Molecular Microbiology (2000) **37**(1), 216–218

MicroCorrespondence

The DB case: pattern matching evidence is not significant

Sir,

The existence of a downstream box (DB) able to enhance translation efficiency in leaderless transcripts has been reported in several *Escherichia coli* genes. This putative element occurs downstream of the start codon, and it is complementary to the 16S rRNA sequence AGUACUUAGUGUUUC, with which it has been proposed to interact. According to the model of Sprengart and Porter (1997, *Mol Microbiol* **24**: 19–28), this should compensate for the lack of stabilizing interaction of the rRNA with a weak or absent Shine–Dalgarno sequence. In particular, some indication about the existence of such elements has been reported: notably, its requirement for the transcription of λ*cI* mRNA (Shean and Gottesman, 1992, *Cell* **70**: 513–522); its role in the expression of the lysyl t-RNA synthetase gene *lysU* (Ito *et al.*, 1993, *Proc Natl Acad Sci USA* **90**: 302–306); and its role in the expression of the cold shock protein CspA (Mitta *et al.*, 1997, *Mol Microbiol* **26**: 321–335).

Nevertheless, several other works have refuted the former analysis (Resch *et al.* ,1996, *EMBO J* **15**: 4740–4748; Tedin *et al.*, 1999, *Mol Microbiol* **31**: 67–77). Following the latter publication, two MicroCorrespondences were published regarding this subject. In one of these, Bläsi and colleagues described a large number of biochemical experiments and arguments against the DB hypothesis (see Bläsi *et al.*, 1999, *Mol Microbiol* **33**: 439–441 and references therein), namely: (i) it is difficult to reconcile how the anti-DB region can be brought into alignment with the mRNA track as shown in the model proposed by Sprengart and Porter (1997, *Mol Microbiol* **24**: 19–28); (ii) the interaction DB–anti-DB has not been supported by chemical footprinting or cross-linking studies; (iii) chemical protection studies have failed to reveal protection of the putative DB; (iv) the DB does not appear to increase the affinity of mRNA for 30S subunits; (v) the anti-DB stretch of rRNA is very well conserved among different species because of the stability of the helix in which it participates, but not because of its primary structure; (vi) finally, although the putative DB of Gram-positive bacteria is rather different, the leaderless λ*cI* mRNA was found to be expressed efficiently *in vivo* in *Bacillus subtilis*. Since the publication of these MicroCorrespondences, O'Connor *et al.* (*Proc Natl Acad Sci USA* **96**: 8973–8978) have presented important results concerning the effect of mutating the anti-DB at the *rrnB* ribosome of *E. coli*. This mutation, although disrupting the DB–anti-DB interaction, did not modify gene expression significantly. Therefore, this work reinforces the idea that DB is irrelevant for gene expression.

Resch *et al.* (1996, *EMBO J* **15**: 4740–4748) have shown that the deletion of the downstream box of the λ*cI* leaderless mRNA does not result in decreased expression levels of the transcript. To this, Etchegaray and Inouye (1999, *Mol Microbiol* **33**: 438–441) responded that the deletion mentioned above created a new DB. This should invalidate the conclusions of Resch *et al.* (1996, *EMBO J* **15**: 4740–4748). Although Etchegaray and Inouye (1999, *Mol Microbiol* **33**: 438–441) admit that cross-linking studies failed to reveal DB–antiDB interactions, they argue that, as the signal exists, an interaction is also likely to exist. In fact, they maintain that their work reveals a correlation between the existence of the signal and gene expression.

In this letter, we contend that such a correlation is not significant, as the DB signal itself is not statistically significant. There is no evidence for the evolutionary selection of the DB signal. In fact, we shall show that the constraints given to define a DB sequence are so poor that any deletion would most likely result in the production of another putative DB site.

We have recently analysed factors affecting translation in *B. subtilis*, using its complete genome, and failed to identify a consensus sequence or a weight matrix for this signal (Rocha *et al.*, 1999, *Nucleic Acids Res* **27**: 3567–3576). As the absence of the ribosomal protein S1 in this organism significantly increases the requirements for a good ribosome binding site (RBS) (Vellanoweth and Rabinowitz, 1992, *Mol Microbiol* **6**: 1105–1114), we expected that this would also be the case for the DB.

In order to check for the statistical significance of the DB pattern found in *lysU*, *lysS* (Ito *et al.*, 1993, *Proc Natl Acad Sci USA* **90**: 302–306) and *tetR*, λ*cI* and P2V (Etchegaray and Inouye, 1999, *Mol Microbiol* **33**: 438–441), we devised the following simple *in silico* experiment. Let us consider one sequence fragment (e.g. the first 30 bases of *lysU*) in which the DB pattern is observed by the authors (Ito *et al.*, 1993, *Proc Natl Acad Sci USA* **90**: 302–306) with *k* exact matches (e.g. *k* is 8 for *lysU*). Let us then generate a large number (5000) of random sequences with the same length and nucleotide composition, and let us count how many of them have a 'best' hit of the DB pattern scoring at least *k* matches. The fraction of this

Table 1. Number of observed matches and respective *P*-values for the five genes mentioned by Ito *et al.* (1993, *Proc Natl Acad Sci USA* **90**: 302–306) and Etchegaray and Inouye (1999, *Mol Microbiol* **33**: 438–441).

		No gaps		With gaps	
Gene	Length	No. of matches	*P*-value	No. of matches	*P*-value
lysU	30	8	0.263	10	0.343
λ*cI*	50	8	0.794	10	0.631
lysS	30	7	0.569	10	0.307
P2V	50	7	0.856	8	0.997
tetR	50	6	0.988	9	0.824

The *P*-values correspond to the probability that the best hit found in a random sequence of the same length and nucleotide composition has at least as many matches as the observed putative DB (see text). The sequences analysed were taken from the beginning of the gene (starting at the start codon) up to the position given in the 'length' column, which corresponds to the sequences analysed by the authors. The analysis accepting insertions and deletions ('with gaps' columns) was performed using a 'pattern fit' variant of the Needleman–Wunsch dynamic programming algorithm (Erickson and Sellers, 1983, *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley, pp 55–91). In order to avoid large gaps, we used a typical gap creation penalty (1) and a very large gap extension penalty (10). All *P*-values are very large, therefore rejecting the significance of the observed hits.

number to the total number of generated sequences yields an estimation of the probability [*P*(best ⩾ 8) for *lysU*] of observing the DB with at least *k* matches on random sequences of the same length and nucleotide composition. Hence, this method provides the *P*-value associated with the occurrence of a DB.

As the DB site is supposed to result from a RNA–RNA interaction, we first performed this test using a simple pattern-searching algorithm, only allowing for matches and mismatches (i.e. ignoring insertions and deletions). One should note that, in this particular case, the probability distribution can also be computed theoretically, and the *P*-value can be evaluated under certain hypotheses (Tatusov *et al.*, 1994, *Proc Natl Acad Sci USA* **91**: 12091–12095).

As Etchegaray and Inouye (1999, *Mol Microbiol* **33**: 438–441) introduced gaps in their analyses, we refined our study further by introducing gaps in the pattern-searching procedure. This is simply achieved by using the 'pattern fit' variant of the Needleman–Wunsch dynamic programming algorithm (Erickson and Sellers, 1983, *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley, pp 55–91).

The results of both analyses are summarized in Table 1. The *P*-values are always very high and, therefore, the occurrence of the DB pattern on these particular sequences cannot be considered as significant. In other terms and to comment Etchegaray's reply to Resch's experiment, whatever the deletion performed, it would have a great chance of producing another putative DB, just because, with such a small number of matches, the DB pattern occurs by chance almost anywhere.

As a last argument, we have counted the number of hits of the DB pattern in the complete genome of *E. coli* allowing up to eight mismatches (seven matches). We found an astonishing number of 543 533 putative DB in *E. coli*, i.e. more than 100 DB per gene! and this number rises to more than one million if one accepts nine mismatches, as is the case for *tetR*. One can hardly imagine how such a ubiquitous signal might regulate ribosome–mRNA interactions and, especially, discriminate the translation start.

More recently, Etchegaray and Inouye (1999, *J Bacteriol* **181**: 5852–5854) have analysed the putative DB of the *E. coli cspB* gene, using a fusion with *lacZ*. This work considers three constructions, one including the DB (pb13), another deleting the DB (pb3) and a third one with an extended DB (pb17)). The fusion that eliminates the DB exhibits a strongly reduced gene expression when compared with the others. Naturally, the authors invoke a DB–anti DB interaction for this observation. However, the definition of DB used in this work is different from that considered by the previous works, exhibiting a shift of 2 bp (i.e. it is UACUUAGUGUUUCAC instead of AGUACUUAGUGUUUC). The best hit of this new definition of DB ignoring U-G pairs (as previously), is eight matches for pb13 and pb17 and seven matches for pb3. According to our previous calculations, none of these signals is statistically significant. Moreover, if one considers the DB signal defined in all the above studies, then the best hit for pb3 is also eight matches. As all the constructs have a similar number of matches to the standard DB, the absence of the signal should not be invoked to explain differences in expressiveness.

Our conclusions strongly suggest that what might be perceived as a complementary signal to the 16S rRNA is the result of nothing but the expected usage of words in the genetic texts of the beginning of genes. We have shown previously that these regions are A-rich (Rocha *et al.*, 1999, *Nucleic Acids Res* **27**: 3567–3576). This probably represents a strategy to avoid stable mRNA structures that would compete with the ribosome and therefore prevent translation initiation (de Smit and Duin, 1994, *J Mol Biol* **235**: 173–184). The proposed anti-DB possesses 7 U out of 15 nucleotides and, therefore, a reasonable complement will most certainly be present at the beginning of most genes.

We tested this hypothesis on the three constructs of Etchegaray and Inouye (1999, *J Bacteriol* **181**: 5852–5854). For this, we considered the 30 bp before and after the start codon for each of the three constructs and folded each of the sequences using the Zuker algorithm with standard parameters (Zuker and Stiegler, *Nucleic Acids Res* **9**: 133–148). The energy associated with the secondary structures is $-9.1\ \mathrm{kcal\ mol^{-1}}$ for pb13 and pb17 and $-11.5\ \mathrm{kcal\ mol^{-1}}$ for pb3. Therefore, the difference in secondary structure is likely to be the cause of the observed effects.

Bläsi *et al*. (1999, *Mol Microbiol* **33**: 439–441) made a strong case against the existence of DB by stressing the biochemical evidence against their existence. We have demonstrated that these elements are not selected for. Hence, we believe that, unless very strong new evidence comes to light, one may conclude that positive correlation between patterns at the beginning of genes and putative DB should be interpreted using alternative explanations.

Eduardo P.C. Rocha,[1,2]* Antoine Danchin[2,3] and Alain Viari[1]

[1]Atelier de BioInformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France. [2]Unité de Regulation de l'Expression Génétique, Institut Pasteur, 28 Rue Dr Roux, 75724 Paris, France. [3]HKU Pasteur Research Centre, Dexter Man Bldg, 8 Sassoon Road, Pokfulam, Hong Kong.

**For correspondence. E-mail erocha@abi.snv.jussieu.fr; Tel. (+33) 1 44 27 65 36; Fax (+33) 1 44 27 63 12.*

Accepted 2 May, 2000.