GeneMark-RC, a Recursive Procedure for Gene Identification in the Genomic Sequence Data with Self-Consistency Evaluation; Its Application to the Analysis of Several Prokaryotic Genomes

Makoto Hirosawa [1] **Katsumi Isono** [2]

hirosawa@kazusa.or.jp isono@biol.kobe-u.ac.jp

[1] Kazusa DNA Research Institute
1532-3 Yana, Kisarazu-shi, Chiba 292, Japan

[2] Department of Biology, Faculty of Science, Kobe University
1-1 Rokkodai, Nada-ku, Kobe 657, Japan

Abstract

Previously, we developed a GeneMark-based procedure, termed GeneMark-RC, and applied it for the identification and classification of ORFs in genomic sequence data, and identified and characterized ORFs in the 1.0 Mb data of the cyanobacterium Synechocystis *sp. strain PCC 6803. In the present study, we have improved the procedure and performed analysis of the whole genomic data of* Synechocystis. *Consequently, we noticed the presence of three distinct classes of ORFs in this organism. The prediction of ORFs by the class-specific GeneMark-RC analysis agreed with 97.9 % of those described for this bacterium. Moreover, 124 additional ORFs were identified. The procedure was similarly applied to the genomic analysis of five other prokaryotes, and 2 to 3 classes of ORFs were recognized in each case. Common features were found among the ORFs identified in the six organisms including* Synechocystis. *Class 1 is composed of most typical ORFs whose GC content is slightly higher than the average, while Class 2 is composed of ORFs with GC contents lower than the average. It was found that ORFs of one species can be detected with the GeneMark-RC parameters obtained from other organisms, and the prediction rate is high when the difference in their GC contents is small. It was also found that ORFs of three species with relatively low GC contents can be nicely detected with the* Synechocystis *matrices of Class 2 ORFs whose GC content is similar to that of the three species. Therefore, although there are two to three classes of ORFs in each species, their di-codon statistics must be rather similar to each other if their GC contents are similar. A notable exception was the case of* Methanococcus jannaschii*, which might reflect the fact that it is an archaebacterium.*

1 Introduction

The advancement in large-scale sequencing has accelerated the production of long contiguous nucleotide sequence data. The genomic sequence data along with the detailed annotations concerning the locations and features of genes/ORFs are currently available for six prokaryotic organisms, namely, *Haemophilus influenzae Rd.* [1], *Mycoplasma pneumoniae* [2], Cyanobacterium *Synechocystis* sp. strain PCC 6803 [3], *Methanococcus jannaschii* [4], *Escherichia coli* [5] and *Mycoplasma pneumoniae* [6]. Strategies for ORF assignment are different from one organism to another, but there are often cases in which ORFs were assigned just because their length exceeds certain threshold values. Also, there are ORFs that were assigned because of their similarity to the data for other organisms stored in the nucleotide and/or protein sequence databases. However, such data themselves might have been assigned because of their length or because of the existence of similar sequences within the databases. Thus, a vicious cycle of ORF assignment will result.

One way to break such a cycle will be to introduce a gene-finding procedure with a satisfactory level of *self-consistency.* We propose here a *self-consistent* set of ORFs to be defined as a set, if, when the statistical parameters derived from it are used in conjunction with a gene-finding procedure such as GeneMark [7] [8], the procedure predicts the set itself. Starting with an initial set of ORFs, a self-consistent set of ORFs can be derived by recursive applications of the gene-finding procedure. When the result of such recursive applications converges, the set of ORFs obtained is said to be self-consistent. In our previous study [9], we introduced a procedure termed GeneMark-RC to find self-consistent sets of ORFs by using GeneMark as a gene-finding tool, although we did not explicitly mention the notion of self consistency. We investigated the effectiveness of the procedure by assigning ORFs in the 1.0 Mb nucleotide sequence data of *Synechocystis*, and showed that GeneMark-RC could also be used for deriving classes of ORFs, which differ in their characteristics in the nucleotide permutation statistics [9]. Here, we propose a gene finding procedure in which we explicitly defined and explored the concept of self-consistency. Once a self-consistent set of ORFs is established for each species by the procedure, additional ORFs will be found by similarity search among species. In this way, more likely protein-coding regions within a genomic sequence data can be assigned.

In the case of *Escherichia coli*, ORFs have been divided into three classes using a clustering method termed *correspondence analysis* [10]. Borodovsky *et al.* [11] showed that the accuracy of prediction by GeneMark was enhanced by applying specific Markov models for the three classes. It would be better if the statistics used for the classification of genes and for the gene-finding are the same, but in this procedure they are different. In our previous study [9], two classes could be differentiated by GeneMark-RC. Two groups of matrices representing respective classes and another set of matrices representing the organism were used to assign ORFs. An apparent prediction accuracy of 98.3 % was achieved by using the three sets of matrices. Unlike ORF prediction based on the classes derived from the correspondence analysis [10], our strategy was to use GeneMark exclusively.

In this paper, we have modified and improved GeneMark-RC to find *extended* self-consistent set of ORFs (see the below). The procedure was accordingly applied to the whole genomic sequence of *Synechocystis* to classify ORFs, in which ORFs were predicted by more than one set of matrices based on the classification. Furthermore, feasibility of the classification of ORFs in five other prokaryotic genomic data was investigated. Finally, the significance of the derived classes was examined.

2 Materials and Methods

Methods to evaluate the self-consistency of a set of ORFs with GeneMark-RC, and its application to the derivation of classes different in their nucleotide statistics will be described below. The details of modification of GeneMark-RC will be described in the next section. The genomic sequences and their annotated ORFs of all species were down loaded through the WWW Microorganism Database of TIGR [12].

GeneMark-RC

Let us assume that a chosen set of ORFs in a contiguous nucleotide sequence S is designated as X, and that Y represents the rest of ORFs in S. Let GM(X, Y) stands for a couple of Markov models, representing statistical patterns in coding and non-coding regions, that have been derived from the training sets X and Y. Parameters for Markov models, transition probabilities, are derived from the three phase dependent oligomer statistics extracted from the protein-coding DNA sequences and phase-independent oligomer statistics from non coding sequences [5]. The order of the model - R - is a parameter of the ORF prediction procedure. GeneMark also uses a parameter T to define the decision making level. If the GeneMark prediction score determined for a given ORF is higher than T, then the ORF is included in the set X[k+1]. The T value is frequently chosen as 0.5. Higher values of T can be used to select more typical ORFs of the organism in question. In what follows, we assume that sequence sets X[k] and Y[k], k = 0,1,2,... are such that the union of X[k] and Y[k] for each particular k produces the whole continuous sequence S. In the GeneMark-RC procedure, the GeneMark program uses matrices GM(X[k], Y[k]) to predict a new (updated) set of ORFs X[k+1]. If, initially, X[0] is assumed to be a set of ORFs annotated as ORFs in the database sequence entries, then X[1] is a set of ORFs predicted by GeneMark using the matrices GM(X[0], Y[0]) and Y[1] is determined as a complement to X[1]. In the next step new matrices are produced from the sets X[1] and Y[1] to obtain new sets X[2] and Y[2]. The process is performed recursively until X[k+1] becomes identical to X[k], and the iteration of finding a stationary set X[k] (Y[k]) by recursive application of GeneMark is termed GeneMark-RC. The stationary set thus derived depends on X[0], T and R. In other words, GeneMark-RC is a procedure to find typical ORFs which mainly compose X[0]. Less typical ORFs are filtered out during the course of iteration and T controls the degree of typicalness in this filtration procedure. In the present study, R=5 was chosen which corresponds to the hexamer statistics in the GeneMark matrices.

Derivation of authentic ORFs of species

To derive authentic ORFs comprising the genome of an organism, a set of well documented ORFs are used as X[0] whenever available. In the present study, such a set of ORFs could be used for the analysis of all bacterial species, but other choices are also possible. For example, ORFs whose length are longer than 300 bp can be used instead. In our previous study with the 1.0 Mb *Synechocystis* genomic sequence, it was confirmed that identical sets of ORFs were derived by starting with either of the possible X[0]s as initial sets. We name this set as the Basic set. X[0] used for deriving the Basic set is named as I-set. In the iteration process, spurious ORFs are expected to be filtered out.

Derivation of ORF classes

While T=0.5 is usually used for deriving the Basic set which exhibits average characteristics of the species, higher value of T can be used for deriving *core classes*, which are expected to have more distinct characteristics of the organism. In this study, T=0.8 was used.

To derive Core-class 1, a set of ORFs from the Basic set with scores higher than T should be used as X[0]. Due to the use of higher T value (0.8), ORFs with typical oligonucleotide statistics of the species are selected in Core-class 1. In the derivation of Core-class 2, a set of ORFs in I-set excluding ORFs belonging to Core-class 1 is used as X[0]. Consequently, a self-consistent set of ORFs harboring non-typical statistics is derived. In this way, Core-classes n (n=2,3,4,....) are derived with a set of ORFs in I-set excluding ORFs belonging to Core-classes m (m = 2,3,..,n-1) as X[0]. The derivation of core ORFs is to continue unless GeneMark-RC fails to find a non-empty set of ORFs.

After deriving self-consistent core classes, each core class was then used for deriving matrices of the class, and ORFs of the class were predicted with the matrices thus obtained. At this stage, T=0.5 was used. The selection of R was decided according to the number of ORFs in the core. For example, R=4 and R=5 were considered to be appropriate for classes with 250 ORFs and 1,000 ORFs in the core, respectively, from our experiences with GeneMark analysis.

3 Results and Discussions

Detection of *Synechocystis* ORFs

The method described above was applied to the whole genomic sequence of *Synechocystis* with 3168 annotated ORFs. A Basic set with 2,888 ORFs was derived after four iterations of GeneMark-RC, and Core-class 1 with 2,059 ORFs was derived after seven iterations of GeneMark-RC. However, Core-class 2 was converged into that of Class 1 after nine iterations of GeneMark-RC. It means that GeneMark-

Table 1. Codon usages in Core-classes

Codon	C1	C2	C3	Codon	C1	C2	C3	Codon	C1	C2	C3	Codon	C1	C2	C3
TTT	2.72	3.82	3.36	TCT	0.69	1.46	1.02	TAT	1.53	2.59	1.88	TGT	0.59	0.63	0.78
TTC	1.05	0.95	1.25	TCC	1.69	1.07	1.53	TAC	1.25	0.94	1.27	TGC	0.37	0.31	0.47
TTA	2.43	3.73	2.66	TCA	0.22	1.02	0.51	TAA	0.11	0.14	0.19	TGA	0.05	0.06	0.10
TTG	3.13	2.02	2.97	TCG	0.37	0.37	0.45	TAG	0.10	0.07	0.13	TGG	1.51	1.48	2.04
CTT	0.82	1.58	1.29	CCT	0.88	1.20	1.17	CAT	1.13	1.23	1.35	CGT	1.03	0.94	1.13
CTC	1.51	0.97	1.31	CCC	2.79	1.30	2.58	CAC	0.75	0.52	0.79	CGC	1.39	0.67	1.08
CTA	1.35	1.55	1.53	CCA	0.69	1.01	1.05	CAA	3.48	3.60	3.58	CGA	0.46	0.65	0.61
CTG	2.24	1.13	1.82	CCG	0.89	0.46	0.83	CAG	2.29	1.58	2.16	CGG	1.58	0.53	1.13
ATT	3.95	4.83	3.72	ACT	1.22	1.93	1.51	AAT	2.26	3.96	2.53	AGT	1.41	1.79	1.49
ATC	1.80	1.62	1.78	ACC	2.89	1.66	2.26	AAC	1.53	1.56	1.45	AGC	1.03	1.00	0.99
ATA	0.24	1.41	0.50	ACA	0.50	1.27	0.82	AAA	2.71	4.09	2.58	AGA	0.28	0.98	0.51
ATG	1.75	1.34	1.62	ACG	0.80	0.58	0.80	AAG	1.18	1.42	1.28	AGG	0.42	0.55	0.49
GTT	1.45	2.26	1.86	GCT	1.93	2.20	1.96	GAT	3.16	3.85	3.02	GGT	2.01	1.97	2.15
GTC	1.11	0.99	1.14	GCC	4.39	1.93	3.27	GAC	1.90	1.53	1.54	GGC	2.48	1.48	2.13
GTA	0.99	1.29	0.98	GCA	0.91	1.44	1.26	GAA	4.57	4.69	3.73	GGA	1.16	1.77	1.35
GTG	3.31	1.25	2.15	GCG	1.74	0.79	1.38	GAG	1.53	1.58	1.69	GGG	1.93	1.11	1.71

C1, C2 and C3 stand for Core-class 1, 2 and 3 respectively.

RC failed to categorize Class 2 as a separate class, although Class 2 could be derived in the analysis of the 1.0 Mb data [9].

For this reason, we modified GeneMark-RC by re-defining the notion of self-consistency. In the modified GeneMark-RC procedure for deriving Core-class n, the ORFs included in Core-class m (m = 1,.., n-1) are excluded in making matrices for assigning ORFs in the next cycle of iteration. With GeneMark-RC with this re-definition of self-consistency, Core-class 2 composed of 319 ORFs and Core-class 3 with 184 ORFs were derived.

Of the annotated ORFs, 2,859 ORFs (90.2 %) were detected in the Basic set of ORFs. The number of false-positive ORFs was as low as 23. The failure in detecting 9.8 % of the annotated ORFs of this organism was interpreted to indicate that the genes of this bacterium would not belong to only one homogeneous class. This was clearly shown by the difference in the GC content of ORF classes. The GC content of the Basic set, and of Core-classes 1, 2 and 3 was 49.4, 50.4, 40.0 and 47.8 %, respectively. Compared with the average GC content of the annotated ORFs, 48.6 %, the average GC content of the Basic set was 0.8 % higher, and that of Core-class 1 was 1.8 % higher. Therefore, the GC content of more typical ORFs can be said to be higher than the average. The GC content of Core-class 2 was about 10 % lower than that of the Basic set. It suggests that a statistically different group of ORFs exists in this organism, and that at least matrices for Class 2 must be prepared to improve the prediction of such ORFs. Class 2 were detected in the 1.0 Mb *Synechocystis* data as well, where the GC content of the core of the class was 41.9 %, namely 1.9 % higher than the value obtained in the present study. However, a correlation was found between both cores, and we considered it likely that they would share the same set of statistical characteristics, because both cores contained ORFs presumed to be of exogenous origin, such as beta lactamase and mercuric resistance operon regulatory protein, and because both contained many transposases, which are also presumed to be of exogenous origin (data not shown).

The codon usage in ORFs belonging to Core-class 2 is biased and has a preference for ATA, TCA and AGA (Table 1), which, in the amino acid sequences of the corresponding gene products, results in the frequent occurrence of Asn and Lys. In the case of Core-class 3 ORFs, on the other hand, Cys and Trp were found to be abundant. However, the significance of this class is not readily clear and it must be investigated further from different points of view.

Table 2. Un-detected Synechocystis ORFs

ORF	Left end	Right end	Length	Similar genes (Accession numbers inGeneBank/EMBL/DDBJ)	Hydoropathies a)
sml0003	146724	146831	108	photosystem II PsbM protein (X04465)	1.18
sml0004	160004	160093	90	hypothetical protein ycf6 (U38804)	1.22
smr0005	467201	467296	96	photosystem I PsaM subunit (X59760)	1.19
ssr1736	502042	502215	174	50S ribosomal protein L32 (U38804)	-1.29
smr0007	571084	571203	120	photosystem II PsbL protein (M33897)	0.19
smr0008	571236	571355	120	photosystem II PsbJ protein (X15767)	1.35
sml0006	831101	831217	117	50S ribosomal protein L36 (U30821)	-0.77
ssl1633	1141803	1142015	213	CAB/ELIP/HLIP superfamily (U30821)	0.61
smr0009	1167333	1167464	132	photosystem II PsbN protein (X58532)	0.56
sml0007	1268189	1268308	120	hypothetical protein ycf32 (Z67753)	0.89
sml0008	1687326	1687448	123	photosystem I subunit IX (L20938)	0.83
smr0010	1823570	1823686	117	PetG subunit of the cytochrome b6/f complex (U30821)	1.00
smr0011	1826764	1826901	138	50S ribosomal protein L34 (Z35718)	-1.41
ssr2802	1961439	1961600	162	ABC transporter (U36795)	0.88
ssl0563	2287334	2287579	246	photosystem I subunit VII (X65170)	0.08
sml0001	2350140	2350256	117	photosystem II PsbI protein (U28040)	0.71
smr0001	2414584	2414679	96	photosystem II PsbT protein (U38804)	1.06
sml0002	2613481	2613600	120	photosystem II PsbX protein (U38804)	1.01
slr0915	2791074	2791526	453	putative endonuclease (U10482)	-0.36
slr0790	3048284	3048634	351	UmuC protein (U13633)	-0.10
smr0004	3458023	3458145	123	photosystem I subunit VIII (L24773)	0.77

a) Average hydopathy among the 3168 annotated ORFs was -0.09.

The *Synechocystis* ORFs were predicted by using the matrices prepared for the three classes together with those for the Basic set. When the order of Class 1, 2 and 3 matrices was set at 5, 4 and 4, respectively, the prediction was found to match 97.9 % (3103/3168) of the ORFs previously assigned for this organism. Among the 65 annotated ORFs that were not predicted in this study, 26 were described as ORFs without similar genes/ORFs in the databases [3], and 18 others were transposases, which may not be expressed in this organism. Therefore, the prediction rate of ORFs may in fact be higher than 97.9 %. The remaining 21 ORFs that could not be detected are listed in Table 2. The length of 19 of the 21 ORFs described above was less than 300 bp. In the case of gene products (proteins) of smaller sizes, the contribution of deviation in the appearance of specific type of amino acids is generally more pronounced than in the case of larger gene products. Such deviation might account for the failure in the prediction of short ORFs, especially in the case of the deduced product of 6 gene/ORFs which harbor hydrophobicity higher than 1.0, and 2 products with hydrophobicity lower than -1.0. Some of the 21 genes/ORFs may not actually be expressed in *Synechocystis*. For example, the length of the ORF termed ssr2802 is much shorter than that of typical ORFs encoding ABC transporter proteins. The same argument may hold for slr0790 as well. It should be noted that the ORF termed slr0915 (453 bp) which escaped the detection by our procedure is a putative endonuclease [13] [3], and is located within an intron of tRNA-fMet. This intron is the only intron detected in this organism.

The detection of protein introns or *inteins* may be facilitated by the information of core-classes. Of the ORFs annotated as possessing similarity to the genes/ORFs of other organisms, 44 are longer than 3000 bp, all of which were detected in the Basic set. Of the 44 ORFs, 38 were included in Core-class 1 and 4 in Core-class 2. None of them was included in Core-class 3. The 4 ORFs in Core-class 2 were slr2046 (fat protein), sll1951 (hemolysin), slr0222 (sensory transduction histidine kinase) and sll0721 (leukotoxin (LtA)). The prediction scores were extremely high for the 42 (38 + 4) ORFs.

Table 3. Number of ORFs contained in derived classes a)

Species	Basic	Class1	Class2	Class3	Annotated
	set	(core)	(core)	(core)	ORFs
Mycoplasma genitalium	482	380	34	--- b)	468
Methanococcus jannaschii	1716	1180	206	--- b)	1680
Haemophilus influenzae Rd.	1725	1204	256	40	1713
Mycoplasma pneumoniae	682	486	96	24	676
Synechocystis	2888	2059	319	184	3168
Escherichia coli	3926	2602	326	429	4285

a) Number of ORFs contained in derived classes, the Basic set, Class 1, 2 and, 3 are shown. b) For Mycoplasma genitalium and Methanococcus jannaschii, Class 3 could not be derived.

Table 4. GC contents of derived classes

Species	Basic	Class1	Class2	Class3	Annotated
	set	(core)	(core)	(core)	ORFs
Mycoplasma genitalium	31.8	31.9	28.3	----	31.8
Methanococcus jannaschii	32.0	32.7	28.7	----	32.0
Haemophilus influenzae Rd.	38.8	39.4	35.7	38.5	38.9
Mycoplasma pneumoniae	41.1	41.6	37.1	39.9	41.1
Synechocystis	49.4	50.4	40.0	47.8	48.6
Escherichia coli	52.5	53.4	46.3	48.0	51.8

However, the remaining two of the long ORFs, namely sll1360 (DNA polymerase III subunit (dnaX)) and sll2005 (DNA gyrase B subunit (gyrB)), couldn't be detected in any Core-classes. In both cases, especially in case of sll1360, a region as long as about 1000 bp has hexamer statistics different from those represented in the matrices for the three derived classes. We found that the two ORFs have been registered in the Intein Database of New England Biolabs [14]. In the database another protein of this organism has also been registered as an intein, which was the first intein detected in this organism [15] and is located within slr0833 (replicative DNA helicase (dnaB)) of 2619 bp in length. The possibility of the ORF being an intein can be suggested in a similar analysis by reducing the threshold of ORF length. So far, only peptide level criteria have been proposed for the prediction of occurrence of likely inteins [16]. Our experience described above suggests that analysis at the nucleotide level is also helpful.

In addition to the ORFs that have already been annotated, our analysis detected a total of new 124 ORFs, of which 10 were detected with the matrices of Class 1, 92 with Class 2 and 38 with Class 3. Two new ORFs thus detected were found to be similar to a putative ORF termed yefM (AE000293 in GenBank) of *Escherichia coli*. They were detected both with Class 2 and 3 matrices, and are located at positions 2074694 through 2074957 and at positions 2085826 through 2086110, respectively.

The classification in other species

Feasibility of the classification by the modified version of GeneMark RC described above was investigated further and we confirmed that it could successfully be applied to the analysis of five other bacterial genomes mentioned in Introduction. We would like to emphasize that even in the case of *Escherichia coli*, which has the longest genomic sequence, i.e. 4.6 Mb, of all bacterial species analyzed, the whole process of classification and detection of ORFs was completed within six hours using a Ultra 2 (Creator Model 1300) of Sun Microsystems. In the case of *Mycoplasma genitalium* harboring the shortest genomic sequence, the total execution time was less ten minutes.

Results of our analysis are shown in Tables 3 and 4. For the two bacterial species with low GC contents (about 32 %), namely *Mycoplasma genitalium* and *Methanococcus jannaschii*, two classes of ORFs were derived. For the other three species, namely *Haemophilus influenzae Rd.*, *Mycoplasma pneumoniae* and *Escherichia coli*, three classes of ORFs were derived. Common characteristics were found among the six species including *Synechocystis*: (1) the GC content of Core-class 1 is slightly higher than that of the Basic set; (2) the GC content of Core-class 2 is much lower than that of Core-class 1 (or the Basic set), and the differences in the GC content between two cores become smaller with the decrease in the average GC content of the species; and (3) the GC content of Core-class 3 is between those of Class 1 and Class 2.

To analyze the features associated with ORFs of different organisms and establish their inter-relationship, ORFs of the six bacterial species were analyzed in a criss-cross manner using the Basic set matrices of each of other species. The value of R was set at 5 throughout the analyses. Therefore, the results can be regarded as being based on hexamer or di-codon statistics.

Table 5. Criss-cross prediction among species

Predicted species	Species of the Basic set matrices					
	[Myco]	[Meth]	[Inf]	[MP]	[Syne]	[E.coli]
Mycoplasma genitalium	98.9	39.3	65.0	69.1	1.7	0.6
Methanococcus jannaschii	82.1	99.3	25.6	12.8	0.0	0.1
Haemophilus influenzae Rd.	44.7	12.6	95.5	67.8	18.7	29.9
Mycoplasma pneumoniae	47.0	6.9	65.1	95.1	51.5	14.6
Synechocystis	10.4	2.5	73.2	85.4	90.2	34.6
Escherichia coli	2.0	0.9	78.7	77.2	62.6	89.8

ORFs of the six representative bacterial species were predicted in a criss-cross manner with each of the Basic set matrices of other species. Myco, Meth, Inf, MP and Syne, respectively, stand for Mycoplasma genitalium, Methanococcus jannaschii, Haemophilus influenzae Rd., Mycoplasma pneumoniae and Synechocystis.

The analyses revealed that ORFs of an organism can also be detected with the matrices of different organisms with higher prediction rate if the difference in their GC contents is smaller (Table 5). For example, the prediction rate for ORFs of *Mycoplasma pneumoniae* (GC: 41.1 %) was higher with the *Haemophilus influenzae Rd.* matrices (GC: 38.8 %) than with the *Mycoplasma genitalium* matrices (GC: 31.8 %), although the two Mycoplasma species are phylogenetically closely related with each other. An exception to this rule was the prediction of *Mycoplasma genitalium* ORFs (GC: 31.8 %). The rate, 39.3 %, obtained with the *Methanococcus jannaschii* matrices (GC: 32.0 %) was much lower than the rate, 65.0 %, obtained with the *Haemophilus influenzae Rd.* matrices (GC: 38.8 %), or the rate, 69.1 %, with the *Mycoplasma pneumoniae* matrices (GC: 41.1 %). Conversely, the prediction of ORFs in other organisms with the *Methanococcus jannaschii* matrices was poorer than that with the *Mycoplasma genitalium* matrices. It might be due to the fact that *Methanococcus jannaschii* is an archeabacterium, while other five species are eubacteria.

The significance of Class 2

As described above, the GC content of Core-class 2 was much lower than that of Core-class 1 in the six bacterial species analyzed. ORFs which are likely to be of exogenous origin were mainly contained in Core-class 2, at least in the case of *Synechocystis*. There are two possible explanations for the abundance of genes/ORFs of putative exogenous origin in Class 2 of *Synechocystis*: (1) such

Table 6. Detection of ORFs in five bacterial species using the Synechocystis matrices

Synechocystis	Num of	Prediction rates of ORFs of other species (%)			
Matrices a)	Detected ORFs	[Annotated ORFs]	[Class 1(core)]	[Class 2(core)]	[Class 3(core)]
# Mycoplasma genitalium					
Basic set	10	1.7(8/466)	1.8(7/380)	0.0(0/34)	
Class1	6	0.9(4/466)	1.1(4/380)	0.0(0/34)	
Class2	479	94.4(440/466)	98.9(376/380)	97.1(33/34)	
Class3	36	7.1(33/466)	7.3(29/380)	2.9(1/34)	
# Methanococcus jannaschii					
Basic set	2	0.0(0/1680)	0.0(0/1180)	0.0(0/206)	
Class1	0	0.0(0/1680)	0.0(0/1180)	0.0(0/206)	
Class2	1672	89.4(1502/1680)	95.5(1127/1180)	90.8(187/206)	
Class3	20	10.1(17/1680)	13.6(16/1108)	0.0(0/206)	
# Haemophilus influenzae Rd.					
Basic set	341	18.7(320/1713)	23.9(288/1204)	7.2(19/256)	7.5(3/40)
Class1	145	8.2(141/1713)	10.2(123/1204)	3.1(8/256)	2.5(1/40)
Class2	1611	87.2(1494/1713)	94.4(1136/1204)	85.4(219/256)	77.5(31/40)
Class3	100	5.7(97/1713)	6.8(82/1204)	2.7(7/256)	7.5(3/40)
# Escherichia coli					
Basic set	2896	62.6(2681/4285)	82.9(2157/2602)	31.9(104/326)	39.4(169/429)
Class1	1524	33.7(1443/4285)	48.7(1267/2602)	12.4(44/326)	12.4(53/429)
Class2	732	11.3(485/4285)	10.9(284/2602)	41.1(134/326)	28.4(122/429)
Class3	103	2.1(88/4285)	2.5(66/2602)	1.5(5/326)	1.6(7/429)

Using the matrices of Synechocystis, ORFs of other species (the annotated ORFs, and ORFs in Core-classes 1, 2 and 3) were predicted.
a) Order 5 were used in all cases.

genes/ORFs have distinct characteristics for some reasons in *Synechocystis* which have resulted in their classification into Class 2; or, (2) such genes/ORFs tend to be inserted into the genomic regions of lower GC-content, and consequently, ORFs located in the lower GC regions were categorized into Class 2. To examine which of the two explanations is more likely in the case of the *Synechocystis* matrices, ORFs of other species (except for *Mycoplasma pneumoniae*) were predicted (Table 6). Results obtained indicate that the second alternative is more likely as will be discussed below.

The annotated ORFs of the three species, *Mycoplasma genitalium*, *Methanococcus jannaschii* and *Haemophilus influenzae Rd.* were predicted by the *Synechocystis* Class 2 matrices at extremely high prediction rates, namely 94.4 %, 89.4 % and 87.2 %, respectively. The prediction rate for Core-class 1 ORFs was higher than that for Class 2 ORFs. On the other hand, the rates obtained with the *Synechocystis* Class 1 matrices were very low and only 0.9 %, 0.0 % and 8.2 %, respectively. The GC contents of the annotated ORFs of the three species were 31.8, 32.0 and 38.9 %, respectively, which are much lower than that of Core-class 1 of *Synechocystis*, 50.4 %, and are closer to the value of its Core-class 2, 40.0 %. In other words, the ORFs of the three species, especially their Core-class 1 ORFs, were more highly predictable with the *Synechocystis* matrices of Class 2 ORFs, which are composed of ORFs of exogenous origins, harboring GC contents similar to those of the three species.

To further confirm the appropriateness of the second explanation mentioned above, the *Synechocystis* data in turn were analyzed with the matrices of four other bacterial species (Table 7). With the matrices of *Mycoplasma genitalium*, *Methanococcus jannaschii* and *Haemophilus influenzae Rd.*, whose GC contents are lower than that of *Synechocystis*, Core-class 2 ORFs were detected with higher prediction rates than Core-class 1 ORFs. In contrast, with the Basic set and Class 1 matrices of

Table 7. Detection of Synechocystis ORFs by applying matrices of other species a)

Matrices a)	Num of	Prediction rates of Synechocystis ORFs (%)			
	Detected ORFs	[Annotated ORFs]	[Class 1(core)]	[Class 2(core)]	[Class 3(core)]
# Mycoplasma genitalium					
Basic	51	10.4(33/3168)	0.3(7/2059)	4.7(15/319)	0.0(0/184)
Class1	94	24.0(76/3168)	0.6(12/2059)	14.1(45/319)	10.9(2/184)
Class2	0	0.0(0/3168)	0.0(0/2059)	0.0(0/319)	0.0(0/184)
# Methanococcus jannaschii					
Basic	105	2.5(80/3168)	0.4(8/2059)	14.1(45/319)	1.6(3/184)
Class1	32	0.8(28/3168)	0.1(3/2059)	6.3(20/319)	0.0(0/184)
Class2	0	0.0(0/3168)	0.0(0/2059)	0.0(0/319)	0.0(0/184)
# Haemophilus influenzae Rd.					
Basic	1463	73.2(2320/3168)	41.6(856/2059)	67.1(214/319)	39.1(72/184)
Class1	510	14.7(466/3168)	14.3(294/2059)	29.2(93/319)	9.2(17/184)
Class2	81	2.3(72/3168)	9.2(19/2059)	12.5(40/319)	0.0(0/184)
Class3	0	0.0(0/3168)	0.0(0/2059)	0.0(0/319)	0.0(0/184)
# Escherichia coli					
Basic	1133	34.6(1096/3168)	51.4(1058/2059)	7.8(25/319)	15.2(28/184)
Class1	426	12.9(410/3168)	17.6(363/2059)	2.2(7/319)	4.9(9/184)
Class2	1066	31.6(1002/3168)	31.6(650/2059)	60.8(194/319)	18.5(34/184)
Class3	1049	31.8(1008/3168)	35.9(739/2059)	39.8(127/319)	21.2(39/184)

a) Order 5 were used in all cases.

Escherichia coli, whose GC content is higher than that of *Synechocystis*, Core-class 2 ORFs were detected with lower prediction rates than Core-class 1 ORFs. Therefore, as far as the *Synechocystis* data are concerned, detection rates apparently correlate with the GC contents of the ORF classes and the matrices used. This correlation, therefore, seems to favor the second explanation mentioned above.

4 Concluding remarks

In the study presented here, we examined the feasibility of prediction and classification of ORFs by a version of GeneMark-RC modified by extending the concept of self-consistency. We analyzed the genomic nucleotide sequence data of six prokaryotic organisms which are currently available in the public databases. Even with the nucleotide sequence data of *Escherichia coli*, which is the longest of all bacterial species analyzed, the whole analysis could be completed in a relatively short time. Therefore, the analytical procedure described here can be practiced in laboratories equipped with ordinary workstations at least as far as small genomes are concerned. Results of criss-cross analyses indicated that the statistical parameters for hexamers or di-codons (i.e. R=5) appear to be very similar among eubacteria. Thus, GeneMark-RC can serve as a potent tool of gene finding in the analysis of the genomic sequence data of at least prokaryotic organisms.

Acknowledgments

First of all, we are grateful to Mark Borodovsky of Georgia Institute of Technology for his generosity in letting us use his program GeneMark, and for his discussion and valuable suggestions on GeneMark analysis. We also thank James D. MacIninch and William Hayes of his laboratory for their help throughout our work using GeneMark.

MH thanks Takakazu Kaneko of Kazusa DNA Research Institute for valuable discussion, and Tetsushi Yada of JST for comments on the final manuscript. He also acknowledges Mitsuru Takanami, Nobuyuki Miyajima and Satoshi Tabata of Kazusa DNA Research Institute for support and encouragement during the course of this work. This work was supported in part by Grants from Kazusa DNA Research Institute Foundation and grant Nos. 06558103 and 07253215 from the Ministry of Education of Japan (to KI).

References

[1] Fleischmann, R. D., Adams, M. D., White, O. *et al.* 1995, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science,* **269,** 496-512.

[2] Fraser, C. M., Gocayne, J. D., White, O., *et al.* 1995, The minimal gene complement of Mycoplasma genitalium, *Science,* **270,** 397-403.

[3] Kaneko, T., Sato, S., Kotani, H. *et al.* 1996, Sequence Analysis of the Genome of the Unicellular Cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.,* **3,** 109-136.

[4] Bult, C., White, O., Olsen G, J. *et al.* 1996, Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschi*, *Science,* **273,** 1058-1073.

[5] University of Wisconsin-Madision, *E.coli* Genome Project, `http://www.genetics.wisc.edu/welcome.html`, 1997.

[6] Himmelreich, R., Hilbert, H., Plagens, H. *et al.* 1997, Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*, *Nucleic Acids Research,* **24,** 4420-4449.

[7] Borodovsky, M. and McIninch, J.D. 1993, GENMARK:Parallel gene recognition for both DNA strands, *Computer Chemistry*, **17**, 123-133.

[8] Hayes,W. and Borodovsky,M. How to interpret anonymous genomes ? Learning Markov models for gene identification in parallel with genomic sequence annotation, *Genome Research*, *Submitted.*

[9] Hirosawa, M., Isono, K., Hayes, W.S. and Borodovsky, M. Gene Identification and Classification in the Synechocystis Genomic Sequence by Recursive GeneMark Analysis, *DNA Sequence*, *In Press.*

[10] Medigue, C., Viari, A., Henaut, A. and Danchin, A. 1993, Colibri: a functional database for the *Escherichia coli* genome. *Microbiological Review*, **57**, 623-654.

[11] Borodovsky, M., McIninch, J.D., Koonin, E.V, Rudd, K.E., Medigue, C. and Danchin, A. 1995, Detection of new genes in a bacterial genome using Markov Models for three gene classes. *Nucleic Acids Research*, **23**, 3554-3562.

[12] TIGR, 1997, Completed microbial genomes, `http://www.tigr.org/tdb/mdb/ mdb.html`

[13] Biniszkiewicz, D., Tchesnavitchene, E. and Shub, D.A. 1995, Self-splicing group I intron in cyanobacterial intiator methionine tRNA: evidence for lateral transfer of introns in bacteria. *EMBO J.*, **13**, 4629-4635.

[14] The New England Biolabs, 1997, Intein Database. `http://www.neb.com/neb/frame_NEB.html`.

[15] Pietrokovski, S. 1996, A new intein in Cyanobacteria and its significance for the spread of inteins. *Trends In Genet.* **12**, 287-288.

[16] Perler, F. B., Olsen, G. J., and Adam, E. 1997, Compilation and analysis of intein sequences. *Nuc. Acids Res.* **25**, 1087-93.