Analysis of Sequence Patterns Surrounding the Translation Initiation Sites on Cyanobacterium Genome Using the Hidden Markov Model

Tetsushi YADA,*1 Takashi SAZUKA,2 and Makoto HIDAWA2

Japan Science and Technology Corporation (JST), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102, JAPAN1 and
Kazusa DNA Research Institute (KDRI), 1532-3 Yana-uchino, Kisarazu, Chiba 292, JAPAN2

(Received 12 November 1996)

Abstract

Sequence patterns surrounding the translation initiation sites of Cyanobacterium were precisely analyzed by the hidden Markov model (HMM) based on the actual translation initiation sites. In a previous study, 72 actual protein coding regions and their translation initiation sites on the genome of Synechocystis sp. strain PCC6803 were determined by Sazuka et al. using protein two-dimensional electrophoresis and microsequencing. In this work, we extracted the sequence patterns surrounding translation initiation sites as HMM using the computer program YEBIS. The constructed HMM could recognize all but one translation initiation site. The HMM contains an AG-rich region (5.7 bp on average), as the Shine-Dalgarno sequence exclusively contains purines, upstream of the translation initiation site (−9.7 position on average) and a CT-rich region (4.2 bp on average) just upstream from the translation initiation site. In addition, we found that the second amino acid (+4,5,6) could be classified into two types, one of which had C as their second codon while another of which has a nucleotide distribution relatively similar to the distribution among amino acids in the 72 proteins. This fact corresponds well to our earlier finding that when the second nucleotide of the second amino acid of a translated protein was C, an initial methionine was processed and that otherwise the methionine was intact with high frequency.

Key words: translation initiation site; hidden Markov model; rare initiation codon

1. Introduction

The complete genome of four species, Haemophilus influenzae Rd.,1 Mycoplasma genitalium,2 Cyanobacterium Synechocystis sp. strain PCC68033 and Methanococcus jannaschii,4 have already been determined and the sequence and potential coding regions of each species, assigned according to certain assumptions, are available from GenBank/EMBL/DDBJ.

However, for the understanding of the complicated physiological process in living organisms, the determination of actual protein coding regions by experiment should be necessary. Sazuka and Ohara determined actual initiation sites and their coding regions of 72 proteins by two dimensional electrophoresis and microsequencing.5 Further, continued determination of initiation sites is possible and important, but it seems to take a long time to complete initiation sites of all the expressed proteins with present technology. Therefore, computer programs are needed to predict translation initiation sites based on the information extracted from the experimentally determined 72 coding regions.

For the prediction and/or analysis of translation initiation sites, the representation and extraction of nucleotide patterns surrounding translation initiation sites is necessary. The representation of certain nucleotide or peptide patterns which carry biological significance, has been represented by regular expression in the biological field for a long time. For example, protein motifs registered in the database Prosite are expressed in it.6 However, in the case of complex patterns like transcriptional or translational controls, the regular expression is not able to represent such patterns precisely.

Recently, representation of biological significant patterns has been intensively studied based on the hidden Markov model by several research groups.7,8 With hidden Markov model (HMM) nucleotide patterns, such as the TATA box of human promoters can easily and concisely be represented as depicted in Fig. 1

HMM is composed of nodes, with nucleotide frequency assigned to them, and transition probabilities between...
Amino-Terminal Sequence Pattern Recognition by HMM

Figure 1. An example of HMM (TATA box in human promoter).

The HMM are composed of nodes, represented by circles, transitions between nodes, represented by arrows. In this figure, the names of the nodes are written in the nodes. Nucleotide frequencies of the nodes are written in boxes near the nodes and transition probabilities are written near the arrows representing the transitions. Node 4 has a self-loop, a transition back to itself, which represents a repetition of A or G represented by this node. The transition from node 4 to node 3 represents a repetition of T, represented by node 3, and A or G, represented by node 4.

nodes. This HMM represents the main pattern of the TATA box as TAT[AG] (TATA or TATG), which is represented by a path composed of Node 1, 2, 3 and 4. Nucleotide frequency in Node 4 results from ambiguity in the fourth amino acid (A or G). A self-loop in Node 4 represents the possible repetition of A or G with a probability of 0.6 after subsequence TATA. A transition from Node 4 to Node 3 represents repetition of T[AG] after TAT[AG] with a probability of 0.2. The transition (probability of 0.2) from Node 4 to the outside of the TATA box represents the end of the TATA box pattern without further repetitions. In this way, variant patterns as well as prototype patterns of biologically significant regions, such as the TATA box, can be expressed by only one HMM.

However, the creation of the target HMM, especially the construction of the network shape of the HMM, requires much computational time if we take a straightforward approach. As a result, an assumption on the network shape based on preliminary analysis is required. Recently, Yada et al., developed a computer program named YEBIS (YEBIS is the acronym for Yet another Environment for contextual analysis of Biopolymer Sequences) which efficiently creates a complex HMM from source sequences without any assumption succeeded in the recognition of HMM corresponding to the human promoter region.

In this study, from 72 nucleotide sequences surrounding the actual translation initiation sites, HMM representing these sequences were created by YEBIS. Based on these created HMM sequences, the surrounding translation initiation sites were analyzed.

2. Materials and Methods

2.1. YEBIS

The YEBIS computer program consists of two modules. The first module creates an HMM network representing a set of given sequences using the Baum-Welch algorithm and Genetic Algorithm. The complexity of the network is controlled by a parameter called balancing factor lambda. With smaller lambda an HMM which is detailed but complicated is created, while with larger lambda an HMM which is simple but understandable is created. The second module detects regions corresponding to the state of the created HMM in a given sequence, using the Viterbi algorithm.

2.2. Sequences surrounding translation initiation sites

Seventy-two nucleotide sequences, whose lengths are 38 bp, surrounding translation initiation sites (from position —25 to 13) were prepared by Sazuka and Ohara. Most of the proteins translated from the respective initiation sites have been detected in their study. The nucleotide sequences are shown in Fig. 2.

2.3. Extraction of HMM representing translation initiation sites

Each of the 72 nucleotide sequences was divided into three domains, namely the upstream (positions —25 to 13), the initiation codon (positions 1 to 3) and the downstream (positions 4 to 13). Then, three data sets, each of which corresponded to the above mentioned domains, were prepared. Together, all three types of data sets comprised 72 entries. The HMM networks of the three domains were created by YEBIS independently.

To make the created HMM more detailed while keeping it understandable, we created the HMM iteratively with lambda value (= 0.10), which keeps the HMM network simple. In the first iteration of YEBIS, an HMM, which is composed of fewer nodes, is generated. Each node corresponds to one of the subdomains in given sequences. In the subsequent iterations, detailed HMMs corresponding to the subdomains are gradually determined. In this way, the three individual HMMs were created and at the same time their subdomains were identified.

Finally, the three created HMMs were concatenated and their parameters were re-optimized by the Baum-Welch algorithm to obtain the final HMM representing the sequences surrounding translation initiation sites.

3. Results

3.1. Modeling of sequences surrounding the translation initiation site by HMM

The created HMM representing sequences surrounding the translation initiation sites was composed of three
domains. Upstream HMM, Initiation codon HMM and Downstream HMM. In Upstream HMM, four distinct regions were identified in the first cycle in the iterative application of YEBIS.

### 3.1.1. Upstream HMM

Overview: From upstream we named the four identified regions, as RegionU1, RegionU2, RegionU3 and RegionU4. Table 1 shows the average length of respective regions which is calculated from the corresponding self-loop translation probability using Markov analysis. RegionU2 was an AG-rich region and RegionU4 was a CT-rich region preceding the translation initiation site mentioned by Sazuka and Ohara.5 The average nucleotide frequencies of respective regions are also listed in Table 1. The concentration of guanine in RegionU2 is remarkable. HMMs corresponding to respective regions are shown in Fig. 3. The Upstream HMM can be obtained by the concatenation of the four HMMs. The significance of each region is described below by interpreting respective HMMs.

#### RegionU1: The average length and nucleotide frequency in RegionU2 was 12.4 bp; nucleotides corresponding to the upstream 11.4 bp were represented by node Ul and the remaining 1.0 bp were represented by node U2 (Fig. 3a). The frequency of nucleotides in U1 (A: 0.29, C: 0.21, G: 0.36) differed from that in all non-coding regions (A: 0.29, C: 0.21, G: 0.36, T: 0.20), showing a lower frequency of guanines. The ratio of the guanines frequency in U1 to the frequency all non-coding regions (A: 0.29, C: 0.21, G: 0.20, T: 0.36) differed from that in all non-coding regions (A: 0.29, C: 0.21, G: 0.36, T: 0.20), showing a lower frequency of guanines. The frequency of nucleotides in U1 (A: 0.29, C: 0.21, G: 0.36) differed from that in all non-coding regions (A: 0.29, C: 0.21, G: 0.36, T: 0.20), showing a lower frequency of guanines.

### Table 1. The average length and nucleotide frequency in four regions upstream of translation initiation site.4a)

| Region    | U1 | U2 | U3 | U4 |
|-----------|----|----|----|----|
| Length    | 12.4| 5.7| 2.7| 4.2|
| Base contents | A: 0.29 | C: 0.21 | G: 0.36 | T: 0.20 |

```
| Nucleotide frequencies and lengths are listed for four regions upstream of the translation initiation sites, RegionU1, RegionU2, RegionU3 and RegionU4. |
```

![Figure 2. Stretches corresponding to RegionU2 in the 72 sequences surrounding translation initiation sites.](https://academic.oup.com/dnaresearch/article-abstract/4/1/1/372213/356407)

The 72 sequences surrounding translation initiation sites, determined by Sazuka and Ohara, are listed. A space is inserted before and after start codons. Stretches of sequences corresponding to RegionU2 identified in this study are written in lower-case letters.

---

4a) Nucleotide frequencies and lengths are listed for four regions upstream of the translation initiation sites, RegionU1, RegionU2, RegionU3, and RegionU4.

---

#### Table 1. The average length and nucleotide frequency in four regions upstream of translation initiation site.4a)

| Region    | U1 | U2 | U3 | U4 |
|-----------|----|----|----|----|
| Length    | 12.4| 5.7| 2.7| 4.2|
| Base contents | A: 0.29 | C: 0.21 | G: 0.36 | T: 0.20 |

```
| Nucleotide frequencies and lengths are listed for four regions upstream of the translation initiation sites, RegionU1, RegionU2, RegionU3, and RegionU4. |
```

---

4a) Nucleotide frequencies and lengths are listed for four regions upstream of the translation initiation sites, RegionU1, RegionU2, RegionU3, and RegionU4.

---

#### Table 1. The average length and nucleotide frequency in four regions upstream of translation initiation site.4a)

| Region    | U1 | U2 | U3 | U4 |
|-----------|----|----|----|----|
| Length    | 12.4| 5.7| 2.7| 4.2|
| Base contents | A: 0.29 | C: 0.21 | G: 0.36 | T: 0.20 |

```
| Nucleotide frequencies and lengths are listed for four regions upstream of the translation initiation sites, RegionU1, RegionU2, RegionU3, and RegionU4. |
```
Amino-Terminal Sequence Pattern Recognition by HMM

Figure 3. HMM of RegionU1 to RegionU4.
(a) HMM of RegionU1, b: HMM of RegionU2, c: HMM of RegionU3, d: HMM of RegionU4.

Figure 4. Initiation codon HMM.
Because every node has only one kind of nucleotide, the nucleotide assigned to a node is directly written in the circle representing the node.

vary from 2 bp (No. 2, 13, 22 and 78) to 18 bp (No. 58). The average length is 5.7 bp. There were 48 sequences in which RegionU2 was 5 bp or longer. The most frequent pattern among the 48 sub-sequences corresponding to RegionU2 was ‘AGGAG,’ which had its complement to 3’ region of 16S rRNA of this species. The prototype pattern was found in 12 of the 48 sequences.

RegionU3: RegionU3 was composed of only A and T with an average length of 2.7 bp (Fig. 3c). Main patterns of RegionU3 (probability of 0.82) were the repetition of T, represented by node U31 (T: 1.00) or the repetition of A and T, represented by a loop composed of U31 and node U32 (A: 1.00) after entering U31 (T: 1.00). Another pattern (probability of 0.18) was ‘A’, represented by node U33 (A: 1.00). It is presumed that RegionU3 is a spacer region between RegionU2 and RegionU4.

RegionU4: The length of RegionU4 was 4.2 bp. The last nucleotide neighboring an initiation codon was strongly biased to pyrimidines (probability of 0.86) as described by Sazuka and Ohara (Fig. 3d). The 5’ side nucleotide of initiation codons was represented by node U43 (A: 0.20, T: 0.80) or node U41 (C: 0.86, G: 0.14) with respective probabilities of 0.64 and 0.34, which were calculated by YEBIS, but not directly described in Fig. 6. U41 and U43 represent the bias to pyrimidines.

3.1.2. Initiation Codon HMM
The initiation codon HMM was a parallel connection of the four HMMs corresponding to four types of initiation codons, ATG, GTG, TTG and ATT, respectively (Fig. 4). Each initiation codon is a concatenation of three nodes, to each of which only one kind of nucleotide is assigned.

Estimated frequencies of initiation codons were, 0.891, 0.066, 0.023 and 0.022, respectively, which were represented as transition probabilities to respective initiation codons. The estimated frequencies were equal to the ac-
Table 2. Nucleotide distribution among 72 experimentally determined coding regions.a)

|                | The second amino acid (methionine processing) | The second amino acid (methionine unprocessing) | All amino acids (a) |
|----------------|-----------------------------------------------|-----------------------------------------------|--------------------|
|                | 1st   | 2nd   | 3rd   | 1st   | 2nd   | 3rd   | 1st   | 2nd   | 3rd   |
| A              | 0.32  | 0.00  | 0.20  | 0.50  | 0.53  | 0.36  | 0.25  | 0.32  | 0.19  |
| C              | 0.07  | 0.82  | 0.32  | 0.25  | 0.11  | 0.32  | 0.20  | 0.24  | 0.32  |
| G              | 0.45  | 0.13  | 0.05  | 0.14  | 0.04  | 0.14  | 0.39  | 0.17  | 0.21  |
| T              | 0.16  | 0.05  | 0.43  | 0.11  | 0.32  | 0.18  | 0.16  | 0.27  | 0.28  |

a) For the 72 experimentally determined coding regions, nucleotide distributions in each codon position were calculated.

The nucleotide distribution among the second amino acids before methionine processing in the case of methionine processing.

The nucleotide distribution among the second amino acids before methionine processing in the case without methionine processing.

d) The nucleotide distribution among all amino acids of the 72 coding regions, exclusive of initiation codons.

3.1.3. Downstream HMM

We found that the second amino acids before methionine processing could be classified into two types which differed in nucleotide distribution.

The first type was represented by nodes D1, D2 and D3 and the probability of the first type was 0.33; the second type was represented by nodes D1, D4 and D5 and the probability of the second type was 0.67 (Fig. 5). The discovery of these two types correlates well with the processing of the initial methionine, which was observed in this species by Sazuka and Ohara. In the translation process, all initiation codons are translated into formylmethionines, a portion of which are processed by methionyl-aminopeptidase after deformylation. They detected the processing of methionine in 44 of the 72 proteins. The nucleotide distribution among the second amino acids with and without methionine processing are shown in Table 2. In the table, the nucleotide distribution among all amino acids in 72 proteins are shown as well.

The nucleotide distribution in D1-D2-D3 (see Fig. 5) had a strong similarity to the nucleotide distribution among the second amino acids in the case of methionine processing, especially with regard to the abundance of cytosine in the second codon. The detection of peculiar nucleotide distribution represented in D1-D2-D3, which carries biological significance, proved the effectiveness of YEBIS to detect biologically significant nucleotide patterns and also suggested predictability of the processing of the initial methionine using the downstream HMM.

In this case, the nucleotide pattern represented by D1-D2-D3 was found in 28 of 72 nucleotide sequences. Translated protein of 25 of the 28 nucleotide sequences had the processed methionine, which was observed by Sazuka and Ohara.

There are two loops of nodes in the HMM, LoopD1, which is composed of D1, D2 and D3 in this order and LoopD2, which is composed of D6, D4 and D5 in this order. D6-D4-D5 represents a triplet of nucleotides, which encodes amino acids. The nucleotide distribution in D6-D4-D5 can be regarded as the distribution among amino acids near initiation sites. The nucleotide distribution in D6-D4-D5 was different from the distribution among all amino acids in 72 proteins with regard to the lower frequency of guanines. The guanine frequencies in D6, D4 and D5 were 0.28, 0.12 and 0.09, respectively (see Fig. 5), while the average guanine frequency in respective codon positions were 0.39, 0.17 and 0.21 (see Table 2) when all amino acids in 72 proteins were considered.

3.2. Validation of the created HMMs

We took each of the 72 sequences as a test set and took the remaining 71 sequences as a data set for each
test set. When the initiation codon of the test sequence was ATG, the parameters used in the HMM for this test were recalculated using the summary statistics from the remaining 71 sequences by YEBIS. When the initiation codon was a rare codon [No.9 (GTG), 11 (TTG), 32 (GTG), 37 (GTG), 44 (GTG), 81 (TTG), 86 (ATT)], other than ATG, the parameters used both in the upstream HMM and the downstream HMM were recalculated by the statistics from the remaining 71 sequences and the parameters used in the initiation codon HMM were not recalculated by considering a small number of rare initiation codons.

We succeeded in the recognition of proper initiation sites in the 71 sequences of 38 bp. The false-negative error was 1.4% (1/72). The false-positive error estimated from 7,200 sequences of 38 bp, which were randomly generated based on the nucleotide frequencies among 72 sequences of 38 bp, was 13.1% (942/7,200).

The translation initiation site that could not be recognized by the HMM was the site in sequence No. 86, which had ATT as its initiation codon. One possible reason for the failure to recognize the site may be that a nucleotide pattern in the upstream of the ATT is different from the nucleotide pattern of the other 71 sequences. It also suggests that ATT may be a different type of rare initiation codon from the type composed of GTG and TTG. However, this hypotheses should be re-examined when enough proteins with ATT as initiation codons have been experimentally determined to make the results statistically robust.

4. Discussion

4.1. Abundance of G in Region U2

As shown in Table 1, the frequency of guanines in the upstream region was lower than Region U2 (C: 0.07, A: 0.46, G: 0.47). As mentioned before, the guanine frequency of Region U1, the upstream of Region U2, was lower than the average guanine frequency in non-coding regions. Also as mentioned before the guanine frequency in amino acids near the initiation site was lower than the average frequency of all the coding regions of the 72 proteins. Therefore, it is presumed that this high accumulation of guanines in Region U2 makes Region U2 distinctive. It is also presumed that the distinguished guanine abundance serves for proper recognition of translation initiation sites in mRNA by ribosomes and also serves to avoid false recognition of translation initiation sites by the creation of pseudo-initiation codons owing to mutations. Especially, relatively lower contents of guanine in Region U1 are presumed to serve to avoid the false recognition in untranslated region of mRNA.

4.2. Relation between Shine-Dalgarno sequence and Region U2

The Shine-Dalgarno sequence has been used in the assignment of translation initiation sites of coding regions. Sazuka and Ohara selected three patterns GAGG, GGAG and AGGA as subsequences of the Shine-Dalgarno sequence and found 35 of 72 sequences contained one of the three patterns. It means we cannot identify 51.4% (37/72) of the translation initiation sites of coding regions in this case. Therefore, it is natural to infer that ribosomes recognize a more general pattern than the Shine-Dalgarno sequence.

Of the 35 sequences to which Sazuka and Ohara assigned GAGG, GGAG or AGGA, 34 sequences had one of the patterns in Region U2. The only exception was sequence 10, although its Region U2 contained another AGGA. Therefore, it is presumed that sequences in Region U2 contain generalized Shine-Dalgarno sequences if not all. However, further investigation is necessary when more translation initiation sites have been experimentally determined.

4.3. Prediction of translation initiation sites

The prediction of precise positions of translation initiation sites will be possible if the precise HMM corresponding to all coding regions and the HMM corresponding to all non-coding regions are added to the HMM extracted in this study. We had already constructed the two necessary HMM in our previous study.8 With the combined HMM of the three we are now investigating the precise prediction of translation initiation sites in Cyanobacterium Synechocystis sp. strain PCC 6803.

Acknowledgments: This work was supported by funds from the Kazusa DNA Research Institute (KDRI) and was performed to examine the effectiveness of a program YEBIS developed in the ALIS (Advanced Life-sciences Information System) project for genome analysis by the Japan Science and Technology Corporation (JST).

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., Adams, M. D. et al. 1995, The minimal gene complement of Mycoplasma genitalium, Science, 270, 397-403.
3. 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.
4. Bult, C., White, O., Olsen, G., Zhou, L., Fleischmann, R. D., Sutton, G. C., Blake, J. A., FitzGerald, L.
M., Clayton, R. A., Gocayne, J. D. et al. 1996, Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschi*, *Science*, 273, 1058–1073.

5. Sazuka, T. and Ohara, O. 1996, Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp. strain PCC 6803 by amino-terminal protein sequencing, *DNA Res.*, 3 225–232.

6. Bairoch, A. 1991, Prosite : A dictionary of protein site and pattern : User manual Release 7.00.

7. Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. 1994, Hidden Markov Models in Computational Biology. Applications to Protein Modeling, *J. Mol. Biol.*, 235, 1501–1531.

8. Yada, T. and Hirosawa, M. 1996, Recognition in cyanobacterium genomic sequence data using the Hidden Markov Model, *In the Proceedings of International Conference on Intelligent System for Molecular Biology (ISMB-96)*, 252–260.

9. Yada, T., Ishikawa, M., Tanaka, H., and Asai, K. 1996, Signal Pattern Extraction form DNA Sequences Using Hidden Markov Model and Genetic Algorithm, *Journal of Japan Information Procession Society*, 37, 1117–1129.

10. Baum, L. E., Petrie, T., Soules, G., and Weiss, M. 1970, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Mat. Stat.*, 41, 164–171.

11. Holland, J. 1992, Adaptation in Natural and Artificial Systems, MIT Press.

12. 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.

13. Feller, W. 1957, An Introduction to Probability Theory and its Applications, John Wiley & Sons.
