Functional Prediction of *B. subtilis* Genes From Their Regulatory Sequences

Tetsushi Yada[1]
Japan Science and Technology Corporation
5-3 Yonbancho, Chiyoda-ku,
Tokyo 102, Japan
yada@tokyo.jst.go.jp

Yasushi Totoki
Information and Mathematical Science Laboratory, Inc.
2-43-1 Ikebukuro, Toshima-ku,
Tokyo 171, Japan
totoki@imslab.co.jp

Takahiro Ishii and **Kenta Nakai**
Institute for Molecular and Cellular Biology, Osaka University
1-3 Yamada-oka, Suita 565, Japan
{ishii,nakai}@imcb.osaka-u.ac.jp

Abstract

In bacterial cells, gene expression is regulated by multiple sigma factors, each of which has its promoter specificity, according to their conditions. Thus, if we can discriminate which sigma factor binds to the upstream region of a given coding sequence, we can predict in what condition it will be expressed. In this paper, we show this approach is feasible for the analysis of *Bacillus subtilis* genome. Based on our collection of known promoter sequences, we prepared 8 predictors to characterize known sigma factors using the hidden Markov model and their prediction accuracies were estimated with a cross-validation test. Furthermore, we predicted the sigma-dependencies for each of 1415 candidate genes in the genome. Our prediction results are experimentally testable and seem useful for the post-sequencing project.

Introduction

Since a number of bacterial genomes have been sequenced, there are great demands for practical computational methods to interpret their biological contents. Although it is relatively easy to locate their candidate genes (Yada & Hirosawa 1996), it is very difficult to infer their function when no similar sequences were found in databases. In this paper, we propose a new approach for the interpretation of bacterial genome sequences: prediction of gene function from its regulatory sequences (a pioneering work has been done in yeast (Fondrat & Kalogeropoulos 1994)).

It is well-known that all bacterial genes are not expressed all the time; their expression is regulated with respect to various conditions. As for its molecular mechanism, the use of multiple sigma factors seems the most fundamental. Sigma factor is a subunit of RNA polymerase that determines its promoter specificity (Lewin 1994). For example, when a bacterium is in the condition of high temperature, it synthesizes a specific heat-shock sigma factor and then RNA polymerases synthesize the mRNAs of genes dependent on this factor. Thus, if we can predict which genes are dependent on a specific sigma factor from the sequence, we can predict the condition of their expression regardless of our knowledge on their coding region. In *Bacillus subtilis*, one of the most well-studied bacteria, 9 sigma factors have been cloned as listed in Table 1 (Haldenwang 1995), in addition to at least 3 other potential sigma factors (N. Ogasawara, personal communication). Unlike *Escherichia coli*, *B. subtilis* undergoes a morphological change called sporulation (Figure 1). When starved, the cell first becomes compartmentalized into two parts, the mother cell and the forespore, and subsequently the latter becomes the dormant spore. During these processes, there is a cascade of gene activation using multiple sigma factors. Thus, several classes of genes must be specific for developmental stages and/or cell types.

Table 1: Functions of Genes Dependent on *Bacillus* Sigma Factors *

σ factor	Function
σ^A	Housekeeping / early sporulation
σ^B	General stress response
σ^C	Expressed in postexponential phase
σ^D	Chemotaxis/autolysin/flagellar related
σ^E	Expressed in early mother cell
σ^F	Expressed in early forespore
σ^G	Expressed in late forespore
σ^H	Expressed in postexponential phase; competence and early sporulation
σ^K	Expressed in late mother cell
σ^L	Degradative enzymes

* Adopted from (Haldenwang 1995) with modification

Although computational recognition of bacterial promoters has been studied for many years, most of them were on *E. coli* promoters dependent on σ^{70}. *Bacillus* promoters for all sigma factors appear to share their basic architecture with this class of promoters; there are two relatively conserved sequence elements around the positions -35 and -10, respectively, and their distance is rather long, varying in several bases.

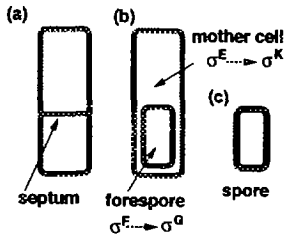

Figure 1: Sporulation processes of *B. subtilis*. (a) First, the septum divides the cell into the mother cell and the forespore. (b) Next, the spore is engulfed and is surrounded by the spore coat. During these processes, some stage- and cell-type-specific sigma factors are activated. (c) After the lysis of the mother cell, the spore is released.

Because of the existence of this large gap, it is difficult to directly apply standard pattern-recognition techniques such as the weight-matrix method. In this study, we used the hidden Markov model, HMM (Krogh *et al.* 1994; Yada, Sazuka, & Hirosawa in printing), with which we can naturally treat the gap region. Since another difficulty for promoter recognition is the existence of extensive false positives (Horton & Kanehisa 1992), we tried to avoid this problem by restricting the search region. In this paper, we report how the discrimination of sigma-factor binding sites can be useful to predict the gene function identified in a large region of genomic sequence.

Materials and Methods

Data Collection and Preprocessing

A set of promoter sequences were collected for each sigma factor from the literature (Table 2). Most of the collected promoters were experimentally verified but some of them were adopted by sequence homology only. To make up for the small size, promoter sequences for homologous factors of other bacteria were also included for σ^B, σ^D, and σ^L. Our collection were used as the training data and are available upon request.

As for the test data, upstream non-coding sequence segments starting from position -1 of any annotated gene (ORF, tRNA, or rRNA) of length 200 bp at most were collected from the 1.4 Mbp genomic sequence determined by the Japanese *Bacillus* genome project. When there is no intergenic sequence in the upstream region, it is treated as "no promoter" (173 cases). Most of them are likely internal genes within operons. The total number of the test data was 1415. Their matching with the training data was examined using the BLASTN program (Altschul *et al.* 1990).

To construct HMMs, collected promoter sequences were multiply aligned using the tree-based round-robin iterative algorithm (Hirosawa *et al.* 1995). Then, the significance of conservation was tested at each position using a χ^2 test. The threshold significance used was

Table 2: Summary of Collected Promoters. In the 2nd column, numbers after the plus sign means the ones adopted from other species; In the 3rd column, typical conserved segments corresponding to the -35 and -10 regions are shown rather arbitrary.

σ factor	Number of data	Consensus pattern -35 spacer (bp) -10
σ^A	142	TTGA 14 TGNTATAATA
σ^B	10+ 1	GTTT 16 GGGTAT
σ^D	11+ 18	MTAAAST 12 TGCCGATAW
σ^E	18	KCATANT 14 CATACANT
σ^F/σ^G	15	GNATAA 17 CANANTA
σ^H	10	AGGATNT 14 GAAT
σ^K	13	ACM 16 CATANNNT
σ^L	3+ 8	TGGCAC 5 TTGCNT

0.1 % for σ^A-dependent promoters, whose number is exceptionally large, while 1.0 % was used for the others.

Hidden Markov Model

Based on the alignment of the maximum segment, an HMM was built for each sigma factor. Its basic architecture is shown in Figure 2. It is a tied, left-to-right HMM without internal loops. Corresponding to position i of the alignment where the conservation is significant, two kinds of states, a match state M_i and a deletion state D_i were prepared. On the other hand, in position j where the conservation is insignificant, an insertion state I_j and a deletion state D_j were prepared. M_i and I_j output the four symbols A, C, G, T. To avoid over-learning, the output symbol distribution of insertion states were tied and their initial values were set to the base frequency of each insignificant region. Other initial parameter-values were also defined from the conditional probabilities calculated from the alignments in the way described in Yada *et al.* (in preparation) and they were further optimized using the Baum-Welch algorithm. Note that our model is designed to restrict the number of adjustable parameters to almost the same level with a weight matrix when transitions to the insertion states are not occurred in the significant regions.

Given a sequence segment, every subsegment whose logarithmic likelihood exceeds a pre-defined cut-off value is reported for each HMM using "local search", based on the Viterbi algorithm. To compare the values for HMMs corresponding to different sigma factors, the likelihood $x_{i,j}$ of a subsegment j detected by HMM i is transformed to a z-score $z^H_{i,j}$ ($z^H_{i,j} = (x_{i,j} - \bar{X}_i)/s_i$, where the mean value $\bar{X}_i$ and the standard deviation s_i were calculated from the genomic data). When HMMs are applied to potential regulatory regions in the genome, this z-score was modulated according to the distance from the starting site of the downstream coding region. We assumed that the distribution of

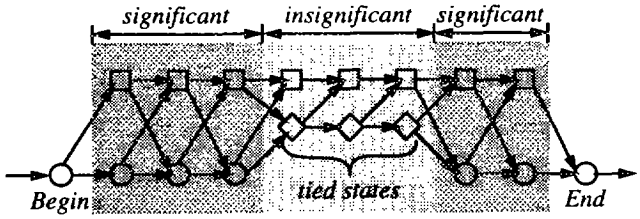

Figure 2: Architecture of HMM. Circles, squares, and diamonds represent match, deletion, and insertion states, respectively.

5' UTR length, *i.e.*, the length of untranslated region between the initiation sites of transcription and translation, follows the Poisson distribution (see Results). Thus, for a subsegment at the position of distance L_j, the z-score for the positional effect is calculated with this formula, $z^D_{i,j} = (L_j - \mu_i)/\sqrt{\mu_i}$, where μ_i is the average distance between the 3' end of HMM i and the 5' end of the downstream coding region. Finally, we defined the score z_i of sigma factor i for a given gene as $z_i = \max_j \{z^H_{i,j} - |z^D_{i,j}|\}$.

Estimation of Predictability

To estimate the discrimination ability for these HMMs, a cross validation test was performed. For σ^A- and σ^D-dependent promoters, one-tenth of the data and 2 sequences, respectively, were randomly chosen as a test set in each trial. These trials were repeated until each sequence is selected 10 times on average. For the other promoters, 1 sequence was used for a trial. The predictability was evaluated both from the sensitivity and the specificity. For the sensitivity evaluation, two criteria were used. In the "Approximate" criterion, cases are counted to be correct when the correct HMM marks a larger likelihood than its minimum value in the training data. In the "Rigorous" criterion, cases are counted to be correct when the correct HMM marks the maximum z-score among the scores of all HMMs and when it satisfies the former condition. For the specificity evaluation, averaged number and standard deviation of apparently falsely-responded HMMs per gene is calculated for each class of data although our data are still likely to include yet-uncharacterized promoters.

Results and Discussion

Estimation of Predictability

Among the sigma factors in Table 1, σ^C was not considered because its *in-vivo* recognition sites are still unknown, while the σ^F- and σ^G-dependent promoters were combined into one class because their recognition sites seem to largely overlap. Table 2 summarizes our data used. Except for σ^A- and σ^D-dependent promoters, their data sizes are rather small and data for σ^D and σ^L largely include the sequences from other species. Some positions other than the so-called -35 and -10 regions were also conserved in σ^A- and σ^D-dependent promoters which are large in size. The conservation degree of the σ^L-dependent promoters appeared to be the highest. Based on the derived alignments, 8 HMMs were constructed. The changes of the parameter values from their initial values were rather small during the optimization procedure.

Table 3: Results of Cross Validation.

σ factor	Sensitivity		Specificity
	Approx.	Rigorous	
σ^A	99.2 % *	83.0 % *	0.99 ± 0.93
σ^B	54.6 %	54.6 %	0.82 ± 0.39
σ^D	87.9 %	83.5 %	0.83 ± 0.84
σ^E	55.6 %	55.6 %	1.33 ± 0.75
σ^F/σ^G	60.0 %	60.0 %	1.73 ± 0.93
σ^H	50.0 %	50.0 %	1.10 ± 0.70
σ^K	61.5 %	61.5 %	1.46 ± 0.50
σ^L	72.7 %	72.7 %	0.73 ± 0.75
Total	85.2 %	75.5 %	1.05

* Standard deviations on approximate and rigorous criteria are 2.7% and 8.3%, respectively.

The cross validation test was performed as described above. The alignments were rather stable in many cases when sequence(s) of test data were excluded. In Table 3, the results for the sensitivity are shown. The results were identical in the two conditions except for σ^A and σ^D and the ranking of sensitivity roughly reflects the ranking of data size. Since we do not know the relative frequency of various promoters in the genome and since we have not characterized all sigma factors yet, the exact estimation of total predictability is difficult. If we simply estimate it by the ratio of correct predictions, they were 85.2 % and 75.5 % for the "Approximate" and "Rigorous" conditions, respectively (for σ^A and σ^D, average numbers were used). Table 3 also shows the result for the specificity evaluation. The σ^L promoters, again, show a marked result probably due to their strong conservation. All of sporulation-specific promoters (σ^E, σ^F/σ^G, and σ^K) show lower specificity. It is likely that their recognition sequences more or less overlap like the σ^F- and σ^G-dependent promoters which we treated as one group. Such similar recognition sequences clearly provide the bacteria with a subtle way of controlling the gene-expression level in sporulation. Considering the possibility of multiple recognition-sites, our result seems rather satisfactory. Moreover, since the σ^A and σ^D, both of which are relatively abundant in data sizes, show better sensitivity, we can expect that the future growth of our training data will improve the predictability significantly.

Table 4: Sample Prediction Result for Each Sigma Factor.

σ factor	Rank	Score	Gene	Description
σ^B	11	2.35	*yfhN*	expression induced by environmental stress
σ^D	16	1.81	*tlpC*	methyl-accepting chemotaxis protein.
σ^E	5	2.68	*nucB*	sporulation-specific extracellular nuclease precursor.
σ^F/σ^G	12	2.20	*yqjW*	similar to *SamB* protein for UV protection and mutation
σ^H	10	1.97	*yyaA*	strong similarity to *spo0J*
σ^K	2	2.72	*gerKA*	spore germination protein *GerKA*
σ^L	2	4.33	*yfjK*	probable acetoin dehydrogenase subunit

Application to Genomic Data

We applied the obtained HMMs to the sequences of upstream region taken from the genomic data. We first examined the positional distribution of 67 promoters involved in both the training data and the genomic data. 85 % of the data were distributed under the 5'UTR length 125 and the average in this range, 45.7, seems to fit the Poisson curve well. Based on this observation and the fact that about 50 bp is needed for HMMs, we used a maximum length 200 bp when we extract the upstream non-coding sequence.

Each upstream sequence was evaluated by 8 HMMs and their z-scores were subtracted by the z-scores representing the positional effect; we confirmed that the result is not drastically changed even if we omit this term (data not shown). In Table 4, a typical result not involved in the training set is shown for each sigma factor except σ^A. For example, since σ^D mainly regulates genes related to chemotaxis, the prediction that the *tlpC* gene is regulated by σ^D seems very likely. Although some genes were apparently contradicting our training data, it is possible that they are only the results for unknown alternative promoters.

Concluding Remarks

Previous attempts to detect the binding sites of various sigma factors have been "at best problematic and sometimes misleading" (Harwood & Wipat 1996). In this study, to some extent, we could overcome this situation (1) by collecting as many promoters as possible and re-evaluating the alignments (2) by using the HMM that is far more elaborate than "eye-balls" (3) by restricting the search area based on the statistical analysis and (4) by performing a systematic analysis of genome-scaled data.

We are now studying the entire *E. coli* genome with the same approach. Further incorporation of our knowledge on other activators/repressors and the collaboration with the groups of *B. subtilis* and *E. coli* post-sequencing projects will be promising for our understanding of the gene-regulation network.

Acknowledgments

The authors thank Naotake Ogasawara for relevant discussions and for letting us use the unpublished data of his group. This project is partly supported by the *ALIS* project for genome analysis of JST and by a Grant-in-Aid (*Genome Science*) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports, and Culture in Japan.

References

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Fondrat, C., and Kalogeropoulos, A. 1994. Approaching the function of new genes by detection of their potential activation sequences in *Saccharomyces cerevisiae:* application to chromosome III. *Current Genet.* 25:396–406.

Haldenwang, W. G. 1995. The sigma factors of *Bacillus subtilis. Microbiol. Rev.* 59:1–30.

Harwood, C. R., and Wipat, A. 1996. Sequencing and functional analysis of the genome of *Bacillus subtilis* strain 168. *FEBS Letters* 389:84–87.

Hirosawa, M.; Totoki, Y.; Hoshida, M.; and Ishikawa, M. 1995. Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput. Appl. Biosci.* 11:13–18.

Horton, P. B., and Kanehisa, M. 1992. An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucl. Acids Res.* 20:4331–4338.

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.

Lewin, B. 1994. *Genes.* Oxford University Press, 5th edition.

Yada, T., and Hirosawa, M. 1996. Gene recognition in Cyanobacterium genomic sequence data using the hidden Markov model. In *Proc. of the 4th Int. Conf. on Intelligent Systems for Molecular Biology,* 252–260. Menlo Park, Calif.: AAAI Press.

Yada, T.; Sazuka, T.; and Hirosawa, M. in printing. Analysis of sequence patterns surrounding the translation initiation sites on Cyanobacterium genome by using hidden Markov model. *DNA Res.*