Nucleic Acids Research, 2001, Vol. 29, No. 17 **3583–3594**

Prediction of rho-independent transcriptional terminators in *Escherichia coli*

Elena A. Lesnik, Rangarajan Sampath, Harold B. Levene, Timothy J. Henderson, John A. McNeil and David J. Ecker*

IBIS Therapeutics, 2292 Faraday Avenue, Carlsbad, CA 92008, USA

Received May 7, 2001; Revised and Accepted June 20, 2001

ABSTRACT

A new algorithm called RNAMotif containing RNA structure and sequence constraints and a thermodynamic scoring system was used to search for intrinsic rho-independent terminators in the *Escherichia coli* K-12 genome. We identified all 135 reported terminators and 940 putative terminator sequences beginning no more than 60 nt away from the 3′-end of the annotated transcription units (TU). Putative and reported terminators with the scores above our chosen threshold were found for 37 of the 53 non-coding RNA TU and for almost 50% of the 2592 annotated protein-encoding TU, which correlates well with the number of TU expected to contain rho-independent terminators. We also identified 439 terminators that could function in a bi-directional fashion, servicing one gene on the positive strand and a different gene on the negative strand. Approximately 700 additional termination signals in non-coding regions (NCR) far away from the nearest annotated gene were predicted. This number correlates well with the excess number of predicted 'orphan' promoters in the NCR, and these promoters and terminators may be associated with as yet unidentified TU. The significant number of high scoring hits that occurred within the reading frame of annotated genes suggests that either an additional component of rho-independent terminators exists or that a suppressive mechanism to prevent unwanted termination remains to be discovered.

INTRODUCTION

Two mechanisms of transcription termination and two classes of termination signals have been described in bacteria: rho-dependent and rho-independent (1). The rho-independent termination signals consist of stable hairpins followed by U-rich regions. The mechanism of intrinsic termination includes pausing of the transcription elongation complex (TEC) at the termination point and formation of the RNA terminator hairpin inside RNA polymerase (RNAP) (1–8). Three nucleic acid binding sites have been characterized in the TEC: a double-stranded DNA binding site (DBS), an RNA:DNA hybrid binding site (HBS), and a single-stranded RNA binding site (RBS) (9,10). Formation of a stable RNA hairpin in combination with dissociation of a weak U-rich RNA:DNA hybrid duplex (11) results in disruption of the protein–nucleic acid interaction in the HBS and RBS of RNAP, destabilization and dissociation of TEC, and release of nascent mRNA (reviewed in 1,12,13).

A few attempts have been made to search for intrinsic rho-independent terminators (14,15). From statistical analysis of identified and predicted intrinsic terminators, d'Aubenton Carafa *et al.* (16) created a model of a typical intrinsic terminator that consisted of an RNA hairpin followed by a 15 nt 'T-region'. The hairpin and thymidine-rich region could be separated by a 'spacer' no longer than 2 nt. An adenosine-rich region was also identified upstream of the hairpin but not included in this terminator model. The scoring system developed included free energy of RNA hairpin formation and an empirical equation considering the number and position of thymidine residues in the T-region. Due to the lack of full genome sequences, annotated coding regions, and clear understanding of the molecular mechanism of intrinsic transcription termination at that time, these searches had limited success. Two recent searches for intrinsic terminators (17,18) used the same terminator model and scoring system (16). No information about hit positions, sequences or gene annotations were published.

For this study, a new algorithm called RNAMotif, which searches nucleic acid databases for RNA structure motifs (19), was used. The motif is defined by a 'descriptor' that contains user-defined constraints on the RNA sequence and structure. Our model of the intrinsic terminator was based on that of d'Aubenton Carafa *et al.* (16), but also included an 11 nt region upstream of the hairpin enriched in adenosine, called the 'A-region'. Our scoring system was based on the assumption that the efficiency of transcription termination depends on both the difference between the stability of the RNA hairpin and the RNA:DNA hybrid duplex within TEC and pausing of TEC at the termination point (7,8,12,20). Thermodynamic parameters for nucleic acid structure formation were used in our scoring system.

The annotated *Escherichia coli* K-12 genome containing 4 639 221 bp (21) was used for testing our novel RNAMotif algorithm and 'thermodynamic' scoring system. There were 1075 hits with scores better than the threshold found near the 3′-end of almost half of the 2592 known and predicted transcription units (TU) including 37 of the 53 TU of non-coding

*To whom correspondence should be addressed. Tel: +1 760 603 2347; Fax: +1 760 431 2768; Email: decker@isisph.com

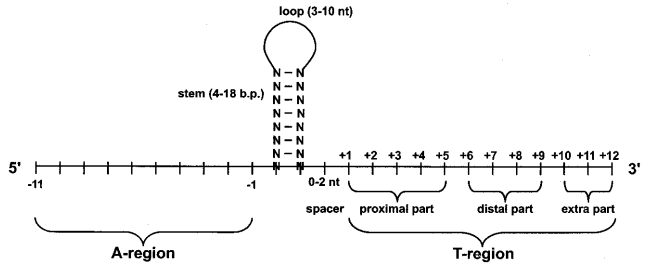

Figure 1. Rho-independent intrinsic terminator construct. Regions include from 5′ to 3′: (i) A-region of 11 nt; (ii) a hairpin with a loop of 3–10 residues and with one of three types of stems: perfect stems of 4–18 base pairs, stems of 9–18 base pairs with internal loop no longer than 20% stem length, or stems of 7–18 base pairs with 1–5 nt bulges in either 5′ or 3′-side of stems; (iii) a spacer of 0–2 nt (any bases except T); (iv) T-region divided into three parts: proximal part of 5 nt, distal part of 4 nt, and extra part of 3 nt. (We did not number spacer nucleotides since no more than 5% putative terminators contained 2 nt spacers and no more than 15% putative terminators contained 1 nt spacers.)

RNAs. We also predicted the existence of approximately 700 additional termination signals in non-coding regions (NCR) more than 60 nt away from the 3′-end of annotated genes and TU.

MATERIALS AND METHODS

Descriptors of intrinsic terminators

The RNAMotif algorithm (19) searches nucleic acid sequence databases for RNA structure motifs. The motif of interest must be defined by a 'descriptor' that contains the desired constraints on the RNA structure and sequence. To create our constraints we analyzed a published list of 147 experimentally identified and proposed terminator sequences (16). The BLAST program (22) identified 135 of these terminator sequences on the *E.coli* K-12 genome (21). Strings of thymidine residues that are a critical part of rho-independent terminators were not present in the remaining 12 sequences. Based on analysis of the 135 terminators, a model of an intrinsic terminator was defined that included the following regions from 5′ to 3′: an 11 nt adenosine-rich region (A-region), a variable-length hairpin, a variable-length spacer and a 12 nt thymidine-rich region (T-region) (Fig. 1). The A-region was added to the previously described model (16) to aid in identification of bi-directional terminators. Due to symmetry, the A-region on one strand becomes the T-region of the terminator on the opposite strand. The T-region was divided into three parts from 5′ to 3′ according to their specific functions in the intrinsic termination process (7): the 5 nt region nearest to the hairpin (proximal-T), the next 4 nt (distal-T), and the final 3 nt (extra-T) (Fig. 1). Structural constraints applied to the terminator hairpin are presented in the legend to Figure 1. Additional sequence constraints on the hairpin and the T-regions included the following: (i) the first base in the hairpin stem cannot be an A; (ii) there can be no less than four GC/CG or GT/TG base pairs in the hairpin stem; (iii) the proximal T-region must contain at least three T residues, no more than one G, and no 5′-TVVTT stretches (V is A, C or G); (iv) the distal T-region cannot have four purines or four cytosines; (v) there must be at least four T residues in the proximal and distal T-region. No special sequence constraints were applied to the A-region, spacer or the extra T-region.

Scoring system

The RNAMotif program calculates a score for each sequence match based upon a user-defined schema. Components of the motif can be independently scored, weighted and combined to create an overall score. We created a scoring system based on the proposed mechanism of intrinsic transcription termination (7). According to this mechanism, RNA hairpin formation inside the TEC is crucial for destabilization and dissociation of TEC. The ability of a hairpin stem to fold 'inside' the TEC strictly depends on disruption of the RNA:DNA hybrid duplex including the spacer (if it exists) and the proximal T-region (Fig. 1). The distal part of T-region plays a role in the slowing and pausing of TEC at the termination point, thus giving the hairpin time to form. The 3′-most part of T-region may enhance the pausing effect and may be important if the distal T-region is too weak to allow effective pausing of RNAP. The competition between stability of two adjacent helices, the RNA hairpin and the RNA:DNA hybrid, results in transcription termination if ΔG^0_{37} of hairpin formation [ΔG^0_{37} (hairpin)] exceeds ΔG^0_{37} of hybrid formation [ΔG^0_{37} (spacer) + ΔG^0_{37} (proximal-T)] and RNAP pauses at termination point. We created a thermodynamic equation to allow quantitation of these effects on termination:

$$\text{Score} = \Delta G^0_{37\,(\text{hairpin})} - [\Delta G^0_{37\,(\text{spacer})} + \Delta G^0_{37\,(\text{proximal-T})}] - 0.5 \times \Delta G^0_{37\,(\text{distal-T})} - 0.01 \times \Delta G^0_{37\,(\text{extra-T})} \quad (\mathbf{1})$$

The free energy of the RNA hairpin formation [$\Delta G^0_{37\,(\text{hairpin})}$] was calculated using the most recent thermodynamic parameters available (23). The free energy for the formation of the RNA:DNA hybrid duplex formed between the coding DNA and the nascent RNA (terms 2–5 in equation **1**) was calculated using nearest-neighbor thermodynamic parameters for RNA:DNA duplex formation (24).

The distal T-region is involved in kinetic process of slowing and pausing of TEC. Substitution of rU in position +8 (Fig. 1) by rC, rG and rA resulted in 90, 80 and 30% inhibition of termination, respectively, compared to rU in that position (7). The free energy of hybrid duplex formation mirrors this observed inhibition, with stability as follows: rGG:dCC > rCC:dGG >> rAA:dTT >> rUU:dAA (24). We used ΔG^0_{37} (distal-T) and ΔG^0_{37} (extra-T) as a penalty for inhibition of TEC pausing by subtracting them from $\Delta G^0_{37\,(\text{hairpin})}$. The base composition effect of the distal T-region on intrinsic termination *in vivo* is weaker than that observed *in vitro* experiments. We found that 21% of the terminators reported had rC or rG at position +8 and 7% had rC and rG in both positions +7 and +8 of the T-region (16). To reflect the reduced effect of the distal and extra parts of the T-region, the ΔG^0_{37} for these regions in equation **1** were multiplied by coefficients of 0.5 and 0.01, respectively. While many terminators have thymidine residues in positions +10 to +12 of the T-region, there are no clear experimental data on the effect of this part of the sequence on termination, therefore we entered a low coefficient in equation **1** for the extra T-region. If experimental data warrants, it would be easy to increase the weight of this term.

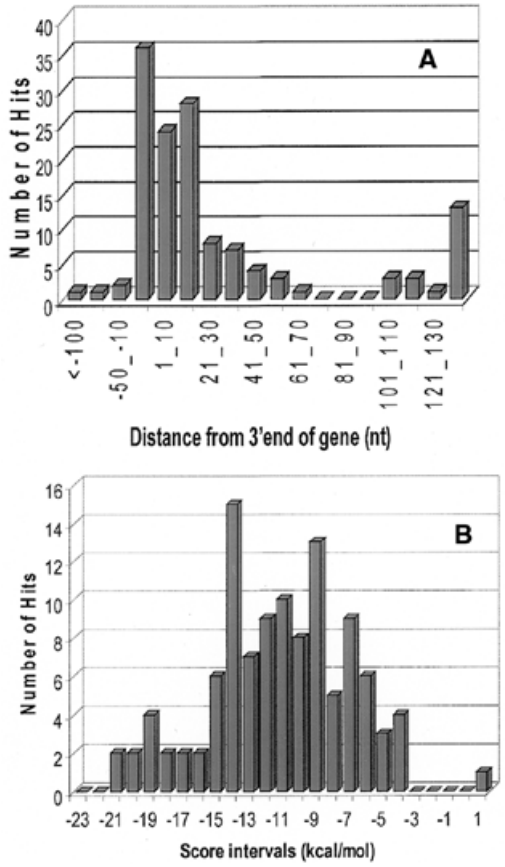

Figure 2. Distance distribution between the start of the hits and the end of nearest annotated genes (**A**) and score distribution (**B**) for the RNAMotif hits of Set A containing 135 reported terminators (16). A translation termination codon of an ORF was considered the end of a protein-encoding gene and the processed end was considered the end of a non-coding RNA gene.

Analysis of identified intrinsic terminators

The set of 135 terminators was used to determine sensitivity of the RNAMotif algorithm, descriptors and thermodynamic scoring system. We identified all 135 reported terminators and analyzed them for their positions relative to annotated genes and score values. The *E.coli* K-12 genome annotation (3) was used to calculate distances between the start of identified terminators and the end of nearest upstream genes. Almost 82% of the terminators started in NCR between 10 nt upstream [a terminator may start 10 nt upstream of the 3′-end of the genes because the 11 nt string (A-region) can be a part of the gene] and 60 nt downstream from the 3′-end of genes (either a translation termination codon or nucleotide corresponding to the processed end of a non-coding RNAs). The other 18% started further than 60 nt downstream of the end of nearest gene or in open reading frames (ORFs) (Fig. 2A). NCRs were defined as sequences between consecutive genes in the operons (intercistronic regions), sequences between consecutive operons and sequences on the antisense strand of coding regions. We considered hits in intercistronic regions as candidates for transcription terminators because termination signals starting in these regions serve frequently as attenuators and are often involved in transcriptional regulation (reviewed in 25–27). We also searched sequences on the antisense strand of coding regions for putative terminators of genes on the opposite strand. Score values calculated by the RNAMotif program for the set of 135 terminators varied from –21.73 to –4.61, with a peak number of hits in a score interval ranging from –10.0 to –14.0 (Fig. 2B). Only one terminator had a score >0 (+0.27); this terminator began 2 nt before the end of the uvrD gene (b3813) (21). The 109 identified terminators with scores less than –4.0 within the distance range –10 to +60 nt from the end of genes were used as a training set (Set A). A score of –4.0 was used as a threshold for this study.

Table 1. RNAMotif hit distribution (score threshold –4.0 kcal/mol)

Description	Distance from 3′-end of gene (nt)	Number of hits meeting the criteria
RNAMotif hits	Total genome	6635
NCR, Set B (putative terminators for annotated genes)	–10[a] _ +60	1075
NCRs	More than +60	2974
Intragenic regions	Less than –10	2586

[a]Negative distance designates a position upstream from the translational stop codons for protein-encoding genes or the processed end of non-protein-encoding RNA genes; positive distance designates a position downstream from the 3′-end of a gene.

Computing pD and pFA for Receiver Operating Characteristics (ROC) curve

An ROC curve is a parametric plot of the probability of detection (pD) versus the probability of false alarm (pFA) for a series of threshold values (scores) created by the detection algorithm. In our case, $pD_{(threshold)}$ was the fraction of all the putative true terminators whose absolute score value was higher than the absolute threshold value. For our purposes, $pFA_{(threshold)}$ represented the fraction of false positive hits (points) that did not represent terminators but had negative scores better (higher in absolute value) than threshold value. To account for the detection performance provided by the RNAMotif algorithm, false positive counts were divided by 8 000 000 to yield the pFA; there are approximately 8 000 000 possible starting points for genes in *E.coli*. To compute pD, the histogram of putative terminator points versus their score was created. We then integrated the number of points in the histogram for each score bin value, starting from the most negative score. Each score bin contained the number of known positives with scores at that defined score interval (–1 kcal/mol). The probability of detection was that number divided by the total number of putative true terminators. The pD climbs from 0 for the 'highest' to 1 for the 'lowest' negative threshold. We computed pFA as we did pD, except that it was divided by 8 000 000 and plotted on a logarithmic scale.

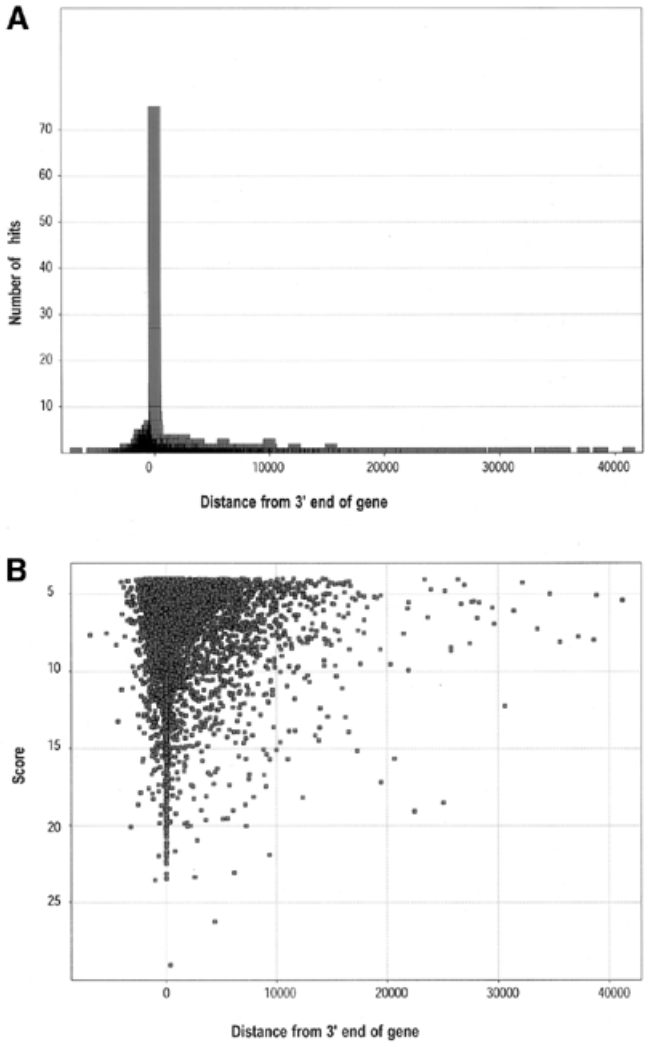

Figure 3. Distance distribution between the start of the hits and the end of nearest annotated genes (**A**) and score distribution (**B**) for the population of RNAMotif hits with scores under threshold –4.0 (ends of genes were designated as coordinate 0 on the *x*-axis). The data were fit by Spotfire software.

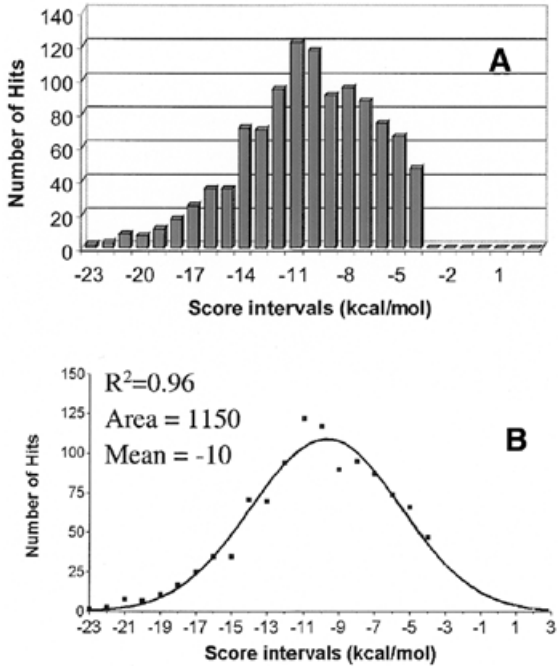

Figure 4. The score distribution histogram (**A**) and the Gaussian curve fitting it (**B**) for RNAMotif hits starting between –10 and +60 nt from the end of annotated genes (Set B). The lowest value of score bins were plotted on the *x*-axis. Each bin is 1 kcal/mol wide. The data were fit by Graphpad Prizm software.

RESULTS

Hit distribution between non-coding and intragenic regions of the *E.coli* genome

The annotated *E.coli* K-12 genome containing 4 639 221 bp (21) was used for testing the RNAMotif algorithm and 'thermodynamic' scoring system. Analysis of the *E.coli* K-12 genome using RNAMotif yielded 6635 hits with scores less than –4.0 (as the score is based upon thermodynamic parameters, a more negative score is better). Of these hits, 4049 began in NCRs while 2586 hits started in ORFs (Table 1). A pronounced peak on the hit distance distribution histogram (Fig. 3A) indicated that highest concentration of the hits was near the end of annotated genes. The score distribution histogram showed that most of the hits starting near the end of genes had good scores (Fig. 3B), while a low percentage of the hits found in non-coding and ORF regions had good scores. Score values for an overwhelming majority of the hits did not exceed –12.0.

General properties of putative terminators identified by RNAMotif

Putative terminators downstream of annotated genes. Based upon the score distribution and the location of the documented terminators of Set A relative to 3′-ends of known genes (Fig. 2), putative terminators should start in the region between 10 nt upstream and 60 nt downstream of the 3′-end of genes and possess scores less than –4.0. Of the 4049 hits in NCR, 1075 hits obey these criteria. These 1075 hits include the terminators in Set A and are called Set B. (The full list of 6635 hits and Set B are available as Supplementary Material. Terminators from Set A are designated as 'old term'.) The score distribution for the set B (Fig. 4A) fit well to a Gaussian curve ($R^2 = 0.96$) with a maximum in the score interval between –10.0 and –11.0 (Fig. 4B). Most (834 hits) began no further than 20 nt downstream from the gene end.

Structural features of putative and identified terminator sets. The comparison of the structural features of hits in Set B with those in Set A revealed general similarity: in both sets, tetraloop hairpins with 6–11 bp stems were predominant. However, there was a broader distribution of structural parameters in Set B than in Set A (Fig. 5). Approximately 60% of the hits in Set B were evenly distributed across stems of 7–11 bp. Only 9% of hairpins in Set B had short stems (4–6 bp), while 22% in Set A did. In Set B, 42% of the hits had long stems (10–13 bp), while only 24% of the hairpins in Set A did (Fig. 5A). Similar trends

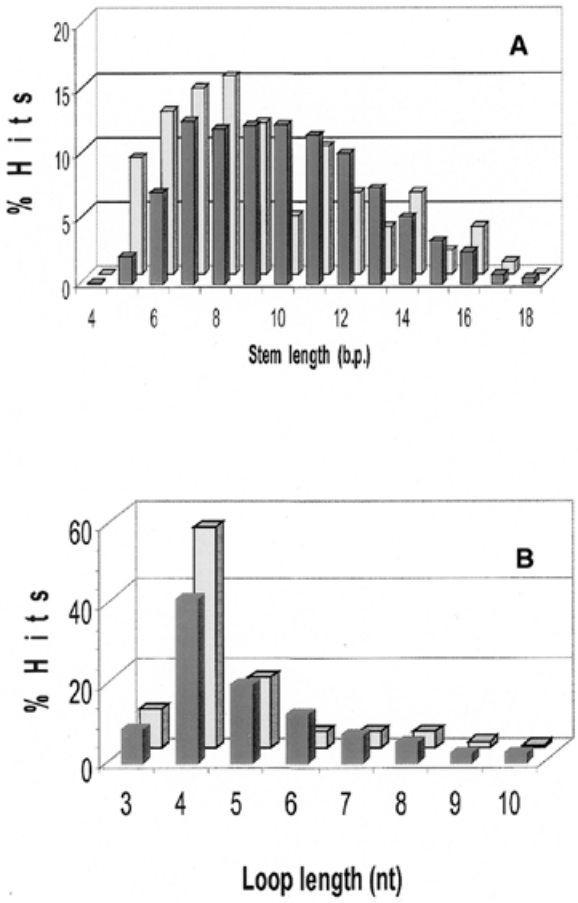

Figure 5. Comparison of hairpin structure parameters for hits in Set A (light bars) and Set B (dark bars): (**A**) stem length distribution; (**B**) loop length distribution.

were observed for loop length distribution (Fig. 5B): a broader profile with a shift to longer loops in Set B compared to Set A. Tetraloops accounted for 56% of the loops in Set A, but only 42% in Set B. In Set B, 30% of the loops contained 6–10 nt, while only 16% of those in Set A did. The two most stable tetraloops, UUCG and GAAA (28), were observed more frequently in Set A than in Set B. While 17% of tetraloops in Set A had UUCG loops, only 7.6% of those in Set B did. The GAAA tetraloop accounted for 5.4% of tetraloops in Set B versus 9.6 % in Set A. However, the thermodynamic stability of the hairpins in Set B was comparable to the stability of those in Set A.

Specific termination signals identified by RNAMotif algorithm

Bi-directional hits. A few intrinsic terminators in *E.coli* that function bi-directionally have been reported (29). Analysis showed that 878 of the hits (439 pairs) could serve as bi-directional terminators. These hits resided on opposite strands and were shifted by no more than 10 nt relative to each other and both had scores better than –4.0 (Fig. 6). Generally, hits in a pair had similar structure and close scores: 105 pairs (24% of the total) had <10% difference in score values. However, due

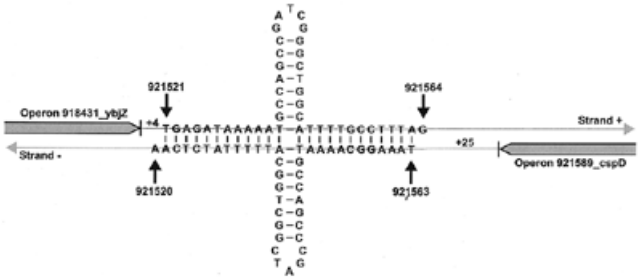

Figure 6. An example of bi-directional putative terminators starting 4 nt (score –11.28) and 25 nt (score –11.95) downstream of annotated genes in (+) and (–) strand, respectively.

to the difference in base pairing between T and G in one strand and between A and C in the opposite strand, significant differences in hairpin structure and score values were also observed (Table 2). Most of the bi-directional hits (94%) resided in NCR. Of these, 92 pairs are included in Set B, i.e. both hits in a pair start near the 3′-end of divergent genes and are putative terminators. In 288 pairs, only one hit was close to the end of an annotated gene, while the other resided in an NCR far from annotated genes. In the remaining pairs, both bi-directional hits resided in NCR far from the annotated genes or in ORF regions. Approximately 60% of the bi-directional hits had 6–10 adenosine residues and an additional 30% had 4–5 adenosines in an 11 nt A-region. (The full list of bi-directional hits is available as Supplementary Material.)

Consecutive hits. Almost one in 10 hits in NCR and one in eight hits in ORFs were followed by one or more additional hits with scores less than –4.0. We found 410 hits in NCR (of these, 134 met the criteria for inclusion in Set B) and 330 hits in ORFs that were followed by one or more additional hits no further than 200 nt from the end of an upstream hit (Table 3). (A full list of clusters of consecutive hits is presented as Supplementary Material.)

Putative terminators for the RNA genes

Putative terminators for rRNA operons. Seven rRNA operons are annotated in the *E.coli* K-12 genome (21). Each operon has two clusters of genes: a cluster of 16S rRNA with one or two tRNA genes followed by a downstream cluster of 23S and 5S rRNAs that may or may not include tRNA genes (30–33). Putative terminators were identified downstream of each of the seven operons following the final gene in the 23S rRNA cluster (either a 5S rRNA or a tRNA) (Table 4). Strong terminators with scores of less than –19.0 and almost identical sequence were found at the end of rrnA, rrnB and rrnD operons. Sequence differences were observed in the loop (a GAAA in the rrnB tetraloop instead of GGAA found in the rrnA and rrnD terminators) and in the second base pair of the hairpin stem (A:T in rrnD terminator instead of G:C in the two others) (Table 4). The putative terminators downstream of tRNA genes in operons rrnH, rrnE and rrnG had scores ranging from –9.8 to –12.0. The hairpins had stems of 8 bp closed by either tetraloops (in rrnH and rrnG) or a pentaloop (rrnE) (Table 4). A putative terminator with a score value below our threshold (–3.18) was found downstream of the last tRNA gene in the

Table 2. Bi-directional putative terminators from Set B

Strand	Score	Hit Start(+)[1] End(-)	Length	End(+)[1] Start(-)	Hit Sequence A string	5′-side stem	Loop	3′-side stem	Spacer	T region	Distance[2] I	II
-	-7.69	658407	41	658448	ctgataaaaaa	cccgctt	tttg	aagcggg		ttttttgtatcg		25
+	-7.54	658409	41	658450	gatacaaaaaa	cccgctt	caaa	aagcggg	.	tttttatcaga		35
-	-7.93	850197	41	850238	aaaaacaaaaa	tccgcc	ccgaga	ggcgga	.	ttttttatatca	1582	-1
+	-10.91	850199	41	850240	gatataaaaaa	tccgcct	ctcg	gggcgga	.	tttttgttttta		10
-	-10.81	2864488	43	2864531	gcacaataaaa	ggtccgat	gccc	atcggacc	.	tttttattaagg		50
+	-9.87	2864490	43	2864533	cttaataaaaa	ggtccgat	gggc	atcggacc	.	ttttattgtgca		1
-	-12.98	2904608	43	2904651	tgatttaaaaa	tgccagcc	tccg	ggctggca	.	ttttttatggtt		13
+	-12.22	2904610	43	2904653	accataaaaaa	tgccagcc	cgga	ggctggca	.	tttttaaatcag		4
-	-9.56	3156897	51	3156948	gttaaaaaata	tccccggcaac	tgacac	gctaccgggga	.	ttttttttatcat	1237	-4
+	-18.31	3156899	51	3156950	tgataaaaaaa	tccccggtagc	gtgtca	gttgccgggga	.	tatttttttaacg	255	-2
-	-8.61	3486529	56	3486585	caggcaaaaga	tgtcggatgcggcg	tgaa	tgccttatccgatg	g	tttagcctgcat		11
+	-4.95	3486534	53	3486587	aggctaaacca	tcggataaggc	attcac	gccgcatccga ca		tcttttgcctga		3
-	-10.66	4637103	41	4637144	gtcaaaaaaaa	cggcgct	tttt	agcgccg	.	tttttatttttc		14
+	-9.91	4637105	41	4637146	aaaaataaaaa	cggcgct	aaaa	agcgccg	.	ttttttttgacg		5

Table 3. Clusters of consecutive RNAMotif hits

Strand	Score	Hit Start	End	A string	Hit sequence 5′-side stem	Loop	3′-side stem	Spacer	T region	Hit residing	5′ Gene	Distance[1]
+	-4.45	3267592	3267629	aaaagtaaaaa	cccgc	cgaa	gcggg	.	tttttacgtaaa	NCR	OPERON(3267304)_b3122	
+	-4.45	3267705	3267742	aaaagcaaaaa	cccgc	cgaa	gcggg	.	tttttacgtaaa	NCR	OPERON(3267304)_b3122	76
+	-4.45	3267818	3267855	aaaagcaaaaa	cccgc	cgaa	gcggg	.	tttttacgtaaa	NCR	OPERON(3267304)_b3122	76
+	-5.76	3267933	3267988	gcagcctaccc	gggttcagtacg	ggc	cgtaccttatgaaccc	c	tatttggccttg	NCR	OPERON(3267304)_b3122	78
-	-8.05	3068171	3068131	gcttatctcaa	ggcccgc	tct	gcgggtc	.	ttttttcgcca	NCR	OPERON(3068185)_fba	
-	-5.71	3068096	3068049	cctacacgtat	cctgcaa	tatattgaat	ttgcagg	.	tttttgtaggcc	NCR	OPERON(3068185)_fba	35
-	-5.25	3068002	3067963	atctgtctgag	ccccgc	tct	gcgggg	a	ttttttcgccaa	NCR	OPERON(3068185)_fba	47
-	-8.51	2944079	2944028	tacggatatga	cagggtgagg	ggtaa	cctcacatcctg	a	ttattgactaag	NCR	OPERON(2944103)_mltA	
-	-5.61	2943953	2943904	gagttatgtag	gcctgataagcg	tag	cgcatcaggc	a	tttatgcgttta	NCR	OPERON(2944103)_mltA	75
-	-13.13	2943915	2943871	ttatgcgttta	gggtgagg	gataa	cctcgccc	.	tttttaattctg	NCR	OPERON(2944103)_mltA	-11
-	-5.40	566689	566638	tgtgcgcggta	gctgaccggg	ctgaact	tccgggaggc	c	tttgccctcagt	NCR	OPERON(573960)_trs5_2	
-	-5.87	566649	566598	ttgccctcagt	ccctgacggcg	caggct	tgccgccacgg	.	tttttacgttaa	NCR	OPERON(573960)_trs5_2	-11
-	-8.57	566518	566466	acaacgctgtc	gcagtgttggcg	gaac	tgctgacgcgtgc	.	ttatccttgtcc	NCR	OPERON(573960)_trs5_2	80
-	-8.66	566281	566216	gacgtgcaatc	tcggtagacatctccag	ttcacgtt	cagaagacgtctgctga	.	ttttgctgttta	NCR	OPERON(573960)_trs5_2	185
+	-9.06	1941558	1941618	gcctcattact	tggcgtcgcgtttggt	ttgtt	gctggacgtgaatcca	.	ttctatgcggtg	ORF	OPERON(1940686)_yebI	
+	-4.40	1941776	1941832	ttacctgttcg	gtgatttgctggc	agtgac	gccagaagatctc	a	tctctattgcga	ORF	OPERON(1940686)_yebI	158
+	-4.40	1941955	1942017	tgctggtgacg	gcattgacgattggtgt	agcg	atgaaatttgtcggtgc	g	ttgattattact	ORF	OPERON(1940686)_yebI	89
+	-5.82	1942066	1942133	cggaacagatg	gctggtgtcgctgttt	tggtgg	ggatggtggcagtgactggc	gg	tttaacctttttc	ORF	OPERON(1940686)_yebI	49
+	-7.13	1942151	1942197	ggcgggtccgt	cggtggtgc	tatgt	gcggcactg	.	ttatttattctc	ORF	OPERON(1940686)_yebI	18
-	-4.28	4239434	4239374	attattggcgg	tattggcgttggttt	agcctca	atgctctcgccaatg	.	tatattgcggaa	ORF	OPERON(4238358)_xylE	
-	-8.42	4239368	4239310	ccagctcatat	tcgcgggaaactggt	ctctttta	accagtttgcga	.	ttattttcgggc	ORF	OPERON(4238358)_xylE	6
-	-11.83	4239278	4239219	gcccgttccgg	tgatgccagctggc	tgaatact	gacggctggcgtta	.	tatgtttgcctc	ORF	OPERON(4238358)_xylE	32

[1]Distance = end of the upstream hit – start of the consecutive hit.

rrnC operon. It consisted of the hairpin with low GC content in the stem (ΔG^0_{37} = –4.6 kcal/mol) closed by an AAA triloop. However, the T-region contained the sequence TTTTTTT-TATCT, which will form a very unstable RNA:DNA heteroduplex (ΔG^0_{37} = –1.8 kcal/mol), thus, the whole construct may function effectively as an intrinsic terminator.

Two groups of identical sequences with scores between –7.0 and –8.5 were identified inside operons rrnH, rrnA and rrnD, and inside operons rrnC, rrnB and rrnE. (data not shown). They started between 15 and 27 nt downstream from tRNA genes at the end of 16S rRNA clusters. These sequences in the spacer regions of rRNA operons are actually *rrn* anti-terminators that mediate efficient transcription through rho-dependent terminators (34–36). Consecutive hits with strong scores (–11.42 to –17.32) were found downstream of the putative terminators in three operons: rrnA, rrnB and rrnD (Table 4, Additional hits). They resided in the NCR between 112 and 118 nt downstream of the end of the putative terminators for rRNA operons.

Putative terminators for tRNA transcription units. There are 86 tRNA genes annotated in the *E.coli* genome (21). Of these, 14 tRNA genes are included in rRNA operons. The other 72 tRNA genes are organized in 36 transcription units of 23 single tRNA genes and 13 operons containing from two to seven tRNA genes (Table 5). We found 23 RNAMotif hits obeying our distance-score requirements for putative terminators. Fourteen hits were found in NCR downstream of single tRNA genes and nine hits were found downstream of the operons.

Table 4. RNAMotif putative terminators for rRNA operons in *E.coli* K-12 genome

Operon	Gene Description	Strand	Score	Hit Start	Length	A-string	Hit Sequence 5′-side stem	Loop	3′-side stem	Spacer	T region	Distance[1] I	II
rrnH	23S+5S+tRNA	+	-9.82	229006	43	ttattaagaag	cctcgagt	taac	gctcgagg	.	ttttttttcgtc		1
rrnC	23S+5S+2 tRNA	+	-3.18	3944653	43	gccagaaatca	tccttagcg	aaa	cgtaagga	.	ttttttttatct	72	-3
rrnA	23S+5S	+	-21.11	4038215	55	atcaaataaaa	gaaaggctcagtc	ggaa	gactgggcctttcg[2]	.	ttttatctgttg	118	-1
rrnB	23S+5S	+	-21.11	4169334	55	atcaaataaaa	cgaaaggctcagtc	gaaa	gactgggcctttcg	.	ttttatctgttg	118	-1
rrnE	23S+5S	+	-11.52	4210741	44	aattagaaaaa	ccccggtc	cataa	ggccgggg	.	ttttttgcatat		2
rrnG	23S+5S	-	-11.42	2724085	43	attatgcgaaa	ggccatcc	tgac	ggatggcc	.	tttttgcattgg		3
rrnD	3S+5S+tRNA+5S	-	-19.31	3421060	55	atcaaataaaa	caaaaggctcagtc	ggaa	gactgggccttttg	.	ttttatctgttg	119	-1
Additional Hits													
rrnA downstr 5S rrfA		+	-11.42	4038388	43	attaagcagaa	ggccatcc	tgac	ggatggcc	.	tttttgcattgg		171
rrnB downstr 5S rrfB		+	-15.25	4169501	57	catcaaattaa	gcagaaggccatcc	tgac	ggatggccttttttgc	g	tttctacaaact		165
rrnD downstr 5S rrfF		-	-17.32	3420893	56	catcaaatcaa	gcgaaaggccatcc	gaaa	ggatggccttttttgc	.	ttttcgaactaa		165

[1]Distance I = start of the hit – start of the 5′ gene (in strand +) or start of the 3′ gene – start of the hit (in strand –); distance II = start of the hit – end of the 5′ gene (in strand +) or end of the 3′ gene – start of the hit (in strand –).
[2]The hits with long t-stretches in 3′-side of hairpin stems are artifacts of the program search. We suggest that those sequences form short hairpins with t-stretches shifted toward T-region.

The analysis of these 23 putative transcription terminators for tRNA genes revealed considerable variation in sequences and structures. The length of terminator regions varied from 37 to 65 nt and scores ranged from –4.36 to –14.54. The hairpins had stems that ranged from 5 to 17 bp with loops of 3–10 nt. Only two identical putative terminators were found, those downstream of the last Lys-tRNA genes in TU4 and TU11 (Table 5). In all other cases, genes encoding the same tRNAs had different putative terminator sequences at different distances from the ends of the genes (Table 5). For nine single tRNA genes and four tRNA operons, no sequences meeting out terminator criteria were found. Either terminator sequences for these genes do not obey the constraints used in this study, or these tRNA transcripts are terminated via another mechanism.

Identical hits with scores of –4.10 were found inside the operons TU4 and TU11, and in NCR upstream of the first Lys-tRNA gene in TU4 (Table 5, Additional hit 4_1). The analysis of these six sequences showed that none were real terminator sequences but instead were intrinsic parts of down-tream Lys-tRNAs genes. The A-regions of these pseudo-terminators began 9 nt upstream of the start of Lys-tRNA genes. The sequences identified as hairpin and T-region of the terminator comprised the 5′-part of the Lys-tRNA gene including the D-loop and the acceptor stem. Thus, no putative terminators were found in intercistronic regions of tRNA operons suggesting that these operons are transcribed into a single pre-tRNA transcript, as has been observed experimentally for some transcripts (37).

Putative terminators for small non-coding RNAs. Putative terminators were found for seven of the 10 small regulatory RNAs known in *E.coli* (38). Some of the regulatory RNAs are processed from long primary transcripts while others are transcripts from single RNA genes (Table 6). Four of the putative terminators were for primary transcripts of single genes (TU 3, 6, 7 and 10) and began 30 or more nt upstream from the 3′-end of genes. Two other putative terminators were found in the NCR downstream of TU4 and TU9. These transcripts are processed post-transcriptionally from long pre-mRNA and these regions may serve as terminators for the long transcript. For three other processed RNAs (TU 1, 2 and 5), no putative terminators were found in the vicinity of the 3′-ends of genes. These RNAs may be processed from a long primary transcript under control of terminators for nearby genes. Three consecutive hits (two of them were identical) were found downstream of the putative terminator for rnpB RNAse P RNA gene (TU9) (Table 6, Additional hits). The results indicate that, in contrast to protein encoding genes, non-coding RNA genes may include all or only part of the terminator sequence within the transcript (Table 6). The terminator sequences may be removed during post-transcriptional processing or may be critical for the function or structure of the RNA.

High scoring hits in non-coding regions

In addition to the 1075 hits of Set B, 2974 hits with score less than –4.0 were found in NCR more than 60 nt downstream from the nearest annotated gene (Table 1). The score distribution histogram for these hits fit two overlapping Gaussian curves (Fig. 7A). A small portion of hits with highest negative scores was fit by a Gaussian curve ($R^2 = 0.94$) with a maximum in a score interval of –10.0 to –12.0 as was observed for Set B (Fig. 7B). This population probably consists of actual termination signals that may function to terminate transcription from the excess of predicted promoters in the *E.coli* genome (39). The vast majority of the hits in NCR (more than 2000) belong to a population that would fit a Gaussian curve with a maximum near score value of +3.0 and are apparently false positive hits.

DISCUSSION

Thermodynamic model for scores of RNAMotif hits

The rho-independent termination pathway includes three consecutive steps: the TEC pausing at termination point, the disruption of hybrid duplex formed between newly synthesized RNA and template DNA strand, and the formation of an RNA hairpin resulting in irreversible destabilization and dissociation of the TEC (7,8,12). A thermodynamic model for intrinsic transcription termination was developed by Yager and von Hippel (1,20), where the net TEC stability [$-\Delta G^0_{37\,(TEC)}$] was described as an algebraic sum of three free energy terms:

$$\Delta G^0_{37\,(TEC)} = \Delta G^0_{37\,(DNA\ bubble)} + \Delta G^0_{37\,(RNA:DNA)} + \Delta G^0_{37\,(TEC\text{-}NA)} \quad (2)$$

where $\Delta G^0_{37\,(DNA\ bubble)}$ is free energy of the formation of the DNA transcription bubble, $\Delta G^0_{37\,(RNA:DNA)}$ is the free energy of

Table 5. RNAMotif putative terminators for tRNA genes in *E.coli* K-12 genome

TU	Gene in TU	Hit Descr.	Start of Gene	Strand	Hit Score	Hit Start	Hit Length	Hit Sequence A-string	5′-side stem	Loop	3′-side stem	Spacer	T region	Distance[1] I	Distance[1] II
1		term	236931	+	-11.32	237020	61	gaaaatgagtt	cagagagccgcaagat	tttta	attttgcggttttttg[2]	.	tatttgaattcc		19
2		no Hit	262095	+											
3		no Hit	563946	+											
4	4_1	no Hit	779777	+											
	4_2	pseudo-term	779988	+	-4.10	780057	53	cccaccaccgg	gtcgttagctc	agttggta	gagcagttgac	.	ttttaatcaatt	69	-6
	4_3	no Hit	780066	+											
	4_4	pseudo-term	780291	+	-4.10	780361	53	ccaccactcgg	gtcgttagctc	agttggta	gagcagttgac	.	ttttaatcaatt	70	-5
	4_5	pseudo-term	780370	+	-4.10	780583	53	cacccgtatgg	gtcgttagctc	agttggta	gagcagttgac	.	ttttaatcaatt		137
	4_6	pseudo-term	780592	+	-4.10	780791	53	tcccgcaaggg	gtcgttagctc	agttggta	gagcagttgac	.	ttttaatcaatt		123
	4_7	term	780800	+	-4.38	780875	38	aatgtaaaaaa	gcgcc	ctaaa	ggcgc	.	ttttttgctatc	75	0
5	5_1	no Hit	1744459	+											
	5_2	no Hit	1744540	+											
6		no Hit	2042571	+											
7		term	2057873	+	-5.92	2057944	55	agccaaattcc	tgaaaagcccgct	tttat	agcgggatttttg[2]	c	tatatctgataa	71	-4
8		term	2060282	+	-10.64	2060351	55	ggagccaaatt	caaaaaagcctgct	ttct	agcaggctttttg[2]	c	tttctaattacc	69	-6
9		term	2284231	+	-5.33	2284314	40	ccaagaaaaaa	ccaacc	cttac	ggttgg	.	ttttttatatc		6
10		no Hit	2464329	+											
11	11_1	no Hit	2518951	+											
	11_2	no Hit	2519071	+											
	11_3	pseudo-term	2519193	+	-4.10	2519264	53	caccacttcgg	gtcgttagctc	agttggta	gagcagttgac	.	ttttaatcaatt	71	-4
	11_4	term	2519273	+	-5.18	2519348	38	aatgtaaaaaa	gcgcc	ctaaa	ggcgc	.	ttttttactatc	75	0
12	12_1	no Hit	2945409	+											
	12_2	no Hit	2945519	+											
	12_3	term	2945629	+	-4.36	2945704	37	caattattgaa	caccc	taac	gggtg	.	ttttttgtttc	75	-1
13		term	3108383	+	-12.41	3108456	53	ccactaattct	taagaacccgcc	cacaa	ggcgggtttttg	c	ttttggatctga	73	-2
14		term	3213239	+	-14.31	3213315	45	gatatagcaaa	ggctgacga	gaaa	tcgtcagcc	.	ttttctttttta		0
15		no Hit	3833849	+											
16	16_1	no Hit	3979988	+											
	16_2	no Hit	3980122	+											
	16_3	no Hit	3980219	+											
	16_4	term	3980348	+	-9.56	3980422	39	ccaattttgaa	ccccgc	ttcg	gcgggg	.	ttttttgttttc	74	-2
17	17_1	no Hit	4172967	+											
	17_2	no Hit	4173051	+											
	17_3	no Hit	4173252	+											
	17_4	term	4173333	+	-7.82	4173447	61	cattcaacaag	tcgggcatgttgcc	tggttgatgt	ggtgatatcaccga	.	tttatccgtgtc		38
18	18_1	no Hit	4389938	+											
	18_2	no Hit	4390050	+											
	18_3	term	4390161	+	-4.03	4390251	48	taaaatatcca	cagcgacgc	gat	gcgttattgctg	g	tttttgttgtct		14
19		term	4493973	+	-6.71	4494061	39	gtatgtaaata	gacctc	aact	gaggtc	.	ttttttatgcc		3
20	20_1	no Hit	696280	-											
	20_2	no Hit	696186	-											
	20_3	no Hit	696088	-											
	20_4	no Hit	695979	-											
	20_5	no Hit	695887	-											
	20_6	no Hit	695765	-											
	20_7	term	695653	-	-7.11	695653	39	acattaaaaaa	gctcgc	ttcg	gcgagc	.	tttttgcttttc	74	0
21		term	925107	-	-13.88	925114	65	caccgccatat	ttaaagaagagctcgtac	gaaa	gtgcgagctttttttttcg[2]	ca	ttttatctgtct	80	-7
22		term	1030848	-	-14.54	1030848	46	aaataagataa	ggggttagc	taaat	gctaacccc	.	ttttctttttgc	87	0
23		term	1096788	-	-12.35	1096795	63	caccgccatat	ttaaagaagagctcgtac	gaaa	gtacgggctttttttttcg[2]	.	tatattgcacac	80	-7
24	24_1	no Hit	1286761	-											
	24_2	no Hit	1286467	-											
25	25_1	no Hit	1990065	-											
	25_2	no Hit	1989937	-											
	25_3	term	1989838	-	-4.61	1989802	37	gtgtcgtgaaa	ccacc	ttcg	ggtgg	.	ttttttttgtgcc		35
26		no Hit	2041490	-											
27		term	2056049	-	-11.82	2055997	52	aattatttttta	cccggttacctt	gtaaa	aaagtgaccggg	.	ttttttgatttc		51
28	28_1	no Hit	2516176	-											
	28_2	term	2516061	-	-10.88	2516062	43	caaatttccaa	ccctcgct	gcaa	agcggggg	.	ttttttgtctct	74	-1
29		no Hit	2783782	-											
30	30_1	no Hit	2816575	-											
	30_2	no Hit	2816495	-											
	30_3	no Hit	2816220	-											
	30_4	no Hit	2816081	-											
	30_5	no Hit	2815806	-											
31		no Hit	2997006	-											
32		no Hit	3315854	-											
33		term	3319713	-	-11.52	3319716	54	accaaattcca	gaaaagagacgct	gaaa	agcgtctttttc[2]	g	ttttggtcctgg	83	-3
34		term	3706245	-	-9.58	3706245	42	aaattcgaaaa	gcctgctc	aac	gagcaggc	.	ttttttgcatct	76	0
35		term	4360129	-	-12.98	4360127	43	attcatataaa	cggacctc	cacg	gaggtccg	.	tttttcgtttca		1
36	36_1	no Hit	4603884	-											
	36_2	no Hit	4603769	-											
	36_3	no Hit	4603647	-											
		Additional Hits													
4	4_1	pseudo-term redundant		+	-4.10	779768	53	cgaaaaagtgg	gtcgttagctc	agttggta	gagcagttgac	.	ttttaatcaatt		155
	4_7	downstr add redundant		+	-4.45	781031	38	aatgtaaaaaa	gcgcc	ctaaa	ggcgc	.	ttttttgctatt		155
9		downstr add		+	-5.73	2284357	55	aattaattcga	taaacagaccgtga	caca	tcacagcctgttta	.	ttttctgttatc		49
11	11_4	downstr add redundant		+	-4.38	2519504	38	aatgtaaaaaa	gcgcc	ctaaa	ggcgc	.	ttttttgctatc		155
13		downstr add		+	-7.30	3108534	67	cttataaagtc	tgggggaattactctcg	ccacgttaa	cgagagtaattttattg[2]	a	tattaatctcct		75
23		downstr add redundant		-	-12.35	1096614	63	caccgccatat	ttaaagaagagctcgtac	gaaa	gtacgggctttttttttcg[2]	.	tatattgcacac		173
		downstr add redundant		-	-11.15	1096433	63	caccgccatat	ttaaagaagagctcgtac	gaaa	gtacgggctttttttttcg[2]	.	tgtattgcacac		354
		downstr add redundant		-	-12.35	1096252	63	caccgccatat	ttaaagaagagctcgtac	gaaa	gtacgggctttttttttcg[2]	.	tatattgcacac		535
32		downstr add		-	-4.42	3315695	36	acagaataact	gggc	tttag	gccc	.	ttttttatgtc		158
33		downstr add		-	-7.76	3319625	60	gttgctttact	cgtagctggtacc	ttgaaacgat	ggtgccggtacg	cc	ttagttataaat		87
34		downstr add redundant		-	-9.58	3706077	42	aaattcgaaaa	gcctgctc	aac	gagcaggc	.	tttttgcatct		167

[1]Distance I = start of hit – start of 5′ gene (in strand +) or start of 3′ gene – start of hit (in strand –); distance II = start of hit – end of 5′ gene (in strand +) or end of 3′ gene – start of hit (in strand –).

[2]The hits with long t-stretches in 3′-side of hairpin stems are artifacts of the program search. We suggest that those sequences form short hairpins with t-stretches shifted toward T-region.

Table 6. RNAMotif putative terminators for small non-coding RNA genes in *E.coli* K-12 genome

TU	Hit Descr.	Start of Gene	Strand	Score	Hit Start	Length	A-string	5′-side stem	Loop	3′-side stem	Spacer	T region	Distance[1] I	II	Gene Name	Comments[2]
1	no Hit	475648	+												ffs_4.5S RNA	processed
2	no Hit	1647406	+												DicF RNA	processed
3	term	2311104	+	-7.89	2311159	42	ccctatttcaa	ccggatgc	ctc	gcattcgg	.	ttttttttaccc	47	-38	micF RNA	primary tr[3]
4	term	2753509	+	-16.94	2754013	49	aagtcctaaga	gcccgcacggc	gcaa	gccctgcgggc	.	ttttttgtgccc		38	ssrA (tmRNA)	processed
5	no Hit	3054003	+												ssrS_6S RNA	processed
6	term	4047479	+	-20.16	4047544	49	aatattttagc	cgccccagtca	gtaa	tgactggggcg	.	ttttttattggg	64	-43	spf spot 42RNA	primary tr
7	term	2023334	-	-15.71	2023285	46	agcaagtttca	tcccgaccc	cctca	gggtcggga	.	ttttttttattgt	49	-35	dsrA	primary tr
8	term	2922539	-	-17.57	2922224	49	cgaaacgaacc	gggagcgctgt	gaat	acagtgctcc	.	ttttttttattcc	315	-46	csrB_RNA	unknown
9	term	3268233	-	-5.81	3267853	37	ttacgtaaaaa	cccgc	ttcg	gcggg	.	tttttgcttttg		3	rnpB_RNaseP RNA	processed
10	term	4155973	-	-9.25	4155894	42	cgtgaactttt	gcggatc	tccag	gatccgc	.	ttttttttgcca	79	-30	OxyS RNA	primary tr
	Additional Hits															
9	downstr redundant		-	-5.81	3267740	37	ttacgtaaaaa	cccgc	ttcg	gcggg	.	tttttgcttttg		116		
	downstr redundant		-	-6.11	3267627	37	ttacgtaaaaa	cccgc	ttcg	gcggg	.	ttttacttttg		229		
	downstr add		-	-5.76	3267398	49	agatggcgtag	gcctgataagcg	tag	cgcatcaggc	a	tttttgcggttg		458		

[1]Distance I = start of the hit – start of the 5′ gene (in strand +) or start of the 3′ gene – start of the hit (in strand –); distance II = start of the hit – end of the 5′ gene (in strand +) or end of the 3′ gene – start of the hit (in strand –).
[2]Comments on RNA processing from Wassarman *et al.* (38).
[3]'primary tr' designates primary transcript.

RNA:DNA hybrid formation and $\Delta G^0_{37\ (TEC\text{-}NA)}$ is the free energy gain from the interaction between RNAP and nucleic acids. The formation of the RNA hairpin at termination point disrupts most of the hybrid structure and reduces interaction between RBS and nascent RNA and presumably between DBS and DNA (8,9,14,15). Therefore, effectiveness of the intrinsic termination is driven mainly by the difference between stabilities of the RNA hairpin [$\Delta G^0_{37\ (hairpin)}$] and the RNA:DNA hybrid duplex [$\Delta G^0_{37\ (RNA:DNA)}$]. Equation **1** used for score calculation was based on this competitive model and includes additional terms reflecting the base composition effects of the T-region on TEC pausing.

Termination signals in NCR

Application of the RNAMotif algorithm resulted in identification of all 135 reported terminators (16). Additional 940 putative terminator sequences obeying distance and score criteria were found near the 3′-end of the annotated TU (Set B). Set B is a homogeneous population that fits to a single Gaussian curve (Fig. 5B). Thus, we believe that these additional 940 hits are functional terminators. If it is so, rho-independent events terminate approximately half of the 2300 annotated TU in the *E.coli* genome in good agreement with predictions (1). The broad shape of the curve (Fig. 5B) indicates that there is a probability that some true terminators with score values worse than our threshold exist. As an example, the putative terminator for rrnC operon had a score of –3.18, slightly worse than our threshold of –4.0.

Almost 3000 hits were identified in NCR, far from annotated TU. This population was fit by two overlapping Gaussian curves (Fig. 7B). Hits with the most negative scores were fit by a Gaussian curve (R^2 = 0.94) very similar to that for Set B, suggesting that this population consists of actual termination signals. Some may be terminators for as yet unidentified genes. Others may be additional terminators that provide a backup to upstream terminators (40) or termination signals used to halt erroneous transcription from spurious promoters. The area under the curve reflects a population of 778 termination signals. If we apply our score threshold of –4.0 to this population the number of expected signals is reduced to approximately 700. The second curve (Fig. 7B) covering more than 2000 hits presumably outlines a population of false positive hits. This raises the issue of whether our constraints were too loose. The application of more restrictive constraints on the T-region resulted in significant reduction of the total number of hits, but

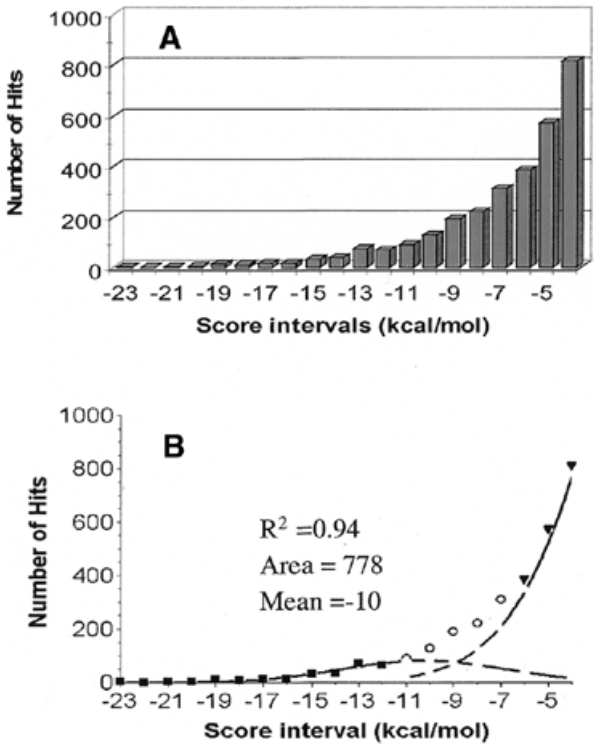

Figure 7. The score distribution histogram (**A**) and the Gaussian curves fitting it (**B**) for RNAMotif hits beginning in NCRs further than 60 nt downstream from the end of annotated genes. The lowest value of score bins were plotted on the *x*-axis. Each bin is 1 kcal/mol wide. The data were fit by Graphpad Prizm software.

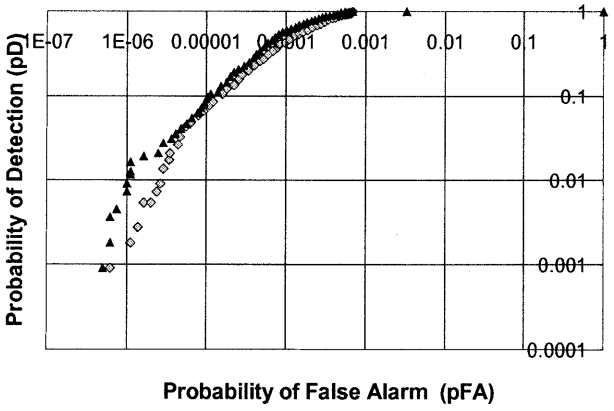

Figure 8. True terminator prediction probability obtained from equation **1** scores (triangles) versus free energy of hairpin formation [$\Delta G^0_{37\ (hairpin)}$] (diamonds).

Table 7. Detection performance at several thresholds for both the score (equation **1**) and free energy of hairpin formation

Threshold	pD	pFA	Number of missed true terminators	Number of retained false positive hits	Probability of prediction of true terminator (%)
Combined parameter score (equation **1**)					
–5.7	0.9	3.5E–04	110	2812	26
–10.4	0.5	7.2E–05	550	577	48
–15.9	0.1	1.1E–05	990	90	56
–20.9	0.01	1.0E–06	1089	8	59
Free energy for a hairpin [$\Delta G^0_{37\ (hairpin)}$]					
–10.0	0.9	1.0E–03	110	4054	20
–14.6	0.5	2.5E–04	550	1014	35
–20.2	0.1	3.2E–05	990	126	48
–24.7	0.01	5.3E–06	1089	21	32

it also reduced the detection of known terminators. In this experiment, constraints were chosen that allowed detection of all known terminators and did not significantly reduce the number of viable terminator candidates.

Evaluation of the prediction probability

When creating and comparing detection algorithms that separate a set of data points into hits and non-hits, it is helpful to have a performance metric. A common method of data evaluation is an ROC curve, discussed in detail in the Materials and Methods. We tested two thermodynamic scoring systems: one represented by equation **1** and second by the free energy of the RNA hairpin formation alone [$\Delta G^0_{37\ (hairpin)}$]. ROC curves for both scoring systems were plotted in logarithmic scale (Fig. 8). The curve corresponding to the score calculated using equation **1** is to the left of that for the hairpin-only score at all points. This indicates that the equation **1** score provides fewer false positives than the hairpin-only score at all pD values.

In Table 7, performance statistics were converted to absolute values for several pD points. Also computed were the percent of points above a given threshold that represent true terminators. The right hand column shows the probability, for a given pD value, that any predicted terminator will be functional when experimentally tested. The equation **1** score significantly improves the odds of success over the hairpin-only score. For example, at the threshold values that yield pD = 0.01, the combined score nearly doubles (59 versus 32%) the chance that a given selection will turn out to be a true terminator when evaluated experimentally.

The ROC curves as presented are adequate for comparing algorithms, as we have done here. To correctly predict the true ROC, one would have to know the set of all actual terminators. The pFA predictions are artificially high because true terminators for unknown genes and other termination signals in NCR are not correctly classified in the ROC analysis. The analysis of hit population residing far from annotated genes in NCR (Fig. 7B) predicts the existence of an additional 700 putative termination signals with the same score distribution as for the Set B members. When the ROC curve is corrected by assuming that these signals are correct (not shown), the pFA estimate for the score calculated using equation **1** improves from 4×10^{-5} to 1×10^{-5} at a pD value of 0.1. At the same pD, the estimate for the probability of a predicted terminator is functional improves from 56 to 81%.

Specificity of termination signals

The results obtained indicate that no more than one-third of 6635 hits with score less than –4.0 found in NCR and ORFs are reported or putative terminators. This raises the question about the role of the other two-thirds of this population. Almost 1000 are the 'consecutive hits' situated no more than 200 nt downstream of the first hit. In NCR, they may enhance the effectiveness of the primary terminator (40), serve as terminators for as yet unidentified small ORFs and RNA genes residing between consecutive terminators or function as processing signals for upstream genes. Some of these apparent terminator sequences are actually part of other signals and structures. For example, six hits with scores between –7.0 and –8.5 (sequences not shown) are previously identified anti-terminators of rho-dependent termination in spacer regions of rRNA (26,34–36,41,42). What the RNAMotif algorithm identified as the hairpin of an intrinsic terminator was actually the 'BoxB' hairpin of the anti-terminator, followed by a T-rich region characteristic of the 'BoxA' signal. Another example is found in the 5′ terminal region of Lys-tRNA genes, where sequences that match our terminator criteria function as the acceptor stem and the D-loop of the tRNA (pseudo-terminators in Table 5, TU4 and TU11). Thus, these are examples of sequences that match terminator constraints but are buried in the context of a larger signal that has a different function.

The question also rises about the role of the hits with good scores residing in ORFs. More than 40 hits with the parameters of very strong intrinsic termination signals (score less than –10.0 and sequences TTTTT or TTTTA in the proximal T-regions) were found in ORFs. While it is tempting to consider these hits to be false positives, we cannot identify any features of these hits that make them less suitable as terminators than terminators with similar scores located in the 'correct' positions near the ends of ORFs. One possibility is that there is a requirement

missing from our search constraints that would separate true functional terminators from false positives. Alternatively, there may be signals outside the terminator sequence that suppresses termination in inappropriate places, as occurs with attenuators. A third possibility is that these terminators actually stop transcription much of the time, producing truncated transcripts that are rapidly degraded. Experimental analysis will be required to distinguish these possibilities.

CONCLUSIONS

The *E.coli* genome consists of more than 2200 annotated TU (18,39). Using the RNAMotif algorithm (21), in combination with descriptors including structure and sequence constraints and a thermodynamic scoring system, putative rho-independent terminators have been identified for 1075 of the annotated genes and operons. Putative terminators for all annotated rRNA operons and small non-coding RNAs as well as for more than 64% of tRNA TU were found. This suggests that the transcription termination of RNA genes occurs mainly via rho-independent mechanism or via combination of both rho-independent and rho-dependent mechanisms as was reported for *rrnG* operon (40). Putative terminators were found for half of the protein encoding ORFs suggesting that both rho-dependent and rho-independent mechanisms participate equally in transcription termination of protein-encoding TU as was predicted (1). We have also predicted the existence of an additional 700 putative termination signals in NCR further than 60 nt downstream from the annotated genes. These numbers are in agreement with the excess number of promoters predicted in the *E.coli* genome (39). Sequences between predicted orphan promoters and terminators in NCR may be candidates for as yet undiscovered genes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr J. Wyatt for her thorough and very useful editing of the manuscript and W. Schuelke for technical support during the manuscript preparation.

REFERENCES

1. von Hippel,P.H. (1998) An integrated model of the transcription complex in elongation, termination and editing. *Science*, **281**, 660–665.
2. Farnham,P.J. and Platt,T. (1981) Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. *Nucleic Acids Res.*, **9**, 563–577.
3. Wilson,K.S. and von Hippel,P.H. (1995) Transcription termination at intrinsic terminators: The role of the RNA hairpin. *Proc. Natl Acad. Sci. USA*, **92**, 8793–8797.
4. Reynolds,R., Bermudez-Cruz,R.M. and Chamberlin,M.J. (1992) Parameters affecting transcription termination by *Escherichia coli* RNA polymerase. I. Analysis of 13 rho-independent terminators. *J. Mol. Biol.*, **224**, 31–51.
5. von Hippel,P.H. and Yager,T.D. (1992) The elongation-termination decision in transcription. *Science*, **255**, 809–812.
6. Uptain,S.M. and Chamberlin,M.J. (1997) *Escherichia coli* RNA polymerase terminates transcription efficiently at rho-independent terminators on single-stranded DNA templates. *Proc. Natl Acad. Sci. USA*, **94**, 13548–13553.
7. Gusarov,I. and Nudler,E. (1999) The mechanism of intrinsic transcription termination. *Mol. Cell*, **3**, 495–504.
8. Yarnell,W.S. and Roberts,J.W. (1999) Mechanism of intrinsic transcription termination and antitermination. *Science*, **284**, 611–615.
9. Zhang,G., Campbell,E.A., Minakhin,L., Richter,C., Severinov,K. and Darst,S.A. (1999) Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell*, **98**, 811–824.
10. Korzheva,N., Mustaev,A., Kozlov,M., Malhotra,A., Nikiforov,V., Goldfarb,A. and Darst,S.A. (2000) A structural model of transcription elongation. *Science*, **289**, 619–625.
11. Martin,F.H. and Tinoco,I.J. (1980) DNA–RNA hybrid duplexes containing oligo(dA:rU) sequences are exceptionally unstable and may facilitate termination of transcription. *Nucleic Acids Res.*, **8**, 2295–2299.
12. Landick,R. (1997) RNA polymerase slides home: pause and termination site recognition. *Cell*, **88**, 741–744.
13. Nudler,E., Gusarov,I., Avetissova,E., Kozlov,M. and Goldfarb,A. (1998) Spatial organization of transcription elongation complex in *Escherichia coli*. *Science*, **281**, 424–428.
14. Brendel,V. and Trifonov,E.N. (1984) A computer algorithm for testing potential prokaryotic terminators. *Nucleic Acids Res.*, **12**, 4411–4427.
15. Brendel,V., Hamm,G.H. and Trifonov,E.N. (1986) Terminators of transcription with RNA polymerase from *Escherichia coli*: what they look like and how to find them. *J. Biomol. Struct. Dyn.*, **3**, 705–723.
16. d'Aubenton Carafa,Y., Brody,E. and Thermes,C. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem–loop structures. *J. Mol. Biol.*, **216**, 835–858.
17. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
18. Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.
19. Macke,T., Ecker,D.J., Gutell,R., Gautheret,D., Case,D. and Sampath,R. (2001) RNA Motif – A new RNA secondary structure definition and discovery algorithm. *Manuscript in preparation.*
20. Yager,T.D. and von Hippel,P.H. (1991) A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*. *Biochemistry*, **30**, 1097–1118.
21. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
24. Sugimoto,N., Nakano,S., Katoh,M., Matsumura,A., Nakamuta,H., Ohmichi,T., Yoneyama,M. and Sasaki,M. (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, **34**, 11211–11216.
25. Landick,T., Turnbough,J.R. and Yanofsky,C. (1996) Transcription attenuation. In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella typhimurium*. American Society for Microbiology, Washington, DC.
26. Richardson,J.P. and Greenblatt,J. (1996) Control of RNA chain elongation and termination. In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella typhimurium*. American Society for Microbiology, Washington, DC.
27. Henkin,T.M. (1996) Control of transcription termination in prokaryotes. *Annu. Rev. Genet.*, **30**, 35–57.
28. Antao,V.P., Lai,S.Y. and Tinoco,I.,Jr (1991) A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.*, **19**, 5901–5905.
29. Postle,K. and Good,R.F. (1985) A bidirectional rho-independent transcription terminator between the *E. coli* tonB gene and an opposing gene. *Cell*, **41**, 577–585.
30. Young,R.A. (1979) Transcription termination in the *Escherichia coli* ribosomal RNA operon rrnC. *J. Biol. Chem.*, **254**, 12725–12731.
31. Brosius,J., Dull,T.J., Sleeter,D.D. and Noller,H.F. (1981) Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*. *J. Mol. Biol.*, **148**, 107–127.

32. Lund,E., Dahlberg,J.E., Lindahl,L., Jaskunas,R., Dennis,P.P. and Nomura,M. (1976) Transfer RNA genes between 16S and 23S rRNA genes in rRNA transcription units of *E. coli*. *Cell*, **7**, 165–177.
33. Schlessinger,D. (1980) In Chambliss,G., Craven,G.R., Davies,J., Davis,K., Kahan,L., Nomura,M. (eds), *Ribosomes: Structure, Function and Genetics*. University Park Press, Baltimore, pp. 767–780.
34. Berg,K.L., Squires,C. and Squires,C.L. (1989) Ribosomal RNA operon anti-termination. Function of leader and spacer region box B-box A sequences and their conservation in diverse micro-organisms. *J. Mol. Biol.*, **209**, 345–358.
35. Squires,C.L., Greenblatt,J., Li,J., Condon,C. and Squires,C.L. (1993) Ribosomal RNA antitermination *in vitro*: Requirement for Nus factors and one or more unidentified cellular components. *Proc. Natl Acad. Sci. USA*, **90**, 970–974.
36. Albrechtsen,B., Squires,C.L., Li,S. and Squires,C. (1990) Antitermination of characterized transcriptional terminators by the *Escherichia coli* rrnG leader region. *J. Mol. Biol.*, **213**, 123–134.
37. Deutscher,M.P. (1984) Processing of tRNA in prokaryotes and eukaryotes. *CRC Crit. Rev. Biochem.*, **17**, 45–71.
38. Wassarman,K.M., Zhang,A. and Storz,G. (1999) Small RNAs in *Escherichia coli*. *Trends Microbiol.*, **7**, 37–45.
39. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Blattner,F.R. and Collado-Vides,J. (2000) RegulonDB(version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 65–67.
40. Albrechtsen,B., Ross,B.M., Squires,C. and Squires,C.L. (1991) Transcriptional termination sequence at the end of the *Escherichia coli* ribosomal RNA *G* operon: complex terminators and antitermination. *Nucleic Acids Res.*, **19**, 1845–1852.
41. Pfeiffer,T. and Hartmann,R.K. (1997) Role of the spacer boxA of *Escherichia coli* ribosomal RNA operons in efficient 23 S rRNA synthesis *in vivo*. *J. Mol. Biol.*, **265**, 385–393.
42. Nodwell,J.R. and Greenblatt,J. (1993) Recognition of boxA antiterminator RNA by the *E. coli* antitermination factors NusB and ribosomal protein S10. *Cell*, **72**, 261–268.