Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models

Tetsushi Yada¹, Mitsuteru Nakao², Yasushi Totoki¹ and Kenta Nakai³,*

¹Genomic Sciences Center, RIKEN, c/o Laboratory of Genome Database, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, ²Institute for Chemical Research, Kyoto University, Gokashou, Uji, Kyoto 611-0011 and ³Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Received on May 14, 1999; accepted on August 12, 1999

Abstract

Motivation: The hidden Markov model (HMM) is a valuable technique for gene-finding, especially because its flexibility enables the inclusion of various sequence features. Recent programs for bacterial gene-finding include the information of ribosomal binding site (RBS) to improve the recognition accuracy of the start codon, using this feature. We report here our attempt to extend the model into the total transcriptional unit, enabling the prediction of operon structures.

Results: First, we improved the prediction accuracy of coding sequences (CDSs) by employing the models of ‘typical’, ‘atypical’ and ‘negative (false-positive)’ classes as well as the models of RBS and its downstream spacer. The sensitivity of exactly predicting the 204 experimentally confirmed CDSs reached 90.2% in an objective test. Based on the prediction result of CDSs, the positions of the promoters and terminators were predicted. Our model could exactly recognize 60% of 390 known transcriptional units. Thus, the accuracy and significance of this prediction problem is far from trivial. We would like to propose this problem as an open theme in bioinformatics because the ongoing or planned post-sequencing projects will produce much data for future improvements.

Availability: The table of predicted transcriptional units of Escherichia coli will be distributed upon request.

Contact: knakai@ims.u-tokyo.ac.jp

Introduction

In this era of large-scale sequencing, there is no doubt that computational methods can be of great use to interpret newly determined DNA sequences. Amongst them, the prediction of coding sequences (abbreviated as CDSs) in bacterial genomes has been one of the most successful endeavors. The main principle of such a prediction is to detect the differences of sequence features, such as three-base periodicity, between coding and non-coding sequences. A classical but still powerful program, GeneMark by Borodovsky’s group, employs the Markov model technique to capture these sequence features effectively (Borodovsky and McIninch, 1993). Later, the hidden Markov model (HMM) technique, which is more powerful for modeling various sequence features, has been widely used in bioinformatics (reviewed in Krogh, 1998). In the field of coding-region prediction, a pioneering work was carried out by Krogh et al. (1994) for bacterial sequences and one triumph was brought by the GENSCAN program which beautifully modeled the eukaryotic gene structure and performs significantly better than previous methods (Burge and Karlin, 1997). Recently, an HMM version of GeneMark, GeneMark.hmm, was also released and it was confirmed that GeneMark.hmm outperforms the original version, especially in its ability to exactly predict the region boundaries (Lukashin and Borodovsky, 1998). Both GENSCAN and GeneMark.hmm employ a special type of HMM architecture known as ‘HMM with duration’ (Rabiner, 1989), where the length distribution of each region including spacers can be integrated. If one further exploits this feature, he or she can naturally include various signals such as promoters in one’s model. However, this possibility has not been pursued especially for bacterial analyses although only the ribosomal binding site (RBS) has been included in some programs to improve the recognition accuracy of the N-terminus of CDSs (Frishman et al., 1998; Lukashin and Borodovsky,
1998). Nevertheless, we believe that the theme of comprehensive modeling of the total bacterial gene structure has its own practical value as well as its theoretical importance in organizing our knowledge because such studies enable us to predict the transcriptional unit (TU) for given open reading frames (ORFs); it is well-known that a bacterial TU can contain multiple CDSs, in other words, a bacterial TU can constitute the operon structure. (Historically, the word operon has not always been used with the same meaning as TU but we do not distinguish between these terms in this paper.) We are entering the post-sequencing age for bacterial genomes; for example, systematic experiments constructing the transcriptional map of the *Bacillus subtilis* genome in various conditions are ongoing (N.Ogasawara, personal communication); DNA chip technology is also beginning to be applied to monitoring the expression of bacterial genes (de Saizieu et al., 1998). Thus, we will have plenty of data for the verification of our theory in the near future. Moreover, if we can predict the operon structure of bacterial genes, it will be useful to deduce the function of uncharacterized ORFs because functionally related bacterial genes are co-transcribed in many cases. In this paper, we report our first attempt to model the gene structure of *Escherichia coli* towards this direction.

**System and methods**

**Collection of annotated sites**

We used the annotation of *E.coli* genome by Blattner et al. (1997), i.e. the information of 4287 CDSs and 2593 TUs, in which 390 of them are described as ‘documented’ (the genome sequence and the specification of coding regions were obtained from http://www.genetics.wisc.edu and three pieces of CDS information which contained in-frame internal stop codon were discarded). To further evaluate the prediction accuracy of start codons, the data of 205 N-terminal start sites of CDSs (Link et al., 1997) were also used (http://twod.med.harvard.edu). To develop the model of promoters, we used the compilation of 441 σ^70^-dependent promoters (Ozoline et al., 1997), kindly sent from the author by e-mail. To construct the score function of ρ-independent terminators, 145 terminators from Carafa et al. (1990) were used.

**Model building**

**Basic architecture.** In Figure 1, the overall architecture of our HMM is shown. Each strand of the genome is modeled as a set of TUs separated by spacers. A TU starts with a promoter and ends with a terminator; it can contain one or more CDSs, each of which has an upstream RBS, part of which is known as the Shine–Dalgarno sequence. CDSs, promoters, and RBSs are further modeled as subHMMs and the probability to being a ρ-independent terminator is calculated beforehand (see below). Spacer regions except the inter-TU spacer which is treated as a dummy are modeled based on their base contents and their approximated length distribution (see ‘Results’). Since we could not collect a sufficient amount of ρ-dependent terminators, the statistics of length distribution of 3′ UTR (untranslated region) was taken only in cases of ρ-independent termination. As described below, the consistency between the two strands was considered in a post-processing step.

**Coding regions (CDSs).** The model of coding regions was constructed from the 4287 regions tabulated by Blattner et al. (1997). The statistics of start codons and the frequency of all possible di-codons were used to build an initial model in a way similar to that built for GeneHacker (Yada and Hirosawa, 1996). Unlike GeneMark.hmm (Lukashin and Borodovsky, 1998), we did not explicitly use the information of length distribution of CDSs but we included the model of RBS (see below) and its downstream spacer in our HMM. Using this model, the coding potential of 4287 CDSs was self-evaluated. Based on the distribution of obtained scores, these CDSs were classified into either ‘typical’ or ‘atypical’ classes (25% from the lowest score were included in the latter class). Then, the model was reconstructed for each class. These HMMs were combined in parallel and were applied to the entire genomic sequence. Note that our model in this stage did not involve the inter-CDS spacers and thus the parsing was done ‘locally’, allowing the detection of overlapped CDSs. For entirely overlapped CDSs, that is, those CDSs that totally include (or are totally included by) other CDSs, we eliminated the one with the lower score. Next, the detected regions other than the original

---

![Fig. 1. Basic architecture of our HMM.](image-url)
4287 regions were regarded as false-positives and were used to construct a ‘negative’ model (models for their dummy RBSs and the downstream spacers were also constructed). Finally, the genome sequence was again scanned by the former ‘typical/atypical’ model and the detected candidates were further checked by the ‘negative’ model. The decision was made using a linear discriminant function constructed between the ‘typical/atypical’ and the ‘negative’ data. To assess the prediction accuracy for unknown sequence data, a cross-validation test was performed. That is, the entire genome was fragmented into 20 pieces and each piece was used for testing while others were used for training. The results for 20 trials were summed.

Ribosomal Binding Sites (RBSs). To obtain the model of RBSs, the upstream 25 bp regions of 4287 CDSs were simultaneously aligned, inhibiting internal gaps according to the algorithm of Hiroswa et al. (1995). Conserved positions were then evaluated by the $\chi^2$ test (cut-off value: 500.0). The base content and the length distribution of the spacer between the RBS and its downstream start codon were then calculated.

Promoters and terminators. The model of $\sigma^{70}$-dependent promoter was constructed as described in Yada et al. (1997) from the 441 samples (the cut-off value of 35.0 was used for the $\chi^2$ test). The models of promoters recognized by other $\sigma$-factors were not included in the current model.

The terminators of E.coli are classified into two groups based on the dependency on the $\rho$-factor. We used a prediction method for $\rho$-independent terminators, which is an extension of Carafa et al.’s method (Carafa et al., 1990). The details of the prediction method will be published elsewhere (M.Nakao and M.Kanehisa, in preparation) but for an outline it uses a linear discriminant function which combines seven variables. These variables reflect the stem–loop feature and the accompanying U-cluster feature of $\rho$-independent terminators. Four variables were obtained from the stem–loop structure calculated using the MFOLD program of Zuker (1989): the free energy of the structure, the stem length, the loop length, and the number of C–G pairs in the stem. The other three variables were characteristics for the uridine-residue stretch located downstream of the stem: the score of the above-mentioned Carafa et al.’s method, the distance between the stem and the stretch, and the length encompassing the whole stem–loop and U-rich region. According to a jackknife (leave-one-out) test, its accuracy was estimated to be almost perfect (99.7% selectivity and 99.8% sensitivity), at least for the 145 positive/4176 negative dataset. It is not easy to estimate its accuracy for genomic sequences, but the error rate (varying from 73 to 91%) for each range of discriminant score was calculated using the number of apparent false-positives observed within the CDSs. Thus, the estimated probability of being a $\rho$-independent terminator is assigned at each position on both strands in a pre-processing calculation.

On the other hand, we could not prepare a reasonable prediction method for the $\rho$-dependent terminators mainly because of the insufficient amount of their known examples. Therefore, we employed a simple rule that a CDS which has a longer spacer region between its stop codon and the downstream RBS than a given cut-off distance should have a terminator. This rule was applied independently from the calculation of $\rho$-independent terminators mentioned above. After some tests, we set the cut-off distance to be 40 bp.

Operons. To model the possibility that a TU can contain multiple CDSs, the distribution of the number of CDSs per TU was examined. In addition, the base content and the length distribution of inter-CDS spacers were calculated for both documented and predicted operons tabulated by Blattner et al. (1997). For a control experiment of the operon-structure prediction, a simple prediction scheme was used: coding region(s) on the same strand are regarded to be within the same TU if the lengths of their spacers are within a cut-off length (see ‘Results’).

Prediction algorithm
Since the prediction accuracy of CDSs is significantly higher than the accuracy of predicting promoters and/or terminators, we adopted a two-step approach: (1) predict CDSs including their preceding RBSs. As described earlier, this prediction was done ‘locally’, i.e. the model of inter-CDS spacers was not used such that the detection of overlapped CDSs is possible. (2) predict TUs by the standard Viterbi algorithm under the constraint of the predicted CDSs. Since this prediction was done ‘globally’ (i.e. the model of inter-TU spacers was included), only one optimal model was obtained for the entire genome. We eliminated the TU with a lower score when it totally included (or was totally included by) another TU. In this prediction of TU structures, we did not separate the testing data because its performance was not sufficient.

Results
Obtained models
The base contents and the length distribution of spacer models are summarized in Table 1 and Figure 2, respectively. Since their sample sizes differ greatly, it is difficult to draw definite conclusions, but all of them appear somewhat A/T-rich in contrast with even uses of four bases in the averaged base content of the entire genome. The length distribution of the spacer between RBS and its downstream start codon shows that RBSs are most likely
located at the positions around 6 bp upstream from the start codon. The consensus representation of the obtained RBSs is ‘(t/a)(a/c)AGGA(g/a)’ (note that they are more precisely represented in our HMM). Both of these results agree well with previous observations (Frishman et al., 1998; Lukashin and Borodovsky, 1998). Significant differences are observed in the length distribution between the 5′ UTR (i.e. the spacer just downstream the promoter), the co-transcribed inter-CDS spacer, and the 3′ UTR (i.e. the spacer just upstream the ρ-independent terminator), although their sample numbers vary and are not plentiful. The inter-CDS spacer lengths are very short in most cases but they sometimes exceed 200 bp while the lengths of 5′ UTRs and 3′ UTRs have a peak at around 25 and 55 bp, respectively. All of these distributions are approximated by the curves shown in the graphs.

Figure 3 depicts the obtained model of σ70-dependent promoters. In addition to the classical −35 and −10 elements, additional A/T-rich upstream elements and the TG-box upstream of the −10 region, both of which are consistent with previous observations (Ozoline et al., 1997), were automatically included in our HMM. On the contrary, significantly conserved regions were not detected in the downstream regions (DSRs) although there has been a report on the effect of DSRs (Brunner and Bujard, 1987). In Figure 4, we show the number distribution of CDSs contained in a TU for both ‘documented’ and ‘predicted’ TUs by Blattner et al. The tendency that the predicted TUs contain fewer CDSs can be seen, justifying the necessity of our study aimed at predicting the TU structure more accurately.

![Figure 2](image2.png)

**Fig. 2.** Length distribution of various spacers. Any overlapping pair of CDSs were merged into one CDS to take the statistics. (a) Spacer after RBS, (b) 5′ UTR spacer, (c) inter-CDS spacer within a TU, (d) 3′ UTR spacer.

<table>
<thead>
<tr>
<th>Spacer</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Amount (nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After RBS</td>
<td>32.9</td>
<td>21.2</td>
<td>19.3</td>
<td>26.6</td>
<td>27 790</td>
</tr>
<tr>
<td>5′ UTR</td>
<td>28.3</td>
<td>22.5</td>
<td>23.6</td>
<td>25.7</td>
<td>39 620</td>
</tr>
<tr>
<td>Within a TU</td>
<td>25.7</td>
<td>23.3</td>
<td>23.2</td>
<td>27.8</td>
<td>18 564</td>
</tr>
<tr>
<td>3′ UTR</td>
<td>24.2</td>
<td>23.4</td>
<td>23.6</td>
<td>28.9</td>
<td>10 373</td>
</tr>
<tr>
<td>Whole genome</td>
<td>24.6</td>
<td>25.4</td>
<td>25.4</td>
<td>24.6</td>
<td>4 639 221</td>
</tr>
</tbody>
</table>

The values for the ‘whole genome’ are taken from the plus strand of the *E.coli* genome.

![Table 1](image1.png)
Prediction of *E. coli* operon structure

**Fig. 3.** Consensus representation of our model for σ70-dependent promoters. The subscript numbers are the percentages of the base content. Bases are shown in capital letters when their content exceed 55%. The numerals in parentheses show the length of variable regions; for example, “t29 (3,5)” means that the length of this content can vary from 3 to 5 bp. Known conserved regions are boxed.

**Estimation of prediction accuracy**

The entire computation time required for the modeling and the prediction was about 24 h using a single processor (Ultra SPARC 250 MHz). In Table 2, the performance of our program to predict CDSs is shown. For ‘answer’ sets, both the annotated CDSs over the entire genome (Blattner *et al.*, 1997) and the experimentally confirmed subset (Link *et al.*, 1997) are used. The results are compared with those of GeneMark.hmm (Lukashin and Borodovsky, 1998). It can be seen that the performance of our program is slightly better than that of GeneMark.hmm. In particular, our program shows 2.7% better accuracy in exactly predicting Link *et al.*’s 204 CDSs the N-terminus of their protein products being experimentally determined. Furthermore, our program produces as much as 5.9% fewer ‘wrong’ genes for Blattner *et al.*’s data even if some of them may turn out to be ‘correct’ in the future.

The performance assessed by the cross-validation procedure is 95.7% (4103/4287) sensitivity (true positive/actual positive) and 95.4% (4103/4299) specificity (true positive/predicted positive) when we consider the CDSs with correctly predicted stop codon to be ‘correct’. These accuracies become about 1–2% higher when the test set is not distinguished from the training set. For the 204 genes of confirmed N-terminus, the sensitivity was 90.2% (184/204) for the ‘exact’ prediction requiring both the start codon and the stop codon to be precisely predicted (note that the specificity cannot be defined in this case).

The result of the prediction of TUs is summarized in Table 3. The 4287 CDSs were mapped into 2381 TUs. For the documented 390 TUs, 60% were correctly predicted. The percentage of TUs with falsely predicted terminators was higher than that of TUs with falsely predicted promoters. For the control experiment, we tried a simple prediction by a criterion that the boundaries of TUs are placed at the position where the inter-CDS spacer length on the same strand exceeds a given cut-off value (we assumed complete knowledge of correct CDS positions). By varying the cut-off values, we obtained an optimized prediction accuracy of 53.3% for the cut-off value of 30 bp. Thus, our prediction level is significantly higher than that of the control experiment even if complete knowledge of CDSs is assumed in the control.

**Discussion**

In this work, we constructed an HMM of the *E. coli* gene structure which allowed us to predict its coding sequences as well as its transcriptional units (operon structures). The prediction of CDSs directly uses the information of upstream RBSs and the distance distribution between RBS and the start codon. Since the parsing of the model is done ‘locally’, overlapping CDSs can be naturally detected. Furthermore, a new idea to construct the model of negative examples, most of which are expected to be false-positives, was effectively used, in addition to the idea of dividing positive CDSs into typical and atypical classes, which was used in GeneMark.hmm and GeneMark-Genesis (Hayes and Borodovsky, 1998; Lukashin and Borodovsky, 1998). Therefore, its performance was excellent; by an objective cross-validation test, both the sensitivity and the specificity exceeded 95% in correctly predicting the stop codons of all CDSs within the genome. For CDSs with experimentally confirmed N-terminus, our HMM outperformed GeneMark.hmm. There are other new programs for the prediction of bacterial CDSs. For example, there is the GLIMMER program which employs an interpolated Markov model (Salzberg *et al.*, 1998). Since its prediction accuracy for *E. coli* genes is not reported, direct comparison is difficult. But the strength of GLIMMER seems to be the simplicity of its training processes rather than the prediction accuracy itself. There is another program, ORPHEUS, which extensively uses the homology information (Frishman *et al.*, 1998). In this case also, direct comparison is difficult, especially because the reported performance has been revised after publication (http://pedant.mips.biochem.mpg.de/orpheus/corrigendum.html). But it seems that the performance of ORPHEUS for the entire *E. coli* genome does not exceed our performance. In conclusion, the
Table 2. Prediction accuracy of CDSs

<table>
<thead>
<tr>
<th>Answer Ref.</th>
<th>No. of genes</th>
<th>Method</th>
<th>Exact prediction (%)</th>
<th>Only 3'end prediction (%)</th>
<th>Missing genes (%)</th>
<th>Wrong genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blattner&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4287</td>
<td>our method</td>
<td>75.0</td>
<td>21.7</td>
<td>3.3</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>4288</td>
<td>GeneMark.hmm</td>
<td>75.4</td>
<td>19.6</td>
<td>5.0</td>
<td>8.2</td>
</tr>
<tr>
<td>Link&lt;sup&gt;b&lt;/sup&gt;</td>
<td>204</td>
<td>our method</td>
<td>90.2</td>
<td>9.3</td>
<td>0.5</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GeneMark.hmm</td>
<td>87.5</td>
<td>12.0</td>
<td>0.5</td>
<td>–</td>
</tr>
</tbody>
</table>

<sup>a</sup>Blattner et al. (1997).  <sup>b</sup>Link et al. (1997).

The classification of prediction results and the values for ‘GeneMark.hmm’ (the version which includes the RBS identification procedure) are taken from Lukashin and Borodovsky (1998).

Table 3. Prediction accuracy of TUs

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both boundaries</td>
<td>59.2% (231/390)</td>
</tr>
<tr>
<td>Divided</td>
<td>7.9% (31/390)</td>
</tr>
<tr>
<td>Only 5’-boundary</td>
<td>17.4% (68/390)</td>
</tr>
<tr>
<td>Only 3’-boundary</td>
<td>11.5% (45/390)</td>
</tr>
</tbody>
</table>

The ‘Divided’ row means the cases where a TU containing two CDSs are predicted to be two TUs; these cases are not counted in ‘Only 5’/3’-boundary’.

Fig. 4. Distribution of numbers of CDSs involved in a TU for both ‘documented’ (—) and ‘predicted’ (- - - - -) TUs listed in Blattner et al. (1997).

Based on the prediction result of CDSs, we further proceeded to the modeling of gene structure, enabling the prediction of transcriptional units (operon structures). Although all of the CDSs listed in Blattner et al. (1997) were provided with the information of predicted operon structure, this is the first attempt to predict objectively the bacterial TUs and to evaluate its accuracy, as far as we know. The difference of CDS number distribution between ‘documented’ and ‘predicted’ operons, shown in Figure 4, partly rationalizes our attempt. This newly set prediction problem turned out to be far from trivial; the prediction accuracy of our method, assessed using the documented operons, was not more than 60%, although this figure is significantly better than the performance based on the inter-CDS distances only. The reasons for this might come from two directions: methodological or biological. Methodologically, our method cannot detect overlapped TUs and only reports the optimal model because it employs a ‘global’ Viterbi parsing strategy. Moreover, the difficulty of detecting signal patterns compared with detecting statistical coding potential has been shown (Staden, 1990). Although many algorithms have been developed to recognize promoters and (ρ-independent) terminators, they were not intended to scan over the whole genome and are likely to yield a number of false-positives. From the biological point of view, our model ignores several important points. For example, it ignores the fact that many promoters are regulated by non-σ<sub>70</sub> sigma factors. The fact that every promoter has its inherent ‘strength’ is also ignored. Such promoter ‘strength’ can vary quite drastically between different promoters and a promoter which is closer to the consensus pattern tends to be ‘strong’. Therefore, we checked the correlation between the promoter scores and the predictability of the corresponding operons; the 390 documented operons were sorted in the descending order of their promoter scores and were divided into four classes containing 100, 100, 100, and 90 data items, respectively. The prediction rate for each class was 64, 68, 62, and 37%, respectively. It can be interpreted that the presence of weak promoters makes the prediction more difficult. In addition, we used a rather unrealistic model of ρ-dependent terminators, ignoring their known sequence features (Alifano et al., 1991). A simple interpretation of the result in Table 3 is that the current state of terminator prediction is more difficult than that of promoter prediction. However, a careless
incorporation of complex models has some dangers in degrading the prediction accuracy. In any case, both the promoters dependent on minor σ factors and the ρ-dependent terminators do not occupy the majority. Perhaps the most important fact we are ignoring now is the regulatory aspects of transcription. A promoter or a terminator does not always work; it can be regulated by some activators or repressors. In fact, according to a Japanese project to construct the transcriptional map of *B. subtilis*, it seems very common that a transcript can be started from, or terminated at, multiple positions (N. Ogasawara, personal communication). Perhaps many of the operons in the ‘documented’ set do not reflect the optimal state of transcription if there are no additional regulatory proteins. Thus, our next challenging theme should be the evaluation of multiple forms of transcripts in terms of the result of ‘local searches’, although for ‘local searches’ there remains the problem of how to distinguish true-positives from false-positives.

Finally, we hope that the prediction of bacterial TUs will become an open problem in the community of bioinformatics and that further studies on this theme will lead us to a better understanding of gene expression through sequence data.

Acknowledgements

This work was partly supported by the Special Coordination Funds for promoting science and technology of STA. K.N. is also supported by Research for the Future Program of JSPS and a Grant-in-Aid for Scientific Research on Priority Areas of MESSC in Japan.

References


