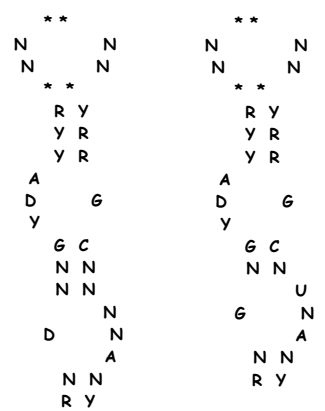
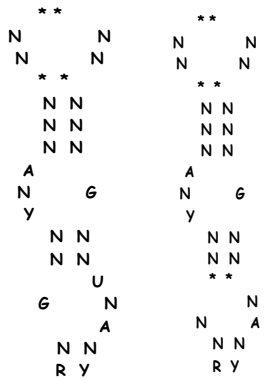
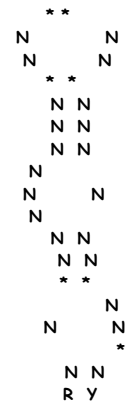
A search for ribonucleic antiterminator sites in bacterial genomes: Not only antitermination?

Noa Gordon, Ronen Rosenblum, Anat Nussbaum-Shochat, Elad Eliahoo and Orna Amster-Choder*

Supplemental Data

	Strict	Medium	Loose
Sequence-Structure Consensus			

Descriptor file	descr
	h5 (len=2,seq="^r")
	ss(tag='iL1',seq="^d$")
	h5 (minlen=2, maxlen=3,seq="g$")
	ss (seq="^yda$")
	h5
	(minlen=3,maxlen=4,seq="^yyr")
	ss(minlen=4,maxlen=6)
	h3 ss(seq="^g$")
	h3 (seq="^c")
	ss(tag='iL2',minlen=2,maxlen=3,seq="a$")
	h3
	score {
	if (length(ss(tag='iL2'))==3) {
	if (ss(tag='iL2')=~"una" &&
	ss(tag='iL1')=="g" &&
	length(h3(index=9))==2)
	ACCEPT;
	else
	REJECT; }}
		descr
		h5(len=2,seq="^r")
		ss(tag='iL1',len=1)
		h5 (minlen=2, maxlen=3)
		ss (seq="^yna$")
		h5 (minlen=3,maxlen=4)
		ss(minlen=4,maxlen=6)
		h3 ss(seq="^g$")
		h3
		ss(tag='iL2',minlen=2,maxlen=3)
		h3
		score {
		if (length(ss(tag='iL2'))==3) {
		if (ss(tag='iL2')=~"una" &&
		ss(tag='iL1')=="g"&&
		length(h3(index=9))==2)
		ACCEPT;
		else
		REJECT;}
		else
		if(length(ss(tag='iL2'))==2) {
		if(ss(tag='iL2')=~"na")
		ACCEPT;
		else
		REJECT;}
		else
		REJECT;}
			descr
			h5 (len=2,seq="^r")
			ss(len=1)
			h5(minlen=2,
			maxlen=3)
			ss (len=3)
			
			h5(minlen=3,maxlen=4)
			ss(minlen=4,maxlen=6)
			h3
			ss(len=1)
			h3
			ss(minlen=2,maxlen=3)
			h3

Fig. S1. The three RNAMotif descriptors used to search for putative RATs.
(A) Graphical representation of the sequence-structure consensuses with the different levels of stringency. The letters are based on standard nucleotide nomenclature.**(B)** The input files submitted to RNAMotif .

A

```
parms
  wc+=gu;

descr
    h5 (minlen=6,maxlen=18,pairfrac=0.8)
     ss (minlen=3,maxlen=10)
    h3
  ss (len=10)
score
{
  ucount=0; gclo=0; i=0; ulo=0;
  for(i=1;i<=length(h5[1]);i++){
    b=h5[1,i,1];
    w=h3[3,$-i+1,1];
    if((b=="g" && w=~"y") || (b=~"y" && w=="g"))
      gclo++;}
  len=length(h3[3]);
  for(i=len-4;i<=len;i++)
    if(h3[3,i,1]=="u") ulo++;
  for(i=1;i<=10;i++)
    if(ss[4,i,1]=="u") ucount++;
  if(gclo<4 || ucount<4 || ulo<3)
   REJECT;
  else
   ACCEPT;
}
```

B

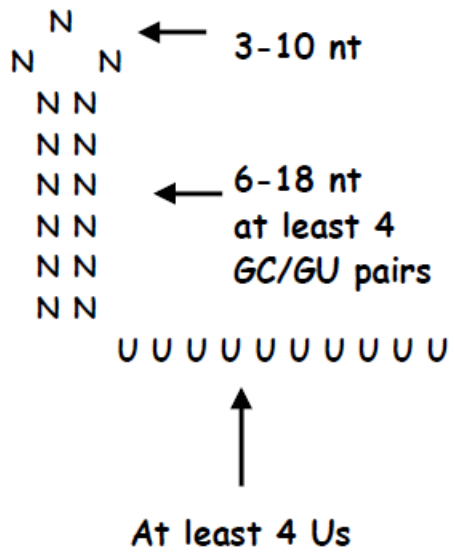


Fig. S2. The RNAMotif descriptor for Rho-independent terminators.
(A) The input files submitted to RNAMotif . (B) Graphical representation of the sequence-structure consensus.

Figure S3. Radioactively labeled RNA probes used in the electrophoretic mobility shift assays.
The radioactively labeled RNA probes were prepared in an *in vitro* transcription reaction, as described in Materials and Methods, and analyzed on a 6% polyacrylamide and 8 M urea gel.
1.RAT-*pafR*, contains the sequence of the putative RAT in the UPEC *pafR* gene (78 b);
2.RAT/ter-*bgl*, contains the sequence of the first RAT and its overlapping Rho-independent terminator in the *E. coli* K12 *bgl* operon (132 b);
3. RAT-*bgl*, contains the sequence of the first RAT in the *E. coli* K12 *bgl* operon (83 b);
4.Control RNA, contains the multiple cloning site of the pGEM vector, from which the different RNA elements were expressed (56 b, negative control).
5.No RNA, DNA template was not added to the in vitro transcription reaction.
M, ΦX174 DNA digested with Hinf I.

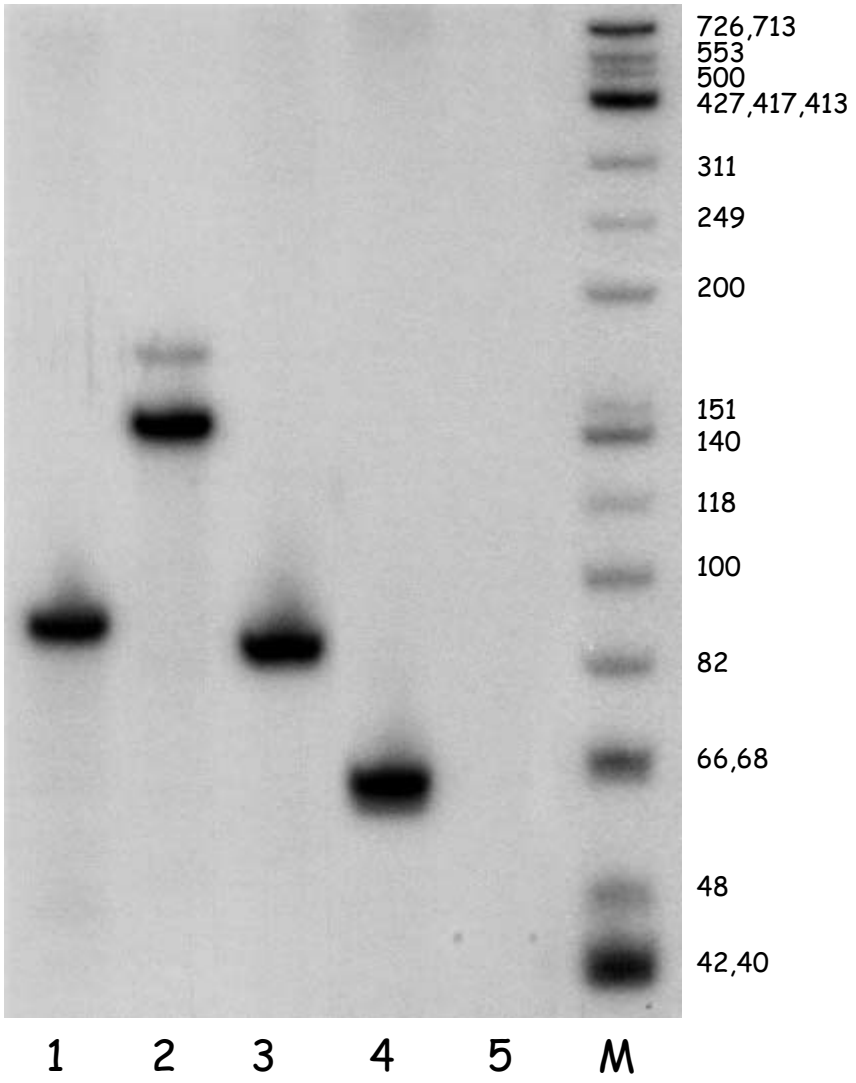


Figure S4. Binding of radioactively labeled RAT-*pafR* to the PafR protein can be competed by unlabeled RAT-*pafR*.
Radioactively labeled RAT-*pafR* was incubated without (lane 1) or with (lanes 2-4) purified MBP-tagged PafR protein. Increasing amounts of unlabeled RAT-*pafR* were added, as indicated above the lanes. Samples were analyzed on 5% non-denaturing acrylamide gel. F, free RNA; B, protein-bound RNA.

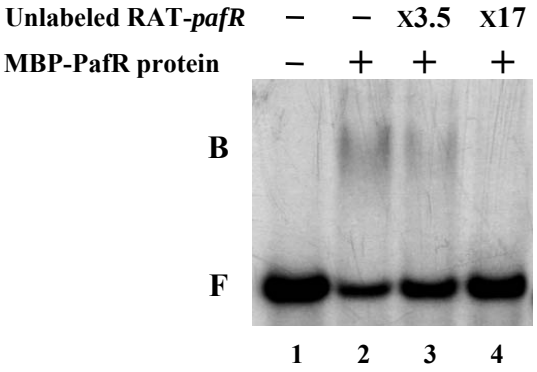


Table S1. Putative RATs identified by the 'strict' and 'medium' descriptors in the genome of *E. coli* K-12 using RNAMotif

Gene & ORF coordinates	Gene product	Gene on the other strand & overlap	RAT descriptor stringency	RAT location and orientation vs. gene	Nearby terminator (stem loop) coordinates & orientation vs. gene	Nearby terminator (stem loop and poly T) coordinates & orientation vs. gene
dmsA 869-3226	anaerobic dimethyl sulfoxide reductase subunit A	*ycaC* no overlap	Medium	783-808 (+)	751-769 (+) 848-868 (+) 799-780 (-)	1620-1655 (+) 299-253 (-) 271-244 (-)
torS Complement (90-2804)	sensor protein TorS (regulator TorR)	*torT* no overlap	Medium	1101-1129 (-)	1103-1080 (-) 1140-1119 (-) 1019-1040 (+) 1123-1106 (+)	1532-1494 (-) 854-886 (+) 1807-1840 (+)
dcp 1-2906	dipeptidyl carboxy-peptidase II	*ydeJ* no overlap	Medium	2500-2473 (-)	2506-2528 (+) 2456-2427 (-)	2532-2567 (+) 2738-2713 (-) 2215-2181 (-)
dicB 2028-2216	inhibition of cell division	*rspB* no overlap	Medium	2147-2174 (+)	2065-2082 (+) 2212-2237 (+) 2062-2042 (-)	2055-2080 (+) 2212-2247 (+) 2341-2315 (-) 2062-2032 (-)
nrdA (*dnaF*) † 1-2286	ribonucleosie diphosphate reductase 1, alpha s.u, B1	*inaA* no overlap	Medium	1007-1036 (+)	901-917 (+) 1086-1107 (+) 1017-1035 (-)	1044-1006 (-) 878-846 (-)
malT 183-2888	positive regulator of *mal* regulon	*yhgJ* no overlap	Medium	45-73 (-) 798-829 (+)	46-13 (-) 40-55 (+) 797-817 (+)	567-595 (+) 1102-1139 (+)
bglG * Complement (4115-4951)	positive regulation of *bgl* operon	-	Strict	5012-5041 (+)	5077-5098 (-) 4974-5020 (+)	5021-4958 (-) 4964-5020 (+)
bglF * Complement (2104-3981)	PTS system β-glucosides, enzyme II	-	Strict	4065-4095 (+)	4066-4026 (-) 4122-4080 (-) 4025-4065 (+)	4076-4026 (-) 4029-4056 (+)
acs 11-1969	Acetyl-CoA synthetase	*nrfA* no overlap	Medium	914-943 (+)	933-961 (+) 964-945 (-)	768-802 (+) 626-599 (-)
yohK 9-704	putative seritonin transporter	*yohI* no overlap	Medium	249-223 (-)	122-138 (+) 214-194 (-)	122-148 (+)
arsC 47-472 yhiS 1101-1883	Arsenate reductase hypothetical protein	*trs5_11* no overlap	Medium	1081-1110 (+)	1003-1025 (+) 1254-1272 (+) 1171-1152 (-) 1068-1047 (-)	1361-1394 (+) 1171-1121 (-)

* known RAT/ter pairs
** new putative RAT/ter pairs
† new putative RAT/ter pairs (terminator precedes RAT)

Table S2. Putative RATs identified by the 'strict' and 'medium' descriptors in the genome of *E. coli* CFT073 using RNAMotif

Gene & ORF coordinates	Gene product	Gene on the other strand & overlap	RAT descriptor stringency	RAT location and orientation vs. gene	Nearby terminator (stem loop) coordinates & orientation vs. gene	Nearby terminator (stem loop and poly T) coordinates & orientation vs. gene
dmsA 100-2544	Anaerobic dimethyl sulfoxide reductase chain A precursor	*ycaC* no overlap	Medium	883-908 (+)	948-968 (+) 987-975 (-)	1720-1685 (-)
dcp 100-2211	dipeptidyl carboxypeptidas	*ydfG* no overlap	Medium	1988-1961 (-)	1994-2016 (+) 1944-1913 (-)	2226-2201 (-) 1703-1673 (-)
malT 100-2805	*malT* Regulatory protein	C4195 C4197 no overlap	Medium	898-929 (+) 145-118 (-)	897-917 (+) 142-117 (-)	667-695 (+) 1202-1239 (+)
bglG * 100-2049	*bgl* operon antiterminator	C4647 overlap	Strict	35-64 (+)	56-112 (+)	56-112 (+)
bglF * 2108-2968	PTS system, β - glucoside-specific enzyme II	C4645 overlap	Strict	981-1011 (+)	1051-1069 (+)	1011-1061 (+)
yjgD 172-588	Hypothetical protein	*argI* no overlap	Medium	45-72 (+)	122-104 (-)	254-220 (-) 203-176 (-)
uxuR 100-873	*uxu* operon transcriptional regulator	C4506 no overlap	Medium	575-600 (+)	326-350 (+) 751-790 (+) 586-560 (-)	786-814 (+) 490-448 (-)
ydhQ 100-1354	Hypothetical protein	*valV* no overlap	Medium	1123-1152 (+)	1073-1091 (+) 1122-1090 (-)	1062-1091 (+) 1230-1192 (-)
hrpA 100-4002	ATP-dependent helicase	*acpD* no overlap	Medium	2828-2802 (-)	2831-2847 (+) 2882-2853 (-)	2565-2593 (+) 3671-3701 (+) 3066-3027 (-)
C3408 *(r1)* ** 100-990	antiterminator R1	C3410 no overlap	Medium	644-669 (+)	669-696 (+) 699-729 (+)	596-634 (+) 735-775 (+)
C0959 † 100-528	Putative Regulatory protein	C0961 no overlap	Medium	407-382 (-)	393-414 (+) 395-376 (-)	395-366 (-)
C0163 † 100-207	Hypothetical protein	*panC* no overlap	Medium	281-253 (-)	245-266 (+) 358-341 (-)	233-269 (+)
C4517 † 236-634	Hypothetical protein	-	Medium	830-805 (-)	784-802 (+) 762-733 (-)	784-809 (+)
C4303 100-273	Putative conserved protein	-	Medium	81-110 (+)	55-27 (-)	176-151 (-)

* known RAT/ter pairs
** new putative RAT/ter pairs
† new putative RAT/ter pairs (terminator precedes RAT)

Table S3. Putative RATs identified by the 'strict' and 'medium' descriptors in the genome of *B. subtilis* using RNAMotif

Gene & ORF coordinates	Gene product	Gene on the other strand & overlap	RAT descriptor stringency	RAT location and orientation vs. gene	Nearby terminator (stem loop) coordinates & orientation vs. gene	Nearby terminator (stem loop and poly T) coordinates & orientation vs. gene
ptsG * 452-2551	phosphotransferase system (PTS) glucose-specific enzyme II^{ABC}	-	Strict	242-271 (+)	278-313 (+)	285-316 (+)
sacB * 464-1885	Levansucrase	*pbpE* no overlap	Strict	282-311 (+)	310-344 (+)	315-347 (+)
sacX * 1615-2994 *sacY* * 3048-3890	negative regulatory protein of SacY transcriptional antiterminator (BglG family)	*gspA* no overlap	Strict	1354-1383 (+)	1395-1435 (+)	1395-1445 (+)
bglP * 441-884	phosphotransferase system (PTS) β-glucoside-specific enzyme II	-	Strict	1436-1465 (+)	1424-1444 (+) 1466-1511 (+) 1466-1436 (-)	1466-1521 (+) 1479-1445 (-)
licT * 252-1085	transcriptional antiterminator (BglG family)	*yxiO* no overlap	Strict	1151-1180 (+)	1182-1227 (+)	1182-1237 (+)
sacT * 136-897 *sacP* * 2008-3390	transcriptional antiterminator (BglG family) enzyme II of PTS	*ywcJ* no overlap	Strict	1887-1916 (+)	1929-1962 (+)	1929-1972 (+)
thiA ** 178-1950	biosynthesis of the pyrimidine moiety of thiamin	*ygaK* no overlap	Medium	1646-1675 (+)	1658-1673 (+) 1686-1717 (+)	1686-1727 (+)
yoeA Complement (65-1456) *yoeB* 1732-3255	involved in proline biosynthesis	No overlap between the two	Medium	1896-1923 (+)	1896-1918 (+) 1925-1956 (+)	1721-1759 (+) 1857-1818 (-)
cydD † 38-1787	ABC membrane transporter (ATP-binding protein)	*yxlA* no overlap	Medium	923-951 (+)	900-923 (+)	884-921 (+) 1036-1080 (+) 1035-1031 (-)
trpB 4527-5729	tryptophan synthase	-	Medium	5186-5216 (+)	5202-5228 (+) 5243-5261 (+)	5076-5105 (+) 5243-5271 (+)
ywhE ** 55-1994	similar to penicillin-binding protein	*ywhF* no overlap	Medium	1589-1619 (+)	1583-1611 (+) 1616-1634 (+)	1616-1654 (+) 1607-1568 (-)
qoxA *ywzA* 334-537	cytochrome aa3 quinol oxidase unknown	*ywzA* no overlap	Medium	57-85 (+)	64-104 (+)	64-114 (+)
ytmJ 73-882	similar to amino acid ABC transporter (binding protein)	*ytlI* no overlap	Medium	880-853 (-)	789-761 (-) 918-895 (-)	985-934 (-) 789-751 (-)
yuxH 1-324	unknown	*yuzC* no overlap	Medium	138-111 (-)	118-80 (-) 233-209 (-)	233-199 (-)
yrpC Complement (15326-16123)	unknown	*yrpE* no overlap	Medium	15576-15541 (-) 17562-17533 (-) 15540-15566 (+)	15430-15408 (+)	15630-15596 (-) 17868-17828 (-) 15227-15256 (+)
yonR Complement (72-398) *yonP* 1371-1565	similar to transcriptional regulator unknown	No overlap between the two	Medium	741-716 (-)	681-651 (-) 854-807 (-)	681-645 (-) 699-740 (+)

* known RAT/ter pairs
** new putative RAT/ter pairs
† new putative RAT/ter pairs (terminator precedes RAT)

	Strict	Medium	Loose
Number of information bits	33-29	26-30	15-19
Probability for finding a random hit	$1/2^{33}$-$1/2^{29}$	$1/2^{26}$-$1/2^{30}$	$1/2^{15}$-$1/2^{19}$
Number of times found in *E. coli* K12	2	11	1651
p-value for *E. coli* K12	$2.30*10^{-11}$ - $1.07*10^{-7}$	$<1*10^{-12}$	1
Number of times found in *E. coli* CFT073	2	16	1980
p-value for *E. coli* CFT073	$4.22*10^{-11}$ – $1.53*10^{-7}$	$5.91*10^{-10}$ – $3.63*10^{-11}$	1
Number of times found in *B. subtilis*	6	17	1528
p-value for *B. subtilis*	$<1*10^{-12}$	$<1*10^{-12}$	1

Table S4. Significance of RAT predictions.
p-values for the predictions, obtained with the three descriptors, were calculated based on the rules of Shannon's information theory (see Material and Methods).

	N - Number of genes in genome	**n -** Number of hits found with the strict & medium descriptor	**M –** Number of genes that reside in the cluster	**k –** Hits within a cluster	**p-value**
***E. coli* K12**	4287	11	*dmsA-torT:* 100 *ydfG-dicB:* 36 *yohI- nrdA:* 94 *malT-yhiS:* 87 *bglG-bglF:* 2	2 2 2 2 2	0.02583263 0.00359615 0.02546886 0.01988188 0.00000599
***E. coli* CFT073**	5377	16	*c0951-dmsA:* 80 *acpD-ydfG-ydhQ:* 214 *c4195-malT-c4303:* 107 *c4520-bglG-bglF:* 125 *yjgD-uxuR:* 91	2 3 3 3 2	0.02024803 0.01984567 0.00292869 0.00454863 0.02578012
B. subtilis	4096	17	*ywhE-sacP-ywzA-sacX-cydD-bglS-bglP:* 156	7	0.00000087

Table S5. RAT clustering significance.
p-values were calculated by applying hypergeometric distribution, as described in Material and Methods.

	E. coli K12	*E. coli* CFT073	*B. subtilis*
Lesnik *et. al.* descriptor	3075	3745	3005
TransTerm	4257	4300	4584
Simple descriptor, no threshold	3898	4789	3849
Simple descriptor, $\Delta G < 0$ threshold	3757	4603	3744
Simple descriptor, $\Delta G < -5$ threshold	2981	3661	3135

Table S6. Rho-independent terminators .
The number of Rho-independent terminators in the three bacterial genomes was predicted by the three different algorithms described in the text.