A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes

Gary J.Sharples and Robert G.Lloyd
Department of Genetics, University of Nottingham, Medical School, Queens Medical Centre, Nottingham, NG7 2UH, UK

Received September 10, 1990; Accepted October 16, 1990

ABSTRACT

We report the discovery of a novel group of highly conserved DNA sequences located within the intergenic regions of the chromosomes of *Escherichia coli, Salmonella typhimurium* and other bacteria. These intergenic repeat units (IRUs) are 124 – 127 nucleotides long and have the potential to form stable stem-loop structures. The location of these sequences within the intergenic regions is variable with respect to known or putative signals for transcription and translation of the flanking genes. Some of the IRU sequences are transcribed, others are probably not. The structure and possible functions of these sequences are discussed in relation to palindromic units and other repeated DNA sequences in bacteria.

INTRODUCTION

The intergenic regions of bacterial chromosomes, and indeed most organisms, contain specific sequences required for control of transcription and translation. These sequences include transcriptional promoters and terminators, translational start and stop signals and binding sites for regulatory proteins. In *Escherichia coli* and *Salmonella typhimurium* these intergenic regions also contain other conserved structures whose functions have not been fully characterised. One interesting example of this latter group is the *p*alindromic *u*nit (PU) or *r*epetitive *e*xtragenic *p*alindromic (REP) sequence (1, 2, 3). These PU sequences constitute a family of repetitive sequences of about 35 nucleotides that exhibit dyad symmetry (3, 4). Between 500 and 1000 copies of the PU sequence are thought to occur on the chromosomes of *E. coli* and *S. typhimurium*, occupying about 1% of the genome (5). PU sequences are always found in transcribed sequences, either located within an intergenic region of an operon or in the 3'-untranslated sequences of a transcription unit (2, 4).

Palindromic units have been postulated to play a role in transcriptional termination or as sites for processing by RNase III. These theories have been discounted, although in certain circumstances PU sequences may well function as terminators (5, 6). Recently it has been shown that the presence of PU sequences in mRNA constitutes a barrier to 3'−5' exonuclease digestion, thus stabilising the message and consequently increasing upstream gene expression (6, 7, 8, 9, 10). However, Gilson et al. (4) have pointed out that this function does not require the high degree of sequence homology observed for PU (of the 170 known *E. coli* PU sequences 80% of nucleotides match the consensus). Recent studies have suggested that this sequence homology is required for binding of specific proteins. DNA gyrase, DNA polymerase I and F8 (an unidentified complex of 5 proteins) all bind to PU *in vitro* (11, 12, 13). These studies led to the proposal that PUs are involved in folding of the bacterial nucleoid into independent supercoiled looped domains (4, 13).

In this paper we describe a novel, highly conserved DNA element located in the intergenic regions of the chromosomes of *E. coli, S. typhimurium* and other bacteria.

RESULTS

The intergenic repeat unit (IRU) we describe here was discovered during analysis of the DNA sequence downstream of the *tls* locus of *E. coli* (14). A computer search of the DNA sequence database (GeneBank release 63) identified seventeen bacterial sequences that showed a high degree of similarity to 3' end of the *tls* sequence. Eight of these were from *E. coli*, five from *S. typhimurium*, two from *Yersinia pseudotuberculosis*, and one each from *Vibrio cholerae* and *Klebsiella pneumoniae* (Table 1). The optimal alignment of all eighteen sequences is shown in Fig. 1. The alignment revealed a very high degree of sequence similarity extending over 127 nucleotides. Several of the sequences appear to have suffered extensive deletions while the initial sequence identified downstream of *tls* is incomplete.

Examination of the 124−127 bp elements shown in Figure 1 revealed in each case several inverted repeat sequences that could lead to the formation of a stable stem-loop with an axis of symmetry (boxed in Fig. 1) centred around an AT rich core. ΔG values for the putative stem-loop structures are shown in Table 1. The proposed secondary structure of the IRU from the *E. coli aceF-lpd* intergenic region is shown in Figure 2. In each case the inverted repeat is best conserved in the central 40 bp section of the sequence with both the paired and unpaired regions of the putative stem being preserved. In addition to the high sequence similarity and the potential to form similar secondary structures, all the sequences show a high percentage of purines in the single-stranded loop regions of the putative stem (Table 1).

* To whom correspondence should be addressed

Table 1. Location and properties of IRU sequences

Strain: locus (min.)	Accession number	base pairs	Percent identity	Free energy	% A+G in ssDNA	References
Eco: rpsB-tsf (4′)	J01684	1023–1149	87.4	−68.9	86.0	(23)
Eco: rplA-J (90′)	J01678	1374–1499	82.5	−35.4	75.9	(33)
Sty: glnA-L (87′)	J01803	77–202	79.4	−37.6	76.3	(18)
Eco: narK-G * (27′)	M17807	70–195	88.9	−46.4	69.8	(34)
Eco: aceF-lpd (2.6′)	V01498	5844–5970	86.6	−72.0	73.1	(35)
Sty: topA-cysB * (28′)	M15040	164–288	82.5	−18.8	69.3	(15)
Eco: hsdR-M * (99′)	X06545	3806–3932	94.5	−48.8	73.2	(36)
Eco: pgk-fda (62′)	X14436	4248–4374	95.2	−45.8	76.7	(37)
Eco: tls (41′)	X53984	692–788	90.7	−18.8	80.0	(14)
Sty: cysJIH * (60′)	M23007	238–357	84.9	−37.2	65.2	(16)
Sty: rpsU-dnaG	M14427	674–799	84.9	−43.4	63.6	(38)
Sty: metE-R (84′)	M17356	48–173	86.5	−51.8	79.2	(39)
Eco: metE-R * (86′)	J04155	147–273	78.6	−48.8	73.6	(40)
Vch: hlyA	Y00557	262–389	82.3	−31.8	84.3	(41)
Yps: pelY	M19399	134–256	71.5	−52.2	64.2	(42)
Eco: hisT-dedA (50′)	J02800	2247–2281	83.7	–	–	(43)
Yps: inv	M17448	418–455	95.2	–	–	(44)
Kpn: fim *	M20720	1379–1478	84.2	−47.4	67.9	(45)

The sequences identified are from *Escherichia coli* K-12 (*Eco*), *Salmonella typhimurium* (*Sty*), *Vibrio cholerae* (*Vch*), *Yersinia pseudotuberculosis* (*Yps*), or *Klebsiella pneumoniae* (*Kpn*). Where known, the genes flanking the IRU are shown along with the map location in minutes. Free energy (ΔG) values were calculated using the estimations of Tinico et al. (46). Revised rules for the calculation of stem-loop free energy gave very similar results (47, 48). The percentage identity to the consensus and the values for the percentage of purines in single-stranded regions of the putative stem were calculated from the sequences shown in Figure 1. Sequences marked with an asterisk are shown in reverse complement in Figure 1. The database searches were conducted using microcomputer software packages from DNASTAR Ltd.

The positions of the IRUs relative to the flanking genes and their transcriptional control signals, where these are known, are shown in Figure 3. Most are located between genes in operons, but their position relative to promoters and terminators is variable and it is not clear whether they are all transcribed. In the case of the *E. coli rpsB-tsf* and *narK-G* regions, the IRUs are clearly transcribed. However, the IRU in the *S. typhimurium topA-cysB* region is located downstream of what appears to be a strong terminator, while the IRUs found in the *metE-R* regions of both *E. coli* and *S. typhimurium* are located between the divergently transcribed *metE* and *metR* genes.

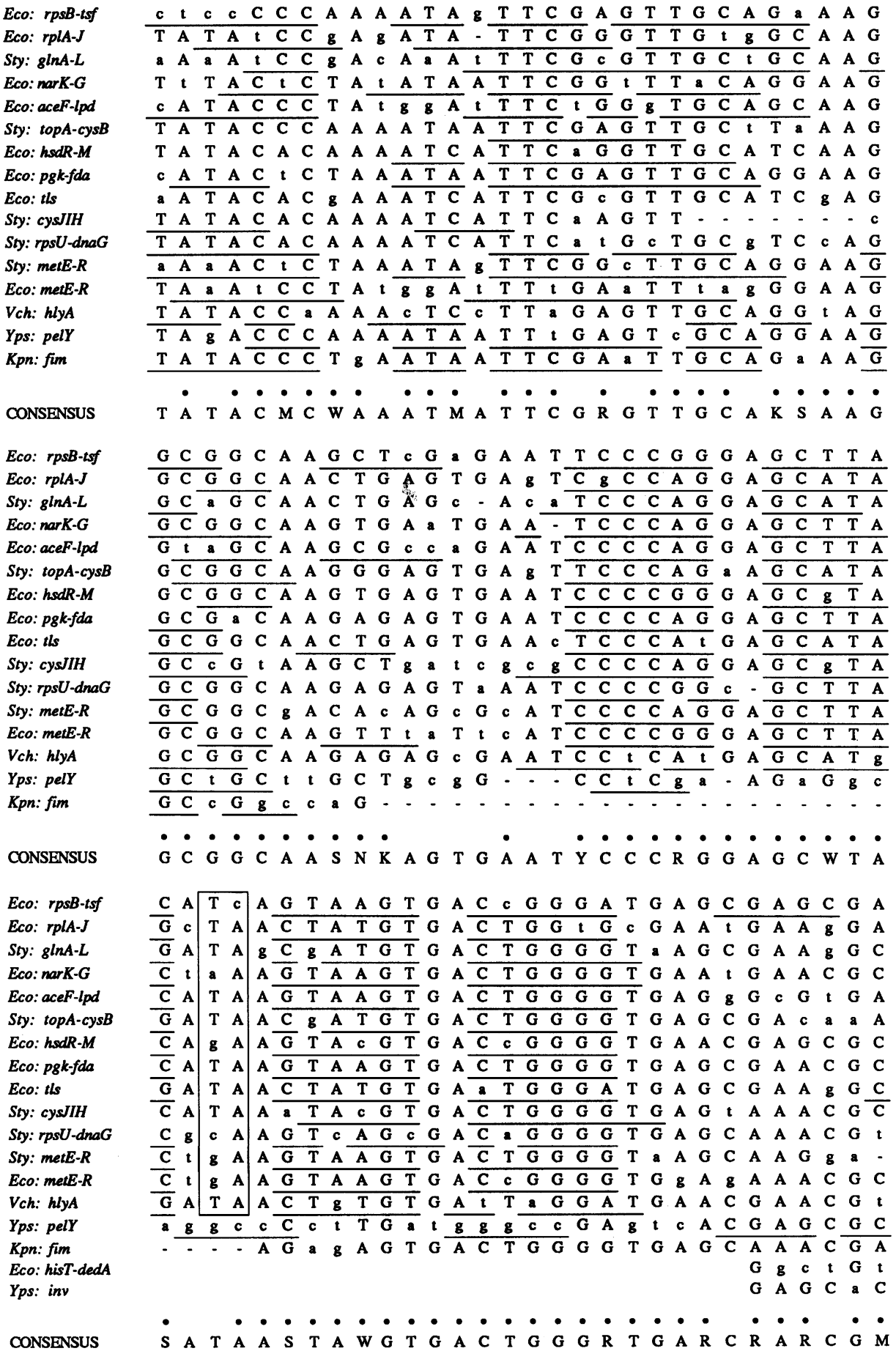
Several of the genetic regions containing an IRU have been sequenced in both *E. coli* and *S. typhimurium*. The *metE-R* region in both organisms contains an IRU at the identical location. However, this is the only case so far where this is true. In *E. coli* the sequences corresponding to the *S. typhimurium topA-cysB, cysJIH* and *rpsU-dnaG* regions are missing a section of DNA that corresponds precisely to the IRU (15, 16, 17). The *glnA-L* (*ntrB*) operon has been sequenced in *S. typhimurium, E. coli* K-12 and *Klebsiella pneumoniae* (18, 19, 20). The major difference between the three operons is the presence of an IRU in the *S. typhimurium* sequence which is completely deleted in *K. pneumoniae* and replaced in *E. coli* by two overlapping palindromic units. The *rplA* region has been sequenced in *Proteus vulgaris* and in the archaebacterium *Halobacterium marismortui* (21, 22). Neither sequence contains an IRU at the position found in the *E. coli rplA* region.

The alignment shown in Figure 1 indicates that a 25 bp sequence near the middle of the IRU in the *K. pneumoniae fim* region has been deleted. A slightly larger region in the centre of the IRU in the *Y. pseudotuberculosis pelY* region appears to have been replaced by an unrelated sequence. The sequences in the *E. coli hisT-dedA* and *Y. pseudotuberculosis inv* regions are included because of their high sequence similarity to the 3′ end of the IRU consensus. Whether or not they are residual fragments of IRU sequences is uncertain.

DISCUSSION

The analysis presented in this paper has revealed the presence of a novel palindromic DNA sequence in the chromosomes of *E. coli*, *S. typhimurium* and other bacteria. Like PU or REP sequences, these IRU sequences are found in the intergenic regions within and between operons and appear capable of adopting a stable secondary structure. They differ from PU sequences in that they are much longer and are not restricted to the 3′ transcribed regions of genes. In addition, although not as numerous as PU sequences, they are found in bacteria other than *E. coli* and *S. typhimurium*.

What function, if any, could these sequences have? Many are found within large multigene operons, which raises the possibility that they may regulate expression of the associated genes. An et al. (23) observed the potential structure of an IRU in the *E. coli rpsB-tsf* intergenic region and suggested that it could be cleaved in mRNA by RNase III. Analysis of a defined RNase III cleavage site revealed ribosome binding sites close to the cleavage position. McConnell (24) suggested that binding of ribosomes to the mRNA 5′ to the cleavage site could modulate the rate of transcriptional termination. Many of the IRUs do contain sequences that match the consensus (AGGAGG) for ribosome binding (25). Indeed, this may account for the high conservation of purines in single-stranded regions. However, from the locations of the IRUs identified here it is not certain that they are all transcribed. Also, in several cases the IRU is in the non-coding strand for the adjacent operon. Therefore, while IRUs and RNase III cleavage sites are similar in their (proposed) secondary structures, though not in their primary sequence, we do not believe that the IRUs are processed by RNase III. This


Eco: rpsB-tsf
Eco: rplA-J
Sty: glnA-L
Eco: narK-G
Eco: aceF-lpd
Sty: topA-cysB
Eco: hsdR-M
Eco: pgk-fda
Eco: tls
Sty: cysJIH
Sty: rpsU-dnaG
Sty: metE-R
Eco: metE-R
Vch: hlyA
Yps: pelY
Kpn: fim
CONSENSUS
Eco: rpsB-tsf
Eco: rplA-J
Sty: glnA-L
Eco: narK-G
Eco: aceF-lpd
Sty: topA-cysB
Eco: hsdR-M
Eco: pgk-fda
Eco: tls
Sty: cysJIH
Sty: rpsU-dnaG
Sty: metE-R
Eco: metE-R
Vch: hlyA
Yps: pelY
Kpn: fim
CONSENSUS
Eco: rpsB-tsf
Eco: rplA-J
Sty: glnA-L
Eco: narK-G
Eco: aceF-lpd
Sty: topA-cysB
Eco: hsdR-M
Eco: pgk-fda
Eco: tls
Sty: cysJIH
Sty: rpsU-dnaG
Sty: metE-R
Eco: metE-R
Vch: hlyA
Yps: pelY
Kpn: fim
Eco: hisT-dedA
Yps: inv
CONSENSUS

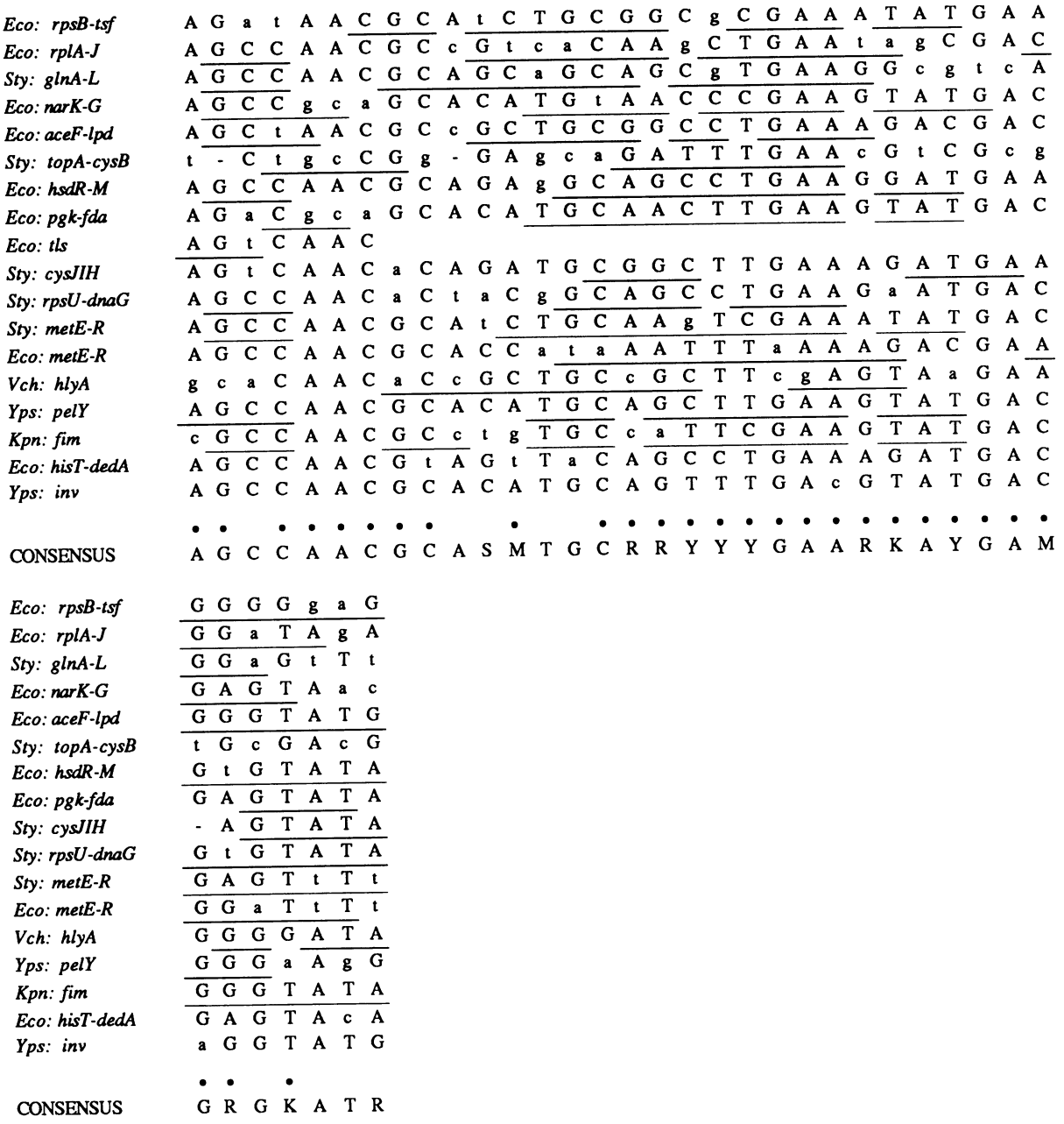


Figure 1. Alignment of inverted repeat units. Sequences are labelled as in Table 1. The consensus is based on the nucleotides present in at least 50% of the sequences. Nucleotides are shown in lower case where they do not match the consensus. The consensus is dotted where it applies to 80% or more of the sequences. Potential double-stranded DNA hairpin regions are underlined. The boxed region designates the centre of the inverted repeat. Y: pyrimidine, R: purine, W: A or T, S: G or C, K: T or G, M: A or C and N: any nucleotide.

conclusion is supported by the observation that RNase III cleavage sites are detectable in the mRNA corresponding to the *dnaG-rpoD* and *rplL-rpoB* intergenic regions, but not at the sites of the IRUs (see Fig. 3).

In those cases where the IRU is clearly transcribed, a stem-loop could no doubt be formed in the mRNA. This structure could interfere with ribosome binding and thereby affect translation. There is a secondary stucture with this property in the mRNA of the *E. coli rplA-J* region (26), but it is distinct from the IRU also found in this interval. The IRU structure could also protect the 3′ end of the message from exoribonuclease digestion, as is the case with PU sequences (7, 8, 9), and increase expression of the upstream gene(s). The fact that in the *glnA-L* region the IRU in *S. typhimurium* is replaced in *E. coli* by two palindromic units provides some support for this possibility.

An alternative possibility for IRUs is that they provide sequence or structural signals within the DNA itself. These signals could be recognised by a specific protein or complex of proteins involved in DNA metabolism, or may provide sites that are used to organise the chromosome within the nucleoid, as has been suggested for PU sequences (4, 13).

The search of the DNA sequence database revealed nine IRUs among the *E. coli* entries. Given that approaching 25% of the genome has been sequenced (27), we predict that the *E. coli* chromosome contains some 30–50 of these IRUs in total. Their frequency may be higher in other bacteria, but the number of

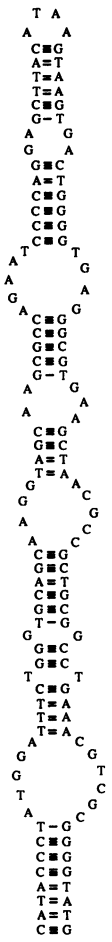


Figure 2. Predicted stem-loop structure of the *E. coli* IRU between *aceF* and *lpd*. Hydrogen bonding is indicated by horizontal bars. A single bar denotes the weak 'pairing' found between T and G base pairs.

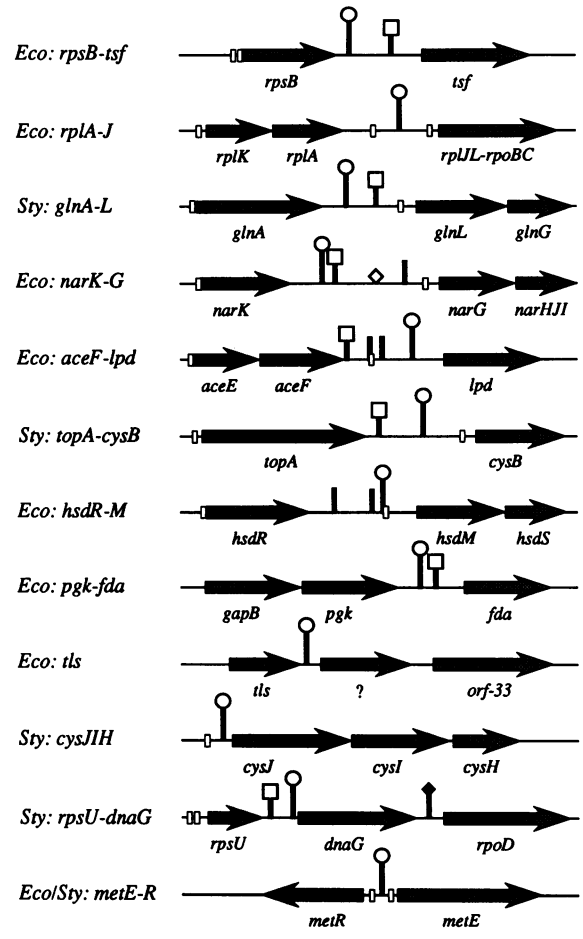


Figure 3. Schematic diagram (not to scale) showing the position of IRU relative to the flanking genes and their transcriptional signals. Coding regions and their directions of transcription are denoted by shaded arrows. Promoter or putative promoter regions are indicated by small open rectangles. Vertical bars represent inverted repeats and have an open square on top to denote a proposed r-independent terminator, an open circle to denote an IRU, and a closed diamond to denote an RNase III cleavage site. A directly repeated sequence between *narK* and *narG* is indicated by an open diamond. The sequence downstream of *tls* in *E. coli* appears to contain a gene of unknown function that encodes a protein of about 18 kDa (unpublished work).

database entries is too low to make accurate predictions. The IRU sequences identified are from Gram negative organisms, all but one of which are members of the *Enterobacteriaceae*. Since there is an IRU at the identical location within the *metE-metR* intergenic regions of *E. coli* and *S. typhimurium* we assume that sequences of this type must have been present in the evolutionary lineage that gave rise to both of these organisms. The presence of IRUs at certain locations in *S. typhimurium* but not at the equivalent locations in *E. coli* demonstrates that these sequences subsequently moved to new locations in the *Salmonella* lineage or were deleted in the *Escherichia coli* lineage. It also suggests that the presence of an IRU is not essential for the proper function of the operon in these cases. If the IRUs have been deleted from *E. coli* then it follows that the deletion events were very precise.

The mechanism by which these palindromic IRUs have spread is not clear. They do not appear to be related to any other repeated elements such as insertion sequences, transposons, or structural RNA genes (28, 29, 30, 31). They are too small to encode a transposase that could mediate their transposition. However, they could be relics of transposable elements or DNA sequences that were spread by functional transposable elements acting *in trans*, though there is no hint of the target site duplication normally associated with transposition. If they are just relics, then it is perhaps surprising that they have been retained in the bacterial lineage that gave rise to both *Escherichia coli* and *Vibrio cholerae*, especially with such a high sequence identity.

The evolution of bacterial genomes has been linked with duplications, inversions and other rearrangements (31, 32). These events are often linked with repeated DNA elements. While we have not been able to link IRUs with any known or postulated rearrangements in *E. coli* or *Salmonella* (32), it is possible that they do affect genome stability by promoting such rearrangements or by providing hot-spots for recombination. The partial IRU sequences detected in *E. coli* and *Y. pseudotuberculosis* provide support for this view.

ACKNOWLEDGEMENTS

This work was supported by grants to RGL from the Science and Engineering Research Council, the Wellcome Trust, and the Medical Research Council.

REFERENCES

1. Clément, J. M. and Hofnung, M. (1981) Cell **27**, 507–514.
2. Higgins, C. F., Ames, G. F.-L., Clément, J. M. and Hofnung, M. (1982) Nature **298**, 760–762.

3. Gilson, E., Clément, J. M., Brutlag, D. and Hofnung, M. (1984) EMBO J. **3**, 1417–1421.
4. Gilson, E., Clément, J. M., Perrin, D. and Hofnung, M. (1987) Trends in Genetics **3**, 226–230.
5. Stern, M. J., Ames, G. F.-L., Smith, N. H., Robinson, E. C. and Higgins, C. F. (1984) Cell **37**, 1015–1026.
6. Higgins, C. F., McLaren, R. S. and Newbury, S. F. (1988) Gene **72**, 3–14.
7. Merino, E., Becerril, B., Valle, F. and Bolivar, F. (1987) Gene **58**, 305–309.
8. Newbury, S. F., Smith, N. H., Robinson, E. C., Hiles, I. D. and Higgins, C. F. (1987) Cell **48**, 297–310.
9. Newbury, S. F., Smith, N. H. and Higgins, C. F. (1987) Cell **51**, 1131–1143.
10. Stern, M., Prossnitz, E. and Ames, G. F.-L. (1988) Mol. Microbiol. **103**, 143–153.
11. Yang, Y. and Ames, G. F.-L. (1988) Proc. Natl. Acad. Sci. USA **85**, 8850–8854.
12. Gilson, E., Perrin, D., Clément, J. M., Szmelcman, S., Dassa, E. and Hofnung, M. (1986) FEBS Lett. **206**, 323–328.
13. Gilson, E., Perrin, D. and Hofnung, M. (1990) Nucl. Acids Res. **18**, 3941–3952.
14. Sharples, G. J., Benson, F. E. and Lloyd, R. G. (1990) Nucl. Acids Res. submitted,
15. Ostrowski, J., Jagura-Burdzy, G. and Kredich, N. M. (1987) J. Biol. Chem. **262**, 5999–6005.
16. Ostrowski, J. and Kredich, N. M. (1989) J. Bacteriol. **171**, 130–140.
17. Burton, Z. F., Gross, C. A., Watanabe, K. K. and Burgess, R. R. (1983) Cell **32**, 335–349.
18. Hanau, R., Koduri, R. K., Ho, N. and Brenchley, J. E. (1983) J. Bacteriol. **155**, 82–89.
19. Miranda-Ríos, J., Sánchez-Pescador, R., Urdea, M. and Covarrubias, A. A. (1987) Nucl. Acids Res. **15**, 2757–2770.
20. MacFarlane, S. A. and Merrick, M. (1985) Nucl. Acids Res **13**, 7591–7606.
21. Sor, F. and Nomura, M. (1987) Mol. Gen. Genet. **210**, 52–59.
22. Arndt, E. and Weigel, C. (1990) Nucl. Acids Res. **18**, 1285.
23. An, G., Bendiak, D. S., Mamelak, L. A. and Friesen, J. D. (1981) Nucl. Acids Res. **9**, 4163–4172.
24. McConnell, D. J. (1979) Nucl. Acids Res. **6**, 3491–3503.
25. Stormo, G. D., Schneider, T. D. and Gold, L. M. (1982) Nucl. Acids Res. **10**, 2971–2996.
26. Christensen, T., Johnsen, M., Fiil, N. P. and Friesen, J. D. (1984) EMBO J. **3**, 1609–1612.
27. Kröger, M., Wahl, R. and Rice, P. (1990) Nucleic Acids Res. **18** (suppl.), 2549–2587.
28. Umeda, M. and Ohtsubo, E. (1989) J. Mol. Biol. **208**, 608–614.
29. Umeda, M. and Ohtsubo, E. (1990) J. Mol. Biol. **213**, 229–237.
30. Umeda, M. and Ohtsubo, E. (1990) Mol. Gen. Genet. **222**, 317–322.
31. Komeni, Y., Adachi, T., Inokuchi, H. and Ozeki, H. (1990) J. Mol. Biol. **212**, 579–598.
32. Riley, M. and Krawiec, S. (1987) In: Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. and Umbarger, H. E. (Eds), *Escherichia coli* and *Salmonella typhimurium* cellular and molecular biology, American Society for Microbiology, Washington DC, pp. 967–981.
33. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. and Dennis, P. P. (1979) Proc. Natl. Acad. Sci. USA **76**, 1697–1701.
34. Li, S. F. and DeMoss, J. A. (1987) J. Bacteriol. **169**, 4614–4620.
35. Stephens, P. E., Lewis, H. M., Darlison, M. G. and Guest, J. R. (1983) Eur. J. Biochem. **135**, 519–527.
36. Loenen, W. A. M., Daniel, A. S., Braymer, H. D. and Murray, N. (1987) J. Mol. Biol. **198**, 159–170.
37. Alefounder, P. R. and Perham, R. N. (1989) Mol. Microbiol. **3**, 723–732.
38. Erickson, B. D., Burton, Z. F., Watanabe, K. K. and Burgess, R. R. (1985) Gene **40**, 67–78.
39. Plamann, L. S. and Stauffer, G. V. (1987) J. Bacteriol. **169**, 3932–3937.
40. Maxon, M. E., Redfield, B., Cai, X.-Y., Shoeman, R., Fujita, K., Fisher, W., Stauffer, G., Weissbach, H. and Brot, N. (1989) Proc. Natl. Acad. Sci. USA **86**, 85–89.
41. Alm, R. A., Stroeher, U. H. and Manning, P. A. (1988) Mol. Microbiol. **2**, 481–488.
42. Manulis, S., Kobayashi, D. Y. and Keen, N. T. (1988) J. Bacteriol. **170**, 1825–1830.
43. Nonet, M. L., Marvel, C. C. and Tolan, D. R. (1987) J. Biol. Chem. **262**, 12209–12217.
44. Isberg, R. R., Voorhis, D. L. and Falkow, S. (1987) Cell **50**, 769–778.
45. Gerlach, G.-F., Allen, B. L. and Clegg, S. (1988) J. Bacteriol. **170**, 3547–3553.
46. Tinico, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. and Gralla, J. (1973) Nature (New Biol.) **246**, 40–41.
47. Nussinov, R. and Tinico, I. (1982) Nucl. Acids Res. **10**, 341–349.
48. Papincolaou, C., Gouy, M. and Ninio, J. (1984) Nucl. Acids Res. **12**, 31–44.