DNA Research **15**, 387–396, (2008) doi:10.1093/dnares/dsn027

MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes

Hideki Noguchi*, Takeaki Taniguchi, and Takehiko Itoh

Advanced Science and Technology Research Group, Mitsubishi Research Institute, Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100-8141, Japan

(Received 19 August 2008; accepted 24 September 2008; published online 21 October 2008)

Abstract

Recent advances in DNA sequencers are accelerating genome sequencing, especially in microbes, and complete and draft genomes from various species have been sequenced in rapid succession. Here, we present a comprehensive gene prediction tool, the MetaGeneAnnotator (MGA), which precisely predicts all kinds of prokaryotic genes from a single or a set of anonymous genomic sequences having a variety of lengths. The MGA integrates statistical models of prophage genes, in addition to those of bacterial and archaeal genes, and also uses a self-training model from input sequences for predictions. As a result, the MGA sensitively detects not only typical genes but also atypical genes, such as horizontally transferred and prophage genes in a prokaryotic genome. In this paper, we also propose a novel approach for analyzing the ribosomal binding site (RBS), which enables us to detect species-specific patterns of the RBSs. The MGA has the ingenious RBS model based on this approach, and precisely predicts translation starts of genes. The MGA also succeeds in improving prediction accuracies for short sequences by using the adapted RBS models (96% sensitivity and 93% specificity for 700 bp fragments). These features of the MGA expedite wide ranges of microbial genome studies, such as genome annotations and metagenome analyses.

Key words: bioinformatics; gene-finding; prokaryote; phage; ribosomal binding site

1. Introduction

Identification of genes on genomic sequences is the indispensable first step in every genome analysis, including individual genome analysis of a single organism and metagenomic analyses. Sequence similarity-based methods of gene predictions enable us to detect reliably the genes if their DNA or amino acid sequences have strong similarities to those of known genes. However, a significant portion of genes has no sequence similarities to known genes, and *ab initio* gene-finding methods are necessary for identifying all genes on newly sequenced microbial genomes, particularly those of uncharacterized or poorly characterized species. Computational gene finding from genomic sequences has a long history[1–3], and a number of tools have been developed for predicting prokaryotic genes. These gene-finding tools have been widely used for annotation processes of prokaryotic genomes.

Although conventional gene-finding tools have achieved extremely high prediction performances, they have some critical limitations. Most conventional tools require predetermined statistical models of the known genes of a target species[4–11] or a long enough input sequence for statistical models to perform self-training[12–16]. This is because the tools

Edited by Masahira Hattori

* To whom correspondence should be addressed. Tel. +81 3-3277-0556. Fax. +81 3-3277-0568. E-mail: nog@mri.co.jp

are designed to predict genes on complete genomes having several million base pairs. However, a target genomic sequence is not always long enough. For example, second-generation DNA sequencers, which have put high throughput sequencing into practice, especially those of microbial genomes[17,18], produce vast amounts of very short sequence reads. The short reads are assembled into some longer contig sequences, but the contigs are usually still short [far shorter than 1 mega bases (Mb)][19–22]. A fosmid clone, which has $\sim$40 kb in insert length, is another example of a short genomic sequence. Moreover, metagenomic analyses produce large amounts of short sequences derived from multiple species' genomes. Most of the conventional gene-finding tools cannot be applied to such sequences. MetaGene[23] (MG) is one of the new tools that is applicable to gene prediction on such short anonymous sequences.

MG is a gene-finding program originally developed for metagenomic sequence data, which is a mixture of (short) sequences derived from various prokaryotic genomes. MG assumes correlations between the GC content and the di-codon frequencies of an input sequence, and enables us to predict genes accurately on short anonymous sequences without any training. MG can be successfully applied to wide varieties of prokaryotic genomic sequences[24–27], but two major limitations exist: one is the lack of a ribosomal binding site (RBS) model, and the other is less sensitivity to atypical genes, whose codon usages are different from those of typical genes. When MG is applied to very short sequences containing one or two partial genes, these limitations are not significant. However, such limitations are undesirable when MG is applied to longer genomic sequences for precise annotations. To overcome these limitations and to improve the usability of the program, we developed a new version of the MG, the MetaGeneAnnotator (MGA). The MGA has statistical models of prophage genes and can automatically detect them in addition to chromosome backbone genes even when input genomic sequences have mosaic structures attributed to lateral gene transfers and/or phage infections. The MGA also has an adaptable the RBS model based on complementary sequences of the 3′ tail of 16S ribosomal RNA, and precisely predicts translation starts of genes even when input genomic sequences are short and anonymous sequences. These features of the MGA remarkably improve prediction accuracies of genes on a wide range of prokaryotic genomes. Here, we report the results of a performance test of the MGA applied to various types of genomic sequences, such as complete genomes, plasmids and their subsequences of various lengths, under conditions of anonymity.

2. Materials and methods

2.1. Construction of prophage gene model

In addition to the bacterial and archaeal gene models of MG[23], prophage models were constructed as follows. Genomic sequences and their annotations for 439 phages were obtained from the RefSeq database[28] (release 27). As a preprocessing, a monocodon usage was calculated from each phage genome, and the Euclidean distances of all pairs of the codon usages were calculated. When the distance between two phages' usages was <0.02, one of them was removed from the dataset because they might have been related (or identical) phages. Then, the codon frequencies of the remaining 244 phages were plotted against their GC contents, and we confirmed that the codon frequencies of phage were highly correlated with their GC contents, as seen in bacterial and archaeal genomes. For gene prediction, the MGA used di-codon frequencies that represent conditional probabilities of codon occurrences providing a previous codon (61×61 frequencies). Because each phage did not have enough genes to calculate di-codon frequencies, phage genomes having about the same range of GC contents were treated as a unit, and then di-codon frequencies were calculated from all genes annotated in the grouped genomes. Finally, a logistic regression analysis was performed in the same manner as the prokaryotic di-codon model construction in MG.

2.2. Procedures for predicting typical and atypical genes

The self-training model for typical genes is constructed as follows. Initially, genes are predicted using an optimal set of the di-codon regression models (bacterial, archaeal or prophage models). Then, these predicted genes are used for the self-training of the di-codon statistics of typical genes. The self-training model is defined as the weighted averages of the di-codon frequencies derived from the predicted genes, and from the regression models used for the initial prediction. A di-codon frequency of the self-training model, f_{self}, is defined by the frequency of the predicted genes, f_{pred}, and of the regression models, f_{reg}, as follows:

$$f_{\mathrm{self}}(a|b) = \frac{k \times f_{\mathrm{reg}}(a|b) + l \times f_{\mathrm{pred}}(a+b)}{k+l}, \qquad (1)$$

where a and b are codons, k and l are the numbers of di-codons used to calculate f_{reg} and f_{pred}, respectively. In the MGA, k was heuristically set to 30 000, which corresponds approximately to the number of di-codons in a 100 kb genomic sequence. The value

was determined by testing prediction performances on the training data of MG[23] and was meant to be enough to avoid overfitting of the self-training model to a few genes on a short input sequence. If a significant number of di-codons are extracted from the predicted genes, the self-training model is nearly equal to the di-codon frequencies of the predicted genes. If not, the self-training model comes closer to the di-codon frequencies derived from the regression models.

After training, four sets (self, bacteria, archaea and prophage) of di-codon frequencies are applied for scoring candidate genes. Unlike the original MG algorithm, each open-reading frame (ORF) is individually scored according to its own GC content in this step to detect atypical genes. Typical genes are expected to score the highest mark with the self-training model, and atypical genes to score the highest mark with one of the other models. Then, a maximal scoring combination of genes is calculated as the definitive prediction. While this procedure (ORF-by-ORF) is sensitive to atypical genes, some more false-positives are included in the prediction. So, the ORF-by-ORF procedure is applied only to the sequences longer than 5000 bp (containing multiple genes). For shorter sequences, the conventional procedure, in which all ORFs are scored by one of the four sets of the di-codon models according to the GC content of the input sequence, is applied.

2.3. *The RBS map analysis and the RBS model construction*

We defined nine hexamers derived from the following sequence, which was complementary to a tail of 16S rRNA, as the potential RBS motifs: G(A/T)(A/T)AGGAGGT(G/A)ATC. Starting from the left, the motifs were named Motif-1, . . ., Motif-9 [e.g. Motif-3 is '(A/T)AGGAG']. An exact match or one-base mismatch sequence of the motifs was sought against an upstream region of a start codon, and the best match motif and location were determined. In the RBS map analysis (see below), upstream sequences of the annotated start codons range from -2 to -21 were used for analysis. In the RBS prediction model, upstream sequences of the predicted start codons (in the previous step) range from -3 to -19 were used for model construction and prediction. The detected sequences were considered to be representative RBSs of the species, and the proportion of genes having representative RBSs (an RBS ratio, w_{RBS}) was stored for the use in scoring RBSs. Then, a two-dimensional frequency distribution of the representative RBSs was calculated to construct the RBS map. For the analysis, distances between the constructed RBS maps were defined by the Euclidean distance, and the neighbor joining method[29] was applied to make clusters of the RBS maps. This RBS map analysis was performed using 591 annotated microbial genomes obtained from the RefSeq database (Supplementary Table S1). As the RBS prediction model, a position weight matrix (PWM) for each motif was constructed using the representative RBS sequences detected earlier. In the prediction process, the RBS scores for all candidate genes were calculated using the constructed PWMs and the frequency distributions of the positions. Here, the RBS score, S_{RBS}, was heuristically weighted using a frequency of a motif m, w_m, and the RBS ratio (w_{RBS}) to reduce noise in less frequently used motifs.

$$S_{\mathrm{RBS}} = \max_{m,j}\left[w_{\mathrm{RBS}} \times w_m \times \sum_{i=1}^{6} \log_2 \frac{p_m(x_{i,j})}{q(x_{i,j})}\right], \qquad (2)$$

where $x_{i,j}$ is an ith nucleotide of a hexamer j appeared in an upstream of a gene, $p_m(x_{i,j})$ is a frequency of $x_{i,j}$ at a position i of a PWM for a motif m, and $q(x_{i,j})$ is a background frequency of $x_{i,j}$ calculated from a GC content of an input sequence. A value of w_m was standardized, and was 1 when a motif m was the most frequently used. For each gene, the best motif, which marked the highest RBS score, was selected, and the RBS score was added to the score of the genes. Then, the optimal combination of genes with the recalculated scores was estimated by the dynamic programming procedure used in MG[23]. All of these steps were iterated until the prediction results stopped changing.

2.4. *Performance evaluation*

Prediction performances of gene-finders were evaluated using datasets, including the MetaGene dataset[23]. The MetaGene dataset consists of nine bacterial and three archaeal genomic sequences (Supplementary Table S2). In addition to these complete sequences, their subsequences (1 Mb, 500, 100, 40, 10, 5, 3 and 1 kb, 700 and 100 bp sequences) having $1\times$ genome coverage (i.e. the total length of the subsequences is equal to the complete genome size) were also used for the evaluation. These sequences were not used for constructing statistical models of the MGA. The ratios of true-positives, including partially matching predictions with correct reading frames, relative to all annotated genes (sensitivity) and to all predicted genes (specificity) were used as indices for the evaluation. In addition, sensitivity to the start codons, in which only exactly matching predictions were counted as true-positives, was also utilized.

3. Results and discussion

3.1. Predicting prophage genes

The MGA is based on the algorithm of the MG and utilizes logistic regression models of the GC content and the di-codon frequencies[23] (di-codon models). In addition to the bacterial and archaeal di-codon models of MG, prophage models are constructed and integrated into the MGA (Fig. 1A). Although the proportions of prophage genes in the prokaryotic genomes are ordinarily not so large, they usually have biologically important functions, such as pathogenicity and niche adaptation, in the organisms. Therefore, detecting prophage genes is fundamental to understanding the genetic background of an organism.

Because most prophage genes have codon frequencies similar to those of bacteria and archaea, MG (and probably other prokaryotic gene finders as well) can predict prophage genes with relatively high accuracies (Supplementary Table S3). However, Fig. 2A and B shows that certain other (non-codon) properties of prophage genes are different from those of prokaryotic genes: prophage genes are generally shorter (~660 bp in average) than bacterial and archaeal genes (~940 bp in average), and most genes are organized in tandem (>90%). This means that gene densities are higher in prophage genomes than in prokaryotic genomes, and most genes are packed in a few operons. These observations and statistics, in addition to the prophage di-codon models, are utilized to predict prophage genes. As a result, the sensitivities of the MGA to prophage genes are remarkably improved (from 88 to 93%) without any decrease in specificity (90%) (Supplementary Table S3).

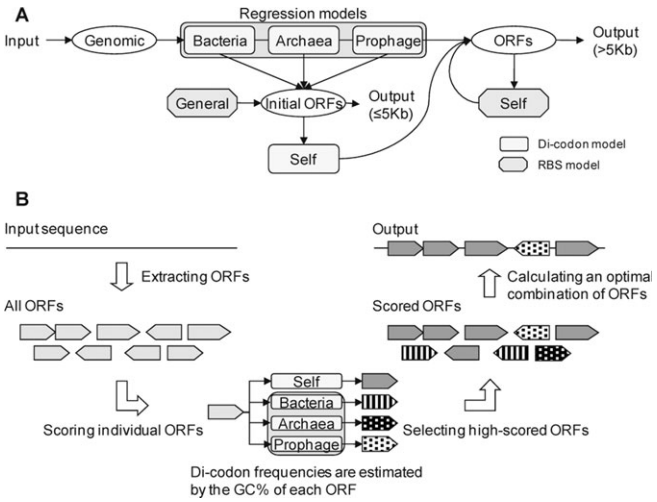

Figure 1. A schematic diagram of the MGA algorithm. (**A**) Prediction protocol of the MGA. (**B**) ORF-by-ORF procedure.

3.2. Predicting atypical genes

MG predicts genes using the di-codon frequencies (and other parameters) estimated by the GC content of an input genomic sequence. That is to say, all genes in the same genomic sequence are predicted by the same set of di-codon frequencies. In this procedure, typical genes can be accurately and specifically predicted, but atypical genes, such as horizontally transferred and prophage genes, cannot be detected because their di-codon frequencies are different from those of typical genes. To overcome this limitation, we employ an ORF-by-ORF procedure, in which each candidate ORF is treated as an individual anonymous sequence (Fig. 1B). This procedure assumes that every ORF has a potentially different origin and contributes to improving the sensitivities of the MGA to atypical genes.

To predict properly the typical genes under the ORF-by-ORF procedure, we arranged a self-training model of di-codon frequencies in addition to the logistic regression models (Fig. 1A). In the self-training model, di-codon frequencies are calculated from the initially predicted genes using the conventional scoring procedure of the MG, and then the weighted averages of di-codon frequencies derived from the predicted genes and from the regression models are calculated as the di-codon frequencies of typical genes. The self-training model fits well to typical

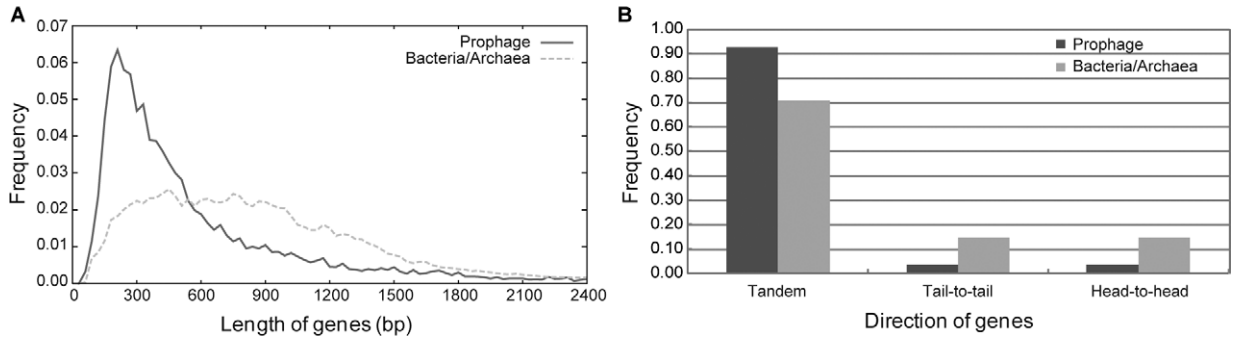

Figure 2. Statistics of prophage genes. (**A**) Frequency distributions of gene lengths in prokaryote and prophage. (**B**) Proportions of the consecutive gene arrangements in prokaryote and prophage.

genes compared with the regression models, and improves both sensitivity and specificity of the MGA to typical genes.

To evaluate the effectiveness of these procedures, prediction performances were tested on the chromosome and plasmid of enterohemorrhagic *Escherichia coli* O157:H7 strain Sakai[30,31] (Supplementary Table S4). Sensitivities of the MGA are extremely higher than those of the MG, especially in S-loops, which are O157:H7 strain-specific regions identified from comparisons with the *E. coli* K12 genome and that contain many horizontally acquired virulence-related genes. Higher sensitivities are also observed for a large virulence plasmid (pO157). Specificities of the MGA are slightly lower than those of MG, but are still higher than those of GeneMarkS[16] and GeneMark.heuristics[32]. These results indicate that our ORF-by-ORF procedure works well for predicting atypical genes and can be applied to genomes having mosaic structures with high specificity.

3.3. *Analyzing species-specific patterns of the RBS*

The other notable feature of the MGA is an adaptable model of the RBS. An RBS, which is also known as the Shine-Dalgarno (SD) sequence[33], is located on the 5′ flanking region of the start codon, and interacts with a part of the 3′ end of 16S ribosomal RNA (rRNA) to control translation initiations of the gene. Although RBSs are complementary to the 3′ tail of the 16S rRNA in every organism, their sequences (motifs) and preferred locations relative to start codons (or 'spacer' lengths) differ slightly from organism to organism. In gene-finding programs, the Gibbs sampling algorithm is widely used for training the motifs and the spacer length distribution of the target species' RBSs[16,17], although this algorithm takes no thought for the observation that the RBSs are complementary to the tail of the 16S rRNA. This approach basically assumes one motif and one frequency distribution of the spacer lengths in

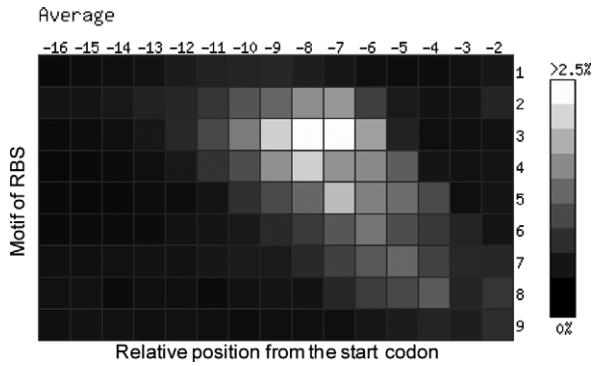

Figure 3. The average RBS map. The horizontal axis represents relative positions from the start codons [equal to −(spacer length+1)], and the vertical axis represents motif numbers.

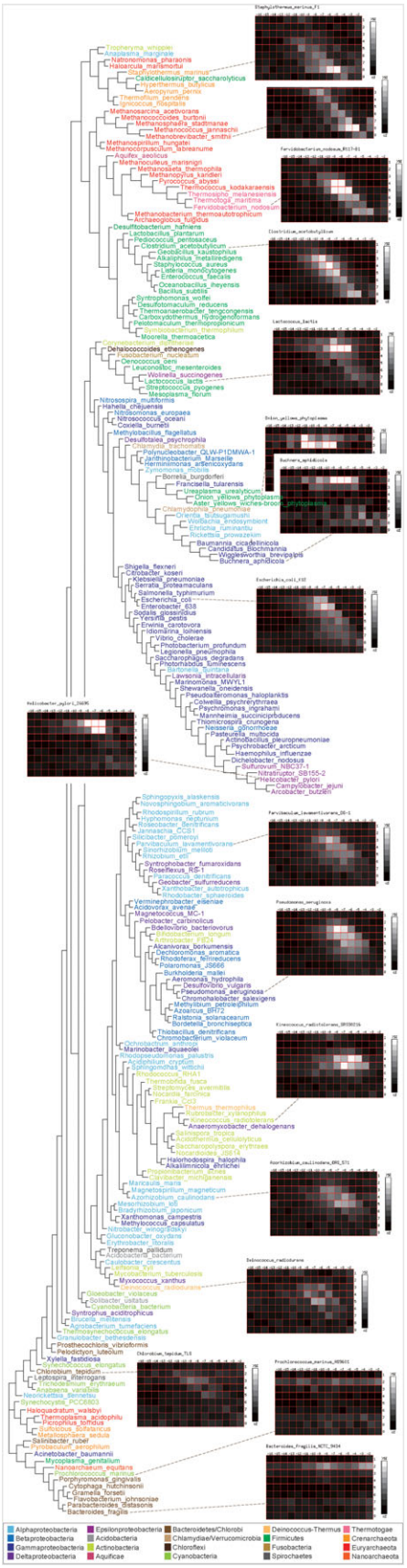

Figure 4. The clustering result of the RBS maps derived from 229 of 591 prokaryotic genomes (one species per genus).

each species. However, our analysis suggests that this assumption is not appropriate for most species.

We examined the upstream sequences of annotated genes from 229 prokaryotic genomes and constructed RBS maps that show a two-dimensional frequency distribution of the best match motif (out of the nine candidate motifs we suggested) and the spacer lengths of the RBSs for each species. The average RBS map (Fig. 3) shows that Motif-3 is most frequently used, but all nine motifs are potential RBSs. The higher the motif number, the shorter the spacer lengths. This is reasonable because it means that the position of the main body of the 16S rRNA is fixed even if the hybridization position of 16S rRNA tail is moved.

The observed patterns of the RBS maps vary from organism to organism, while phylogenetically related species show similar patterns (Figs 4 and 5). Although some species such as *Helicobacter pylori* (Fig. 5A) and *Buchnera aphidicola*, predominantly use Motif-2 and -3 and are therefore congruous with the one motif assumption described earlier many other species show broader distributions. For example, some *Firmicutes*, including *Clostridium* (Fig. 5B), and *Thermotogae* indicate broad and clear patterns of the RBS maps. Some archaea, including methanogens (Fig. 5C), also indicate broad patterns, but the preferred motifs are different between these bacteria and archaea (e.g. *Clostridium acetobutylicum* prefers Motif-3 and -4, but *Methanobrevibacter smithii* prefers Motif-8.). Overall, bacterial species tend to prefer motifs of 3′ side of a tail of 16S rRNA, while archaeal ones tend to prefer motifs of 5′ side of the tail. Only very weak signals of the RBS motifs are found in some species belonging to *Bacteroidetes* and *Cyanobacteria* (Fig. 5D). In these species, no other significant motif is found. These results suggest that our RBS map with nine fixed motifs is effective for capturing the species-specific pattern of the RBSs. Hence, we used this two-dimensional frequency distribution and the PWMs of the nine RBS motifs as an RBS model of the MGA. Parameters of the RBS model are estimated from upstream sequences of predicted genes. To predict the RBSs on very short input sequences (having no training data), a general model of the RBS was manually constructed, based on the average RBS map and was integrated into the MGA (Fig. 1A).

3.4. *Prediction performances on long genomic sequences*

The prediction performances of the MGA and conventional gene-finding tools based on unsupervised learning, such as GeneMarkS[16] and Glimmer3[17],

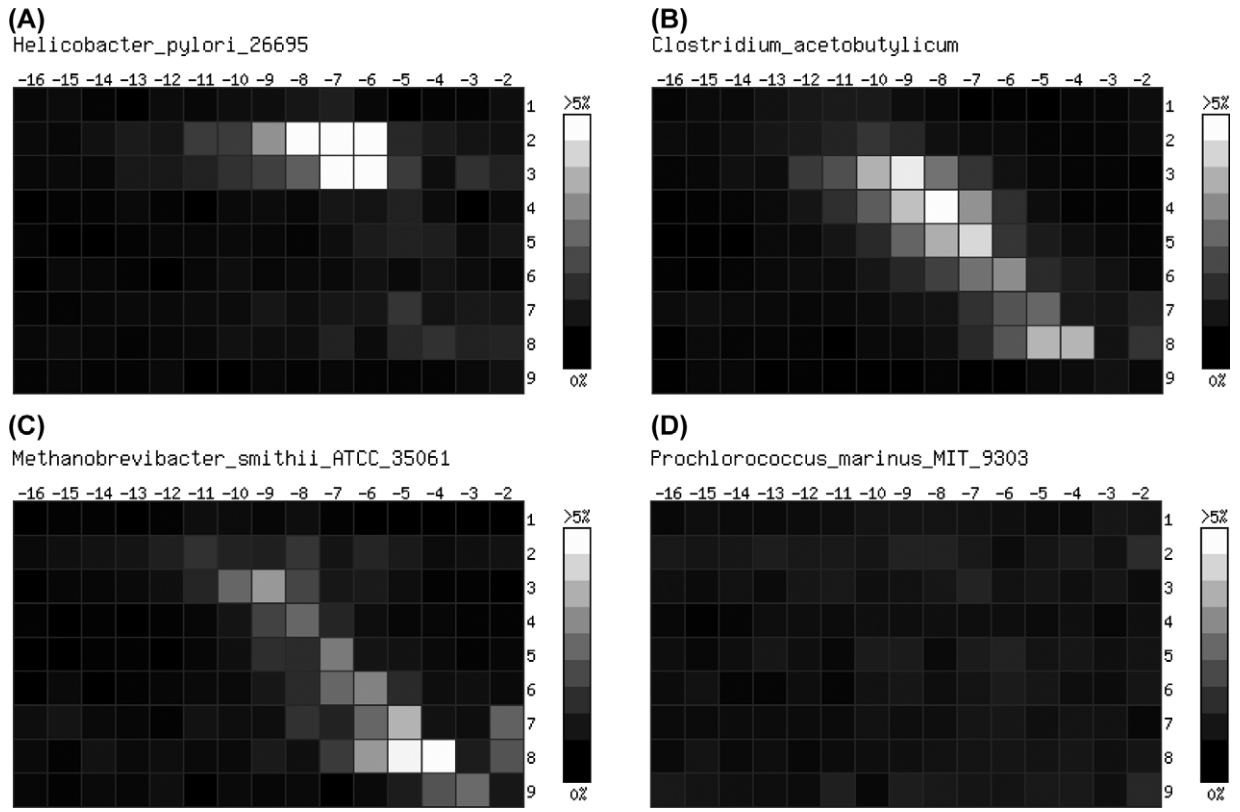

Figure 5. The RBS maps for four species. (**A**) *Helicobacter pylori* (**B**) *Clostridium acetobutylicum* (**C**) *Methanobrevibacter smithii* (**D**) *Prochlorococcus marinus*.

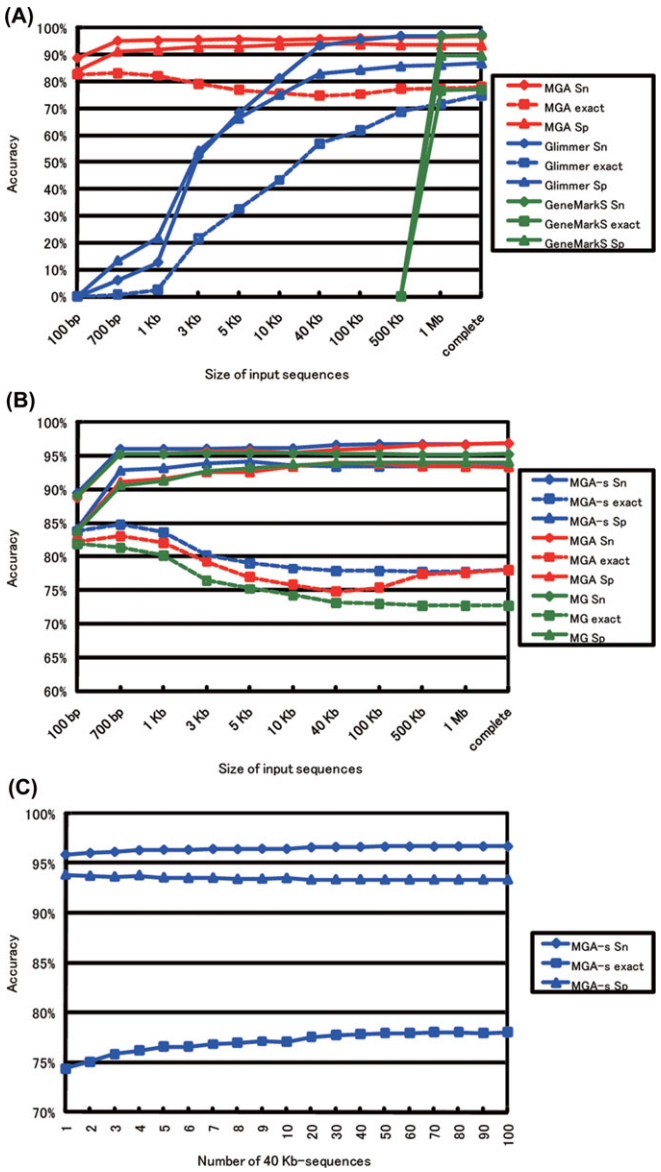

Figure 6. Prediction performances of gene finders on the MetaGene dataset. (**A**) Accuracy comparisons of the MGA, GeneMarkS and Glimmer3. In the Glimmer3 prediction, a script 'g3-iterated.csh' is used. (**B**) Accuracy comparisons of the MGA and MG. In the MGA prediction, two different running options, which treat multiple input sequences individually (MGA) or as a unit (MGA-s), are used. (**C**) Relationship between accuracies and number of 40 kb-sequences in the MGA-s prediction. Sn, exact and Sp indicate sensitivity, sensitivity to start codons and specificity, respectively.

were evaluated on various datasets. Fig. 6A shows the prediction accuracies on the MetaGene dataset, which consists of nine bacterial and three archaeal genomic sequences (Supplementary Table S2) and their subsequences having 1× genome coverage (i.e. the total length of the subsequences is equal to the complete genome size).

For complete genomes and 1 Mb subsequences, all prediction tools indicate almost identical sensitivities (~97%), while specificity is significantly higher in the MGA (93%) compared with the others (90% in GeneMarkS and 86–87% in Glimmer3). In other words, the sensitivities of the MGA are potentially higher than the others at the same specificity level. Sensitivities to start codons are also identical in the MGA (78%) and GeneMarkS (77%), but Glimmer3 shows lower values (72–75%), although both GeneMarkS and Glimmer3 utilize the Gibbs sampling procedure to train their RBS models. In contrast, the mean sensitivity to start codons in Glimmer3 is better than that in GeneMarkS on the other dataset (Table 1), which consists of six complete genomes (one archaea and five bacteria) having relatively broad distributions of the RBS maps. The performance of the MGA is stable and exceeds the others also in this dataset, especially in *Clostridium acetobutylicum*. In comparison with the original MG (Fig. 6B), the MGA remarkably improves sensitivities to both genes and start codons without reducing specificities. These results indicate that our simple RBS model works well for detecting various types of the RBS.

3.5. *Prediction performances on short genomic sequences*

For sequences shorter than 1 Mb, the MGA retained high accuracies in every index (Fig. 6A). Both sensitivities and specificities of Glimmer3 are relatively high when input sequences are longer than 40 kb, but the performance of the start codon prediction is rapidly degraded as the input sequences become shorter. This is because the Gibbs sampling algorithm requires a significant number of positive (RBS) sequences to detect the correct motif. A 40 kb-sequence has <40 genes (or RBSs) on average, and the sensitivity to start codons declines to 57% in Glimmer3. GeneMarkS does not accept a shorter input sequence than 1 Mb, probably because it has the same weak point as Glimmer3. Unlike the RBS models of these tools, the MGA assumes only nine hexamers as candidate's RBSs, and relatively few sequences are needed to estimate the parameters of the RBS model. As a result, the MGA requires just 500 kb (or ~500 genes) to adapt the RBS model fully to the input sequence, and its sensitivity to start codons is sufficiently high (75%) even in 40 kb sequences. Furthermore, Fig. 6B and Table 2 show that the general RBS model of the MGA also works well for predicting the start codons of genes on very short sequences. Although most genes on 700 bp-subsequences lack their 5′ sequences (including start codon and RBSs), the results also indicate that the RBS model contributes to improving prediction specificities by deselecting false-positive translation starts.

Table 1. Prediction performances on the complete genomes

Species	GC%	RBS%	MGA		GeneMarkS		Glimmer3	
			Sn (exact) (%)	Sp (%)	Sn (exact) (%)	Sp (%)	Sn (exact) (%)	Sp (%)
S. marinus	35.7	85.4	99.4 (87.8)	94.5	99.6 (87.2)	92.5	99.8 (87.6)	90.8
C. acetobutylicum	30.9	93.7	98.3 (92.1)	96.1	98.5 (74.1)	92.8	98.0 (90.9)	94.5
F. nodosum	35.0	90.2	99.6 (91.2)	94.8	99.8 (90.6)	92.8	99.7 (91.1)	94.0
L. lactis	35.3	81.1	98.5 (88.0)	95.1	98.9 (88.4)	92.7	98.2 (86.2)	93.2
D. radiodurans	67.0	47.9	97.8 (63.5)	93.6	96.3 (56.7)	93.1	96.5 (58.3)	92.1
A. caulinodans	67.3	64.8	99.2 (66.2)	95.4	98.8 (61.5)	95.8	98.6 (63.6)	93.6
Average			98.7 (80.2)	95.0	98.5 (74.3)	93.4	98.2 (78.0)	93.5

RBS%, the RBS ratio (the proportion of genes having representative RBSs); Sn, sensitivity to genes; (exact), sensitivity to start codons; Sp, specificity.

Table 2. Prediction performances on 700 bp subsequences (1 × genome coverage)

Species	GC%	RBS%	MGA-s		MGA		MG	
			Sn (exact) (%)	Sp (%)	Sn (exact) (%)	Sp (%)	Sn (exact) (%)	Sp (%)
M. jannaschii	31.4	87.6	98.3 (79.3)	95.8	97.7 (80.3)	94.1	97.8 (82.4)	94.1
A. fulgidus	48.6	61.7	96.7 (82.9)	94.1	95.7 (81.7)	93.5	95.8 (81.5)	93.7
N. pharaonis	63.4	39.6	97.4 (88.3)	97.1	97.1 (87.1)	94.5	97.1 (86.2)	93.0
B. aphidicola	26.3	60.9	98.4 (91.5)	93.6	98.6 (91.7)	93.2	98.2 (90.9)	92.7
P. marinus	31.2	21.0	95.2 (88.8)	93.0	94.9 (87.4)	92.3	95.5 (87.6)	92.7
W. endosymbiont	34.2	40.1	93.8 (85.8)	74.5	93.6 (82.9)	72.7	93.1 (80.8)	76.0
H. pylori	39.2	78.3	96.8 (88.1)	95.1	93.5 (82.9)	92.4	92.6 (77.7)	92.7
B. subtillis	43.5	92.6	97.3 (88.8)	94.5	93.9 (82.2)	92.9	92.3 (73.5)	92.5
E. coli	50.8	77.6	96.4 (83.5)	94.6	95.0 (82.9)	94.0	95.3 (81.2)	93.2
C. tepidum	56.5	45.4	88.8 (75.3)	93.5	87.6 (73.7)	90.6	88.1 (73.2)	89.6
C. jeikeium	61.4	72.8	95.7 (83.9)	95.1	94.9 (82.8)	93.3	94.0 (78.5)	91.4
B. pseudomallei 1	67.7	56.2	96.6 (83.1)	93.7	96.9 (82.9)	90.6	96.8 (81.2)	87.9
B. pseudomallei 2	68.5	56.3	96.2 (83.9)	91.6	96.4 (82.8)	89.0	96.6 (81.2)	85.7
			96.0 (84.9)	92.8	95.1 (83.2)	91.0	94.9 (81.2)	90.4

RBS%, the RBS ratio (the proportion of genes having representative RBSs); Sn, sensitivity to genes; (exact), sensitivity to start codons; Sp, specificity; MGA-s, MGA with −s option in which multiple sequences are treated as a unit.

3.6. *Advantage of self-training using a set of genomic sequences*

If multiple (short) input sequences can be assumed as the genomic sequences of the same species, prediction accuracies on the sequences are improved by self-training of the models as well as on a long-genomic sequence (the MGA-s in Fig. 6B and C). Fig. 6C shows the relationships between prediction accuracies and the number of 40 kb-sequences treated as a unit. Fig. 6C also suggests that a total of about 500 kb (10−20×40 kb) are needed for full adaptation of the RBS model, but the prediction accuracies steadily improve if the number of input sequences are increased. When a sufficient amount of sequences are available, the MGA provides prediction performances comparable to the complete genome analyses, even if each sequence is very short (Table 2). So, if multiple contig sequences are obtained by sequencing a single species' genome, or if metagenomic sequences are phylogenetically classified into groups using some classification methods[34,35], genes on the sequences can be more precisely predicted by the MGA.

3.7. *Conclusion*

As mentioned, the MGA successfully overcomes the limitations of the MG, and archives high prediction accuracies especially in the start codon predictions. Although some gene-finding tools advocating high sensitivity to start codons, such as GeneMarkS and Glimmer3 tend to sacrifice specificities for improving sensitivities, the RBS model of the MGA enables the sensitive detection of start codons without reducing specificities. Our RBS model is based on previous knowledge about the RBS and 16S rRNA, and requires

little training data for estimating its parameters. As a result, the MGA can precisely predict genes even on short genomic sequences unlike the other tools. Both typical and atypical genes can be sensitively and precisely detected while keeping high specificity. The MGA can detect not only chromosome backbone genes but also prophage genes and provides a complete set of genes on a genomic sequence. The MGA also provides information about the selected dicodon model (bacteria, archaea, prophage or self) for predicting each gene, and the information is helpful for further analyses of genes because it reflects statistical differences among the genes.

In addition to the precise prediction ability of the MGA, the RBS map analysis proposed here is helpful for genome annotations and is useful for analyzing translation initiation mechanisms and their evolutions. It is important for annotators to comprehend a specific RBS pattern of a target species and its related species. The MGA can automatically extract the pattern, and outputs information on RBSs in addition to location information on genes. We believe that the MGA accelerates not only metagenomic analyses but also the annotation processes of all kinds of prokaryotic and phage genomes.

Availability

MetaGeneAnnotator are freely available at http://metagene.cb.k.u-tokyo.ac.jp.

Acknowledgements: The original MetaGene was developed at Toshihisa Takagi laboratory (University of Tokyo). We thank Prof Tetsuya Hayashi (Miyazaki University) and Prof Ken Kurokawa (Tokyo Institute of Technology) for stimulating discussions.

Supplementary Data: Supplementary Data is available online at www.dnaresearch.oxfordjournals.org.

References

1. Fickett, J. W. 1981, Recognition of protein coding regions in DNA sequences, *Nucleic Acids Res.*, **10**, 5303–5318.
2. Gribskov, M., Devereux, J. and Burgess, R. R. 1984, The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression, *Nucleic Acids Res.*, **12**, 539–549.
3. Staden, R. 1984, Measurements of the effects of that coding for a protein has on a DNA sequence and their use for finding genes, *Nucleic Acids Res.*, **12**, 551–567.
4. Borodovsky, M. Y., Sprizhitskii, Y. A., Golovanov, E. I. and Aleksandrov, A. A. 1986, Statistical patterns in primary structures of functional regions in the *E. coli* genome: III. Computer recognition of coding regions, *Mol. Biol.*, **20**, 1145–1150.
5. Borodovsky, M. Y. and McIninch, J. D. 1993, GeneMark: parallel gene recognition for both DNA strands, *Comput. Chem.*, **17**, 123–153.
6. Krogh, A., Mian, I. S. and Haussler, D. 1994, A hidden Markov model that finds genes in *E.coli* DNA, *Nucleic Acids Res.*, **22**, 4768–4778.
7. Salzberg, S. L., Delcher, A. L., Kasif, S. and White, O. 1998, Microbial gene identification using interpolated Markov model, *Nucleic Acids Res.*, **26**, 544–548.
8. Lukashin, A. V. and Borodovsky, M. 1998, GeneMark.hmm: new solutions for gene finding, *Nucleic Acids Res.*, **26**, 1107–1115.
9. Delcher, A. L., Harmon, D., Kasif, S., White, O. and Salzberg, S. L. 1999, Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, **27**, 4636–4641.
10. Yada, T., Nakao, M., Totoki, Y. and Nakai, K. 1999, Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov model, *Bioinformatics*, **15**, 987–993.
11. Yada, T., Totoki, Y., Takagi, T. and Nakai, K. 2001, A novel bacterial gene-finding system with improved accuracy in locating start codons, *DNA Res.*, **8**, 97–106.
12. Hayes, W. S. and Borodovsky, M. 1998, How to interpret an anonymous bacterial genome: machine learning approach to gene identification, *Genome Res.*, **8**, 1154–1171.
13. Audic, S. and Claverie, J. M. 1998, Self-identification of protein-coding regions in microbial genomes, *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 10026–10031.
14. Baldi, P. 2000, On the convergence of a clustering algorithm for protein-coding regions in microbial genomes, *Bioinformatics*, **16**, 367–371.
15. Besemer, J., Lomsadze, A. and Borodovsky, M. 2001, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes, *Nucleic Acids Res.*, **29**, 2607–2618.
16. Delcher, A. L., Bratke, K. A., Powers, E. C. and Salzberg, S. L. 2007, Identifying bacterial genes and endosymbiont DNA with Glimmer, *Bioinformatics*, **23**, 673–679.
17. Schuster, S. C. 2008, Next-generation sequencing transforms today's biology, *Nat. Methods*, **5**, 16–18.
18. Hall, N. 2007, Advanced sequencing technologies and their wider impact in microbiology, *J. Exp. Biol.*, **210**, 1518–1525.
19. Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. 2007, SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing, *Genome Res.*, **17**, 1697.
20. Hernandez, D., François, P., Farinelli, P., Østerås, M. and Schrenzel, J. 2008, De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer, *Genome Res.*, **18**, 802.
21. Butler, J., MacCallum, I., Kleber, M., et al. 2008, ALLPATHS: de novo assembly of whole-genome shotgun microreads, *Genome Res.*, **18**, 810.
22. Zerbino, B. R. and Birney, E. 2008, Velvet: algorithms for de novo short read assembly using de Bruijn graph, *Genome Res.*, **18**, 821.

23. Noguchi, H., Park, J. and Takagi, T. 2006, MetaGene: prokaryotic gene finding from environmental genome shotgun sequences, *Nucleic Acids Res.*, **34**, 5623–5630.
24. Kurokawa, K., Itoh, T., Kuwahara, T., et al. 2007, Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes, *DNA Res.*, **14**, 169–181.
25. Raes, J., Foerstner, K. U. and Bork, P. 2007, Get the most out of your metagenome: computational analysis of environmental sequence data, *Curr. Opin. Microbiol.*, **10**, 490–498.
26. Schmeisser, C., Steele, H. and Streit, W. R. 2007, Metagenomics, biotechnology with non-culturable microbes, *Appl. Microbiol. Biotechnol.*, **75**, 955–962.
27. Pop, M. and Salzberg, S. L. 2008, Bioinformatics challenges of new sequencing technology, *Trends Genet.*, **24**, 142–149.
28. Pruitt, K. D., Tatusova, T. and Maglott, D. R. 2007, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, **35**, D61–65.
29. Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**, 406–425.
30. Hayashi, T., Makino, K., Ohnishi, M., et al. 2001, Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12, *DNA Res.*, **8**, 11–22.
31. Ohnishi, M., Kurokawa, K. and Hayashi, T. 2001, Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors?, *Trens Microbiol.*, **9**, 481–485.
32. Besemer, J. and Borodovsky, M. 1999, Heuristic approach to deriving models for gene finding, *Nucleic Acids Res.*, **27**, 3911–3920.
33. Shine, J. and Dalgarno, L. 1974, The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: Complementary to nonsense triplets and ribosome binding sites, *Proc. Natl. Acad. Sci. U. S. A.*, **71**, 1342–1346.
34. McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. and Rigoutsos, I. 2007, Accurate phylogenetic classification of variable-length DNA fragments, *Nat. Methods*, **4**, 63–72.
35. Krause, L., Diaz, N. N., Goesmann, A., et al. 2008, Phylogenetic classification of short environmental DNA fragments, *Nucleic Acids Res.*, **36**, 2230–2239.