NAIST-IS-DD0561014

Doctoral Dissertation

The entire organization of operons on the *Bacillus subtilis* genome

Hirokazu Kobayashi

February, 7, 2008

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering.

Thesis Committee:
Professor Shigehiko Kanaya
Professor Naotake Ogasawara
Professor Kotaro Minato
Associate Professor Ken Kurokawa

The entire organization of operons on the *Bacillus subtilis* genome

Hirokazu Kobayashi

Abstract

In the post-genomic era, comprehension of cellular processes and systems requires global and non-targeted approaches to handle vast amounts of biological information. The present study predicts transcription units (TUs) in *Bacillus subtilis*, based on an integration of *B. subtilis* experimental data involving DNA sequence and transcriptome analyses. First, co-expressed gene clusters are predicted by calculating the Pearson correlation coefficients of adjacent genes for all the genes in a series that are transcribed in the same direction with no intervening gene transcribed in the opposite direction. Transcription factor (TF) binding sites are then predicted by detecting statistically significant TF binding sequences on the genome using a position weight matrix. This matrix is a convenient way to identify sites that are more highly conserved than others in the entire genome because any sequence that differs from a consensus sequence has a lower score. I identify genes regulated by each of the TFs by comparing gene expression between wild-type and TF mutants using a one-sided test. By applying the integrated approach to 11 σ factors and 17 TFs of *B. subtilis*, I am able to identify fewer candidates for genes regulated by the TFs than were identified using any single approach, and also detect the known TUs efficiently. Using these TU data, I predicted the genome-wide operon structure in the *B. subtilis* genome by comparative genomic analysis of 55 gram positive bacteria. This integrated approach showed determining to appropriate boundaries of operons efficiently and I identified some internal operons. Furthermore, I took another operon prediction approach by a support vector machine (SVM)-based classification algorithm. SVM is one of the supervised machine learning methods. I used five types of features: intergenic distances, expression profiles, two comparative genomic data and gene functional data, and separated all consecutive gene pairs into operon pairs and boundary pairs on a 10-dimensional feature hyperplane and efficiently identified operon structures. Finally I combined the results of these approaches, and demonstrated new operon structures in the *B. subtilis* genome

Keywords: operon, microarray, position weight matrix, *Bacillus subtilis*, false discovery rate, comparative genomics, support vector machine

*Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0561014, February, 7, 2008

Contents

1 Preface

The first discovered operon was the lactose operon in *Escherichia coli*, which consists of three genes, *lacZ*, *lacY* and *lacA* [Jacob and Monod 1961]. These three adjacent genes are required for the transport of lactose and expressions of these genes are induced in response to abundance/scarcity of lactose in *E. coli* and some other bacteria. The *lac* operon is regulated by several factors, one of which is the availability of lactose as an energy source. Control of the *lac* genes was the first elucidation of the genetic regulatory mechanism controlled by the *lac* repressor and positive regulatory elements. The *lac* repressor is made up of four identical polypeptides, homotetramer. Part of the molecule has sites that enable it to recognize and bind to the 24 base pairs of the *lac* operator. Another part of the repressor contains sites that bind to lactose. When lactose unites with the repressor, it causes a change in the shape of the molecule, so that it can no longer remain attached to the DNA sequence of the operator and prevents transcription. On binding of lactose to the repressor, the *lac* repressor leaves the operator, and three genes are expressed together. The genes of the operon are *lacY* coding for the lactose permease that transports lactose into the cell, *lacZ* coding an intracellular enzyme, β-galactosidase that cleaves the disaccharide lactose into glucose and galactose, and *lacY* coding a transport protein, β-galactoside permease that pumps lactose into the cell. Specific control of the *lac* genes depends on the presence of the substrate lactose in the growth medium. These genes are organized into an operon such that they are oriented in the same direction immediately adjacent on the chromosome and are co-transcribed into a single mRNA molecule, and the mRNA is then translated into protein, only when lactose is present. This elucidation of the control of expression of genes in the *lac* operon brought the Nobel Prize in Physiology or Medicine for Jacques Monod and Francois Jacob in 1965.

Mechanisms of gene transcription in the upstream of protein coding regions by gene regulatory proteins are summarized in Figure 1. An operon is one or more structural genes, and all the genes in an operon are expressed as a single unit and are transcribed into one mRNA. In the present study, a set of genes transcribed into an identical mRNA is referred to as a **transcription unit** (**TU**) or an **operon**. At the upstream of the operon, there are an operator and a promoter that are gene expression regulation sites. An operator is a segment of DNA that regulates the activity of the transcription by interacting with a specific gene regulatory protein and exits in both the upstream and downstream of the promoter. For example, the lambda repressor and *lac* repressor bind upstream and downstream of promoters respectively and inactivate particular transcriptions

[Pabo and Sauer 1984; Matthews and Nichols 1998]. In negative control of an operon, a repressor protein binds to the operator and prevents transcription of the genes composing the operon. Most repressor proteins are produced by regulator genes but they are unable to bind to the operators in their normal conformation. When a ligand binds to (Fig. 1A1) or is released from the repressor protein (Fig. 1A2), the conformation of the repressor protein changes and the repressor protein can bind to the operator. The activated repressor protein binds to the operator and prevents transcription (Fig. 1A). When the repressor protein is unbound from the operator by the presence or removal of inducer molecules, its conformation changes and the transcription of the genes begin in reverse. Operons can also be positively controlled (Fig. 1B). When ligands bind to (Fig. 1B1) or are released from the activator protein (Fig. 1B2), the conformation of activator protein changes and the protein binds to the operator, and then the activator stimulates transcription by binding the RNA polymerase holoenzyme, and operates it. In this study, I regarded these proteins, repressor and activator, to be **transcription factors** (**TFs**). A promoter site is recognized by RNA polymerase which then binds and initiates transcription. In the process, the RNA polymerase core enzyme reads the signal of transcription beginning in the DNA with the assistance of detachable subunits called σ-factors. Therefore, transcription of the genes is promoted with the presence of σ-factors.

After determining complete the *Bacillus subtilis* genome sequence [Kunst et al. 1997], characterization of operons provides the basic knowledge to understand the gene organization of genomes and regulatory networks. The organization of operons is considered to facilitate the efficient coordinated regulation and association of functionally related protein products. A number of bioinformatic approaches have been carried out for operon prediction. The purpose of this thesis is to clarify organization of TUs and gene-regulated relations in the whole genome of *B. subtilis* based on an integrated approach consisting of genome sequences and transcriptome analyses. In chapter 2, I review the current progress of operon prediction techniques in bioinformatics. In the *B. subtilis* genome, about 300 operons consisting of approximately 1,000 genes have been experimentally determined and 18 σ-factors and 384 TFs are predicted by computational analyses [DBTBS]. Among the genes that encode the TFs, there exist 270 genes concerning transcription regulation in the *B. subtilis* genome, consisting of 18 σ-factors, 201 TFs with the Helix-Turn-Helix motif, 35 TFs with a two-component system, 16 RNA-binding-regulators, and 16 other factors including anti-terminators. The final goal of the

gene regulation networks in *B. subtilis* is to clarify relations between transcription regulation between those and targeted genes, and to understand whole gene expression regulation systems. To attain this purpose, I examine a gene regulation network using σ-factors and TFs whose DNA-binding sites in transcription regulation are determined. In chapter 3, I discuss an approach for prediction of TUs based on the integrated approach for TU prediction and general properties for TUs in the *B. subtilis* genome, and examine conservation of TUs in the *B. subtilis* genome in comparison to the order of genes in 55 Gram-positive bacteria and predict the genome-wide operon structure in the *B. subtilis* genome by comparative genomic analysis of 55 Gram-positive bacteria. In chapter 4, I apply support vector machine (SVM) to predict operons using a large amount of microarray data published by NCBI and operon compositions in the *B. subtilis* genome. Chapter 4 also contains the discussion on advantages and disadvantages between the integrated and the SVM approaches. Finally, in chapter 5, I summarize the perspective of bioinformatics focused on operon prediction based on the whole regulation networks of *B. subtilis* genes.

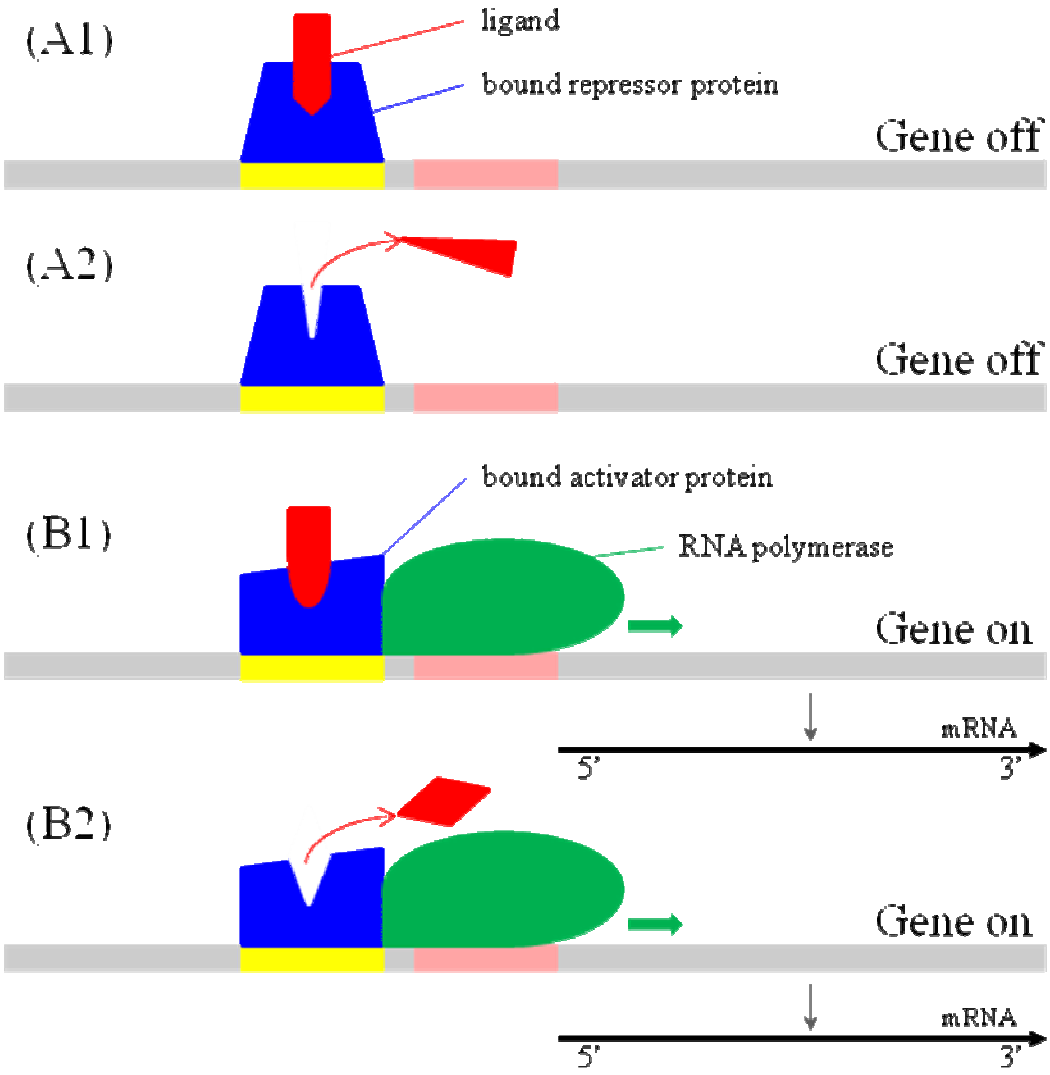


Figure 1. Mechanisms of gene transcription in upstream of protein coding regions by gene regulatory proteins. (A) Negative regulation of gene transcription by repressor protein binding to DNA (A1) with addition of ligand, and (A2) with removal of ligand. (B) Positive regulation of gene transcription by activator protein binding to DNA (B1) with addition of ligand, and (B2) with removal of ligand.

2 Bioinformatic approaches for operon prediction

A number of bioinformatic approaches have been carried out for operon prediction from genome sequences, high-throughput expression profile experiments such as cDNA microarray and GeneChip. Bioinformatic approaches for operon prediction can be classified into five categories, (i) intergenic distance between genes and (ii) detecting transcription control signals, that is, existence of promoters and terminators, (iii) the conservation of operon structures based on comparative genomics, (iv) co-expression relations and (v) integrated approaches.

2.1 Intergenic distances

Statistical analysis of experimentally determined operons has established that genes within an operon tend to have much shorter intergenic distances than those at the borders of the operon. A log-likelihood function of intergenic distance for predicting operons was developed in comparison of intergenic distances within-operon and between-operon gene pairs for the *Escherichia coli* genome, and identified well the known *E. coli* operon [Salgado et al. 2000]. This method has used for the most generally applicable and successful pairwise operon prediction method in many prokaryote genomes.

2.2 Detection of transcription control signals

Existence of promoters and terminators is a clue for operon prediction. Operon prediction has been carried out by distinguishing co-directional TU boundaries from intergenic TU boundaries in operons by densities of signa-70 promoter like signals because of the trend that in *E. coli,* regions upstream of TU contain higher densities of sigma-70 promoter-like signals than both coding regions and the intergenic region downstream of the convergent transcribed genes [Huerta and Collado-Vides 2003; Janga et al. 2006]. The recognition rates of the two types of experimentally determined boundaries using the densities of tri-nucleotide signatures are 78% for *E. coli*, and 72% for *B. subtilis*. Construction of a Hidden Markov Model (HMM) based on known promoters and terminators in *E. coli* makes it possible to predict 60% of known operons [Yada et al. 1999]. Nucleotide sequence patterns at the initiation and termination sites of transcription conserved across multiple genomes makes it possible to predict operons but these sequence motifs are short and can be highly variable, limiting the prediction capability of this method [Thieffry et al.1998].

2.3 Conservation of operon structure based on comparative genomics

There has been a variety of operon prediction methods, including those that use comparative genomic strategies. These methods utilized nucleotide or peptide sequences [Wang et al. 2004; Edwards et al. 2005]. If I use such a method alone, the data it depends on are only available for a selected subset of bacterial species, so that it is difficult to detect the independent results of query genome data. Functions of genes are conserved across related genomes [Overbeek et al. 1999]. Many sets of genes occur in conserved orders on multiple genomes across long stretches of evolutionary time, representing candidate operons. A comparative genome analysis on 34 prokaryotic species led to the conclusion that more than 7600 pairs of genes are highly likely to belong to the same operon [Ermolaeva et al. 2001]. Through taking a certain distance between adjacent genes into consideration, this method allows highly confident prediction of operons in a wide variety of species but using this method, a large portion of the known operon in *E. coli* cannot be predicted. The fairly low sensitivity of this method is due to the low conservation at the operon level between phylogenetically distant genomes [Wolf et al. 2001].

2.4 Function similarity of genes in operons

Genes in an operon tend to encode enzymes that catalyze successive reactions in metabolic pathways. This concept was applied, using 42 prokaryotic species, to predict operons [Zheng et al. 2002]. This method is useful for the purpose of estimating putative operon structures, but cannot predict operons that include unknown function genes.

2.5 Co-expression relations

When microarray gene expression data has become available, the accuracy of operon prediction has been greatly elevated. Sabatti et al. [2002] compiled data from 72 cDNA microarray experiments for *E. coli* and predicted operons based on a Bayesian classification using the correlation between expression ratios of adjacent genes. This method allows a significant refinement of the sequenced-based predictions described above. Yamanishi et al. [2003] applied a generalized kernel canonical correlation analysis to group genes, which share similarities with respect to position within genome and gene expression. However, this method was restricted to a subset of *E. coli* genes that comprised known metabolic pathways.

2.6 Integrated approaches for operon prediction

The prediction rate of operons using both DNA sequence and microarray expression data has been better than those performed using only one type of data. For example, Bockhorst et al. [2003] presented a probabilistic machine-learning approach to predicting operons based on Bayesian networks using diverse evidence sources including gene coordinate, operon length, promoter and terminator signal, codon usage frequency and cDNA microarray expression data. Expression profiles of genes are dynamic; on the other hand, genome information including gene organization, and intergenic DNA sequences is static. Multi-factorial data such as transcriptome and genome information should be combined for predicting operons and thus development of element technology in bioinformatics toward combining the multi-factorial data has been needed for operon prediction. Previous studies have demonstrated that multiple factors of information including intergenic distances [Salgado et al. 2000], transcription control signals [Huerta and Collado-Vides 2003], conservation of operon structure [Ermolaeva et al. 2001], function similarity of genes [Zheng et al. 2002] and co-expression gene relations [Sabatti et al. 2002] are effectively utilized for predicting operon structures in genomes. Integration of these multiple factors could lead to predicting operon structures effectively, and the way to do so is an important subject in this study. Consequently, in the present study, we try to develop a method for identifying operons in *B. subtilis* by integration of these factors.

3 The integrated approach to identification of operons

3.1 Transcription unit prediction by integration of *B. subtilis* data

3.1.1 Introduction

Recent progress in genome projects has generated a vast amount of nucleotide sequence data, and analyses of gene expression by global approaches have started to broaden understanding of cell systems. As a useful model for systems biology and genomics, many studies use *B. subtilis*, a spore-forming Gram-positive bacterium whose genome sequence has been determined [Kunst et al. 1997]. The ultimate goal of post-genome analysis is to specify transcriptional regulation in the entire genome. Computational algorithms to locate TUs have been developed based on analysis of signal sequences that are located at the boundaries of TUs from promoters to terminators, homologous gene pairs on other genomes, intergenic distance, functional categories, and gene clusters conserved among various species [Ermolaeva et al. 2001; Huerta et al. 1998; Salgado et al. 2000; Wang et al. 2004; Westover et al. 2005; Yada et al. 1999]. In the present study, a string of one or more genes co-transcribed is defined as a TU [Salgado et al. 2000].

Identification of σ-factors and TFs and their binding sites on their target genes is an important element of transcriptome analysis in the post-genome-sequencing era. Various approaches have been taken to identify specific DNA-binding sites of σ-factors and TFs. DNA-binding specificities have traditionally been determined by experimental techniques such as DNase I footprinting and electromobility shift assay [Ogasawara et al. 1985; Schujman et al. 2003]. More recently, the binding sites have been predicted by computational techniques such as hidden Markov models (HMMs) [Moreno-Campuzano et al. 2006] and position-weight matrices (PWMs) [Stormo 2000; Qiu et al. 2003]. The PWM has one column for each position in the binding site and one row for each nucleotide. Each of the matrix elements is proportional to the relative frequency of the corresponding nucleotide at each position, and the score for a particular site is the sum of the matrix values for the sequence. Therefore, PWM is often used to predict nucleotide-protein binding sites and is used in the TRANSFAC database, which covers many known TFs and binding sites [Heinemeyer et al. 1998]. This approach is a convenient way to identify positions that are more highly conserved than others in a whole genome, because any sequence that differs from a consensus sequence has a lower score. The accuracy of detecting promoter sequences thus

depends on the conservation of these binding sites.

I can now use complete genomic DNA sequences from several species and analyze massive amount of data on differential gene expression in microarray experiments [Kobayashi et al. 2001]. Using microarrays in various conditions, I can obtain co-expression patterns for adjacent genes, which is an important property for determining TUs.

In the present chapter, I identify the TUs in *B. subtilis* using a combination of (i) a bioinformatics approach, using PWM methods that identify σ-factor and TF binding sites by detecting statistically significant binding sequences on the genome; and (ii) two DNA microarray analyses, one to predict co-expressed gene clusters by calculating Pearson correlation coefficients of expression profiles for neighboring genes, and the other to determine genes regulated by each of the TFs in the units by comparing gene expression between wild-type and TF deletion mutants in the genome.

3.1.2 Data

3.1.2.1 Microarray data

For prediction of co-expressed gene clusters, I used time-series gene expression data in *B. subtilis* 168. It was grown in 13 different time-series growth conditions: anaerobic growth; competent medium; cold-shock experiments; DSM medium; DGG medium; glucose-limited medium; heat-shock experiment; LB medium; minimum-glucose medium; sodium-shock conditions; phosphate-starvation medium; and SOS stress experiments.

Expression analysis of σ-factor and TF deletion mutants were applied to gene expression data from inactivate mutants of σ-factor and TF which were grown at 37^{o}C in different medium conditions: LB medium for *sigB, L, M, W, X, araR, ctsR, hrcA, iolR, lmrA, rocR, sinR, xylR* deletion mutants; LB medium with trace elements for *fur* and *perR* deletion mutants; DSM medium for *sigD, E, F, G, H, K, treR* deletion mutants; DSM medium with 2% Gln and 5% glucose for, respectively, *glnR* and *resD* deletion mutants; MC medium for the *comK* deletion mutant; MGM medium for the *ccpC* deletion mutant; and MGM with adenine and guanine for the *purR* deletion mutant.

3.1.2.2 Genomic data

The genome sequence of *B. subtilis* and the annotation files were obtained from National center for biotechnology (NCBI) ftp site [NCBI FTP]. The leading and lagging strands were scanned based on whether they were transcribed in the same directions or in different directions.

3.1.3 Methods

3.1.3.1 Methodology

Establishment of method to predict TUs in genome could aid in better understanding the role of unknown function genes. Then I made TU predictions by an integrated procedure (Fig. 2).

First, I selected directions [Edwards et al. 2005], series of genes transcribed in the same direction with no intervening genes in the opposite direction in the *B. subtilis* genome.

Second, co-expressed gene clusters were determined by correlating expression profiles between neighboring genes transcribed in the same directon (Fig. 2(1)). Co-expression between neighboring genes was estimated using a t-test of the Pearson correlation coefficient. To predict co-expressed gene clusters in the *B. subtilis* genome, I used 98 cDNA microarray data sets in 13 different time-series growth conditions.

Third, I regarded genes having promoters predicted by PWM as the start of the TUs (Fig. 2(2)). I examined the binding sequences of 11 σ factors and 17 TFs, which are known to possess multiple binding sites within 300 bp upstream of an open reading frame for all 4,225 genes of *B. subtilis* by PWM.

Fourth, I derived significant expression change data from σ-factor and TF deletion mutant microarray data to identify genes regulated by each of the σ-factors and TFs in the directon (Fig. 2(3)). In these analyses, I applied a one-sided test to examine genes whose expression changed significantly in the microarray data, and found candidates for the σ-factor and TF regulated genes.

Finally, the three approaches were integrated. I regarded σ-factor and TF regulative gene strings in co-expressed gene clusters beginning with detected promoters as TUs (Fig. 2(4)).

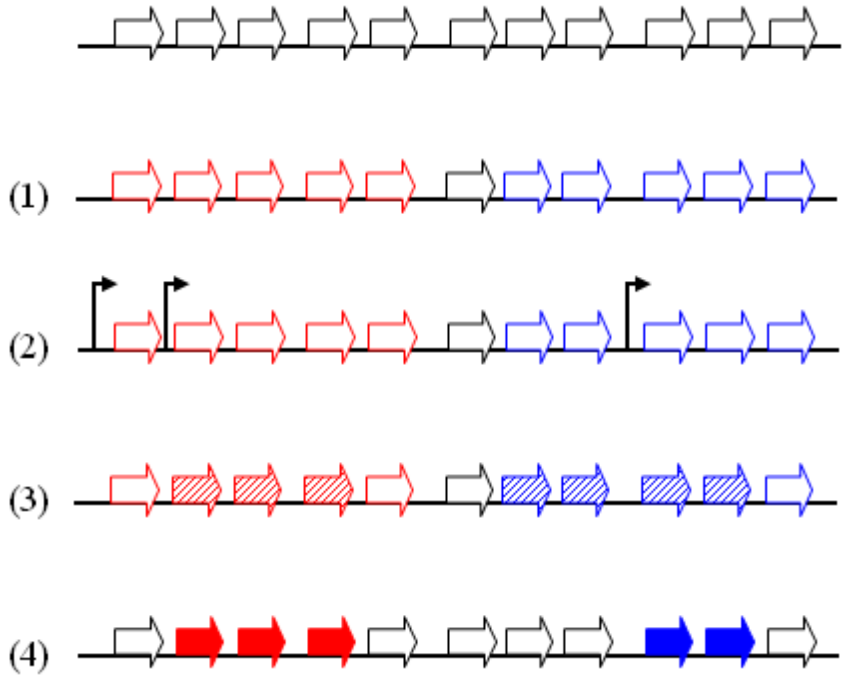


Figure 2. Outline of procedure for elucidating TUs by integrated analyses in gene organization on the genome. (1) Co-expressed gene clusters predicted by correlating expression profiles between neighboring genes; (2) σ-factor and TF binding site (thin arrows) prediction by PWM as the start of a TU in the putative co-expressed gene clusters; (3) detection of significantly expressed genes (stripe thick arrows) by comparison between σ-factor and TF deletion mutant arrays and wild ones; (4) identification of genes composing putative TUs (filled thick arrows) by integration of the three analyses.

3.1.3.2 Normalization in microarray experiments

Gene expression levels are evaluated by scanning the fluorescence intensity for each spot, and there is usually some experimental variation that occurs in every microarray experiment. It is, therefore, important to minimize experimental variation, and although several methods of microarray normalization have been developed [Quackenbush 2002; Yang et al. 2002], there are usually some false-positive data arising when analyzing gene expression data collected via microarrays.

Normalization of the logarithmic ratio of expression intensity between target (R_i) and control (G_i) experiments was carried out based on MA plots [Dudoit et al. 2002], which can show the intensity-dependent ratio of raw microarray data using TREBAX software [TREBAX]. The plots differed in the axes used. The MA plot used M_i ($\log_{10}(R_i/G_i)$) as the y-axis and A_i ($\log_{10}\sqrt{R_iG_i}$) as the x-axis. By plotting values of A_i on the abscissa and M_i on the ordinate of a coordinate system, it was possible to evaluate the bias error with respect to the average logarithmic intensities, and partial averages (PA_i) of the intensities were calculated on the assumption that the intensities of two experiments for most of genes was equivalent. PAv(s) was an average of logarithmic of expression intensities in *s*th width of 0.1 on the abscissa to obtain the following:

$$PAv(s) = \frac{\sum_{k=1}^{Ns} M_k}{\min(s)<Ai<\max(s)\ N_s}.$$

Here, min(s) and max(s) are minimum and maximum values of average logarithmic ratio of expression intensity between target and control experiments in the *s*th compartment, and Ns represents total number of the plots in the *s*th compartment. I then used the average, point by point, to correct the logarithmic ratio of expression intensity (M_i). The normalized log ratio M''_i was estimated as the difference between M_i and baseline M'_i as shown in Figure 3. Here, using the relation between M_i and A_i ($M_i = f(A_i) + \varepsilon_i$, where ε_i is the difference between M_i and $f(A_i)$ for *i*th gene for the MA plot), the baseline for the *i*th gene was estimated by $M'_I = f(A_i)$. Genes whose signal intensity for targeted or control experiments was regarded as zero were eliminated from the present analysis. With this methodology, it is assumed that there was no large error due to expression intensity in the majority of the spots.

The MA plots before and after the adjustments were shown in Figure 4. These plots before

the normalization tend to be dispersed especially in low intensity signal area (Fig. 4A). This is because stability of two types of fluorescent chemical marking cDNA: Cy3 and Cy5, differ according to the coloring effects respectively, and the effectiveness of the labelling process varies since the dyes embedment levels in cDNA differ and thus the image processing leads to a fluorescence intensity map containing two types of errors, i.e. random noise and bias. These plots after the normalization demonstrates that the bias error of the signal-specific to the florescence-material has been reduced (Fig. 4B).

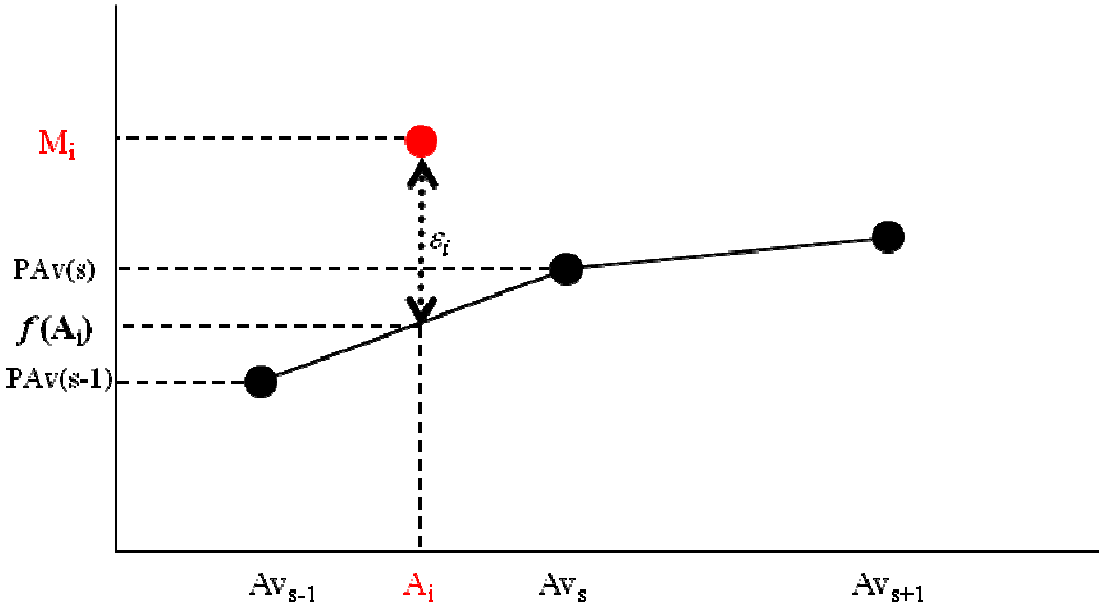


Figure 3. Bias reduction by partial averages on MA-plot.

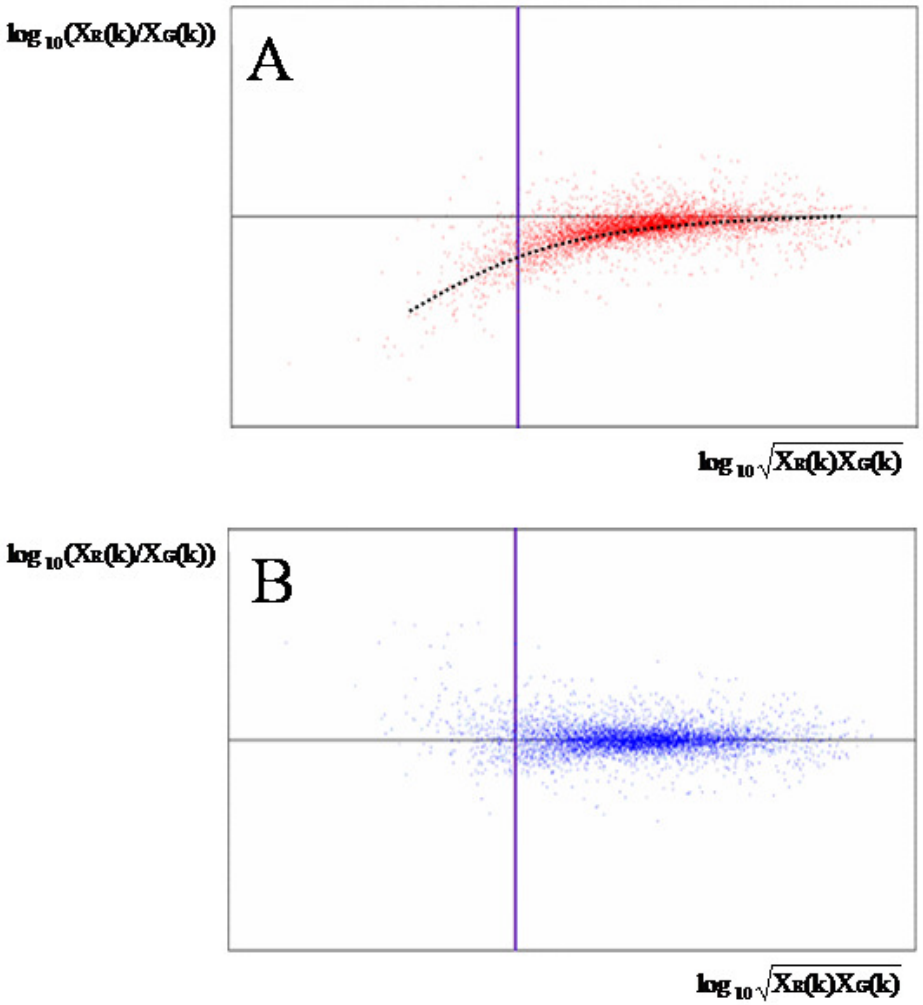


Figure 4. MA plots (A) before and (B) after the normalization of the logarithmic ratio of expression intensity by TREBAX.

3.1.3.3 Prediction of co-expressed gene clusters

If multiple genes are transcribed into a single mRNA, correlations of expression profiles among these genes can be observed. So, I try to develop a procedure for predicting TUs using correlations of expression profiles between genes in the same directon. Figure 5 shows that co-expressed gene clusters were predicted based on expression profiles and genomic locations (Fig. 5). The expression profile of the *i*th position gene is represented by vector $\mathbf{x}_i$, consisting of logarithmic ratios for microarray experiments:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iN})$$

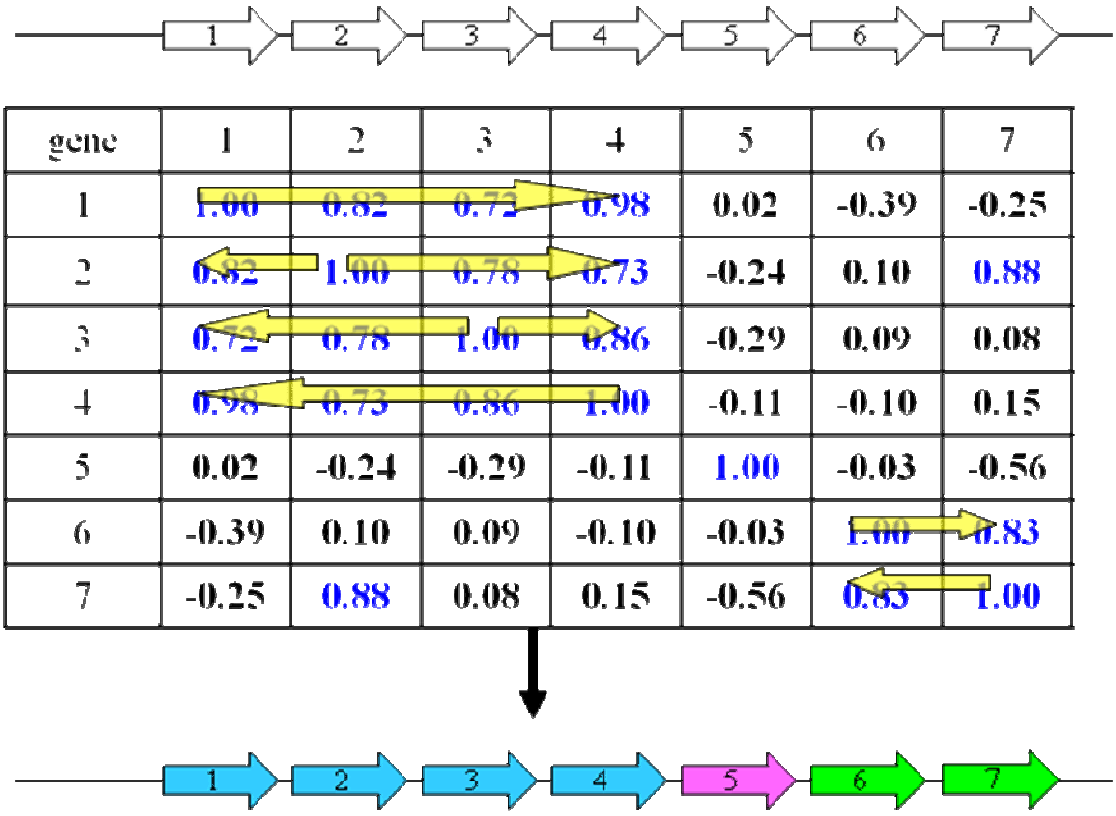
where N is the number of microarray experiments.

The procedure for predicting co-expressed gene clusters is as follows: I selected a directon. The genes were denoted by $g_1, g_2, \ldots g_i, \ldots, g_M$ from their 5' to 3' termini. Here, g_i and g_{i+1} (i = 1, 2, …, M-1) are adjacent genes on the same DNA strand. First, Pearson correlation coefficients (r_{st}) were estimated for all pairs of vectors $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ (s = 1, 2, …, M; t = 1, 2, …, M) as follows:

$$r_{st} = \frac{\sum_{j=1}^{N} (x_{sj} - \overline{x_s})(x_{tj} - \overline{x_t})}{\sqrt{\sum_{j=1}^{N} (x_{sj} - \overline{x_s})^2 \sum_{i=1}^{N} (x_{tj} - \overline{x_t})^2}} .$$

Here, $\overline{x_s}$ and $\overline{x_t}$ are average values of each vector.

Second, a pair of genes was assigned to a candidate group G_s as follows: All the genes $g_{s+1}, g_{s+2}, \ldots, g_{s+Ts}$, whose correlations $r_{s(s+1)}, r_{s(s+2)}, \ldots, r_{sTs}$ were statistically significant in a t-test at the 5% significance level, were classified into G_s. In the same manner, all the genes $g_{s-1}, g_{s-2}, \ldots, g_{s-Us}$, whose correlations $r_{s(s-1)}, r_{s(s-2)}, \ldots, r_{sUs}$ were statistically significant in a t-test at the 5% significance level, were also classified into G_s. Thus, altogether T_s+U_s+1 genes were classified into group G_s. By comparing all members of group G_s (s = 1, 2, …, M), I counted the number of groups consisting of identical members among G_s (s = 1, 2, …, M) and selected the group having the highest count as the first co-expressed gene cluster T_1. After excluding the T_1 genes from all the groups (g_1 to g_M), I selected the next-highest identical group as the next co-expressed gene cluster T_2. This procedure was carried on until the number of members in the cluster T_v (v = 1, 2, …, M) was zero, or until all positions j (j = 1, 2,…, M) were occupied by genes belonging to the cluster T_v.



gene	1	2	3	4	5	6	7
1	1.00	0.82	0.72	0.98	0.02	-0.39	-0.25
2	0.82	1.00	0.78	0.73	-0.24	0.10	0.88
3	0.72	0.78	1.00	0.86	-0.29	0.09	0.08
4	0.98	0.73	0.86	1.00	-0.11	-0.10	0.15
5	0.02	-0.24	-0.29	-0.11	1.00	-0.03	-0.56
6	-0.39	0.10	0.09	-0.10	-0.03	1.00	0.83
7	-0.25	0.88	0.08	0.15	-0.56	0.83	1.00

Figure 5. Prediction method of co-expressed gene clusters. In the gene matrix of Pearson correlation coefficients, blue figures indicate statistically significant values in a t-test at the 5 % significance level, and a yellow arrow indicates a series of genes with statistical significance between a gene at the starting edge of the arrow and genes in the arrow shaft (upper figure). A series of the same color genes indicates co-expressed gene cluster (lower figure).

3.1.3.4 Expression analysis of σ-factor and TF deletion mutants of *B. subtilis*

3.1.3.4.1 Detection of significant expressed genes

The normalized fluorescence intensity data were analyzed using a one-sided test to compare the results of the σ-factor and TF deletion mutant to the control samples, and genes whose expression exceeded the threshold were regarded as σ-factor and TF regulated genes. In lower one-sided tests, I considered genes of decreased expression as being up-regulated by the σ-factor and TF, whereas genes of increased expression were considered as down-regulated by the σ-factor and TF in upper one-sided tests.

3.1.3.4.2 False discovery rate procedure

For detecting proportion of false positive data in the significant data of σ-factor and TF mutant arrays identified in one-sided tests, I used the false-discovery rate (FDR), an alternative approach to multiple testing [Benjamini and Hochberg 1995]. On the assumption that I conducted m multiple tests, the null hypothesis that each gene is differentially expressed is true for m_0 tests, and the alternative hypothesis is true for m_1 ($= m - m_0$). Among the m_0 null hypotheses, U hypotheses were declared false-negative and V ($= m_0 - \mathrm{U}$) hypotheses were declared true-positive. Among the m_1 alternative hypotheses, T hypotheses were called true-negative and S ($= m_1 - \mathrm{T}$) hypotheses were called false-positive. R ($= \mathrm{V} + \mathrm{S}$) is the total number of hypotheses rejected and an observable random variable. The FDR was defined as

$$\pi_0 = \begin{cases} E(\dfrac{\mathrm{V}}{\mathrm{R}}) & (\mathrm{R} > 0) \\ 0 & (\mathrm{R} = 0) \end{cases}$$

where E(V/R) means expectation values of V/R. I thus regarded R (1 - π_0) as the number of true active genes (Table 1).

Table 1. Number of intrinsic errors from m null hypothesis tests of a set of genes

	Declared non-significant	Declared significant	Total
Null hypothesis	**U**	**V**	m_0
Alternative hypothesis	**T**	**S**	m_1
Total	m - **R**	**R**	m

3.1.3.5 Identification of promoter sequences by PWM

To estimate σ-factor and TF binding sites, I used position-weight matrices (PWMs). PMWs make it possible to search for sequences highly homologous to those known to be recognized by σ-factors and TFs using PWM. First, I prepared datasets of training sequences consisting of experimentally determined promoters from DBTBS and a reference [Sonenshein et al. 2001], which were aligned on the basis of their consensus regions. PWMs for individual σ-factors and TFs were constructed by the frequencies F_{Ak}, F_{Tk}, F_{Gk}, and F_{Ck} of the four nucleotides (X = A, T, G, C) in the kth position, including the consensus regions and the five bases upstream and downstream.

I determined the score by multiplying all the frequencies corresponding to a given sequence. Second, the thresholds for the binding sites were determined as follows: 4,000 DNA sequences respectively comprising 300 nucleotides were generated randomly taking the GC content of *B. subtilis* into consideration. The threshold was defined by the upper value of the lowest 95% of the maximum scores in individual DNA sequences. Third, within the 300-nucleotide sequence upstream of the protein-coding region, individual σ-factor and TF binding sites were predicted by the maximum PWM score above the threshold because about 95% of these binding sites were known to exist in the regions. I chose optimal matrices for each random sequence, and regarded sequences that exceeded the threshold as being regulated by the σ-factor and TF. Therefore, I used these sequences to search for other sequences highly similar to those recognized by kth gene regulatory protein. This was done by calculating scores, $S(k)_{p(k)}$, or $S(k)_{p(k)\,q(k)}$ for the partial sequences composing $p(k)$ nucleotides for single biding site, or for $p(k)$ and $q(k)$ nucleotides for two binding sites in the stretch of 300 nucleotides upstream of the protein-coding regions of all *B. subtilis* genes. Here, the index k corresponds to kth gene regulatory protein. The scores were estimated using the following equations.

$$S(k)_{p(k)} = \prod_{i=1}^{p(k)} F(k)_{Xi} \qquad \text{(for a single binding site),}$$

$$S(k)_{p(k)q(k)} = \prod_{m=1}^{p(k)} F(k)_{Xm} + F(k)_{int(r)} + \prod_{n=1}^{q(k)} F(k)_{Xn} \qquad \text{(for two binding sites).}$$

Here, x_i is any nucleotide, A, T, G or C and $F(k)_{int(r)}$ is the frequency at interval between these sites,

and *int(r)* is nucleotide number. Sequences whose scores exceeded a threshold were regarded as σ-factor and TF binding sites.

In the case that the PMW for *k*th gene regulatory protein consisting of 5 nucleotides is given by

$$\begin{pmatrix} F(k)_{A1} & F(k)_{A2} & F(k)_{A3} & F(k)_{A4} & F(k)_{A5} \\ F(k)_{T1} & F(k)_{T2} & F(k)_{T3} & F(k)_{T4} & F(k)_{T5} \\ F(k)_{G1} & F(k)_{G2} & F(k)_{G3} & F(k)_{G4} & F(k)_{G5} \\ F(k)_{C1} & F(k)_{C2} & F(k)_{C3} & F(k)_{C4} & F(k)_{C5} \end{pmatrix}.$$

$S(k)_{p(k)}$ for a nucleotide sequence ATGCG is calculated by,

$$\begin{aligned} S(k)_{p(k)} &= \prod_{i=1}^{5} F(k)_{Xi} \\ &= F(k)_{A1} F(k)_{T2} F(k)_{G3} F(k)_{C4} F(k)_{G5} \end{aligned}$$

3.1.4 Results and Discussion

3.1.4.1 The integrated strategy for TU prediction

There are at least 18 different σ factors that direct RNA polymerase, and a large number of sequence-specific DNA binding proteins that play various roles in controlling gene expression, as promoter activators or repressors in *B. subtilis* [Fujita and Fujita 1987; Henikoff et al. 1988]. PWMs for 11 σ-factors and 17 TFs were constructed as shown in Figure 6. Therefore new consensus sequences for the 11 σ-factors and 17 TFs are offered in this study. Consensus sequences for the σ-factors in Table 2 are fairly well consistent with the previous reports in spite of using different training data [Sonenshein et al. 2001], In consensus sequences for 11 σ-factors, these sequences for σD, σL and σW are the same formations between my results and the previous data by Sonenshein et al [2001]. The consensus sequences for σE and σK at the -35 region in this study are the same as the previous data, and the sequences at the -10 region in my results and the past data are "CATAyawT" and "cATAcanT" respectively in σE, and "CATAtnmT" and "CATAnnnT" respectively in σK (Table 2) [Sonenshein et al. 2001]. In these sequences for σE and σK at the -10 region, the 5^{th} and 7^{th} bases in my result differ from those in the past data. In the other σ-factors, these sequences at both the -35 and -10 regions differ slightly from the previous data about a couple of bases, but these different bases mostly exist in less conserved positions of the consensus sequences (Table 2). 77.6% of the bases in the consensus sequences and 95.7% of the bases in the highly conserved positions of the previous results are conserved in my results. Comparing my results with the past data, I used more σ-factor binding sequences for my training data set than those used in the past results, so my results could be proper results particularly in less conserved positions of the consensus sequences.

In the σ-factor and TF binding site searches using PWM, I found putative binding sites regulated by each σ-factor and TF below the thresholds and detected many binding sites in the known binding sites identified by experiments (Table 3). In addition, I use these data for the feature analysis, and Figure 7 shows a comparison of coverage (Fig. 7A) and sensitivity (Fig. 7B) between the 1% and 5% thresholds. I was able to narrow down the candidates for σ-factor and TF binding sites to 26.1% of the candidates (i.e. from 431 to 110 sites) when I changed the threshold from 5% to 1% (Fig. 7A). On the 5% threshold, I identified an average of 78% known binding sites, and an average of 69% binding sites on the threshold of 1% (Fig. 7B). Thus, the average

difference of detecting known binding sites is 9%, corresponding to 3.5 binding sites, by changing the threshold from 5% to 1%. Furthermore, in most gene regulatory proteins I could efficiently narrow down candidates for the binding site, and found that the number of known binding sites detected below each of the thresholds hardly changed. Therefore, I took these PWM analyses at the threshold of 1%.

In several gene regulatory proteins, particularly those having a large number of binding sites, the number of detected known binding sites decreased, as I took the lower threshold. This may be because the TFs have multiple binding sites, upstream of protein-coding regions, whose binding specificities are different [Zheng and Losick 1990; Hamoen et al. 2001]. In the detection of genes regulated by σ-factors and TFs, I found candidate up-regulated genes for 11 σ-factor and 17 TFs, and candidate down-regulated genes for 17 TFs (Table 4). This statistical procedure to identify differentially expressing genes involves a multiple comparison problem because I perform as many hypothesis tests as the number of the candidate genes in microarray. Therefore, I used the false discovery rate (FDR) procedure to remove false-positive data from the candidates of significant expression change data and narrow the candidates for genes regulated by each of the σ-factors and TFs [Pawitan et al. 2005]. FDR is a concept of testing error that is defined as the expected value of the proportion of the inactive genes among the detected genes [Jung 2005]. This enabled us to improve the prediction of candidates for genes regulated by σ-factors and TFs to an average of 33% (range, 12%-49%). In the analyses of genes regulated by each of the σ-factors and TFs, I identified significant microarray data by calculating the *P*-value for each gene, and used the FDR process to eliminate pseudo-positive samples in the detected data. I decreased the candidates for genes regulated by the σ-factors and TFs using this method (Table 4) and was able to efficiently narrow the TU candidates regulated by each of the σ-factor and TFs and detect known operons in the combination approach with FDR control. FDR is a natural scale for work on statistical problems of controlling error rates, and my proposed method could be a useful predictor for searching for TUs on whole genomes.

I then integrated these analyses, and present a comparison of coverage (Fig. 8A) and sensitivity (Fig. 8B) between two integrated conditions (i.e. a 5% threshold at PWM and a 5% threshold in the deleted mutant array without FDR, and a 1% threshold at PWM and a 5% threshold in the deleted mutant array with FDR).

I was able to narrow down the candidate genes composing TUs at the 5% PWM without FDR

to 24.5% of the candidates (i.e. from 194 to 50 genes) when I changed the condition from the PWM 5% threshold without FDR to the PWM 1% threshold with FDR, with 87.1% of genes detected at 5% PWM without FDR also being detected at 1% PWM with FDR. Thus, the candidates can be effectively narrowed without remarkable loss of regulation-known genes under the condition of 1% PWM with FDR. The σL, PerR, and PurR TUs were efficiently detected. Regarding the σL TUs in particular, I could narrow down the 63 TU candidates for the PWM 5% threshold to 9 candidates for the PWM 1% threshold with FDR control without any loss of sensitivity. The detected TUs for the 1% PWM with FDR and known TUs regulated by each of the TFs are listed in Table 5.

Previous predictions of TUs or operons have been based on homologous gene pairs on other genomes, intergenic distance, functional categories and conserved gene clusters among multiple organisms [Salgado et al. 2000; Ermolaeva et al. 2001; Wang et al. 2004; Westover et al. 2005]. In contrast, this approach has the advantage of allowing us to predict gene cluster, without having to depend on homologous gene arrangement in other genomes or functional classification of orthologous genes. I can therefore identify gene clusters at different phases of the life cycle and under various circumstances. The method may be applicable to understanding new functional clusters on the whole genome. However, some identified gene clusters differed from known operons in the number of genes that composed the units. A number of opportunities exist for improving my prediction of gene clusters. Recent studies have used high-density tiling arrays and identified expression in short chromosomal regions [Ishkanian et al. 2004; Bertone et al. 2004]. Such arrays may be helpful to recognize borders of each gene cluster.

As a result, I can detect variously-sized TUs, so the entire TU structure on the *B. subtilis* genome can be estimated on the basis of these predicted TUs. At the 1% PWM with FDR, I can pick 2,183 genes composing 892 TUs, which include known operons, from the complete *B. subtilis* genome. The average size of the polycistronic TU is 3.71 genes, which is comparable in size to those in *Staphylococcus aureus* (3.47 genes) [Wang et al. 2004] and in *E. coli* K12 (3.41 genes) [Salgado et al. 2000]. Distribution of the TUs to the number of genes is almost identical between *B. subtilis* and *S. aureus* (Fig. 9) [Wang et al. 2004]. Thus the operon organization of those two Gram-positive bacteria are fundamentally identical and are approximated by power-law equations, where the correlation of the double logarithm linear relation between the numbers of genes and of TUs composed by the genes is -0.98 in *S. aureus* and -0.97 in *B. subtilis*.

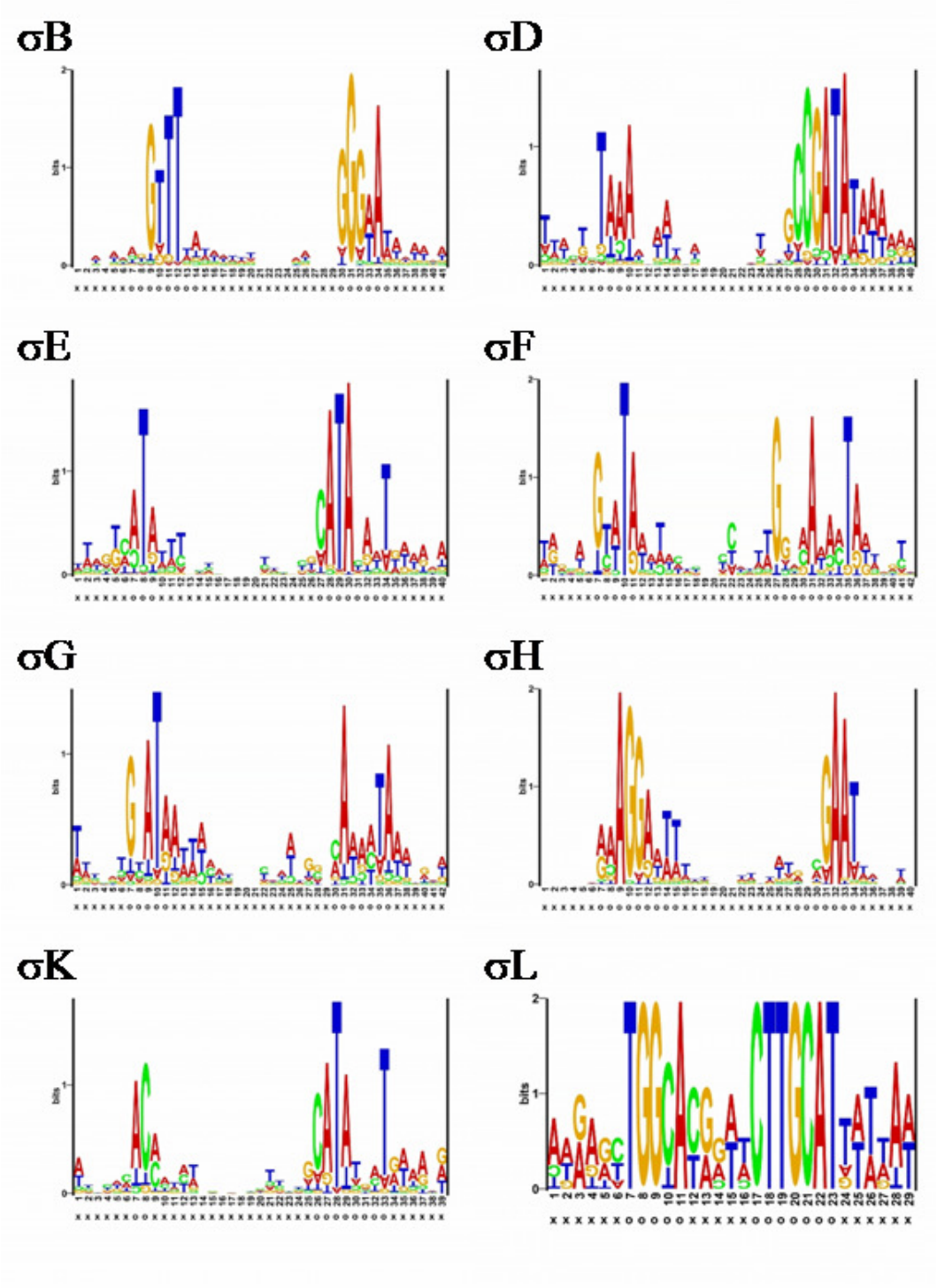


Figure 6. Sequence logos of σ-factor and TF binding sites. The consensus regions are indicated by O in the bottom logos.

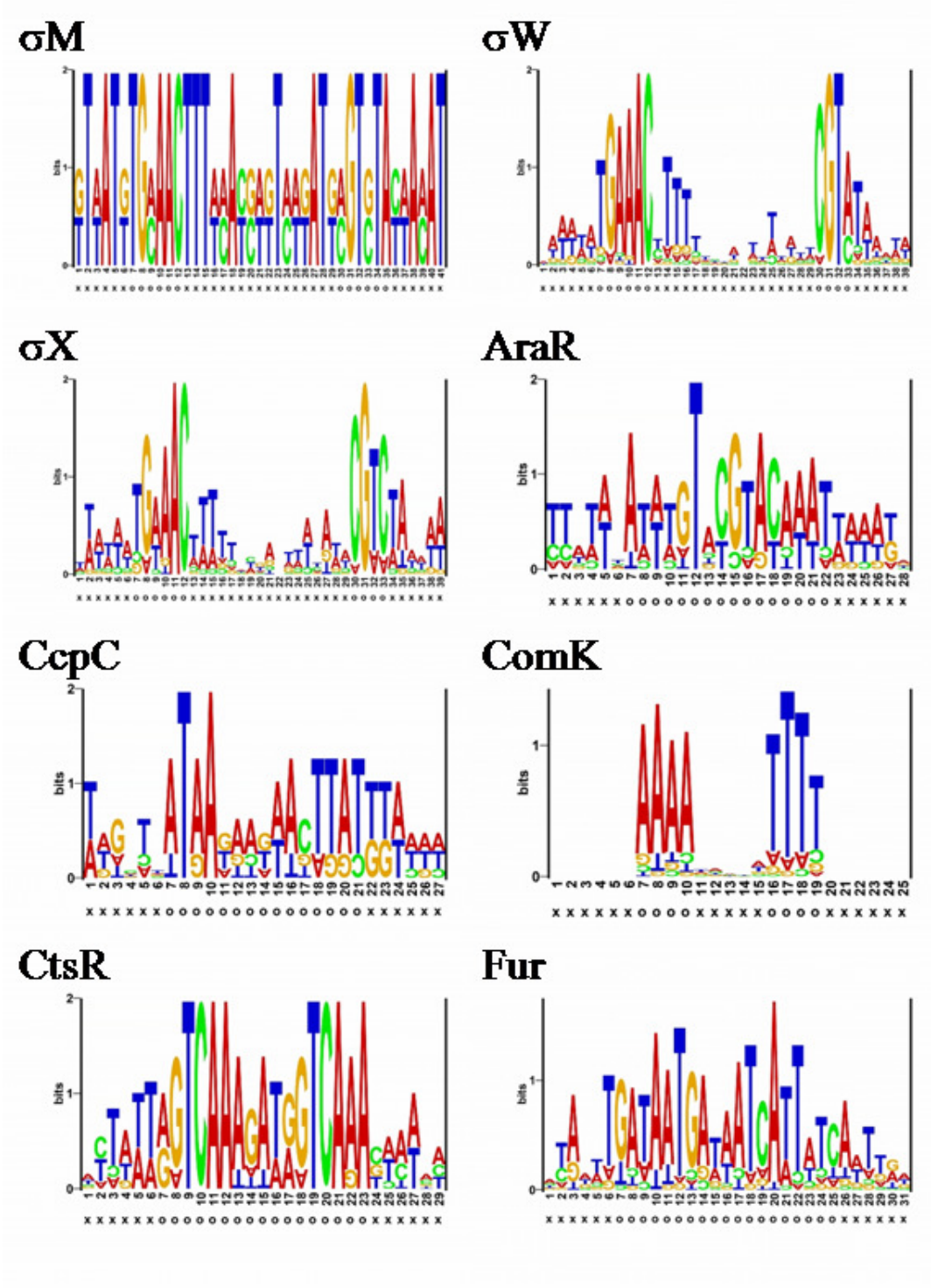


Figure 6. (continued)

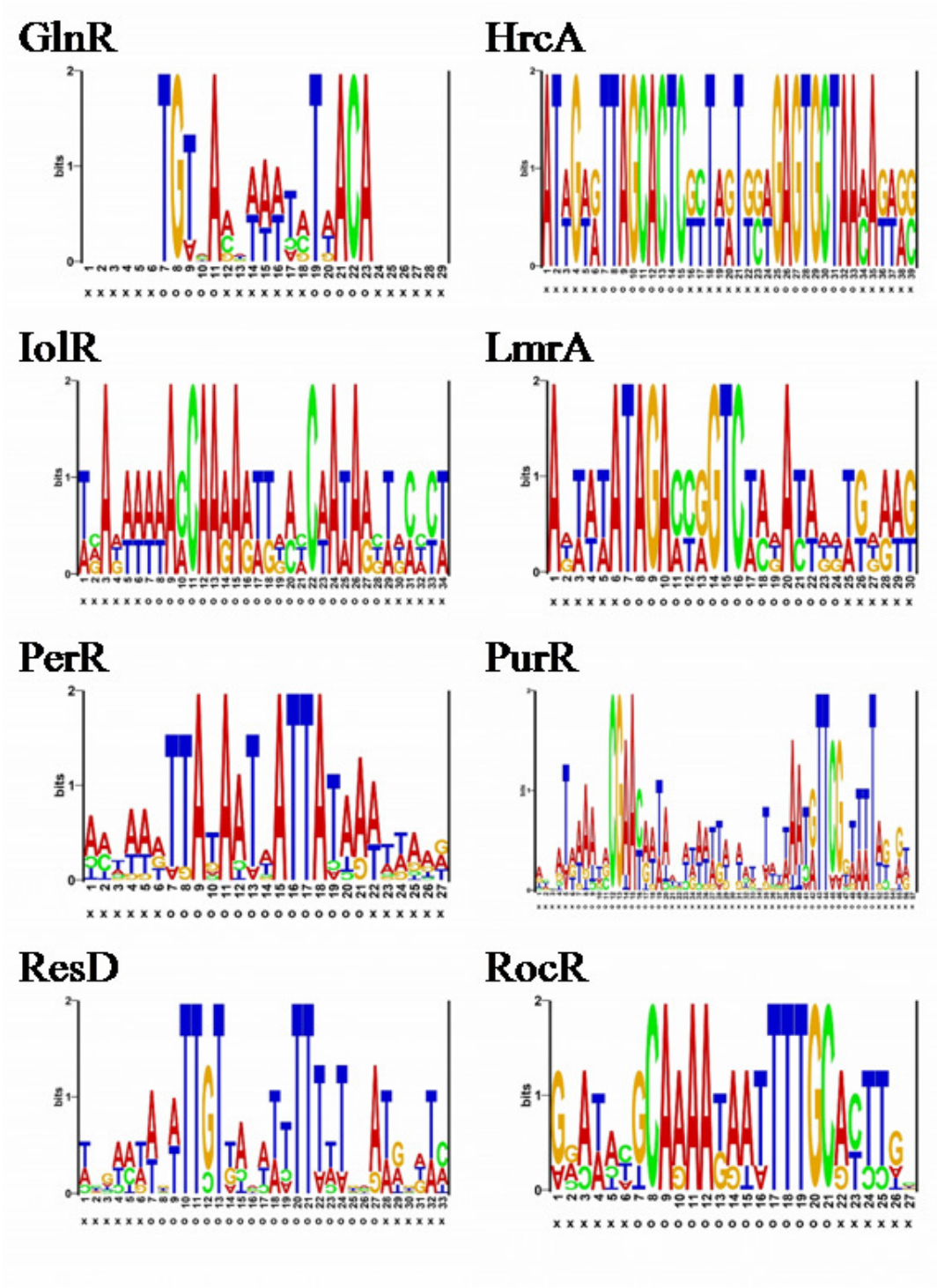


Figure 6. (continued)

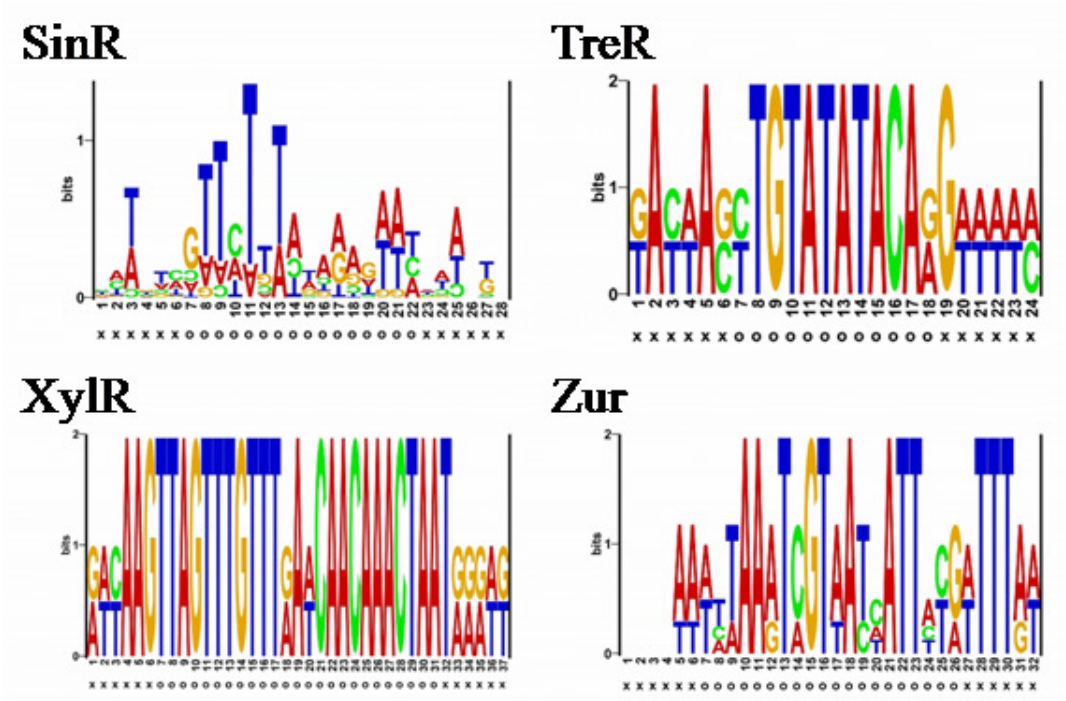


Figure 6. (continued)

Table 2. Consensus sequences for each σ-factor and TF based on the PWM training data in this study.

	Consensus sequence[a]	Binding sites[b]
σB	akGTTTwa 10-20 GGGWAw	95
	(rGGwTTrA 12-15 GGgtAt)	(72)
σD	TAAA 12-16 gCCGATAT	34
	(TAAA 14-16 gCCGATAT)	(17)
σE	ATA 14-17 CATAyawT	68
	(Ata 16-18 cATAcanT)	(35)
σF	GtWTA 14-15 GgnmAwAmTA	15
	(GywTA 15 GgnrAnAnTw)	(13)
σG	GyATA 15-18 mAwaaTA	36
	(gnATr 15 cAtnnTA)	(33)
σH	RAAGGAwTT 7-12 rnmGAAT	51
	(RnAGGAwWW 11-12 RnnGAAT)	(16)
σK	AC 16-17 CATAtnmT	32
	(AC 16-18 CATAnnnT)	(27)
σL	TGGCA 5 CTTGCAT	6
	(TGGcA 5 CTTGCAT)	(6)
σM	TGMAAC 17 MGTSTA	2
	(TGCAAC 16-17 CGTGta)	(1)
σW	TGAAAC 16-17 CGTA	35
	(TGAAAC 16-17 CGTa)	(22)
σX	TGWAAC 16-17 CGTC	16
	(tGtAAC 16-17 CGwC)	(6)
AraR	ATWTGTaCGTACAAAT	8
CcpC	ATAAkaakWAcTTAT	5
ComK	AAAA 5-6 TTTT	68
CtsR	RGTCAAAGATRGTCAAA	7
Fur	GATAATGAwWATCATTwTC	24
GlnR	TGTnA 7 TwACA	6
HrcA	TTAGCACTC 9 GAGTGCTAA	2
IolR	WWAMCAARARWKnMnCWAWARn	3
LmrA	TAGAMYRGTCWMnAYWnn	3
PerR	TTAtAATwATTATAA	11
PurR	WAAtnCGAAYRWTA 16-17 WAATRTTCGkTWWT	10
ResD	WnWTTGTkanwWTTTTwTnnA	6
RocR	GCAAAAKAATTTTGC	5
SinR	gTTmTtWakaRrdWWw	11
TreR	YTGTATATACAR	2
XylR	TTAGTTTGTTTRAWCAACAAACTAA	2
Zur	WnTAAATCGTAATnATTnYG	4

[a] Capital letters indicate highly conserved positions; lowercase letters indicate less conserved positions. Inner figures indicate nucleotide numbers at the interval between two binding sites. Abbreviations: R, A or G; Y, C or T; M, A or C; K, G or T; S, G or C; W, A or T.

[b] The figures indicate numbers of the PWM training data.

Characters in parentheses indicate the consensus sequences for σ-factors published by Sonenshein et al. in 2001.

Table 3. The number of the σ-factor and TF binding sites predicted by position weight matrix

		sigB	sigD	sigE	sigF	sigG	sigH	sigK	sigL	sigM	sigW	sigX	AraR	CcpC	ComK
Candidates[a]	5%	1073	260	1407	326	338	1024	823	451	44	183	155	260	348	761
	1%	284	86	381	94	90	265	197	46	17	71	67	55	71	236
Detected promoters[b]	5%	77/105	26/34	43/68	14/15	25/36	46/51	19/32	5/6	1/2	31/35	15/16	8/8	5/5	19/32
	1%	44/105	25/34	20/68	9/15	18/36	35/51	12/32	5/6	1/2	29/35	12/16	8/8	3/5	14/32
		CtsR	Fur	GlnR	HrcA	IolR	LmrA	PerR	PurR	ResD	RocR	SinR	TreR	XylR	Zur
Candidates[a]	5%	446	472	253	83	340	125	362	371	557	352	478	149	300	339
	1%	96	142	72	19	72	10	92	78	157	78	106	39	104	79
Detected promoters[b]	5%	6/7	19/24	4/6	1/2	1/2	1/3	10/11	8/10	5/6	5/5	8/11	2/2	2/2	4/4
	1%	6/7	19/24	4/6	1/2	1/2	1/3	10/11	8/10	5/6	5/5	7/11	2/2	2/2	4/4

[a]These columns list candidate numbers of the σ-factor and TF binding sites detected by this analysis. 1% and 5% refer to thresholds of this analysis.

[b]These columns consist of Q_i / Q_k, where Q_i is the number of known promoters identified by this analysis, and Q_k is the number of all known binding sites regulated by the σ-factors and TFs.

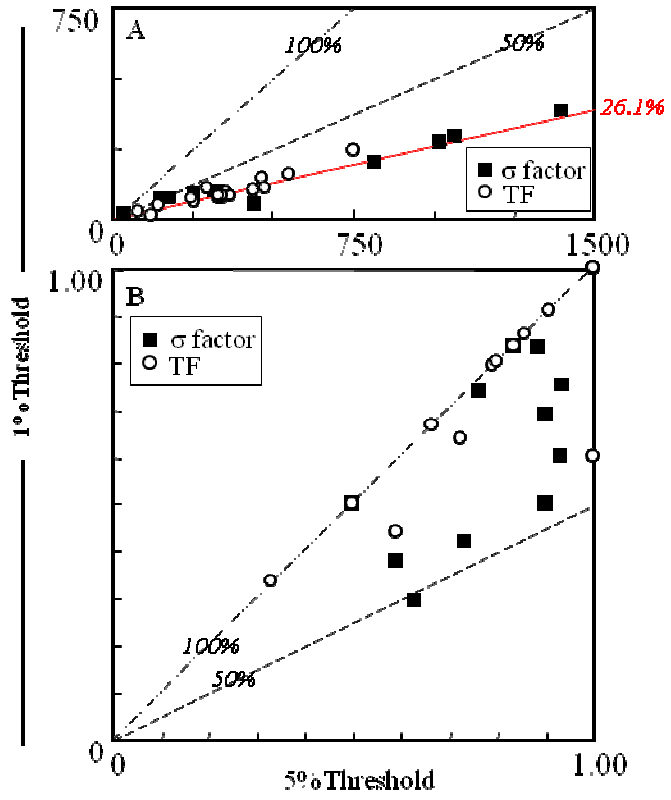


Figure 7. Comparison of coverage and sensitivity in each of the σ-factors and TFs. The coverage of binding sites on the *B. subtilis* genome (A), and the sensitivity of known binding site detection by PWM below the 1% threshold vs. that below 5% threshold (B).

Table 4. Quantification of significantly expressed genes by σ-factors and TFs

Up[a]	sigB	sigD	sigE	sigF	sigG	sigH	sigK	sigL	sigM	sigW	sigX
5%TH[b]	860	624	1115	568	492	465	689	1051	792	900	585
FDR[c]	660	424	915	369	293	265	490	853	591	703	385
(%)[d]	76.7	67.9	82.1	65.0	59.6	57.0	71.1	81.2	74.6	78.1	65.8

Up[a]	AraR	CcpC	ComK	CtsR	Fur	GlnR	HrcA	IolR	LmrA	PerR	PurR	ResD	RocR	SinR	TreR	XylR	Zur
5%TH[b]	410	415	444	521	894	766	552	427	405	1003	588	847	666	481	575	734	738
FDR[c]	212	217	245	322	696	568	354	230	207	806	390	596	466	284	378	537	538
(%)[d]	51.7	52.3	55.2	61.8	77.9	74.2	64.1	53.9	51.1	80.4	66.3	70.4	70.0	59.0	65.7	73.2	72.9
Down[a]	AraR	CcpC	ComK	CtsR	Fur	GlnR	HrcA	IolR	LmrA	PerR	PurR	ResD	RocR	SinR	TreR	XylR	Zur
5%TH[b]	404	607	575	576	557	476	529	458	507	862	481	805	710	526	690	743	657
FDR[c]	206	409	376	377	359	278	331	261	309	665	283	608	510	329	493	546	457
(%)[d]	51.0	67.4	65.4	65.5	64.5	58.4	62.6	57.0	60.9	77.1	58.8	75.5	71.8	62.5	71.4	73.5	69.6

[a] **Up** and **Down** refer to genes that are, respectively, up- and down-regulated by the σ-factor and TF.

[b] These rows list the number of genes having significant expression change in σ-factor and TF deletion mutants of *B. subtilis* by one-sided test under the 5% threshold.

[c] These rows list the numbers of inactive genes removed using FDR from those that are declared active. I regarded TH ($1 - p_0$) as the threshold under the FDR procedure in this analysis.

[d] The bottom rows list percentages of significantly expressed genes remaining after the FDR procedure.

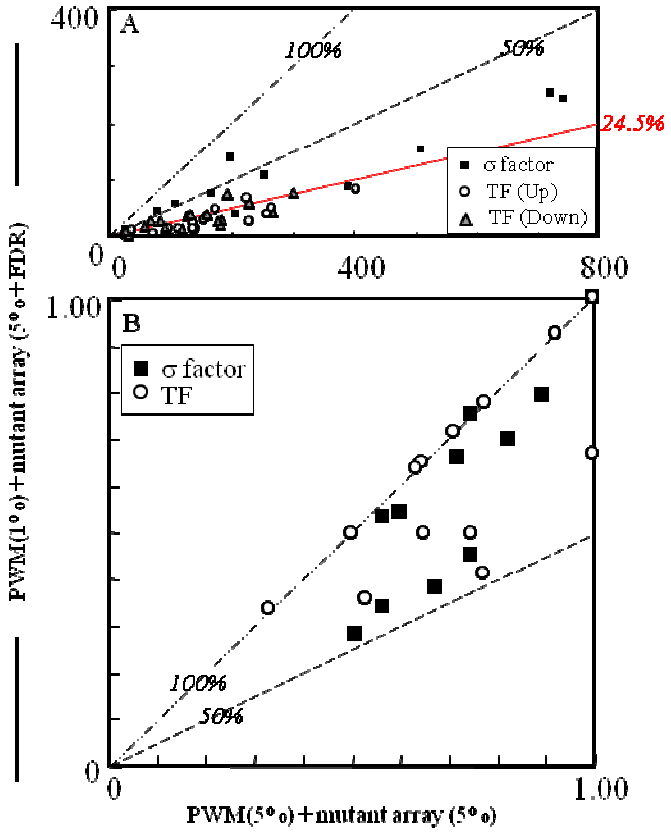


Figure 8. The coverage of genes composing TUs on the *B. subtilis* genome (B1), and the sensitivity of detection of genes composing known TUs by this integrated analysis below the 5% threshold for PWM and the deletion mutant array without FDR vs. that below the 1% threshold for PWM and the deletion mutant array with FDR (B2).

Table 5. Transcription units regulated by the σ-factors and TFs on the *Bacillus subtilis* genome

TF[a]	Identified transcption units[b]
σB (246) **U**	***aldY***, *aprE,* ***bmrU***, ***bofCcsbXyrbE***, ***cdd***, *citCZ,* ***clpP***, ***csbA***, ***csbB***, ***ctc***, *dapAymfA, dnaJKgrpEhrcAhemNlepA,* ***dps***, *glgPA, glnRA,* ***gsiBydbBC***, ***gspA***, ***gtaB***, *iolIH*, ***katX***, ***nadE***, ***opuE***, *paiB, panBbirA,* ***phoRP***, *pksIJKLMNP, ppsCB,* ***rsbVWsigBrsbX*** *ydcFGH,* ***sigAdnaGyqxD***, ***spoVG***, *sunAyolF,* ***trxA***, *uvrB, uvrC, xkdIJKMNO, xkdRS, yaaH,* ***yaaI***, *yabNOPQ,* ***yacHIclpCsmsyacKLM***, ***ybyB***, ***ycbP***, *ycdD, ycdFG, yckC,* ***ycnH***, ***ydaB***, ***ydaC***, ***ydaG***, ***ydaJ***, *ydaKLMN,* ***ydaP***, ***ydaST***, ***ydbD***, ***ydbP***, *ydcOPQR, yddGH, ydfB,* ***ydfO***, ***ydgC***, ***ydhK***, *yebE,* ***yerD***, *yfhD,* ***yfhE***, ***yfhF***, ***yfhJKLM***, ***yfkJIH***, ***yflA***, ***yflS***, *yflT,* ***ygxB***, ***yhaR***, ***yhaUTS***, ***yhcM***, ***yhdF***, ***yhdN***, *yhdP,* ***yheK***, *yhfP,* ***yhxD***, ***yjbCD***, *yjcE,* ***yjfByjgB***, ***yjgCD***, *ykgA,* ***ykgBA***, *ykuL, ykzA, yllBylxAftsL, ynfC,* ***yocB***, ***yocK***, *yosCB,* ***yotK***, ***yoxByoaA***, *ypbR, yphB, yphPilvDypgRQbsaA,* ***ypuCB***, ***ypuD***, *yqeGF, yqfDC,* ***yqgZ***, ***yqhA***, ***yqhB***, ***yqhQP***, *yqiSR,* ***yqxL***, *yrhK,* ***yrvD***, ***yrvIrelA***, ***ytaB***, ***ytkL***, ***ytxJHG***, *yueKJIH,* ***yunG***, *yurMN,* ***yuzA***, *yvaA,* ***yvaK***, ***yvbG***, ***yvgN***, ***yvgO***, ***yvgW***, ***yvgZ***, ***yvrE***, ***yvyD***, ***ywdD***, ***ywjC***, ***ywlE***, ***ywmE***, ***ywmG***, *ywnJspoIIQ,* ***ywtCBAywsC***, *ywtG,* ***yxbG***, ***yxcC***, ***yxiSkatB***, *yxjHG,* ***yxkO***, ***yxnA***, ***yxzFyxlJ***, ***yycD***, ***yycE***
σD (154) **U**	***cheV***, *cotN,* ***degR***, ***dltABCDE***, *dps,* ***epr***, *etfABysiBA,* ***fliIJylxFfliKylxGflgEfliLMYcheYfliZPQRflhBAF-ylxHcheBAWCDsigDylx*** *LrpsB,* ***fliTSDyvyC***, ***hag, lytCBA***, ***lytD***, ***mcpCykwC***, ***motBA***, ***nucAtlpC***, *phrIyddM,* ***sigA***, ***tlpBmcpAtlpAmcpB***, *xkdA, ybdO,* ***yfmTS***, ***yhdD***, *yhfUV,* ***yjbJ***, ***yjcPQ*** *RSyjdA,* ***yjfB***, *ykoVW,* ***ylqB***, *ylxPrbfA,* ***yoaH***, *yocJ, yokFE, yolBA, yomK, yonR, yopJI, yoqLKJ, yorZYXWV-mtbPyorTSRQPONMLKJIHGFEDCBAyoqZY, yosRQPONMLKJIHG,* ***yscB***, *yvaQ, yviFE****flgLKyvyGflgMyvyFcomFC***, *yvzD, ywcHG,* ***yxkC***
σE (315) **U**	*aspShisS, citMyflN, comPX, cotAgabP,* ***cotE***, *cotZYXW,* ***ctaA***, ***cwlJ***, *enopgmtpipgkgapyvbQ, fruRBA, gerBABBBC,* ***gerM***, *glcTptsG,* ***glgPADCB***, *glnHMP, homyutH,* ***mblspoIIIDusd***, ***nucB***, *odhBA, opuAAABAC, opuCDCCCBCA,* ***phoRP***, ***prkA***, *purA, recRyaaL****bofA*** *rrnA, rplBrpsSrplVrpsCrplPrpmCrpsQrplNXErpsNHrplFRrpsErpmDrplOsecYadkmapinfA-rpmJrpsMKrpoArplQybxAybaEFtruArplMrpsIybaJK****cwlD***, *rpsPylqCDEtrmDrplS,* ***sigG***, ***spmBAdacB***, ***spoIID***, ***spoIIIAHIIIAGIIIAFIIIAEIIIADIIIACIIIABIIIAA***, ***spoIIM***, ***spoIVA***, ***spoIVCA***, ***spoIVCB***, ***spoIVFBI***, ***spoVB***, ***spoVD***, ***spoVEmurG***, ***spoVK***, ***spoVM***, ***spoVR***, *spsKJIGF, sspEygaB, tig, tkt, trePARyfkO, uppglyA,* ***valSysxE***, ***yaaH***, *yabGvegsspF,* ***yabMNO***, ***yabPQ***, *yacD, ybaL,* ***ybaN***, *ybbCDEF,* ***ycgFG***, ***ydcA***, ***ydcC*** *dal, ydhF****phoB***, *ydjDE,* ***yfhS***, *yfjS, yfkRQ, yhaUTS,* ***yhaX***, ***yhbH***, *yhcOP, yheBA, yheM, yhjM, yisCDEFGH,* ***yisO***, *yjbFG,* ***yjbX***, ***yjmEFGHIJ***, ***yknT***, *yknU, yknWXYZ, yktA,* ***ykvI***, *ykvS,* ***ykvUV***, *ylaJ, ylaK, ylbBCD,* ***ylbJ***, *ylbKL, ylbQyllA, yndL, yneNtlp,* ***yngDEFGHI***, *yoaU,* ***yoaW***, *yotK, ypbF, yphB,* ***ypjB***, *yptA, yqeVUTdnaJKgrpEhrcAhemNlepA,* ***yqfDC***, ***yqfZY***, ***yqiQmmgEDCBA***, *yqjXW,* ***yqxAspoIIP***, *yqzG,* ***yrbA***, *yrkC, yrzF, ysdC, ysnD, ysnE, ytaGFmutMpolA, ytcAB, yteV, ytlQ, ytvI, ytxC, yugO, yuiH, yunB, yuzC, ywcA, ywfKpta, ywlD, ywqFE, ywrJcotB, yxeC, yxiQ, yxjFEDC, yyaD*
σF (119) **U**	***bofC***, *enopgmtpipgkgapyvbQ,* ***gerAAABAC***, ***katX***, *lonB, miaA, relAaptyrvED, rpoBCybxFrpsLGfustufAybaCrpsJrplCDWBrpsS-rplVrpsCrplPrpmCrpsQrplNXErpsNHrplFRrpsErpmDrplOsecYadkmapinfArpmJrpsMKrpoArplQ ybxAybaEFtruArplMrpsIybaJKcwlD,* ***sigG***, *sinI, spo0AI,* ***spoIIQ***, ***spoIVB***, *spoVAFVAEVADVACVABVAA****sigFspoIIABIIAAdacF***, ***sspE***, *thiA,* ***tlp***, *yabKmfdspoVTyabMNOPQ, yerQyefA, yetF,* ***yhcNOP***, *yhfW, ykvI, ylbBCD, ynaC, yncC, yqxAspoIIP****gpr***, *yqzG, ytfJI, ythA,* ***ytkD***, *yuxHyueKJI, ywfN,* ***ywhE***, *ywlB****spoIIR***

Table 5. (Continued)

TF[a]		Identified transcption units[b]
σG (95)	**U**	*adhByraED, bofC, cotK, csgAybxH, enopgmtpipgkgapyvbQ,* ***gerAAABAC*** *,* ***gerBABBBC*** *,* ***gerD*** *, gpr, malSytnP, sigG, splAB, spoIVB, spoIVCA,* ***spoVAFVAEVADVACVABVAA*** *sigFspoIIABIIAAdacF,* ***sspB*** *,* ***sspC*** *, sspD, sspE, sspF, yabKmfd* ***spoVT*** *yabMNOPQ, ybaKcwlD,* ***ycxEgdh*** *, ydbI, yfhD, yfhS, yfjQ, yfjS, yfjU,* ***yhcN*** *OP, yhcQ, yhfM, yhjGH, ykjAykkA, ykoG, ykuJKykzF, ykuV, ykvV, ylaJ, ylaM, ymfJ, ypeB* ***sleB*** *, yqfS, yqzG, ysfA, yteJIsspA, ythA, yvsH, ywfBA, ywfI, ywhE, yxjA*
σH (279)	**U**	*accAyttI, araQP,* ***ccdA*** *,* ***citG*** *, codY, dal, deaDyxiMLKJIyxzGyxiHGyxzCyxiFyxxG, dgkAyqfG, divIC,* ***ftsA*** *,* ***glgPAD*** *, glnA,* ***gltP*** *,* ***kinA*** *,* ***lytE*** *,* ***minDC*** *, mtrAhbs, opuABA, panCB,* ***phrC*** *,* ***phrE*** *,* ***phrF*** *,* ***phrG*** *,* ***phrK*** *, pksJKLMNP, ppsCBA* ***pbpyoxAyoeA*** *, proJH, rapG, rplWBrpsSrplVrpsCrplPrpmCrpsQrplNXErpsNHrplFRrpsErpmDrplOsecYadkmapinfArpmJrpsM KrpoArplQ-ybxAybaEFtruArplMrpsIybaJKcwlD, rtp,* ***sigAdnaG*** *,* ***spo0AyqxM*** *,* ***spo0F,*** *spoVAFVAEVADVACVABVAA* ***sigFspoIIABIIAA*** *,* ***spoVG*** *gcaDprs,* ***spoVStdh*** *kblymcBA, srfACAycxA,* ***ureCBA*** *,* ***vpr*** *, yabPQ, yclF,* ***ycsN*** *, ydgE, ydhG, ydjO,* ***yesLMNOPQRSTUVWXYZyetA*** *lplABCDyetF, yfhD, yfjDC,* ***ygaI*** *, ygxA,* ***yhaR*** *, yhbI,* ***yhcR*** *, yheM,* ***yisK*** *,* ***yisT*** *, yjmABCDEFGHIJ, ykgA, ykkAB, yknYZ,* ***ykoM*** *,* ***ykrQ*** *dat, ylmDEF,* ***ymaH*** *ymzC* ***, ynaD*** *, yneF, yngL,* ***yobO*** *, yocKLM, yodL,* ***yojL*** *, yonV, yoqDCBA, yosTS, yotKJIH, yozNyocN, yphPilvDypgRQ* ***bsaA*** *,* ***ypiB*** *, ypmTSRQ, yppD, yqcG, yqeB, yqgY, yqiG, yqiQmmgE, yqjBA, yrpD,* ***yrvJ*** *, yteVUTSR, ytiBA,* ***yttP*** *,* ***ytxJHG*** *, yusJKLM,* ***yuxI*** *yukJ,* ***yuxL*** *, yuxO, yvaX,* ***yveRQPONMLK*** *, yvqHI,* ***yvyD*** *,* ***ywfF*** *, ywhG, ywkBAtdkrpmErhoywjImurZywjHfbaA, yxiB, yycR*
σK (117)	**U**	*atpCDGAH, bglS,* ***cgeAB*** *,* ***cgeEDC*** *,* ***cotA*** *,* ***cotC*** *,* ***cotD*** *,* ***cotE*** *,* ***cotF*** *,* ***cotG*** *,* ***cotH*** *,* ***cotM*** *,* ***cotT*** *,* ***cotZY*** *,* ***cwlC*** *, fliQR,* ***ftsY*** *, gbsBA, ggaB, pgi, pksSymzBymaE, racEysmB* ***gerE*** *, ribH,* ***spoIVCB*** *,* ***spoIVFBI*** *, spoVEmurGBdivIB,* ***spoVFAVFBasddapGA*** *ymfA* ***, spoVK*** *, spsKJIGF,* ***yabG*** *, ybdBD, yceFG,* ***yfhP*** *, yfnHGFED, yhcP, yisCDEFGHI, yjgCD, yloCDHI,* ***yoaN*** *,* ***yobW*** *, yodHI, yonU, ypbH,* ***yqeE*** *, yqzDC,* ***yrbBA*** *, ytlABCD, ytsD, ytxOcotSytxN, yueG, yugNM,* ***yurS*** *, yuzE, yvcSR, yvdO, yvfH, ywqM, ywrJ* ***cotB*** *, yxbDC*
σL (39)	**U**	***acoABCL*** *RyfjU, bfmBBBABBAA* ***yqiVUTS*** *, fnrnarKargS,* ***rocFED*** *,* ***sacClevGFED*** *, yuxHyueKJ,* ***yweB*** *, ywfEDCBA* ***rocCBA*** *, ywiE*
σM (10)	**U**	*divIC,* ***yhdKLM*** *, ypuA, yrhJIH, ywhB, ywoA*
σW (77)	**U**	***abh*** *,* ***divICyabR*** *, pbpX,* ***sigWybbM*** *, xpaCyaaN,* ***ybfOP*** *, ycdB,* ***yceCDEFGH*** *,* ***ydbST*** *, ydcB,* ***ydjF*** *,* ***ydjOPyeaA*** *, yebC, yfhLM,* ***yjbCD*** *,* ***yjoB*** *rapA,* ***yknWXYZ*** *,* ***yndN*** *,* ***yoaF*** *,* ***yoaG*** *,* ***yob1J*** *, yozO, yqfDC* ***yqfBAyqeZ*** *, yqgZ, yrhH, ysdByscB,* ***yteJI*** *, ythB, ythQ* ***ythP*** *,* ***yuaIGF*** *, yveGF* ***racXpbpE*** *,* ***yvlDCBA*** *,* ***ywaC, ywbN, ywnJ, ywoA, ywrE, yxjI, yxzE***
σX (58)	**U**	***abh*** *,* ***csbB*** *yfhO, degAyisS,* ***divIC*** *,* ***dltABCDE*** *,* ***lytR*** *,* ***pbpX*** *,* ***pssAybfMpsd*** *ybfN,* ***rapD*** *, slp,* ***yceCDEFGH*** *, ycgQR, ydaJKLMN,* ***ydcB*** *,* ***yjbCD*** *, yngC, yobIJ, yppBponA, ypuA,* ***ypuNsigX*** *,* ***yrhH*** *, yrvEDCsecDF, ytnAasnB,* ***ywbN*** *,* ***ywnJ*** *,* ***ywoA*** *, ywtFE, yxeED*
AraR (30)	**U**	*gpsAyphC, rpoArplQ, yeeABC, yoaJ, yorZ, ypuIHG*
	D	***araE*** *,* ***araR*** *, prsAyhaK,* ***xsa*** *, ysfEcstA* ***abfAaraQPNMLDBAabnA*** *, yydK*
CcpC (26)	**U**	*degUS, ggaBA, grlBA, yjcM,* ***ykuM*** *, ykwC, ynzFG, yorWV*
	D	***citB*** *yneN,* ***citZ*** *, mraYmurDspoVEmurG, spo0B, ykrL, yndEF, yurL, yybF*
ComK (116)	**U**	***addBA*** *, bofCcsbX,* ***comC*** *,* ***comECEBEA*** *,* ***comGGGFGEGDGCGBGA*** *,* ***comK*** *,* ***cspB*** *,* ***glcRywpH*** *,* ***hipO*** *, levGFED,* ***medyjzA*** *,* ***ninnucA*** *,* ***nrgAB*** *,* ***pta*** *,* ***rapH*** *,* ***recA*** *, rplBrpsS,* ***rpsFyyaF*** *,* ***smf*** *,* ***topA*** *, xpaCyaaN, ybdBD,* ***ybdK*** *,* ***ycbL*** *,* ***yhjB*** *, yhxD,* ***yhzC*** *,* ***yjbF*** *yjbG, yjdA, ylnF,* ***yndG*** *,* ***ynzC*** *, yomK, yozO, yqgML,* ***yqzG*** *,* ***ysxA*** *, yvcT,* ***yvrP*** *, yweA,* ***ywfL*** *, ywnJspoIIQ,* ***ywzA*** *,* ***yyaAgidBAthdF***
	D	*aroHBFcheR, atpEB, citG, dnaXyaaK, dppEykfABC, purT,* ***rok*** *, sdhA, sigA, spo0A, yddRS, ydeM, yhjR, ykuJK, ykvJK, ylbBCD, ypfD, yqgQP, yrhJI, ysxD, ytgBA, ytpI, yvdS, ywnH, yyaJ*

Table 5. (Continued)

TF[a]		Identified transcption units[b]
CtsR (68)	**U**	*cotN,* ***ctsRyacHIclpCsmsyacK*** *LM, gltBA, motBA* ***clpE*** *, tlpBmcpAtlpAmcpB, yhcP, ykcBC, yolJIsunTA, yopN, yorSR, yqzC, yviFEflgLKyvyGflgMyvyF, ywjC, yycJI*
	D	***clpP*** *, motAclpE, ndhFybcC, spoVStdhkblymcBA, ykhAhmp, yloIpriA, ynbBglnR, yosDC, yraFadhB, yrbC, yulEDCByuxG, yuxLthrB, yvcBA*
Fur (135)	**U**	*ccdA, dnaGyqxD, gcaDprs, metBypfP,* ***nasFE*** *, pdhBCD, pdpnupCdra, ybaRS, ycdHIyceA, ycgJ, ycgRS, ydbL, ydhMNOPQRST,* ***ydhU*** *, yfnA, yjdA, ymcBA, yncB, yptA, yqfON, yuzC, yvbTU, ywhG, yxnBasnH, yydD, yydJ*
	D	*ahpCF, fhuCGB,* ***fhuD*** *, hemLBDCXA, mrgA, tig,* ***ybbAfeuCBAybbB*** *,* ***ybbBCDEFHIJK*** *, ycgT, yclNOPQ, ydbN,* ***yfhC*** *, yfiY, yfiZyfhA, yfjR,* ***yfkM*** *, yfmFEDC,* ***yhfQ*** *,* ***ykuNOP*** *, ykvW,* ***yoaJ*** *, yolKJIsunTA, yorJIHGFE, yppQP, ytsP,* ***yuiI*** *,* ***yukLMdhbFBECA*** *, yumC, yurUVWXY,* ***yusV*** *, yvgZ,* ***ywbONML*** *,* ***ywjBA*** *,* ***yxeB***
GlnR (53)	**U**	***glnRA*** *, spoIVCA, yclJK, yexApurLQFMNHD, yhcGHI, yqhN, yrkLK, yyaC, yybMLKJ*
	D	*alsT,* ***nasA*** *,* ***nasB*** *,* ***nasC*** *,* ***nasFED*** *, nrgA, tnrA,* ***ureCBA*** *, xkdNO, ykzBykoL, yoaZ, yqjED, yqkKJ, ytpB, yunGH, yunIJKLM*
HrcA (19)	**U**	***yqeVUTdnaJKgrpEhrcA*** *, yurXY*
	D	***groESEL*** *, ybaRS, ydiKL, yerQyefA, ytqAB*
IolR (32)	**U**	*cmk, yjbCD, yndB, yobJ, yokFE, ypiBA, yybNMLKJ*
	D	*ahpCF,* ***iolRS*** *, msmRE, pdpnupCdradeoR, proSpolC, ydhOP,* ***ydjK*** *, yojA, ywfBA*
LmrA (7)	**U**	***lmrBA***
	D	*yfkCBA,* ***yxaHG***
PerR (127)	**U**	*citAyhdF, dacAyaaDE, murEmraYmurDspoVEmurGBdivIBylxW, oppABCDFyjbB, sigWybbM, ybdKL, yclNOPQ, yluAcdsAyluBCproSpolCylxSnusAylxRQinfBylxPrbfAtruBribCrpsOpnpAylxYymxG, ynzD, yoaEF, yolKJIsunTA, yorGF, ytcFgapB, yueCByukABC, yusA, ywjA, yxeB, yxkC*
	D	***ahpCF*** *, citR, gltBA,* ***hemLBDCXA*** *, hemY,* ***katA*** *,* ***mrgA*** *, murEmraYmurDspoVEmurGBdivIBylxWXsbp, ppsCBApbpyoxAyoeA, yaaH, ydbO, ydeLM, yfkM, yfmJ,* ***ygaG*** *,* ***ykvW*** *, ylmAB, yncB, yobL, yokJI, yozIyobE, yphPilvD, yppQ,* ***yqkL*** *, yrhED, yrhF, yukLMdhbFBECA*
PurR (81)	**U**	*clpCsms, flhPO, infArpmJ, lrpA, nusAylxRQinfB,* ***purRyabJ*** *, rplErpsN, yacBCD, ycgFG, ydeST, ykuG, ylqFrnhylqGH, yopJI, yorDC, yxkD*
	D	*appDFAB, atpBIupp* ***glyA*** *, gltP,* ***pbuXxpt*** *, ppsApbpyoxAyoeA,* ***purA, purEKBCyexApurLQFMNHD*** *, recNahrCyqxCyqiEDCBfolD* ***yqhZ*** *, yaaDE, ydaR,* ***yebB*** *, yfjBA, yjdJ, ykaAykbA,* ***ytiP*** *,* ***yumD***
ResD (159)	**U**	*codY,* ***ctaA*** *, cysH, deaDyxiMLKJIyxzGyxiHGyxzCyxiFyxxGwapAyxxFyxiE,* ***fnr*** *, hemLBDCXA,* ***hmp*** *,* ***nasFED*** *, smbAfrryluAcdsAyluBCproSpolCylxSnusAylxRQinfBylxPrbfAtruBribCrpsOpnpAylxYymxG, ybbAfeuCBAybbB,* ***yclJ*** *, yeeAB, yerQyefA, yjdB, yobB, yocE, ypuI, yqgS, yraON, yrkON, ytbQbioIBDFAW, yttB, ytzFGytgP, yvbT, yxaLK*
	D	*adaB, bprspoIIGAsigEG, ccdA, dppABCDEykfABCD, odhBA, oppAB, rocFED, spoIIEyabST, spoVStdh, ungywdF, ybbEFHIJK, ybxG, yclM, yczG, ydcMN, yddT, yeeAB, yfnA, yhjGH, yitK, yknA, ylbA, yomL, yosQP, yrhG, yrhJI, ysnBA, yurL, yutJ, yvaDEF, ywbBA, yweA ywfCBArocCBA, ywhA, yxaBA, yybP, yydC*
RocR (39)	**U**	***rocFED*** *,* ***rocR*** *, rpmErhoywjImurZywjHfbaA, trpFC, ycgJ, yomK, ywfCBA* ***rocCBA*** *, ywiEnarIJHG*
	D	*appBCyjbA, bglA, comGGGF, fhuD, oppBCDFyjbB, tdkrpmE, ybgHJ, yitM, yodPargE, ytcQ*

Table 5. (Continued)

TF[a]		Identified transcption units[b]
SinR (95)	**U**	*cheV, fliIJylxFfliKylxGflgEfliLMYcheYfliZPQRflhBAFylxHcheBAWCDsigDylxLrpsB, lytCBA, odhBA, sucCD, tlpBmcpAtlpAmcpB, wprA, yaaDE, yopR, yqiDCBfolD, yrkA*
	D	***aprE***, ***comK***, ***cotNsipWyqxM***, *glcRywpH, kinB,* ***sigFspoIIABI***, *slr,* ***spo0A***, ***spoIIGAsigEG***, *spoIIIJrnpA, spsKJIG, yesOP, yfmPO, ykrL,* ***ykuW***, *yrdBA, ytgDC, yurK,* ***yveTSRQPONMLK***, *yvfEDCBA*
TreR (30)	**U**	*atpIupp, pksS, ypfBA*
	D	*araNMLD, csn, murFydbRST, narA, pyrDF, tagHG,* ***trePAR*** *yfkO, xkdKM, yfkRQ, yhdI, yqjTS*
XylR (54)	**U**	*sinI, xylR, yceCDEF, ygaE, yocJ, yqiYX, yviEflgLKyvyG*
	D	*mtlAD, pyrRP, xtmBxkdEFG,* ***xylAB***, ***xynB***, ***xynP***, *ycgLM, ydaJK, yfnHG, yhaGserC, yhfEF, ykrTU, ymfFGHIJ, ytxEDccpA, yulByuxG, yusKL, yvaN, ywdJI*
Zur (54)	**U**	*amhX, divIC, dps, yabFksgA, yceI, yefB, yfiO, yloW, yolJIsunTA, yomL, yopKJIHGFEDC, yqcKJ, yuxOcomA, yybKJ*
	D	*brnQazlDCB, citR, ctrArpoE,* ***ycdHIyceA***, ***yciABC***, *yfhC, yhzA, yodDE, yraO, yrpE, ytiBA, yuxOcomA, yvgQR*

[a] **U** and **D** refer to genes that are up- and down-regulated, respectively, in the presence of the σ-factor and TF. Figures in parentheses indicate number of genes regulated by the σ-factors and TFs.

[b] All data were identified under 1% threshold in PWM and 5% threshold in deletion mutant array with FDR. Among these genes, the known genes regulated by the σ-factor and TF are indicated in bold.

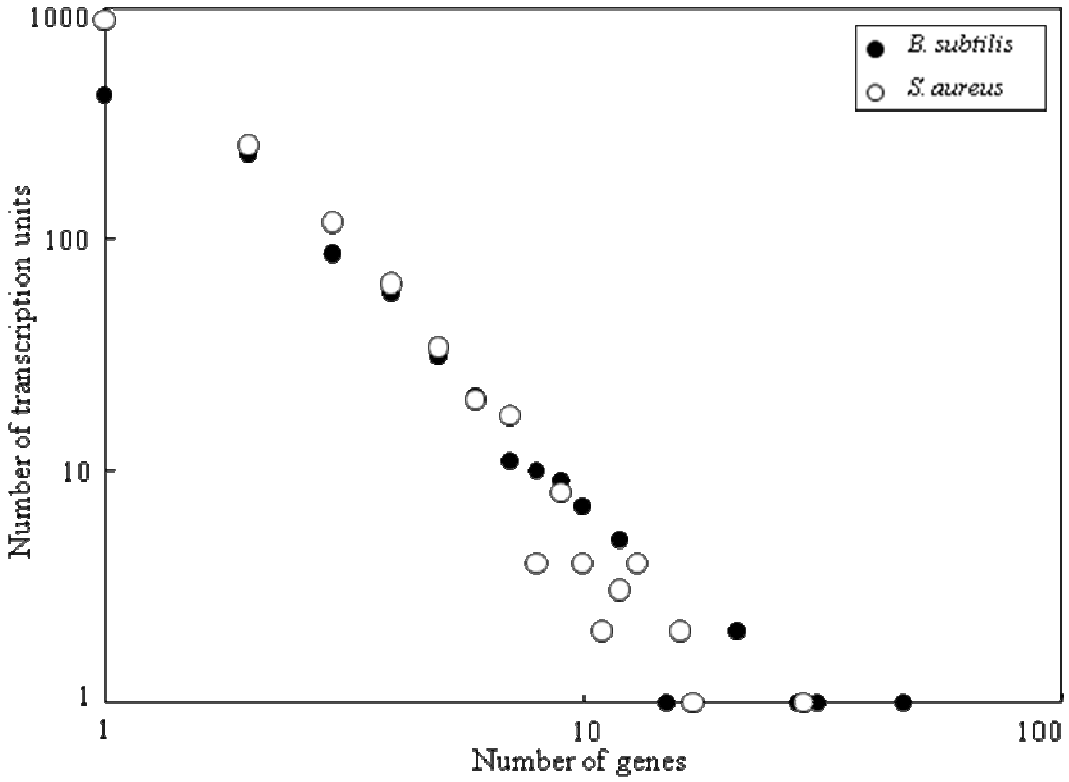


Figure 9. The relationship between the number of TUs detected in the present study and the number of genes composing the TUs.

3.1.4.2 Functions of TUs in *B. subtilis*

3.1.4.2.1 COG functional classification of genes regulated by σ-factors and TFs

Using the TU data, I examined the transcriptional regulation of genes by 11 σ factors whose promoter sequences have been characterized. The properties of individual σ factors are as follows: five σ factors (σE, σF, σG, σH, σK) regulate sporulation through morphological stages that involve the conversion of the growing cell to a two-cell sporangium, which ultimately proceeds to a single spore; σB mediates the general stress response, and more than 150 protein-coding genes for general stress belong to the σB regulon [Hecker et al. 1996]; σL mediates cold-shock adaptation and regulation of the acetoin catabolic pathway [Wiegeshoff et al. 2006]; σD regulates flagellar synthesis, motility, and chemotaxis [Marquez-Magana and Chamberlin 1994]; σM mediates salt resistance [Horsburgh and Moir 1999]; and σX and σW play modulatory roles in extracytoplasmic function [Turner and Helmann 2000]. All the regulative relations of the 11 σ factors and 17 TFs to targeted genes are listed in Table 5, making it possible to characterize individual σ factors according to the genes they target. Therefore, I classified genes belonging to each of the TUs into 19 COG (clusters of orthologous groups of proteins) functional categories [Tatusov et al. 1997] for estimating the general roles of the σ factors and TFs in cellular processes (Fig. 10).

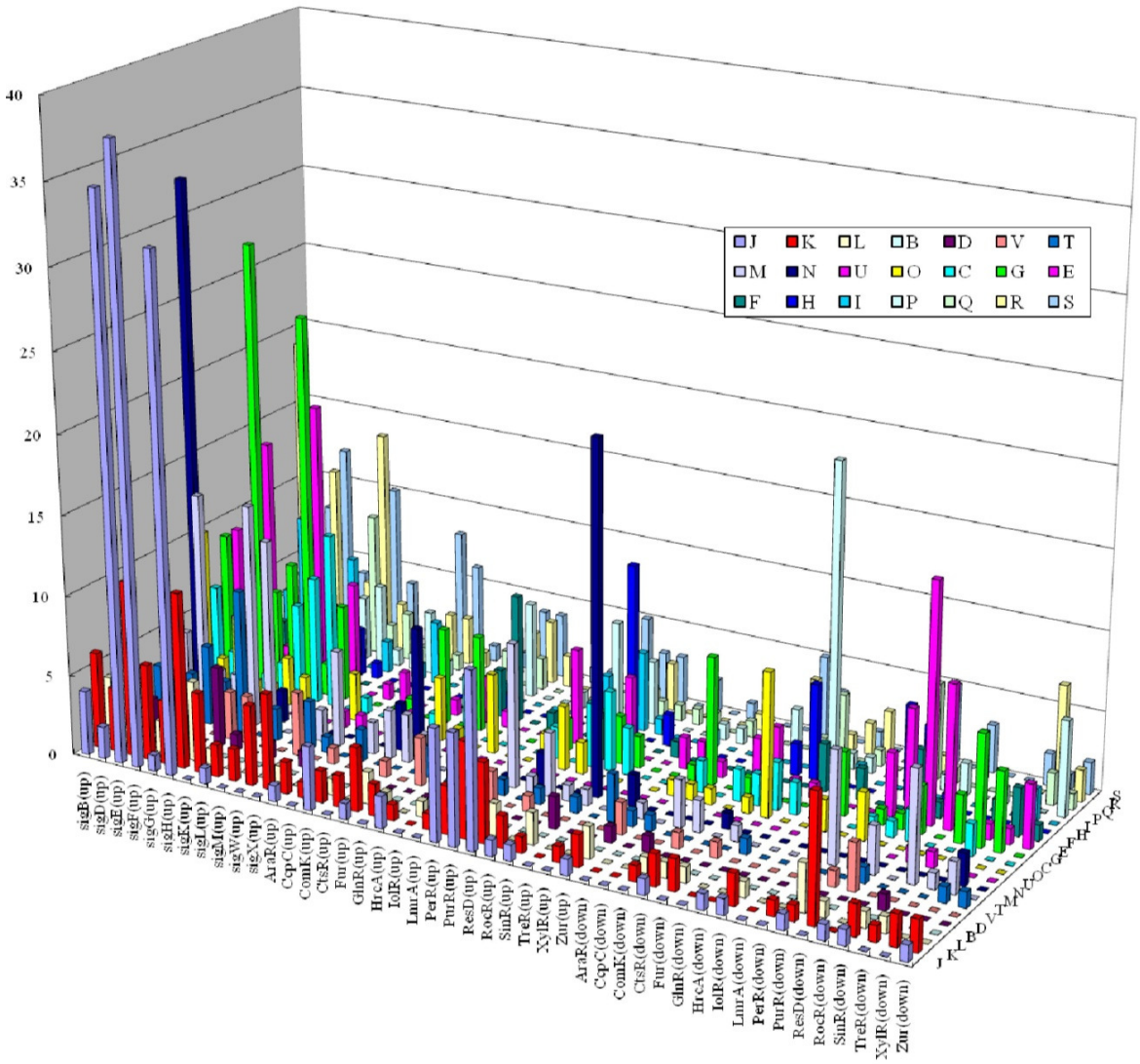


Figure 10. Classification of genes regulated by each σ-factor and TF by COG functional categories.

One-letter abbreviations use used for the functional categories: J, translation, ribosomal structure and biogenesis; K, transcription; L, DNA replication, recombination and repair; B, Chromatin structure and dynamics; D, cell division and chromosome partitioning; V, Defense mechanisms; T, signal transduction mechanisms; M, cell envelope biogenesis and outer membrane; N, cell motility and secretion; U, Intracellular trafficking and secretion; O, posttranslational modification and protein turnover, chaperones; C, energy production and conversion; G, carbohydrate transport and

metabolism; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; H, coenzyme metabolism; I, lipid metabolism; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown.

3.1.4.2.2 Functional relationships between categories of genes regulated by σ-factor and TFs

The similarity of the roles in cellular process between individual σ-factor and TFs was estimated using Pearson correlation coefficients for the number of genes belonging to each of the COG categories (Fig. 11).

The five σ factors associated with regulation of the sporulation process can be classified into three groups corresponding to the sporulation process Stage 0-III (σH, σF and σE) characterized by category [J; translation, ribosomal structure and biogenesis], Stage IV (σG) characterized by the category [G; carbohydrate transport and metabolism], and Stage V (σK) characterized by category [M; cell envelope biogenesis and outer membrane]. Gene expression under the σG control occurs in the prespore, and the main functions are to protect the spore from several hazardous conditions, high osmotic pressure [Tovar-Rojo et al. 2003], UV radiation and dry heat [Setlow 1995], and to prepare the spore for germination and outgrowth [Hilbert and Piggot 2004]. In this process, σG regulates carbohydrate content in the cell, for example, by activating expression of the glucose dehydrogenase operon [Nakatani et al. 1989], controlling metabolism of the tricarboxylic acid cycle [Magill et al. 1996] and glucose uptake [Lorca et al. 2003]. σK is synthesized and becomes active in the mother cell, and directs formation of the spore coat and spore maturation [Hilbert and Piggot 2004]. Therefore, these previous experimental studies are consistent with the present results. Moreover, I can observe that each TF in a cluster has one of the frequently detected functional categories (Fig. 11). The AraR protein is well known as a negative regulator of the L-arabinose metabolic operon [Sa-Nogueira and Mota 1997], and most of the genes negatively regulated by AraR belong to [G] (Fig. 10). Almost all the genes up-regulated by SinR are in category [N], which consists of proteins controlling cell motility and secretion, while the down-regulated genes belong to category [M], which consists of proteins operating cell-wall and membrane biogenesis (Fig. 11). ComK synthesis is regulated by a series of reactions that involve quorum sensing; SinR is one of the activators in this cascade, acting negatively on *rok* transcription [Hoa et al. 2002], and is known to be a potent repressor of biofilm formation [Kearns et al. 2005]. Thus, the analysis presented here agrees well with previous experimental data and enables us to assess the roles of the σ factors and TFs in cellular processes.

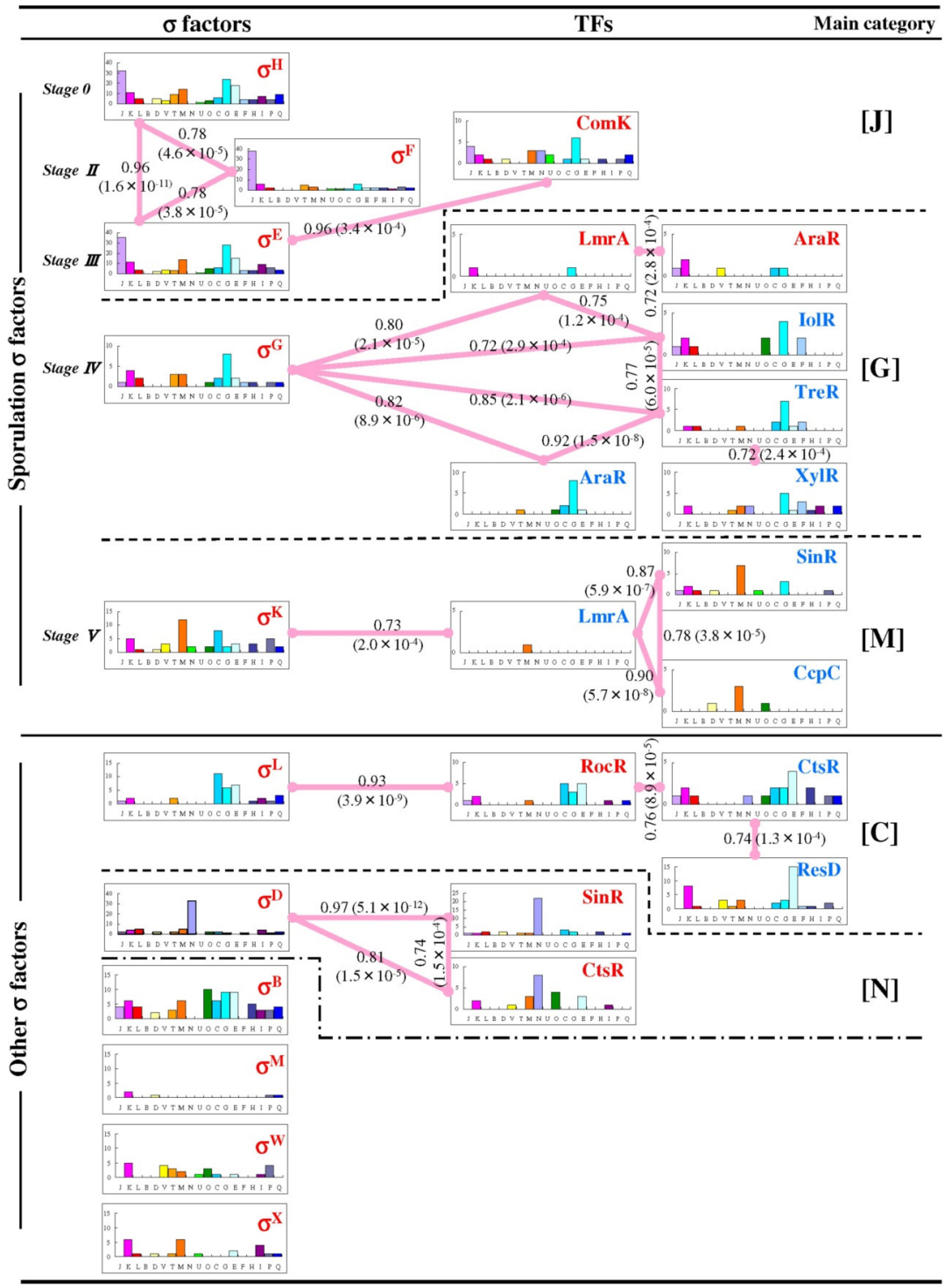


Figure 11. Clustering of σ-factor and TFs based on the functional similarity of genes composing the TUs. For

clustering of σ-factor and TFs, Pearson correlation coefficients among σ-factor and TFs were calculated using the frequencies of genes belonging to each of the COG categories. I regarded groups of σ-factors and TFs that have correlation coefficients above 0.70 (solid lines) as the clusters. Pearson correlation coefficients and P-values in the parentheses are represented on the solid lines. The main category represents the COG category into which the most genes regulated by each of the σ-factors and TFs in the clusters are classified. All data were identified under PWM with 1% threshold and FDR control. One-letter abbreviations show in Figure 10.

3.1.4.2.3 PCA for functions of genes composing TUs

In addition, the genes targeted by σ factors and TFs are classified into 36 categories based on functional classification of the *B. subtilis* protein-encoding genes [Sonenshein et al. 2001] to examine the role similarities among them based on *B. subtilis*-specific gene functions such as the endospore-formation process. I then show the projection of σ factors and TFs in the largest two principal components (Fig. 12A) and factor loadings of individual categories, indicating the contribution of the category frequencies to the two principal components based on the frequencies of the 36 categories (Fig. 12B). I observe a small cluster composed of σD, CtsR and SinR (a broken line circle in Fig. 12A), which is consistent with the result in Figure 11. Here, σD is the σ28-form subunit of RNA polymerase, and many σD-dependent genes are known to be necessary for flagellar synthesis and motility functions [Mirel and Chamberlin 1989]. In addition, CtsR controls the expression of heat-shock proteins that are required for stress tolerance and growth at high temperature [Derre et al. 1999], and play essential roles in competence development and motility [Msadek et al. 1998]; SinR also regulates the development of genetic competence and motility [Guillen et al. 1989]. Thus, the roles of these three TFs in cellular processes are associated with motility, and those are plotted in the same region of the cluster characterized by category [1.6, motility and chemotaxis] (Fig. 12AB). This result shows that roles of σ-factors and TFs can be estimated by the principal component analysis (PCA) based on comprehensive searches for functions of genes composing these TUs.

It can also be seen in another cluster composed of Fur, Zur, IolR, PurR, RocR, and GlnR (a broken line circle in Fig. 12A). Fur and Zur regulate the expression of ABC transporters and both TFs control iron and zinc uptake and homeostasis pathways in response to available metals [Gaballa and Helmann 1998; Ollinger et al. 2006]. IolR and PurR also control transport systems. IolR regulates genes encoding inositol transporters and inositol uptake [Yoshida et al. 1997], while PurR regulates purine transport, metabolism, and biosynthetic pathways [Ebbole and Zalkin 1989]. In this cluster, RocR and GlnR relate to controlling nitrogen sources: RocR controls arginine catabolism [Gardan et al. 1995] and the arginase pathway in which arginine is converted to glutamate [Belitsky and Sonenshein 1999], while GlnR regulates responses to nitrogen availability, such as nitrogen metabolism [Fisher 1999] and assimilation [Magasanik 1982].

Based on these previous studies, this result shows that I can cluster together homeostatic regulation σ-factors and TFs (Fig. 12A). Moreover, σ factors that regulate sporulation (σE, σF, σG,

σH, σK) tend to exist near the y-axis in the region of lower first-principal component (PC1) values with negative PC2 values, and TreR, SinR, and CcpC are also plotted near the σ factors (Fig. 12A). TreR regulates trehalose as the sole carbon and energy source of *B. subtilis* during spore outgrowth [Kennett and Sueoka 1971], while SinR controls regulatory genes involved in the early stages of sporulation [Cervin et al. 1998]. Thus, sporulation-related TFs tend to have lower PC1 values and negative PC2 values, which may be evidence that category [1.9; sporulation] and [1.4; germination] are plotted in the area (Fig. 12B). Therefore, CcpC is known to be a regulator of the tricarboxylic acid cycle genes [Hanson and Cox 1967], but may also have a function in regulating sporulation genes. These results demonstrate that the combined approach is very useful for specifying all of the TUs on the *B. subtilis* genome.

In the present study, I identified various sizes of TUs regulated by each σ-factor and TF and part of TUs overlapped by some genes. There are two types of overlap pattern between adjacent TUs. One type is that some short TUs compose parts of a long TU. In the structure, I detected gene clusters consisting of part of well-known operons (*yabPQ* regulated by σE and *divIC-yabR* regulated by σX in *yabMNOPQ-divIC-yabR* operon, *nasDEF* regulated by GlnR in *nasBCDEF* operon, *yjmEFGHIJ* regulated by σE in *yjmABCDEFGHIJ* operon, *spoVE-murG* regulated by σE in *murE-mraY-murD-spoVE-murG-murB-divIB-ylxWX-sbp* operon, *xynB* regulated by XylR in *ynaJ-xynB* operon and *yoxB-yoaA* regulated by σB in *yoxCB-yoaA* operon mentioned in Table 5). Part of them are known to be regulated by internal promoters and to constitute functional components [Gao et al. 2006], for instance, *yabPQ* regulated by σE that plays an important role in synthesis of the spore cortex and coat [Asai et al. 2001], and *divIC-yabR* regulated by σX which is essential for the initiation of vegetative septum formation [Huang and Helmann 1998; Levin and Losick 1994] in *yabMNOPQ-divIC-yabR* operon. Therefore, these gene clusters separated by internal promoters are likely to be functional units. The other type is that tail region of upstream TU overlaps with head region of downstream TU. These overlapping regions should belong to upstream or downstream TU, and therefore there is promoter for the downstream TU at edge of upstream in the region or there is terminator for the upstream TU at edge of downstream in the region.

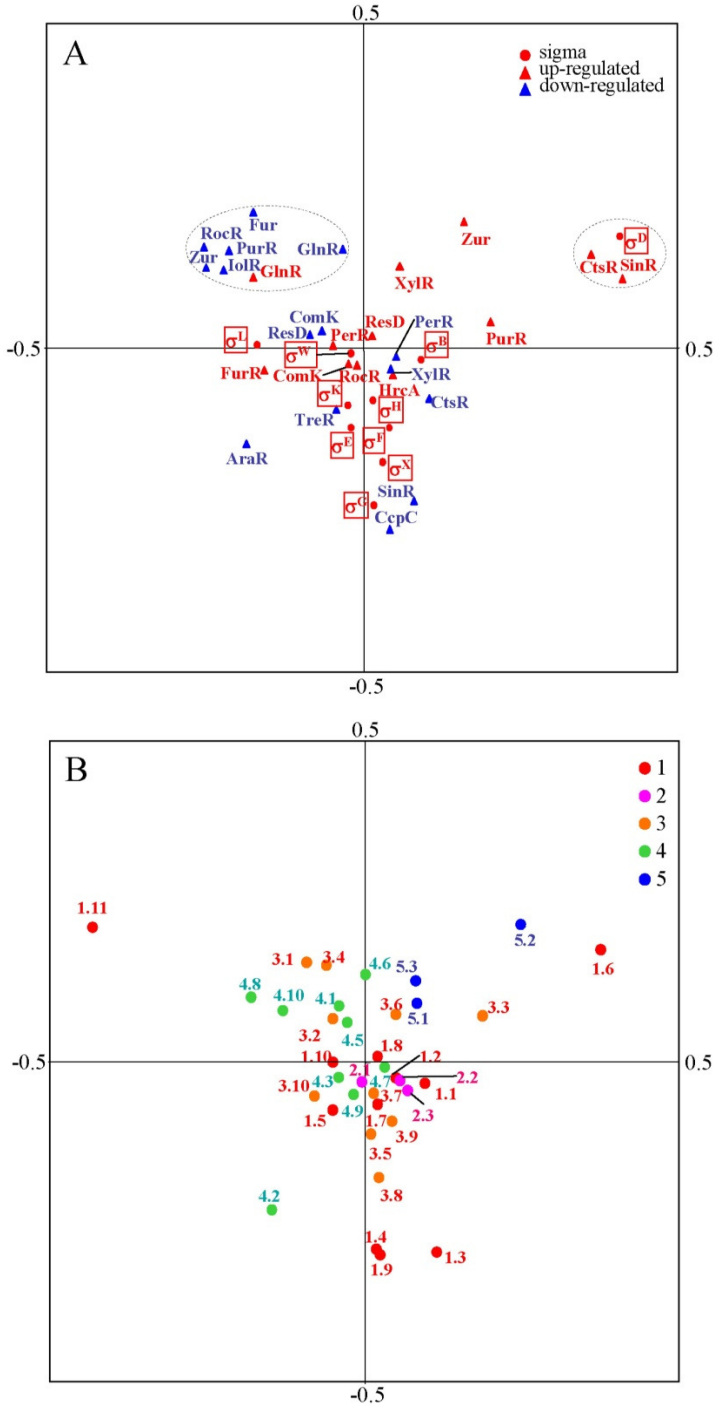


Figure 12. Score plots for PCA of σ-factors and TFs and functional categories. (A) Score plot based on frequencies of σ-factor and TF functions. (B) Score plot based on frequencies of functional classification of the *B. subtilis*

protein-encoding genes. Abbreviations for the functional categories: 1, cell envelope and cellular processes [1.1, adaptation to atypical conditions; 1.2, cell division; 1.3, cell wall; 1.4, germination; 1.5, membrane bioenergetics (electron transport chain and ATP synthase); 1.6, motility and chemotaxis; 1.7, protein secretion; 1.8, sensors (signal transduction); 1.9, sporulation; 1.10, transformation/competence; 1.11, transport/binding proteins and lipoproteins]; 2, cytochromes [2.1, cytoplasmic; 2.2, membrane-bound; 2.3, other cytochromes]; 3, information pathway [3.1, competence regulatory; 3.2, detoxification; 3.3, DNA packaging and segregation; 3.4, DNA replication; 3.5, DNA restriction/modification, repair and recombination; 3.6, protein folding; 3.7, protein modification; 3.8, protein synthesis; 3.9, RNA modification; 3.10, RNA synthesis]; 4, intermediary metabolism [4.1, antibiotic production; 4.2, carbohydrates and related molecules; 4.3, degradation; 4.4, general function prediction; 4.5, metabolism of amino acids and related molecules; 4.6, metabolism of coenzymes and prosthetic groups; 4.7, metabolism of lipids; 4.8, metabolism of nucleotides and nucleic acids ; 4.9, metabolism of phosphate; 4.10, metabolism of sulfur]; 5, other functions [5.1, antibiotic production; 5.2, phage-related functions; 5.3, transposon and insertion elements].

3.2 Identification of *B. subtilis* operon using comparative genomic approach

3.2.1 Introduction

There has been a variety of clues for operon prediction, and many studies have used phylogenetic information as a promising means of predicting operon structure [Bergman et al. 2007; Edwards et al. 2005]. In prokaryote genomes, gene orders and orientations are conserved across a variety of genomes and these adjacent genes that are found in the same order in multiple genomes are likely to be co-transcribed [Ermolaeva et al. 2001]. The conservation of gene grouping is likely to occur by sharing the gene clusters among evolutionarily related organisms and by horizontal transfer of the genes, and thus the formation of the operons remains widely the same and gene orders in the operons are conserved across species.

In many cases of phylogenetic analyses including this study, the basic local alignment search tool (BLAST) has been frequently used for searching the most similar sequences from the DNA and protein sequence database [Altschul et al. 1997]. BLAST is a heuristic program that is written to compare protein or DNA queries with protein or DNA databases. This approach to finding the sequences in the databases most similar to a query is performed to calculate the pairwise score of the alignment between the query and every sequence in the databases [Altschul et al. 1990].

In this study, I used the *blastp* program [Altschul et al. 1997], which compares protein queries to protein databases and detected orthologous genes in 54 Gram-positive bacteria except *B. subtilis*. Using these data and my TU data, I determined the appropriate boundaries of TUs and predicted genome-wide operon structure in the *B. subtilis* genome.

3.2.2 Data

A set of 55 Gram-positive bacterium genomes were obtained from NCBI [NCBI FTP].

3.2.3 Methods

3.2.3.1 Identification of successive homologous gene clusters

Each of the gene sets of the bacterium genomes was compared to that of the *B. subtilis* genome and reversely the gene set of *B. subtilis* genome was compared to those of the 54 query genomes by BLASTP with the BLOSUM62 matrix and an E-value cut-off of 10^{-5}, and the genes which have the highest sequence similarity in the both results as the orthologous genes were considered (Table 6). A gene with no homologous in *B. subtilis* genome was considered absent in this genome.

3.2.3.2 Identification of *B. subtilis* operons

I considered genes that were the orthologs and were in series in the same direction with no intervening gene on a pair of the genomes as the successive homologous gene clusters. When plural operon candidates overlapped in *B. subtilis* genome, I compared the orthologous clusters with each of the operon candidates and counted numbers of preserved species. When the upstream candidate had higher score than the other, I regarded the upstream one as an operon. When the downstream candidate had higher score than the other, I supposed that there was internal promoter which regulated the downstream candidate and the promoter divided the upstream one and regarded the forward part of the upstream one and the downstream one as operons. Finally, I selected the suitable operons by these procedures and identified *B. subtilis* operon on the genome.

Table 6. Query genomes for comparative genomic analysis

No.	Name	BLOSUM 62 No.	Protein	No.	Name	BLOSUM 62 No.	Protein
1	*Bacillus subtilis*	NC_000964	4105	29	*Lactobacillus_brevis_ATCC_367*	NC_008497	2185
2	*Bacillus halodurans*	NC_002570	4066	30	*Lactobacillus_casei_ATCC_334*	NC_008526	2751
3	*Bacillus anthracis Ames*	NC_003997	5311	31	*Lactobacillus_gasseri_ATCC_33323*	NC_008530	1755
4	*Bacillus cereus ATCC 14579*	NC_004722	5234	32	*Pediococcus_pentosaceus_ATCC_25745*	NC_008525	1755
5	*Bacillus_thuringiensis_konkukian*	NC_005957	5117	33	*Enterococcus_faecalis_V583*	NC_004668	3113
6	*Bacillus licheniformis ATCC 14580*	NC_006270	4152	34	*Oenococcus_oeni_PSU-1*	NC_008528	1691
7	*Bacillus_clausii_KSM-K1*	NC_006582	4096	35	*Leuconostoc_mesenteroides_ATCC_8293*	NC_008531	1970
8	*Oceanobacillus iheyensis*	NC_004193	3500	36	*Clostridium acetobutylicum*	NC_003030	3672
9	*Geobacillus_kaustophilus_HTA426*	NC_006510	3498	37	*Clostridium perfringens*	NC_003366	2660
10	*Staphylococcus aureus N315*	NC_002745	2588	38	*Clostridium tetani E88*	NC_004557	2373
11	*Staphylococcus epidermidis ATCC 1222*	NC_004461	2419	39	*Carboxydothermus_hydrogenoformans_Z-290.*	NC_007503	2620
12	*Staphylococcus haemolyticus*	NC_007168	2676	40	*Desulfitobacterium_hafniense_Y51*	NC_007907	5060
13	*Staphylococcus saprophyticus*	NC_007350	2446	41	*Syntrophomonas_wolfei_Goettingen*	NC_008346	2504
14	*Listeria monocytogenes*	NC_003210	2846	42	*Thermoanaerobacter tengcongensis*	NC_003869	2588
15	*Listeria innocua*	NC_003212	2968	43	*Moorella thermoacetica*	NC_007644	2465
16	*Listeria welshimeri SLCC5334*	NC_008555	2774	44	*Mycoplasma genitalium*	NC_000908	477
17	*Lactococcus lactis*	NC_002662	2321	45	*Mycoplasma pneumoniae*	NC_000912	689
18	*Streptococcus_pyogenes_MGAS10270*	NC_008022	1987	46	*Mycoplasma pulmonis*	NC_002771	782
19	*Streptococcus pneumoniae TIGR4*	NC_003028	2105	47	*Mycoplasma penetrans*	NC_004432	1037
20	*Streptococcus agalactiae 2603*	NC_004116	2124	48	*Mycoplasma gallisepticum*	NC_004829	726
21	*Streptococcus mutans*	NC_004350	1960	49	*Mycoplasma mycoides*	NC_005364	1016
22	*Streptococcus thermophilus CNRZ1066*	NC_006449	1915	50	*Mycoplasma mobile*	NC_006908	633
23	*Lactobacillus plantarum*	NC_004567	3009	51	*Mycoplasma hyopneumoniae 232*	NC_006360	691
24	*Lactobacillus_johnsonii_NCC_533*	NC_005362	1821	52	*Mycoplasma_synoviae_53*	NC_007294	672
25	*Lactobacillus_acidophilus_NCFM*	NC_006814	1864	53	*Mycoplasma_capricolum_ATCC_27343*	NC_007633	812
26	*Lactobacillus_sakei_23K*	NC_007576	1879	54	*Ureaplasma urealyticum*	NC_002162	614
27	*Lactobacillus_salivarius_UCC118*	NC_007929	1717	55	*Mesoplasma_florum_L1*	NC_006055	682
28	*Lactobacillus_delbrueckii_bulgaricus*	NC_008054	1562				

Protein refers to total number of proteins in the bacterium.

3.2.4 Results and Discussion

Conservation of gene order is one of the principal types of context information. On the *B. subtils* genome, there are 3,123 consecutive gene pairs in the same directions and some gene pairs are operon pairs (OPs) in polycistronic operons. Comparing interspecies conservation of *B. subtilis* known OPs (KOPs) to that of the other consecutive gene pairs in the same directions among Gram-positive bacteria, 59.8% orthologous gene pairs of KOPs existed in other Gram-positive bacteria but 60.3 % of consecutive gene pairs except KOPs are only in the *B. subtilis* genome (Fig. 13(B)). The average number of conserved species in KOPs among Gram-positive bacteria was approximately twice as many as the number in the consecutive gene pairs in the same directions except KOPs (Fig. 13(A)). Therefore, the conservation of gene order is a key to finding relationships between neighbor genes, and this taxonomical information is likely to support determination of operon organization in a target genome.

I used comparative genomic analysis for determining appropriate boundaries of TUs identified by the integrated analysis and determined genes belonging to these overlapping regions using complete genome sequence data of other Gram-positive bacteria (Table 6). Based on the analysis, I regarded the non-overlapping TUs and these determined TUs as *B. subtilis* operons and showed the operon map (Appendix A). Applying this approach, I narrowed the candidate genes composing operons down to 999 genes from 2,183 genes composing TUs (Table 5, Appendix B), and the difference between operons predicted in the present study and known operons in the *B. subtilis* genome indicates that most of the predicted operons are consistent with those reported (Fig. 14). This additional application identified 528 *B. subtilis* operons (Fig. 15A2) and I then searched functional constitutions of operons predicted in this study and experimentally derived operons by DBTBS functional classification of their gene products (Fig. 15). The results show that most operons categorized by the four major DBTBS classifications tend to be composed of mono-functional gene clusters. When these operons are categorized by the DBTBS classifications, 86% are composed of mono-functional genes and 12 % have two functions (Fig. 15A) and I then search detailed functional formations for the mono-functional operons (Appendix B). In the mono-functional operons categorized by the detailed functional classifications, a few operons are di-functional in structure but 94% or more in each of the classifications are made of the same functional genes. This result supports that operon structure is constructed by gathering the same functional genes in taxonomical evolution and suggests that operons work as functional units.

There are a few multi-functional operons, but the operons having the largest functions are tri-functional operons (Fig. 15A2).

In the tri-functional operons, I can detect some operons that are of different structures from experimentally derived operons; *trePARyfkO, acoABCLRyfjU, etfABysiBA and yolKJIsunTA*. The detected operon, *trePARyfkO*, is composed of the *tre* operon [Schock and Dahl 1996], and the *yfkO* gene. In the *tre* operon, it is deduced that TreP acts as a specific enzyme in trehalose transport and phosphorylation [Klein et al. 1995], *treA* encodes phospho-α–(1,1)-glucosidase which hydrolyzes trehalise-6-phosphate in the growth phase [Gotsche and Dahl 1995], and treR encodes a specific Tre repressor involved in regulation of the expression of the *tre* operon [Schock and Dahl 1996]. Thus the *tre* operon regulates the ability to utilize trehalose. The function of *yfkO* composing the detected operon is not well understood, but I can assume by sequential analysis that *yfkO* encodes nitroreductase which is an enzyme lowering the activation energy for a reduction reaction of nitro groups, so that YfkO may act on trehalose metabolism via catalyzing nitro substrates.

In the *acoABCLRyfjU* operon, the four gene string, *acoABCL*, is known as the *aco* operon that encodes the E1α, E1β, E2, E3 subunits of the acetoin dehydrogenase multi-enzyme complex. Acetoin is a carbon source for growth and acts as an external carbon storage material in exponential growth [Huang et al. 1999]. Transcription of the *aco* operon is regulated by carbon catabolite repression *via* CcpA which is a negative regulator of carbon utilization genes and is a positive effector of genes in biosynthesis and secretion of metabolites [Henkin 1996]. The CcpA exerts negative regulation binding to the catabolite-responsive element of *acoA* [Miwa et al. 2000] and directly regulates transcription of *acoR* located immediately downstream from the *aco* operon [Ali et al. 2001]. AcoR is a regulator of the expression of the *aco* operon in the presence of σL and induces acetoin expression [Ali et al. 2001]. Furthermore, *yfjU* at the end of the *acoABCLRyfjU* operon encoding small, acid-soluble protein unique to spores of *B. subtilis* is not found for the gene expression in the vegetative phase and appears to be expressed in the growth phase as the expression of the *aco* operon. Then previous structural analysis showed that there is no obvious transcription terminator between the immediately upstream gene, *acoR* and *yfjU* and there is a transcription terminator candidate sequence with symmetry following the gene, so they suggested that these genes may be co-transcribed [Cabrere-Hernandez et al. 1999]. Therefore, the *acoABCLRyfjU* operon is constructed by different functional genes (Fig. 15B), but these genes

may be assembled for regulation of the catabolic pathway in the growth phase and the two genes, *acoRsspH*, of the detected operon may work as an internal operon [Gao et al. 2006], following the previous study [Huang et al. 1999]. My result supports northern blotting data in the BSORF database showing *acoABCLRyfjU* and *acoRsspH* transcripts [BSORF].

Recent study by BLASTP sequence similarity search presumed the four genes, *etfABysiBA*, that encoded α- and β-subunits of electron transfer flavoproteins involved in fatty acid β-oxidation, a enzyme belonging to the enoyl coenzyme A hydratase / isomerase family, and a TF to negatively regulate the expression of *lcfA* operon genes respectively [Barabesi et al. 2007] and then each product of these gene acts in fatty acid degradation pathway [Matuoka et al. 2007]. Furthermore, these genes are known to compose the *lcfA* operon which consists of five genes, *etfABysiBAlcfA* [Wipat et al. 1996], and northern blotting analysis of *etfABysiBAlcfA* transcripts detected several sizes of transcripts, *lcfA*, *ysiA*, *ysiABetfBA* and *etfABysiBAlcfA* [Matuoka et al. 2007]. Therefore, my result can confirm that there is an internal operon, *etfABysiBA*, in the *lcfA* operon. As for *sunATyolIJK*, northern bolt analyses using *sunA*- and *sunT*-specific probes showed that the *sunA* gene is transcripted as a monocistronic mRNA and the *sunT*-probe hybridized to an approximately 4.3-kb transcript corresponding to the polycistronic transcription of *sunTyolIJK* [Serizawa et al. 2005]. These data suggest that these two transcripts are independently regulated by YvrH as a positive regulator, but YvrH binding sites in the upstream of *sunA* and *sunT* have not been found [Serizawa et al. 2005], the negative regulator of these genes, Rok binds to the promoter region of *sunA* only [Albano et al. 2005], and these transcripts have functional relation of sublancin 168. Therefore, the five gene cluster, *sunATyolIJK*, may comprise an operon including the internal operon, *sunTyolIJK*.

(A)

	Average number of conserved species
KOP	**13.9**
Gene pairs except KOP	**7.3**

(B)

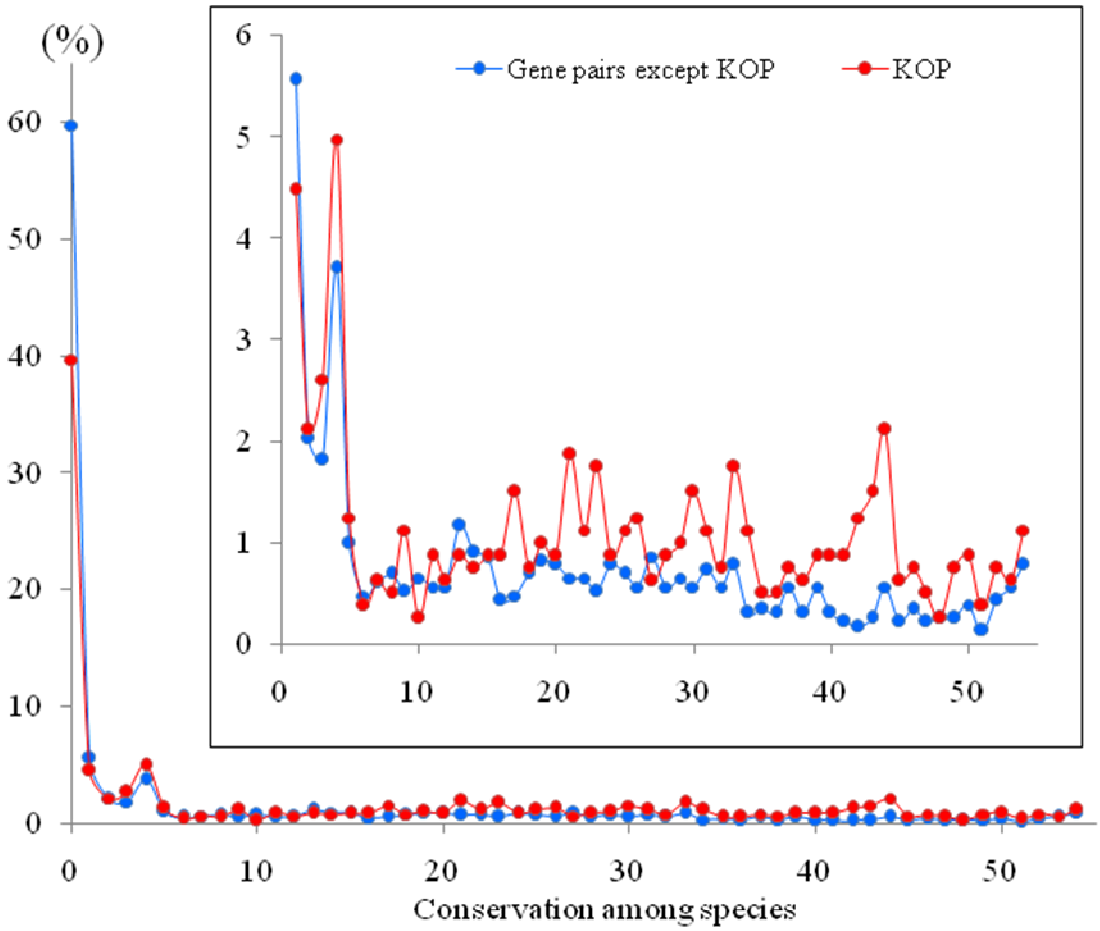


Figure 13. Conservation of consecutive gene pairs in the same directions among Gram-positive bacteria. (A) Average number of conserved species of orthologous gene pairs. (B) Conservation ratio of *B. subtilis* gene pairs between *B. subtilis* and 54 Gram-positive bacteria.

KOP refers to a known operon pair.

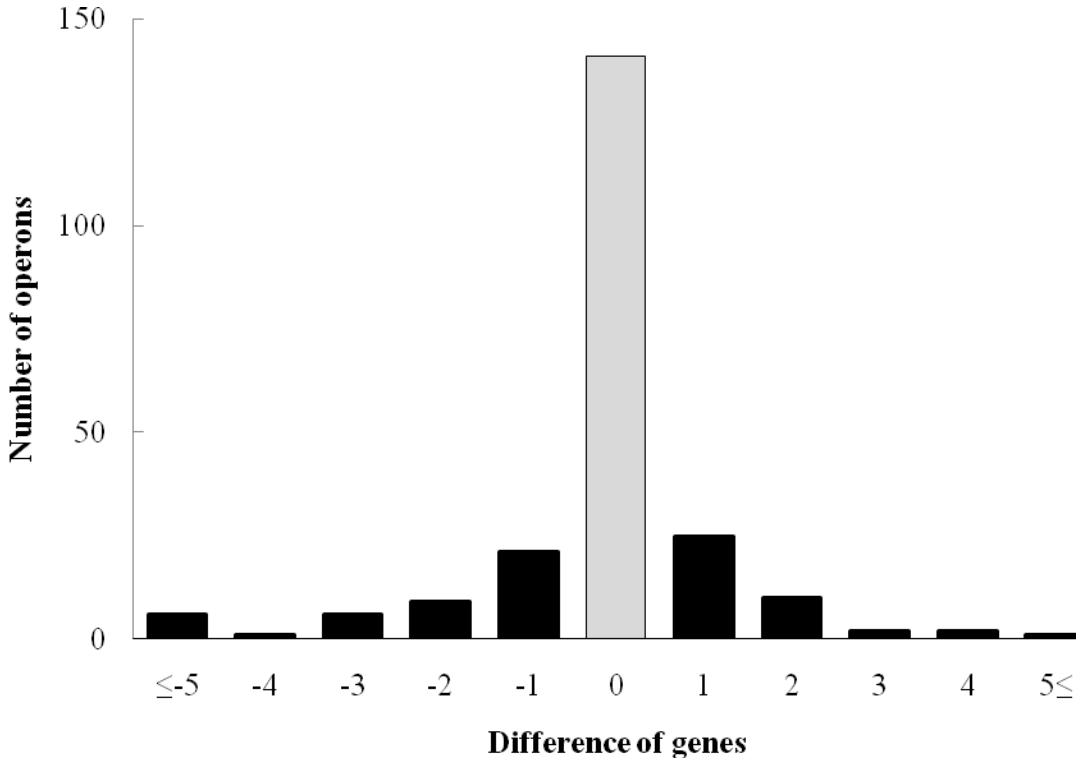


Figure 14. Comparing operons predicted in the integrated approach with known *B. subtilis* operons, which start with the same promoters as the operons. 62.9% of the known operons matched the predicted operons (i.e., 141 predicted operons in the 224 known *B. subtilis* operons); when I included operons lacking or gaining one gene of known operons, this figure rose to 83.5%. The x-axis indicates difference in gene number between predicted operons and known *B. subtilis* operons, and the y-axis indicates the number of predicted operons.

(A1)

1	2	3	4	
164	16	19	2	1
	147	18	5	2
		98	4	3
			45	4

(A2)

M	D	T	total
454	64	10	528

(B)

Detected operons	Functional categories
trePARyfkO	0231
acoABCLRyfjU	222231
kdgTAKR	1223
etfABysiBA	1123
ytrFEDCBA	102203
rbsRKDACBywsB	3211110
gntRKPZ	3212
ycsFGIJOK	400130
ycbABCDEFGHJ	002202304
yolKJIsunTA	32304

(C1)

1	2	3	4	5	6	7	8	9	10	
22	0	0	0	0	1	0	2	0	0	1
	13	0	0	0	0	0	1	0	0	2
		3	0	0	0	0	0	0	1	3
			19	0	0	0	1	0	0	4
				14	0	0	0	0	0	5
					2	0	0	0	0	6
						5	1	0	0	7
							63	2	0	8
								6	0	9
									8	10

(C2)

1	2	3	4	5	6	7	
56	3	0	1	1	0	0	1
	35	2	0	1	0	0	2
		12	0	0	0	0	3
			12	0	1	0	4
				16	0	0	5
					3	0	5
						3	7

(C3)

1	2	3	4	5	6	7	8	9	
1	0	0	0	2	0	0	0	0	1
	10	0	0	0	0	0	0	0	2
		3	1	0	0	0	0	0	3
			4	0	0	0	0	0	4
				51	0	1	0	0	5
					1	0	0	0	6
						14	0	0	7
							7	0	8
								3	9

(C4)

1	2	3	4	5	6	
19	0	0	0	0	1	1
	12	0	0	0	0	2
		1	0	0	0	3
			8	0	0	4
				0	0	5
					4	6

Figure 15. Categorization of detected and known operons by DBTBS gene functional classification. (A1) Number of mono- or di-functional operons by DBTBS gene function. (A2) Total number of each functional operon. M, D and T refer to mono-, di- and tri-functional operon respectively. (B) Tri-functional operons classified by DBTBS gene function. (C) Detailed categorization of single functional operons by DBTBS gene function. (C1) Categorization in Cell envelope and cellular processes. (C2) Categorization in intermediary metabolism. (C3) Categorization in information pathways. (C4) Categorization in other functions. Bold faces indicate abbreviations for functional categories.
Abbreviations: 0, unknown; 1 Cell envelope and cellular processes [1.1 Cell wall, , 1.2 Transport/binding proteins and lipoproteins, 1.3 Sensors (signal transduction), 1.4 Membrane bioenergetics (electron transport chain and ATP synthase), 1.5 Mobility and chemotaxis, 1.6 Protein secretion, 1.7 Cell division, 1.8 Sporulation, 1.9 Germination, 1.10 Transformation/competence]; 2 Intermediary metabolism [2.1 Metabolism of carbohydrates and related molecules, 2.2 Metabolism of amino acids and related molecules, 2.3 Metabolism of nucleotides and nucleic acids, 2.4 Metabolism of lipids, 2.5 Metabolism of coenzymes and prosthetic groups, 2.6 Metabolism of phosphate, 2.7 Metabolism of sulfur]; 3

Information pathways [3.1 DNA replication, 3.2 DNA restriction/modification and repair, 3.3 DNA recombination, 3.4 DNA packaging and segregation, 3.5 RNA synthesis, 3.6 RNA modification, 3.7 Protein synthesis, 3.8 Protein modification, 3.9 Protein folding]; 4 Other functions [4.1 Adaptation to atypical conditions, 4.2 Detoxification, 4.3 Antibiotic production, 4.4 Phage-related functions, 4.5 Transposon and IS, 4.6 Miscellaneous].

4 Operon Prediction by Support Vector Machine

4.1 Introduction

SVM is a computer based algorithm that belongs to a family of generalized linear classifiers and is closely related to neural networks. This is a set of related supervised learning methods used for classification and regression. When there exist many decision boundaries separating the samples into two classes, SVMs find the one that achieves the maximum margin between them using statistical learning theory.

Recently, this supervised learning technique has been utilized to perform in multiple area of biological analysis. For instance, SVMs examined the gene expression profile from ovarian cancer tissues, normal ovarian tissues and other normal tissues [Furey et al. 2000]. As a result of the computational analysis, they were able to classify tissue and cell types based on data sets, and discovered and confirmed to be cancer cells some that were mistakenly labeled. Brown et al. [2000] demonstrated that SVMs could accurately classify genes into some functional categories based on microarray expression profiles and identify the functions of unannotated yeast genes. Another analysis showed exact translation initiation sites in order to identify protein sequences from nucleotide sequences by SVM with an appropriate kernel function [Zien et al. 2000]. In this way, SVMs have been recently used in bioinformatics as classifiers for biological objects: proteins, DNA sequences, microarray expression profiles and mass spectra [Noble 2004].

In this study, I employed SVM as a classification tool for *B. subtilis* operon prediction using features based on a large amount of data: (i) intergenic distance data, (ii) expression data, (iii) comparative genomic data, and (iv) *B. subtilis* specific functional data. I also examine discrimination capacity of boundaries between intra operons and inter operons by combination of the four features of (i) - (iv).

4.2 Data

4.2.1 Microarray Data

For Pearson correlation coefficients, I used 98 time-series gene expression data in *B. subtilis* 168 grown in 13 different time-series growth conditions: anaerobic growth; competent medium; cold-shock experiments; DSM medium; DGG medium; glucose-limited medium; heat-shock

experiment; LB medium; minimum-glucose medium; sodium-shock conditions; phosphate-starvation medium; and SOS stress experiments. I also used 394 gene expression data of *B. subtilis* σ-factor and TF deletion mutants which were *sigB, D, E, F, G, H, K, L, M, W, X, abrB, acoR, ahrC, alsR, arsR, azlB, codY, comA, cspB, deoR, exuR, fnr, fruR, glcT, gntR, gutR, hpr, hutP, kdgR, lacR, lexA, mntR, msmR, mtrB, pksA, rbsR, spo0J, ybbH, ysiA, yvaN, yvrH, yybA, yybE, yydK, araR, ccpC, citT, comK, ctsR, fur, glnR, hrcA, iolR, lmrA, perR, purR, resD, rocR, sinR, treR, xylR, zur, ccpA, degU, gerE, phoP, spo0A, spoIIID* and *spoVT* deletion mutant. Furthermore, I used 312 public gene expression data from Gene Expression Omnibus (GEO) in NCBI which were GPL188, GPL1353, GPL2760, GPL3645, GPL3655, GPL3665, GPL3668, GPL3671, GPL3949, GPL3950, GPL4694 and GPL6031 in GEO accession numbers [GEO].

4.2.2 Genomic Data

The genome sequence of *B. subtilis* and the annotation files were obtained from the NCBI ftp site [NCBI FTP]. The leading and lagging strands were scanned based on whether they were transcribed in the same directions or in different directions. The 4225 genes in the linear chromosome were binned into 3000 same-strand pairs and 1224 opposite-strand pairs.

4.2.3 Comparative Genomic Data

Comparative genomic data was obtained from OperonDB [Ermolaeva et al. 2001]. I used two types of data sets; one data set contained probability that a gene pair is located in the same operon and the other data set was the number of other genomes that have the same pair of genes located in the same directon.

4.2.4 Functional Data

Functional data was obtained from DBTBS. I used 32 functional classifications for sorting 4,224 gene pairs into three groups: same functional pairs, different functional pairs and unknown pairs.

4.3 Methods

4.3.1 Microarray Data Normalization

Details of all protocols are described in chapter **3.2.2**.

4.3.2 Calculation of Pearson correlation coefficients of neighboring genes

Pearson correlation coefficients of neighboring genes were calculated using a 4225 (genes) × 804 (samples) matrix.

4.3.3 Intergenic Distance Calculation

Intergenic distance in base pairs between the genes in every gene pair ($gene_I$ - $gene_{II}$) was calculated as $distance_{I-II} = gene_{II}\_start - gene_I\_end - 1$. Negative intergenic distance implies an overlap between the translated regions of the two genes.

4.3.4 Supervised Classification

4.3.4.1 Training Set: Positive and Negative Classes

The training set consists of 805 known operon pairs (KOPs) based on northern bolts, primer extension analyses and nuclease protection assays from DBTBS and BSORF databases. The set of gene pairs that comprise the negative class was created using the genes composing known monocistronic operons and their immediately upstream on downstream genes. These gene pairs are referred to as known boundary pairs (KBPs) and a set of 272 KBPs constitutes the negative class.

4.3.4.2 Model Training and Selection

Binary SVM classifiers were trained for operon prediction using five different features: Pearson correlation coefficient between neighboring genes having the same direction, intergenic distance, two comparative genomic data, and gene function data. Intergenic distance is measured in base pairs, whereas the Pearson correlation coefficient is bound between -1 and 1. Due to the large difference in the range of these features, scaling was performed by discretizing the intergenic distances into six bins corresponding to $d \leq 0$, $0 < d \leq 20$, $20 < d \leq 50$, $50 < d \leq 100$, $100 < d \leq 300$ and $d > 300$ bp. Therefore, I used vectors in ith gene pair with 10 components called features:

$$\mathbf{x}_i = (r^i, d_1^i, d_2^i, d_3^i, d_4^i, d_5^i, d_6^i, P_{operon}^i, P_{directon}^i, f^i).$$

Here r is the Pearson correlation coefficient of neighboring genes, d_n (n = 1, 2, …, 6) is binary data made of six bins based on the intergenic distances, P_{operon} is the probability that a gene pair is

located in the same operon based on OperonDB, $P_{directon}$ is the probability that a gene pair is located in the same directon based on OperonDB and f is binary data based on the DBTBS gene function. The discrimination rule established during training can result in over-fitting whereby the classifier cannot accurately discriminate test/unseen data. Leave-one-out cross-validation was therefore performed to estimate the performance of the model in classifying an independent dataset that was not used for training.

4.3.4.3 Support Vector Machines

The SVM algorithm is the kernel learning algorithm [Vapnik 1998]. It performs robust pattern recognition using well-established concepts in optimization theory. In the 10-dimensional feature space, I limited two-class classification problems and identified the two classes with the class (+) for OPs and (–) for BPs. A training set of a number of patterns (S):

$$S = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}, y_k \in \{-1,+1\},$$

is given. The training patterns are used to build a decision function $f(x)$, that is a scalar function of an input pattern x. New patterns are classified according to the sign of the decision function:

$f(x) > 0 \Rightarrow x \in$ class (+)

$f(x) < 0 \Rightarrow x \in$ class (–)

$f(x) = 0$, decision boundary.

Decision functions are simple weighted sums of the training patterns plus a bias. In this notation:

$$f(x) = (w \cdot x) + b ,$$

where w is the weight vector and b is a bias value and both are determined by running a learning procedure on a training set S, and $(w \cdot x)$ indicates an inner product of w by x.

In this study, I used SVM^{light}, an implementation of SVM in C used for model training and evaluation [Joachims 1999]. Radial basis function (RBF) on non-linear mapping, was used for classification. The RBF transforms the data using the nonlinear function:

$$K(x_1, x_2) = \exp(-\gamma \| x_1 - x_2 \|^2) ,$$

where γ determines RBF width. For the radial kernel function, the parameters γ (–g) and the cost function (–c) were selected using the leave-one-out model selection (looms) procedure [Lee and Lin 2000]; I used 128.0 as the constant for soft margin, and 1.024 as the constant for RBF. The algorithm calculates the leave-one-out error rates for a range of parameters and outputs the one with the minimum error rate.

4.3.4.4 Leave-one-out Approach

Leave-one-out cross-validation is an iterative approach where each gene pair in the training set of '*n*' gene pairs is left out in one iteration. The model is trained with (*n - 1*) gene pairs and used to classify the nth gene pair. In each iteration, the true class of the pair (whether it is a KOP or NBP) is compared with the predicted class. The performance of the model is then evaluated using different metrics.

4.4 Results and Discussion

4.4.1 The features for SVM

The intergenic distance between a pair of adjacent genes is a powerful feature for operon prediction [Salgado et al., 2000; Moreno-Hagelsieb and Collado-Vides, 2002]. I used 805 KOP and 272 KBP, and showed distributions of KOP and KBP with respect to intergenic distance (Fig. 16). 28.2 % of KOPs have an intergenic distance less than 0 bp; i.e., the end of one gene overlaps the start of the next, and 88.9 % of KOPs have a distance of less than 100 bp. In the range, KOPs have two peaks around intergenic distances of 0 and 15 bp. In contrast, 1.1 % of KBPs overlap between the neighboring genes. Genes belonging to the same operon tend to exhibit small intergenic distance having two sharp peaks. On the other hand, intergenic distance distribution of genes that do not belong to the same operon is represented by a gentle curve having the maximum between 100 and 200 bp (Fig. 16). Observing the intergenic distance distribution of all pairs of *B. subtilis* genes, it can be said that this distribution is probably represented by the combined shape.

In this research, I used original features: the Pearson correlation coefficient and identity of function by DBTBS functional classification between adjacent genes with the same direction. The average value of the correlation coefficients of KOPs was about twice as much as that of KBPs (Table 7). Moreover, the number of gene pairs with the same function was about twice as much as the number with different function in KOPs, whereas 85 % of KBPs (in case of function known genes) were of different functions (Table 7). The feature of correlation is typically used in many operon predictions [Charaniya et al. 2007; de Hoon et al. 2004], and I could detect the bias of correlation between KOPs and KBPs in this research using huge microarray data and therefore this feature is useful for operon-boundary discrimination by SVM. KOPs were likely to be composed of same function pairs based on DBTBS functional classification; the same functional

pairs were about twice as much as the different functional pairs in KOPs, and KBPs had a strong tendency to be of different functional pairs. Some studies used functional similarity for operon prediction [Price 2005], but I used more detailed functional classification (32 classifications) than those used in other studies, so this functional classification could be a useful feature.

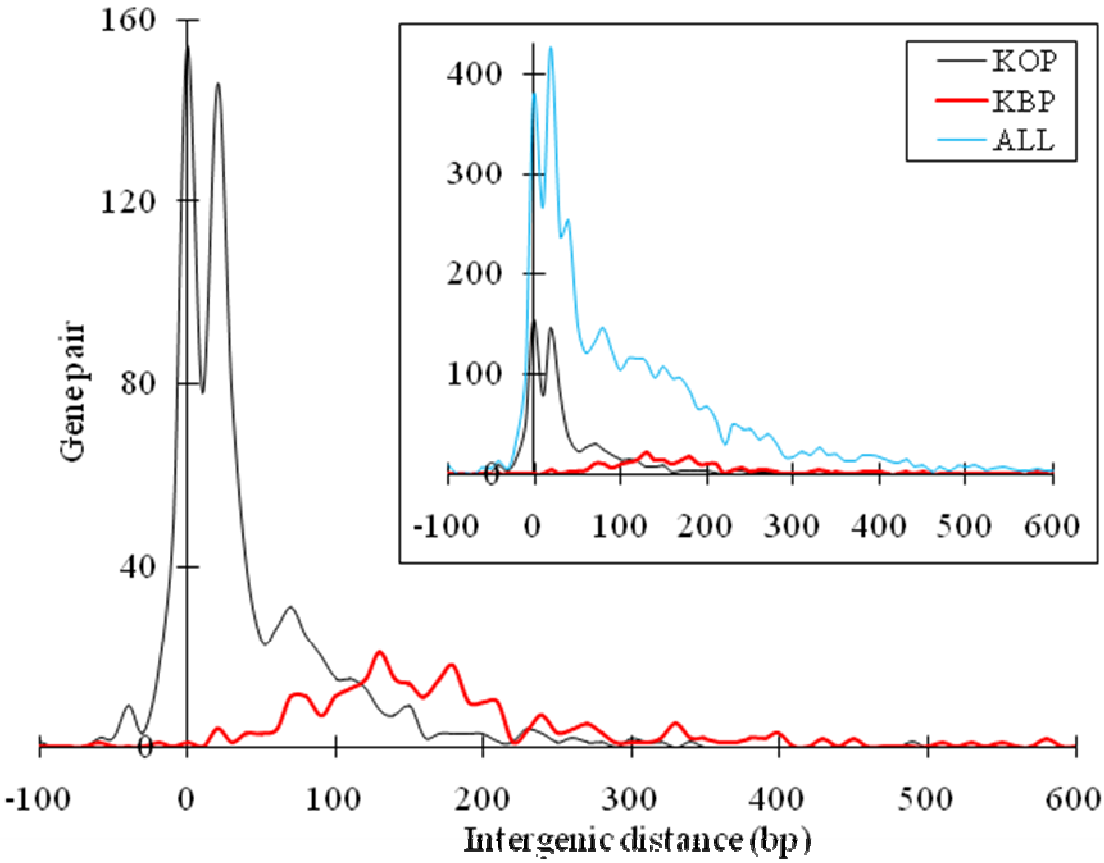


Figure 16. Frequency intergenic distance distributions of KOP (known operon pairs), KBP (known boundary pairs) and ALL (all *B. subtilis* gene pairs).

Table 7. Average value of Pearson correlation coefficients and DBTBS functional composition of KOP and KBP.

	Correlation	Function	
		same	different
KOP	0.66	334	170
KBP	0.32	11	62

4.4.2 Kernel functions for SVM

For the SVM classifier, I applied two kernel functions: linear kernel and radial basic function (RBF) kernel. In the linear classification by SVM, I perfectly detected KOPs in the range of intergenic distance below 100 bp and KBPs in the range of them with 100 bp or more, but I could not detect KOPs and KBPs in the other range at all (Fig.17). The performance of the linear classifier was strongly affected by feature of the intergenic distances. I made 6 bins on the basis of intergenic distance and used the features for the classifier, and then 88.9 % of KOPs belong to the four bins of intergenic distances with 100 bp or less and 80.5 % of KBPs reversely belong to the two bins of them with 100 bp or more. Compared with the other features, the distribution of KOPs and KBPs was biased across 100 bp on the feature of intergenic distance, and thus this feature had strong effect for the performance of classifier (Table 8).

Next, I discriminated OPs from BPs using non-linear SVM with RBF kernel. This kernel function played the role of the dot product in feature spaces. I separated 3,000 gene pairs into OP and BP on 10 dimensional feature hyperplane, and I could detect perfectly KOPs in the range of intergenic distance below 50 bp, 98.1 % of KOPs in the four bins of intergenic distances with 100 bp or less, and 50.1 % of KOPs in the two bins of these with 100 bp or more. Meanwhile, I detected 58.6 % of KBPs in the four bins of intergenic distances with 100 bp or less, and 93.9 % of KBPs in the two bins of these with 100 bp or more. Using only intergenic distance as the feature for SVM, the identification of KOPs and KBPs by the linear SVM and the RBF kernel SVM was similar, but the detection of KOPs and KBPs by non-linear SVM were considerably improved by RBF kernel function when more features were applied as input vectors. In the range of non-detecting KOPs and KBPs by linear classification, the non-linear SVM classifier especially identified KOPs and KBPs on account of effect of the features except intergenic distance (Table 8), and reversely the detection ratio of KOPs decreased by 1.95 % and the ratio of KBPs decreased at 6.07 % in the range of full-detecting KOPs and KBPs by linear classification. With allowance for soft margin in the RBF kernel SVM, I slightly failed the detections of known gene pairs which were identified by linear SVM [Tarca et al. 2007]. This may be because SVM is sensitive to noise, so the non-linear classifier was affected by noise in accordance with increasing features, but the detection ratio of total KOPs and KBPs certainly improved using non-linear classifier compared to liner classifier and thus this method could be more reliable than linear SVM.

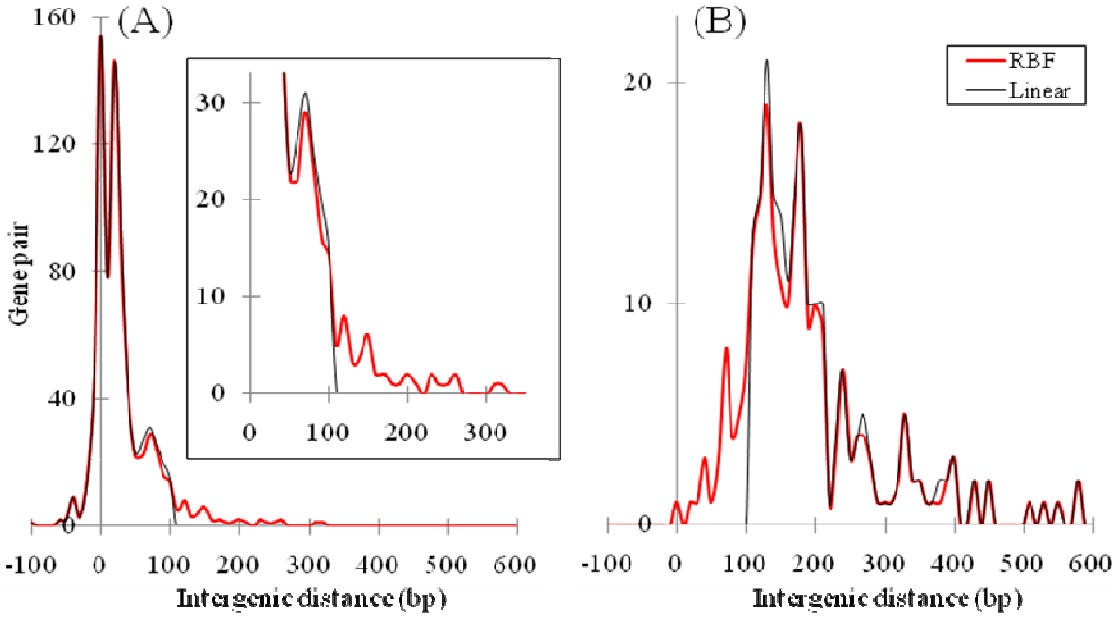


Figure 17. Comparison between linear classification and non-linear classification using radial basic function. (A) Distribution of identified KOPs. (B) Distribution of identified KBPs.

bp represents base pair.

Table 8. Comparison of different classifiers based on intergenic distances.

		$d \leq 0$	$0 < d \leq 20$	$20 < d \leq 50$	$50 < d \leq 100$	$100 < d \leq 300$	$d > 300$
OP (%)	[a] linear	100	100	100	100	0.0	0.0
	[b] RBF	100	100	98.6	89.7	50.0	60.0
BP (%)	[a] linear	0.0	0.0	0.0	0.0	100	100
	[b] RBF	33.3	25.0	71.4	61.4	93.4	97.0

d represents intergenic distance using number of base pairs.

[a] linear refers to linear SVM classification.

[b] RBF refers to non-linear SVM classification using radial basic function kernel.

4.4.3 Binary classification by non-linear SVM

For separating OPs and BPs in the *B. subtilis* genome, I used SVMs to binary classification of the gene pairs using the five types of features. I combined these features and estimated the contribution for separating OPs and BPs by detection ratios of KOPs and KBPs (Fig. 18). In the case of using all features, I identified 92.9% of KOPs and 86.4% of KBPs.

When only intergenic distance was used for classification of training set, 89.2% of KOPs and 78.7% of KBPs were classified as OPs and BPs respectively. Meanwhile, 92.1% of KOPs and 46.3% of KBPs were classified as OPs and BPs respectively, when only Pearson correlation coefficient was used for classification, and when the other features except intergenic distance and Pearson correlation coefficient were used for classification, every gene pairs were sorted into OPs. These results show that each feature significantly affects classification of gene pairs into OPs, but intergenic distance is likely to have critical contribution of detecting BPs. Moreover, I used SVM classifier with combination of these features, and then over 75% of KBPs were classified as BPs, when intergenic distance and Pearson correlation coefficient or the other features are combined, but I identified only 56.6% of KOPs in the case of using the features except intergenic distance. Therefore, the feature of intergenic distance critically influences the SVM classifier for separating OPs and BPs from the *B. subtilis* gene pairs as the results of previous studies showed that intergenic distance can be effectively used for operon prediction in other prokaryotes [Westover et al., 2005; Moreno-Hagelsieb and Collado-Vides, 2002]. The other features secondarily affect the classification and improve the detection ratios of the KOPs and KBPs with combination of the four types of features.

Using the classification results by SVM, one or more OPs that were composed of consecutive genes on the same strand were regarded as polycistronic operons. I detected 2,669 genes composing 806 operons, which include known operons, from the complete *B. subtilis* genome. This method separates all genes into two groups, so I identified precise structures of operons from many consecutive genes with the high classification accuracy (Fig. 19). For model organisms like *B. subtilis*, SVM could be useful classifier for detecting whole operons on genomes because we can use many kinds of information about the organisms.

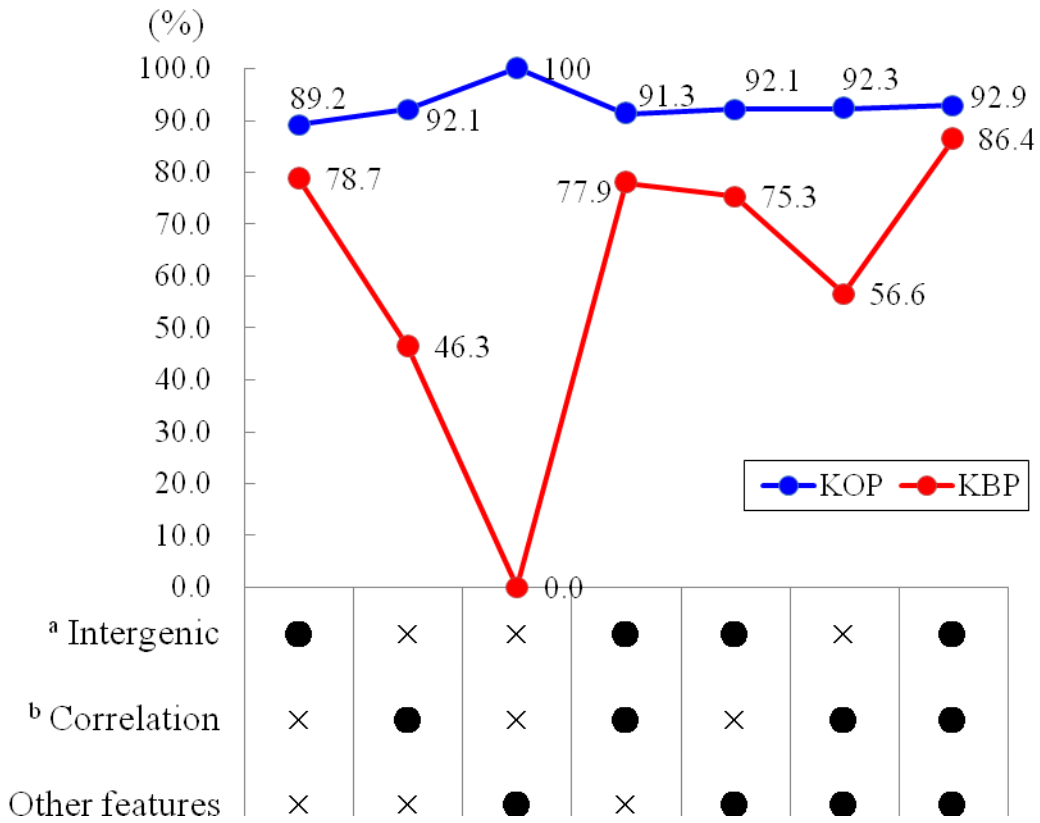


Figure 18. Effect of particular features for non-linear SVM classification.

[a] Intergenic refers to a feature of intergenic distance (bp).

[b] Correlation refers to a feature of Pearson correlation coefficient.

● represents used feature and × represents non-used feature in the SVM classification.

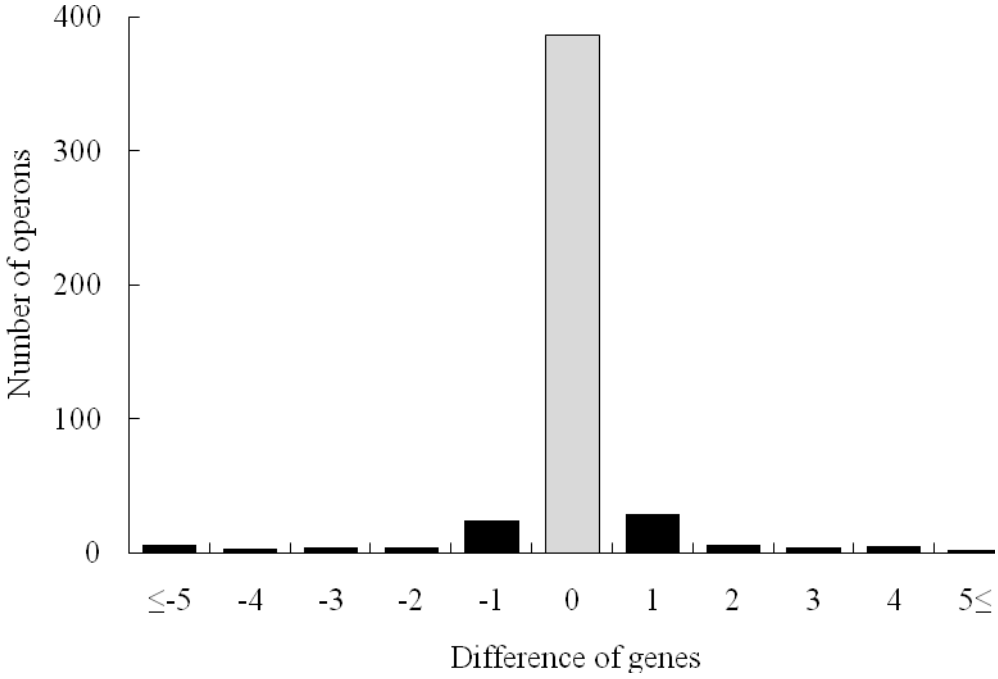


Figure 19. Comparing operons predicted by non-linear SVM with known *B. subtilis* operons, which start with the same promoters as the operons. 81.6% of the known operons matched the predicted operons (i.e., 387 predicted operons in the 474 known *B. subtilis* operons); when I included operons lacking or gaining one gene of known operons, this figure rose to 92.8%. The x-axis indicates difference in gene number between predicted operons and known *B. subtilis* operons, and the y-axis indicates the number of predicted operons.

4.4.4 Comparison of other operon predictions

Several studies have predicted operons in *B. subtilis* whole genome and have used some methods including SVM. Regardless of supervised learning, almost all approaches used intergenic distance as a feature in operon prediction and thus each approach was characterized by other features and various algorithms. In the previous operon prediction method using SVM, Zhang et al. [2006] applied the KEGG pathway data, the conserved gene data among 98 bacteria, the phylogenic profiles and the domain interaction data by Pfam database. I cannot simply compare my result to their result, since they used 310 KOPs and 121 KBPs as training set for SVM which was fewer than half compared to my training set, but my detection of KOPs improved by 2.9 % from their result (Table 9). Some studies applied Bayesian classifier using gene expression data [de Hoon et al. 2004] and GO similarity score [Dam et al. 2007] as characteristic features. Most methods have also been used successfully for operon prediction incorporating different approaches but there was no method detecting operons completely (Table 9). Compared with other predictions of *B. subtilis* operons, my approach is the best for identification of KOPs. This result suggests that the combination of the features used in proposed method is appropriate for detecting operon pairs. On the other hand, my approach is inferior in identification of KBPs to a couple of other results (Table 9). This is because the features for my approach were operonic information and thus I may need to add specific and precise information of boundary pairs as the feature for SVM for raising the precision of detection of boundary pairs.

Table 9. Comparison of operon predictions with previous methods.

	This research	de Hoon (2004)	Price (2005)	Zhang (2006)	DAM (2007)
OP (%)	92.9	88.8	77.0	90.0	89.5
BP (%)	86.4	87.9	88.0	86.0	90.8

5 Conclusion

Various technical innovations for investigating the structures of micro-organisms and their workings have recently brought us many kinds of information about micro-organisms including bacteria, and we can use this information all over the world with the rapid progress of communication engineering. A part of these data, however, contain noise caused by the properties of materials and investigating techniques, and thus technologies for extracting the proper biological information are demanded. In bacterial operon analyses, many studies have demonstrated various approaches for identifying operon structures on genomes using clues about the relationship between genes: intergenic distances [Salgado et al. 2000], transcription control signals [Huerta and Collado-Vides 2003], phylogenetic information based on comparative genomic analyses [Bergman et al. 2007], functional relatedness of two adjacent genes [Westover et al. 2005], and global changes in gene expression [Sabatti et al. 2002]. Moreover, some studies have adopted integration of these clues and efficiently identified operons on genomes [de Hoon et al. 2004; Dam et al. 2007]. In this study, I tried to demonstrate two methods for identifying operons in *B. subtilis* by integration of these clues.

The first approach is an integrated approach for identification of TUs based on statistical and PWM algorithmic methods, and comparative genomic analysis. The integrated approach used two types of cDNA microarray data: time series expression profiles using the cells grown in different growth conditions, and gene expression data of σ-factor and TF deletion mutants, and σ-factor and TF binding sequential data, and thus I efficiently detected genes composing TUs regulated by each σ-factor and TF in the *B. subtilis* genome, and thus I predicted gene regulation networks. In addition, using the comparative genomic analysis in the integrated method, I was able to detect operons candidates and identified σ-factor and TFs regulating operons including internal operons. Moreover, this integrated analysis demands no training data, and thus this method is useful for organisms for which the amount of accumulated data about operons is small.

In the other approach, I used an SVM-based supervised classification approach to predict operon structure for the whole genome. After the determination of the complete genome sequence of the *B. subtilis* genome, various information about this organism, for example, expression data, sequence data, and functional data has rapidly increased, and thus a vast amount of information is available to detect most operonic gene pairs in the genome. As a result, SVM is a very effective approach for identifying operon boundaries in *B. subtilis*, i.e. this method is very useful for operon

prediction for organisms for which huge amount of operonic data is available.

Comparing polycistronic operons detected by the integrated approach with those detected by SVM, 85.8% of genes composing polycistronic operons detected by the integrated approach were identified as the genes detected by SVM (Fig. 20). Figure 20 also shows that 2,811 genes compose polycistronic operons on the *B. subtilis* genome, so one third of genes in the *B. subtilis* genome are monocistronic operons, considering the gene number of *B. subtilis* (Fig. 20). Among overlapping operons identified by both of the approaches, the number of operons identified by the SVM approach (SVM operons) encompassing operons identified by the integrated approach (INT operons) is more than twice the number of INT operons encompassing SVM operons (Table 10). SVM is a binary classification method and every gene pair in the genome is divided into two groups: operon pair and boundary pair, and thus I could detect longer operons than the integrated approach (Table 10). The integrated approach is based on statistical analyses for TFs and gene expression data, and thus this method is useful for detailed structures in operons regulated by each TF, and then I could find internal operons in the overlap regions. In these regions, I also identified 116 operons with identical composition by both analyses and detected 64 new operons structures in these operons (Table 10, Table 11). There are a few reports about the function of the genes in newly identified operons, but the operons including known functional genes are mostly found to be composed of the same or similar functional genes: *uxaCyjmB, gltBA, spoIIIAHIIIAGIIIAFIIIAEIIIADIIIACIIIABIIIAA, yqxAspoIIPgpr, yusKL* and *uppglyA*, and thus these newly identified operons would be reliable (Table 10).

Therefore, both of these methods should respectively contribute to studies of predicting operons in the bacterial genomes and which method to apply to a particular case depends on the availability of data type and data quantity. Furthermore, combining these results could reliably predict operon structures on the genomes.

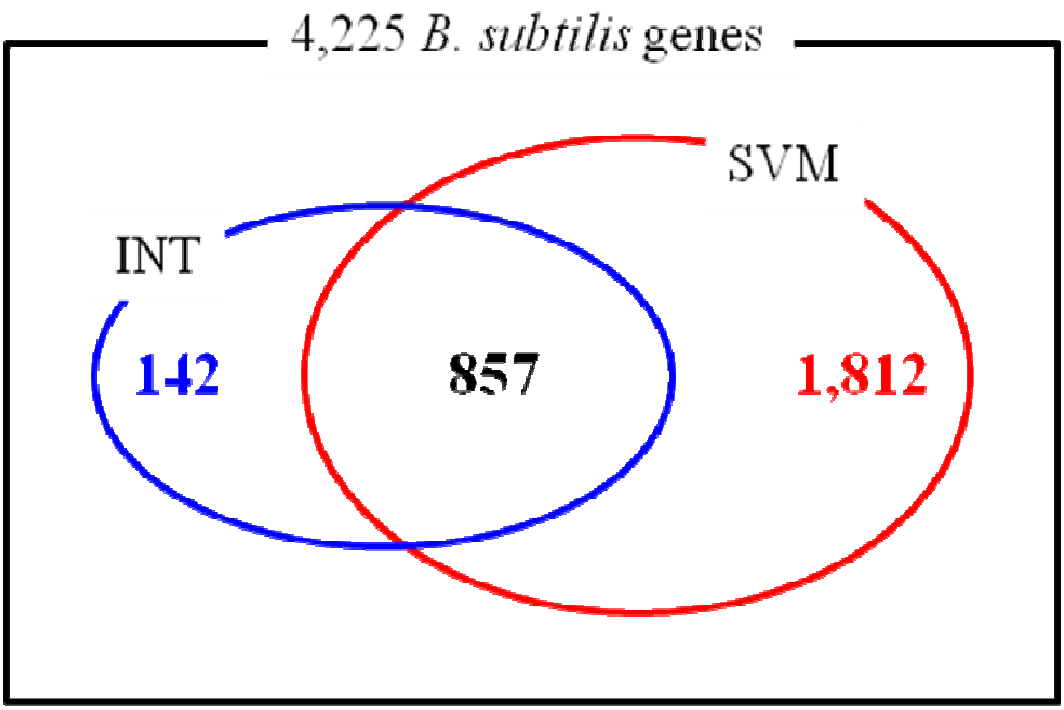
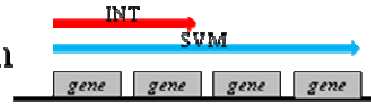
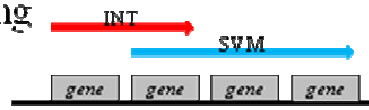
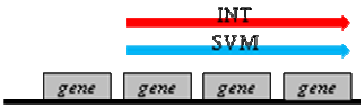


Figure 20. Formation of genes belonging to polycistronic operons identified by the integrated and SVM approaches.

Each figure refers to number of genes identified by these approaches.

INT refers to the integrated approach and SVM refers to support vector machine approach.

Table 10. Comparing gene composition in operons identified by the integrated and SVM approaches in overlap regions on the *B. subtilis* genome.

Gene composition of operons identified by INT and SVM in the overlap regions	Sum
SVM operon encompassing INT operon 	98
INT operon encompassing SVM operon 	44
A part of genes overlapping between both operons 	56
Identical composition 	116

INT operon refers to operon identified by the integrated approach and SVM operon refers to operon identified by support vector machine approach.

Table 11. Operons identified by both the integrated and SVM approaches.

Operon	[a] Function
dnaXyaaK	3.1-0
spoIIEyabST	1.8-0-0
rpoArplQ	0-0
ybbAfeuCBAybbB	3.5-1.2-1.2-1.2-0
pssAybfMpsdybfN	2.4-2.6-2.4-0
ycdFG	0-0
opuAAABAC	0-0-0
ycgQR	0-0
yclJK	0-0
yclNOPQ	0-0-0-0
ydaJKLMN	0-0-0-0-0
ydiKL	0-0
yesLMN	0-0-0
yfmFEDC	0-0-0-0
yfkRQ	0-0
yfiZyfhA	0-0
yhaUTS	0-0-0
yhaGserC	0-0
yhfEF	0-0
yisRdegA	0-0
yjgCD	0-0
uxaCyjmB	2.1-1.2
ykaAykbA	0-0
ykcBC	0-0
ylbKL	0-0
ylmDEF	0-0-0
ymcBA	0-0
citByneN	2.1-1.4
gltBA	2.2-2.2
proJH	0-0
yoxByoaA	0-3.8
yozNyocN	0-0
yodDE	0-0
yoqLKJ	0-0-0
yolBA	0-0
yokFE	0-0
panCB	0-0
aroHBFcheR	0-0-0-0
gpsAyphC	0-0

[a] These numbers indicated gene functions by DBTBS at respective positions in the operon.

Bold faces represent known operons.

Table

Operon	[a] Function
ypfBA	0-0
spoIIIAHIIIAGIIIAFIIIAEIIIADIIIACIIIABIIIAA	1.8-1.8-1.8-1.8-1.8-1.8-1.8-1.8
yqgML	0-0
yqzDC	0-0
yqfON	0-0
yqxAspoIIPgpr	1.9-1.8-0
yrhJI	3.5-4.2
yrhED	0-0
aspShisS	0-0
accAD	0-2.4
ytfJI	0-0
ytqAB	0-0
gbsBA	0-0
yugNM	0-0
yuxIyukJ	0-0
yurUVcsdyurXY	0-0-0-0-0
yusKL	2.4-2.4
fhuCGB	0-0-0
yvaDEF	0-0-0
opuCDCCCBCA	0-0-0-0
yvcBA	0-0
tagHG	0-0
uppglyA	2.2-2.3
tdkrpmE	3.7-0
ywbBA	0-0
yaaDE	4.2-2.2
xpaCyaaN	2.6-4.2
purRyabJ	3.5-2.3
sigWybbM	3.5-0
ybgHJ	2.2-0
lmrBA	3.5-0
ycdHIyceA	1.2-0-1.2
yceCDEFGH	4.2-4.2-4.2-4.2-0-4.2
glcUgdh	1.2-2.1
mtlAD	0-2.1
ydbST	0-0
ydhFphoB	2.6-0
yhdKLsigM	3.5-0-0
gerPFPEPDPCPBPA	1.9-1.9-1.9-1.9-1.9-1.9

Table

Operon	[a] Function
kinEogt	1.3-3.2
motBA	1.5-1.5
fruRKA	3.5-2.1-0
glnRA	3.5-2.2
xylAB	2.1-2.1
parEC	3.4-3.4
odhBA	2.1-2.1
yodHI	0-0
pbuXxpt	2.3-1.2
recUponA	3.3-1.1
ypeBsleB	1.9-1.8
rsiXsigX	3.5-3.5
yqjXW	3.2-0
bkdBABAAlpdVbukbcdptb	2.4-2.4-2.4-2.4-2.4-2.4-2.4-2.4
sigAdnaG	3.1-3.5
yqfBAyqeZ	0-0-0
yqeVUTdnaJKgrpEhrcAhemNlepA	3.7-2.5-3.5-4.1-3.1-3.1-3.8-0-0
hemLBDCXA	2.5-2.5-2.5-2.5-2.5-2.5
yteJsppA	3.8-0
ytxEDccpA	3.5-1.5-1.5
ythQP	0-0
ytlABCD	0-0-0-0
glgPADCB	2.1-2.1-2.1-2.1-2.1
dhbFBECA	2.5-2.5-2.5-2.5-2.5
gerAAABAC	1.9-1.9-1.9
enopgmtpiApgkgapAcggR	3.5-2.1-2.1-2.1-2.1-2.1
yvlDCBA	0-0-0-0
fliTSDyvyC	1.5-1.5-1.5-1.5
lytCBA	1.6-1.1-1.1
gerBABBBC	1.9-1.9-1.9
flhPO	1.5-1.5
narIJHG	1.4-1.4-1.4-1.4
dltABCDE	1.1-1.1-1.1-1.1-1.1
yxzFyxlJ	3.2-0
yxaBA	0-0
ahpCF	4.2-4.2
rocFED	2.2-1.2-2.2
yybNMLKJ	0-0-0-0-0-0

Acknowledgements

I would like to express my special thanks to my supervisor, Professor Shigehiko Kanaya, for his guidance and support during my graduate studies, and to Professor Naotake Ogasawara and Assistant Professor Kazuo Kobayashi for the offer of huge microarray data and their appropriate comments.

I am highly thankful to Mr. Joe Akitomi and Mr. Nobuyuki Fuji for assists of my integrated analysis and to Ms. Yoko Shinbo for support of my comparative genomic analysis.

I am deeply grateful to Professor Wei-Shou Hu and Mr. Salim Charaniya of University of Minnesota for their guidance and support of my support vector machine analysis.

I would also like to thank Associate Professor Ken Kurokawa and Assistant Professor Md. Altaf-Ul-Amin for valuable comments.

References

BSORF: http://bacillus.genome.jp/.

GEO: http://www.ncbi.nlm.nih.gov/geo/index.cgi.

NCBI FTP: http://www.ncbi.nlm.nih.gov/Ftp/.

DBTBS: http://dbtbs.hgc.jp/.

TREBAX: http://kanaya.naist.jp

Albano M, Smits WK, Ho LT, Kraigher B, Mandic-Mulec I, Kuipers OP, Dubnau D: **The Rok protein of Bacillus subtilis represses genes for cell surface and extracellular functions**. *J Bacteriol* 2005, **187**(6):2010-2019.

Ali NO, Bignon J, Rapoport G, Debarbouille M: **Regulation of the acetoin catabolic pathway is controlled by sigma L in Bacillus subtilis**. *J Bacteriol* 2001, **183**(8):2497-2504.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

Asai K, Takamatsu H, Iwano M, Kodama T, Watabe K, Ogasawara N: **The Bacillus subtilis yabQ gene is essential for formation of the spore cortex**. *Microbiology* 2001, **147**(Pt 4):919-927.

Barabesi C, Galizzi A, Mastromei G, Rossi M, Tamburini E, Perito B: **Bacillus subtilis gene cluster involved in calcium carbonate biomineralization**. *J Bacteriol* 2007, **189**(1):228-235.

Belitsky BR, Sonenshein AL: **An enhancer element located downstream of the major glutamate dehydrogenase gene of Bacillus subtilis**. *Proc Natl Acad Sci U S A* 1999, **96**(18):10290-10295.

Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing**. *J Roy Stat Soc B Met* 1995, **57**((1)):289-300.

Bergman NH, Passalacqua KD, Hanna PC, Qin ZS: **Operon prediction for sequenced bacterial genomes without experimental information**. *Applied and environmental microbiology* 2007, **73**(3):846-854.

Bockhorst J, Craven M, Page D, Shavlik J, Glasner J: **A Bayesian network approach to operon**

prediction. *Bioinformatics* 2003, **19**(10):1227-1235.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines**. *Proc Natl Acad Sci U S A* 2000, **97**(1):262-267.

Cabrera-Hernandez A, Sanchez-Salas JL, Paidhungat M, Setlow P: **Regulation of four genes encoding small, acid-soluble spore proteins in Bacillus subtilis**. *Gene* 1999, **232**(1):1-10.

Cervin MA, Lewis RJ, Brannigan JA, Spiegelman GB: **The Bacillus subtilis regulator SinR inhibits spoIIG promoter transcription in vitro without displacing RNA polymerase**. *Nucleic Acids Res* 1998, **26**(16):3806-3812.

Charaniya S, Mehra S, Lian W, Jayapal KP, Karypis G, Hu WS: **Transcriptome dynamics-based operon prediction and verification in Streptomyces coelicolor**. *Nucleic Acids Res* 2007, **35**(21):7222-7236.

Dam P, Olman V, Harris K, Su Z, Xu Y: **Operon prediction using both genome-specific and general genomic information**. *Nucleic Acids Res* 2007, **35**(1):288-298.

De Hoon MJ, Imoto S, Kobayashi K, Ogasawara N, Miyano S: **Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information**. *Pacific Symposium on Biocomputing* 2004:276-287.

Derre I, Rapoport G, Devine K, Rose M, Msadek T: **ClpE, a novel type of HSP100 ATPase, is part of the CtsR heat shock regulon of Bacillus subtilis**. *Mol Microbiol* 1999, **32**(3):581-593.

Deuel TF, Prusiner S: **Regulation of glutamine synthetase from Bacillus subtilis by divalent cations, feedback inhibitors, and L-glutamine**. *J Biol Chem* 1974, **249**(1):257-264.

Dudoit S, Fridlyand J, Speed T: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *J Am Stat Ass* 2002, **97**(457):77-87.

Ebbole DJ, Zalkin H: **Bacillus subtilis pur operon expression and regulation**. *J Bacteriol* 1989, **171**(4):2136-2141.

Edwards MT, Rison SC, Stoker NG, Wernisch L: **A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context**. *Nucleic Acids Res* 2005, **33**(10):3253-3262.

Edwards MT, Rison SC, Stoker NG, Wernisch L: **A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context**.

Nucleic Acids Res 2005, **33**(10):3253-3262.

Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes**. *Nucleic Acids Res* 2001, **29**(5):1216-1221.

Fisher SH: **Regulation of nitrogen metabolism in Bacillus subtilis: vive la difference!** *Mol Microbiol* 1999, **32**(2):223-232.

Fujita Y, Fujita T: **The gluconate operon gnt of Bacillus subtilis encodes its own transcriptional negative regulator**. *Proc Natl Acad Sci U S A* 1987, **84**(13):4524-4528.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, **16**(10):906-914.

Gaballa A, Helmann JD: **Identification of a zinc-specific metalloregulatory protein, Zur, controlling zinc transport operons in Bacillus subtilis**. *J Bacteriol* 1998, **180**(22):5815-5821.

Gao G, Le D, Huang L, Lu H, Narumi I, Hua Y: **Internal promoter characterization and expression of the Deinococcus radiodurans pprI-folP gene cluster**. *FEMS microbiology letters* 2006, **257**(2):195-201.

Gardan R, Rapoport G, Debarbouille M: **Expression of the rocDEF operon involved in arginine catabolism in Bacillus subtilis**. *J Mol Biol* 1995, **249**(5):843-856.

Gotsche S, Dahl MK: **Purification and characterization of the phospho-alpha(1,1)glucosidase (TreA) of Bacillus subtilis 168**. *J Bacteriol* 1995, **177**(10):2721-2726.

Guillen N, Weinrauch Y, Dubnau DA: **Cloning and characterization of the regulatory Bacillus subtilis competence genes comA and comB**. *J Bacteriol* 1989, **171**(10):5354-5361.

Hanson RS, Cox DP: **Effect of different nutritional conditions on the synthesis of tricarboxylic acid cycle enzymes**. *J Bacteriol* 1967, **93**(6):1777-1787.

Hecker M, Schumann W, Volker U: **Heat-shock and general stress response in Bacillus subtilis**. *Mol Microbiol* 1996, **19**(3):417-428.

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA *et al*: **Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL**. *Nucleic Acids Res* 1998, **26**(1):362-367.

Henikoff S, Haughn GW, Calvo JM, Wallace JC: **A large family of bacterial activator proteins**. *Proc Natl Acad Sci U S A* 1988, **85**(18):6602-6606.

Henkin TM: **The role of CcpA transcriptional regulator in carbon metabolism in Bacillus**

subtilis. *FEMS microbiology letters* 1996, **135**(1):9-15.

Hilbert DW, Piggot PJ: **Compartmentalization of gene expression during Bacillus subtilis spore formation**. *Microbiol Mol Biol Rev* 2004, **68**(2):234-262.

Hoa TT, Tortosa P, Albano M, Dubnau D: **Rok (YkuW) regulates genetic competence in Bacillus subtilis by directly repressing comK**. *Mol Microbiol* 2002, **43**(1):15-26.

Horsburgh MJ, Moir A: **Sigma M, an ECF RNA polymerase sigma factor of Bacillus subtilis 168, is essential for growth and survival in high concentrations of salt**. *Mol Microbiol* 1999, **32**(1):41-50.

Huang M, Oppermann-Sanio FB, Steinbuchel A: **Biochemical and molecular characterization of the Bacillus subtilis acetoin catabolic pathway**. *J Bacteriol* 1999, **181**(12):3837-3841.

Huang X, Helmann JD: **Identification of target promoters for the Bacillus subtilis sigma X factor using a consensus-directed search**. *J Mol Biol* 1998, **279**(1):165-173.

Huerta AM, Collado-Vides J: **Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals**. *J Mol Biol* 2003, **333**(2):261-278.

Huerta AM, Salgado H, Thieffry D, Collado-Vides J: **RegulonDB: a database on transcriptional regulation in Escherichia coli**. *Nucleic Acids Res* 1998, **26**(1):55-59.

Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins.** *J Mol Biol* 1961, **3**:318-356.

Janga SC, Lamboy WF, Huerta AM, Moreno-Hagelsieb G: **The distinctive signatures of promoter regions and operon junctions across prokaryotes**. *Nucleic Acids Res* 2006, **34**(14):3980-3987.

Joachims T: **Making Large-Scale SVM Learning Practical**: MIT Press; 1999.

Jung SH: **Sample size for FDR-control in microarray data analysis**. *Bioinformatics* 2005, **21**(14):3097-3104.

Kearns DB, Chu F, Branda SS, Kolter R, Losick R: **A master regulator for biofilm formation by Bacillus subtilis**. *Mol Microbiol* 2005, **55**(3):739-749.

Kennett RH, Sueoka N: **Gene expression during outgrowth of Bacillus subtilis spores. The relationship between gene order on the chromosome and temporal sequence of enzyme synthesis**. *J Mol Biol* 1971, **60**(1):31-44.

Klein W, Horlacher R, Boos W: **Molecular analysis of treB encoding the Escherichia coli enzyme II specific for trehalose**. *J Bacteriol* 1995, **177**(14):4043-4052.

Kobayashi K, Ogura M, Yamaguchi H, Yoshida K, Ogasawara N, Tanaka T, Fujita Y: **Comprehensive DNA microarray analysis of Bacillus subtilis two-component regulatory systems**. *J Bacteriol* 2001, **183**(24):7365-7370.

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al*: **The complete genome sequence of the Gram-positive bacterium Bacillus subtilis**. *Nature* 1997, **390**(6657):249-256.

Lee JHL, C. J.: **Automatic model selection for support vector machines.** In. Taipei: Department of Computer Science and Information Engineering, National Taiwan University; 2000.

Levin PA, Losick R: **Characterization of a cell division gene from Bacillus subtilis that is required for vegetative and sporulation septum formation**. *J Bacteriol* 1994, **176**(5):1451-1459.

Lorca G, Winnen B, Saier MH, Jr.: **Identification of the L-aspartate transporter in Bacillus subtilis**. *J Bacteriol* 2003, **185**(10):3218-3222.

Madan Babu M, Teichmann SA, Aravind L: **Evolutionary dynamics of prokaryotic transcriptional regulatory networks**. *J Mol Biol* 2006, **358**(2):614-633.

Magasanik B: **Genetic control of nitrogen assimilation in bacteria**. *Annual review of genetics* 1982, **16**:135-168.

Magill NG, Cowan AE, Leyva-Vazquez MA, Brown M, Koppel DE, Setlow P: **Analysis of the relationship between the decrease in pH and accumulation of 3-phosphoglyceric acid in developing forespores of Bacillus species**. *J Bacteriol* 1996, **178**(8):2204-2210.

Marquez-Magana LM, Chamberlin MJ: **Characterization of the sigD transcription unit of *Bacillus subtilis***. *J Bacteriol* 1994, **176**(8):2427-2434.

Matsuoka H, Hirooka K, Fujita Y: **Organization and function of the YsiA regulon of *Bacillus subtilis* involved in fatty acid degradation**. *J Biol Chem* 2007, **282**(8):5180-5194.

Matthews KS, Nichols JC: **Lactose repressor protein: functional properties and structure.** *Prog Nucleic Acid Res Mol Biol*. 1998, **58**:127-64.

Mirel DB, Chamberlin MJ: **The Bacillus subtilis flagellin gene (hag) is transcribed by the sigma 28 form of RNA polymerase**. *J Bacteriol* 1989, **171**(6):3095-3101.

Miwa Y, Nakata A, Ogiwara A, Yamamoto M, Fujita Y: **Evaluation and characterization of catabolite-responsive elements (cre) of Bacillus subtilis**. *Nucleic Acids Res* 2000,

28(5):1206-1210.

Moreno-Campuzano S, Janga SC, Perez-Rueda E: **Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes--a genomic approach**. In: *BMC genomics.* vol. 7; 2006: 147.

Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes**. *Bioinformatics* 2002, **18 Suppl 1**:S329-336.

Msadek T, Dartois V, Kunst F, Herbaud ML, Denizot F, Rapoport G: **ClpP of Bacillus subtilis is required for competence development, motility, degradative enzyme synthesis, growth at high temperature and sporulation**. *Mol Microbiol* 1998, **27**(5):899-914.

Nakatani Y, Nicholson WL, Neitzke KD, Setlow P, Freese E: **Sigma-G RNA polymerase controls forespore-specific expression of the glucose dehydrogenase operon in Bacillus subtilis**. *Nucleic Acids Res* 1989, **17**(3):999-1017.

Noble WS: **Support vector machine applications in computational biology**: MIT Press; 2004.

Ogasawara N, Moriya S, Yoshikawa H: **Structure and function of the region of the replication origin of the Bacillus subtilis chromosome. IV. Transcription of the oriC region and expression of DNA gyrase genes and other open reading frames**. *Nucleic Acids Res* 1985, **13**(7):2267-2279.

Ollinger J, Song KB, Antelmann H, Hecker M, Helmann JD: **Role of the Fur regulon in iron transport in Bacillus subtilis**. *J Bacteriol* 2006, **188**(10):3664-3673.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci U S A* 1999, **96**(6):2896-2901.

Pabo CO, Sauer RT: **Protein-DNA recognition.** *Annu Rev Biochem.* 1984, **53**:293-321.

Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies**. *Bioinformatics* 2005, **21**(13):3017-3024.

Price MN, Huang KH, Alm EJ, Arkin AP: **A novel method for accurate operon predictions in all sequenced prokaryotes**. *Nucleic Acids Res* 2005, **33**(3):880-892.

Qiu P, Qin L, Sorrentino RP, Greene JR, Wang L, Partridge NC: **Comparative promoter analysis and its application in analysis of PTH-regulated gene expression**. *J Mol Biol* 2003, **326**(5):1327-1336.

Quackenbush J: **Microarray data normalization and transformation**. *Nat Genet* 2002, **32 Suppl**:496-501.

Sabatti C, Rohlin L, Oh MK, Liao JC: **Co-expression pattern from DNA microarray experiments as a tool for operon prediction**. *Nucleic Acids Res* 2002, **30**(13):2886-2893.

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in Escherichia coli: genomic analyses and predictions**. *Proc Natl Acad Sci U S A* 2000, **97**(12):6652-6657.

Sa-Nogueira I, Mota LJ: **Negative regulation of L-arabinose metabolism in Bacillus subtilis: characterization of the araR (araC) gene**. *J Bacteriol* 1997, **179**(5):1598-1608.

Schock F, Dahl MK: **Expression of the tre operon of Bacillus subtilis 168 is regulated by the repressor TreR**. *J Bacteriol* 1996, **178**(15):4576-4581.

Schujman GE, Paoletti L, Grossman AD, de Mendoza D: **FapR, a bacterial transcription factor involved in global regulation of membrane lipid biosynthesis**. *Dev Cell* 2003, **4**(5):663-672.

Serizawa M, Kodama K, Yamamoto H, Kobayashi K, Ogasawara N, Sekiguchi J: **Functional analysis of the YvrGHb two-component system of Bacillus subtilis: identification of the regulated genes by DNA microarray and northern blot analyses**. *Bioscience, biotechnology, and biochemistry* 2005, **69**(11):2155-2169.

Setlow P: **Mechanisms for the prevention of damage to DNA in spores of Bacillus species**. *Annual review of microbiology* 1995, **49**:29-54.

Sonenshein AL, Hoch JA, Losick RM: **Bacillus subtilis and Its Closest Relatives: from Genes to Cells**: ASM Press; 2001.

Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**(1):16-23.

Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S: **Machine learning and its applications to biology**. *PLoS computational biology* 2007, **3**(6):e116.

Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families**. *Science* 1997, **278**(5338):631-637.

Thieffry D, Salgado H, Huerta AM, Collado-Vides J: **Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12**. *Bioinformatics* 1998, **14**(5):391-400.

Tovar-Rojo F, Cabrera-Martinez RM, Setlow B, Setlow P: **Studies on the mechanism of the osmoresistance of spores of Bacillus subtilis**. *Journal of applied microbiology* 2003, **95**(1):167-179.

Turner MS, Helmann JD: **Mutations in multidrug efflux homologs, sugar isomerases, and antimicrobial biosynthesis genes differentially elevate activity of the sigma(X) and sigma(W) factors in Bacillus subtilis**. *J Bacteriol* 2000, **182**(18):5202-5210.

Vapnik VN: **Statistical Learning Theory**: Wiley-Interscience; 1998.

Wang L, Trawick JD, Yamamoto R, Zamudio C: **Genome-wide operon prediction in Staphylococcus aureus**. *Nucleic Acids Res* 2004, **32**(12):3689-3702.

Westover BP, Buhler JD, Sonnenburg JL, Gordon JI: **Operon prediction without a training set**. *Bioinformatics* 2005, **21**(7):880-888.

Wiegeshoff F, Beckering CL, Debarbouille M, Marahiel MA: **Sigma L is important for cold shock adaptation of Bacillus subtilis**. *J Bacteriol* 2006, **188**(8):3130-3133.

Wipat A, Carter N, Brignell SC, Guy BJ, Piper K, Sanders J, Emmerson PT, Harwood CR: **The dnaB-pheA (256 degrees-240 degrees) region of the Bacillus subtilis chromosome containing genes responsible for stress responses, the utilization of plant cell walls and primary metabolism**. *Microbiology* 1996, **142 (Pt 11)**:3067-3078.

Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context**. *Genome research* 2001, **11**(3):356-372.

Yada T, Nakao M, Totoki Y, Nakai K: **Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models**. *Bioinformatics* 1999, **15**(12):987-993.

Yamanishi Y, Vert JP, Nakaya A, Kanehisa M: **Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis**. *Bioinformatics* 2003, **19 Suppl 1**:i323-330.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucleic Acids Res* 2002, **30**(4):e15.

Yoshida KI, Aoyama D, Ishio I, Shibayama T, Fujita Y: **Organization and transcription of the myo-inositol operon, iol, of Bacillus subtilis**. *J Bacteriol* 1997, **179**(14):4591-4598.

Zhang GQ, Cao ZW, Luo QM, Cai YD, Li YX: **Operon prediction based on SVM**. *Computational biology and chemistry* 2006, **30**(3):233-240.

Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S: **Computational identification of operons in microbial genomes**. *Genome research* 2002, **12**(8):1221-1230.

Zien A, Ratsch G, Mika S, Scholkopf B, Lengauer T, Muller KR: **Engineering support vector machine kernels that recognize translation initiation sites**. *Bioinformatics* 2000, **16**(9):799-807.

Appendix

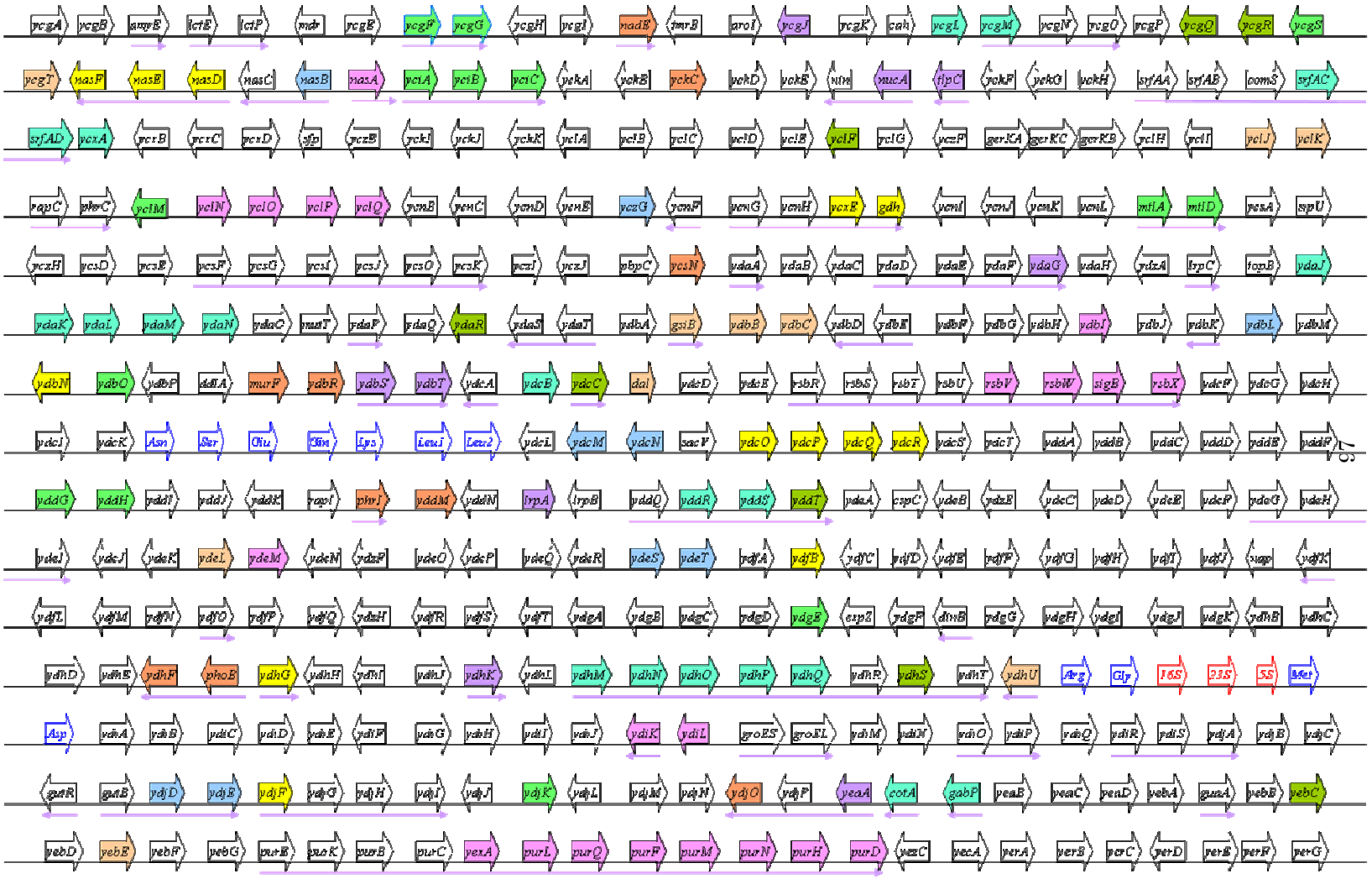
Appendix A. Operon map on the *B. subtilis* genome identified by the integrated analysis
Purple thin arrows refer to known *B. subtilis* operon structures (DBTBS).

Appendix B. Functional conformation of *B. subtilis* operons identified by the integrated analysis.
[a] CDF is an abbreviation of compound degree of functions.
[b] These numbers indicated gene functions by DBTBS at respective positions in the operon.

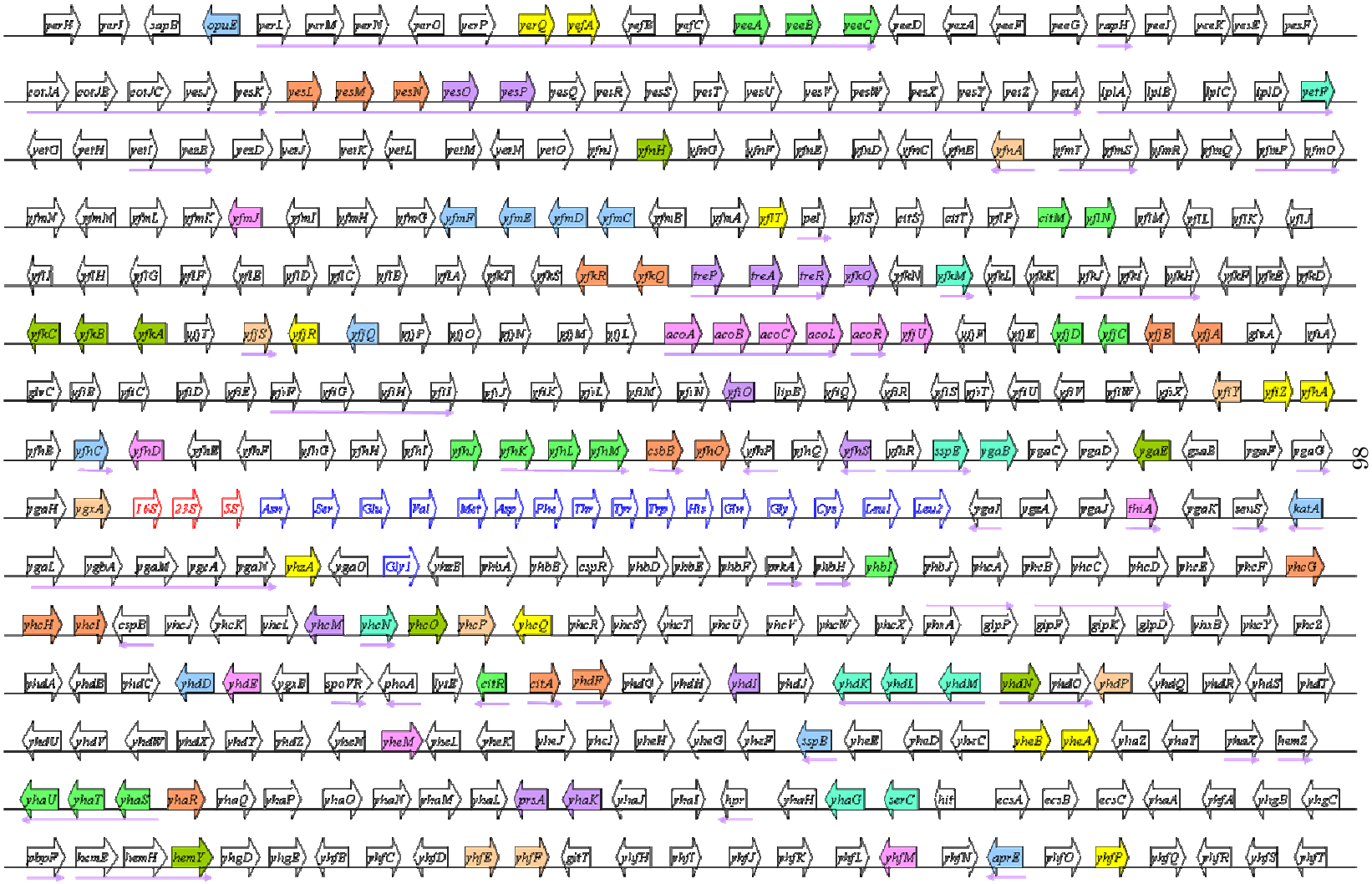
Appendix C. Classification of operon pairs and boundary pairs by support vector machine analysis.
Yellow columns indicate operon pairs and blue columns indicate boundary pairs.


dnaA dnaN yaaA recF yaaB gyrB gyrA 16S Ile Ala 23S 5S yaaC guaB dacA yaaD yaaE serS Ser1 yaaF yaaG yaaH yaaI yaaJ ser
dnaX yaaK recR yaaL bofA 16S Ile Ala 23S 5S cylB rpaC yaaN yaaO tmk yaaQ yaaR holB yaaT yabA yabB razA rubC abrB metS
yabD yabE yabF ksgA yabG veg sspF yebH purR yabJ spoVG gcaD prs etc spoVC yabK mfd spoVT yabM yabN yabO yabP yabQ divIC yabR
Met1 Glu1 spoIIE yabS yabT yacA hprT ftsH yacE yacC yacD cysK yabB yabA yabC sul folA folK yazE yacF lysS 16S 23S 5S Val
Thr Lys Leu1 Gly Leu2 Arg Pro Ala 16S 23S 5S ctsR yacH yac1 clpC sms yacK yacL yacM yacN gltX cysE cysS yazC yacO yacP
sigH rpmG secE yazG rplK rplA rplJ rplL yhxB rpoB rpoC yhxF rpsL rpsG fus tufA yhaC rpsJ rplC rplD rplW rplB rpsS rplV rpsC
rplP rpmC rpsQ rplN rplX rplE rpsN rpsH rplF rplR rpsE rpmD rplO secY adk map infA rpmJ rpsM rpsK rpoA rplQ yhxA yhaE yhaP
truA rplM rpsI yhaJ yhaK cwlD yhaL gerD khaA yhaM 16S 23S 5S Asn Thr Gly Arg Pro Ala 16S 23S 5S 16S 23S 5S
yhaR yhaS yhbA feuC feuB feuA yhbB yhbC yhbD yhbE yhbF yhbH yhbI yhbJ yhbK Glu3 Val1 Thr1 Tyr1 Gln1 sigW yhbM yhbF yhbR yhbT
glmS yhbU alkA adaA adaB ndhF ybcC ybcD ybcF ybcH ybc1 ybcL ybcM ybcO ybcP ybcQ ybcS ybcT ybdA ybdB ybdD ybdE ybdG ybdI ybdK
ybdL ybdM ybdN ybdO ybxG csgA ybxH ybxI ybdT ybyE ybeC gtpQ gtpT ybeP ybfA ybjE ybjB ybjF ybjG ybjH ybjI purT ispr ybjJ ybjK
pssA ybfM psd ybjN ybfO ybfP ybfQ gltP ybfS ybfT ybgA ybgB ybgE ybgF ybgG ybgH ybgJ ycbA ycbB ycbC ycbD ycbE ycbF ycbG ycbH
ycbJ yczA ycbK ycbL ycbM ycbN ycbO yebP cwlJ ycbR phoD yczE ycbT pcp ycbU lmrB lmrA yccC lipA yczC yccF yccG yccH matA matB
yceK yceA yceB yceC yceD yceE yceF yceG yceH yce1 yceJ yceK opuAA opuAB opuAC amhX

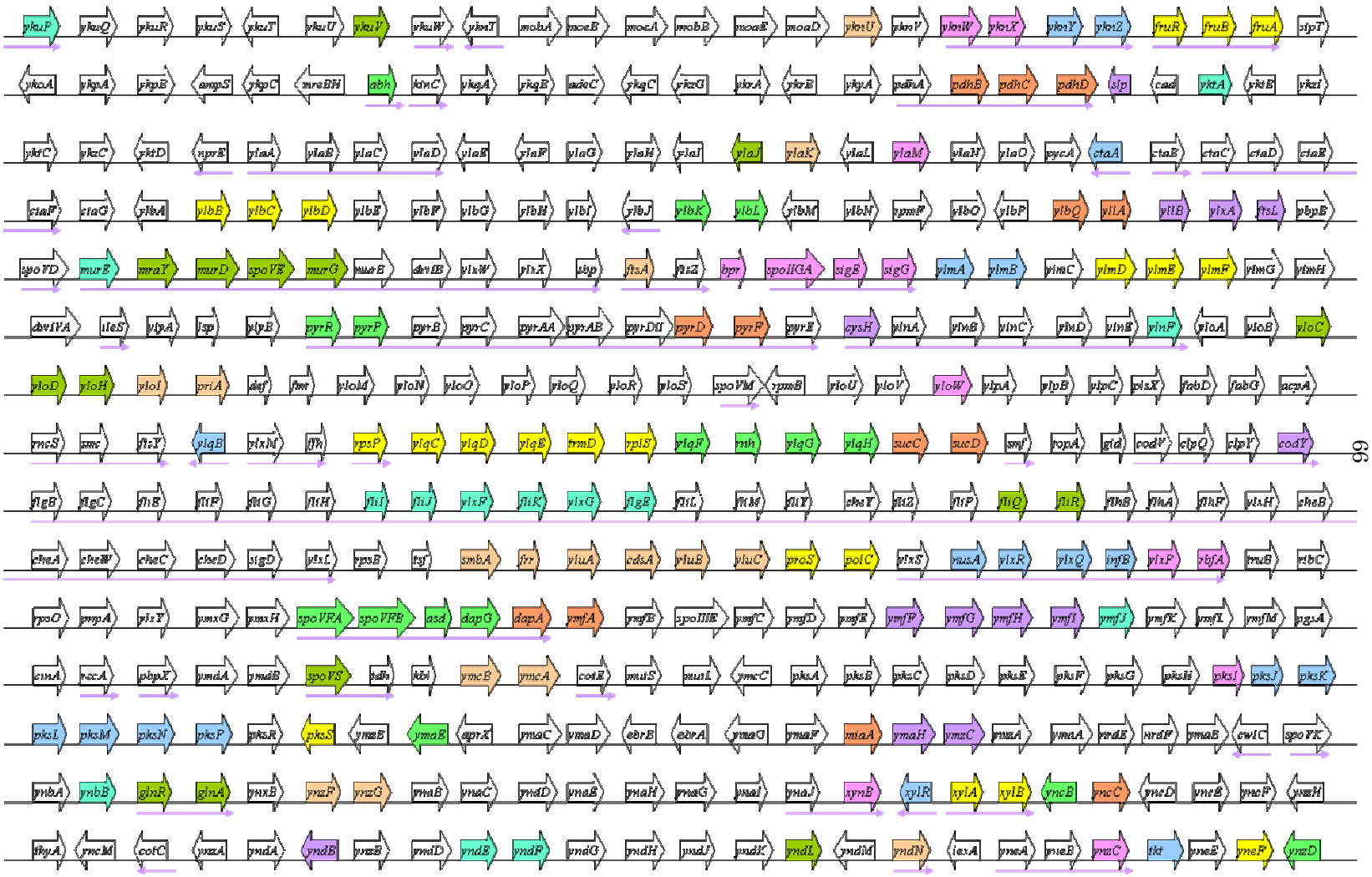
Appendix A.



Appendix A. (Continued)



Appendix A. (Continued)



Appendix A. (Continued)

100

Appendix A. (Continued)

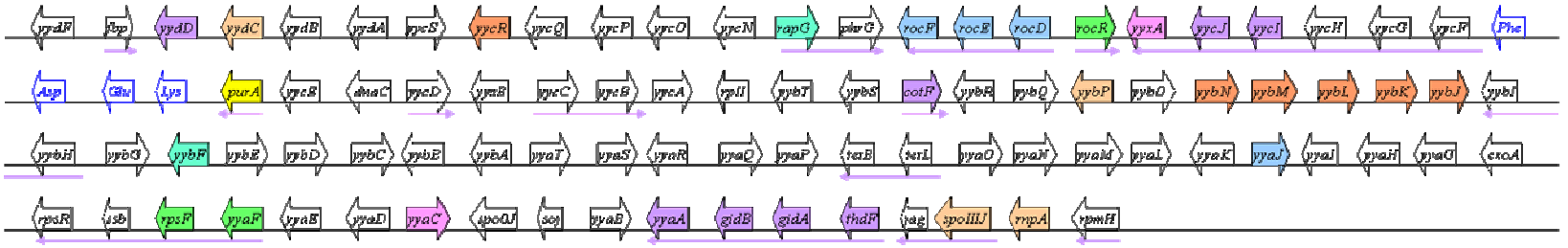
Appendix A. (Continued)

Appendix A. (Continued)

Appendix A. (Continued)

Appendix A. (Continued)

Appendix A. (Continued)



Appendix A. (Continued)

Appendix B.

Operon	CDF[a]	Function[b]
dnaAN	**1**	3.1-3.1
yaaArecFyaaBgyrB	**1**	0-3.3-0-3.4
gyrA	**1**	3.4
dacA	**0**	0
yaaDE	**2**	4.2-2.2
serS	**1**	3.7
yaaH	**1**	1.8
dnaXyaaK	**0**	0-0
recRyaaLbofA	**1**	0-0-1.8
csfB	**1**	4.6
xpaCyaaN	**2**	2.6-4.2
abrB	**1**	3.5
yabGvegsspF	**2**	1.8-4.6-1.8
purRyabJ	**2**	3.5-2.3
spoVGgcaDprs	**2**	1.8-1.1-2.3
ctc	**1**	4.1
yabKmfdspoVTyabMNOPQ	**2**	0-0-3.5-1.2-0-0-0-0
divIC	**1**	1.7
spoIIEyabST	**1**	1.8-0-0
ftsH	**1**	1.7
yacBCD	**0**	0-0-0
cysK	**1**	2.2
pabBACsulfolAKyazByacFlysS	**2**	2.5-2.5-2.5-2.5-2.5-2.5-0-3.5-3.7
ctsRyacHI	**2**	3.5-4.1-4.1
clpCsms	**2**	4.1-3.2
yacLMN	**0**	0-0-0
gltXcysESyazCyacOP	**2**	3.7-2.2-3.7-0-3.6-0
sigH	**1**	3.5
rpoBCybxFrpsLGfustufA	**1**	3.5-0-0-0-0-0-0
rplWBrpsSrplVrpsCrplPrpmCrpsQrplNXE-rpsNHrplFRrpsErpmDrplOsecYadkmap	**0**	0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0
infArpmJ	**0**	0-0
rpoArplQ	**0**	0-0
truA	**0**	0
ybaKcwlD	**1**	0-1.1
ybaL	**0**	0
gerD	**1**	1.9
ybaN	**1**	2.1
ybaRS	**0**	0-0
ybbAfeuCBAybbB	**2**	0-1.2-1.2-1.2-3.5
ybbCDEF	**0**	0-0-0-0
ybbK	**0**	0
sigWybbM	**1**	3.5-0
alkA	**1**	3.2
adaB	**1**	3.2
ndhFybcC	**0**	0-0
ybdBD	**0**	0-0
ybdKL	**1**	1.3-0
ybdO	**0**	0
ybxG	**1**	2.1

Appendix B. (Continued)

csgAybxH	**1**	1.8-0
ybyB	**0**	0
glpQT	**2**	2.4-1.2
purT	**1**	2.3
pssAybfMpsdybfN	**1**	2.4-2.6-2.4-0
ybfOP	**1**	4.2-0
gltP	**0**	0
ybfST	**0**	0-0
ybgB	**0**	0
ybgE	**1**	2.2
ybgHJ	**1**	0-2.2
ycbABCDEFGHJ	**3**	0-0-2.1-2.1-0-2.1-3.5-0-4.2
yczAycbK	**0**	0-0
ycbP	**0**	0
cwlJ	**1**	1.1
phoD	**1**	2.6
lmrBA	**1**	0-3.5
yccC	**1**	2.2
ycdD	**0**	0
ycdFG	**0**	0-0
ycdHIyceA	**1**	1.2-0-1.2
yceCDEFGH	**1**	4.2-4.2-4.2-4.2-0-4.2
yceI	**0**	0
opuAAABAC	**0**	0-0-0
amhX	**1**	3.8
amyE	**1**	2.1
lctEP	**1**	1.4-0
ycgFG	**0**	0-0
nadE	**1**	2.5
ycgJ	**0**	0
ycgLM	**0**	0-0
ycgQR	**0**	0-0
ycgS	**0**	0
ycgT	**1**	1.4
nasFED	**1**	2.5-2.2-2.2
nasB	**1**	2.2
nasA	**0**	0
yciABC	**1**	0-0-1.2
yckC	**0**	0
nucAtlpC	**2**	2.3-1.5
srfACADycxA	**1**	4.3-4.3-0
yclF	**0**	0
yclJK	**0**	0-0
rapCphrC	**1**	1.8-1.8
yclM	**1**	2.2
yclNOPQ	**0**	0-0-0-0
yczG	**1**	3.5
ycnF	**1**	3.5
ycnGH	**1**	2.2-2.2
ycxEgdh	**2**	1.2-2.1
mtlAD	**1**	0-2.1
ycsFGIJOK	**3**	4.2-0-0-1.8-3.5-0

Appendix B. (Continued)

ycsN	**0**	0
ydaA	**1**	3.5
ydaG	**1**	4.1
lrpC	**1**	3.5
ydaJKLMN	**0**	0-0-0-0-0
ydaP	**1**	2.1
ydaR	**1**	1.2
ydaST	**0**	0-0
gsiBydbBC	**1**	4.1-0-0
ydbDE	**1**	4.2-0
ydbI	**0**	0
ydbL	**0**	0
ydbN	**0**	0
ydbO	**1**	1.2
murFydbR	**0**	0-0
ydbST	**0**	0-0
ydcB	**0**	0
ydcC	**0**	0
dal	**0**	0
rsbVWsigBrsbX	**2**	4.1-4.1-3.5-4.1
ydcMN	**0**	0-0
ydcOPQR	**0**	0-0-0-0
yddGH	**0**	0-0
phrIyddM	**1**	1.8-0
lrpA	**0**	0
yddRS	**0**	0-0
yddT	**0**	0
ydeGHI	**0**	0-0-0
ydeL	**0**	0
ydeM	**0**	0
ydeST	**0**	0-0
ydfB	**0**	0
ydfK	**0**	0
ydfO	**0**	0
ydgE	**0**	0
dinB	**1**	3.2
ydhD	**1**	1.8
ydhFphoB	**1**	0-2.6
ydhG	**0**	0
ydhK	**0**	0
ydhMNOPQ	**0**	0-0-0-0-0
ydhS	**0**	0
ydhU	**1**	4.2
ydiKL	**0**	0-0
groESEL	**1**	3.9-3.9
ydiOP	**1**	3.2-3.2
ydiRSydjA	**1**	0-3.2-0
gutR	**1**	3.5
ydjDE	**0**	0-0
ydjF	**1**	4.1
ydjK	**0**	0
ydjO	**0**	0

Appendix B. (Continued)

yeaA	**0**	0
cotAgabP	**1**	1.8-1.2
guaA	**1**	2.3
yebC	**0**	0
yebE	**0**	0
yexApurLQFMNHD	**1**	2.3-2.3-2.3-2.3-2.3-2.3-2.3-0
opuE	**0**	0
yerQyefA	**0**	0-0
yeeABC	**0**	0-0-0
rapH	**0**	0
cotJAJBJC	**1**	1.8-1.8-1.8
yesLMN	**0**	0-0-0
yesOP	**0**	0-0
yetF	**0**	0
yetIyezB	**0**	0-0
yfnH	**1**	2.1
yfnA	**0**	0
yfmTS	**2**	2.1-1.5
yfmPO	**1**	3.5-0
yfmJ	**1**	1.4
yfmFEDC	**0**	0-0-0-0
yflT	**0**	0
pel	**1**	2.1
citMyflN	**1**	2.1-0
yfkRQ	**0**	0-0
trePARyfkO	**3**	0-2.1-3.5-1.4
yfkM	**0**	0
yfkJIH	**1**	3.8-0-0
yfkCBA	**0**	0-0-0
yfjS	**1**	2.1
yfjR	**1**	2.4
yfjQ	**0**	0
acoABCLRyfjU	**3**	2.1-2.1-2.1-2.1-3.5-1.8
yfjDC	**0**	0-0
yfjBA	**0**	0-0
glvAyfiAglvC	**1**	2.1-0-0
yfiFGHI	**1**	3.5-0-0-0
yfiO	**1**	3.5
yfiY	**0**	0
yfiZyfhA	**0**	0-0
yfhC	**0**	0
yfhD	**0**	0
yfhJKLM	**2**	0-1.7-0-2.1
csbByfhO	**1**	4.1-0
yfhP	**0**	0
yfhS	**0**	0
sspEygaB	**1**	1.8-0
ygaE	**0**	0
ygaG	**1**	4.2
ygxA	**0**	0
ygaI	**1**	1.8
thiA	**1**	2.5

Appendix B. (Continued)

senS	**1**	3.5
katA	**1**	4.2
ygaLygbAygaMygcAygaN	**1**	0-0-0-2.7-0
yhzA	**0**	0
prkA	**1**	3.8
yhbH	**0**	0
yhbI	**1**	3.5
yhcGHI	**0**	0-0-0
cspB	**1**	4.1
yhcM	**0**	0
yhcN	**0**	0
yhcO	**0**	0
yhcP	**0**	0
yhcQ	**0**	0
yhxAglpP	**2**	2.5-3.5
glpFKD	**2**	1.2-2.1-2.1
yhdD	**1**	1.1
yhdE	**0**	0
spoVR	**1**	1.8
phoA	**1**	2.6
citR	**1**	3.5
citAyhdF	**1**	2.1-2.1
yhdI	**0**	0
yhdKLM	**1**	0-0-3.5
yhdN	**1**	2.1
yhdP	**0**	0
yheM	**1**	3.2
sspB	**1**	1.8
yheBA	**0**	0-0
yhaX	**0**	0
hemZ	**1**	2.5
yhaUTS	**0**	0-0-0
yhaR	**0**	0
prsAyhaK	**0**	0-0
hpr	**1**	3.5
yhaGserC	**0**	0-0
pbpF	**1**	1.1
hemY	**1**	2.5
yhfEF	**0**	0-0
yhfM	**0**	0
aprE	**1**	2.2
yhfP	**0**	0
yhfUV	**1**	0-1.5
yhfW	**1**	1.4
yhzC	**0**	0
comK	**1**	3.5
yhxD	**1**	2.1
yhjGH	**0**	0-0
yhjI	**0**	0
yhjJKL	**0**	0-0-0
yhjM	**0**	0
yhjR	**0**	0

Appendix B. (Continued)

addBA	**1**	3.3-3.3
sbcD	**1**	3.3
yisCDEFGH	**1**	1.9-1.9-1.9-1.9-1.9-1.9
yisI	**0**	0
yisK	**0**	0
wprA	**1**	1.1
yisO	**1**	2.2
degAyisS	**0**	0-0
yisZyitAB	**1**	2.7-2.7-2.7
yitK	**0**	0
yitM	**0**	0
argCJBDcarABargF	**1**	2.2-2.2-2.2-2.2-2.2-2.2-2.2
medyjzA	**1**	1.10-1.10
yjaXY	**1**	2.4-2.4
appDFAB	**1**	0-0-1.2-1.2
trpS	**1**	3.7
oppAB	**1**	1.2-0
yjbCD	**0**	0-0
mecA	**1**	1.1
yjbFG	**0**	0-0
yjbJ	**1**	4.4
tenAIyjbRSTUV	**2**	3.5-3.5-2.2-2.5-2.5-2.5-2.5
yjbX	**0**	0
cotZYXW	**1**	1.8-1.8-1.8-1.8
yjcIJ	**1**	2.2-2.2
yjcM	**0**	0
yjcPQR	**0**	0-0-0
yjcSyjdA	**1**	0-2.4
yjdB	**0**	0
yjdJ	**0**	0
cotT	**1**	1.8
yjfB	**0**	0
yjgB	**0**	0
yjgCD	**0**	0-0
yjmAB	**2**	2.1-1.2
yjmEF	**1**	2.1-2.1
yjoBrapA	**1**	1.7-1.8
xkdA	**1**	4.4
xre	**1**	3.5
xkdBCDxtrA	**1**	4.4-4.4-4.4-4.4
xtmBxkdEFG	**1**	4.4-4.4-4.4-4.4
xkdIJ	**1**	4.4-4.4
xkdKM	**1**	4.4-4.4
xkdNO	**0**	0-0
xkdRS	**0**	0-0
ykaAykbA	**0**	0-0
ykcBC	**0**	0-0
htrA	**1**	4.1
dppABCDEykfABCD	**2**	3.8-1.2-1.2-1.2-1.2-0-0-0-0
ykgA	**0**	0
ykhA	**0**	0
hmp	**1**	1.4

Appendix B. (Continued)

ykjA	**0**	0
ykkA	**0**	0
proBA	**0**	0-0
yklA	**0**	0
ykzA	**1**	4.2
yknA	**1**	2.3
ykoG	**0**	0
tnrA	**1**	3.5
ykzBykoL	**0**	0-0
ykoM	**0**	0
ykoVW	**0**	0-0
ykoZ	**1**	3.5
sspD	**1**	1.8
ykrL	**0**	0
ykrQdat	**2**	1.3-3.2
ykrTU	**0**	0-0
ykrV	**1**	2.2
ykrWXYZ	**1**	2.1-0-0-0
spo0E	**1**	1.8
motBA	**1**	1.5-1.5
clpE	**1**	4.1
ykvI	**0**	0
ykvJK	**0**	0-0
ykvPQ	**1**	0-2.1
ykvS	**0**	0
ykvUV	**1**	1.8-1.4
ykvW	**0**	0
glcTptsG	**1**	0-1.2
splAB	**2**	3.5-1.8
mcpC	**1**	1.5
ykwC	**0**	0
kinA	**1**	1.3
cheV	**1**	1.5
ykuD	**1**	1.8
ykuJKykzF	**0**	0-0-0
ykuL	**0**	0
ykuM	**1**	3.5
ykuNOP	**1**	1.4-0-1.4
ykuV	**0**	0
ykuW	**1**	1.1
yknT	**1**	1.8
yknU	**0**	0
yknWX	**0**	0-0
yknYZ	**0**	0-0
fruRBA	**2**	3.5-2.1-0
abh	**1**	3.5
kinC	**1**	1.3
pdhBCD	**1**	2.1-2.1-2.1
slp	**0**	0
yktA	**0**	0
nprE	**1**	2.2
ylaABCD	**1**	0-0-3.5-0

Appendix B. (Continued)

ylaJ	**0**	0
ylaK	**0**	0
ylaM	**0**	0
ctaA	**1**	1.4
ctaB	**1**	1.4
ctaCDEF	**1**	1.4-1.4-1.4-1.4
ylbBCD	**1**	2.1-0-0
ylbJ	**0**	0
ylbKL	**0**	0-0
ylbQyllA	**0**	0-0
yllBylxAftsL	**1**	0-0-1.7
spoVD	**1**	1.1
murE	**1**	1.1
mraYmurDspoVEmurG	**1**	1.1-1.1-1.8-1.1
ftsA	**1**	1.7
bprspoIIGAsigEG	**2**	0-1.8-3.5-3.5
ylmAB	**0**	0-0
ylmDEF	**0**	0-0-0
ileS	**0**	0
pyrRP	**2**	3.5-1.2
pyrDF	**1**	2.3-2.3
cysH	**1**	2.2
ylnF	**1**	2.5
yloCDH	**0**	0-0-0
yloIpriA	**0**	0-0
spoVM	**1**	1.8
yloW	**0**	0
rncSsmcftsY	**2**	3.6-3.4-1.6
ylqB	**0**	0
ylxMffh	**1**	0-1.6
rpsPylqCDEtrmDrplS	**1**	3.7-0-0-0-0-0
ylqFrnhylqGH	**0**	0-0-0-0
sucCD	**1**	2.1-2.1
smf	**1**	1.1
codY	**1**	3.5
fliIJylxFfliKylxGflgE	**1**	1.5-1.5-0-1.5-1.5-1.5
fliQR	**1**	1.5-1.5
smbAfrryluAcdsAyluBC	**0**	0-0-0-0-0-0
proSpolC	**0**	0-0
nusAylxRQinfB	**1**	3.5-0-3.7-3.7
ylxPrbfA	**1**	0-3.7
spoVFAVFBasddapG	**2**	1.8-1.8-2.2-2.2
dapAymfA	**1**	2.2-0
ymfFGHI	**1**	0-0-0-2.4
ymfJ	**0**	0
recA	**1**	3.3
pbpX	**1**	1.1
spoVS	**1**	1.8
ymcBA	**0**	0-0
cotE	**1**	1.8
pksI	**0**	0
pksJKLMNP	**0**	0-0-0-0-0-0

Appendix B. (Continued)

pksS	**0**	0
ymaE	**0**	0
miaA	**0**	0
ymaHymzC	**0**	0-0
cwlC	**1**	1.1
spoVK	**1**	1.8
ynbB	**0**	0
glnRA	**2**	3.5-2.2
ynzFG	**0**	0-0
xynB	**1**	2.1
xylR	**1**	3.5
xylAB	**1**	2.1-2.1
yncB	**0**	0
yncC	**0**	0
cotC	**1**	1.8
yndB	**0**	0
yndEF	**1**	1.9-1.9
yndL	**1**	4.4
yndN	**1**	4.2
ynzC	**0**	0
tkt	**1**	2.1
yneF	**0**	0
ynzD	**0**	0
ccdA	**1**	1.4
cotM	**1**	1.8
citByneN	**2**	2.1-1.4
tlp	**1**	1.8
grlBA	**1**	0-3.4
ynfC	**1**	3.4
alsT	**0**	0
bglC	**0**	0
yngC	**0**	0
yngDEFGHI	**1**	0-2.1-2.4-2.4-2.5-2.4
yngL	**0**	0
ppsB	**0**	0
pbpyoxA	**1**	1.1-0
yoeA	**0**	0
gltBA	**2**	2.2-3.5
gltC	**0**	0
proJH	**0**	0-0
rtp	**1**	2.4
yoxByoaA	**1**	3.8-0
yoaBCD	**0**	0-0-0
yoaE	**0**	0
yoaF	**0**	0
yoaG	**0**	0
yoaH	**1**	1.5
yoaJ	**1**	2.1
yoaN	**0**	0
yoaU	**0**	0
yoaW	**0**	0
yoaZ	**0**	0

Appendix B. (Continued)

yobB	**0**	0
yozIyobE	**0**	0-0
phrK	**1**	1.8
yobIJ	**0**	0-0
yobL	**0**	0
yobO	**1**	4.4
yobW	**1**	1.8
yocE	**1**	2.4
yocFG	**2**	1.3-3.5
yocH	**1**	1.1
yocJ	**1**	1.4
yocKLM	**1**	4.1-0-0
yozNyocN	**0**	0-0
yozO	**0**	0
odhBA	**1**	2.1-2.1
yojA	**0**	0
yodDE	**0**	0-0
yodF	**0**	0
yodHI	**0**	0-0
yodL	**0**	0
yodPargE	**0**	0-0
cgeEDC	**1**	0-1.8-1.8
cgeAB	**1**	1.8-1.8
yotKJIH	**0**	0-0-0-0
sspC	**0**	0
yosTS	**0**	0-0
yosP	**0**	0
yosONMLKJIHG	**0**	0-0-0-0-0-0-0-0
yosDC	**0**	0-0
yosB	**0**	0
yorZYXWVmtbP-yorTSRQPONMLKJIHGFEDCBAyoqZY	**0**	0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0
yoqLKJ	**0**	0-0-0
yoqDCBA	**0**	0-0-0-0
yopR	**0**	0
yopN	**0**	0
yopKJIHGFEDC	**0**	0-0-0-0-0-0-0-0
yonV	**0**	0
yonU	**0**	0
yonR	**0**	0
yomL	**0**	0
yomK	**0**	0
yolKJIsunTA	**3**	3.9-2.1-3.9-0-4.3
yolF	**0**	0
yolBA	**0**	0-0
yokJI	**0**	0-0
yokFE	**0**	0-0
yppQ	**0**	0
yppP	**1**	3.5
ypmTSRQ	**0**	0-0-0-0
ypmPilvA	**1**	2.2-0
yphPilvD	**1**	2.2-0

Appendix B. (Continued)

bsaA	**1**	2.2
metBypfP	**0**	0-0
degR	**0**	0
ypbR	**0**	0
pbuXxpt	**2**	1.2-2.3
kdgTAKR	**3**	1.2-2.1-2.1-3.5
yptA	**0**	0
cotD	**0**	0
yppD	**1**	1.8
yppBponA	**2**	3.3-1.1
panCB	**0**	0-0
birA	**0**	0
ypjB	**0**	0
qcrCBA	**1**	1.4-1.4-1.4
ypiBA	**0**	0-0
trpFC	**1**	2.2-2.2
aroHBFcheR	**0**	0-0-0-0
mtrAhbs	**1**	0-3.4
spoIVA	**1**	1.8
gpsAyphC	**0**	0-0
yphB	**1**	1.8
ypfD	**0**	0
cmk	**0**	0
ypfBA	**0**	0-0
ypeBsleB	**1**	1.8-1.9
ypbH	**0**	0
ypbF	**0**	0
aroC	**0**	0
ypuNsigX	**1**	3.5-3.5
resEDCBA	**2**	1.3-3.5-1.4-1.4-1.4
spmBAdacB	**1**	1.8-1.8-1.1
ypuI	**0**	0
ypuHG	**0**	0-0
ribH	**1**	2.5
sipS	**1**	1.6
ypuCB	**0**	0-0
ppiB	**1**	3.8
ypuA	**0**	0
spoVADVACVABVAAsigFspoIIABIIAAdacF	**2**	1.8-1.8-1.8-1.8-3.5-1.8-1.8-1.1
pnpdrm	**1**	2.3-2.3
yqkL	**1**	3.5
spoIIM	**1**	1.8
yqkKJ	**0**	0-0
ansBA	**1**	2.2-2.2
ansR	**1**	3.5
yqjXW	**1**	0-3.2
yqjTS	**1**	0-2.5
yqjED	**0**	0-0
yqjBA	**0**	0-0
yqiYX	**0**	0-0
bfmBBBABBAAyqiVUTS	**1**	2.4-2.4-2.4-2.4-2.4-2.4-2.4
yqiR	**1**	3.5

Appendix B. (Continued)

yqiQmmgE	**1**	2.1-0
mmgCBA	**1**	2.4-2.4-2.4
yqiG	**1**	1.4
spo0AIVB	**2**	3.5-1.8
yqiDCBfolD	**1**	0-0-0-2.5
yqhZ	**1**	3.5
accCB	**1**	2.4-2.4
spoIIIAHIIIAGIIIAFIIIAEIIIADIIIACIIIABIII	**1**	1.8-1.8-1.8-1.8-1.8-1.8-1.8-1.8
efpyqhTS	**0**	0-0-0
yqhQP	**0**	0-0
yqhN	**0**	0
sinI	**1**	1.8
cotN	**1**	1.8
yqzG	**0**	0
comGGGFGEGDGCGBGA	**1**	1.10-1.10-1.10-1.10-1.10-1.10-1.10
yqxL	**0**	0
yqhA	**1**	4.1
yqgZ	**0**	0
yqgY	**0**	0
yqgS	**0**	0
yqgQP	**0**	0-0
yqgML	**0**	0-0
yqzDC	**0**	0-0
yqfZY	**1**	0-1.1
yqfON	**0**	0-0
sigAdnaG	**1**	3.5-3.1
yqxD	**0**	0
glySQ	**1**	3.7-3.7
cdd	**1**	2.3
dgkAyqfG	**0**	0-0
yqfDC	**1**	1.8-0
yqfBAyqeZ	**0**	0-0-0
yqeVUTdnaJKgrpEhrcAhemNlepA	**3**	0-0-3.8-4.1-3.9-4.1-3.5-2.5-3.7
yqxAspoIIPgpr	**1**	0-1.8-1.9
comECEBEA	**1**	1.10-1.10-1.10
yqeGF	**0**	0-0
yqeB	**0**	0
nucB	**1**	2.3
spoIVCB	**1**	3.5
spoIVCA	**1**	1.8
yqcKJ	**1**	0-3.5
phrE	**1**	1.8
yqcG	**0**	0
spoIIIC	**1**	3.5
yrkON	**0**	0-0
yrkLK	**0**	0-0
yrkC	**0**	0
brnQazlDCB	**1**	0-0-0-3.5
yrpD	**0**	0
yrpE	**0**	0
yraO	**0**	0
yraN	**0**	0

Appendix B. (Continued)

csn	**0**	0
yraF	**1**	1.8
adhB	**1**	2.1
levGFED	**0**	0-0-0-0
yrhK	**0**	0
yrhJI	**2**	4.2-3.5
yrhH	**1**	2.1
yrhG	**0**	0
yrhF	**0**	0
yrhED	**0**	0-0
yrhBAyrrUT	**1**	2.2-2.2-2.3-0
alaS	**0**	0
glnHMP	**1**	1.2-1.2-1.2
aspShisS	**0**	0-0
yrvJ	**1**	1.1
relAaptyrvED	**2**	2.3-2.3-3.3-0
secDF	**0**	0
spoVB	**1**	1.8
csbXyrbE	**0**	0-0
yrzF	**0**	0
yrbC	**0**	0
yrbBA	**1**	1.8-1.8
nadCB	**1**	2.5-2.5
nifSyrxA	**2**	2.5-3.5
pheAB	**1**	2.2-2.2
spo0B	**1**	1.8
spoIVFBIVFA	**1**	1.8-1.8
minDC	**1**	1.7-1.7
comC	**1**	1.1
valSysxE	**1**	3.7-0
hemLBDCXA	**1**	2.5-2.5-2.5-2.5-2.5-2.5
ysxD	**0**	0
ysxClonA	**1**	4.6-4.1
lonB	**1**	4.1
tig	**1**	3.9
leuDCBAilvCNB	**1**	2.2-2.2-2.2-2.2-2.2-2.2-2.2-2.2
ysnD	**0**	0
ysnE	**1**	2.2
ysnF	**0**	0
ysnBA	**0**	0-0
gerM	**1**	1.9
racEysmBgerE	**1**	0-0-3.5
sdhA	**1**	2.1
lysC	**1**	2.2
uvrC	**1**	3.2
trxA	**1**	1.4
xsa	**1**	2.1
etfABysiBA	**3**	1.4-1.4-2.4-3.5
lcfA	**1**	2.4
pheTS	**1**	3.7-3.7
ysfA	**1**	1.8
araQP	**0**	0-0

Appendix B. (Continued)

araD	**1**	2.1
abnA	**1**	2.1
ysdC	**0**	0
ysdB	**0**	0
yscB	**0**	0
thrS	**1**	3.7
ytxC	**0**	0
ytcFgapB	**1**	2.2-2.1
ytaGFmutMpolA	**0**	0-0-0-0
phoRP	**2**	1.3-3.5
citCZ	**1**	2.1-2.1
ytvI	**0**	0
accAyttI	**1**	2.4-0
ytpI	**0**	0
ytnMhipOribRytnJIytmOhisPytmMLKJI	**2**	0-2.3-2.5-4.2-0-4.2-0-0-0-0-0-0
ytkL	**0**	0
argHG	**0**	0-0
ackA	**1**	2.1
ytfJI	**0**	0-0
yteJI	**1**	0-3.8
sspA	**1**	1.8
yttP	**0**	0
ytsP	**0**	0
rpsD	**1**	3.7
tyrS	**1**	3.7
acsA	**1**	2.1
acuABC	**1**	2.1-2.1-2.1
ytxEDccpA	**2**	1.5-1.5-3.5
aroA	**1**	2.2
ytxJHG	**1**	4.1-4.1-4.1
ytpRQP	**0**	0-0-0
malSytnP	**1**	2.1-0
ytlQ	**0**	0
ytiP	**0**	0
ythQP	**0**	0-0
ytzFGytgP	**0**	0-0-0
yteVUTSR	**0**	0-0-0-0-0
ytcQ	**0**	0
ytbQbioIBDFAW	**1**	0-2.5-2.5-2.5-2.5-2.5-2.5
msmRE	**0**	0-0
leuS	**1**	3.7
yttB	**1**	1.2
yttA	**0**	0
ytsD	**0**	0
ytrFEDCBA	**3**	1.2-0-2.1-2.1-0-3.5
ytqAB	**0**	0-0
ytpB	**0**	0
ytnAasnB	**1**	0-2.2
ytlABCD	**0**	0-0-0-0
dps	**1**	4.1
ytiBA	**0**	0-0
ythA	**0**	0

Appendix B. (Continued)

ythB	**0**	0
ytgBA	**1**	0-1.2
ytfDmenEB	**1**	2.5-2.5-2.5
ytcAB	**0**	0-0
ytxOcotSytxN	**1**	0-1.8-1.8
glgPADCB	**1**	2.1-2.1-2.1-2.1-2.1
yuaIGF	**1**	0-4.6-0
gbsBA	**0**	0-0
yuxG	**1**	1.5
tlpB	**1**	1.5
mcpAtlpA	**1**	1.5-1.5
mcpB	**1**	1.8
yugO	**0**	0
yugNM	**1**	0-2.1
pgi	**1**	2.1
yuzA	**1**	3.6
kinB	**1**	1.8
pbpDyuxK	**0**	0-0
yufOPQ	**0**	0-0-0
yufTUVDCB	**1**	4.2-4.2-4.2-4.2-4.2-4.2
yuxO	**0**	0
comA	**1**	3.5
comPX	**1**	1.3-1.10
degQ	**1**	4.1
yuzC	**0**	0
yuxH	**0**	0
yueKJIH	**1**	2.5-0-0-0
yueG	**0**	0
yuzE	**0**	0
yueCByukABC	**0**	0-0-0-0-0
ald	**1**	2.2
yuxIyukJ	**0**	0-0
yukLMdhbFBECA	**1**	0-0-2.5-2.5-2.5-2.5-2.5
yuiI	**0**	0
yuiH	**0**	0
yumC	**0**	0
yumD	**1**	2.3
paiB	**0**	0
yutJ	**0**	0
yuxLthrB	**0**	0-0
homyutH	**0**	0-0
yunB	**0**	0
yunGH	**1**	0-2.3
yunIJKLM	**2**	3.5-0-0-2.3-2.3
yurBCDEF	**1**	2.3-2.3-2.3-2.3-2.3
yurK	**0**	0
yurL	**1**	2.2
yurMN	**0**	0-0
yurS	**0**	0
yurUVWXY	**0**	0-0-0-0-0
yusA	**0**	0
yusKL	**1**	2.4-2.4

Appendix B. (Continued)

yusM	**1**	2.2
yusV	**0**	0
mrgA	**1**	4.1
yvqAB	**2**	3.5-1.3
citG	**1**	2.1
gerAAABAC	**1**	1.9-1.9-1.9
yvqHI	**0**	0-0
yvrE	**1**	4.6
yvrP	**0**	0
fhuCGB	**1**	0-0-1.2
fhuD	**0**	0
yvsH	**1**	1.8
yvgO	**0**	0
yvgQR	**1**	2.7-2.7
yvgUV	**1**	3.9-3.9
yvgZ	**0**	0
yvaA	**0**	0
yvaDEF	**0**	0-0-0
yvaN	**0**	0
yvaQ	**0**	0
yvaX	**0**	0
opuCDCCCBCA	**0**	0-0-0-0
enopgmtpipgkgapyvbQ	**2**	2.1-2.1-2.1-2.1-2.1-3.5
araE	**0**	0
araR	**1**	3.5
yvbT	**0**	0
yvbU	**0**	0
yvbW	**0**	0
yvfOlacAyvfMLK	**1**	2.1-2.1-0-0-0
yvfH	**0**	0
yveTSRQPONMLK	**2**	2.1-0-4.1-4.1-2.1-2.1-4.1-4.1-4.1-4.1
yveGFracXpbpE	**2**	0-0-2.2-1.1
sacByveBA	**1**	2.1-2.1-0
yvdS	**0**	0
yvdO	**0**	0
clpP	**1**	4.1
yvdMLKJIHGFE	**0**	0-0-0-0-0-0-0-0-0
yvcT	**0**	0
yvcSR	**0**	0-0
yvcBA	**0**	0-0
hisIFAHBDGZ	**0**	0-0-0-0-0-0-0-0
yvlDCBA	**0**	0-0-0-0
uvrB	**1**	3.2
csbA	**1**	4.6
yvzD	**0**	0
yvyD	**1**	3.5
fliTSDyvyC	**1**	1.5-1.5-1.5-1.5
hag	**1**	1.5
yviFEflgLKyvyGflgMyvyF	**1**	0-0-1.5-1.5-1.5-1.5-1.5
comFC	**1**	1.1
degUS	**2**	3.5-1.3
tuaHGFEDCBA	**1**	1.1-1.1-1.1-1.1-1.1-1.1-1.1-1.1-1.1

Appendix B. (Continued)

lytCBA	**1**	1.1-1.1-1.6
lytR	**1**	3.5
gtaB	**1**	1.1
ggaB	**1**	1.1
ggaA	**1**	1.1
tagHG	**0**	0-0
tagFED	**1**	1.1-1.1-1.1
tagAB	**1**	1.1-1.1
tagC	**1**	1.1
lytD	**1**	1.1
pmi	**1**	2.1
gerBABBBC	**1**	1.9-1.9-1.9
ywtFE	**1**	3.5-0
ywtCBAywsC	**1**	0-4.1-4.1-4.1
rbsRKDACBywsB	**3**	3.5-2.1-1.2-1.2-1.2-1.2-0
alsDS	**1**	2.1-2.1
ywrJcotB	**1**	0-1.8
cotH	**1**	1.8
cotG	**1**	1.8
ywrE	**0**	0
ywrD	**1**	2.2
ywrCB	**0**	0-0
ywqM	**0**	0
ywqFE	**1**	2.1-0
glcRywpH	**1**	3.5-3.1
rapD	**1**	1.8
flhPO	**1**	1.5-1.5
mblspoIIIDusd	**2**	1.1-3.5-1.8
nrgA	**1**	1.2
ywoA	**0**	0
ywnJ	**0**	0
spoIIQ	**1**	1.8
ywnH	**0**	0
mta	**1**	3.5
ureCBA	**1**	2.2-2.2-2.2
ywmGF	**0**	0-0
narA	**1**	2.5
ywmE	**0**	0
spoIID	**1**	1.8
atpCDGAH	**1**	1.4-1.4-1.4-1.4-1.4
atpEB	**1**	1.4-1.4
uppglyA	**1**	2.3-2.2
ywlGF	**1**	0-2.1
ywlD	**0**	0
ywlBspoIIR	**1**	0-1.8
ywkBA	**1**	0-2.1
tdkrpmE	**1**	0-3.7
rhoywjImurZywjHfbaA	**1**	0-2.1-0-2.1-2.1
spo0F	**1**	3.5
ctrArpoE	**2**	2.3-3.5
ywjC	**0**	0
ywjA	**0**	0

Appendix B. (Continued)

ywiE	**0**	0
narIJHG	**1**	1.4-1.4-1.4-1.4
fnrnarKargS	**2**	3.5-1.2-3.7
phrF	**1**	1.8
ywhG	**1**	2.2
ywhE	**1**	1.1
ywhB	**0**	0
ywhA	**1**	3.5
thrZ	**1**	3.7
ywfN	**1**	1.8
ywfKpta	**2**	3.5-2.1
ywfI	**0**	0
ywfA	**0**	0
rocA	**1**	2.2
yweB	**1**	2.2
yweA	**0**	0
spsKJIGF	**1**	1.8-1.8-1.8-1.8-1.8
ywdJI	**0**	0-0
ungywdF	**2**	2.1-3.2
ywdAsacAP	**1**	2.5-0-2.1
ywcHG	**2**	2.2-4.2
qoxDCBA	**1**	0-1.4-0-0
galTK	**1**	0-2.1
ywcA	**0**	0
ywbONML	**0**	0-0-0-0
thiCKywbI	**1**	0-2.5-2.5
ywbHGFE	**1**	3.5-0-0-0
ywbBA	**0**	0-0
epr	**0**	0
sacXY	**1**	2.2-2.1
gspA	**1**	3.5
tyrZ	**0**	0
ywaC	**1**	2.2
dltABCDE	**2**	2.5-1.1-1.1-1.1-1.1
licHACBR	**1**	2.2-2.1-0-0-0
yxzFyxlJ	**1**	3.5-0
katX	**1**	3.2
yxlGFEDCsigY	**0**	0-0-0-0-0-0
yxkO	**0**	0
cydDCBA	**1**	0-0-0-1.4
yxzE	**1**	4.1
yxkHmsmXyxkF	**0**	0-0-0
yxkD	**1**	2.1
yxkC	**0**	0
galE	**0**	0
yxjI	**0**	0
yxjHG	**0**	0-0
yxjFEDC	**1**	0-2.1-2.4-2.4
yxjA	**0**	0
yxiSkatB	**0**	0-0
yxiQ	**1**	4.2
bglS	**1**	2.1

Appendix B. (Continued)

deaDyxiMLKJIyxzGyxiHGyxzCyxiFyxxG	**2**	0-3.6-2.1-0-0-0-0-0-0-0-0-0
yxiE	**0**	0
bglHPyxxE	**2**	0-2.1-1.2
yxiB	**0**	0
hutPHUIGM	**2**	2.1-3.5-2.2-2.2-2.2-2.2
nupCdradeoR	**2**	2.3-1.2-2.3
yxeED	**0**	0-0
yxeC	**0**	0
yxeB	**0**	0
yxdMLK	**0**	0-0-0
iolIH	**1**	2.1-2.1
iolRS	**2**	2.2-3.5
htpG	**0**	0
yxbG	**0**	0
yxbDC	**1**	2.1-0
yxnBasnH	**0**	0-0
yxaLK	**1**	0-3.8
yxaHG	**0**	0-0
yxaBA	**0**	0-0
gntRKPZ	**3**	3.5-2.1-1.2-2.1
ahpCF	**1**	4.2-4.2
bglA	**1**	2.1
yydK	**1**	3.5
yydJ	**0**	0
fbp	**1**	2.1
yydD	**0**	0
yydC	**0**	0
yycR	**1**	2.1
rapG	**1**	1.8
rocFED	**2**	2.2-1.2-2.2
rocR	**1**	3.5
yyxA	**1**	4.1
yycJI	**0**	0-0
purA	**1**	2.3
yycD	**0**	0
yycCB	**0**	0-0
cotF	**1**	1.8
yybP	**0**	0
yybNMLKJ	**0**	0-0-0-0-0
yybIH	**0**	0-0
yybF	**1**	1.2
tetBL	**1**	0-4.2
yyaJ	**0**	0
rpsFyyaF	**2**	3.7-4.6
yyaC	**0**	0
yyaAgidBAthdF	**2**	1.8-1.7-1.7-4.2
spoIIIJrnpA	**1**	1.8-0
rpmH	**1**	3.7

Appendix C.

Gene Pair	
dnaA	*dnaN*
dnaN	*yaaA*
yaaA	*recF*
recF	*yaaB*
yaaB	*gyrB*
gyrB	*gyrA*
guaB	*dacA*
dacA	*yaaD*
yaaD	*yaaE*
yaaE	*serS*
dck	*dgk*
dgk	*yaaH*
yaaH	*yaaI*
dnaX	*yaaK*
yaaK	*recR*
recR	*yaaL*
yaaL	*bofA*
csfB	*xpaC*
xpaC	*yaaN*
yaaN	*yaaO*
yaaO	*tmk*
tmk	*yaaQ*
yaaQ	*yaaR*
yaaR	*holB*
holB	*yaaT*
yaaT	*yabA*
yabA	*yabB*
yabB	*yazA*
yazA	*yabC*
metS	*yabD*
yabD	*yabE*
yabE	*rnmV*
rnmV	*ksgA*
ksgA	*yabG*
yabG	*veg*
veg	*sspF*
sspF	*ispE*
ispE	*purR*
purR	*yabJ*
yabJ	*spoVG*
spoVG	*gcaD*
gcaD	*prs*
prs	*ctc*
ctc	*spoVC*
spoVC	*yabK*
yabK	*mfd*
mfd	*spoVT*
spoVT	*yabM*
yabM	*yabN*
yabN	*yabO*
yabO	*yabP*
yabP	*yabQ*
yabQ	*divIC*
divIC	*yabR*
spoIIE	*yabS*
yabS	*yabT*
yabT	*yacA*
yacA	*hprT*
hprT	*ftsH*
ftsH	*yacB*

Gene Pair	
yacB	*yacC*
yacC	*yacD*
yacD	*cysK*
cysK	*pabB*
pabB	*pabA*
pabA	*pabC*
pabC	*sul*
sul	*folB*
folB	*folK*
folK	*yazB*
yazB	*yacF*
yacF	*lysS*
ctsR	*mcsA*
mcsA	*mcsB*
mcsB	*clpC*
clpC	*radA*
radA	*yacK*
yacK	*yacL*
yacL	*yacM*
yacM	*yacN*
yacN	*gltX*
gltX	*cysE*
cysE	*cysS*
cysS	*yazC*
yazC	*yacO*
yacO	*yacP*
yacP	*sigH*
sigH	*rpmGB*
rpmGB	*secE*
secE	*nusG*
nusG	*rplK*
rplK	*rplA*
rplA	*rplJ*
rplJ	*rplL*
rplL	*ybxB*
ybxB	*rpoB*
rpoB	*rpoC*
rpoC	*ybxF*
ybxF	*rpsL*
rpsL	*rpsG*
rpsG	*fusA*
fusA	*tufA*
tufA	*ybaC*
ybaC	*rpsJ*
rpsJ	*rplC*
rplC	*rplD*
rplD	*rplW*
rplW	*rplB*
rplB	*rpsS*
rpsS	*rplV*
rplV	*rpsC*
rpsC	*rplP*
rplP	*rpmC*
rpmC	*rpsQ*
rpsQ	*rplN*
rplN	*rplX*
rplX	*rplE*
rplE	*rpsN*
rpsN	*rpsH*
rpsH	*rplF*

Gene Pair	
rplF	*rplR*
rplR	*rpsE*
rpsE	*rpmD*
rpmD	*rplO*
rplO	*secY*
secY	*adk*
adk	*map*
map	*infA*
infA	*rpmJ*
rpmJ	*rpsM*
rpsM	*rpsK*
rpsK	*rpoA*
rpoA	*rplQ*
rplQ	*ybxA*
ybxA	*ybaE*
ybaE	*ybaF*
ybaF	*truA*
truA	*rplM*
rplM	*rpsI*
rpsI	*ybaJ*
ybaJ	*ybaK*
ybaK	*cwlD*
cwlD	*ybaL*
ybaR	*ybaS*
ybbA	*feuC*
feuC	*feuB*
feuB	*feuA*
feuA	*ybbB*
ybbB	*ybbC*
ybbC	*ybbD*
ybbD	*ybbE*
ybbE	*ybbF*
ybbF	*ybbH*
ybbH	*ybbI*
ybbI	*ybbJ*
ybbJ	*ybbK*
sigW	*ybbM*
ybbM	*ybbP*
ybbP	*ybbR*
ybbR	*ybbT*
ybbT	*glmS*
glmS	*ybbU*
adaA	*adaB*
adaB	*ndhF*
ndhF	*ybcC*
ybcC	*ybcD*
ybcD	*ybcF*
ybcF	*ybcH*
ybcH	*ybcI*
ybcI	*ybcL*
ybcL	*ybcM*
ybcM	*ybcO*
ybcO	*ybcP*
ybcP	*ybcS*
ybcS	*ybcT*
ybcT	*ybdA*
ybdA	*ybdB*
ybdB	*ybdD*
ybdD	*ybdE*
ybdE	*ybdG*

Gene Pair	
ybdG	*ybdJ*
ybdJ	*ybdK*
ybdK	*ybdL*
ybdM	*ybdN*
ybdO	*ybxG*
ybxG	*csgA*
csgA	*ybxH*
glpQ	*glpT*
ybeF	*ybfA*
ybfA	*ybfB*
ybfE	*ybfF*
ybfF	*ybfG*
ybfG	*ybfH*
ybfH	*ybfI*
purT	*mpr*
mpr	*ybfJ*
ybfJ	*ybfK*
ybfK	*pssA*
pssA	*ybfM*
ybfM	*psd*
psd	*ybfN*
ybfN	*ybfO*
ybfO	*ybfP*
ybfP	*ybfQ*
gltP	*gamP*
gamP	*gamA*
ybgA	*ybgB*
ybgB	*ybgE*
ybgF	*ybgG*
ybgG	*ybgH*
ybgH	*ybgJ*
ycbA	*ycbB*
ycbB	*ycbC*
ycbC	*ycbD*
ycbD	*ycbE*
ycbE	*ycbF*
ycbF	*ycbG*
ycbG	*ycbH*
ycbH	*ycbJ*
ycbJ	*yczA*
yczA	*ycbK*
ycbK	*ycbL*
ycbL	*ycbM*
ycbM	*ycbN*
ycbN	*ycbO*
cwlJ	*ycbR*
ycbR	*phoD*
phoD	*tatAD*
tatAD	*tatCD*
lmrB	*lmrA*
yccC	*lip*
yccG	*yccH*
natA	*natB*
natB	*yccK*
ycdB	*ycdC*
rapJ	*ycdF*
ycdF	*ycdG*
ycdG	*ycdH*
ycdH	*ycdI*
ycdI	*yceA*

Gene Pair	
yceC	*yceD*
yceD	*yceE*
yceE	*yceF*
yceF	*yceG*
yceG	*yceH*
yceH	*yceI*
yceJ	*yceK*
opuAA	*opuAB*
opuAB	*opuAC*
ycgA	*ycgB*
ycgB	*amyE*
amyE	*ldh*
ldh	*lctP*
ycgE	*ycgF*
ycgF	*ycgG*
ycgI	*nadE*
ycgK	*cah*
cah	*ycgL*
ycgL	*ycgM*
ycgM	*ycgN*
ycgN	*ycgO*
ycgO	*ycgP*
ycgQ	*ycgR*
ycgR	*ycgS*
nasF	*nasE*
nasE	*nasD*
nasD	*nasC*
nasC	*nasB*
nasA	*yciA*
yciA	*yciB*
yciB	*yciC*
yckA	*yckB*
yckC	*yckD*
yckD	*yckE*
nin	*nucA*
nucA	*tlpC*
tlpC	*hxlB*
hxlB	*hxlA*
hxlR	*srfAA*
srfAA	*srfAB*
srfAB	*comS*
comS	*srfAC*
srfAC	*srfAD*
srfAD	*ycxA*
ycxB	*ycxC*
sfp	*yczE*
yczE	*yckI*
yckI	*yckJ*
yckJ	*yckK*
yckK	*yclA*
yclB	*yclC*
yclC	*yclD*
yclD	*yclE*
gerKA	*gerKC*
gerKC	*gerKB*
yclH	*yclI*
yclJ	*yclK*
yclK	*rapC*
rapC	*phrC*
yclN	*yclO*

Gene Pair	
yclO	*yclP*
yclP	*yclQ*
ycnB	*ycnC*
ycnC	*ycnD*
ycnD	*ycnE*
gabT	*gabD*
gabD	*glcU*
glcU	*gdh*
ycnI	*ycnJ*
ycnJ	*ycnK*
ycnL	*mtlA*
mtlA	*mtlD*
mtlD	*ycsA*
ycsA	*sipU*
ycsD	*ycsE*
ycsE	*ycsF*
ycsF	*ycsG*
ycsG	*ycsI*
ycsI	*kipI*
kipI	*kipA*
kipA	*kipR*
kipR	*ycsK*
ycsK	*yczI*
pbpC	*ycsN*
ycsN	*mtlR*
mtlR	*ydaB*
ydaD	*ydaE*
ydaE	*ydaF*
ydaF	*ydaG*
ydaG	*ydaH*
lrpC	*topB*
topB	*ydaJ*
ydaJ	*ydaK*
ydaK	*ydaL*
ydaL	*ydaM*
ydaM	*ydaN*
ydaN	*ydaO*
ydaO	*mutT*
mutT	*ydaP*
ydaP	*ydaQ*
mntH	*ydaS*
ydaS	*ydaT*
ydbA	*gsiB*
gsiB	*ydbB*
ydbB	*ydbC*
ydbD	*dctB*
dctS	*dctR*
dctR	*dctP*
dctP	*ydbI*
ydbI	*ydbJ*
ydbJ	*ydbK*
ydbK	*ydbL*
ydbL	*ydbM*
ddl	*murF*
murF	*ydbR*
ydbR	*ydbS*
ydbS	*ydbT*
acpS	*ydcC*
ydcC	*alr*
alr	*ydcD*

Appendix C. (Continued)

Gene Pair	
ydcD	*ydcE*
ydcE	*rsbR*
rsbR	*rsbS*
rsbS	*rsbT*
rsbT	*rsbU*
rsbU	*rsbV*
rsbV	*rsbW*
rsbW	*sigB*
sigB	*rsbX*
rsbX	*ydcF*
ydcF	*ydcG*
ydcG	*ydcH*
ydcH	*ydcI*
ydcI	*ydcK*
ydcL	*ydcM*
ydcM	*ydcN*
sacV	*ydcO*
ydcO	*ydcP*
ydcP	*ydcQ*
ydcQ	*ydcR*
ydcR	*ydcS*
ydcS	*ydcT*
ydcT	*yddA*
yddA	*yddB*
yddB	*yddC*
yddC	*yddD*
yddD	*yddE*
yddE	*yddF*
yddF	*yddG*
yddG	*yddH*
yddH	*yddI*
yddI	*yddJ*
rapI	*phrI*
phrI	*yddM*
yddQ	*yddR*
yddR	*yddS*
yddS	*yddT*
yddT	*ydeA*
ydeA	*cspC*
ydeB	*ydzE*
ydzE	*ydeC*
ydeF	*ydeG*
ydeG	*ydeH*
ydeH	*ydeI*
ydeJ	*ydeK*
ydeL	*ydeM*
ydeN	*ydzF*
ydeS	*ydeT*
ydeT	*ydfA*
ydfA	*ydfB*
ydfH	*ydfI*
ydfI	*ydfJ*
nap	*ydfK*
ydfK	*ydfL*
ydfL	*ydfM*
ydfN	*ydfO*
ydfO	*ydfP*
ydfP	*ydfQ*
ydzH	*ydfR*
cotP	*ydgA*

Gene Pair	
ydgA	*ydgB*
ydgC	*ydgD*
ydgD	*ydgE*
expZ	*ydgF*
ydgF	*dinB*
ydgG	*ydgH*
ydgI	*ydgJ*
ydhC	*ydhD*
ydhD	*ydhE*
ydhF	*phoB*
ydhH	*ydhI*
ydhJ	*ydhK*
ydhM	*ydhN*
ydhN	*ydhO*
ydhO	*ydhP*
ydhP	*ydhQ*
ydhQ	*ydhR*
ydhR	*ydhS*
ydhS	*ydhT*
thiL	*ydiB*
ydiB	*ydiC*
ydiC	*ydiD*
ydiD	*gcp*
ydiG	*ydiH*
ydiH	*tatAY*
tatAY	*tatCY*
ydiK	*ydiL*
groES	*groEL*
groEL	*ydiM*
ydiM	*ydiN*
ydiN	*ydiO*
ydiO	*ydiP*
ydiP	*ydiQ*
ydiQ	*ydiR*
ydiR	*ydiS*
ydiS	*ydjA*
ydjA	*ydjB*
ydjB	*ydjC*
gutB	*gutP*
gutP	*ydjE*
ydjE	*pspA*
pspA	*ydjG*
ydjG	*ydjH*
ydjH	*ydjI*
ydjM	*ydjN*
ydjO	*ydjP*
ydjP	*yeaA*
yeaA	*cotA*
cotA	*gabP*
yeaB	*yeaC*
yeaC	*yeaD*
yeaD	*yebA*
yebA	*guaA*
guaA	*pbuG*
pbuG	*yebC*
yebC	*yebD*
yebD	*yebE*
yebE	*yebG*
yebG	*purE*
purE	*purK*

Gene Pair	
purK	*purB*
purB	*purC*
purC	*purS*
purS	*purQ*
purQ	*purL*
purL	*purF*
purF	*purM*
purM	*purN*
purN	*purH*
purH	*purD*
yecA	*yerA*
yerA	*yerB*
yerB	*yerC*
pcrB	*pcrA*
pcrA	*ligA*
ligA	*yerH*
yerH	*yerI*
sapB	*opuE*
gatC	*gatA*
gatA	*gatB*
yerP	*yerQ*
yerQ	*yefA*
yefC	*yeeA*
yeeA	*yeeB*
yeeB	*yeeC*
yeeD	*yezA*
yezA	*yeeF*
yeeG	*rapH*
rapH	*yeeI*
yeeI	*yeeK*
yeeK	*yezE*
yezE	*yesE*
yesE	*yesF*
yesF	*cotJA*
cotJA	*cotJB*
cotJB	*cotJC*
cotJC	*yesJ*
yesJ	*yesK*
yesK	*yesL*
yesL	*yesM*
yesM	*yesN*
yesN	*yesO*
yesO	*yesP*
yesP	*yesQ*
yesQ	*yesR*
yesR	*yesS*
yesS	*yesT*
yesT	*yesU*
yesU	*yesV*
yesV	*yesW*
yesW	*yesX*
yesX	*yesY*
yesY	*yesZ*
yesZ	*yetA*
yetA	*lplA*
lplA	*lplB*
lplB	*lplC*
lplC	*lplD*
lplD	*yetF*
yetG	*yetH*

Gene Pair	
yetI	*yezB*
yezB	*yezD*
yezD	*yetJ*
yetJ	*yetK*
yetO	*yfnI*
yfnI	*yfnH*
yfnH	*yfnG*
yfnG	*yfnF*
yfnF	*yfnE*
yfnE	*yfnD*
yfnC	*yfnB*
yfnB	*yfnA*
yfmT	*yfmS*
yfmS	*yfmR*
yfmR	*yfmQ*
yfmQ	*yfmP*
yfmP	*yfmO*
yfmO	*yfmN*
yfmL	*yfmK*
yfmJ	*yfmI*
yfmH	*yfmG*
yfmF	*yfmE*
yfmE	*yfmD*
yfmD	*yfmC*
yfmC	*yfmB*
yfmA	*yflT*
yflT	*pel*
pel	*yflS*
yflS	*citS*
citS	*citT*
citT	*yflP*
yflP	*citM*
citM	*yflN*
yflN	*yflM*
yflJ	*yflI*
yflI	*yflH*
yflH	*yflG*
yflD	*yflC*
yflC	*yflB*
yflB	*yflA*
yfkT	*yfkS*
yfkS	*yfkR*
yfkR	*yfkQ*
treP	*treA*
treA	*treR*
treR	*yfkO*
yfkL	*yfkK*
yfkJ	*yfkI*
yfkI	*yfkH*
yfkE	*yfkD*
yfkC	*yfkB*
yfkB	*yfkA*
yfjT	*yfjS*
yfjR	*yfjQ*
yfjP	*yfjO*
yfjO	*yfjN*
yfjN	*yfjM*
yfjM	*yfjL*
yfjL	*acoA*
acoA	*acoB*

Gene Pair	
acoB	*acoC*
acoC	*acoL*
acoL	*acoR*
acoR	*sspH*
yfjF	*yfjE*
yfjE	*yfjD*
yfjD	*yfjC*
yfjC	*yfjB*
yfjB	*yfjA*
malA	*yfiA*
yfiA	*malP*
malP	*yfiB*
yfiB	*yfiC*
yfiC	*yfiD*
yfiD	*yfiE*
yfiE	*yfiF*
yfiF	*yfiG*
yfiG	*yfiH*
yfiH	*yfiI*
yfiI	*yfiJ*
yfiJ	*yfiK*
yfiK	*yfiL*
yfiL	*yfiM*
yfiM	*yfiN*
lipB	*yfiQ*
yfiR	*yfiS*
yfiU	*yfiV*
yfiW	*yfiX*
yfiZ	*yfhA*
yfhA	*yfhB*
yfhB	*yfhC*
yfhD	*yfhE*
yfhE	*yfhF*
yfhG	*yfhH*
yfhH	*yfhI*
yfhJ	*yfhK*
yfhK	*yfhL*
yfhL	*yfhM*
yfhM	*csbB*
csbB	*yfhO*
fabL	*sspE*
sspE	*ygaB*
ygaB	*ygaC*
ygaC	*ygaD*
ygaE	*gsaB*
ygaF	*perR*
perR	*ygzB*
ygzB	*ygxA*
ygzA	*ygaJ*
ygaJ	*thiC*
ssuB	*ssuA*
ssuA	*ssuC*
ssuC	*ssuD*
ssuD	*ygaN*
ygaN	*yhzA*
yhbA	*yhbB*
yhbB	*cspR*
cspR	*yhbD*
yhbD	*yhbE*
yhbE	*yhbF*

Gene Pair	
yhbF	*prkA*
prkA	*yhbH*
yhbH	*yhbI*
yhbI	*yhbJ*
yhbJ	*yhcA*
yhcA	*yhcB*
yhcB	*yhcC*
yhcC	*yhcD*
yhcD	*yhcE*
yhcE	*yhcF*
yhcF	*yhcG*
yhcG	*yhcH*
yhcH	*yhcI*
yhcN	*yhcO*
yhcO	*yhcP*
yhcR	*yhcS*
yhcU	*yhcV*
yhcV	*yhcW*
yhcW	*yhcX*
yhcX	*yhxA*
yhxA	*glpP*
glpP	*glpF*
glpF	*glpK*
glpK	*glpD*
glpD	*yhxB*
yhxB	*yhcY*
yhcY	*yhcZ*
yhcZ	*yhdA*
lytF	*yhdE*
yhdE	*ygxB*
citA	*yhdF*
yhdF	*yhdG*
yhdG	*yhdH*
yhdK	*yhdL*
yhdL	*sigM*
yhdN	*yhdO*
yhdP	*yhdQ*
yhdU	*yhdV*
yhdV	*yhdW*
yhdX	*yhdY*
yhdY	*yhdZ*
nhaC	*nhaX*
yheJ	*yheI*
yheI	*yheH*
yheG	*yheF*
yheF	*sspB*
sspB	*yheE*
yheE	*yheD*
yheD	*yheC*
yheB	*yheA*
yhaZ	*yhaY*
yhaX	*hemZ*
yhaU	*yhaT*
yhaT	*yhaS*
yhaR	*yhaQ*
yhaQ	*yhaP*
yhaP	*yhaO*
yhaO	*yhaN*
yhaN	*yhaM*
yhaM	*yhaL*

Appendix C. (Continued)

Gene Pair	
prsA	*yhaK*
yhaK	*yhaJ*
hpr	*yhaH*
yhaH	*yhaG*
yhaG	*serC*
serC	*hit*
ecsA	*ecsB*
ecsB	*ecsC*
yhaA	*yhfA*
yhfA	*yhgB*
yhgB	*yhgC*
pbpF	*hemE*
hemE	*hemH*
hemH	*hemY*
hemY	*yhgD*
yhgD	*yhgE*
yhfE	*yhfF*
glrT	*yhfH*
yhfI	*yhfJ*
yhfJ	*yhfK*
yhfK	*yhfL*
yhfO	*yhfP*
yhfP	*yhfQ*
yhfR	*yhfS*
yhfS	*yhfT*
yhfT	*yhfU*
yhfU	*hemAT*
hemAT	*yhfW*
yhjB	*yhjC*
yhjC	*yhjD*
yhjE	*sipV*
sipV	*yhjG*
yhjG	*yhjH*
glcP	*yhjJ*
yhjJ	*yhjK*
yhjK	*yhjL*
yhjN	*yhjO*
yhjQ	*yhjR*
addB	*addA*
addA	*sbcD*
sbcD	*yirY*
yirY	*yisB*
gerPF	*gerPE*
gerPE	*gerPD*
gerPD	*gerPC*
gerPC	*gerPB*
gerPB	*gerPA*
gerPA	*yisI*
yisI	*yisJ*
yisK	*yisL*
yisL	*wprA*
asnO	*yizA*
yisQ	*yisR*
yisR	*degA*
yisS	*yisT*
yisU	*yisV*
yisZ	*yitA*
yitA	*yitB*
yitC	*yitD*
yitE	*yitF*

Gene Pair	
yitF	*yitG*
yitG	*yitH*
yitH	*yitI*
yitI	*yitJ*
yitJ	*yitK*
yitM	*yitN*
yitN	*yitO*
yitO	*yitP*
yitQ	*yitR*
yitR	*nprB*
yitT	*ipi*
yitV	*yitW*
yitW	*yitY*
yitY	*yitZ*
yitZ	*argC*
argC	*argJ*
argJ	*argB*
argB	*argD*
argD	*carA*
carA	*carB*
carB	*argF*
argF	*yjzC*
yjaU	*yjaV*
yjaV	*med*
med	*comZ*
fabHA	*fabF*
fabF	*yjaZ*
yjaZ	*appD*
appD	*appF*
appF	*appA*
appA	*appB*
appB	*appC*
appC	*yjbA*
oppA	*oppB*
oppB	*oppC*
oppC	*oppD*
oppD	*oppF*
oppF	*yjbB*
yjbB	*yjbC*
yjbC	*yjbD*
mecA	*yjbF*
yjbF	*yjbG*
yjbH	*yjbI*
yjbI	*yjbJ*
yjbJ	*yjbK*
yjbL	*yjbM*
yjbM	*yjbN*
yjbN	*yjbO*
yjbQ	*tenA*
tenA	*tenI*
tenI	*goxB*
goxB	*thiS*
thiS	*thiG*
thiG	*thiF*
thiF	*yjbV*
yjbV	*fabI*
fabI	*yjbX*
cotZ	*cotY*
cotY	*cotX*
cotX	*cotW*

Gene Pair	
cotW	*cotV*
yjcA	*yjcB*
yjcB	*yjcC*
yjcD	*yjcE*
yjcE	*yjcF*
yjcF	*yjcG*
yjcG	*yjcH*
yjcI	*yjcJ*
yjcK	*yjcL*
yjcN	*yjcO*
yjcO	*yjcP*
yjcP	*yjcQ*
yjcQ	*yjcR*
yjcR	*yjcS*
yjcS	*yjdA*
manR	*manP*
manP	*manA*
manA	*yjdF*
yjdG	*yjdH*
yjdJ	*ctaO*
ctaO	*cotT*
yjfA	*yjfB*
yjgA	*yjgB*
yjgC	*yjgD*
yjgD	*yjhA*
yjhA	*yjhB*
yjiB	*yjiC*
yjiC	*yjjA*
yjkA	*yjkB*
yjkB	*yjlA*
yjlB	*yjlC*
yjlC	*yjlD*
yjlD	*uxaC*
uxaC	*yjmB*
yjmB	*yjmC*
yjmC	*yjmD*
yjmD	*uxuA*
uxuA	*yjmF*
yjmF	*exuT*
exuT	*exuR*
exuR	*uxaB*
uxaB	*uxaA*
yjnA	*yjoA*
yjoB	*rapA*
rapA	*phrA*
yjqB	*yjqC*
xkdA	*xre*
xkdB	*xkdC*
xkdC	*xkdD*
xkdD	*xtrA*
xtrA	*xpf*
xpf	*xtmA*
xtmA	*xtmB*
xtmB	*xkdE*
xkdE	*xkdF*
xkdF	*xkdG*
xkdG	*xkdH*
xkdH	*xkdI*
xkdI	*xkdJ*
xkdJ	*xkdK*

Gene Pair	
xkdK	*xkdM*
xkdM	*xkdN*
xkdN	*xkdO*
xkdO	*xkdP*
xkdP	*xkdQ*
xkdQ	*xkdR*
xkdR	*xkdS*
xkdS	*xkdT*
xkdT	*xkdU*
xkdU	*xkdV*
xkdV	*xkdW*
xkdW	*xkdX*
xkdX	*xepA*
xepA	*xhlA*
xhlA	*xhlB*
xhlB	*xlyA*
spoIISB	*spoIISA*
spoIISA	*pit*
pit	*ykaA*
ykaA	*ykbA*
ykcA	*ykcB*
ykcB	*ykcC*
proG	*dppA*
dppA	*dppB*
dppB	*dppC*
dppC	*dppD*
dppD	*dppE*
dppE	*ykfA*
ykfA	*ykfB*
ykfB	*ykfC*
ykfC	*ykfD*
ykgB	*ykgA*
ykhA	*hmp*
ykjA	*ykkA*
ykkA	*ykkB*
ykkB	*ykkC*
ykkC	*ykkD*
ykkD	*ykkE*
ykkE	*proB*
proB	*proA*
proA	*yklA*
guaD	*metE*
metE	*ispA*
ykoC	*ykoD*
ykoD	*ykoE*
ykoE	*ykoF*
ykoG	*ykoH*
ykoH	*ykoI*
ykoI	*ykoJ*
ykoJ	*ykzD*
ykzD	*ykoK*
ykzB	*ykoL*
ykoL	*ykoM*
ykoS	*ykoT*
ykoU	*ykoV*
ykoV	*ykoW*
ykoX	*ykoY*
ykoY	*sigI*
sigI	*ykrI*
sspD	*ykrK*

Gene Pair	
ykrL	*ykrM*
ykrM	*ykzE*
kinE	*ogt*
ykrS	*ykrT*
ykrT	*ykrU*
ykrV	*ykrW*
ykrW	*ykrX*
ykrX	*ykrY*
ykrY	*ykrZ*
spo0E	*eag*
motB	*motA*
motA	*clpE*
ykvI	*ykvJ*
ykvJ	*ykvK*
ykvK	*ykvL*
ykvL	*ykvM*
ykvO	*ykvP*
ykvP	*ykvQ*
ykvQ	*ykvR*
ykvT	*ykvU*
ykvU	*ykvV*
ykvV	*ykvW*
ykvW	*ykvY*
ykvY	*ykvZ*
ykvZ	*glcT*
glcT	*ptsG*
ptsG	*ptsH*
ptsH	*ptsI*
ptsI	*splA*
splA	*splB*
mcpC	*ykwC*
ykuA	*kinA*
ykyB	*ykuC*
ykuC	*ykuD*
ykuD	*ykuE*
ykuF	*ykuG*
ykuG	*ykuH*
ykuH	*ykuI*
ykuI	*ykuJ*
ykuJ	*ykuK*
ykuK	*ykzF*
ykzF	*ykuL*
ykuL	*ccpC*
ccpC	*ykuN*
ykuN	*ykuO*
ykuO	*ykuP*
ykuP	*ykuQ*
ykuQ	*ykuR*
ykuR	*ykuS*
ykuU	*ykuV*
ykuV	*rok*
mobA	*moeB*
moeB	*moeA*
moeA	*mobB*
mobB	*moaE*
moaE	*moaD*
moaD	*yknU*
yknU	*yknV*
yknV	*yknW*
yknW	*yknX*

Gene Pair	
yknX	*yknY*
yknY	*yknZ*
yknZ	*fruR*
fruR	*fruK*
fruK	*fruA*
fruA	*sipT*
ykpA	*ykpB*
ampS	*ykpC*
ykpC	*mreBH*
abh	*kinC*
kinC	*ykqA*
ykqA	*ykqB*
ykqB	*adeC*
ykqC	*ykzG*
ykyA	*pdhA*
pdhA	*pdhB*
pdhB	*pdhC*
pdhC	*pdhD*
slp	*speA*
ykzI	*yktC*
yktC	*ykzC*
yktD	*nprE*
ylaA	*ylaB*
ylaB	*ylaC*
ylaC	*ylaD*
ylaE	*ylaF*
ylaG	*ylaH*
ylaI	*ylaJ*
ylaM	*ylaN*
ylaN	*ftsW*
ftsW	*pycA*
ctaB	*ctaC*
ctaC	*ctaD*
ctaD	*ctaE*
ctaE	*ctaF*
ctaF	*ctaG*
ylbB	*ylbC*
ylbC	*ylbD*
ylbD	*ylbE*
ylbE	*ylbF*
ylbF	*ylbG*
ylbG	*ylbH*
ylbH	*ylbI*
ylbK	*ylbL*
ylbN	*rpmF*
rpmF	*ylbO*
ylbQ	*yllA*
yllA	*yllB*
yllB	*ylxA*
ylxA	*ftsL*
ftsL	*pbpB*
pbpB	*spoVD*
spoVD	*murE*
murE	*mraY*
mraY	*murD*
murD	*spoVE*
spoVE	*murG*
murG	*murB*
murB	*divIB*
divIB	*ylxW*

Appendix C. (Continued)

Gene Pair	
ylxW	*ylxX*
ylxX	*sbp*
sbp	*ftsA*
ftsA	*ftsZ*
ftsZ	*bpr*
bpr	*spoIIGA*
spoIIGA	*sigE*
sigE	*sigG*
sigG	*ylmA*
ylmA	*ylmB*
ylmB	*ylmC*
ylmC	*ylmD*
ylmD	*ylmE*
ylmE	*ylmF*
ylmF	*ylmG*
ylmG	*ylmH*
ylmH	*divIVA*
divIVA	*ileS*
ileS	*ylyA*
ylyA	*lspA*
lspA	*ylyB*
ylyB	*pyrR*
pyrR	*pyrP*
pyrP	*pyrB*
pyrB	*pyrC*
pyrC	*pyrAA*
pyrAA	*pyrAB*
pyrAB	*pyrK*
pyrK	*pyrD*
pyrD	*pyrF*
pyrF	*pyrE*
pyrE	*cysH*
cysH	*cysP*
cysP	*sat*
sat	*cysC*
cysC	*ylnD*
ylnD	*ylnE*
ylnE	*ylnF*
yloB	*yloC*
yloC	*ylzA*
ylzA	*gmk*
gmk	*yloH*
yloH	*yloI*
yloI	*priA*
priA	*def*
def	*fmt*
fmt	*yloM*
yloM	*yloN*
yloN	*prpC*
prpC	*prkC*
prkC	*yloQ*
yloQ	*rpe*
rpe	*yloS*
yloS	*spoVM*
yloU	*yloV*
yloV	*sdaAB*
sdaAB	*sdaAA*
sdaAA	*recG*
recG	*ylpC*
ylpC	*plsX*

Gene Pair	
plsX	*fabD*
fabD	*fabG*
fabG	*acpA*
acpA	*rnc*
rnc	*smc*
smc	*ftsY*
ylxM	*ffh*
ffh	*rpsP*
rpsP	*ylqC*
ylqC	*ylqD*
ylqD	*rimM*
rimM	*trmD*
trmD	*rplS*
rplS	*ylqF*
ylqF	*rnhB*
rnhB	*ylqG*
ylqG	*ylqH*
ylqH	*sucC*
sucC	*sucD*
sucD	*smf*
smf	*topA*
topA	*gid*
gid	*codV*
codV	*clpQ*
clpQ	*clpY*
clpY	*codY*
codY	*flgB*
flgB	*flgC*
flgC	*fliE*
fliE	*fliF*
fliF	*fliG*
fliG	*fliH*
fliH	*fliI*
fliI	*fliJ*
fliJ	*ylxF*
ylxF	*fliK*
fliK	*ylxG*
ylxG	*flgE*
flgE	*fliL*
fliL	*fliM*
fliM	*fliY*
fliY	*cheY*
cheY	*fliZ*
fliZ	*fliP*
fliP	*fliQ*
fliQ	*fliR*
fliR	*flhB*
flhB	*flhA*
flhA	*flhF*
flhF	*ylxH*
ylxH	*cheB*
cheB	*cheA*
cheA	*cheW*
cheW	*cheC*
cheC	*cheD*
cheD	*sigD*
sigD	*ylxL*
ylxL	*rpsB*
rpsB	*tsf*
tsf	*pyrH*

Gene Pair	
pyrH	*frr*
frr	*uppS*
uppS	*cdsA*
cdsA	*dxr*
dxr	*yluC*
yluC	*proS*
proS	*polC*
polC	*ylxS*
ylxS	*nusA*
nusA	*ylxR*
ylxR	*ylxQ*
ylxQ	*infB*
infB	*ylxP*
ylxP	*rbfA*
rbfA	*truB*
truB	*ribC*
ribC	*rpsO*
rpsO	*pnpA*
pnpA	*ylxY*
ylxY	*mlpA*
mlpA	*ymxH*
ymxH	*spoVFA*
spoVFA	*spoVFB*
spoVFB	*asd*
asd	*dapG*
dapG	*dapA*
dapA	*ymfA*
ymfA	*tepA*
tepA	*spoIIIE*
spoIIIE	*ymfC*
ymfC	*ymfD*
ymfD	*ymfE*
ymfE	*ymfF*
ymfF	*ymfG*
ymfG	*ymfH*
ymfH	*ymfI*
ymfI	*ymfJ*
ymfJ	*ymfK*
ymfK	*ymfL*
ymfL	*ymfM*
ymfM	*pgsA*
pgsA	*cinA*
cinA	*recA*
recA	*pbpX*
pbpX	*ymdA*
ymdA	*ymdB*
ymdB	*spoVS*
spoVS	*tdh*
tdh	*kbl*
kbl	*ymcB*
ymcB	*ymcA*
ymcA	*cotE*
cotE	*mutS*
mutS	*mutL*
ymcC	*pksA*
pksA	*pksB*
pksB	*pksC*
pksC	*pksD*
pksD	*pksE*
pksE	*acpK*

Gene Pair	
acpK	*pksF*
pksF	*pksG*
pksG	*pksH*
pksH	*pksI*
pksI	*pksJ*
pksJ	*pksL*
pksL	*pksM*
pksM	*pksN*
pksN	*pksR*
pksS	*ymzB*
ymzB	*ymaE*
ymaE	*aprX*
ymaC	*ymaD*
ebrB	*ebrA*
ebrA	*ymaG*
ymaF	*miaA*
miaA	*ymaH*
ymaH	*ymzC*
ymzC	*ymzA*
ymzA	*ymaA*
ymaA	*nrdE*
nrdE	*nrdF*
nrdF	*ymaB*
spoVK	*ynbA*
ynbA	*ynbB*
ynbB	*glnR*
glnR	*glnA*
glnA	*ynxB*
ynxB	*ynzF*
ynzF	*ynzG*
ynzG	*ynaB*
ynaB	*ynaC*
ynaC	*ynaD*
ynaD	*ynaE*
ynaE	*ynaF*
ynaF	*ynaG*
ynaG	*ynaI*
ynaI	*xynP*
xynP	*xynB*
xylA	*xylB*
yncE	*yncF*
yncM	*cotC*
cotC	*tatAC*
ynzB	*yndD*
yndD	*yndE*
yndE	*yndF*
yndF	*yndG*
yndG	*yndH*
yndH	*yndJ*
yndJ	*yndK*
yndK	*yndL*
yneA	*yneB*
yneB	*ynzC*
ynzC	*tkt*
tkt	*yneE*
yneE	*yneF*
ccdA	*yneI*
yneI	*yneJ*
yneK	*cotM*
cotM	*sspP*

Gene Pair	
sspP	*sspO*
citB	*yneN*
yneN	*sspN*
sspN	*tlp*
tlp	*yneP*
yneP	*yneQ*
yneR	*yneS*
yneT	*parE*
parE	*parC*
alsT	*bglC*
bglC	*ynfE*
ynfF	*xynD*
yngA	*yngB*
yngB	*yngC*
yngD	*yngE*
yngE	*yngF*
yngF	*yngG*
yngG	*yngH*
yngH	*yngI*
yngI	*yngJ*
yngJ	*ynzE*
ynzE	*yngK*
yngK	*yngL*
yngL	*ppsE*
ppsE	*ppsD*
ppsD	*ppsC*
ppsC	*ppsB*
ppsB	*ppsA*
ppsA	*dacC*
dacC	*yoxA*
yoxA	*yoeA*
yoeB	*nSL-Arg1*
yoeC	*yoeD*
gltB	*gltA*
proJ	*proH*
proH	*rtp*
rtp	*yoxD*
yoxC	*yoxB*
yoxB	*yoaA*
yoaB	*yoaC*
yoaC	*yoaD*
yoaE	*yoaF*
yoaJ	*yoaK*
pelB	*yoaM*
yoaN	*yoaO*
yoaO	*yoaP*
yozF	*yoaR*
yoaS	*yozG*
yozG	*yoaT*
yoaW	*yoaZ*
pps	*xynA*
yozI	*yobE*
yobF	*yozJ*
rapK	*phrK*
yobH	*yozK*
yozK	*yozL*
yobI	*yobJ*
yobJ	*yobK*
yobK	*yobL*
yobL	*yobM*

Gene Pair	
yobN	*yobO*
csaA	*yobQ*
yobQ	*yobR*
yobR	*yobS*
yobS	*yobT*
yobT	*yobU*
yobU	*yobV*
yozB	*yocB*
yocC	*yocD*
yocD	*des*
des	*yocF*
yocF	*yocG*
yocH	*yocI*
yocI	*yocJ*
yocJ	*yocK*
yocK	*yocL*
yocL	*yocM*
yozN	*yocN*
yozO	*yozC*
dhaS	*sqhC*
sqhC	*sodF*
odhB	*odhA*
odhA	*yojO*
yojO	*yojN*
yojN	*yojM*
yojM	*yojL*
yojL	*yojK*
yojK	*yojJ*
yojI	*yojH*
yojG	*yojF*
yojF	*yojE*
yojC	*yojB*
yojB	*yojA*
yodD	*yodE*
yodH	*yodI*
yodJ	*deoD*
deoD	*yodL*
yodL	*yodM*
yodM	*yozD*
yozD	*yodN*
yodN	*yozE*
yozE	*kamA*
kamA	*yodP*
yodP	*yodQ*
yodQ	*yodR*
yodR	*yodS*
yodS	*yodT*
yodT	*cgeE*
cgeE	*cgeD*
cgeD	*cgeC*
cgeA	*cgeB*
yotN	*yotM*
yotK	*yotJ*
yotJ	*yotI*
yotI	*yotH*
yotH	*yotG*
yotG	*yotF*
yotF	*yotE*
yotE	*yotD*
yotD	*yotC*

Appendix C. (Continued)

Gene Pair	
yotC	*yotB*
yosZ	*yosX*
yosX	*yosW*
yosW	*yosV*
yosT	*yosS*
yosS	*yosR*
yosR	*yosP*
yosP	*yosQ*
yosQ	*yosN*
yosN	*yosM*
yosM	*yosL*
yosL	*yosK*
yosK	*yosJ*
yosJ	*yosI*
yosI	*yosH*
yosH	*yosG*
yosG	*yosF*
yosF	*yosE*
yosE	*yosD*
yosD	*yosC*
yosC	*yosB*
yorZ	*yorY*
yorY	*yorX*
yorX	*yorW*
yorW	*yorV*
yorV	*mtbP*
mtbP	*yorT*
yorT	*yorS*
yorS	*yorR*
yorR	*yorQ*
yorQ	*yorP*
yorP	*yorO*
yorO	*yorN*
yorN	*yorM*
yorM	*yorL*
yorL	*yorK*
yorK	*yorJ*
yorJ	*yorI*
yorI	*yorH*
yorH	*yorG*
yorG	*yorF*
yorF	*yorE*
yorE	*yorD*
yorD	*yorC*
yorC	*yorB*
yorB	*yorA*
yorA	*yoqZ*
yoqZ	*yoqY*
yoqY	*yoqX*
yoqW	*ligB*
yoqS	*yoqR*
yoqR	*yoqP*
yoqL	*yoqK*
yoqK	*yoqJ*
yoqJ	*yoqI*
yoqI	*yoqH*
yoqH	*yoqG*
yoqG	*yoqF*
yoqF	*yoqE*
yoqE	*yoqD*

Gene Pair	
yoqD	*yoqC*
yoqC	*yoqB*
yoqB	*yoqA*
yoqA	*yopZ*
yopZ	*yopY*
yopY	*yopX*
yopX	*yopW*
yopW	*yopV*
yopV	*yopU*
yopU	*yopT*
yopR	*yopQ*
yopQ	*yopP*
yopP	*yopO*
yopO	*yopN*
yopN	*yopM*
yopM	*yopL*
yopL	*yopK*
yopK	*yopJ*
yopJ	*yopI*
yopI	*yopH*
yopH	*yopG*
yopG	*yopF*
yopF	*yopE*
yopE	*yopD*
yopD	*yopC*
yopC	*yopB*
yopB	*yopA*
yopA	*yonX*
yonX	*yonV*
yonV	*yonU*
yonU	*yonT*
yonT	*yonS*
yonS	*yonR*
yonP	*yonO*
yonO	*yonN*
yonN	*yonK*
yonK	*yonJ*
yonH	*yonG*
yonG	*yonF*
yonF	*yonE*
yonE	*yonD*
yonD	*yonC*
yonC	*yonB*
yonB	*yonA*
yonA	*yomZ*
yomZ	*yomY*
yomY	*yomX*
yomX	*yomW*
yomW	*yomV*
yomV	*yomU*
yomU	*yomT*
yomT	*yomS*
yomS	*yomR*
yomR	*yomQ*
yomQ	*yomP*
yomP	*yomO*
yomO	*yomN*
yomN	*yomM*
yozP	*yomL*
yomK	*yomJ*

Gene Pair	
yomJ	*yomI*
yomI	*yomH*
yomH	*yomG*
yomG	*yomF*
yomF	*yomE*
yomE	*yomD*
yomD	*blyA*
blyA	*bhlA*
bhlA	*bhlB*
bdbB	*yolJ*
yolJ	*bdbA*
bdbA	*sunT*
sunT	*sunA*
sunA	*yolF*
yolF	*uvrX*
uvrX	*yolD*
yolB	*yolA*
yolA	*yokL*
yokL	*yokK*
yokK	*yokJ*
yokJ	*yokI*
yokI	*yokH*
yokH	*yokG*
yokF	*yokE*
yokE	*yokD*
yokC	*yokB*
yokA	*ypqP*
yppQ	*msrA*
ypnP	*ypmT*
ypmT	*ypmS*
ypmS	*ypmR*
ypmR	*ypmQ*
ypmQ	*ypmP*
ypmP	*ilvA*
yplP	*yplQ*
ypkP	*dfrA*
dfrA	*thyB*
thyB	*ypjQ*
ypjQ	*ypjP*
ypjP	*ypiP*
ypiP	*yphP*
yphP	*ilvD*
ilvD	*ypgR*
ypgR	*ypgQ*
ypgQ	*bsaA*
metA	*ugtP*
ugtP	*cspD*
ypeQ	*ypeP*
ypdP	*ypdQ*
ypdQ	*sspL*
ypcP	*ypbS*
ypbS	*ypbR*
ypbR	*ypbQ*
ypbQ	*bcsA*
bcsA	*pbuX*
pbuX	*xpt*
xpt	*ypwA*
ypwA	*kdgT*
kdgT	*kdgA*
kdgA	*kdgK*

Gene Pair	
kdgK	*kdgR*
kduI	*kduD*
ypvA	*yptA*
yptA	*ypsC*
ypsB	*ypsA*
ypsA	*cotD*
cotD	*yprB*
yprB	*yprA*
yprA	*ypqE*
yppE	*yppD*
recU	*ponA*
ypoC	*nth*
nth	*dnaD*
dnaD	*asnS*
asnS	*aspB*
aspB	*ypmB*
ypmB	*ypmA*
ypmA	*dinG*
dinG	*panD*
panD	*panC*
panC	*panB*
panB	*birA*
birA	*cca*
cca	*ypjH*
ypjH	*ypjG*
ypjG	*mgsA*
mgsA	*dapB*
dapB	*ypjD*
ypjB	*ypjA*
ypjA	*qcrC*
qcrC	*qcrB*
qcrB	*qcrA*
qcrA	*ypiF*
ypiF	*ypiB*
ypiB	*ypiA*
ypiA	*aroE*
aroE	*tyrA*
tyrA	*hisC*
hisC	*trpA*
trpA	*trpB*
trpB	*trpF*
trpF	*trpC*
trpC	*trpD*
trpD	*trpE*
trpE	*aroH*
aroH	*aroB*
aroB	*aroF*
aroF	*cheR*
cheR	*ndk*
ndk	*hepT*
hepT	*menH*
menH	*hepS*
hepS	*mtrB*
mtrB	*mtrA*
mtrA	*hbs*
hbs	*spoIVA*
spoIVA	*yphF*
yphF	*yphE*
yphE	*gpsA*
gpsA	*yphC*

Gene Pair	
yphC	*seaA*
seaA	*yphA*
yphA	*ypgA*
ypgA	*ypfD*
ypfD	*cmk*
cmk	*ypfB*
ypfB	*ypfA*
ypfA	*ypeB*
ypeB	*sleB*
sleB	*ypdC*
ypdC	*ypdA*
ypdA	*gudB*
gudB	*ypbH*
ypbH	*ypbG*
ypbG	*ypbF*
ypbF	*ypbE*
ypbE	*ypbD*
ypbD	*recQ*
recQ	*ypbB*
ypaA	*ypzE*
aroC	*rsiX*
rsiX	*sigX*
sigX	*resE*
resE	*resD*
resD	*resC*
resC	*resB*
resB	*resA*
resA	*rluB*
rluB	*spmB*
spmB	*spmA*
spmA	*dacB*
dacB	*ypuI*
ypuI	*ypuH*
ypuH	*ypuG*
ribT	*ribH*
ribH	*ribA*
ribA	*ribE*
ribE	*ribD*
ribD	*ypuE*
ypuE	*ypuD*
ypuD	*sipS*
ypuC	*ypuB*
lysA	*spoVAF*
spoVAF	*spoVAE*
spoVAE	*spoVAD*
spoVAD	*spoVAC*
spoVAC	*spoVAB*
spoVAB	*spoVAA*
spoVAA	*sigF*
sigF	*spoIIAB*
spoIIAB	*spoIIAA*
spoIIAA	*dacF*
dacF	*punA*
punA	*drm*
drm	*ripX*
ripX	*fur*
fur	*spoIIM*
spoIIM	*yqkK*
yqkK	*mleA*
mleA	*mleN*

Gene Pair	
mleN	*ansB*
ansB	*ansA*
yqxK	*nudF*
yqkC	*yqkB*
yqkB	*yqkA*
yqkA	*yqjZ*
yqjZ	*yqjY*
yqjY	*yqjX*
yqjX	*yqjW*
yqzH	*yqjV*
yqjV	*yqjU*
yqjT	*coaA*
coaA	*dsdA*
dsdA	*yqjQ*
yqjQ	*yqjP*
yqjM	*yqjL*
yqjI	*yqjH*
yqzJ	*yqjG*
yqjG	*yqjF*
yqjE	*yqjD*
yqjD	*yqjC*
yqjC	*yqjB*
yqjB	*yqjA*
yqjA	*yqiZ*
yqiZ	*yqiY*
yqiY	*yqiX*
yqiX	*yqiW*
bmrU	*bmr*
bmr	*bmrR*
bkdB	*bkdAB*
bkdAB	*bkdAA*
bkdAA	*lpdV*
lpdV	*buk*
buk	*bcd*
bcd	*ptb*
ptb	*bkdR*
yqiQ	*mmgE*
mmgE	*mmgD*
mmgD	*mmgC*
mmgC	*mmgB*
mmgB	*mmgA*
mmgA	*yqiK*
yqiK	*yqiI*
yqiI	*yqiH*
spo0A	*spoIVB*
spoIVB	*recN*
recN	*ahrC*
ahrC	*yqxC*
yqxC	*dxs*
dxs	*yqiD*
yqiD	*yqiC*
yqiC	*yqiB*
yqiB	*folD*
folD	*nusB*
nusB	*yqhY*
yqhY	*accC*
accC	*accB*
accB	*spoIIAH*
spoIIAH	*spoIIAG*
spoIIAG	*spoIIAF*

Appendix C. (Continued)

Gene Pair	
spoIIIAF	*spoIIIAE*
spoIIIAE	*spoIIIAD*
spoIIIAD	*spoIIIAC*
spoIIIAC	*spoIIIAB*
spoIIIAB	*spoIIIAA*
spoIIIAA	*yqhV*
yqhV	*efp*
efp	*yqhT*
yqhT	*yqhS*
yqhS	*yqhR*
yqhQ	*yqhP*
yqhO	*mntR*
mntR	*yqhM*
gcvPB	*gcvPA*
gcvPA	*gcvT*
yqhH	*yqhG*
yqhG	*sinI*
sinI	*sinR*
tasA	*sipW*
sipW	*yqxM*
yqzE	*comGG*
comGG	*comGF*
comGF	*comGE*
comGE	*comGD*
comGD	*comGC*
comGC	*comGB*
comGB	*comGA*
comGA	*yqxL*
yqgV	*yqgU*
yqgU	*yqgT*
yqgT	*yqgS*
yqgS	*glcK*
glcK	*yqgQ*
yqgQ	*yqgP*
yqgP	*yqgO*
yqgO	*yqgN*
yqgN	*rpmGA*
rpmGA	*yqgM*
yqgM	*yqgL*
yqzD	*yqzC*
pstBB	*pstBA*
pstBA	*pstA*
pstA	*pstC*
pstC	*pstS*
pstS	*pbpA*
pbpA	*yqgE*
yqgE	*sodA*
sodA	*yqgC*
yqfZ	*yqfY*
zur	*yqfU*
yqfS	*yqfR*
yqfQ	*yqfP*
yqfO	*yqfN*
yqfN	*cccA*
cccA	*sigA*
sigA	*dnaG*
yqxD	*yqfL*
yqfL	*yqzB*
yqzB	*glyS*
glyS	*glyQ*

Gene Pair	
glyQ	*recO*
recO	*era*
era	*cdd*
cdd	*dgkA*
dgkA	*yqfG*
yqfG	*yqfF*
yqfF	*phoH*
phoH	*yqfD*
yqfD	*yqfC*
yqfC	*yqfB*
yqfB	*yqfA*
yqfA	*yqeZ*
yqeZ	*yqeY*
yqeY	*rpsU*
yqeV	*yqeU*
yqeU	*yqeT*
yqeT	*dnaJ*
dnaJ	*dnaK*
dnaK	*grpE*
grpE	*hrcA*
hrcA	*hemN*
hemN	*lepA*
lepA	*yqxA*
yqxA	*spoIIP*
spoIIP	*gpr*
yqeN	*comEC*
comEC	*comEB*
comEB	*comEA*
yqeM	*yqeL*
yqeL	*yqeK*
yqeK	*yqeJ*
yqeJ	*yqeI*
yqeI	*aroD*
aroD	*yqeH*
yqeH	*yqeG*
yqeF	*cwlH*
spoIVCA	*arsC*
arsC	*arsB*
arsB	*yqcK*
yqcK	*arsR*
arsR	*yqcI*
rapE	*phrE*
phrE	*yqzI*
yqzI	*yqcG*
yqcG	*yqcF*
yqxJ	*yqxI*
yqxI	*cwlA*
cwlA	*yqxH*
yqxH	*yqxG*
yqxG	*yqcE*
yqcE	*yqcD*
yqcD	*yqcC*
yqcC	*yqcB*
yqcB	*yqcA*
yqcA	*yqbT*
yqbT	*yqbS*
yqbS	*yqbR*
yqbR	*yqbQ*
yqbQ	*yqbP*
yqbP	*yqbO*

Gene Pair	
yqbO	*yqbN*
yqbM	*yqbL*
yqbL	*yqbK*
yqbK	*yqbJ*
yqbJ	*yqbI*
yqbI	*yqbH*
yqbH	*yqbG*
yqbG	*yqbF*
yqbF	*yqbE*
yqbE	*yqbD*
yqbD	*yqbC*
yqbC	*yqbB*
yqbB	*yqbA*
yqbA	*yqaT*
yqaT	*yqaS*
yqaS	*yqaR*
yqaR	*yqaQ*
yqaO	*yqaN*
yqaN	*yqaM*
yqaM	*yqaL*
yqaL	*yqaK*
yqaK	*yqaJ*
yqaJ	*yqaI*
yqaI	*yqaH*
yqaH	*yqaG*
yqaG	*yqdA*
yqdA	*yqaF*
yqaD	*yqaC*
yqaB	*spoIIIC*
yrkS	*yrkR*
yrkR	*yrkQ*
yrkQ	*yrkP*
yrkO	*yrkN*
yrkN	*yrkM*
yrkL	*yrkK*
yrkK	*yrkJ*
yrkJ	*yrkI*
yrkI	*yrkH*
yrkH	*yrkG*
yrkG	*yrkF*
yrkF	*yrkE*
yrkE	*yrkD*
yrkD	*yrkC*
yrkC	*yrkB*
yrkB	*bltR*
blt	*bltD*
yrkA	*yrdR*
trkA	*czcD*
czcD	*yrdN*
yrdK	*brnQ*
brnQ	*azlD*
azlD	*azlC*
azlC	*azlB*
azlB	*yrdF*
yrdF	*cypA*
cypA	*yrdD*
yrdD	*yrdC*
yrdC	*yrdB*
yrdB	*yrdA*
yrdA	*aadK*

Gene Pair	
yrpB	*yrpC*
yrpC	*yrpD*
yrpE	*sigZ*
yraO	*yraN*
yraL	*yraK*
yraJ	*yraI*
yraI	*yraH*
yraG	*yraF*
yraF	*adhB*
adhB	*yraE*
yraE	*yraD*
adhA	*yraA*
sacC	*levG*
levG	*levF*
levF	*levE*
levE	*levD*
levD	*levR*
yrhO	*sigV*
sigV	*yrhM*
yrhM	*yrhL*
yrhL	*yrhK*
yrhJ	*yrhI*
yrhI	*yrhH*
yrhH	*yrzI*
yrzI	*yrhG*
yrhG	*yrhF*
yrhE	*yrhD*
yrhC	*yrhB*
yrhB	*yrhA*
yrhA	*mtn*
mtn	*yrrT*
yrrS	*yrrR*
yrrR	*greA*
greA	*udk*
udk	*yrrO*
yrrO	*yrrN*
yrrN	*yrrM*
yrrM	*yrrL*
yrrL	*yrzB*
yrzB	*yrrK*
yrrK	*yrzL*
yrzL	*alaS*
alaS	*yrrI*
glnQ	*glnH*
glnH	*glnM*
glnM	*glnP*
yrrD	*yrrC*
yrrC	*yrrB*
yrrB	*trmU*
trmU	*yrvO*
yrvO	*yrzC*
yrvM	*aspS*
aspS	*hisS*
hisS	*yrzK*
yrvI	*relA*
relA	*apt*
apt	*yrvE*
yrvE	*yrvD*
yrvD	*yrvC*
yrvC	*secDF*

Gene Pair	
secDF	*yrzD*
yrbF	*tgt*
tgt	*queA*
queA	*ruvB*
ruvB	*ruvA*
ruvA	*bofC*
bofC	*csbX*
csbX	*yrbE*
yrzF	*yrzG*
yrzG	*yrzH*
yrzH	*yrbD*
yrbC	*coxA*
coxA	*safA*
safA	*nadA*
nadA	*nadC*
nadC	*nadB*
nifS	*yrxA*
pheA	*pheB*
pheB	*obg*
obg	*spo0B*
spo0B	*rpmA*
rpmA	*ysxB*
ysxB	*rplU*
rplU	*spoIVFB*
spoIVFB	*spoIVFA*
spoIVFA	*minD*
minD	*minC*
minC	*mreD*
mreD	*mreC*
mreC	*mreB*
mreB	*radC*
radC	*maf*
maf	*spoIIB*
spoIIB	*comC*
comC	*folC*
folC	*valS*
valS	*ysxE*
ysxE	*spoVID*
spoVID	*hemL*
hemL	*hemB*
hemB	*hemD*
hemD	*hemC*
hemC	*hemX*
hemX	*hemA*
ysxC	*lonA*
lonA	*lonB*
lonB	*clpX*
clpX	*tig*
tig	*ysoA*
ysoA	*leuD*
leuD	*leuC*
leuC	*leuB*
leuB	*leuA*
leuA	*ilvC*
ilvC	*ilvH*
ilvH	*ilvB*
ysnD	*ysnE*
ysnE	*ysnF*
ysnB	*ysnA*
ysnA	*rph*

Gene Pair	
rph	*gerM*
gerM	*racE*
racE	*ysmB*
ysmB	*gerE*
gerE	*ysmA*
ysmA	*sdhB*
sdhB	*sdhA*
sdhA	*sdhC*
lysC	*ask*
ask	*uvrC*
uvrC	*trxA*
trxA	*xsa*
xsa	*etfA*
etfA	*etfB*
etfB	*ysiB*
ysiB	*ysiA*
ysiA	*lcfA*
lcfA	*yshE*
yshE	*mutSB*
mutSB	*yshC*
yshC	*yshB*
yshB	*yshA*
pheT	*pheS*
pheS	*ysgA*
sspI	*ysfB*
ysfB	*ysfC*
ysfC	*ysfD*
ysfE	*cstA*
cstA	*abfA*
abfA	*araQ*
araQ	*araP*
araP	*araN*
araN	*araM*
araM	*araL*
araL	*araD*
araD	*araB*
araB	*araA*
araA	*abnA*
abnA	*ysdC*
ysdA	*rplT*
rplT	*rpmI*
rpmI	*infC*
yscA	*yscB*
ysbB	*ysbA*
ysbA	*lytT*
lytT	*lytS*
thrS	*ytxC*
ytxC	*ytxB*
ytxB	*dnaI*
dnaI	*dnaB*
dnaB	*ytcG*
ytcG	*speD*
speD	*gapB*
gapB	*ytcD*
ytbD	*ytbE*
ytaG	*ytaF*
ytaF	*mutM*
mutM	*polA*
polA	*phoR*
phoR	*phoP*

Appendix C. (Continued)

Gene Pair	
phoP	*mdh*
mdh	*icd*
icd	*citZ*
citZ	*ytwI*
ytzA	*pyk*
pyk	*pfkA*
pfkA	*accA*
accA	*accD*
accD	*ytsJ*
ytsJ	*dnaE*
ytoI	*ytnM*
ytnM	*ytnL*
ytnL	*ribR*
ribR	*ytnJ*
ytnJ	*ytnI*
ytnI	*ytmO*
ytmO	*ytmN*
ytmN	*ytmM*
ytmM	*ytmL*
ytmL	*ytmK*
ytmK	*ytmJ*
ytmJ	*ytmI*
ytkL	*ytkK*
ytkK	*ytzD*
ytzD	*argH*
argH	*argG*
argG	*moaB*
moaB	*ackA*
ackA	*ytxK*
ytxK	*tpx*
tpx	*ytfJ*
ytfJ	*ytfI*
ytfI	*yteJ*
yteJ	*sppA*
ytcJ	*ytcI*
ytcI	*sspA*
sspA	*ytbJ*
ytbJ	*nifZ*
tyrS	*acsA*
acuA	*acuB*
acuB	*acuC*
ytxE	*ytxD*
ytxD	*ccpA*
ccpA	*aroA*
aroA	*ytxJ*
ytxJ	*ytxH*
ytxH	*ytxG*
ytxG	*murC*
murC	*ytpT*
ytpT	*ytpS*
ytpS	*ytpR*
ytpR	*ytpQ*
ytpQ	*ytpP*
malS	*ytnP*
ytnP	*ytmQ*
ytmP	*amyX*
amyX	*ytlR*
ytlR	*ytlQ*
ytlQ	*ytlP*
ythQ	*ythP*

Gene Pair	
ytzF	*ytzG*
ytzG	*ytgP*
ytfP	*opuD*
yteV	*yteU*
yteU	*yteT*
yteT	*yteS*
yteS	*yteR*
yteR	*yteQ*
yteQ	*yteP*
ytcQ	*ytcP*
ytcP	*ytbQ*
ytbQ	*bioI*
bioI	*bioB*
bioB	*bioD*
bioD	*bioF*
bioF	*bioA*
bioA	*bioW*
bioW	*ytaP*
msmR	*msmE*
msmE	*amyD*
amyD	*amyC*
amyC	*melA*
ytwF	*leuS*
leuS	*ytvB*
ytsD	*ytsC*
ytsC	*ytsB*
ytsB	*ytsA*
ytsA	*ytrF*
ytrF	*ytrE*
ytrE	*ytrD*
ytrD	*ytrC*
ytrC	*ytrB*
ytrB	*ytrA*
ytrA	*ytzC*
ytqA	*ytqB*
ytpB	*ytpA*
ytnA	*asnB*
asnB	*metK*
ytmB	*ytmA*
ytlA	*ytlB*
ytlB	*ytlC*
ytlC	*ytlD*
ytkD	*ytkC*
ytkC	*dps*
dps	*ytkA*
ytkA	*luxS*
ytiB	*ytiA*
ythA	*ythB*
ythB	*ythC*
mntD	*mntC*
mntC	*mntB*
mntB	*mntA*
mntA	*menC*
menC	*menE*
menE	*menB*
menB	*ytxM*
ytxM	*menD*
menD	*menF*
ytcA	*ytcB*
ytcB	*ytcC*

Gene Pair	
ytxO	*cotS*
cotS	*cotSA*
ytaA	*ytaB*
glgP	*glgA*
glgA	*glgD*
glgD	*glgC*
glgC	*glgB*
yuaI	*yuaG*
yuaG	*yuaF*
yuaD	*gbsB*
gbsB	*gbsA*
yuaC	*yuaB*
yuaB	*yuaA*
yuaA	*yubG*
yubF	*yubE*
yubE	*yubD*
yubB	*yubA*
yulE	*yulD*
yulD	*yulC*
yulC	*yulB*
yulB	*yuxG*
yuxG	*tlpB*
tlpB	*mcpA*
mcpA	*tlpA*
tlpA	*mcpB*
yugU	*yugT*
yugT	*yugS*
yugS	*yugP*
yugN	*yugM*
yugM	*pgi*
pgi	*yugK*
yugK	*yugJ*
yugI	*alaT*
alaT	*alaR*
patB	*kinB*
kinB	*kapB*
yuxJ	*pbpD*
pbpD	*yuxK*
yufL	*yufM*
yufM	*yufN*
yufN	*yufO*
yufO	*yufP*
yufP	*yufQ*
yufQ	*maeN*
mrpA	*mrpB*
mrpB	*mrpC*
mrpC	*mrpD*
mrpD	*mrpE*
mrpE	*mrpF*
mrpF	*mrpG*
yuxO	*comA*
comA	*comP*
comP	*comX*
comX	*comQ*
comQ	*degQ*
yuxH	*yueK*
yueK	*yueJ*
yueJ	*yueI*
yueI	*yueH*
yueH	*yueG*

Gene Pair	
yueG	*yueF*
yuzF	*yueE*
yueE	*yueD*
yueD	*yueC*
yueC	*yueB*
yueB	*yukA*
yukA	*yukB*
yukB	*yukC*
yukC	*yukD*
yukD	*yukE*
yukF	*ald*
ald	*yuxI*
yuxI	*yukJ*
dhbF	*dhbB*
dhbB	*dhbE*
dhbE	*dhbC*
dhbC	*dhbA*
dhbA	*yuiI*
yuiI	*yuiH*
yuiF	*yuiE*
yuiC	*yuiB*
yuiB	*yuiA*
yuiA	*yumB*
paiB	*paiA*
paiA	*yutM*
yutM	*dapF*
dapF	*yutK*
yutK	*yuzB*
yuxL	*thrB*
thrB	*thrC*
thrC	*hom*
hom	*yutH*
yutF	*yutE*
yutE	*yutD*
yunB	*yunC*
yunC	*yunD*
yunD	*yunE*
yunE	*yunF*
yunF	*yunG*
yunG	*pucH*
pucR	*pucJ*
pucJ	*pucK*
pucK	*pucL*
pucL	*pucM*
pucE	*pucD*
pucD	*pucC*
pucC	*pucB*
pucB	*pucA*
pucA	*yurG*
yurG	*yurH*
yurL	*yurM*
yurM	*yurN*
yurN	*yurO*
yurO	*yurP*
yurP	*yurQ*
yurQ	*yurR*
sspG	*yurS*
yurT	*yurU*
yurU	*yurV*
yurV	*csd*

Gene Pair	
csd	*yurX*
yurX	*yurY*
yurY	*yurZ*
yurZ	*yusA*
yusA	*yusB*
yusB	*yusC*
yusC	*yusD*
yusD	*yusE*
yusE	*yusF*
yusF	*yusG*
yusG	*gcvH*
gcvH	*yusI*
yusI	*yusJ*
yusJ	*yusK*
yusK	*yusL*
yusL	*yusM*
yusN	*yusO*
yusO	*yusP*
yusQ	*yusR*
yusR	*yusS*
yusU	*yusV*
yusV	*yusW*
yusW	*yusX*
yusX	*yusY*
yusZ	*mrgA*
cssR	*cssS*
cssS	*yuxN*
gerAA	*gerAB*
gerAB	*gerAC*
yvqC	*yvqE*
yvqE	*yvqF*
yvqF	*yvqG*
yvqG	*yvqH*
yvqH	*yvqI*
yvqI	*yvqJ*
yvqJ	*yvqK*
yvqK	*yvrA*
yvrA	*yvrB*
yvrB	*yvrC*
yvrE	*yvrG*
yvrG	*yvrH*
yvrH	*yvrI*
oxdC	*yvrL*
yvrN	*yvrO*
yvrO	*yvrP*
yvrP	*fhuC*
fhuC	*fhuG*
fhuG	*fhuB*
yvsH	*sspJ*
yvsG	*yvgJ*
yvgL	*yvgM*
yvgP	*yvgQ*
yvgQ	*yvgR*
yvgR	*yvgS*
yvgS	*yvgT*
yvgT	*bdbC*
bdbC	*bdbD*
bdbD	*yvgW*
yvgW	*yvgX*
yvgX	*yvgY*

Gene Pair	
yvgY	*yvgZ*
yvaB	*yvaC*
yvaC	*yvaD*
yvaD	*yvaE*
yvaE	*yvaF*
yvaF	*yvaG*
smpB	*rnr*
rnr	*yvaK*
yvaK	*secG*
yvaN	*yvaO*
yvaO	*yvaP*
yvaP	*yvaQ*
opuBD	*opuBC*
opuBC	*opuBB*
opuBB	*opuBA*
yvaV	*yvaW*
yvaW	*yvaX*
yvaX	*yvaY*
yvaZ	*yvbA*
yvbA	*opuCD*
opuCD	*opuCC*
opuCC	*opuCB*
opuCB	*opuCA*
yvbH	*yvbI*
eno	*pgm*
pgm	*tpiA*
tpiA	*pgk*
pgk	*gapA*
gapA	*cggR*
cggR	*araE*
yvbT	*yvbU*
yvbW	*yvbX*
yvbX	*yvbY*
yvbY	*yvfW*
yvfW	*yvfV*
yvfV	*yvfU*
yvfU	*yvfT*
yvfT	*yvfS*
yvfS	*yvfR*
rsbQ	*rsbP*
yvfO	*lacA*
lacA	*yvfM*
yvfM	*yvfL*
yvfL	*yvfK*
yvfK	*lacR*
lacR	*yvfI*
yvfF	*yvfE*
yvfE	*yvfD*
yvfD	*yvfC*
yvfC	*yvfB*
yvfB	*yvfA*
yvfA	*yveT*
yveT	*yveS*
yveS	*yveR*
yveR	*yveQ*
yveQ	*yveP*
yveP	*yveO*
yveO	*yveN*
yveN	*yveM*
yveM	*yveL*

Appendix C. (Continued)

Gene Pair	
yveL	*yveK*
slr	*pnbA*
padC	*yveG*
yveG	*yveF*
yveF	*racX*
racX	*pbpE*
sacB	*yveB*
yveB	*yveA*
yveA	*yvdT*
yvdT	*yvdS*
yvdS	*yvdR*
yvdQ	*yvdP*
pgcM	*malL*
malL	*yvdK*
yvdK	*yvdJ*
yvdJ	*yvdI*
yvdI	*yvdH*
yvdH	*yvdG*
yvdG	*yvdF*
yvdF	*yvdE*
yvdE	*yvdD*
yvdB	*yvdA*
yvcS	*yvcR*
yvcR	*yvcQ*
yvcQ	*yvcP*
yvcP	*yvcN*
yvcN	*crh*
crh	*yvcL*
yvcL	*yvcK*
yvcK	*yvcJ*
yvcJ	*yvcI*
yvcI	*trxB*
trxB	*yvcE*
yvcE	*yvcD*
yvcD	*yvcC*
yvcC	*yvzA*
yvzA	*yvcB*
yvcB	*yvcA*
yvcA	*hisI*
hisI	*hisF*
hisF	*hisA*
hisA	*hisH*
hisH	*hisB*
hisB	*hisD*
hisD	*hisG*
hisG	*hisZ*
yvpB	*yvpA*
yvoF	*hprP*
hprP	*yvoD*
yvoD	*lgt*
lgt	*hprK*
nagA	*nagB*
nagB	*yvoA*
cypX	*yvmC*
yvmB	*yvmA*
yvlD	*yvlC*
yvlC	*yvlB*
yvlB	*yvlA*
yvzB	*uvrA*
uvrA	*uvrB*

Gene Pair	
uvrB	*csbA*
yvkC	*yvkB*
yvkB	*yvkA*
yvjD	*yvzD*
yvzD	*yvjB*
yvjB	*ftsX*
ftsX	*ftsE*
ftsE	*cccB*
cccB	*yvjA*
yvjA	*prfB*
prfB	*secA*
secA	*yvyD*
yvyD	*fliT*
fliT	*fliS*
fliS	*fliD*
fliD	*yvyC*
yvyC	*hag*
hag	*csrA*
csrA	*yviF*
yviF	*yviE*
yviE	*flgL*
flgL	*flgK*
flgK	*yvyG*
yvyG	*flgM*
flgM	*yvyF*
yvyF	*comFC*
comFC	*comFB*
comFB	*comFA*
comFA	*yviA*
yviA	*degU*
degU	*degS*
yvyE	*yvhJ*
tagO	*tuaH*
tuaH	*tuaG*
tuaG	*tuaF*
tuaF	*tuaE*
tuaE	*tuaD*
tuaD	*tuaC*
tuaC	*tuaB*
tuaB	*tuaA*
tuaA	*lytC*
lytC	*lytB*
lytB	*lytA*
ggaB	*ggaA*
ggaA	*tagH*
tagH	*tagG*
tagG	*tagF*
tagF	*tagE*
tagE	*tagD*
tagA	*tagB*
tagB	*tagC*
lytD	*pmi*
gerBA	*gerBB*
gerBB	*gerBC*
ywtF	*ywtE*
ywtD	*ywtC*
ywtC	*ywtB*
ywtB	*ywtA*
ywtA	*ywsC*
rbsR	*rbsK*

Gene Pair	
rbsK	*rbsD*
rbsD	*rbsA*
rbsA	*rbsC*
rbsC	*rbsB*
rbsB	*ywsB*
ywsB	*ywsA*
ywrO	*alsD*
alsD	*alsS*
alsR	*ywrK*
ywrJ	*cotB*
cotB	*cotH*
cotG	*ywrF*
ywrF	*ywrE*
ywrC	*ywrB*
ywrB	*ywrA*
ywqO	*ywqN*
ywqL	*ywqK*
ywqK	*ywqJ*
ywqJ	*ywqI*
ywqI	*ywqH*
ywqH	*ywqG*
ywqG	*ywqF*
ywqF	*ywqE*
ywqE	*ywqD*
ywqD	*ywqC*
ywqB	*ywqA*
ywpJ	*glcR*
glcR	*ywpH*
ywpH	*ywpG*
mscL	*ywpB*
flhP	*flhO*
flhO	*mbl*
mbl	*spoIIID*
spoIIID	*usd*
ywoH	*ywoG*
ywoF	*ywoE*
ywoE	*ywoD*
ywoD	*ywoC*
ywoC	*ywoB*
nrgA	*nrgB*
nrgB	*ywoA*
ywnJ	*spoIIQ*
ywnH	*ywnG*
ywnB	*ywnA*
ywnA	*ureC*
ureC	*ureB*
ureB	*ureA*
ureA	*csbD*
csbD	*ywmF*
ywmF	*rapB*
rapB	*moaA*
moaA	*fdhD*
fdhD	*ywmE*
ywmE	*ywmD*
ywmD	*ywmC*
ywmC	*spoIID*
spoIID	*murAA*
murAA	*ywmB*
ywmB	*ywzB*
atpC	*atpD*

Gene Pair	
atpD	*atpG*
atpG	*atpA*
atpA	*atpH*
atpH	*atpF*
atpF	*atpE*
atpE	*atpB*
atpB	*atpI*
atpI	*upp*
upp	*glyA*
glyA	*ywlG*
ywlG	*ywlF*
ywlF	*ywlE*
ywlE	*ywlD*
ywlD	*ywlC*
ywlC	*ywlB*
ywlB	*spoIIR*
ywkF	*ywkE*
ywkE	*prfA*
ywkD	*ywkC*
ywkB	*ywkA*
ywkA	*tdk*
tdk	*rpmE*
rpmE	*rho*
rho	*ywjI*
ywjI	*murAB*
murAB	*ywjH*
ywjH	*fbaA*
fbaA	*spo0F*
pyrG	*rpoE*
rpoE	*acdA*
acdA	*ywjF*
ywjE	*ywjD*
ywjD	*ywjC*
ywjB	*ywjA*
ywjA	*ywiE*
ywiE	*narI*
narI	*narJ*
narJ	*narH*
narH	*narG*
narG	*arfM*
fnr	*narK*
narK	*argS*
argS	*ywiB*
sboA	*sboX*
sboX	*albA*
albA	*albB*
albB	*albC*
albC	*albD*
albD	*albE*
albE	*albF*
albF	*albG*
ywhL	*ywhK*
rapF	*phrF*
phrF	*ywhH*
speB	*speE*
ywhD	*ywhC*
ywhA	*thrZ*
thrZ	*mmr*
mmr	*ywgB*
ywgB	*ywgA*

Gene Pair	
ywgA	*ywfO*
ywfO	*ywzC*
ywfM	*ywfL*
ywfL	*ywfK*
ywfK	*pta*
ywfI	*ywfH*
ywfG	*ywfF*
ywfF	*ywfE*
ywfE	*ywfD*
ywfD	*ywfC*
ywfC	*ywfB*
ywfB	*ywfA*
ywfA	*rocC*
rocC	*rocB*
rocB	*rocA*
rocA	*rocG*
rocG	*yweA*
yweA	*spsL*
spsL	*spsK*
spsK	*spsJ*
spsJ	*spsI*
spsI	*spsG*
spsG	*spsF*
spsF	*spsE*
spsE	*spsD*
spsD	*spsC*
spsC	*spsB*
spsB	*spsA*
ywdK	*ywdJ*
ywdJ	*ywdI*
ung	*ywdF*
ywdF	*ywdE*
ywdE	*ywdD*
ywdC	*thiD*
ywdA	*sacA*
sacA	*sacP*
sacT	*ywcI*
ywcH	*nfrA*
nfrA	*rodA*
qoxD	*qoxC*
qoxC	*qoxB*
qoxB	*qoxA*
galT	*galK*
galK	*ywcD*
ywcD	*ywcC*
ywcB	*ywcA*
ywbO	*ywbN*
ywbN	*ywbM*
ywbM	*ywbL*
ywbL	*thiE*
thiE	*thiM*
thiM	*ywbI*
ywbH	*ywbG*
ywbG	*ywbF*
ywbF	*ywbE*
ywbB	*ywbA*
epr	*sacX*
sacX	*sacY*
gspA	*ywaF*
ywaE	*tyrZ*

Gene Pair	
tyrZ	*ywaD*
ywaC	*menA*
dltA	*dltB*
dltB	*dltC*
dltC	*dltD*
dltD	*dltE*
dltE	*ywaA*
licH	*licA*
licA	*licC*
licC	*licB*
licB	*licR*
licR	*yxzF*
yxzF	*yxlJ*
katX	*yxlH*
yxlG	*yxlF*
yxlF	*yxlE*
yxlE	*yxlD*
yxlD	*yxlC*
yxlC	*sigY*
yxkO	*cydD*
cydD	*cydC*
cydC	*cydB*
cydB	*cydA*
yxkJ	*yxkI*
yxkI	*yxzE*
yxkH	*msmX*
msmX	*yxkF*
galE	*yxkA*
yxjO	*yxjN*
yxjN	*yxjM*
yxjM	*yxjL*
yxjL	*pepT*
pepT	*yxjJ*
yxjJ	*yxjI*
yxjI	*yxjH*
yxjH	*yxjG*
yxjF	*scoB*
scoB	*scoA*
scoA	*yxjC*
yxjC	*yxjB*
yxiT	*yxiS*
yxiS	*katE*
bglS	*licT*
licT	*yxiP*
deaD	*yxiM*
yxiM	*yxiL*
yxiL	*yxiK*
yxiK	*yxiJ*
yxiJ	*yxiI*
yxiI	*yxzG*
yxzG	*yxiH*
yxiH	*yxiG*
yxiG	*yxzC*
yxzC	*yxiF*
yxiF	*yxxG*
yxxG	*wapA*
wapA	*yxxF*
yxxF	*yxiE*
yxiE	*bglH*
bglH	*bglP*

Appendix C. (Continued)

Gene Pair	
bglP	*yxxE*
yxxE	*yxxD*
yxxD	*yxiD*
yxiD	*yxiC*
yxiC	*yxiB*
yxiB	*yxiA*
hutP	*hutH*
hutH	*hutU*
hutU	*hutI*
hutI	*hutG*
hutG	*hutM*
pdp	*nupC*
nupC	*dra*
dra	*deoR*
deoR	*yxxB*
yxxB	*yxeR*
yxeR	*yxeQ*
yxeQ	*yxeP*
yxeP	*yxeO*
yxeO	*yxeN*
yxeN	*yxeM*
yxeM	*yxeL*
yxeL	*yxeK*
yxeK	*yxeJ*
yxeJ	*yxeI*
yxeI	*yxeH*
yxeH	*yxeG*
yxeG	*yxeF*
yxeE	*yxeD*
yxeA	*yxdM*
yxdM	*yxdL*
yxdL	*yxdK*
yxdK	*yxdJ*
yxdJ	*fbaB*
fbaB	*iolI*
iolI	*iolH*
iolH	*idh*
idh	*iolF*
iolF	*iolE*
iolE	*iolD*
iolD	*iolC*
iolC	*iolB*
iolB	*mmsA*
iolR	*iolS*
iolS	*yxcE*
yxcE	*yxcD*
yxcD	*csbC*
htpG	*yxcA*
yxbD	*yxbC*
yxbB	*yxbA*
yxbA	*yxnB*
yxnB	*asnH*
asnH	*yxaM*
yxaL	*yxaJ*
yxaH	*yxaG*
yxaG	*yxaF*
yxaB	*yxaA*
gntR	*gntK*
gntK	*gntP*
gntP	*gntZ*

Gene Pair	
gntZ	*ahpC*
ahpC	*ahpF*
bglA	*yyzE*
yydJ	*yydI*
yydI	*yydH*
yydH	*yydG*
yydG	*yydF*
yydD	*yydC*
yydC	*yydB*
yydB	*yydA*
yycR	*yycQ*
yycQ	*yycP*
yycP	*yycO*
yycO	*yycN*
rapG	*phrG*
rocF	*rocE*
rocE	*rocD*
yyxA	*yycJ*
yycJ	*yycI*
yycI	*yycH*
yycH	*yycG*
yycG	*yycF*
purA	*yycE*
yycE	*dnaC*
yycC	*yycB*
yycB	*yycA*
rplI	*yybT*
yybT	*yybS*
yybO	*yybN*
yybN	*yybM*
yybM	*yybL*
yybL	*yybK*
yybK	*yybJ*
yybI	*yybH*
yybE	*yybD*
yybD	*yybC*
yybB	*yybA*
yyaT	*yyaS*
yyaQ	*yyaP*
tetB	*tetL*
yyaO	*yyaN*
yyaN	*yyaM*
yyaM	*yyaL*
maa	*yyaH*
yyaH	*ccpB*
ccpB	*exoA*
exoA	*rpsR*
rpsR	*ssb*
ssb	*rpsF*
rpsF	*yyaF*
yyaF	*yyaE*
yyaE	*yyaD*
spo0J	*soj*
yyaA	*gidB*
gidB	*gidA*
gidA	*thdF*
thdF	*jag*
jag	*spoIIIJ*
spoIIIJ	*rnpA*
rnpA	*rpmH*