Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 372 (2006) 171–181
GENE
SECTION
FUNCTIONAL GENOMICS
www.elsevier.com/locate/gene

Prediction of non-coding and antisense RNA genes in *Escherichia coli* with Gapped Markov Model

Nozomu Yachie [a,b], Koji Numata [a,b], Rintaro Saito [a,c,*], Akio Kanai [a,c], Masaru Tomita [a,b,c]

[a] *Institute for Advanced Biosciences, Keio University, Tsuruoka, 997-0035, Japan*
[b] *Bioinformatics Program, Graduate School of Media and Governance, Keio University, Fujisawa, 252-8520, Japan*
[c] *Department of Environmental Information, Keio University, Fujisawa, 252-8520, Japan*

Received 19 October 2005; received in revised form 2 December 2005; accepted 28 December 2005
Available online 24 March 2006
Received by T. Gojobori

Abstract

A new mathematical index was developed to identify and characterize non-coding RNA (ncRNA) genes encoded within the *Escherichia coli* (*E. coli*) genome. It was designated the GMMI (Gapped Markov Model Index) and used to evaluate sequence patterns located at the separate positions of consensus sequences, codon biases and/or possible RNA structures on the basis of the Markov model. The GMMI was able to separate a set of known mRNA sequences from a mixture of ncRNAs including tRNAs and rRNAs. Consequently, the GMMI was employed to predict novel ncRNA candidates. At the beginning, possible transcription units were extracted from the *E. coli* genome using consensus sequences for the sigma70 promoter and the rho-independent terminator. Then, these units were evaluated by using the GMMI. This identified 133 candidate ncRNAs, which contain 29 previously annotated small RNA genes and 46 possible antisense ncRNAs. Furthermore 12 transcripts (including five antisense RNAs) were confirmed according to the expression analysis. These data suggests that the expression of small antisense RNAs might be more common than previously thought in the *E. coli* genome.

Keywords: Bioinformatics; Markov model; Small RNA (sRNA); Sigma70 promoter; Rho-independent terminator

1. Introduction

Genome sequencing projects have identified numerous examples of non-coding RNA (ncRNA) genes (Eddy, 2001). ncRNA molecules do not encode protein products, but are thought to have structural, regulatory or catalytic properties. Along with well-known RNA molecules, such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), many classes of ncRNAs have been identified by the experimental and bioinformatics approaches (Eddy, 2001). Among bacteria, particularly *Escherichia coli* (*E. coli*), many studies of small RNAs (sRNAs) have focused on the intergenic regions (Wassarman et al., 1999; Argaman et al., 2001; Carter et al., 2001; Rivas et al., 2001; Wassarman et al., 2001; Chen et al., 2002; Tjaden et al., 2002). Recent articles have estimated that approximately 1000 sRNA candidates of unknown function are encoded in the *E. coli* genome (Hershberg et al., 2003; Vogel et al., 2003). However, only 62 of these have been detected experimentally through analysis of RNA expression in *E. coli*. Vogel et al. performed global sRNA identification using the shotgun cloning approach known as experimental RNomics. They suggested that 5% of the final contigs mapped to the strand complementary to coding regions within the *E. coli* genome (Vogel et al., 2003). Although several small antisense RNAs (saRNAs) are known to be transcribed from loci which differ from their target mRNAs in *E. coli* (*trans*-antisense) (Carpousis, 2003), few specific examples transcribed in the opposite direction from the same

Abbreviations: ncRNA, non-coding RNA; sRNA, small RNA; saRNA, small antisense RNA; tRNA, transfer RNA; *E. coli*, *Escherichia coli.*

* Corresponding author. Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan. Tel.: +81 235 29 0522; fax: +81 235 29 0525.
E-mail address: rsaito@sfc.keio.ac.jp (R. Saito).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.12.034

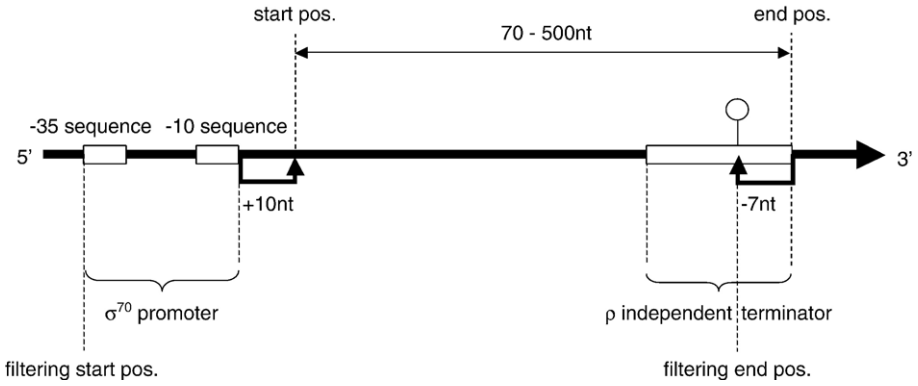

Fig. 1. Transcription-unit model. The transcription-unit model is defined as the region between the sigma70 promoter motif +10 ('start pos.') and the end of the rho-independent-terminator motif ('end pos.'), which ranges in size from 70 to 500 nt. The 'filtering start pos.', which is the original start position of the promoter motif, and the 'filtering end pos.', which is 7 nt upstream from the original end of the terminator motif, are defined after filtering out those that overlap with ORF regions on the same DNA strand.

genomic loci (*cis*-encoded saRNAs) have been documented (Kawano et al., 2002; Vogel et al., 2003). Moreover, all previous sRNA predictions based on the bioinformatics approach have focused only on the intergenic regions as defined by gene annotation. A bioinformatics based approach for the prediction of saRNAs has yet to be reported.

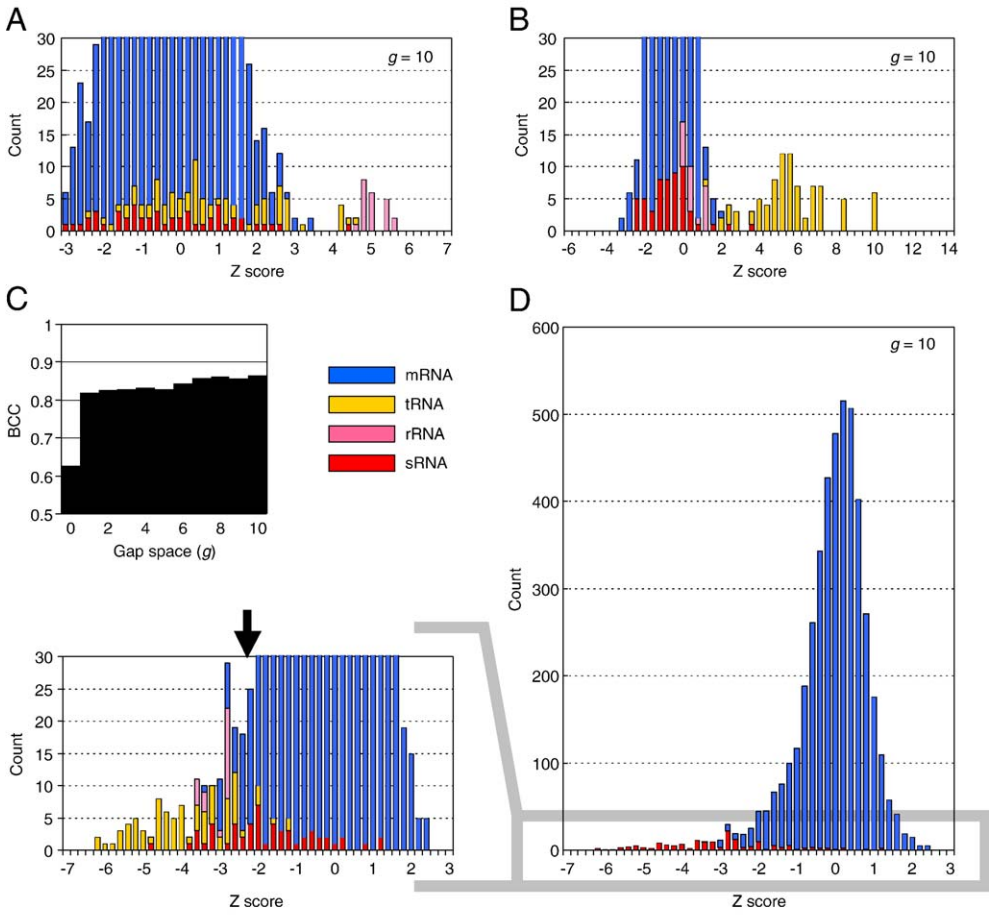

Fig. 2. Isolation of each RNA category using the GMMI. The histograms represent the distributions of the *Z* score of GMMI for each RNA category (A, B and D). The Markov-order (*e*) was 2 and the gap space parameter (*g*) was 10. (A) The GMMI is trained using tRNA sequences. (B) The GMMI is trained using rRNA sequences. (C) The best correlation coefficients (BCC) for the quality measure of every gap space parameter (*g*) where GMMI is trained using mRNAs. (D) Separation of the ncRNAs from the mRNA category. The GMMI is trained using mRNA sequences. An enlargement of the histogram is presented on the left. The arrow indicates the threshold value (−2.34) that isolates ncRNAs with over 80% sensitivity and selectivity.

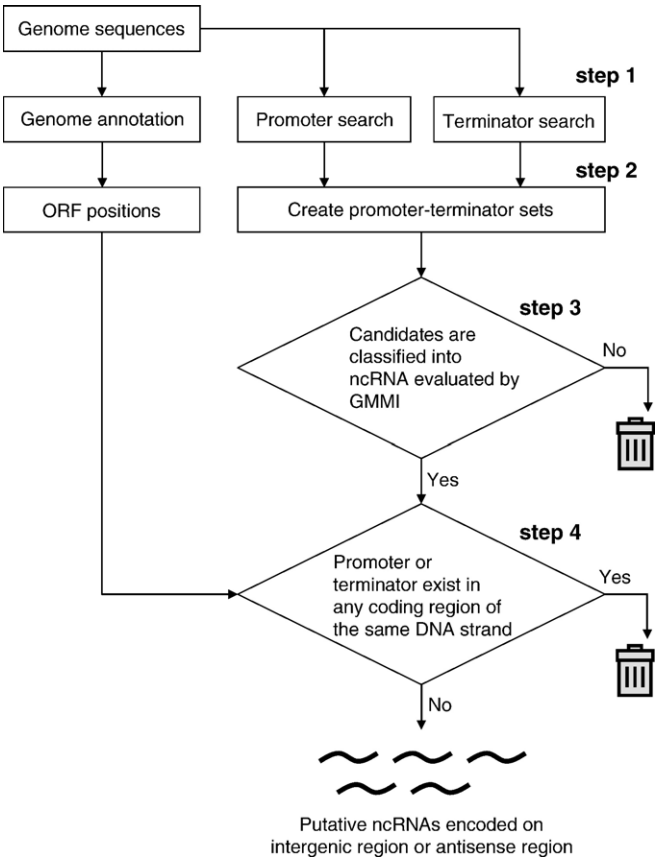

Fig. 3. Schematic representation of the computational procedure used to extract putative ncRNAs.

We have developed a new approach for the detection of small ncRNA which we have designated the Gapped Markov Model Index (GMMI). This allows the detection of ncRNAs not only in intergenic regions, but also in the antisense strands of coding regions. The GMMI evaluates whether an RNA sequence is coding or non-coding on the basis of the Markov model. The successful identification of previously known ncRNA sequences from the total RNA sequences was used to validate the GMMI. Subsequently, the GMMI allowed the identification of novel ncRNA candidates within predicted transcription units, and some of these molecules were encoded as antisense RNAs. We detected the expression of some ncRNA candidates, including saRNAs using RT-PCR and Northern hybridization.

2. Materials and methods

2.1. Gapped Markov Model Index (GMMI)

A novel approach designated the Gapped Markov Model Index (GMMI) was developed to mathematically distinguish ncRNA sequences from other mRNA sequences. The aim was to evaluate the specificity of a RNA sequence ('test sequence') by the certain set of RNAs ('training set') with the combinations of specific nucleotide sequence patterns located at separate positions of consensus sequences, codon biases and/or its possible structural motifs, based on the Markov model. The biochemical functions of ncRNAs such as tRNA and rRNA molecules are thought to depend on not only their sequence motifs but also their higher order structures. As nucleotides at separate positions can interact with one another, RNA sequence, structure and spatial relationships, such as hydrogen bonding should be considered. Let R be the RNA sequence to be evaluated ('test sequence'), which is presented as a string of n characters: $R=r_1r_2\ldots r_k\ldots r_n$ ($r_k \subset$ (A, C, G, U), $1 \le k \le n$). The segment from r_i to r_j in the sequence can be presented as: $R_{i,j}=r_ir_{i+1}\ldots r_j$ ($1 \le i \le j \le n$). The GMMI score (G) with the constant numbers of the Markov-order (e) and the gap space parameter (g) of R is defined as the product of the conditional probability of observing each nucleotide $P(r_k|R_{k-f-e,k-f-1})$, given a sequence pattern composed of e nt located in f nt upstream ($0 \le f \le g$) divided by the frequency of each nucleotide $P(r_k)$ in the 'training set':

$$G = \left\{ \prod_{k,f} \frac{P(r_k|R_{k-f-e,k-f-1})}{P(r_k)} \right\}^{\frac{1}{n}}$$

The score G evaluates the likelihood of the given sequence being related to our Gapped Markov Model which is constructed using the 'training set'. A score below 1 indicates that a given sequence has a specific pattern that is not observed in the training set, whereas a score above 1 indicates concordance (for more details, see Supplementary materials).

2.2. Sequence data

The complete *E. coli* genome sequence and a set of mRNAs, documented ncRNAs and sRNAs, were downloaded from GenBank via the National Center for Biotechnology Information (NCBI) ftp server. In addition, details of 62 previously identified sRNAs of *E. coli* K12MG1655 were obtained from the literature (Hershberg et al., 2003; Vogel et al., 2003). Open reading frames (ORFs) were categorized as mRNAs. The RNA list we constructed consisted of 4311 mRNAs (ORFs) and 163 ncRNAs (86 tRNAs, 22 and 55 sRNAs). Recently, researchers have compiled a set of 62 sRNA genes in *E. coli* and we adopted 55 of these, excluding those from transcription units that were neither predicted nor identified experimentally (Hershberg et al., 2003; Vogel et al., 2003).

2.3. Prediction of putative transcription units

At the genome level within *E. coli* K12MG1655 we initially predicted transcription units between the sigma70 promoters and rho-independent terminators. The prediction of putative transcription units was performed using a recently reported method (Chen et al., 2002). The program *pftools*2.2, obtained from the ftp server of the Swiss Bioinformatics Institute, was used to identify the putative sigma70 promoters; a profile describing the major class of *E. coli* sigma70 promoters was included in the software. We used the reported threshold of 50, which is capable of identifying approximately 70% of

Table 1
Intergenic ncRNA candidates

Name	Minutes	Strand	Left boundary	Right boundary	Length	Adjacent genes
NC004	4.89	−	227,062	227,343	282	*yaeD*+/*yafB*+
NC005	4.93	+	228,966	229,048–229,058	83–93	*yaeD*+/*yafB*+
NC007	5.1	+	236,915	237,080	166	*dnaQ*+/*yafT*+
NC009	6.82	+	316,420–316,554	316,626	73–207	*ykgA*−/*ykgB*−
NC010	7.12	−	330,773	330,958	186	*betT*+/*yahA*+
NC011	9.29	+	431,272	431,448	177	*tsx*−/*yajI*−
NC012	9.92	−	460,539	460,641	103	*lon*+/*hupB*+
NC013	12.01	+	557,197–557,214	557,328	115–132	*folD*−/*sfmA*+
NC014	12.28	−	569,952	570,038–570,045	87–94	*ybcK*+/*ybcL*+
NC015	12.61	+	584,926–585,013	585,219	207–294	*ompT*−/*envY*−
NC016	16.21	+	752,115–752,189	752,395	207–281	*ybgD*−/*gltA*−
NC019	16.8	+	779,739	780,109	371	*ybgF*+/*nadA*+
NC020	16.83	−	780,911–781,067	781,243	177–333	*ybgF*+/*nadA*+
NC021	18.42	+	854,993	855,137	145	*ybiS*−/*ybiT*+
NC022	19.39	+	899,841	899,956–900,055	116–215	*artJ*−/*artM*−
NC023	19.94	−	925,104–925,114	925,225–925,337	112–234	*clpA*+/*infA*−
NC025	23.64	−	1,096,423–1,096,795	1,096,897–1,096,937	103–515	*ycdU*+/*ycdW*+
NC027	25.77	−	1,195,651	1,195,827–1,196,071	177–421	*icdA*+/*ymfD*−
NC030	26.09	+	1,210,552	1,210,856–1,210,864	305–313	*mcrA*+/*ycgW*−
NC032	26.21	−	1,216,280	1,216,367–1,216,383	88–104	*ymgC*+/b1168+
NC034	26.41	−	1,225,360	1,225,440	81	*minC*−/*ycgJ*+
NC035	27.06	+	1,255,698	1,255,895	198	b1202−/*ychF*−
NC037	31.24	+	1,449,530	1,449,620	91	*tynA*−/*maoC*−
NC038	31.53	+	1,463,146	1,463,221–1,463,398	76–253	b1400+/yi222−
NC039	31.53	−	1,463,154	1,463,301–1,463,317	148–164	b1400+/yi222−
NC040	31.54	+	1,463,146–1,463,296	1,463,398	103–253	b1400+/yi222−
NC041	32.1	−	1,489,517	1,489,624	108	*cybB*+/*ydcA*+
NC044	33.74	+	1,565,327–1,565,393	1,565,515	123–189	b1490−/b1491−
NC046	34.4	+	1,596,200	1,596,325	126	*ydeK*−/*ydeV*−
NC048	35.27	−	1,636,282–1,636,293	1,636,395	103–114	b1551−/*cspI*−
NC050	35.97	−	1,669,100	1,669,214	115	*ynfM*+/*asr*+
NC051	37.49	+	1,739,231–1,739,268	1,739,378	111–148	*ydhC*+/*cfa*+
NC052	37.99	−	1,762,744	1,762,848	105	*ydiC*−/b1685−
NC053	38.88	−	1,804,162	1,804,289	128	b1722−/*pfkB*+
NC056	41.02	+	1,903,333–1,903,411	1,903,517	107–185	b1820+/b1821+
NC059	43.97	+	2,040,100	2,040,182	83	b1973+/b1974+
NC060	44.02	+	2,042,565	2,042,774	210	b1976+/b1978+
NC061	44.31	−	2,055,997	2,056,130	134	b1983+/*yeeO*−
NC063	44.4	+	2,060,082–2,060,276	2,060,398–2,060,405	123–324	*nac*−/*erfK*−
NC064	44.91	+	2,083,596	2,083,717	122	*yeeE*−/*yeeF*−
NC065	46.36	+	2,151,204	2,151,377	174	b2073−/b2074+
NC066	46.69	+	2,166,152–2,166,429	2,166,509	81–358	b2085−/b2086+
NC068	49.23	+	2,284,229	2,284,353	125	*yejM*+/*yejO*−
NC069	53.04	−	2,460,719	2,460,827–2,461,092	109–374	*fadL*+/b2345+
NC072	53.2	−	2,468,523–2,468,526	2,468,598	73–76	b2352+/b2353+
NC073	53.49	−	2,481,604	2,481,743–2,481,798	140–195	*emrK*−/*evgA*+
NC076	53.7	+	2,491,623–2,491,661	2,491,746	86–124	b2374−/b2375−
NC077	54.23	−	2,516,049–2,516,062	2,516,271	210–223	b2395−/*yfeC*+
NC078	54.29	+	2,518,865–2,518,946	2,519,316–2,519,385	371–521	*gltX*−/*xapR*−
NC079	55.8	−	2,588,772–2,588,779	2,589,005–2,589,064	227–293	*acrD*+/*yffB*+
NC082	57.96	−	2,689,214	2,689,355	142	*yfhK*−/*purL*−
NC083	58.15	+	2,697,660–2,698,056	2,698,138	83–479	*yfhL*+/*acpS*−
NC084	59.35	+	2,753,612	2,754,061	450	*smpB*+/*intA*+
NC087	60.28	−	2,796,832	2,796,960	129	*stpA*−/b2670+
NC089	63.48	+	2,945,350–2,945,408	2,945,740	333–391	*mltA*−/b2817−
NC093	66.79	+	3,098,850	3,098,919	70	*ansB*−/*yggN*−
NC094	67	+	3,108,325–3,108,385	3,108,501–3,108,591	117–267	*yqgA*+/*yghD*−
NC095	69.26	+	3,213,241	3,213,359	119	*ygjF*−/*yqjH*−
NC096	71.47	−	3,315,695–3,315,746	3,315,933–3,316,010	188–316	b3170−/*argG*+
NC097	72.39	−	3,358,426	3,358,513–3,358,567	88–142	*gltD*+/*gltF*+
NC098	73.45	−	3,407,693	3,407,767–3,408,031	75–339	*prmA*+/*yhdG*+
NC099	73.74	−	3,420,887–3,421,065	3,421,330	266–444	*yhdZ*+/b3279+
NC101	74.74	+	3,467,665–3,467,686	3,467,764	79–100	*yheB*−/*tufA*−
NC102	78.79	−	3,655,593	3,655,672–3,655,904	80–312	*hdeD*+/*yhiE*+

Table 1 (*continued*)

Name	Minutes	Strand	Left boundary	Right boundary	Length	Adjacent genes
NC103	78.79	−	3,655,593–3,655,654	3,655,725–3,655,904	72–312	*hdeD*+/*yhiE*+
NC106	79.15	−	3,672,054	3,672,228	175	*yhjD*+/*yhjE*+
NC107	79.88	−	3,706,245	3,706,321	77	*dppA*−/*yhjW*−
NC111	81.19	−	3,766,712	3,766,784–3,766,866	73–155	*yibG*+/b4549+
NC114	84.98	−	3,942,630	3,942,911	282	*yiePs*−/*yifDA*−
NC115	85.78	+	3,979,979	3,980,460	482	*yifK*+/*aslB*+
NC116	85.78	−	3,979,984	3,980,295	312	*yifK*+/*aslB*+
NC117	85.87	+	3,983,895–3,983,935	3,984,256	322–362	*aslA*−/*hemY*−
NC119	87	−	4,036,401	4,036,682	282	*hemG*+/*mobB*−
NC120	87.04	+	4,038,191	4,038,264–4,038,274	74–84	*hemG*+/*mobB*−
NC121	89.79	−	4,165,949	4,166,028–4,166,158	80–210	*murI*+/*murB*+
NC122	89.83	−	4,167,523	4,167,804	282	*murI*+/*murB*+
NC123	89.87	+	4,169,310	4,169,383–4,169,393	74–84	*murI*+/*murB*+
NC124	90.72	−	4,208,925	4,209,206	282	*purH*−/*yjaA*+
NC125	90.76	+	4,210,713	4,210,784	72	*purH*−/*yjaA*+
NC126	91.8	−	4,258,959	4,259,184	226	*yjbM*+/*yjbN*+
NC127	91.8	+	4,258,840–4,259,054	4,259,130	77–291	*yjbM*+/*yjbN*+
NC128	93.98	−	4,360,127	4,360,202	76	*cadC*−/*yjdC*−
NC129	94.07	+	4,364,372	4,364,454	83	*dcuA*−/b4139−
NC131	95.6	−	4,435,129	4,435,243–4,435,325	115–197	*cysQ*+/*ytfI*+
NC132	96.86	+	4,493,956	4,494,099	144	b4269−/*intB*+
NC133	98.16	+	4,553,924–4,553,945	4,554,181	237–258	b4325−/b4326+

Intergenic ncRNA candidates are listed along with their position, strand, length and the names of the adjacent genes. Candidates located on the plus strand are indicated by '+' and candidates located on the minus strand are indicated by '−' in the *E. coli* genome. The left boundary denotes the transcription start position on the '+' strand or the transcription stop position on the '−' strand. The right boundary denotes the transcription start position on the '−' strand or the transcription stop position on the '+' strand. Where multiple positions were detected, the range is indicated. The length is equal to the absolute value of the 'left boundary' minus the 'right boundary'. The intergenic region is defined by the both sides of 'adjacent genes'. For example, NC013 is encoded in the intergenic region delineated by the *folD* gene on the left-hand side, which is transcribed from the '−' strand, and the *sfmA* gene on the right-hand side, which is transcribed from the '+' strand.

previously annotated promoters (Chen et al., 2002). In order to identify putative rho-independent terminators, we predicted their sequences and structural motifs using the RNAMotif software (Macke et al., 2001). The structural specifications of the motif model for the program and the optimal cut-off score were obtained from previous work (Lesnik et al., 2001).

2.4. Identification of putative ncRNA genes using the GMMI

We used 4311 ORF sequences as the training set for our model. The GMMI score was then calculated for every putative transcription unit. The size of a transcription unit was defined as the distance between the transcription start position, which is 10 nt downstream of the 3′ end of the sigma70 promoter (Fig. 1, 'start pos.'), and the transcription stop position, which is at the 3′ end of the rho-independent terminator ('end pos.'). We limited the length of transcription units according to the sizes of the previously identified sRNAs at between 70 and 500 nucleotides. We extracted those having a Z score of the GMMI ($e=2$ and $g=10$) below the threshold −2.34 (see Results section). We then discarded candidates in which the 5′ ends of the sigma70 promoters ('filtering start pos.'), or position −7 with respect to the 3′ ends of the rho-independent terminators ('filtering end pos.'), overlapped any ORF regions on the same DNA strand. This computational schema and the GMMI program is implemented by Perl version 5 and runs under the G-language GAE which provides a generic analysis workbench for bioinformatics (Arakawa et al., 2003). The source codes are available upon request.

2.5. Bacterial culture, RNA isolation

E. coli strain K12 MG1655 was used in this study. Cells were grown anaerobically at 37 °C in L broth (10 g tryptone, 5 g yeast extract and 0.5 g NaCl/l). After an overnight incubation, 10 ml of the bacterial culture (stationary phase) was transferred to 100 ml of fresh medium and incubated to an approximate optical density of 600 nm. Cells were harvested in two different growth phases; late log ($0.7<O.D._{600}<1.0$) and early stationary phase ($1.0<O.D._{600}<2.0$). Total RNA for RT-PCR was isolated using the TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. For Northern blot hybridization, *E. coli* pellets were initially treated with the RNAprotect Bacteria Reagent (Qiagen GmbH, Hilden, Germany) for stabilizing RNA and then total RNA was isolated essentially using the RNeasy Midi Kit (Qiagen), except that phenol-chloroform extraction of RNA was performed for purifying RNA instead of using the RNeasy Midi column.

2.6. RT-PCR

RT-PCR for the analysis candidate ncRNA expression was carried out using the ReverTra Dash Kit (TOYOBO Biochemicals, Japan). PCR was performed for 25 cycles at 98 °C (10 s), 55 °C (2 s) and 74 °C (30 s). The PCR products were separated

Table 2
Antisense ncRNA candidates

Name	Minutes	Strand	Left boundary	Right boundary	Length	Sense genes
NC001	0.18	−	8471	8575	105	*talB*+ (M)
NC002	3.13	−	145,382	145,470	89	*yadE*+ (M)
NC003	4.13	−	191,751	191,935	185	*pyrH*+ (N)
NC006	5.03	+	233,493	233,843	351	*dniR*− (M)
NC008	6.31	+	293,013–293,138	293,508	371–496	*yagL*− (N), *yagM*− (C)
NC017	16.33	−	757,821	758,150	330	b0725+ (C), *sucA*+ (N)
NC018	16.73	−	776,452–776,456	776,537	82–86	*tolA*+ (M)
NC024	20.56	+	953,999–954,005	954,097–954,104	93–106	*ycaO*− (C)
NC026	25.21	+	1,169,665	1,169,745	81	*mfd*− (C)
NC028	25.8	+	1,197,016–1,197,275	1,197,437	163–422	*ymfE*− (M)
NC029	25.83	−	1,198,645	1,198,751	107	*lit*+ (M)
NC031	26.16	+	1,213,898	1,214,112	215	b1163− (M)
NC033	26.29	−	1,219,712	1,219,823–1,220,056	112–345	b1169+ (M)
NC036	30.88	+	1,432,950	1,433,214	265	*ynaF*− (C)
NC042	32.42	−	1,504,268	1,504,343	76	*ydcN*+ (M)
NC043	33.41	+	1,550,245	1,550,428	184	*yddM*− (C)
NC045	34.23	+	1,588,443	1,588,652	210	b1506− (N)
NC047	35.23	−	1,634,837	1,635,085–1,635,206	249–370	*ydfO*+ (N)
NC049	35.47	−	1,645,927	1,646,313	387	*dicA*+ (N)
NC054	40.21	−	1,865,821	1,866,092	272	*yeaG*+ (M)
NC055	40.87	−	1,896,373	1,896,538	166	b1815+ (N)
NC057	41.9	−	1,944,079	1,944,183–1,944,219	105–141	*yebB*+ (N)
NC058	43	+	1,994,949–1,994,956	1,995,086	131–138	*yecC*− (C)
NC062	44.35	+	2,057,869	2,057,991	123	*cbl*− (C)
NC067	47.62	−	2,209,374–2,209,375	2,209,572	198–199	*yehR*+ (M)
NC070	53.12	−	2,464,731	2,464,921	191	*intC*+ (M)
NC071	53.15	−	2,465,791	2,466,224	434	b2350+ (N)
NC074	53.54	−	2,483,994	2,484,232–2,484,410	239–417	*evgS*+ (M)
NC075	53.6	+	2,487,019	2,487,246	228	*yfdE*− (N)
NC080	57.09	+	2,648,981	2,649,059	79	b2520− (M)
NC081	57.7	+	2,677,127	2,677,477	351	*yphF*− (N)
NC085	59.47	+	2,759,270	2,759,665	396	*yfjK*− (C)
NC086	59.6	−	2,765,252	2,765,409	158	*yfjO*+ (C)
NC088	60.54	−	2,808,664	2,808,743	80	*ygaH*+ (C)
NC090	64.49	−	2,991,938	2,992,101–2,992,111	164–174	b2854+ (N)
NC091	65.14	−	3,022,243	3,022,327	85	*ygfO*+ (N)
NC092	66.16	+	3,069,309	3,069,486	178	*pgk*− (C)
NC100	73.99	−	3,432,775	3,432,935	161	*fmt*+ (C), *sun*+ (N)
NC104	78.8	−	3,656,170	3,656,242–3,656,368	73–199	*yhiE*+ (M)
NC105	78.96	+	3,663,407	3,663,802	396	*yhiX*− (N)
NC108	80.17	−	3,719,671	3,719,812–3,719,834	142–164	t150+ (C)
NC109	80.5	+	3,734,659–3,734,695	3,735,077	383–419	*bax*− (N)
NC110	81.16	−	3,765,221	3,765,463–3,765,491	243–271	*yibJ*+ (M)
NC112	81.88	+	3,798,670–3,798,700	3,798,769	70–100	*rfaJ*− (M)
NC113	81.96	+	3,802,677–3,802,746	3,803,001	256–325	*rfaS*− (M), *rfaP*− (C)
NC118	86.23	+	4,000,752	4,000,888	137	*yigG*− (N)
NC130	95	−	4,407,633	4,407,737–4,407,794	105–162	*yjfI*+ (N)

Antisense ncRNA candidates are listed along with their position, strand, length (for details, see the footnote to Table 1), and the names of the genes on the sense strand. (N), (C) and (M) indicate the amino-terminal, carboxy-terminal and middle section of a sense gene, respectively. For example, NC113 is encoded on two antisense regions, the middle of *rfaS* and the carboxy-terminal of *rfaP*.

using 3% NuSieve 3 : 1 agarose gel electrophoresis (Cambrex Bio Science Rockland, CA, USA) and stained using ethidium bromide (EtBr).

2.7. Northern blot hybridization

Total RNA (20 g per lane) was separated by electrophoresis in 6% polyacrylamide gel containing 8 M Urea and transferred to positively charged nylon membrane Hybond+ (Amersham Biosciences, Piscataway, NJ, USA) by electroblotting. The specific oligonucleotide probes (45 mer) for each gene were labelled with BrightStar Psoralen-Biotin, nonisotopic labelling kit (Ambion, Austin, TX, USA). Membranes were hybridized with these probes using a NorthernMax Hybridization kit (Ambion). The washing temperature was 42 °C for genes NC051 and NC065, and 45 °C for genes NC086 and NC092, respectively. Detection procedures followed the instructions of BrightStar BioDetect (Ambion), a nonisotopic detection kit. Images were

Table 3
Documented sRNAs in our candidates

Name	Promoter score	Terminator score	*Z* score	Annotations
NC005	52.65	−9.82	−3.63	*aspU*
NC007	65.1	−10.82	−2.46	*aspV*
NC019	58.58	−4.1	−3	*lysT*, *valT*, *lysW*
NC023	74	−16.88	−4.05	*serW*
NC025	77.56	−15.35	−3.21	*serX*
NC043	51.46	−8.03	−2.53	C0362
NC052	50.27	−13.26	−2.38	*rydB*/tpe7/IS082
NC060	74	−17.99	−3.1	*asnT*
NC061	72.81	−11.82	−3.96	*asnW*
NC062	64.51	−5.82	−2.44	*asnU*
NC063	62.14	−8.78	−3.49	*asnV*
NC068	62.14	−5.33	−2.36	*proL*
NC077	72.81	−10.88	−3.01	*alaX*, *alaW*
NC078	69.25	−5.18	−2.96	*valU*, *valX*, *valY*, *lysV*
NC082	53.83	−11.95	−2.73	tke1/*sroF*
NC084	59.76	−19.94	−2.95	*ssrA*
NC089	59.76	−6.36	−2.85	*metZ*, *metW*, *metV*
NC094	66.29	−12.41	−3.61	*pheV*
NC095	58.58	−16.31	−2.6	*ileX*
NC096	69.85	−7.47	−3.46	*metY*
NC099	52.65	−18.42	−3.29	*thrV*, *rrfF*, *rrfD*
NC107	69.25	−9.58	−2.43	*proK*
NC115	60.95	−12.56	−3.19	*argX*, *hisR*, *leuT*, *proM*
NC117	54.43	−14.87	−2.53	*sraJ*/*ryiA*/k19
NC120	52.65	−20.21	−3.21	*rrfA*
NC123	52.65	−20.21	−3.13	*rrfB*
NC125	52.65	−11.52	−3.11	*rrfE*
NC128	65.1	−12.98	−3.93	*pheU*
NC132	55.02	−6.71	−2.68	*leuX*

Previously reported sRNAs, according to GenBank annotations and Hershberg et al. (9), are listed along with their promoter score, terminator score and *Z* score of GMMI, together with the *E. coli* genome annotations. The promoter score represents the scores from *pftool*2.2 and the terminator score represents the scores from RNAMotif (for further details, see the Materials and methods section). Five sRNAs (NC043, NC052, NC082, NC084 and NC117) are documented in Hershberg et al. and the others are tRNAs and rRNAs documented in GenBank annotations. Separations made using a comma indicate correspondence with multiple sRNAs, whereas separations made using a slash (or solidus) indicate synonymous sRNAs.

visualized and analyzed with a Molecular Imager FX Pro (Bio-Rad Laboratories, Hercules, CA, USA).

2.8. Oligonucleotides

The oligonucleotides for RT-PCR analysis and Northern analysis were designated using Primer3, which is implemented on the *E. coli* genome database *coli*Base (http://colibase.bham.ac.uk). The list of oligonucleotides is provided as Supplementary material.

3. Results

3.1. Analysis of previously annotated ncRNA sequences using the GMMI

3.1.1. Test the GMMI using tRNAs and rRNAs

The GMMI performance test was provided using a set of known ncRNA sequences referred to as the 'training set', together with an independent set of RNAs designated the 'test set'. We could then investigate whether the GMMI correctly identified ncRNAs that were absent from the training set. Two sets of well-known ncRNAs, tRNAs and rRNAs (Fig. 2A and B) were used to train our model. Due to the limited number of sequences, the following procedure was adopted. Initially, one sequence was removed from the set and the remainders were used to train the model. The excluded sequence was then used to test the model. The procedure was repeated to evaluate each sequence in the set. The Markov-order (e) of 2 and the gap space parameter (g) of 10 were used. Accordingly, the majority of the sequences in the 'training set' category had a GMMI score above 1, whereas those in the other categories had GMMI scores below 1 (data not shown). The average *Z* score for tRNA was 5.88 and the average *Z* score for the sequences in other categories was −0.12 when the tRNA category was used as the training set. In contrast, when the rRNA category was employed as the training set, the average *Z* scores for rRNA and for the other categories were 5.17 and −0.03, respectively. Thus we suggest that GMMI was able to model sequence patterns specific to tRNAs and rRNAs, and successfully separated tRNAs and rRNAs from other RNA molecules based on GMMI scores.

3.1.2. Parameter optimization for ncRNAs isolation by training mRNAs

In order to predict all of the ncRNA families, we chose mRNAs, the negative controls of ncRNA, as the 'training set' in the GMMI calculations. Because very small number of documented sRNAs (small RNAs) might not have any common features for the efficient training and might contain possible tiny-peptide-coding RNAs (4), we avoided training such class of RNAs. Intergenic regions were also avoided, because there may be un-annotated coding regions.

We needed to define the threshold, which would allow the GMMI to identify ncRNA sequences efficiently. The randomly selected 10% of 4311 mRNA sequences were used as the 'training set'. The remaining 90% of the mRNA sequences, in addition to the tRNA, rRNA and sRNA (ncRNA) sequences were used as the 'test set'. To investigate the efficiency of our Gapped Markov Model and to estimate the optimal gap space parameter (g), we measured how known ncRNAs could be isolated from mRNAs depending on the parameter of g by the best correlation coefficient (BCC), which is defined as max CC (t), where CC(t) is correlation coefficient (Burset and Guigo, 1996) calculated according to the numbers of modeled/non-modeled RNAs having GMMI score above/below the threshold t (Fig. 2C). The parameters of 0 to 10 were tested according to the order of sequence length of documented ncRNAs (~500). The separative power of gapped model ($g>0$) was substantially higher rather than the non-gapped model ($g=0$). For the separation of mRNAs and the others, $e=2$ and $g=10$ were adopted in the GMMI calculation (Fig. 2D). While 97% of the mRNAs had GMMI scores above 1, 97% of the ncRNAs had GMMI scores below 1 (data not shown). The *Z* score average of the mRNAs was 0.11, and that of the ncRNAs was −2.94. Therefore, we clearly separated scores for

previously annotated ncRNA sequences from the total set of RNA sequences. We then calculated the correlation coefficients, to determine the threshold value for the identification of ncRNA sequences. The BCC for the quality measure of the isolation was 0.82 and the corresponding threshold score was −2.34; achieving 82.0% sensitivity, 83.5% selectivity and 99.5% specificity.

3.2. Extraction of novel candidate ncRNAs

Novel ncRNA candidates were predicted using the following strategy (Fig. 3). We predicted an initial total of 14,068 sigma70 promoters and 6606 rho-independent terminators (step 1). As the length of previously identified sRNAs transcription units ranged from 70 to 500 nt, we chose 4579 of the putative transcription units which agreed with these sizes (step 2). We trained the GMMI using 4311 ORF sequences and extracted 433 transcriptional units with a Z score of −2.34 or less as initial candidates (step 3). This set of initial candidates contained redundant fragments that overlapped on the same DNA strand. Therefore, 184 sequences were extracted by sorting overlapping candidates as follows. We initially selected the innermost DNA region between the sigma70 promoter and the rho-independent terminator, restricting the size to above 70 nt, then we chose the best scores according to the GMMI, *pftool*2.2 and RNAMotif. After deselecting candidates that overlapped any ORFs (as indicated in GenBank annotations) on the same DNA strand, we finally identified 133 novel ncRNA candidates, denoted NC001 to NC133 (step 4). Among these 133 candidates, 87 were encoded by regions that were bordered by two genes (intergenic regions; Table 1), 46 were on the complementary strand of an annotated ORF (*cis*-antisense; Table 2). They included 29 previously annotated ncRNAs (Table 3).

3.3. Experimental validations of ncRNA candidates

RNA expression analysis using RT-PCR against total RNA from the *E. coli* late log phase was initially performed to detect expression of corresponding genomic region of our novel ncRNA candidates. This detected nine major PCR products (from ncRNA candidates NC008, NC020, NC028, NC030, NC051, NC079, NC086, NC109 and NC116) in an RT-dependent manner (Fig. 4). These PCR products matched each predicted sizes according to the locations of the primer pairs. We also confirmed the nucleotide sequences of the products after subcloning. Two PCR products generated from primers directed to candidates, NC011 and NC065 had some unspecific size amplicons, however, the remaining amplicons were the predicted size and their nucleotide sequences provided further confirmation that they matched the candidate ncRNAs. We also detected the documented sRNA *dsrA* (Sledjeski and

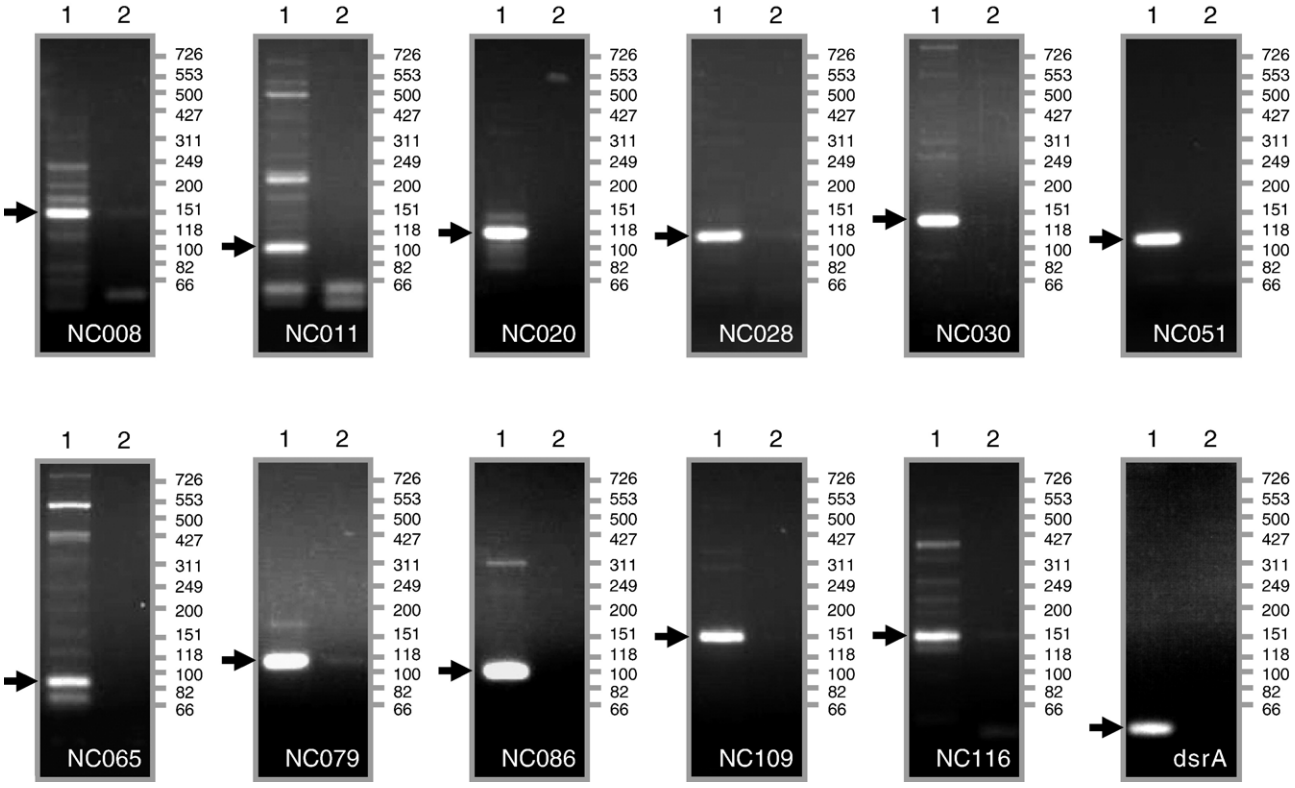

Fig. 4. RT-PCR analysis of the putative ncRNAs in log phase. PCR products were separated using 3% NuSieve agarose-gel electrophoresis and stained with EtBr. The arrows indicate the predicted size according to the designated primers. The lane 1 contains the product of the PCR reaction (RT+). The lane 2 represents the product of the same reaction without reverse transcriptase (RT−, negative control). DsrA is a documented sRNA for the positive control. The size (bp) of the φX174/*Hinf* I marker (STRATAGENE, USA) is shown on the right. The same results were obtained in at least two independent experiments for each gene.

Gottesman, 1995). These results show that genomic region of some candidate ncRNAs are actually transcribed in *E. coli*, although they are previously unknown. Notably, NC008, NC028, NC086 and NC109 were located on the complementary strands of annotated ORFs. This suggests that these RNAs might regulate their neighbour mRNAs through possible antisense-mediated mechanism.

A signal detected by RT-PCR confirms that the target region is transcribed but does not reflect the size of the transcript. Therefore we also performed Northern hybridization analyses with strand specific probes from the selected ncRNAs candidates against total RNA from the *E. coli* late log and early stationary phases. The expressions of three novel small transcripts, NC051, NC065 and NC092, were detected (Fig. 5). NC092 was differentially expressed in log and stationary phase, which is in agreement with previous sRNAs in *E. coli* (Argaman et al., 2001; Wassarman et al., 2001; Vogel et al., 2003). The documented sRNA IS061 (Chen et al., 2002) was also detected (data not shown). The sizes of three transcripts were larger than our predictions. Because we selected innermost DNA region between the sigma70 promoter and the rho-independent terminator as candidates, it is likely to be transcribed from different combination of promoter and terminator. The predicted transcription unit of NC092 transcript overlaps with carboxy-terminal of neighbouring *pgk* mRNA on the opposite strand by 13 nt suggesting that NC092 may act as small RNA regulator by base pairing with *pgk* mRNA.

We could not get any positive signals for some transcripts that have not previously been identified by the Northern analysis (data not shown). Although the absence of detection might be caused by their unstabilities, low abundances and/or certain culture/growth condition, total of 11 transcripts, including four antisense RNAs, were detected by RT-PCR under the single set of culture/growth condition. The expression of other transcripts may be detectable in further studies using different experimental conditions.

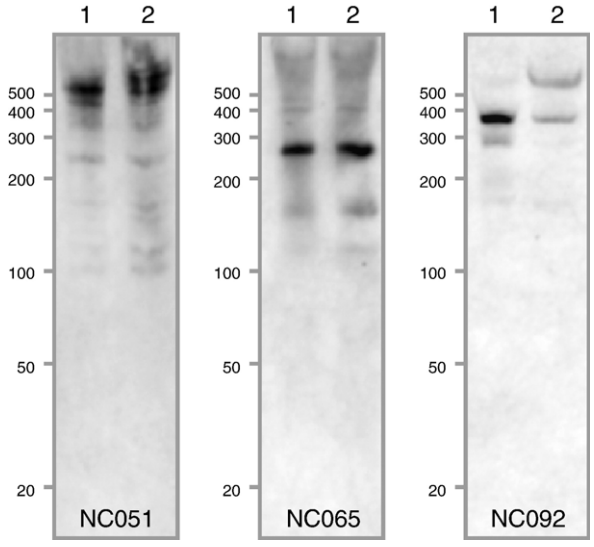

Fig. 5. Northern blot analysis of new small RNAs in *E. coli*. Total RNAs were prepared from either stationary phase (lane 1) or log phase (lane 2). Membranes were hybridized with gene specific oligonucleotides for NC051, MC065 and NC092. Numbers at left indicate length of marker RNA (BioDynamics, Japan) in base. The same results were obtained in at least two independent experiments for each gene.

4. Discussion

This report described a novel index (GMMI), which evaluates the specificity of sequence pattern with the combinations of patterns located on separate positions of consensus sequences, codon biases and/or consisting possible structural motifs, based on the Gapped Markov Model. Training sets comprising tRNAs and rRNAs were used to test the GMMI performance, and successfully separated the sequences in the training set from other categories of RNA, as expected. This might have been due to strong biases in the training set resulting from the domain sequences and/or the high stabilities of these structures (for example, the cloverleaf structure of tRNA). The successful isolation of ncRNAs from a total set of RNAs implies that the GMMI is useful for the detection of novel ncRNAs. Despite the fact that the scores for the majority of the sRNAs fell within the range of most ncRNAs, some of sRNAs were distributed within the range of the mRNAs. We therefore suspect that the sRNA category might include small peptide-coding RNAs. Most of these sRNAs have not been proved to function as ncRNAs in previous computational or experimental studies.

Twenty-nine previously annotated sRNA genes were included among our candidates. In previous studies, tRNAs (Lowe and Eddy, 1997; Rivas and Eddy, 2001; Sakakibara, 2003), microRNAs (miRNAs) (Grad et al., 2003; Lai et al., 2003; Lim et al., 2003a, b), small nucleolar RNAs (snoRNAs) (Liang et al., 2002; Chen et al., 2003) and the other characteristic ncRNA families have been modeled using complicated methods, such as stochastic context-free grammars, hidden Markov models, and/or comparative genomics with calculation of the structural formation and its free energy. However, our present results imply that a simpler approach based on the Markov model is useful in isolating the ncRNA cluster, which constitutes a superfamily of functional RNA molecules. Although most of the 29 candidates matched transcribed regions of tRNAs and rRNAs, our set also included five sRNA-encoding genes recently identified: *rydB*/tpe7/IS082, *ssrA*, *sraJ*/*ryiA*/k19, C0362 and tke1/*sroF*. High expression of RydB causes decreased expression of the sigma38 factor RpoS (Wassarman et al., 2001). *SsrA* releases a stalled ribosome from the end of 'broken' mRNA, which lacks a stop codon, by acting both as a tRNA and an mRNA via a '*trans*-translation' mechanism (transfer-messenger RNA; tmRNA) (Muto et al., 1996). *SraJ* is subject to RNase III processing (Argaman et al., 2001) and an interaction with the Hfq protein has been reported (Wassarman et al., 2001). The functions of the remaining two genes, C0362 and tke1, are not known.

According to our results, 46 (approximately 40%) of our candidates were encoded on the complementary DNA strands of annotated ORFs (*cis*-antisense). Antisense RNAs are diffusible regulatory RNAs that bind to the complementary sequences of mRNAs in order to control their biological

function at the post-transcriptional level. Our experimental results showed that four of novel antisense transcripts were detected using RT-PCR. Although these bands may not correspond to the predicted size of the transcription units, expressions and their sizes of two transcripts including the one detected in RT-PCR analyses are confirmed by Northern hybridization. Computational methods have not yet been established to predict *cis*-encoded saRNAs from the bacterial genome sequence. Because most of the genome-based methods previously used to predict ncRNAs or sRNAs have been applied only in intergenic regions, antisense transcripts overlapping with neighbour ORFs in the complementary strands could not be detected. Although several saRNAs transcribed from different loci in the target mRNAs (*trans*-antisense) have been described in the *E. coli* genome (Carpousis, 2003), few small *cis*-encoded saRNAs have been documented (Kawano et al., 2002; Vogel et al., 2003; Kawano et al., 2005). For example, the LdrD peptide and the RdlD antisense regulatory RNA act as a toxin–antitoxin pair (Kawano et al., 2002). In addition, the interaction between two newly identified sRNAs, RyeB (Wassarman et al., 2001) and SraC/RyeA (Argaman et al., 2001; Wassarman et al., 2001), has been reported to result in RNase-III-dependent cleavage (Vogel et al., 2003). However, the existence of large amounts of such transcripts has been predicted using the recent shotgun-cloning approach (experimental RNomics), which suggested that 5% of all sRNAs are *cis*-antisense (Vogel et al., 2003). Likewise, Carpusis pointed out that *cis*-encoded saRNAs with extensive complementarity to coding regions would have been overlooked using previous computational methods (Carpousis, 2003). Among eukaryotes, many chromosomal *cis*-antisense transcripts are involved in aspects of gene-expression regulation, such as genomic imprinting (Reik and Walter, 2001) and circadian clocks (Crosthwaite, 2004). However, in bacterial species, *cis*-antisense transcripts are principally known in extrachromosomal elements (for example, transposons, plasmids and bacteriophages). An abundance of chromosomal small *cis*-antisense RNAs and their regulating mechanisms in *E. coli* or other bacterial species might be expected on the basis of our results. The products of mRNAs encoded on the opposite strand of our 46 antisense candidates included two ATP-binding proteins and eight metabolic enzymes (Supplementary Table 3). It is possible that some of these antisense candidates can regulate metabolic pathways at the post-transcriptional level.

NC092, newly detected by Northern hybridization, overlaps with 3′ end of *pgk* in opposite strand. Previously identified sRNA GadY regulates the expression of GadX mRNA, and the possible regulation mechanism via base pairing between 3′ sequences of GadY and GadX has been suggested (Opdyke et al., 2004). We hereto suggest that NC092 acts as small RNA regulator by base paring with 3′ sequence of mRNAs encoded in opposite strand. Since two RNAs are encoded on same loci and overlapped, any mutations arising in the double strand DNA may not affect their regulatory mechanisms.

In summary, here we present the novel ncRNA candidates in *E. coli*, isolated by the prediction of transcription units followed by the GMMI-based evaluation and the prediction of these units. Although the experimental validation of the candidates is needed more widely, our list will be informative for the further RNomics approaches such as tilling array. Five *cis*-antisense small transcripts (one is detected by Northern hybridization) were verified under the certain culture/growth conditions. We showed that our computational approach may enable the discovery of novel sRNAs, including *cis*-antisense RNAs, in the genomes of bacteria without requiring comparisons with other species or other experimental resources. Finally we suggest that the *cis*-antisense RNAs might be more common than previously thought, some of which may have crucial roles such as RNA metabolisms within *E. coli* and other bacterial species.

Acknowledgements

We are grateful to Satoshi Harada, Asako Sato and Dr. Yoko Fukuda for excellent experimental support, Yuka Watanabe for assistance and performing computational tests of the GMMI, and Dr. Takeshi Ara and Dr. Tomoya Baba for helpful advice. The authors also thank Dr. Elena Lesnik for providing the descriptor file for the rho-independent terminator prediction.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2005.12.034.

References

Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y., Tomita, M., 2003. G-language genome analysis environment: a workbench for nucleotide sequence data mining. Bioinformatics 19, 305–306.

Argaman, L., et al., 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr. Biol. 11, 941–950.

Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–367.

Carpousis, A.J., 2003. Degradation of targeted mRNAs in *Escherichia coli*: regulation by a small antisense RNA. Genes Dev. 17, 2351–2355.

Carter, R.J., Dubchak, I., Holbrook, S.R., 2001. A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res. 29, 3928–3938.

Chen, S., et al., 2002. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. Biosystems 65, 157–177.

Chen, C.L., Liang, D., Zhou, H., Zhuo, M., Chen, Y.Q., Qu, L.H., 2003. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. Nucleic Acids Res. 31, 2601–2613.

Crosthwaite, S.K., 2004. Circadian clocks and natural antisense RNA. FEBS Lett. 567, 49–54.

Eddy, S.R., 2001. Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. 2, 919–929.

Grad, Y., et al., 2003. Computational and experimental identification of *C. elegans* microRNAs. Mol. Cell 11, 1253–1263.

Hershberg, R., Altuvia, S., Margalit, H., 2003. A survey of small RNA-encoding genes in *Escherichia coli*. Nucleic Acids Res. 31, 1813–1820.

Kawano, M., Oshima, T., Kasai, H., Mori, H., 2002. Molecular characterization of long direct repeat (LDR) sequences expressing a stable mRNA encoding for a 35-amino-acid cell-killing peptide and a *cis*-encoded small antisense RNA in *Escherichia coli*. Mol. Microbiol. 45, 333–349.

Kawano, M., Reynolds, A.A., Miranda-Rios, J., Storz, G., 2005. Detection of 5′- and 3′-UTR-derived small RNAs and *cis*-encoded antisense RNAs in *Escherichia coli*. Nucleic Acids Res. 33, 1040–1050.

Lai, E.C., Tomancak, P., Williams, R.W., Rubin, G.M., 2003. Computational identification of *Drosophila* microRNA genes. Genome Biol. 4, R42.

Lesnik, E.A., Sampath, R., Levene, H.B., Henderson, T.J., McNeil, J.A., Ecker, D.J., 2001. Prediction of rho-independent transcriptional terminators in *Escherichia coli*. Nucleic Acids Res. 29, 3583–3594.

Liang, D., et al., 2002. A novel gene organization: intronic snoRNA gene clusters from *Oryza sativa*. Nucleic Acids Res. 30, 3262–3272.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., Bartel, D.P., 2003a. Vertebrate microRNA genes. Science 299, 1540.

Lim, L.P., et al., 2003b. The microRNAs of *Caenorhabditis elegans*. Genes Dev. 17, 991–1008.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., Sampath, R., 2001. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. 29, 4724–4735.

Muto, A., Sato, M., Tadaki, T., Fukushima, M., Ushida, C., Himeno, H., 1996. Structure and function of 10Sa RNA: *trans*-translation system. Biochimie 78, 985–991.

Opdyke, J.A., Kang, J.G., Storz, G., 2004. GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. J. Bacteriol. 186, 6698–6705.

Reik, W., Walter, J., 2001. Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. 2, 21–32.

Rivas, E., Eddy, S.R., 2001. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2, 8.

Rivas, E., Klein, R.J., Jones, T.A., Eddy, S.R., 2001. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr. Biol. 11, 1369–1373.

Sakakibara, Y., 2003. Pair hidden Markov models on tree structures. Bioinformatics 19, i232–i240.

Sledjeski, D., Gottesman, S., 1995. A small RNA acts as an antisilencer of the H-NS-silenced rcsA gene of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 92, 2003–2007.

Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E., Rosenow, C., 2002. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. Nucleic Acids Res. 30, 3732–3738.

Vogel, J., et al., 2003. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. Nucleic Acids Res. 31, 6435–6443.

Wassarman, K.M., Zhang, A., Storz, G., 1999. Small RNAs in *Escherichia coli*. Trends Microbiol. 7, 37–45.

Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G., Gottesman, S., 2001. Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev. 15, 1637–1651.