



Conservation analysis of small RNA genes in *Escherichia coli*

Yong Zhang, Zhihua Zhang, Lunjiang Ling, Baochen Shi and Runsheng Chen*

Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Received on June 24, 2003; revised on August 23, 2003; accepted on September 29, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: Small RNA (sRNA) genes in *Escherichia coli* have been in focus recently, as 44 out of 55 experimentally confirmed sRNA genes have been precisely located in the genome. The object of this study is to analyze quantitatively the conservation of these sRNA genes and compare it with the conservation of protein-encoding genes, function-unknown regions and tRNA genes.

Results: The results show that within an evolutionary distance of 0.26, both sRNA genes and protein-encoding genes display a similar tendency in their degrees of conservation at the nucleotide level. In addition, the conservation of sRNA genes is much stronger than function-unknown regions, but much weaker than tRNA genes. Based on the conservation of studied sRNA genes, we also give clues to estimate the total number of sRNA genes in *E.coli*.

Contact: crs@sun5.ibp.ac.cn

Supplementary information: Supplementary information is available at <http://www.bioinfo.org.cn/SM/sRNAconservation.htm>

INTRODUCTION

Small RNA (sRNA) genes, the prevalent term of non-coding RNA genes in bacteria (Storz, 2002), have drawn much attention in later years because of their abundance and diversity. Studies of sRNA have focused on *Escherichia coli* K-12 because of its important role as a model organism. Up to now, 13 sRNAs in *E.coli* have been thoroughly studied (Wassarman *et al.*, 1999; Urbanowski *et al.*, 2000; Majdalani *et al.*, 2001; Masse and Gottesman, 2002), and most of them were discovered by chance. These sRNAs function as central regulators in response to diverse environmental changes (Wassarman, 2002) and are involved in regulating various biological processes, including RNA processing, mRNA stability, translation, protein stability and protein secretion (Wassarman *et al.*, 1999).

With the availability of the complete genome sequence of *E.coli* K-12 (Blattner *et al.*, 1997), it has become possible to predict sRNA genes on the genomic scale. Six groups have recently published different methods for predicting sRNA genes in *E.coli* (Argaman *et al.*, 2001; Carter *et al.*, 2001; Rivas *et al.*, 2001; Wassarman *et al.*, 2001; Chen *et al.*, 2002; Tjaden *et al.*, 2002). These approaches are based on different methods or sources of information, such as sequence conservation (Argaman *et al.*, 2001; Rivas *et al.*, 2001; Wassarman *et al.*, 2001), transcription signals (Argaman *et al.*, 2001; Chen *et al.*, 2002), structure conservation (Rivas *et al.*, 2001), DNA array data (Wassarman *et al.*, 2001), machine learning (Carter *et al.*, 2001) and whole genome array (Tjaden *et al.*, 2002). Except for one pure theoretical approach, the groups have used various experimental approaches to verify some predicted sRNA genes. Currently, in addition to these 13 thoroughly studied sRNA genes, there are 55 known sRNA genes in *E.coli* (Hershberg *et al.*, 2003). However, the total number of sRNA genes in *E.coli* is still an unsolved problem (Eddy, 2001).

Although one group has attempted to analyze the conservation of these known sRNA genes in other bacteria, no quantitative conclusion was drawn except that most of sRNA genes are not conserved beyond *Yersinia pestis* (Hershberg *et al.*, 2003). Meanwhile, no research has focused on the comparison between the conservation of sRNA genes and that of other genomic regions. In this work, we introduce a quantitative approach to analysis of the conservation of sRNA genes in *E.coli*, compared with the conservation of protein-encoding genes, function-unknown regions and tRNA genes. The results also give clues for estimating the total number of sRNA genes in *E.coli*.

DATA AND METHODS

All protein-encoding and intergenic regions in *E.coli* K-12 were created based on the gene annotations in the EcoGene12 release of the EcoGene database (Rudd, 2000). A protein-encoding region was defined as a genomic region, which contains an open reading frame (ORF) on either of the two strands, whereas other parts of genome were defined as intergenic

*To whom correspondence should be addressed.

regions. Function-unknown regions are intergenic regions that have no functional regions located inside. Functional regions include sRNA genes, tRNA genes, rRNA genes, IS elements, promoters, terminators, regulatory regions and repeats. In this work, the sRNA gene dataset included 44 sRNA genes for each of which either the 5' or the 3' end was experimentally determined, or the 3' end was predicted (Hershberg *et al.*, 2003). Forty two out of 44 sRNA genes are completely located in intergenic regions, while the remaining two sRNA genes, *sraL* and *sraB*, slightly overlap with protein-encoding regions (7 and 5 bp separately). The locations of tRNA genes, rRNA genes and IS elements were derived directly from the EcoGene database. The data on promoters were drawn from promEC (Hershberg *et al.*, 2001), a database of *E.coli* mRNA promoter sequences. A total of 471 experimentally identified mRNA transcriptional start sites are listed in promEC, and sequences spanning nucleotides -75 to +25 relative to the transcriptional start sites were regarded as promoters. The sequences and map positions of terminators and regulator regions (e.g. bending, attachment sites) were collected from ECDC (Wahl and Kroger, 1995; Kroger and Wahl, 1996), the *E.coli* database collection. In ECDC, some sRNA genes are included in regulator regions. In this work, these sRNA genes were eliminated from the dataset of regulator regions. The sequences of terminators and regulator regions were compared with the *E.coli* genome sequence using the BLAST program (Altschul *et al.*, 1997). For each terminator or regulator region, when the sequence had a complete alignment with a genomic region, and the sequence's map position was accordant with the genomic region's location, it was regarded as being located in the genome. Totally 166 out of 186 terminator sequences and 16 out of 28 regulator regions were located in the *E.coli* genome. Because certain types of repeats are hypothesized as functional, in this work repeats were also considered as functional regions. Locations of intergenic repeats were collected from the EcoGene database (Rudd, 1999). The release of these intergenic repeats corresponds to the EcoGene12 release of gene annotations. The number and average length of these functional regions are shown in Table 1.

16S rRNA, a classical molecule used for the reconstruction of microbial phylogeny, was chosen to calculate the relative evolutionary distances between *E.coli* K-12 and other organisms studied in this work. Seventy eight genomes, including 67 bacteria and 11 archaea, with their 16S rRNA genes, were obtained from NCBI (<ftp://ftp.ncbi.nlm.nih.gov>) and are shown in the Supplementary materials. The program CLUSTALW (Thompson *et al.*, 1994) was used to do multiple sequence alignment for all 78 16S rRNA genes, and then the program DNADIST in the PHYLIP package with all default settings was used to calculate a distance matrix based on the alignment of the rRNA genes. The distances between 16S rRNA gene of *E.coli* K-12 and other 16S rRNA genes were defined as the evolutionary distances between *E.coli* K-12 and other organisms, as shown in the Supplementary

Table 1. Average length and number of known functional regions located in intergenic regions of *E.coli*

Name	Average length (bp)	Number
sRNA	157	44
tRNA	78	86
rRNA	1458	22
IS elements	898	35
Promoter	101	379
Terminator	31	147
Regulatory region	188	11
Repeats	38	2013

For each kind of functional region, the value of average length was calculated based on all sequences of this kind discovered and located in the genome of *E.coli*. The value of number was the amount of sequences of this kind completely located in intergenic regions or located in intergenic regions more than half in length.

materials. These calculated evolutionary distances correspond well to other microbial phylogenetic data, i.e. the 18 gamma-subdivision proteobacteria are considered closest to *E.coli* K-12 and the 11 archaea in the most outlying groups.

The genome sequence of *E.coli* K-12 was used as query to compare with each selected organism by program BLAST (Altschul *et al.*, 1997) and then the genome sequence of each selected organism was used as query to compare with *E.coli* K-12 in reverse. In order to increase the sensitivity and length of alignment, the parameter wordsize was set to 7 (rendering BLAST very time-consuming), and the penalty for a mismatch was set to -2. For comparing the conservation of sRNA genes and other regions, all 'orthologous' sequences between *E.coli* K-12 and other 77 organisms were identified. Here 'orthologous' sequences were defined as follows: (a) a pair of homologous regions in two organisms identified by BLAST (*E*-value lower than 0.01); (b) the value of identity in this region must be over 70% and the length at least 20 bp and (c) each sequence in this pair must be the best hit for the corresponding sequence in the other genome.

In this work, we focused on orthologous sequences between *E.coli* and other organisms in order to analyze the conservation of sRNA genes in *E.coli*. We separately compared the locations of orthologous sequences in *E.coli* to the locations of sRNA genes, protein-encoding regions, function-unknown regions and tRNA genes, and counted the base pair number in each kind of region which overlapped with orthologous sequences between *E.coli* and another organism. Reasonably, *E.coli* tends to share more orthologous sequences with larger genome. To remove the effect of genome size, we calculated the ratio of smaller genome size to *E.coli* genome size, and then divided those counted base pair numbers by this calculated ratio. If the genome size of *E.coli* was the relative smaller one, that ratio was equal to 1. For comparison between different regions, the numbers should also be normalized by the total base pairs of related kind of region. In this work, these calculated proportions were defined as the degrees of

conservation of related regions between those organisms and *E.coli*.

RESULTS AND DISCUSSION

For each organism, the base pair numbers of orthologous sequences shared with each kind of region in *E.coli* and the related degrees of conservation are shown in the Supplementary materials. Not unexpectedly, the degree of conservation of each kind of region depended on the evolutionary distance from *E.coli*. Similar to protein-encoding and function-unknown regions, the falling trend in the degrees of conservation of sRNA genes was approximately a negative exponential function of the evolutionary distance. For each of these three kinds of regions, the open interval of evolutionary distance, i.e. from 0 to the point where the degree of conservation drops below 0.001, was considered as the region where the negative exponential function is in effect. That is, (0, 0.2658) was found to be the effective open interval of sRNA genes, (0, 0.2914) the effective open interval of protein-encoding regions and (0, 0.1577) the effective open interval of function-unknown regions. Different from these three kinds of regions, tRNA genes tend to be much more conserved than other regions. Therefore, in this work tRNA genes were used as a more conserved control for the sRNA genes. In order to compare quantitatively the conservation of sRNA genes, protein-encoding regions, function-unknown regions and tRNA genes, the effective open interval of evolutionary distance of the tRNA genes had also to be determined. Here, (0, 0.2914), the largest open interval of other three kinds of regions, was chosen as the tRNA genes effective open interval. In Figure 1, the degrees of conservation in the effective open intervals of the four kinds of regions are shown in a half-logarithmic coordinate, and the parameters of linear fit are shown in Table 2. Except for the tRNA genes, the trends in the degrees of conservation of the other three kinds of regions were linear in a half-logarithmic coordinate (correlation coefficient $|R| > 0.95$) in their effective open intervals. Furthermore, the slope value of the line fit by sRNA genes was similar to that of protein-encoding regions, and they were both intervenient between the slope values of tRNA genes and function-unknown regions. In general, below an evolutionary distance of 0.26, the trend in the degrees of conservation of sRNA genes could not be distinguished from that of protein-encoding regions. In addition, with the exception of few organisms, the declining tendency in the degrees of conservation of protein-encoding regions could be extended to the boundary of bacteria and archaea. The data from at least some organisms also indicated that given more sRNA gene data are made available in the near future, the trend in the degrees of conservation of sRNA genes might also be extended similarly far (see Supplementary materials). The results indicated that within the domain of bacteria, both sRNA genes and protein-encoding genes in *E.coli* statistically tended

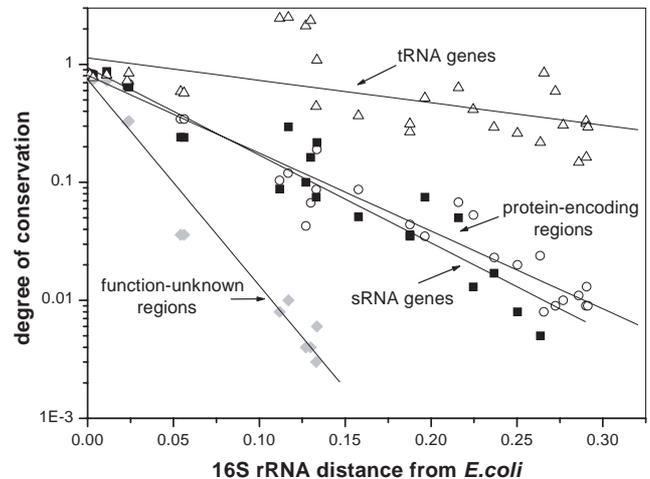


Fig. 1. Linear fit of degrees of conservation versus evolutionary distance for four kinds of genomic regions in a half-logarithmic coordinate. Every point indicates the degree of conservation of certain kind of region between *E.coli* and the corresponding organism. Different symbols present sRNA genes (closed squares), protein-encoding regions (open circles), function-unknown regions (gray squares) and tRNA genes (open triangles) separately. Each line presents a linear fit of the region annotated by an arrow, and indicates the tendency of variation in the degree of conservation with increasing evolutionary distance.

to show the same variation in degrees of conservation at the nucleotide level.

Although several groups have developed different approaches to predict sRNA genes in *E.coli*, the total number of sRNA genes is still an interesting and puzzling question (Eddy, 2001). In this work, we have attempted to use the conservation of sRNA genes to find an approximate estimate of this number. All studied 44 sRNA genes are completely or mostly located in intergenic regions, making it reasonable to assume there are still some sRNA genes hidden in the intergenic regions, or more exactly, in function-unknown regions. If the declining tendency of degrees of conservation with increasing evolutionary distance of undiscovered sRNA genes is presumed to be similar to that of the 44 studied sRNA genes, the estimated total number of sRNA genes would be acquired by regarding all base pairs of orthologous sequences located in function-unknown regions as being contributed by undiscovered sRNA genes. To avoid the influence of significant homology or lack of data, six organisms within moderate evolutionary distance ranging from 0.11 to 0.14 were used in the estimation. For each of the six organisms, we calculated the ratio of base pair number of orthologous sequences located in function-unknown regions to undiscovered sRNA genes' total base pair number, which is unknown and denoted as x , and then removed the effect of genome size. Corresponding to the definition of degree of conservation in the Methods section, these calculated proportions, which are functions of x , were

Table 2. Parameters of the linear fit in Figure 1

Regions	A	A error	B	B error	R	SD	P
sRNAs	-0.02626	0.07424	-7.43427	0.49323	-0.95676	0.20256	<0.0001
Protein-encoding regions	-0.09265	0.05473	-6.59847	0.28267	-0.97441	0.15982	<0.0001
Function-unknown regions	-0.12896	0.07597	-17.4572	0.88975	-0.98477	0.17954	<0.0001
tRNAs	0.05494	0.09287	-1.90093	0.47963	-0.59274	0.27118	0.00044

The two parameters *A* and *B* are intercept and slope values, respectively, in the linear fit expressions: \log_{10} (degree of conservation) = $A + B \times$ (evolutionary distance). *R* is the correlation coefficient; SD is the standard deviation; *P* is the probability that *R* is zero.

defined as the degrees of conservation of undiscovered sRNA genes. For each of the six organisms, a value of undiscovered sRNA genes' total base pair number was calculated under the following formula: \log_{10} (degree of conservation) = $A + B \times$ (evolutionary distance), where *A* and *B* were the known sRNA genes' parameters of linear fit in Table 2. Thus, six different values of unknown sRNA genes' total base pair numbers were calculated separately. Meanwhile, the average length of these unknown sRNA genes was hypothesized as same as that of 44 studied sRNA genes. Therefore, together with 44 studied sRNA genes, six numbers of total sRNA genes were estimated, from 118 to 260, with average 182. This approach to estimate the total number of sRNA genes in *E.coli* was based on several assumptions. Among them, some presumed features of undiscovered sRNA genes were reasonable extrapolated from the features of studied 44 sRNA genes, such as being located in function-unknown regions, having the same average length and trend in the degrees of conservation with studied sRNA genes. As to the assumption that all base pairs of orthologous sequences located in function-unknown regions are contributed by undiscovered sRNA genes, it would result in overestimating the number of undiscovered sRNA genes because there must be some orthologous sequences, which are contributed by non-sRNA regions. Only organisms within moderate evolutionary distance were taken into consideration might avoid the influence of significant homology and ensure the estimated number was within an acceptable range. Totally, 150 sRNA genes have been confirmed by experimental approaches, or predicted by at least two different studies in *E.coli* (Hershberg *et al.*, 2003), which is a number comparable with our estimate. Therefore, our simple method to estimate the total number of sRNA genes could give a referential upper limit to the total number of sRNA genes in *E.coli*.

Our study is based on the limited available dataset of sRNA genes in *E.coli*. Fortunately, with the development of experimental and computational Rnomics (Filipowicz, 2000; Huttenhofer *et al.*, 2002), it is reasonable to anticipate that large amount of sRNA genes in different organisms will be found in the next few years. The conservation of sRNA genes will be more thoroughly studied when sRNA genes in more bacteria can be collected.

ACKNOWLEDGEMENTS

We thank Dr Zhenyu Xuan of Cold Spring Harbor Laboratory for providing helpful discussions and critical review of the manuscript. This work was supported by the National Knowledge Innovation Program of the Chinese Academy of Sciences grant no. KSCX2-2-07 and KJCX1-08, the National '863' High-tech Program grant no. 2002AA231031, National Key Basic Research & Development Program (973) grant no. 2002CB713805 and the Bioinformatics Program of Institute of Computing Technology of the Chinese Academy of Sciences grant no. IIP2003-5.

REFERENCES

- 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.
- Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.
- Blattner,F.R., Plunkett,G., III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
- Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.*, **29**, 3928–3938.
- Chen,S., Lesnik,E.A., Hall,T.A., Sampath,R., Griffey,R.H., Ecker,D.J. and Blyn,L.B. (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, **65**, 157–177.
- Eddy,S. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
- Filipowicz,W. (2000) Imprinted expression of small nucleolar RNAs in brain: time for RNomics. *Proc. Natl Acad. Sci. USA*, **97**, 14035–14037.
- Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813–1820.
- Hershberg,R., Bejerano,G., Santos-Zavaleta,A. and Margalit,H. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.*, **29**, 277.

- Huttenhofer, A., Brosius, J. and Bachelier, J.P. (2002) RNomics: identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.*, **6**, 835–843.
- Kroger, M. and Wahl, R. (1996) Compilation of DNA sequences of *Escherichia coli* K12 (ECD and ECDC; update 1995). *Nucleic Acids Res.*, **24**, 29–31.
- Majdalani, N., Chen, S., Murrow, J., St John, K. and Gottesman, S. (2001) Regulation of RpoS by a novel small RNA: the characterization of RprA. *Mol. Microbiol.*, **39**, 1382–1394.
- Masse, E. and Gottesman, S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 4620–4625.
- Rivas, E., Klein, R.J., Jones, T.A. and Eddy, S.R. (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.
- Rudd, K.E. (1999) Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.*, **150**, 653–664.
- Rudd, K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
- Storz, G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E. and Rosenow, C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res.*, **30**, 3732–3738.
- Urbanowski, M.L., Stauffer, L.T. and Stauffer, G.V. (2000) The *gcvB* gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in *Escherichia coli*. *Mol. Microbiol.*, **37**, 856–868.
- Wahl, R. and Kroger, M. (1995) ECDC—a totally integrated and inter-actively usable genetic map of *Escherichia coli* K12. *Microbiol. Res.*, **150**, 7–61.
- Wassarman, K.M. (2002) Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. *Cell*, **109**, 141–144.
- Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G. and Gottesman, S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.
- Wassarman, K.M., Zhang, A. and Storz, G. (1999) Small RNAs in *Escherichia coli*. *Trends Microbiol.*, **7**, 37–45.