Predicting the Operon Structure of Bacillus subtilis Using Operon Length, Intergene Distance, and Gene Expression Information

M.J.L. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano

Pacific Symposium on Biocomputing 9:276-287(2004)

PREDICTING THE OPERON STRUCTURE OF BACILLUS SUBTILIS USING OPERON LENGTH, INTERGENE DISTANCE, AND GENE EXPRESSION INFORMATION

M.J.L. DE HOON[1], S. IMOTO[1], K. KOBAYASHI[2],
N. OGASAWARA[2], S. MIYANO[1]
[1] *Human Genome Center, Institute of Medical Science, University of Tokyo*
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2] *Graduate School of Biological Science, Nara Institute of Science and Technology,*
8916-5 Takayama, Ikoma, Nara 630-0101, Japan

We predict the operon structure of the *Bacillus subtilis* genome using the average operon length, the distance between genes in base pairs, and the similarity in gene expression measured in time course and gene disruptant experiments. By expressing the operon prediction for each method as a Bayesian probability, we are able to combine the four prediction methods into a Bayesian classifier in a statistically rigorous manner. The discriminant value for the Bayesian classifier can be chosen by considering the associated cost of misclassifying an operon or a non-operon gene pair. For equal costs, an overall accuracy of 88.7% was found in a leave-one-out analysis for the joint Bayesian classifier, whereas the individual information sources yielded accuracies of 58.1%, 83.1%, 77.3%, and 71.8% respectively. The predicted operon structure based on the joint Bayesian classifier is available from the DBTBS database (`http://dbtbs.hgc.jp`).

1 Introduction

In prokaryotes, open reading frames (ORFs) belonging to the same operon are transcribed together into a single mRNA molecule. To understand gene regulation in prokaryotic organisms, as a first step it is important to determine the operon structure of their genomes. In addition, as genes in the same operon are likely to be functionally related, the inferred operon structure may reveal the role of currently unknown genes.

The distance between two adjacent genes on the same strand of DNA tends to be shorter if they belong to the same operon, and longer if they belong to different operons. Using a list of experimentally known operons, we can determine the discriminant value of the intergenic distance at which an adjacent gene pair is more likely to be an operon pair than a non-operon pair. For the *Escherichia coli* genome, operon pair predictions using the intergenic distance information were 82% accurate.[1,2]

An alternative method of operon prediction is based on gene expression measurements. Using cDNA microarray technology, the expression levels can be measured simultaneously for all genes in the genome by measuring the corre-

sponding mRNA concentrations. In time course gene expression experiments, the expression levels are measured at several time points following a change in the environment of the organism, such as an increase in the temperature or the salt concentration. In gene disruptant experiments, the steady-state gene expression levels are measured for an organism in which the expression of a specific gene has been disrupted. As genes belonging to the same operon are transcribed into a single mRNA molecule, the degree of similarity in the gene expression profiles of two adjacent genes can be used to assess the likelihood that the gene pair belongs to the same operon. When applied to operon prediction in *Escherichia coli* using a collection of 72 cDNA microarray experiments to calculate the similarity in gene expression, a sensitivity of 82% was found.[3]

Sabatti *et al.*[3] postulated that gene experiments that perturb a large number of genes offer more information for operon prediction than confined perturbations. Time-course gene expression data may therefore be more suitable for operon prediction than gene disruptant expression data, as changes in the environment of an organism in a time-course experiment are likely to affect a larger number of genes than the disruption of a single gene in a gene disruptant experiment.

In practice, the distribution functions of both the intergenic distance and the measured similarity in gene expression exhibit a large degree of overlap for operon gene pairs and non-operon gene pairs, and the choice between operon and non-operon may become ambiguous. The reliability of operon prediction can be improved by considering the intergenic distance and the similarity in gene expression together in a Bayesian posterior probability, which resulted in a sensitivity of 88% for the *Escherichia coli* genome.[3] For these predictions, a constant (uninformative) prior was used.

To find the true Bayesian posterior probability, we would have to consider the relative abundance of operon pairs in comparison to non-operon pairs. This will give us a base line rate of finding operon gene pairs among the adjacent gene pairs, depending on the average number of genes per operon. Within a Bayesian framework, we can consider this rate as the prior probability of a gene pair to belong to the same operon, while the intergenic distance and gene expression information are used to calculate the Bayesian posterior probability.

As on average an operon in the *Bacillus subtilis* contains more than two genes, there are more operon gene pairs than non-operon gene pairs. Including the prior probability will therefore lead to a more accurate prediction for operon pairs, a less accurate prediction for non-operon pairs, and a higher overall prediction. To guard against a less accurate prediction for non-operon pairs, we can consider the relative cost of misclassification as an operon pair compared to the cost of misclassification as a non-operon pair. For example, if we want to

verify experimentally the operon boundaries by considering all candidate non-operon gene pairs, the cost of misclassifying a non-operon pair as an operon pair would be relatively high, and we might consider to classify a gene pair as a tentative operon pair even if the posterior probability is somewhat lower than 50%.

Here, we use the combination of intergenic distance and similarity in gene expression from 99 gene disruptant experiments and 75 time-course expression measurements to predict the operon structure in *Bacillus subtilis*. From a list of 635 known operons,[4,5,6] we found 582 operon pairs and 91 non-operon pairs. Using these data, we predicted the operon structure of *Bacillus subtilis*, and assessed the overall prediction accuracy and the relative contributions of operon length, intergenic distance, and expression information.

2 Operon structure predictors

2.1 Operon length

Table 1 shows the distribution of the operon length based on our list of 635 known operons. To infer a base line rate for adjacent gene pairs to belong to the same operons, we would like to fit a statistical model to these measured operon lengths. The simplest statistical model consistent with the data is the geometric distribution:[7]

$$\Pr\left[\text{operon contains } n \text{ genes}\right] = p^{n-1}\left(1-p\right) \tag{1}$$

Accordingly, we regard operons as being produced by a Bernoulli process with probability p, as shown in Figure 1. A Bernoulli process is the discrete equivalent of a Poisson process, and is the only discrete distribution without memory. Biologically, it means that a priori there is a probability p for each intergenic region to contain a terminator sequence to mark the end of an operon, independent of its length. Using Eq. 1, we can calculate the probability p from the average operon length $\bar{n}$:

$$p = \frac{\bar{n}-1}{\bar{n}}, \tag{2}$$

where $\bar{n} = 2.39$ is determined from Table 1, leading to a prior probability $p = 0.581$ of finding an operon pair. Figure 2 shows the distribution of the measured operon lengths, as well as the geometric distribution fitted to it. Note that except for singletons, any known operon will contribute to the set of known operon pairs, while non-operon pairs can only be found if two adjacent operons both happen to be known. Estimating p directly from the number of known operon pairs and known non-operon pairs would therefore lead to a severely biased estimate.

Table 1: Number of genes per operon, calculated from the list of 635 known operons.

Length	Frequency	Length	Frequency	Length	Frequency
1	279	6	19	11	0
2	170	7	14	12	1
3	70	8	7	13, 14, 15	0
4	35	9	5	16	1
5	31	10	2	31	1

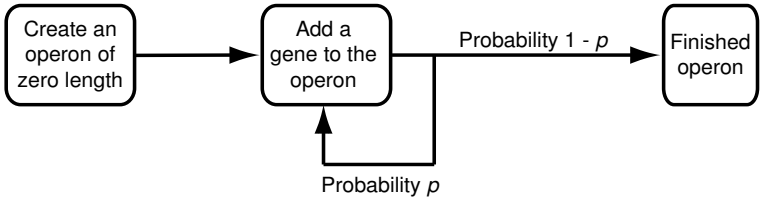

Figure 1: The distribution of the operon length can be described in terms of a Bernoulli process with probability p.

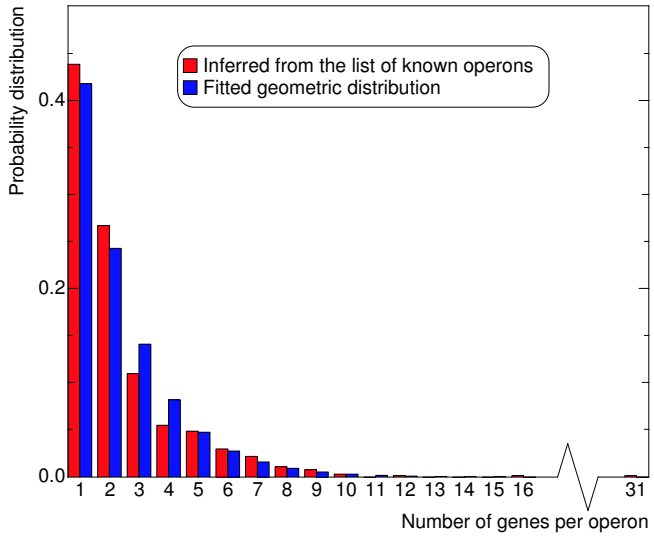

Figure 2: The distribution of the operon length, as determined from the list of 635 known operons.

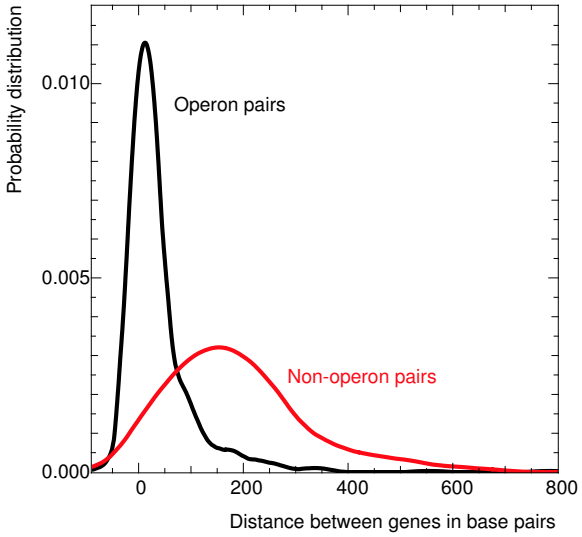

Figure 3: The distribution function of the distance in base pairs between adjacent genes for operon pairs and non-operon pairs.

2.2 Intergenic distance

Using the list of known operon and non-operon pairs, we estimated the probability density distribution of the distance between the genes, measured in base pairs, using an estimation procedure based on the Epanechnikov kernel.[8] As some genes partially overlap each other, the intergenic distance is allowed to be negative. Figure 3 shows the inferred probability distribution for operon pairs and non-operon pairs. Whereas the intergenic distance on average is considerably less for operon pairs than for non-operon pairs, there is a substantial overlap between the two distribution functions, highlighting the need for additional predictors to distinguish operon pairs from non-operon pairs.

2.3 Gene expression data

As genes that belong to the same operon are transcribed into a single mRNA molecule, we expect their measured expression profiles to be highly similar. In cluster analysis,[9] the Pearson correlation and the Euclidean distance[10] are commonly used to assess the similarity in gene expression profiles. In operon prediction from gene expression data, the Pearson correlation is typically used. However, the theory of discriminant analysis[11] suggests that the Euclidean

Table 2: The time points at which expression measurements were made for the eight time-course experiments of *Bacillus subtilis*.

Experiment	Measurement time points in minutes
Cold shock	0, 5, 10, 30, 60, 120
Competence	0, 60, 120, 180, 240, 300, 360
Glucose, glutamine added during sporulation	0, 60, 120, 180, 240, 300
Glucose limitation	0, 60, 125, 180, 240
Heat shock	0, 5, 10, 30, 60
Increased aminoacid availability	0, 30, 60, 120, 210, 300, 420, 540
Phosphate, glucose starvation	0, 60, 120, 180, 240, 300, 360, 420
Phosphate limitation	0, 55, 115, 175, 235, 295
Salt stress	0, 5, 10, 30, 60
Sporulation	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 510, 540

distance would be optimal, given that the expression profiles of gene pairs in the same operon are equal rather than merely correlated. Here, we will apply both the Euclidean distance and the Pearson correlation to evaluate their effectiveness in separating operon pairs from non-operon pairs.

We consider the gene expression data measured at 75 time points total in eight time-course experiments, described in Table 2, together with 99 gene disruptant experiments, listed in Table 3. Genes with more than 50% missing data were removed for the leave-one-out analysis described below. Furthermore, in each disruptant experiment the measured expression levels for the disrupted gene were marked as missing. Global normalization was applied to the remaining genes.

Figures 4 and 5 show the distribution functions of the Pearson correlation and the Euclidean distance for known operon and non-operon gene pairs. To guarantee that the probability density function vanishes for distances less than zero, a mirroring technique was used in which the negative of each data point was added to the data set. The probability density function estimated from the padded data set was subsequently multiplied by two and set to zero for negative distances. For the Pearson correlation r, the same mirroring technique was used for $r = 1$; for $r = -1$, no mirroring was needed as both probability density functions were already zero. Both figures show a considerable amount of overlap between the distribution functions for operon pairs and non-operon pairs, although the Pearson correlation achieves a slightly better separation.

Table 3: Disrupted gene in each experiment. The genes *degU*, *sigF*, *sigW*, and *veg* were each disrupted in two experiments, as indicated here.

abh	*citR*	*yjmH*	*iolR*	*paiB*	*sigF*	*sigY*	*tnrA*	*yufL*
abrB	*citT*	*yqkL*	*ycsO*	*ygaG*	*sigF*	*sigZ*	*treR*	*yugG*
acoR	*codY*	*gerE*	*lacR*	*phoP*	*sigG*	*sinR*	*veg*	*yurK*
ahrC	*comA*	*glcR*	*levR*	*purR*	*sigH*	*soj*	*veg*	*yvkB*
alsR	*comK*	*glcT*	*lexA*	*pyrR*	*ykoZ*	*splA*	*xylR*	*yvrH*
ansR	*cspB*	*glnR*	*lmrA*	*rocR*	*sigL*	*spo0A*	*ybbH*	*ywaE*
araR	*ctsR*	*gntR*	*lrpA*	*sacT*	*yhdM*	*spo0J*	*ybfA*	*yyaA*
azlB	*ydbG*	*gutR*	*lrpC*	*senS*	*sigV*	*spoIIIC*	*yesS*	*yybA*
ccpA	*degU*	*hpr*	*yqhN*	*sigB*	*sigW*	*spoIIID*	*yhjM*	*yybE*
yyaG	*degU*	*hrcA*	*mtrB*	*sigD*	*sigW*	*spoVT*	*yotL*	*yydK*
ykuM	*deoR*	*hutP*	*paiA*	*sigE*	*sigX*	*tenA*	*ytzE*	*yqfV*

2.4 Bayesian classifier

From the estimated distribution functions $f_{\mathrm{OP}}(d)$, $f_{\mathrm{NOP}}(d)$ of the intergenic distance d for known operon pairs (OP) and known non-operon pairs (NOP), and the estimated distribution functions $g_{\mathrm{OP}}(D)$, $g_{\mathrm{NOP}}(D)$ of the dissimilarity D between two expression profiles, we construct the joint Bayesian classifier

$$p_{\mathrm{posterior}}(d, D) = \frac{p \cdot f_{\mathrm{OP}}(d) \cdot g_{\mathrm{OP}}(D)}{p \cdot f_{\mathrm{OP}}(d) \cdot g_{\mathrm{OP}}(D) + (1-p) \cdot f_{\mathrm{NOP}}(d) \cdot g_{\mathrm{NOP}}(D)}. \quad (3)$$

With the prior probability p calculated from the average operon length (Eq. 2), the joint Bayesian classifier is equal to the posterior probability of finding an operon pair. The prediction accuracy will be higher for operon pairs than for non-operon pairs, due to the former being more abundant than the latter in the *Bacillus subtilis* genome, as parameterized by p. With the uninformative prior ($p = \frac{1}{2}$) proposed previously,[3] Eq. 3 is no longer the true Bayesian posterior probability. The uninformative prior leads to an equal accuracy for operon and non-operon pairs, but to a lower overall accuracy.

Usually, a gene pair is predicted to belong to the same operon if the posterior probability is more than $\frac{1}{2}$, and to different operons if the posterior probability is less than $\frac{1}{2}$. Instead, we propose to classify a gene pair as an operon pair if the posterior probability surpasses a certain discriminant value p_{D} which is not necessarily equal to 0.5. This allows us to tune the relative accuracy of finding operon pairs or non-operon pairs by choosing the parameter p_{D} appropriately, depending on how the operon predictions will be used. For example, for terminator sequence prediction we may want to include all gene

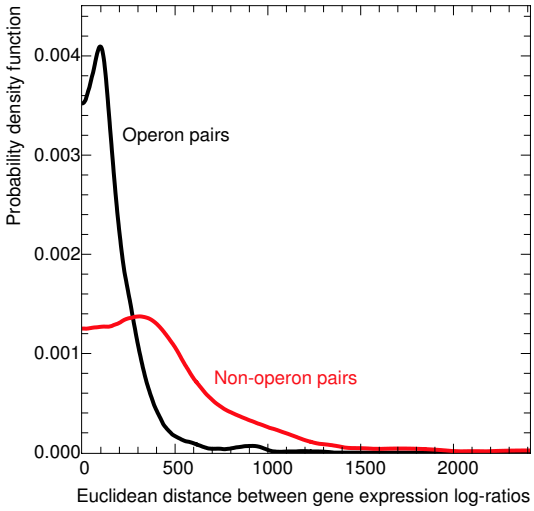

Figure 4: The probability density function of the measured Euclidean distance between the expression log-ratios for known operon and known non-operon gene pairs, as calculated from the combined gene disruptant and time-course gene expression data.

pairs that have a posterior probability of 30% or more of being a non-operon pair ($p_{\mathrm{D}} = 0.7$), as requiring a posterior probability of 50% will cause us to miss many potential terminator sequences.

3 Prediction accuracy

The operon prediction accuracy was assessed using a leave-one-out analysis, in which each of the known operon or non-operon pairs was consecutively ignored in the learning phase, followed by a prediction of the operon status of the gene pair that was left out. Using only the operon length information, the Bayesian classifier reduces to the prior probability for all gene pairs. Consequently, all gene pairs are predicted to be operon pairs, resulting in a 100% prediction accuracy for operon pairs, a 0% accuracy for non-operon pairs, and an 58.1% overall prediction accuracy, corresponding to the prior probability p.

Table 4 shows the accuracy of predictions based on the intergenic distance, the gene expression data, and on the joint Bayesian classifier, using a discriminant $p_{\mathrm{D}} = \frac{1}{2}$ for the posterior probability. The intergenic distance, at an accuracy of 83.1%, is a somewhat more reliable predictor of the operon structure than the gene expression data, which yielded an accuracy of 79.9%. As

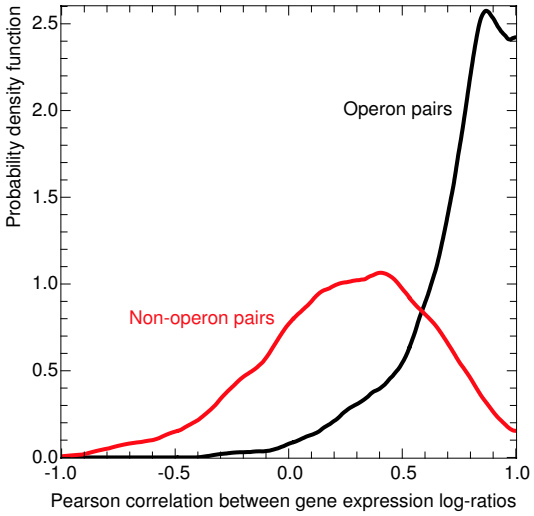

Figure 5: The distribution of the measured Pearson correlation between the expression log-ratios for known operon and known non-operon gene pairs, as calculated from the combined gene disruptant and time-course gene expression data.

expected, the joint Bayesian classifier surpasses each of the separate predictors, reaching an accuracy of 88.7%. Here, the similarity in the gene expression profiles was assessed using the Pearson correlation r by defining $D \equiv 1 - r$. The Euclidean distance yielded a marginally lower prediction accuracy of 88.6% for the joint Bayesian classifier. The time-course gene expression data achieved a better prediction accuracy (77.3% based on 75 expression measurements) than gene disruptant experiments (71.8% based on 99 expression measurements). This is consistent with the conjecture by Sabatti *et al.*[3] that gene expression experiments affecting a large number of genes are more suitable for operon prediction. The combined expression data of the time-course and the gene disruptant experiments achieved an improved prediction accuracy of 79.9%.

As in this analysis the cost of misclassifying an operon pair is regarded to be equal to the cost of misclassifying a non-operon pair, the discriminant value for the posterior probability was chosen to be 50%. The prediction accuracy of non-operon pairs can be improved at the expense of a less accurate prediction for operon pairs by increasing the discriminant value p_{D}, and vice versa. Figure 6 shows the prediction accuracy of the joint Bayesian classifier as a function of the discriminant probability p_{D}. The optimal overall accuracy is achieved for a discriminant probability less than 0.5, which reflects the fact

Table 4: The accuracy of operon prediction in *Bacillus subtilis*, based on a leave-one-out analysis. The discriminant value for the posterior probability was set to 50%.

Predictor	Operon pairs	Non-operon pairs	Overall accuracy
Intergenic distance	82.1%	89.0%	83.1%
Gene expression, overall	80.1%	79.1%	79.9%
Time-course experiments	76.8%	80.2%	77.3%
Gene disruptant experiments	69.9%	83.5%	71.8%
Joint Bayesian classifier	88.8%	87.9%	88.7%

that operon pairs are more abundant than non-operon pairs in the *Bacillus subtilis* genome.

Next, we used the joint Bayesian classifier to predict the operon structure of the complete *Bacillus subtilis* genome, using the Pearson correlation to assess the similarity in the expression profiles. The predicted operon structure is available from the DBTBS database[5] in terms of the posterior probability, enabling users to assess the reliability of each prediction, as well as to choose the discriminant value p_{D} corresponding to their interests.

In addition to the predictors described above, we examined the viability of determining the operon structure by finding the σ^{A} transcription factor binding site and the terminator sequence motif. For all regions between adjacent gene pairs on the same strand of DNA, we calculated the motif score using the Position Specific Score Matrix for the σ^{A} binding site.[5] The terminator sequence motif was predicted using dtp, a prediction tool for finding rho-independent transcription terminators.[12] Neither of these predictors produced a clear distinction between operon pairs and non-operon pairs, and were therefore not included in the joint Bayesian classifier. Note that in both cases the aim of the predictor is to find where a motif is located in a given sequence segment, rather than whether a given sequence segment contains the motif. It may therefore be possible to construct better sequence analysis tools for the specific task of operon structure prediction.

4 Conclusion

We predicted the operon structure of the *Bacillus subtilis* genome by combining operon length, intergenic distance, and gene disruptant and time-course gene expression experiments at an estimated overall accuracy of almost 89%. The intergenic distance information was the most accurate single predictor (83.1%), followed by the time-course gene expression data (77.3%) and the

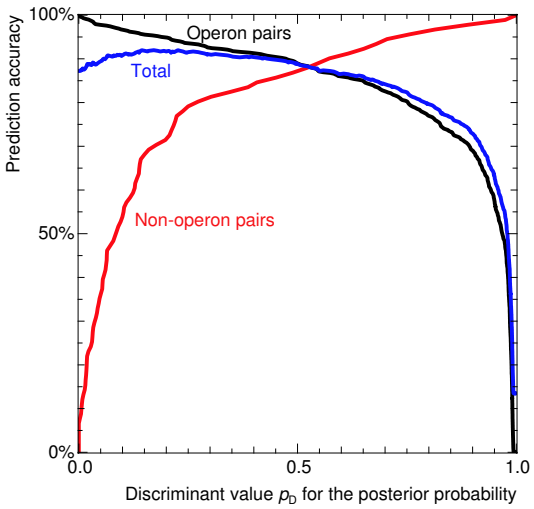

Figure 6: The prediction accuracy as a function of the choice for the discriminant probability p_{D}. A large value of p_{D} corresponds to a high cost of misclassifying a non-operon gene pair.

gene disruptant data (71.8%). The average operon length was considered in order to determine the base line probability of finding an operon pair. The distribution of the operon length was modeled by a geometric distribution, which means that a priori there is an equal probability of finding a terminator sequence between any pair of adjacent genes, irrespective of the length of the operons in which those genes are located. The predicted operon structure is available from the DBTBS database.[5]

In the leave-one-out analysis, we found that assessing the expression similarity using the Euclidean distance does not yield a better separation between operon and non-operon pairs than the Pearson correlation. This is somewhat surprising from the viewpoint of discriminant analysis. The superior results of the Pearson correlation may be due to the error structure in gene expression measurements, or to hitherto unexplained dependencies in the expression level of two adjacent genes in different operons. Similarity measures may exist that are even more suitable for operon prediction than the Pearson correlation.

Acknowledgments

We would like to thank Yuko Makita and Mitsuteru Nakao of the University of Tokyo for assisting us with the σ^{A} and terminator sequence motif prediction.

References

1. G. Moreno-Hagelsieb and J. Collado-Vides. A powerful non-homology method for the prediction of operons in prokaryotes. In *Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB 2002), Bioinformatics Supplement 1*, pages S329–S336, 2002.
2. H. Salgado, G. Moreno-Hagelsieb, T.F. Smith, and J. Collado-Vides. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA*, 97:6652–6657, 2000.
3. C. Sabatti, L. Rohlin, M.-K. Oh, and J.C. Liao. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, 30:2886–2893, 2002.
4. S. Okuda, S. Kawashima, and M. Kanehisa. Database of operons in *Bacillus subtilis*. In *Genome Informatics*, volume 13, pages 496–497, 2002.
5. Y. Makita and K. Nakai. DBTBS: Database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Research, submitted*, 2003. http://dbtbs.hgc.jp.
6. A.L. Sonenshein, J.A. Hoch, and R. Losick. *Bacillus subtilis and its closest relatives: From genes to cells*. ASM Press, Washington, DC, 2001.
7. J.H. Zar. *Biostatistical Analysis*. Prentice-Hall, London, 4 edition, 1999.
8. B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hill, London, 1986.
9. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
10. D.K. Slonim, P. Tamayo, J.P. Mesirov, T. Golub, and E.S. Lander. Class prediction and discovery using gene expression data. In *RECOMB 2000*, pages 263–272, 2000.
11. M. S. Bartlett and N. W. Please. Discrimination in the case of zero mean differences. *Biometrika*, 50:17–21, 1963.
12. T. Yada, M. Nakao, Y. Totoki, and K. Nakai. Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. *Bioinformatics*, 15(12):987–993, 1999.