitor on DOS-based PC. The basic algorithm to arrive at the secondary structure was developed earlier by Jacobson *et al.*[11]. It accepts the RNA or DNA sequences in GenBank format[12] file as input and converts the DNA into RNA automatically. The program then searches for secondary structures in the user specified region of interest. Once the secondary structures are located, it then calculates the stability of the secondary structures. Change in the Gibbs free energy due to formation of secondary structure is calculated by considering the contributions of the loop, H-bond and helix formation energy[13]. Subsequently, it sorts energies due to secondary structure formation in an ascending order. Next, it has an option to display the ten best stable structures present in the region of study. This is an additional feature not available in related programs.

It would be preferred if one could theoretically alter a nucleotide involved in a secondary structure and see the effect of the change in the calculated free energy. Keeping this in mind, we have included a novel mutation module which can change the sequence of the RNA and calculate its corresponding free energy. This module chooses the hydrogen-bonded stem region of the most stable structure and changes one nucleotide (point mutation) at the base of the stem, in such a way, that the stability of the secondary structure is reduced. The nucleotide which replaces the original one does not form a hydrogen bond, as in the native secondary structure. A built-in feature of this operation is that while analysing the coding region, the predicted mutation does not change the amino acid sequence, i.e. it uses the degenerate codons of the amino acids. If need be, a subset of the degenerate codons that do not contain the rare codons for a particular organism can be used. This process of the unfolding the secondary structure is reiterated until the molecule is totally unfolded.

We cloned the gene with the wild-type sequence into the expression vector pTrc 99C (ref. 14). This was done by amplifying the gene from the second codon of the wild-type sequence. The sequence of the forward primer F1 was 5′ AGG GGG ATG CTG CCC CTC TTT GA. The reverse primer R1 including the native stop codon was designed with a Sal I site to facilitate the cloning of the amplified product directionally into the vector. The initiating methonine codon of the gene came from the filled-in Nco I site of the vector. The PCR amplification was done using the commercial Taq polymerase from Cetus. The conditions of amplification were denaturation at 94°C for 30 s, annealing at 55°C for 30 s and extension at 72°C for 2.5 min. Taking care of the high GC content of the Taq polymerase gene, the PCR was done in 0.5X PCR buffer. This construct (clone I) was introduced into the DH5α strain of *E. coli* and tested for expression by induction with isopropyl-β-D-thiogalactopyranoside (IPTG). For the mutant Taq polymerase the sequence of the forward primer F2, as suggested by the mutation module of ANARCI was 5′ CCG TTA TCG CGA GGG ATG CTG CCG TTG TTT GAG CCC. The three altered codons are underlined. Using this primer and the reverse primer R1 we reamplified the 2.5 kb gene from the wild-type construct. Because of the design of the primer, the amplified product included a NruI site (TCG/ CGA) at its 5′ end to facilitate cloning. The cleavage of the amplified product with NruI resulted in a blunt 5′ end starting with the second codon CGA. This fragment was cloned into the vector pTrc 99C which was digested with NcoI and Klenow filled-in, thereby providing the initiator codon AUG as before. The resulting construct (clone II) retained in-frame the code for the native amino acid sequence of Taq polymerase but without the predicted secondary structure encompassing the initiating codon in the 5′ region of the transcript.

The Taq polymerase was amplified from *Thermus aquaticus* genomic DNA using the primer pairs F1–R1 and F2–R1 and cloned into the vector pTrc 99C. Colonies containing the recombinant plasmid, i. e. having either the wild-type DNA sequence (clone I) or with the insert having the altered synonymous DNA sequence (clone II) were grown at 37°C until the absorbance at 600 nm reached 0.5. These were then induced with IPTG for 4 h. Aliquots equivalent to 0.5 ml cells of about 1 O.D. were pelleted and then dispersed in 0.5 ml of 1.1X PCR buffer, 50 µl of 1% lysozyme solution was then added and the cells incubated at 37°C for 10 min. The bath temperature was then elevated to 85°C with a gradual increase of about 3°C per min. The lysed cells were then tranferred to an ice bath. Most of the *E. coli* proteins precipated under this condition. The soluble supernatant was recovered by spinning in a microfuge at 14,000 r.p.m. for 10 min. The supernatant was then concentrated 10-fold by ultrafiltration in a microcon 30 (Amicon) filter. The concentrated supernatant was then checked for expression by SDS-PAGE.

In Table 1, we have listed the total number of genes analysed for their expression levels and the Gibbs free energy change, (with a cut-off at – 6.0 kcal/mole) due to the formation of secondary structures of different genes. It was observed that genes expressed at high levels in *E. coli*, such as ribosomal protein coding genes, do not have stable secondary structures around the translational initiation region ($\Delta G > -6.0$ kcal/mole). On the other hand, genes which are moderately and poorly expressed in *E. coli*, like *trpE* and *trpR* do have stable secondary structures in some cases ($\Delta G < 6.0$ kcal/mole). This observation supports the notion that highly expressed genes prefer minimum negative control to regulate their expression, while poorly expressed genes may have more negative control at the transcriptional or translational level, or both. It is well known that the tryptophan operon is regulated at the translational level by secondary structure formation[15], as has been observed in the case of *trpE* and *trpR* genes.

Though our data set is limited (due to not having unequivocal quantitative experimental information related to expression level of genes), the information that energy higher than – 6.0 kcal/mole would not hinder expression, is valuable. The stabilization of secondary structures occurs in the presence of water and the same order of energy difference is required to stabilize the inter-water hydrogen bonds. Hence, it has been noted that – 6.0 kcal/mole can be considered as the cut-off for stable secondary structures.

Till date there are no published reports of hyper-expression of the wild-type Taq polymerase in *E. coli* although successful hyper-expression has been reported with an altered N-terminal sequence wherein the wild-type sequence Met.Arg ... is changed to Met.Asn.Ser ... of the protein[16–17]. When the wild-type construct (clone I) was introduced into the *E. coli* strain DH5α and tested for expression by induction with IPTG no discernable induced protein band of 94 kDa was visible on Commasie blue stained SDS-PAGE (Figure 1). The secondary structure in the 5′ region of the transcript was analysed by ANARCI taking into consideration that the longest strech that this program can handle is about 75 nucleotides. The region chosen encompassed the entire length of the untranslated region along with about 35 nucleotides in the coding of Taq polymerase mRNA. In the predicted mRNA structure (Figure 2 *a*), the first and second amino acids are involved in a stable loop whose calculated free energy value is – 9.8 kcal/mole. It can be hypothesized that the formation of such a loop not only makes the initiation codon (AUG) inaccessible to the f-Met-tRNA but also blocks the codon anti-codon interaction at the second codon, thus hampering efficient initiation of translation and elongation. Based on an analysis of data in the literature, de Smit and van Duin[18] proposed that structures stronger than – 6 kcal/mole usually showed reduced translation efficiency. In order to test the hypothesis that disruption of secondary structure of mRNA in this region may enhance the translation efficiency, we introduced breaks in the stable secondary structure using the mutation module. The mutation module was used in such a way that only the degenerate codon could break the structure. This procedure is extremely important as the amino acid sequence remains unaltered as well as the sequence motif of the 5′ untranslated region remains unchanged. This exercise led to a change in three codons in the N terminal region of the gene. The second codon AGG (Arg) was changed to CGA, the sixth codon was changed from CCC (Pro) to CCG and the seventh codon CTC (Leu) was changed to TTG. While designing the primer F2 these imputs were taken into account. These three alterations which disrupted the stable loop kept the amino acid sequence unchanged. We also took into account the codon usage while designing this mutant. Since the frequency of usage of the second codon was implicated in the efficiency of elongation[19], and hence expression, we have taken this factor in the mutant design. The wild-type and the mutant codons at the second position are AGG and CGA respectively, both are among

Table 1. Identification of stem–loop structure in the 5′ end of mRNA in *E. coli*

Total no. of genes	Expression level		Free energy (kcal/mole)		Source
	High	Low	> – 6.0	< – 6.0	
52	30	22	50	2*	Medigue *et al.*[9]
17	10	7	15	2#	Thanraj and Pandit[10]

**PNP* and *deoD* genes
#trpE and *trpR* genes

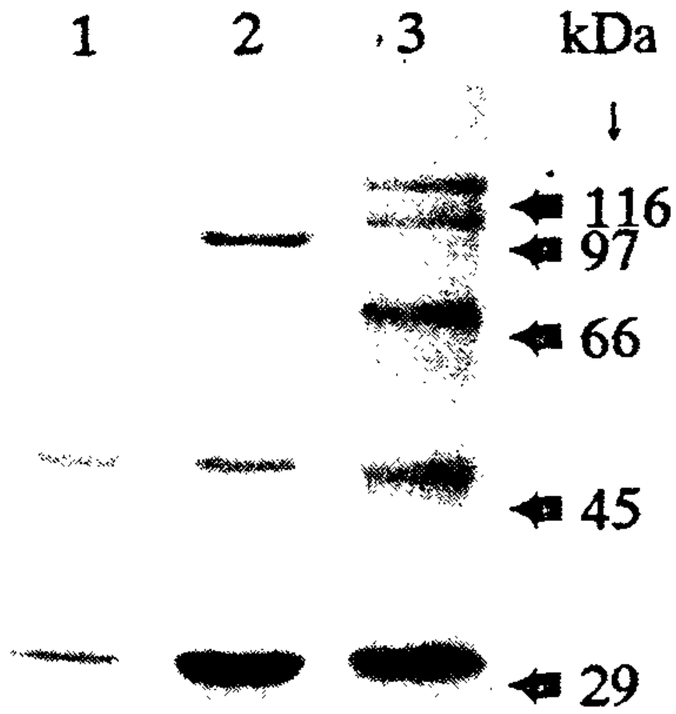

Figure 1. Normalized 100 µl lysates from clones I and II were loaded in lanes 1 and 2, respectively. The expressed Taq polymerase band clearly visible in lane 2 is around 94 kDa. Molecular markers are loaded in lane 3.

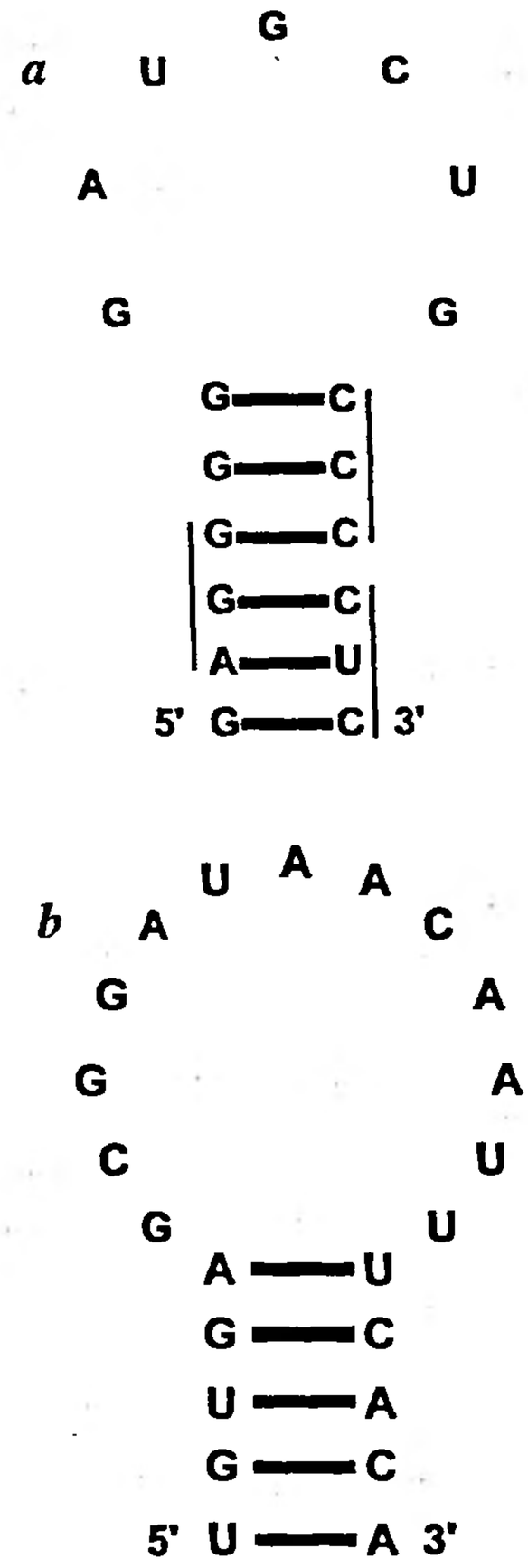

Figure 2. Most stable secondary structure in the 5′ region of the Taq polymerase transcript. *a* The stem–loop structure in the wild-type sequence (clone I). The 2nd, 6th and 7th codons that are involved in stabilizing the structure and which are changed in the clone II are overlined; *b* The stem–loop structure in the mutant allele (clone II). It only involves the 5′ non-coding region of the transcript. The sequence that is involved in stabilizing the structure comes from the vector.

the least preferred Arg codons in *E. coli*[20]. If the formation of the loop down-regulated the expression, opening of the loop using alternative synonymous codons should alleviate the down-regulation of this expression. The stem–loop structure shown in Figure 2 *b* formed at the 5′ region of the clone II transcript is an inherent feature of the vector nucleotide sequence which is unstable ($\Delta G < -6$ kcal/mole) and does not involve either the RBS or the initiating AUG codon.

Figure 1 shows a distinct band of 94 kDa protein in the lane containing the lysate of clone II, whereas no such band was visible in the clone I. However, when the gel was scanned by a laser densitometer the specific band of 94 kDa was detected in both the lanes and clone II had about 24-fold more Taq polymerase than clone I. The

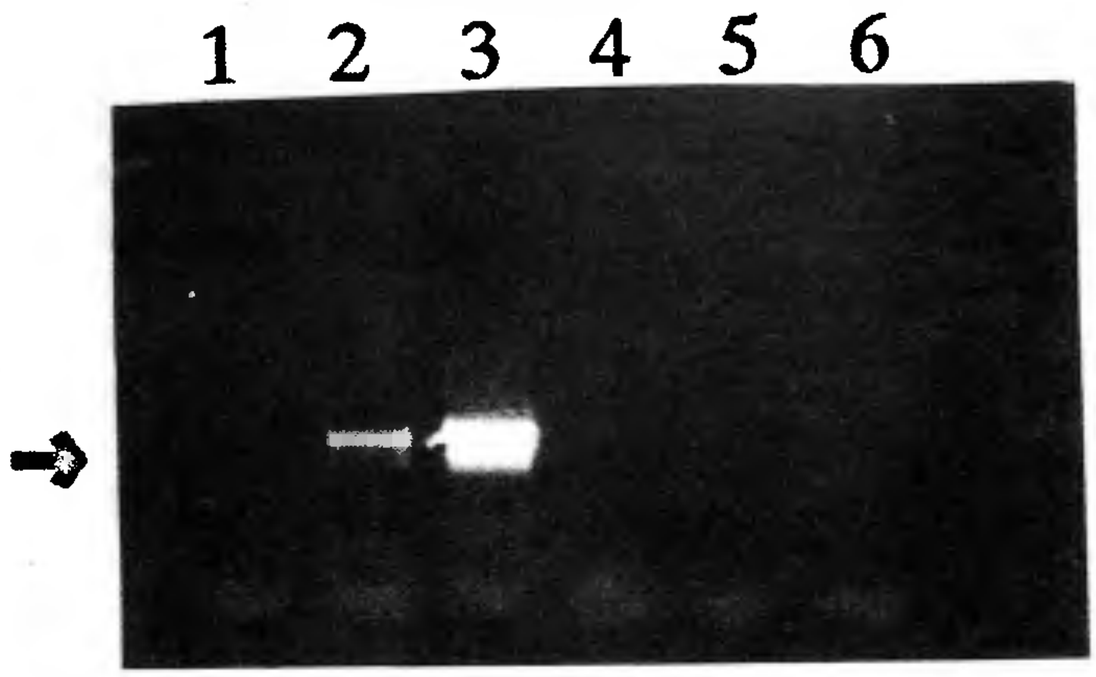

Figure 3. Activity of the expressed Taq polymerase: Normalised 3, 9 and 27 µl of lysates from clones I and II were used to PCR amplify a specific fragment of *E. coli* alkaline phosphatase gene. Lanes 1–3 indicate the amplification by clone II lysates and lanes 4–6 indicate the amplification by clone I lysate.

identitity of the 94 kDa band as Taq polymerase was established by demonstrating that only the soluble supernatant from clone II could replace the Taq polymerase enzyme in a PCR assay (Figure 3). The results indicated that the supernatant from clone II amplified a fragment of alkaline phosphatase gene from *E. coli* DNA whereas a similar extract from clone I was unable to do so. Since both the clones were identical except for the three synonymous codon substitutions, it may be concluded that the observed 24-fold increase in the expression of clone II was due to the proposed unfolding of the 5′ end of the mRNA sequence.

This study indicates that highly expressed *E. coli* genes do not have secondary structures in the 5′ end of mRNA and the cut-off free energy change that determines the stability of the hairpin loop is about – 6.0 kcal/mole at room temperature. Using ANARCI, point mutations that unfold the stem–loop can be designed in a way such that amino acid sequence of the expressed protein does not change. In earlier studies, the stable folds that negatively affected translation always included the S–D sequence[17]. This led to the conclusion that the down-regulation of the translation was due to the failure of hydrogen bonding between the S–D sequence and the 16S rRNA. In the present case of wild-type Taq polymerase gene cloned in *E. coli*, the S–D sequence comes from the vector and lacks any stable secondary structure, the fold starts in the coding sequence and only the third base of the initiation codon and the second codon are involved in the stable fold. This probably means that the binding of fMet–tRNA with the AUG codon and the codon anti-codon pairing of the second amino acid are impaired. This may also explain why it is not possible to express Taq polymerase with the wild-type sequence in any expression vector. Since the coding region (from the very first codon) is involved in the secondary structure, vector sequences

which bring in the 5′ untranslated region cannot possibly break the stable fold. On the other hand, when the N-terminal sequence is altered as in two cases of expression of Taq polymerase[14,15] we see expression in two different vectors. This may also indicate that the secondary structure in the mRNA after the third codon does not hinder translation initiation. Finally it may be reasonable to speculate that the wild-type Taq polymerase gene is expressed in the native strain *Thermus aquaticus* because of its thermophillic nature. At temperatures above 70°C, the secondary structure is unfolded and the translational block is released.

To get a copy of ANARCI write to Indira Ghosh at indira.ghosh@astra.in.astra.com.

1. McCarthy, J. E. G. and Gualerzi, C., *Trends Genet.*, 1990, **6**, 78–85.
2. de Smit, M. H. and van Duin, J., *J. Mol. Biol.*, 1994, **235**, 173–184.
3. Olsthoorn, R. C. L., Zoog, S. and van Duin, J., *Mol. Microbiol.*, 1995, **15**, 333–339.
4. Ramesh, V., De, A. and Nagaraja, V., *Protein Eng.*, 1994, **7**, 1053–1057.
5. de Smit, M. H. and van Duin, J., *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 7668–7672.
6. Tobias, J. W., Shrader, T. E., Rocap, G. and Varshavsky, A., *Science*, 1991, **254**, 1374–1377.
7. Varshavsky, A., *Cell*, 1992, **69**, 725–735.
8. Brandi, A., Pietroni, P., Gualerzi, C. O. and Pon, C. L., *Mol. Microbiol.*, 1996, **19**, 231–240.
9. Medigue, C., Viari, A., Henaut, A. and Danchin, A., *Microbiol. Rev.*, 1993, **57**, 623–654.
10. Thanraj, T. A. and Pandit, P. M. W., *Nucleic Acids Res.*, 1989, **17**, 2973–2985.
11. Jacobson, A. B., Good, L., Simonetti, J. and Zuker, M., *Nucleic Acids Res.*, 1984, **12**, 45–52.
12. Burks, C., Cinkosky, M. J., Fischer, W. M., Gilna, P., Hayden, J. E. D., Keen, G. M., Kelly, M., Kristofferson, D. and Lawrence, J., *Nucleic Acids Res.*, 1992, **20**, 2065–2069.
13. Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilsons, T. and Turner, D., *Proc. Natl. Acad. Sci. USA*, 1986, **83**, 9373–9377.
14. Amann, E., Ochs, B. and Abel, K.J., *Gene*, 1988, **69**, 301–315.
15. Lewin, B., *Genes 3/e*, John Wiley & Sons, 1987, p. 255.
16. Engelke, D. R., Krikos, A., Bruck, M. E. and Ginsberg, D., *Anal. Biochem.*, 1991, **191**, 396–400.
17. Desai, U. J. and Pfaffle, P. K., *Biotechniques*, 1995, **19**, 780–784.
18. de Smit, M. H. and van Duin, J., *J. Mol. Biol.*, 1994, **244**, 144–150.
19. Looman, A. and Knippenberg, P., *FEBS Lett.*, 1987, **197**, 315–320.
20. Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T., *Nucleic Acids Res.*, 1992, **20**, 2111–2118.

Received 6 January 1999; revised accepted 8 March 1999.

Anther dimorphism, differential anther dehiscence, pollen viability and pollination success in *Caesalpinia pulcherrimma* L. (Fabaceae)

B. S. Sarala*, R. Lokesha*† and R. Vasudeva**

*Department of Genetics and Plant Breeding, College of Agriculture, P.B. No. 24, Raichur 584 101, India

**Department of Forest Biology, College of Forestry, Banavasi Road, Sirsi 581 401, India

Self-compatible flowers of *Caesalpinia pulcherrimma* L. (Fabaceae) exhibit dimorphism in respect of size of anthers, in their dehiscence time and filament movement. All the ten anthers borne in a flower do not anthesize simultaneously; seven small anthers dehisce at early hours of the day (< 8.30 h) and three large anthers at mid-day. The pollen grains of small anthers were viable during early hours of the day, lose viability across time on the day of anthesis and were not viable the next day. In contrast, pollen grains of large anthers were viable till early hours of the next day. Stigma is receptive for 24 to 25 h. However, the pollen capture by stigma begins only after mid-day when the style has deflected upwards and positioned close and parallel to nectar guide. Removal of large anthers alone reduced fruit set compared to removal of small anthers alone. The adaptive advantage of this mechanism has been studied from the point of pollination success and discussed under the light of psychophilic syndrome of *Caesalpinia pulcherrimma*.

TOTAL pollen production of a plant can be hierarchically divided and may be presented sequentially to reduce the risk of pollen removal by an individual pollinator during a single visit[1–3]. This tactic results in dispersal of pollen grains to more pollinators and subsequently to several different stigma. In addition, staggered presentation of pollen will prevent weather-related deterioration on pollen. Further, scheduling of pollen presentation has been shown to effect male reproductive success[4].

Harder and Thomson's model[3] predicts that the pollen presentation to be synchronous with low pollinator visitation rates and staggered with high pollinator visitation rates. However, in contrast to predictions, *Erythronium grandiflorum* (Liliaceae) was found to stagger pollen presentation despite its low visitation rates[5].

Staggered pollen presentation is achieved by either altering time of anthesis within an inflorescence or anther dehiscence within a flower coupled with gradual squeezing of pollen from an anther pore[3]. Percival[6,7] observed that in 52 of 81 Welsh species (44.2%) anthers did not dehisce simul-

†For correspondence