

Gene Recognition in Cyanobacterium Genomic Sequence Data Using the Hidden Markov Model

Tetsushi Yada

The Japan Information Center of
Science and Technology (JICST)
5-3 Yonbancho, Chiyoda-ku, Tokyo 102, JAPAN
yada@jicst.go.jp

Makoto Hirosawa

Kazusa DNA Research Institute
1532-3 Yanauchino, Kisarazu, Chiba 292, JAPAN
hirosawa@kazusa.or.jp

Abstract

We have developed a hidden Markov model (HMM) to detect the protein coding regions within one megabase contiguous sequence data, registered in a database called GenBank in eight entries, of the genome of cyanobacterium, *Synechocystis* sp. strain PCC6803. Detection of the coding regions in the database entry was performed by using HMM whose parameters were determined by taking the statistics from the rests of the entries. This HMM has states modeling the di-codons and their frequencies within coding regions and those modeling its base contents in the intergenic regions. Results of the cross-validation showed that the HMM recognized 92.1% of coding regions assigned in sequence annotation. In addition, it suggested 94 potential new coding regions whose length are longer than 90 bases. The recognition accuracy calculated at the level of individual bases was 90.7% for the coding regions and 88.1% for the intergenic regions. This corresponds to a correlation coefficient for coding region recognition of 0.784. Comparison with its prediction accuracy with that by GeneMark showed that the HMM has the same level of prediction accuracy as GeneMark on average. Since we can extend the HMM to utilize information such as SD sequences, the prediction accuracy of the HMM will be enhanced. It was observed that correlation was positive between the prediction rate of the coding regions and the G+C content at the third position of the codon. This suggests the possibility that the prediction rate of coding regions in the cyanobacteria sequence can be enhanced by improving the present HMM into that reflects the classification of coding regions based on the G+C content.

Introduction

The advancement in large-scale sequencing has accelerated the production of long contiguous nucleotide sequences. Today, more than fifty contiguous nucleotide sequences, whose length are longer than 100 kb, are available through World Wide Web (WWW) (Robinson 1996). These determined contiguous nucleotide

demands researchers to develop effective algorithms to detect protein coding regions in the nucleotide sequences.

One way to identify the coding regions is to use a similarity search against databases. However, studies on large-scale sequencing suggested that only a fraction of the coding regions can be identified through a similarity search (Koonin, Bork, & Sander 1994; Wilson *et al.* 1995).

Algorithms that utilize codon usage bias within coding regions to discriminate the regions from intergenic regions have been studied by many researchers (Fickett & Tung 1992; Borodovsky, Rudd, & Koonin 1994). Some of the algorithms utilize sequence patterns outside both ends of coding regions as well (Uberbacher & Mural 1991; Soderlund *et al.* 1992; Guigò *et al.* 1992; Solovyev, Salamov, & Lawrence 1994; Dong & Searls 1994; Krogh, Mian, & Haussler 1994; Snyder & Stormo 1995). This approach, based on the Hidden Markov Model (HMM) (Krogh, Mian, & Haussler 1994), is on this line.

HMM, which is a stochastic model, can be described as a network composed of nodes and directed paths connecting them. Nodes and directed paths represent the states and transition between the states, respectively. Several parameters, such as transition probabilities, output symbol distributions and initial state distributions, are attached to each node and directed path.

One of the advantages of the HMM approach is that we can naturally construct an HMM network by simply borrowing the structure of the coding regions in a nucleotide sequence. We can design HMMs, each of which corresponds to a gene component, such as a start codon, protein coding region, stop codon and intergenic region. After assignment of parameters representing such as the information on codon usage bias, consensus sequence and base contents to HMM network components, these components are integrated to construct an HMM network representing nucleotide sequences embedded with coding regions. These parameters for given sequences are estimated or determined by an expectation maximization algorithm (EM algo-

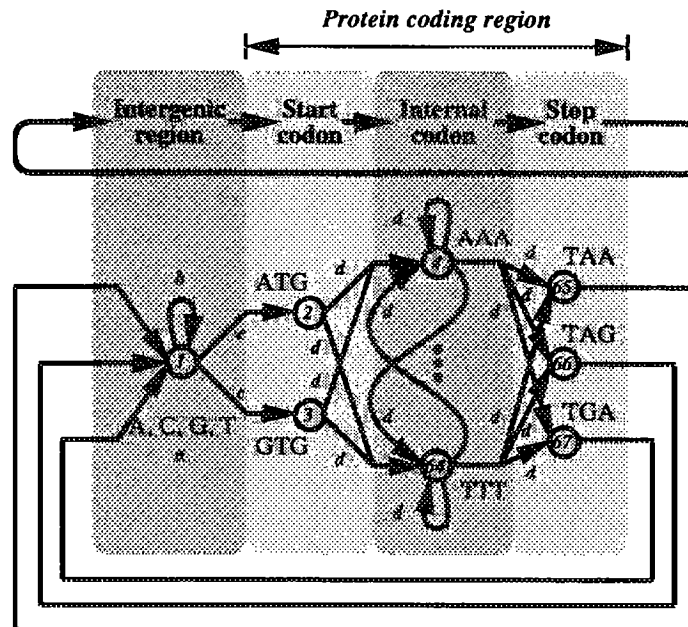


Figure 1: An HMM representing the genome structure of cyanobacteria.

rithm) (Baum *et al.* 1970) or statistical analysis.

The Viterbi algorithm (Levinson, Rabiner, & Sondhi 1983) enables us to find, in an HMM network, the most probable path, parts of which corresponding to coding regions can be detectable in the algorithm. Successful detection of coding regions, with high accuracy in an HMM approach, has enabled us to understand the genomic structure properly. We decided to develop a new gene finding algorithm based on HMM because of the advantages previously stated.

We selected a contiguous sequence of a unicellular cyanobacterium, *Synechocystis* sp. strain PCC6803 to construct an HMM representing sequence with genomic structure. We carried out tests to detect regions having a potential protein-coding capacity to prove the effectiveness of our algorithm.

There are two main reasons for our selection. The first is that the sequence is as long as 1,003,450 nt (D63999 ~ D64006 in GenBank (GenBank 1995)) and was determined by a single institute, Kazusa DNA Research Institute. Its length is enough for us to statistically analyze the effectiveness of the gene finding algorithms we developed. The fact that the sequence was determined by a single institute leads to a uniform quality of sequencing, then to suitability for the evaluation of the algorithms. The second is that determination of the whole genome sequence (3.6 Mb in total) will be finished by this March and the sequence will be available soon. The sequence will enable us to represent and analyze the whole genome organization of the organism in HMM.

Method and Materials

HMM Network

We developed an HMM network representing the genome structure of cyanobacteria (Figure 1). The network is a directed cycle whose components are the intergenic region, start codon, internal codon and stop codon. The order of the components in the cycle is the same as stated above. The protein coding region consists of start, internal and stop codons. The transitions from the intergenic region to the start codon and from the stop codon to the intergenic region correspond to the start and end points of the coding region, respectively.

The HMM has states modeling the di-codons and their frequencies within the coding regions and those modeling its base contents in the intergenic regions. The statistics, based on di-codons, has been shown to be important for the discrimination of the regions from intergenic regions (Uberbacher & Mural 1991; Fickett & Tung 1992; Guigò *et al.* 1992; Borodovsky, Rudd, & Koonin 1994). The statistics based on the di-codons is one of the features that was not employed in the HMM developed by Krogh *et al.* (Krogh, Mian, & Haussler 1994) but introduced newly in our HMM. Krogh *et al.* proposed an HMM based on mono-codon statistics. In the "Results and Discussion", we discuss the advantage of di-codon statistics over mono-codon statistics in the cyanobacteria coding regions.

Intergenic region is represented by state S_1 in the HMM. The state outputs characters A, C, G and T, and possesses transitions to itself and to start codons. There are two kinds of start codons in cyanobacteria

genome, which are represented by S_2 and S_3 , respectively. S_2 produces a ATG codon, and S_3 produces a GTG codon. After the output of a start codon, transition to internal codons occurs. Internal codons are represented by $S_4 \sim S_{64}$. These 61 states correspond to codons from AAA to TTT exclusive stop codons (see below). A state of internal codon possesses transitions to all internal codons including itself and the stop codons. Three kinds of stop codons, TAA, TAG and TGA, are known in cyanobacteria genome and are represented by $S_{65} \sim S_{67}$ in the HMM. After output of a stop codon, transition back to the intergenic region occurs to complete one cycle of the transitions.

Training and Test Sets

A contiguous sequence, whose length is 1,003,450 nt, of a unicellular cyanobacterium, *Synechocystis* sp. strain PCC6803 was selected for making both training and test data for the analysis. The sequence was divided in eight entries in GenBank (accession numbers D63999–D64006) with potential protein coding regions in the annotations (Kaneko *et al.* 1995; Hirosawa, Kaneko, & Tabata 1995; Hirosawa 1996). We respected the eight divisions of the sequence and created eight training data sets and test sets.

Each of the eight entries was assigned as a test set, and as its training set, the seven remaining entries were assigned. A training set, such as assembled sequence (D64000–D64006), was used to determine the HMM parameters and a test set, such as D63999, was used to validate the recognition ability of HMM based on parameters derived from the training set. The parameters include those for the coding regions and those for the intergenic regions by respecting the annotations of the coding regions in the entries. With these parameters, we modeled a nucleotide sequence embedded with coding regions.

We did not modeled the coding regions in the complementary direction in this study. However, extension to model coding regions in the complementary direction is straightforward. Each entry was used twice as a test set: a test set to evaluate detection of coding in the normal direction and that to evaluate detection in its complementary direction.

Parameter Determination

From genomic structural points of view, the HMM parameters can be classify into four groups labeled with $a \sim d$ in Figure 1. We determined the parameters of HMM for examination of given test set by the statistical analysis described below.

- a This implies base frequency in intergenic regions. We assigned the frequency to the output symbol distribution at state S_1 .
- b This represents average length of intergenic regions. Using Markov analysis (Feller 1957), we converted the length into corresponding probability, then, assigned the probability to transition probability from

S_1 to S_1 . To be concrete, (1) defining $S_2 \sim S_{67}$ states as an absorption state; (2) the probability b is given by the following equation:

$$b = \frac{l - 1.0}{l} \quad (1)$$

where l is the average length of intergenic regions.

- c These imply start codon frequencies. We divided the probability $1.0 - b$ into two being proportional to the frequencies of start codons, then assigned them to corresponding transition probabilities.
- d These imply the di-codon frequencies. We assigned the frequencies to corresponding transition probabilities.

Gene Recognition

In our algorithms, coding regions were detected in test sets using HMM by the following procedures:

1. An HMM representing a given training set was developed as described in the above subsections. When we used the HMM to predict the coding regions in the corresponding test set, we had to determine the threshold of logarithmic likelihood which discriminates the coding regions from the others. For that purpose, logarithmic likelihood, which represents the likelihood of coding region candidates as being the true coding region, for the annotated coding regions in a given training set was computed by using the Viterbi algorithm. We defined the threshold as the value corresponding to lowest logarithmic likelihood among the annotated coding regions in the training set.
2. The most probable path in the HMM was determined by the Viterbi algorithm for each test set and the possible coding regions on the path were listed.
3. Logarithmic likelihood for the possible coding regions in each test set was calculated. Coding regions, whose logarithmic likelihood were more than the threshold for the test set, were selected as the predicted coding regions in the test set.

Results and Discussion

Appropriateness of di-codon Statistics

By using χ^2 test, we validated the advantage of di-codon statistics over mono-codon statistics in representing internal coding regions of cyanobacteria. Mono-codon and di-codon correspond to zero-th and first order stationary Markov process, respectively. Assuming a zero-th order Markov process, the relative χ^2 value is given by the following equation:

$$\chi^2 = \sum_{i,j}^{61} \frac{(p_{ij} - \pi_{ij})^2}{\pi_{ij}} \quad (2)$$

where

$$\pi_{ij} = (p_i)(p_j) \quad (3)$$

Table 1: Calculation of the performance measures at base level.

	Predicted positives (PP)	Predicted negatives (PN)
Actual positives (AP)	True positives (TP)	False negatives (FN)
Actual negatives (AN)	False positives (FP)	True negatives (TN)
Sensitivity (S_n)	$S_n = TP/(TP + FN)$	
Specificity (S_p)	$S_p = TP/(TP + FP)$	
Correlation coefficient (CC)	$CC = [(TP)(TN) - (FP)(FN)]/\sqrt{(PP)(PN)(AP)(AN)}$	

Table 2: Recognition accuracy at gene level.

Test	Exact recognition		Approximate recognition	
1	65.5%	(57/87)	94.2%	(82/87)
2	77.1%	(98/127)	92.1%	(117/127)
3	77.3%	(75/97)	94.8%	(92/97)
4	76.1%	(67/88)	96.6%	(85/88)
5	68.9%	(62/90)	88.9%	(80/90)
6	64.1%	(75/117)	88.0%	(103/117)
7	60.9%	(67/110)	85.5%	(94/110)
8	81.7%	(85/104)	98.1%	(102/104)
Total	71.5%	(586/820)	92.1%	(755/820)

with p_{ij} and p_i are the respective frequencies of di-codons and mono-codons observed in the regions. 61 is the number of codons other than stop codons. The adverse influence of a large number of samples on χ^2 test is reduced by employing the relative χ^2 formula (Wonnacott & Wonnacott 1972) instead of original χ^2 formula.

As a result of calculation based on the zero-th order Markov process, we could obtained the following relation:

$$\chi^2 = 0.4816 > \frac{\chi_{0.05}^2}{m} \quad (4)$$

where the degree of freedom is 3720(= 61 × 61 − 1). m , which indicates the number of di-codon samples in the regions, is 285,039. Equation 4 shows that the zero-th order Markov process of codon usage is rejected based on the significance at the 5% level. It can be stated that an HMM network representing the zero-th order Markov process of codon usage (Krogh *et al.* model) dose not have the ability to represent the first order codon usage (our model).

Theoretically, employment of statistics higher than the first order codon usage is more desirable. However, it is impossible to take the option because of insufficient data size, even after all coding regions in cyanobacteria genome (3.6 M bp) becomes available. Roughly speaking, several tens times as large as current data size is required for reliable statistics of the second order codon usage.

Taking into account this situation, an HMM network representing the first order codon usage is the most suitable for reliable gene recognition of this species.

Prediction Quality

The genomic sequence data of *Synechocystis* were analyzed by the HMM to find the location of the protein coding region in the nucleotide sequence. From HMM and biological points of view, we evaluated the quality of the coding region prediction in comparison with the annotated coding regions in the eight entries. Further, we compared our results with the coding region prediction of the same specie by GeneMark developed by Borodovsky *et al.* (Borodovsky, Rudd, & Koonin 1994).

HMM Aspect We analyzed the results of the prediction in two levels, the gene level and base level. In the gene level, whether the annotated coding regions were detected or not was focused on. The evaluation was carried out as regards both exact and approximate recognition. Exact recognition stands for exact identification of both 5' and 3' ends of the coding regions. Approximate recognition stands for exact identification of only 3' end of the coding regions.

As regards the base level, more a detailed evaluation was carried out by focusing on every single base. First, we counted the four values, the number of bases correctly predicted as the coding region (true positive: TP), falsely predicted (false positive: FP), correctly excluded (true negative: TN), and falsely excluded (false negative: FN). Then, we computed the indices shown in Table 1. Statistical metrics, such as sensitivity, specificity and correlation coefficient are among them.

The results of the evaluation of the coding region

Table 3: Recognition accuracy at base level.

Test	True positives		True negatives	
1	92.4%	(102,856/111,284)	88.6%	(123,731/139,654)
2	93.4%	(112,893/120,899)	90.1%	(154,564/171,643)
3	92.1%	(83,778/90,957)	85.3%	(102,858/120,633)
4	93.2%	(86,157/92,445)	85.2%	(110,192/129,371)
5	85.2%	(82,751/97,144)	86.7%	(111,829/128,984)
6	88.3%	(103,117/116,813)	88.4%	(134,069/151,585)
7	83.8%	(97,299/116,100)	92.2%	(143,143/155,176)
8	97.4%	(112,605/115,636)	87.1%	(129,475/148,576)
Total	90.7%	(781,456/861,278)	88.1%	(1,009,861/1,145,622)

Table 4: Protein coding regions not predicted by the HMM.

Entry No.	CDS (Length)
D63999	sll0161 (303), ssl0296 (255), slr0361 (750), slr0373 (381), sll0327 (420)
D64000	slr0770 (321), slr0344 (1422), sll0315 (684), slr0353 (366), ssl0601 (183), slr0013 (528), sll0010 (381), slr0021 (834), sll0210 (981), sll0200 (360)
D64001	sll0409 (963), sll0406 (486), sll0404 (1479), slr0070 (993), slr0489 (456)
D64002	slr0417 (2583), sll0371 (1566), slr0427 (1251)
D64003	slr0896 (1386), slr0915 (453), slr0573 (444), slr0575 (555), sll0539 (1194), slr0829 (2490), sll0834 (1695), sll0832 (456)
D64004	sll0487 (1209), slr0514 (888), slr0517 (615), slr0518 (534), sll0595 (432), slr0617 (1356), slr0099 (1020), sll0107 (621), slr0106 (705), slr0108 (1257), slr0109 (957), sll0095 (1704), slr0115 (726), ssl0172 (270)
D64005	slr0309 (1422), slr0317 (864), slr0319 (915), slr0786 (825), sll0800 (399), slr0794 (3165), slr0797 (1929), slr0799 (324), sll0787 (969), sll0780 (303), slr0698 (465), sll0678 (438), sll0675 (315), sll0668 (381), sll0667 (660), sll0657 (1098)
D64006	sll0039 (360), sll0033 (1506)

analysis can be seen in Table 2 and Table 3. Table 2 and Table 3 correspond to the evaluation of the prediction quality in the gene level and that in the base level, respectively.

In the gene level, the HMM recognized exactly the location of 71.5% of the coding regions annotated in the eight entries, and was approximately 92.1%. Coding regions of 20.6% were predicted to start at more internal position than their counterparts in the annotations. Generally speaking, assignment of 3' end for the coding region in the databases is reliable because the coding regions definitely stop at any stop codon. However, assignment of 5' is not so reliable. When there are multiple candidates of the start codons for a given coding region, analysis of the intergenic region proceeding to the possible coding regions serves to correctly assign 3' position of the coding region. Or some experimental analysis is necessary. It is often the case, these necessary analysis are not conducted but

the longest possible coding regions are assigned.

If we take into account the above situation, we do not necessarily regard negatively the fact 5' position of 20.6% of the prediction coding regions is different from 5' position of their counterparts in the annotations. Rather, we can take our assignments of 5' position as alternative candidates of 5' positions of these coding regions.

In the base level, the true positive index and true negative index were 90.7% and 88.1%, respectively. Sensitivity, specificity and correlation coefficient for gene recognition were 0.907, 0.852 and 0.784.

The results of these analysis both in the gene and base levels clearly indicate that the HMM is very effective for predicting the coding regions within the cyanobacteria genome. Contribution of the usage of di-codon statistics to the highly successful prediction by our algorithm should be investigated.

Table 5: Protein coding regions newly predicted by the HMM.

Entry No.	Strand ⁽¹⁾	Region
D63999	D	12805..12927, 23515..23700
	C	20493..20582, 62888..63100, 82881..82997, 98477..98590, 120008..120103
D64000	D	14508..14684, 21856..21951, 38808..38969, 92602..92709, 117095..117286, 141128..141337
	C	24538..24738, 68473..68580, 92687..92794, 123800..124021
D64001	D	14076..14339, 25176..25391, 37093..37281, 65571..65675, 85927..86028
	C	4726..5019, 5025..5324, 14496..14789, 41292..41558, 57621..57779, 71536..71676, 74482..74601, 92612..92851
D64002	D	1507..1728, 5412..5555, 13856..14140, 50379..50510, 50826..50981, 51755..52042, 77450..77575, 96628..96777, 104152..104364, 109847..110056, 110059..110292
	C	13922..14089, 31413..31631, 51331..51609, 66625..66843, 73997..74125, 90628..90804, 93918..94121, 94124..94339, 110388..110606
D64003	D	4733..4930, 13080..13184, 18935..19024, 19622..19906, 41072..41263, 53252..53347, 61979..62095, 83987..84202, 101805..102053
	C	19540..19677, 23358..23504, 40736..40954, 59028..59222, 60364..60618, 61289..61408, 61524..61751, 89107..89370, 104369..104587
D64004	D	81341..81577, 88392..88598, 109633..109737
	C	36569..36667
D64005	D	3844..4038, 24774..24869, 71976..72113, 81283..81450, 81455..81619, 95459..95752, 115226..115336
	C	6028..6285, 39521..39706, 41453..41689, 57374..57499, 64392..64682, 113474..113716, 113772..113867
D64006	D	89860..90090, 96622..96714, 96835..97077, 121394..121525
	C	49502..49624, 73076..73204, 98271..98435, 101410..101502

(1) Characters "D" and "C" stand for direct and complementary strand, respectively.

Biological Aspect It is important to check whether annotated coding regions assigned with high reliability were predicted by our algorithm or not. Among the annotated coding regions in the contiguous sequence, half of the coding regions was assigned because of their similarity to entries in databases. Among these coding regions, 25 coding regions had exact counterparts in the databases at the time the sequence was submitted (Kaneko *et al.* 1995). The HMM successfully detected all but one of the reliable annotated coding regions (GeneMark couldn't detect one more coding region). This high rate of detection of the reliable annotated coding regions (96%) demonstrates effectiveness of the HMM.

Slr0915 in D64003, a coding region escaped detection, is identical to U10482 of the GenBank. However, U10482 is described as an putative endonuclease and its existence has not been studied yet. The un-detection of slr0915 does not necessarily lead to a failure of the HMM in detecting coding regions.

In addition, the HMM suggested 94 potential new

coding regions whose length are longer than 90 bases (Table 5). The detection of nearly 100 coding regions suggests high detectability of the algorithms. A process for investigating whether the suggested coding regions are bona fides or not, is going on now, by analyzing the results of similarity search against databases, such as SwissPort, PIR and GenBank.

Comparison with GeneMark Hirose *et al.* reported that GeneMark, whose parameters were optimized for cyanobacteria genomic data, recognized approximately the location of 92% of the coding regions annotated in the eight entries (Hirose *et al.* 1995). On the other hand, the HMM recognized approximately 92.1% of the regions. Comparison of these accuracies indicates that the HMM possesses recognition ability equivalent to GeneMark.

We investigated further by focusing on recognition accuracy of coding regions shorter than 300 bases. Eight coding regions are annotated in the eight entries. Table 6 is the list of the eight coding regions and results of recognition by the HMM and GeneMark. The

Table 6: Coding regions shorter than 300 bases and results of recognition by the HMM and GeneMark.

Entry No.	CDS	Length	HMM	GeneMark
D63999	ssl0563	270	Yes	Yes
	ssr0562	246	Yes	No
D64000	ssl1426	204	Yes	No
	ssl0601	183	No	No
	ssl0020	294	Yes	Yes
D64002	ssr0330	228	Yes	Yes
D64003	ssr1398	198	Yes	Yes
	ssr1399	216	Yes	Yes

HMM successfully detected all but one coding regions. On the contrary, GeneMark failed in detection of two more regions. Since sample number is small, it is not fair to maintain the HMM is better model to detect short coding regions. However, it can be maintained that the HMM is a promising method for revealing the existence of unknown short coding regions.

Improvements of the Model

We propose some improvements of the present HMM. Improvements described below will increase the reliability of gene recognition, and enable us to understand the genomic structure more in detail.

Influence of G+C Content on Prediction Performance Researches of analysis on gene recognition programs developed so far have shown that the prediction rate of coding regions in a given nucleotide sequence has a correlation with the G+C content of the sequence (Lopez, Larsen, & Prydz 1994; Snyder & Stormo 1995). It is valuable to check whether it is also the case in our algorithm using HMM for further improvement of our algorithm.

Figure 2 shows a correlation between the G+C content at the third position of the codon and the prediction rate for all sets of test data. It is known that the G+C content of a nucleotide sequence reflects the G+C content at the third position of the codon in the coding region (Aota & Ikemura 1986). Figure 2 shows that the correlation is also positive between the prediction rate of coding regions and the G+C content in our analysis of the cyanobacterium coding regions.

This correlation suggests that the codon usage bias in the annotated coding region within A+T rich coding regions is different from the one within the G+C rich coding regions and that the difference results in under-prediction of the annotated coding regions in A+T rich sequence domains. It is known that genome is composed of domains with a different G+C content level, in each of which, the G+C content is fairly homogeneous and length of the domains is rather long (Ikemura & Aota 1988). It is desirable to represent, in HMM, this domain structure having dependence on the G+C content. It is expected that the HMM reflecting the domain structure will enhance the prediction

rate of the coding regions. For example, slr0915, annotated regions with high reliability but not detected in the present HMM, may be detected by the HMM.

Figure 3 is an HMM, we propose, consists of two HMM components, an HMM for G+C rich domains and an HMM for A+T rich domains. The two components are connected to each other by directed paths which represent the transition from the G+C rich domain to the A+T rich domain and vice versa.

Both components are an HMM whose network structure is the same as in Figure 1. The HMM parameters, such as the codon usage bias and base content, are determined so as to model the specificity in respective regions. We are now developing an algorithm classifying coding regions based on the G+C contents and other available information to develop and evaluate the proposed HMM.

Models for Intergenic Regions, Frame Shift Errors and Overlapping Genes The present HMM has only one state as intergenic states, because 5' ends of coding regions are not known accurately. However, extension of the HMM to include intergenic states representing sequence patterns specific to the outside of both ends of coding regions, in addition to states representing all over intergenic regions, is straightforward. According to personal communication from T. Sazuka (Kazusa DNA Research Institute), nearly 100 accurate 5' ends of coding regions will be published soon. The information will enable us to design a more complicated intergenic model. To realize the model, we are investigating the methodology based on the combination of statistical analysis (Pesole *et al.* 1992) and the method for network determination of HMM (Yada *et al.* 1996).

Cyanobacteria sequence data we used can be regarded as reliable and its quality is homogeneous, because the sequence was determined by reading both strands at a single institute (Kotani *et al.* 1994). The present HMM have, therefore, no states for detecting frame shift errors. However, the implementation of states for detecting frame shift errors into the HMM is important to broaden its applicable species. For the purpose, frame shift model which Krogh *et al.* pro-

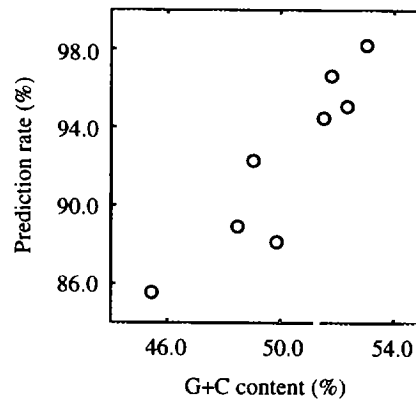


Figure 2: Correlation between the G+C content and prediction rate.

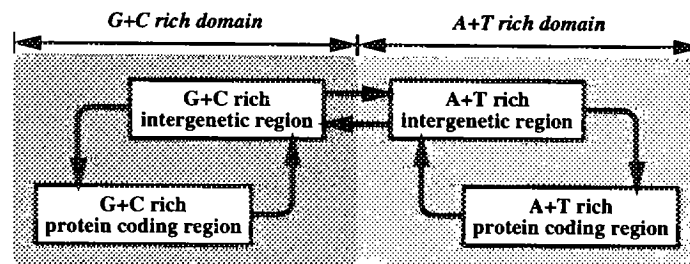


Figure 3: An HMM reflecting classification of genes based on the G+C content.

posed is one of possible ways.

Overlapping genes, whose overlap length is various, are sometimes observed in cyanobacteria genome. Krogh *et al.* proposed an HMM representing overlapping genes. The HMM can just handle overlaps of one or four bases long and didn't handle other type of overlaps in the same framework. We investigated more general approach and found that dynamic programming (DP) obtaining sub-optimal solutions (Knight & Myers 1995) is applicable to the Viterbi algorithm. This approach enables us to identify alternative splice sites in primate genome, as well as overlapping genes.

Acknowledgments

This work was carried out as a part of ARIS (Advanced Research Information System) project for genome analysis by the Japan Information Center of Science and Technology (JICST), and was supported in part by fund from Kazusa DNA Research Institute.

References

Aota, S., and Ikemura, T. 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* 14:6345-6355.
 Baum, L. E.; Petrie, T.; Soules, G.; and Weiss, M. 1970. A maximization technique occurring in the sta-

tistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41:164-171.

Borodovsky, M.; Rudd, K. E.; and Koonin, E. V. 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.* 22:4756-4767.

Dong, S., and Searls, D. B. 1994. Gene structure prediction by linguistic methods. *Genomics* 23:540-551.

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, 2 edition.

Fickett, J. W., and Tung, C.-S. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* 20:6441-6450.

GenBank. 1995. Genetic sequence data bank, release 92.0. Technical report, BBN Laboratories, U.S.A.

Guigò, R.; Kundsén, S.; Drake, N.; and Smith, T. 1992. Prediction of gene structure. *J. Mol. Bio.* 226:141-157.

Hirosawa, M.; Kaneko, T.; Tabata, S.; McIninch, J. D.; Hayes, W. S.; Borodovsky, M.; and Isono, K. 1995. Computer survey for likely genes in the one megabase contiguous genomic sequence data of *Synechocystis* sp. strain PCC6803. *DNA Res.* 2:239-246.

- Hirosawa, M.; Kaneko, T.; and Tabata, S. 1995. Cyanobase: Visual representation of information on the genome of cyanobacterium *Synechocystis* sp. strain PCC6803 through WWW. In *Proc. of Genome Informatics Workshop VI*, 102–103.
- Hirosawa, M. 1996. Cyanobase. <http://www.kazusa.or.jp/cyano/cyano.html>.
- Ikemura, T., and Aota, S. 1988. Global variation in G+C content along vertebrate genome DNA. *J. Mol. Biol.* 203:1–13.
- Kaneko, T.; Tanaka, A.; Sato, S.; Kotani, H.; Sazuka, T.; Miyajima, N.; Sugiura, M.; and Tabata, S. 1995. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. i. sequence features in the 1 Mb region from map positions 64% to 92% of the genome. *DNA Res.* 2:153–166.
- Knight, J. R., and Myers, E. W. 1995. Super-pattern matching. *Algorithmica* 13:211–243.
- Koonin, E.; Bork, P.; and Sander, C. 1994. Yeast chromosome III: New gene function. *EMBO J.* 13:493–503.
- Kotani, H.; Kaneko, T.; Matsubayashi, T.; Sato, S.; Sugiura, M.; and Tabata, S. 1994. A physical map of the genome of a unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. *DNA Res.* 1:303–307.
- Krogh, A.; Mian, I. S.; and Haussler, D. 1994. A hidden Markov model that finds gene in E.coli DNA. *Nucleic Acids Res.* 22:4768–4778.
- Levinson, S. E.; Rabiner, L. R.; and Sondhi, M. M. 1983. An introduction to the application of the theory of probabilistic function of a Markov process to automatic speech recognition. *Bell Syst. Tech. J.* 62:1035–1074.
- Lopez, R.; Larsen, F.; and Prydz, H. 1994. Evaluation of the exon predictions of the GRAIL software. *Genomics* 24:133–136.
- Pesole, G.; Prunella, N.; Liuni, S.; Attimonelli, M.; and Saccone, C. 1992. WORDUP: An efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Res.* 20:2871–2875.
- Robinson, K. 1996. The 100 kb club. <http://golgi.harvard.edu/100kb/>.
- Snyder, E. E., and Stormo, G. D. 1995. Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* 248:1–18.
- Soderlund, C.; Schanmugam, P.; White, O.; and Fields, C. 1992. gm: a tool for exploratory analysis of DNA sequence data. In *Hawaii Int. Conf. on System Sciences*, 653–662.
- Solovyev, V. V.; Salamov, A. A.; and Lawrence, C. B. 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* 22:5156–5163.
- Uberbacher, E. C., and Mural, R. J. 1991. Locating protein-coding region on human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. U.S.A.* 88:11261–11265.
- Wilson, R.; Ainscough, R.; Anderson, K.; Baynes, C.; Berks, M.; Bonfield, J.; Burtoni, J.; Connell, M.; Copsey, T.; and Cooper, J. 1995. 2.2Mb of continuous nucleotide sequence from chromosome III of *C. elegans*. *Nature (London)* 368:32–38.
- Wonnacott, T. H., and Wonnacott, R. J. 1972. *Introductory Statistics*. John Wiley & Sons, 2 edition.
- Yada, T.; Ishikawa, M.; Tanaka, H.; and Asai, K. 1996. Extraction of hidden Markov model representations of signal patterns in DNA sequences. In *Proc. of Pacific Symp. on Biocomputing '96*, 686–696.