Nucleic Acids Research Advance Access published October 24, 2007

Nucleic Acids Research, 2007, 1–15
doi:10.1093/nar/gkm501

Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*

Salim Charaniya[1], Sarika Mehra[3], Wei Lian[1], Karthik P. Jayapal[1], George Karypis[2] and Wei-Shou Hu[1,*]

[1]Department of Chemical Engineering and Materials Science and [2]Department of Computer Science and Engineering, University of Minnesota, 421 Washington Avenue SE, Minneapolis, MN 55455-0132, USA and [3]Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076, India

Received May 18, 2007; Accepted June 7, 2007

ABSTRACT

***Streptomyces* spp. produce a variety of valuable secondary metabolites, which are regulated in a spatio-temporal manner by a complex network of inter-connected gene products. Using a compilation of genome-scale temporal transcriptome data for the model organism, *Streptomyces coelicolor*, under different environmental and genetic perturbations, we have developed a supervised machine-learning method for operon prediction in this microorganism. We demonstrate that, using features dependent on transcriptome dynamics and genome sequence, a support vector machines (SVM)-based classification algorithm can accurately classify >90% of gene pairs in a set of known operons. Based on model predictions for the entire genome, we verified the co-transcription of more than 250 gene pairs by RT-PCR. These results vastly increase the database of known operons in *S. coelicolor* and provide valuable information for exploring gene function and regulation to harness the potential of this differentiating microorganism for synthesis of natural products.**

INTRODUCTION

Transcriptional regulation is perhaps the most fundamental control in gene expression. Many functionally related genes are often co-regulated, meaning that their expression is coordinated temporally or even spatially in response to the need of the organism in a given environmental condition. In prokaryotes these co-regulated genes are often organized in their genomes into physical clusters called operons. An operon thus consists of more than one adjacent gene expressed as a transcription unit. Operons allow an organism to simultaneously express the genes that are needed for cell survival under the same condition, providing a control circuit that is both simple and economical. In some cases, however, there is also a need to fine-tune the expression of individual genes in an operon under some circumstances. This is accomplished by alternative regulation of genes, which are normally co-regulated in one operon (1). Transcription of a unit encoding a single gene or an operon is controlled by a promoter and a terminator. Alternative regulation in an operon is accomplished by one or more alternative promoters or internal transcription terminator.

In the past few years numerous bacterial genomes have been completely sequenced and the number is steadily increasing. Identifying potential operons in those genomes facilitates the functional annotation of the genes involved and is important in elucidating the regulation of those genes. Several approaches have been previously used for operon predictions. Most methods rely on features based on the genome structure or the functional similarity of genes of interest. Since adjacent genes in an operon often are physically closer to each other than those not in the same operon, intergenic distance provides information about the likelihood that two adjacent genes may be on the same operon (2). The conservation of gene order in multiple organisms is also taken into account (3). Additionally, the similarity of codon usage (the frequency with which synonymous codons encode amino acids in neighboring genes) is also used for operon predictions (4). Since genes on the same operon are co-regulated, at least under the conditions when alternative regulation is not in play, their transcription profiles are likely to be well correlated. Identifying adjacent genes whose transcription levels are well correlated also provides much information on the likelihood of their being in the same operon (5).

*To whom correspondence should be addressed. Tel: +1 612 625 0546; Fax: +1 612 626 7246; Email: acre@cems.umn.edu
Present Address:
Wei Lian, Abbott Bioresearch Center, 100 Research Drive, Worcester, MA 01605 USA

Unsupervised Bayesian methods using features based on genome sequence and functional similarity have been reported for operon prediction in all sequenced prokaryotes (3,4,6). An empirical scoring method has also been reported previously (7). Since these methods do not require a training set, they are advantageous for organisms where little or no information about known operons is available. Alternatively, machine-learning approaches have also been used to train models based on databases of known operons. Studies have shown that log-likelihoods derived from distribution of intergenic distance in a set of known operons can be used for operon prediction in several prokaryotes (2,8). Naïve Bayesian classifier as well as C5.0, a decision tree-based algorithm, have been reported for predicting operons in *Escherichia coli* (9,10). A support vector machines (SVM)-based model has recently been reported for operon prediction in *E. coli* and *Bacillus subtilis* (11). Few reported methods have combined transcriptome data and genome sequence for predicting operon structure. A hidden Markov model based on expression data alone has been reported for *E. coli* (12). Bayesian methods that combine similarity of transcript profiles with information based on genome sequence have been previously used for operon prediction in *E. coli* and *B. subtilis* (5,13,14).

Streptomyces coelicolor, with an 8.7 Mbp linear chromosome and approximately 7800 predicted ORFs, has one of the largest completely sequenced bacterial genomes. The genome encodes 20 secondary metabolite gene clusters including clusters for three antibiotics—actinorhodin (Act), undecylprodigiosin (Red) and calcium-dependent antibiotic (CDA) (15). It belongs to the genus *Streptomyces* whose members are widespread in soil, have complex multicellular lifecycle, and produce nearly two-thirds of reported naturally occurring antibiotics and a variety of other natural compounds including anti-tumor agents and immunosuppressants (16). With its complex life cycle and capacity to produce numerous antibiotics, *S. coelicolor* leads a very dynamic life cycle of vegetative growth and sporulation, undergoing changes from primary metabolism to secondary metabolite (antibiotic) production. Changes in environmental factors and growth conditions, or perturbations through genetic mutations often result in major changes in its transcriptome profile (17–19). Like many other bacteria, genes of the same functional class or in the same pathway are often organized into operons. However, with only around 50 previously reported operons, its operon structure is not well characterized. The recent development of whole genome microarray for *S. coelicolor* has generated an increasing amount of transcriptome data obtained from different mutants and/or different culture conditions. This data can potentially be used to discern transcriptional co-regulation and identify operons; thus, facilitating gene annotation and providing valuable functional information.

In this study, we employed genome-wide temporal transcriptome data from several strains and culture conditions, information about intergenic distance and transcription terminator predictions, and applied a SVM-based model for operon prediction in the entire genome of *S. coelicolor*. The model predicts more than 2000 gene pairs as being co-transcribed, of which 250 were subsequently experimentally verified. This report demonstrates the application of SVM for operon prediction and verification.

MATERIALS AND METHODS

Microarray data

Strains and culture conditions. *Streptomyces coelicolor* A3(2) strain M145 (prototroph, SCP1$^-$, SCP2$^-$) and mutant strains of two regulatory genes were used—YSK3225 (M145 Δ*absA1::apr*) and YSK4425 (M145 Δ*afsS::apr*). The strains were grown in batch cultures in liquid medium as described in an earlier report (19).

Probe preparation and microarray hybridization. Temporal transcriptome profiling was performed using a whole genome DNA microarray of *S. coelicolor* that has probes for 7579 genes (19). Cell samples were taken at different time points along the culture for transcriptome profiling. RNA extraction, cDNA synthesis, microarray hybridization, washing, scanning and image analysis was performed as described elsewhere (19). Genomic DNA (gDNA) was used as a common reference for all the hybridizations. Details for all the protocols are available at http://hugroup.cems.umn.edu/Protocols/protocol.htm.

Microarray data compilation and processing. The time series microarray data comprise 67 cell samples from three different strains—27 samples from wild type (M145) (GEO accession numbers: GSE8084, GSE8086, GSE8107) and 40 samples from two mutant strains - YSK3225 (GSE8108, GSE8109) and YSK4425 (GSE8110, GSE8160). The data was arranged as three sets (set 1–3) as shown in Supplementary Table S1. All hybridizations were performed using genomic DNA as a reference (cDNA:gDNA). The data was normalized by quantile normalization method, which assumes that the overall distribution of total mRNA is the same for different RNA samples (19,20).

Transcriptome data publicly available in the Stanford Microarray Database (SMD), comprising time series experiments reported by Karoonuthaisiri *et al*. (21) on two *S. coelicolor* strains—M145 and M600, under different stress conditions was also compiled. In these experiments hybridization was performed by pairing two cDNA samples (cDNA:cDNA) with one being $t = 0$ h cDNA sample, used as a reference in most cases. The data from 61 samples was arranged as three different sets (set 4–6) depending on the type of experiment, strain and growth medium used (Supplementary Table S1). Additionally temporal transcriptome data reported by Huang *et al*. (18) on *S. coelicolor* A3(2) M145, J1501 and several mutant strains was also compiled. This dataset, which included 48 cDNA:cDNA measurements and 30 cDNA:gDNA measurements was arranged as three sets (set 7–9) as shown in Supplementary Table S1. The experiments with genomic DNA as reference were quantile normalized. Several genes in each array sample were flagged 'absent' due to low intensity, small spot

diameter or low spot regression coefficient. Samples with >25% genes flagged, were discarded before further processing.

Similarity between the transcript levels of genes in every pair was calculated by the Pearson correlation coefficient (r). For calculation of Shannon entropy, the samples with cDNA:gDNA measurements were standardized by dividing the cDNA/gDNA ratio for every gene by the cDNA/gDNA ratio of that gene in the first time point of M145 in set 1.

Genome organization

The genome sequence of *S. coelicolor* and the annotation files were obtained from The Sanger Institute (ftp://ftp.sanger.ac.uk/pub/S_coelicolor/). The leading and lagging strands were scanned and pairs of genes were grouped based on whether they were transcribed in the same directions (same-strand gene pairs) or in different directions (opposite-strand gene pairs). The 7825 genes in the linear chromosome were binned into 4965 same-strand pairs and 2859 opposite-strand pairs as shown in Figure 1.

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

Prediction of transcription terminators

The presence of rho-independent transcription terminator in the intergenic region of every gene pair was predicted by the TransTerm algorithm (22). The algorithm searches for mRNA motifs that potentially form a hairpin structure and are followed by a short uracil-rich region both within and between the genes. The stability of the hairpin structure and the presence of the U-rich region are characterized by a score that is used to estimate a confidence score/probability of the presence of terminator at a particular site in the genome. Using a confidence level of 0.9 that has been reported to identify 95% of known terminators in *E. coli* (22), we searched all the gene pairs in *S. coelicolor* for which the probability of the presence of terminator in the intergenic region is 0.9 or higher.

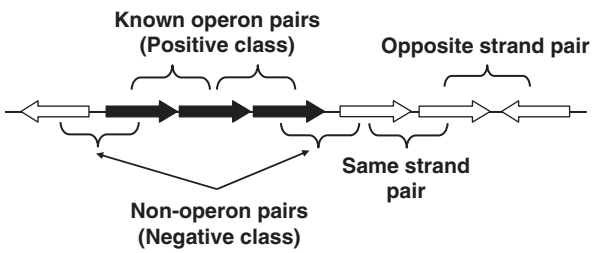

Figure 1. Definition of known operon pairs (KOPs), non-operon pairs (NOPs), same-strand pairs and opposite-strand pairs. Closed-block arrows indicate genes in a known operon. Open-block arrows represent genes with unknown operon status.

Experimental verification of operons

Culture condition, RNA extraction and cDNA synthesis. *Streptomyces coelicolor* M145 wild-type spores were grown in batch culture in modified R5 liquid medium (17), as described elsewhere (19), and samples were withdrawn periodically for RNA extractions. The mycelia was fragmented in liquid nitrogen using mortar and pestle and total RNA was extracted using RNeasy Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. Residual genomic DNA was digested using Turbo DNA-free™ kit (Ambion, Austin, TX, USA) according to the protocol suggested by the manufacture for rigorous DNase treatment. Total RNA was suspended in 50 µl of nuclease-free water and stored at −80°C until further use.

Equal amounts of RNA from four samples corresponding to exponential, late exponential, transition and stationary phase, were pooled and reverse transcribed using random hexamers and Superscript™ III (Invitrogen, Carlsbad, CA, USA) at 50°C for 1 h according to manufacturer's protocol. Fifty nanogram of random hexamer was used for every 5 µg of total RNA. A negative control was also done without the addition of the reverse transcriptase enzyme. Thereafter, the RNA was digested by addition of RNase H (Invitrogen) and incubation at 37°C for 20 min. cDNA was stored at −20°C until further use.

PCR. Gene-specific primers used for whole genome microarray construction (19), were used for RT-PCR based verification of transcripts. To confirm that a pair of adjacent genes is on the same mRNA transcript, the 5′ primer of the first gene and the 3′ primer of the second gene were combined to form a primer pair at a working concentration of 5 µM for each primer. The length of the amplicon for this primer pair was obtained from the chromosomal location of the primers, obtained by blasting the primer sequences against a database of *S. coelicolor* genome, and ranged from 300 bp to 2.5 kb for the gene pairs that were tested.

cDNA from 100 ng of pooled RNA was used as template for every PCR reaction. PCR was also performed on an equivalent amount of negative control from cDNA synthesis to check for any residual genomic DNA contamination in the RNA samples. The PCR conditions were as follows: 5 min of initial denaturation at 95°C, 40 cycles of amplification—denaturation for 30 s at 94°C, annealing for 30 s at a temperature between 60 and 64°C depending on the melting temperature of the primers, and extension at 72°C for 150 s. The final extension was done at 72°C for 5 min. The total reaction volume was 50 µl and 20 µl was analyzed on 1% (w/v) agarose gel.

Supervised classification

Support Vector Machines (SVM). SVM are a class of kernel-based machine-learning methods that use the principle of structural risk minimization to identify a decision function that separates objects from two classes with maximum margin (23,24). SVM^{light}, an implementation of SVM in C was used for model training and

evaluation (25). Two of the various kernel functions, linear and radial basis function (RBF), were used for classification. A linear kernel (k) measures the similarity between two training objects ($\mathbf{x}_1$ and $\mathbf{x}_2$) as a dot product in the input feature space, $k(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2'$. The radial kernel function transforms the data using the non-linear function, $k(\mathbf{x_1},\mathbf{x_2}) = \exp(-\gamma\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$, where γ determines RBF width. For radial kernel function, the parameters γ (–g) and the cost function (–c) were selected using the leave-one-out model selection (looms) procedure (26). The algorithm calculates the leave-one-out error rates for a range of parameters and outputs the one with minimum error rates.

Training set—Positive and negative classes. The training set consists of 49 known operons compiled from literature. An additional six known operons in *Streptomyces lividans* were also included in the training set. *S. lividans* and *S. coelicolor* are close relatives with 99.6% similarity in their 16s rRNA sequences (27), and common structural and genetic organization (28). To increase the number of known operons in the training set, an additional eight operons reported in other *Streptomyces* spp. (*S. griseus, S. antibioticus, S. ambofaciens, S. ramocissimus, S. thermoviolaceus* and *Streptomyces* sp. NRRL 5331) with a conserved gene order in *S. coelicolor* were also included in the training set. Here, 27, 12, 17, 2 and 5 of the known operons have 2, 3, 4, 5 and 6 or more genes, respectively.

The gene pairs formed by consecutive genes in the known operons were referred to as *known operon pairs* (KOPs), as shown in Figure 1. The resulting 149 KOPs constitute the positive class of the training set. The set of gene pairs that comprise the negative class was created as follows. The first gene of every known operon and the gene immediately upstream, as well as the last gene in every known operon and the gene immediately downstream form *non-operon pairs* (NOPs) (Figure 1). The resulting set of 122 NOPs constitutes the negative class. Nine of the known operons have internal regulation with one or more internal promoters or a transcriptional terminator. For these operons, the pair of genes on either side of the internal control element was not considered as a KOP.

Model training and selection. Binary SVM classifiers were trained for operon prediction using three different features—intergenic distance, correlation of transcript profiles and transcription terminator predictions. Intergenic distance is measured in base pairs and varies from −26 to 811 bp in the training set, whereas 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 seven bins corresponding to $d \leqslant 0$, $0 < d \leqslant 20$, $20 < d \leqslant 50$, $50 < d \leqslant 100$, $100 < d \leqslant 200$, $200 < d \leqslant 300$ and $d > 300$ bp.

The discrimination rule established during training can result in *overfitting* whereby the classifier cannot accurately discriminate test/unseen data. Leave-one-out and k-fold cross-validation was thus performed to estimate the performance of the model in classifying an independent dataset that was not used for training (i.e. assess its generalizability) (29).

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 NOP) is compared with the predicted class. The performance of the model is then evaluated using different metrics.

Evaluation metrics. The following metrics were used to compare the performance of different classifiers.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{False positive rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{Total error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

$$\text{F} - \text{factor} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

where, TP (true positives) = Number of KOPs accurately classified as operon pairs by the model.

FN (false negatives) = Number of KOPs falsely classified as non-operon pairs by the model.

FP (false positives) = Number of NOPs falsely classified as operon pairs.

TN (true negatives) = Number of NOPs accurately classified as non-operon pairs.

Recall quantifies the sensitivity of the model—how many KOPs can be predicted as operon pairs by the model and precision quantifies the specificity of the model—how many of the operon pairs predicted from the training set (KOPs and NOPs) are KOPs. F-factor combines the two metrics to quantify the overall performance of the model. The F-factor can range from 0 to 1 with 1 corresponding to an ideal classifier.

K-fold cross-validation. A stratified 5-fold cross-validation procedure was implemented to compare the performance of classifiers with different features. In this procedure, the training set was randomly divided into five subsets, where each subset was stratified such that it contains the same proportion of KOPs and NOPs as the original training set. Four subsets were used for training the model which was then used to assign a score (s) to every gene pair in the 5th test subset. The procedure was repeated five times. This 5-fold cross-validation was performed five times (5×5) and the true class of the gene pairs in each of the 25 test subsets and their scores were then used to generate receiver operating characteristics (ROC) graphs. An ROC curve is a plot of recall as a function of FPR. Using the test subsets, 25 ROC graphs were generated for each classifier. Instead of merging the 25 ROC graphs to one large set and calculating a single ROC curve for each classifier, we used the vertical averaging procedure

described by Fawcett (30). This procedure combines the 25 ROC graphs to estimate the average recall and its standard deviation (SD) at different FPRs. Briefly, for a fixed value of FPR, each of the 25 ROC graphs are scanned and the maximum recall or true positive rate at that FPR is chosen, using interpolation if necessary. These values are used to compute the average recall and draw confidence intervals (±SD) at the fixed FPR. The FPR can be increased from 0 to 1 in small step sizes to get the average ROC curve. Area under ROC curve (AUC) was used as a scalar measure for comparing the performance of different classifiers—the AUC for a random classifier is 0.5 and that of an ideal classifier is 1.

RESULTS

Known operon pairs have shorter intergenic distance

As described in the Materials and Methods section, from a set of known operons we obtained 149 known operons pairs (KOPs) and 122 non-operon pairs (NOPs). The density distribution of the intergenic distances in KOPs and NOPs is shown in Figure 2. For KOPs, the distribution has a sharp peak around intergenic distance of 0 bp. Sixty-seven (45%) KOPs have an intergenic distance less than 0 bp indicative of a translational overlap between the genes. Fifty-seven of these gene pairs have an overlap of 4 bp. Among them 35 have ATGA as the overlapping sequence, where ATG corresponds to start codon for the second gene and TGA is the stop codon for the first gene. The overlapping sequence in other 22 pairs is GTGA. Since *S. coelicolor* has 72% GC content, GTG is also a commonly observed translational start codon. An overlap of 1 bp between the start and the stop codons of adjacent genes was also observed among five of the KOPs. In contrast, only six (5%) NOPs have an overlap in the intergenic distance.

Although a short intergenic distance is a strong indication of co-transcription, a significant fraction of genes in KOPs are separated by intermediate to large intergenic distance. In the training set, 33 (22%) and 21 (14%) KOPs have an intergenic distance that is greater than 50 and 100 bp, respectively. If only intergenic distance is used for operon prediction based on this training set, using a distance threshold of 50 bp, 116 (78%) KOPs can be classified accurately. However, 19 (16%) NOPs will also be falsely classified as being co-transcribed. If the threshold is increased to 100 bp, 128 (86%) KOPs can be correctly classified with a large false positive rate of 32%.

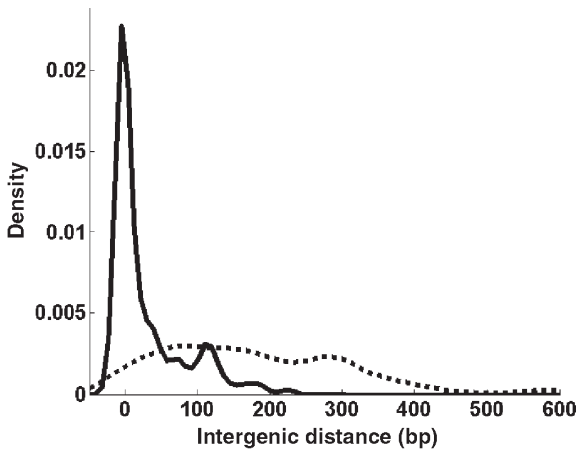

Figure 2. Density distribution of intergenic distance in KOPs and NOPs. (continous line) KOPs; (dashed line) NOPs.

Genes in known operons have greater expression correlation

Genes within an operon are likely to have a higher similarity in their transcript levels compared to genes that are not co-transcribed. Similarity of transcript profiles can be measured using several metrics such as Euclidean distance, cosine function and Pearson correlation coefficient (r). Among these metrics, it has been previously observed that Pearson correlation achieves the best separation between KOPs and NOPs (13).

Temporal transcriptome data obtained from 206 cell samples were divided into nine different sets depending on the experimental design, strains and culture conditions used (Supplementary Table S1). For every KOP, the Pearson correlation between the transcript levels of the adjacent genes was calculated for each of the nine sets, and the number of sets in which the correlation exceeds 0.7 was counted. The KOPs were divided into 10 groups according to the number of sets (0,1,2,...9) in which transcript correlation exceeds 0.7. Figure 3a shows the distribution of the KOPs in different groups. Only one out of 149 KOPs has transcript correlation $r > 0.7$ in all the nine sets. The error in measurement of transcript level due to noise, may have contributed to the relatively low correlation between genes in KOPs. The presence of as yet-unidentified site for internal regulation (internal promoter or transcription terminator), or differential mRNA degradation could also potentially reduce the similarity in transcript level of genes in a KOP. Nonetheless, 58 (39%) KOPs have transcript correlation $r > 0.7$ in four or more sets. In contrast, only six (5%) NOPs have transcript correlation $r > 0.7$ in four or more sets (Figure 3b). Further, 78 (64%) NOPs do not satisfy the correlation threshold of 0.7 in any of the nine sets, in contrast to only 18 (12%) KOPs. This separation between KOPs and NOPs is evident even at higher correlation thresholds. Thirty-two KOPs have transcript correlation $r > 0.8$ in four or more sets in contrast to only one NOP.

To confirm that the higher Pearson correlation in KOPs is not by chance, the correlation between the transcript levels of genes in 20 000 randomly selected pairs was also calculated for all the nine sets. Only 5% of randomly selected gene pairs have $r > 0.7$ in four or more sets (Figure 3c). This indicates that the higher degree of correlation between the transcript levels of genes in KOPs can be used for operon prediction.

Transcription terminators

Using TransTerm which identifies rho-independent transcription terminators, none of the KOPs were found to have a transcription terminator predicted in the intergenic region with a probability of 90% or higher. In contrast, 16 NOPs have a predicted transcription terminator with

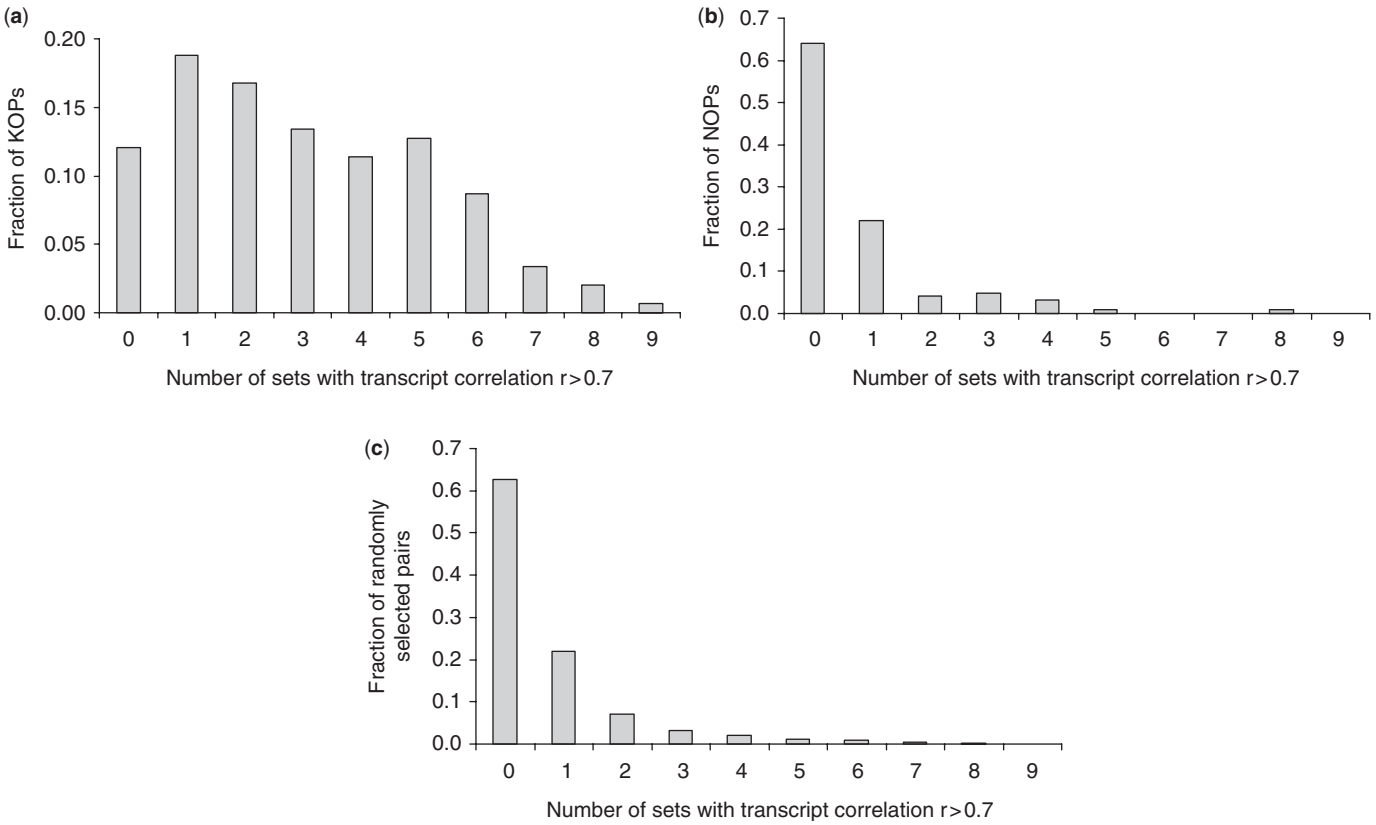

Figure 3. Comparison of Pearson correlation between transcript levels of adjacent genes in KOPs and NOPs. The microarray experiments were divided into nine sets and correlation between transcript levels of adjacent genes in every pair was calculated for each set. The histogram of the number of sets in which correlation exceeds 0.7 in (**A**) Known operon pairs (KOPs); (**B**) Non-operon pairs (NOPs); (**C**) Randomly selected pairs.

90% or greater likelihood, of which nine have a probability greater than 99%. Thus, the probability of presence of a transcription terminator in the intergenic region of gene pairs can also be used as a discriminatory feature for operon prediction.

Binary classification results

Using SVMs as a supervised classification tool, binary classifiers were designed to discriminate KOPs and NOPs using different combinations of features. As described in the Materials and Methods section, leave-one-out and k-fold cross-validation was used for evaluation and selection of the *best* classifier.

Leave-one-out cross-validation results. The performance of different classifiers is shown in Table 1. If only intergenic distance is used for classification of training set (classifier I), 82% of KOPs can be accurately classified as operon pairs with a precision of 86%. However, 16% of NOPs are misclassified as operon pairs. Discretization of distance (classifiers II and III) results in a small reduction in recall (78%) with comparable precision and false positive rates (FPR). If only transcriptome data is used for classification, a radial SVM model (classifier V) with recall and precision of 80 and 82%, respectively, performs marginally better than linear SVM model (classifier IV). However, with an F-factor of 0.838 the performance of distance-based classifier I is slightly better than the transcriptome-based classifier V (F-factor = 0.810). Terminator predictions alone can differentiate only 16 (13%) NOPs due to the presence of a predicted terminator site in their intergenic region. However, the remaining 87% NOPs cannot be differentiated from KOPs resulting in a large FPR (classifier VI in Table 1).

When intergenic distance and transcriptome data are combined, the performance of the linear (classifier VII) as well as the radial SVM classifier (classifier VIII) improves significantly with recall and precision of 90 and 88%, respectively. With an F-factor of 0.89, the classifiers VII and VIII that combine transcriptome data with intergenic distance are better than any of the classifiers that use only one feature (classifier I–VI). The radial model based on all the three features (classifier X) has a marginal improvement in recall (92%) and precision (89%) compared to classifier VII and VIII. Among the various combinations of feature sets and kernel functions, the radial classifier X has the highest recall and precision (Table 1).

Increasing transcriptome data improves prediction accuracy. The performance of a classifier based on transcriptome data is profoundly affected by the diversity of experimental conditions under which microarray experiments are performed. To demonstrate this, we trained an SVM classifier based on transcriptome data

Table 1. Comparison of different classifiers using leave-one-out cross-validation

Classifier	Kernel function	Feature(s)	Recall (%)	Precision (%)	Total error rate (%)	False positive rate (%)	F-factor
I	Radial ($\gamma = 0.01$)	Distance	82	86	17	16	0.838
II	Linear	Distance (discretized)	78	86	19	16	0.817
III	Radial ($\gamma = 0.0025$)		78	86	19	16	0.817
IV	Linear	Transcriptome	78	82	22	21	0.798
V	Radial ($\gamma = 0.02$)		80	82	21	21	0.810
VI	Linear	Terminator prediction	100	58	39	87	0.734
VII	Linear	Distance and transcriptome	90	88	12	15	0.890
VIII	Radial ($\gamma = 0.25$)		90	88	12	15	0.890
IX	Linear	Distance, transcriptome and terminator prediction	90	88	12	15	0.887
X	Radial ($\gamma = 0.25$)		92	89	11	14	0.904

from the time course experiment of M145 wild type in $R5^-$ liquid medium only (set 1 in Supplementary Table S1). The classifier has a recall and precision of 60 and 71%, respectively. In contrast, the radial classifier (classifier V) based on all transcriptome data, has a significantly higher recall and prediction of 80 and 82%, respectively. Thus, addition of microarray experiments performed with different strains and culture conditions can improve the accuracy of operon predictions significantly.

K-fold cross-validation results. In order to compare different feature sets and their combinations, a 5-fold cross-validation (see Materials and Methods section) was performed on classifiers I (intergenic distance), V (transcriptome data), VIII (intergenic distance and transcriptome data) and X (all features). Since the classifier VI based on terminator predictions alone has a large FPR, we did not include it in this comparative study.

ROC graphs were generated for each classifier, as described in the Materials and Methods section. As shown in Figure 4, the classifier V based on transcriptome data results in significant improvement compared to a random classifier (depicted by a diagonal 45° line). Sixty percent of KOPs can be accurately classified with a FPR of 10% indicating that correlation between transcript profiles of adjacent genes can indeed be used for operon prediction. The radial SVM classifier I based on intergenic distance alone has similar recall and FPR as classifier V based on transcriptome data. Combination of these two features in classifier VIII results in a sharp increase in recall. At a FPR of 10% it can classify 75% of KOPs compared to 60% by classifier I. Addition of terminator predictions to intergenic distance and transcriptome data results in a small but noticeable improvement in classification accuracy (classifier X).

A comparison of the AUC of the four classifiers is shown in Table 2. With an AUC of 0.81, there is no significant difference between the distance-based classifier I and the transcriptome-based classifier V (P-value = 0.65, Wilcoxon signed rank test). Discretization of intergenic distance did not result in any decrease or increase in the AUC (data not shown). The radial SVM classifier VIII combining intergenic distance and transcriptome data has an AUC of 0.89, which is significantly greater than the AUC of distance-based classifier I (P-value = 1.1×10^{-4}). The radial classifier X combining all the three features has the largest AUC of 0.91 and is marginally better than classifier VIII (P-value = 6.7×10^{-3}).

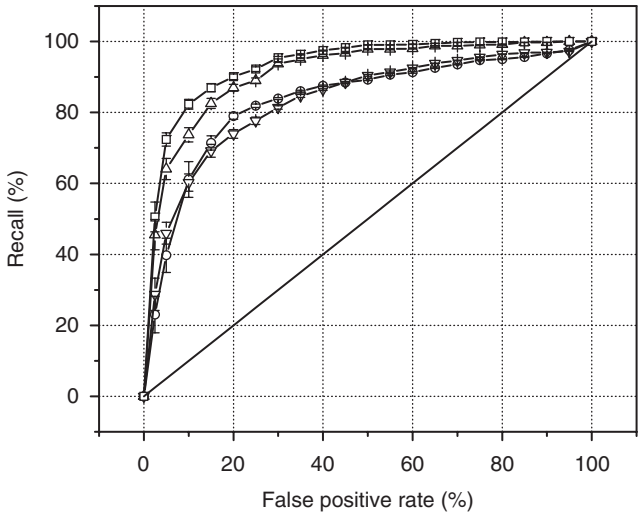

Figure 4. Comparison of different classifiers by ROC curve. False positive rate is the percentage of non-operon pairs (NOPs) misclassified as operon pairs and recall is the percentage of known operon pairs (KOPs) correctly classified as operon pairs. The ROC curves were generated for each classifier by a 5-fold cross-validation as described in the text. (Open circle) classifier I; (Inverted triangle) classifier V; (Open triangle) classifier VIII; (Open square) classifier X.

The radial SVM classifier X was used to assign a score (s) to every gene pair in the training set. A positive score suggests a high likelihood that the adjacent genes are co-transcribed. At a score threshold of zero, 140 (94%) KOPs with a positive score were correctly classified. These 140 KOPs were divided into groups according to different range of scores, corresponding to increasing level of confidence—Group 1 ($0 \leq s < 1$), Group 2 ($1 \leqslant s < 1.2$) and Group 3 ($s \geqslant 1.2$). Among the positive-scoring KOPs, 37, 51, and 52 fall into the three groups, respectively. At higher score threshold in group 3, almost all the KOPs have transcript correlation $r > 0.7$ in at least one set of transcriptome data. Further, 32 (62%) gene pairs in group 3 have $r > 0.7$ in four or more sets. However the transcript correlation between adjacent genes reduces at lower scores in group 1 and group 2. Only seven (19%) of the 37 gene pairs in group 1 have correlation $r > 0.7$ in four or more

Table 2. Comparison of different classifiers by 5-fold cross-validation. The null hypothesis was tested by comparing the AUC of the 25 ROC graphs for each classifier by Wilcoxon signed rank test

Classifier	Feature(s)	Average AUC	*P*-value	Null hypothesis
I	Distance	0.81	–	
V	Transcriptome	0.81	6.5×10^{-1}	$AUC_V - AUC_I = 0$
VIII	Distance and transcriptome	0.89	1.1×10^{-4}	$AUC_{VIII} - AUC_I = 0$
X	Distance, transcriptome and terminator prediction	0.91	1.2×10^{-5}	$AUC_X - AUC_I = 0$
			6.7×10^{-3}	$AUC_X - AUC_{VIII} = 0$

sets. Thus, the score of a KOP reflects the degree of correlation between the transcript levels of adjacent genes—higher score indicative of a stronger correlation.

We also examined whether the extent of perturbation of a gene is important for determining its operon status. Using Shannon entropy of the expression level of a gene across all the microarray experiments as a measure of its perturbation, we found that the average entropy of a pair of genes in group 3 is greater than that in group 1 (*P*-value = 0.04, Kolmogorov–Smirnov test). This suggests that the operon status of adjacent genes with a higher degree of temporal variation in their transcript profiles can, in many cases, be determined with greater confidence.

Identification of transcription units

Prediction of an entire transcription unit requires identification of intracistron genes as well as the genes at the cistron boundary. The genes in a pair with negative score have a low probability of co-transcription and are hence likely to have a cistron boundary between them. To examine the accuracy of our model to predict complete operons, we compared our classification results with known operons in the training set. Twenty-three known operons are dicistronic. All of them have a positive score indicating that they were successfully identified. Moreover, for all these 23 dicistrons, the identified cistron size is two, implying that the cistron boundaries were identified correctly.

Since many operons have more than two genes, it is important to identify adjacent gene pairs that are expressed as one transcription unit. Thirty-one known operons have more than two genes. Of these 9, 15, 2 and 5 operons have 3, 4, 5 and 6 or more genes, respectively. We examined adjacent KOPs and grouped them into operons if their score was greater than zero. Nineteen of those identified polycistrons have the same number of genes as that of the known operon, indicating that all the internal gene pairs as well as the cistron boundaries were correctly identified. Among these 19 operons, 5, 11, 1 and 2 have 3, 4, 5 and 6 or more genes, respectively. Interestingly, among the 31 operons with more than two genes, four have a larger number of genes than the size that has been reported suggesting that additional genes at the cistron boundaries are potentially co-transcribed. As will be described in the experimental verification section, the prediction of those additional genes was verified for three operons by RT-PCR.

For eight of the known operons that have been reported to have more than two genes, the identified cistron size was less than the number of genes reported. They were incorrectly predicted as each consisting of two transcription units because one of the internal KOPs has a negative score. Supplementary Table S2 provides a list of the known operons along with the accuracy of our model to identify them.

Operons with internal regulation

The prediction of operons with internal control elements such as internal promoters, transcription factor binding sites and transcription terminators is a challenging task (4,31). In the training set, nine known operons have been suggested to have internal regulation. Among these, four operons, *litQR* (32), *rsbB-rsbA-sigB* (33,34), *trpCXBA* (35), *ushY-ushX-sigH* (36–38), have internal promoters. Additionally, the *rspO-pnp* operon has an intergenic transcription terminator (39,40). Another dicistronic operon *SCO3661-SCO3660* is induced by heat shock although constitutive expression of *SCO3660* has also been observed (41). The *galTEK* galactose operon and the *recAX* operon involved in SOS response have been characterized in *S. lividans*. The galactose operon has two promoters, one upstream of *galT*, which is induced by galactose and another upstream of *galE* that is constitutively expressed (42). In the *recAX* operon, the *recA* gene is expressed constitutively at a basal level whereas *recA-recX* transcript is observed in response to DNA damage (43). The *rpsL-rpsG-fus-tuf1* operon has an internal promoter upstream of *tuf1* gene in *S. ramocissimus*. However, this promoter sequence is highly conserved among various *Streptomyces* spp. including *S. coelicolor* suggesting the possibility of a common regulatory mechanism (44,45).

The transcript level of the genes in these operons may not be correlated due to internal regulation. We examined the features of the adjacent genes in these operons. In particular, for each of these operons, we examined the transcript correlation and intergenic distance of the pair of genes on either side of the internal regulation site. In three of these nine operons, the gene pair flanking the regulation site has correlation $r < 0.7$ in all the nine sets of transcriptome data. A notable exception is the genes in the dicistron, *rpsO-pnp*, which are strongly correlated ($r > 0.7$) in two of the nine sets of transcriptome data despite a recent report that identified the presence of an intergenic stem-loop structure, which acts as a site for RNase III processing and cleavage (40). Also,

the *rsbA-sigB* gene pair in *rsbB-rsbA-sigB* operon has transcript correlation $r > 0.7$ in two sets, although a developmentally regulated internal promoter has been reported in the *rsbA-sigB* intergenic region (33,34).

Interestingly, six of these nine gene pairs have a large intergenic distance ($d > 150$ bp). Adjacent genes in the same operon are rarely separated by intergenic distance exceeding 200 bp (3,46). Therefore, a high degree of transcript correlation, orthology or functional similarity or combination of all these features is essential to predict the presence of a read-through transcript across these gene pairs. In only two of the nine operons, *galTEK* and *trpCXBA*, the gene with an upstream internal promoter is separated from its upstream neighboring gene by short intergenic distance ($d < 25$ bp).

Operon predictions for entire genome

Using a combination of transcriptome data obtained from several strains and culture conditions, and other features from the genome sequence, the SVM model was successful in classifying 94% of KOPs at a score threshold of zero. None of the features could achieve such a high degree of accuracy when used alone. We therefore used the SVM classifier with all the features for predicting the operon status of all same-strand pairs in *S. coelicolor* genome.

Overall analysis. The entire genome was arranged into 4965 pairs of genes in the same orientation (same-strand pairs) and 2859 pairs of genes in opposite orientation. Excluding the 149 KOPs, the 4816 same-strand gene pairs were further analyzed for co-transcription. The features, intergenic distance, correlation of transcript profiles and the likelihood of a transcription terminator were calculated for everyone of those pairs. TransTerm was used for prediction of transcription terminators (22). Among the 2498 transcription terminators predicted in *S. coelicolor* genome, only 169 in the intergenic region of same-strand pairs with probabilities greater than 0.9 were retained.

The radial SVM classifier X was used to identify the same-strand gene pairs that have a high likelihood of co-transcription. Based on the features, the classifier predicts a score for every gene pair. The score distribution of these gene pairs is shown in Table 3. A total of 2012 of the 4816 same-strand pairs with unknown operon status have a positive score suggesting a high probability of co-transcription. Among these, 1369, 301 and 342 gene pairs fall into groups 1 ($0 < s < 1$), 2 ($1 \leqslant s < 1.2$) and 3 ($s \geqslant 1.2$), respectively. Both transcript correlation and intergenic distance play an important role in predicting a positive score for these gene pairs. At higher threshold in group 3, almost all the gene pairs have a transcript correlation $r > 0.7$ in at least 1 set, and 173 (51%) have $r > 0.7$ in four or more sets. As expected, the transcript correlations are somewhat lower among the gene pairs in group 1 and group 2 with lower scores (Table 3). Moreover, the percentage of gene pairs with short intergenic distance ($d < 25$ bp) is higher in group 3 compared to group 1. Operon predictions for all the same-strand pairs in the genome are available in Supplementary Table S4.

Table 3. Distribution of scores of same-strand gene pairs with unknown operon status

Score	Number of gene pairs	Number of pairs with $r > 0.7$[a] in at least 1 set	Number of pairs with short intergenic distance ($d < 25$ bp)
$s < -1$	1452	161 (11%)	3 (<1%)
$-1 \leqslant s < 0$	1352	597 (44%)	123 (9%)
$0 \leqslant s < 1$	1369	658 (48%)	1074 (78%)
$1 \leqslant s < 1.2$	301	230 (76%)	264 (88%)
$s \geqslant 1.2$	342	329 (96%)	307 (90%)
Total	4816	1975	1771

[a]r is correlation between transcript profiles of the adjacent genes in a same-strand pair

Among the 4816 same-strand pairs, 1452 pairs have score less than -1. The transcript correlations among these 1452 gene pairs are significantly lower than the gene pairs with positive score, as shown in Table 3. Further, < 1% of these 1452 pairs have short intergenic distance ($d < 25$ bp). Thus, the likelihood that adjacent genes in these pairs are co-transcribed is low; in other words, a cistron boundary is likely to exist in the intergenic region of those gene pairs.

Functional analysis. Genes involved in the same biochemical pathway/function tend to cluster together in prokaryotic genomes (47,48), and are regulated similarly at transcription level. We therefore performed functional analysis to test if the genes in same-strand pairs with high score are functionally related. The protein classification scheme originally described by Monica Riley (49), and subsequently adapted for *S. coelicolor* was used (http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml). Among the 7825 genes in the genome, 2371 (30.3%) encode hypothetical proteins without any known function and an additional 565 (7.2%) genes have putative assignments and do not belong to any functional class. Similarly, 3264 (41.7%) genes are not categorized in any of the Gene Ontology classes. Thus, 4889 (62.5%) genes were assigned to 175 functional classes according to the scheme of Monica Riley (49).

Among the 149 KOPs in the training set, the adjacent genes in 121 pairs are functionally annotated. Ninety-two (76%) of these 121 pairs of adjacent genes belong to the same functional class. In contrast, out of the 72 NOPs in which both genes are annotated, only eight (11%) share the same functional class. We examined the functional relatedness of genes in same-strand pairs grouped according to their scores. At higher score threshold in group 3, the genes in 67% of the pairs belong to the same functional class. However, the functional similarity between adjacent genes decreases at lower score thresholds in group 1 and 2, as shown in Table 4. This trend of decreasing functional similarity is more vivid when we examine gene pairs with negative score. Only 106 (18%) pairs of adjacent genes with score less than -1 share the same functional class. This sharp difference in functional similarity of gene pairs with positive and negative score

Table 4. Functional analysis of same-strand gene pairs

Score	Number of gene pairs	Number of annotated gene pairs	Number of pairs in same functional class
$s<-1$	1452	605	106 (18%)
$-1 \leqslant s<0$	1352	521	117 (22%)
$0 \leqslant s<1$	1369	667	317 (48%)
$1 \leqslant s<1.2$	301	169	100 (59%)
$s \geqslant 1.2$	342	206	137 (67%)

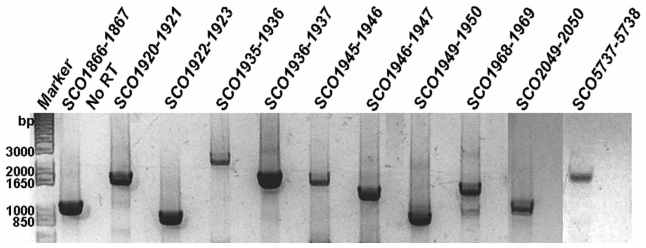

Figure 5. Experimental verification of co-transcription of adjacent genes by RT-PCR. RNA isolation and RT-PCR was performed as described in the Materials and Methods section. Primers were used to amplify across adjacent genes and the products were analyzed by gel electrophoresis. The expected size of the amplicons in bp is: *SCO1866-1867*—1035; *SCO1920-1921*—1637; *SCO1922-1923*—885; *SCO1935-1936*—2298; *SCO1936-1937*—1633; *SCO1945-1946*—1659; *SCO1946-1947*—1353; *SCO1949-1950*—964; *SCO1968-1969*—1426; *SCO2049-2050*—980; *SCO5737-5738*—1615. For every gene pair, a negative control (No RT) in which the RT enzyme was not added was also performed. The negative control is shown next to each RT reaction.

is consistent with our operon predictions. Among the functional classes shared by pairs of adjacent genes with positive score, the most abundant classes are summarized in Supplementary Table S3. The class of transport/binding proteins is the most abundant followed by the genes involved in secondary metabolism and its subclass polyketide synthases (PKS).

Experimental verification

To confirm the co-transcription of predicted gene pairs, RT-PCR was performed on some pairs using primers that amplify across their intergenic region. To allow for a large number of primer sets to be used readily, we employed the primers previously used for construction of whole-genome *S. coelicolor* microarray. Based on the success rate of amplification in preliminary RT-PCR experiments, the verification was limited to only those gene pairs whose amplicon size was no larger than 2.5 kb. This list of gene pairs was further constrained by considering only those whose transcript profiles have a correlation $r > 0.7$ in at least one of the nine sets. With those criteria 114, 91 and 163 gene pairs in group 1 ($0 \leqslant s<1$), 2 ($1 \leqslant s<1.2$) and 3 ($s \geqslant 1.2$), respectively were selected for verification. A number of examples of gene pairs verified by RT-PCR are shown in Figure 5.

Overall 250 (68%) of the 368 gene pairs tested were verified to be on the same operon. The distribution of gene pairs according to different range of scores is listed in Table 5. At lower scores in group 1, a transcript was detected in 59% of the gene pairs tested. The percentage of gene pairs verified to be on the same operon increases to 67 and 75 at higher scores in groups 2 and 3, respectively. The range of intergenic distance of the gene pairs that were verified by RT-PCR is also shown in Table 5. A total of 106 (87%) of the 122 verified pairs with $s \geqslant 1.2$ have short intergenic distance ($d<25$ bp). At lower threshold of $s < 1.2$, 13 of the 128 verified pairs have intergenic distance exceeding 100 bp.

Table 5. RT-PCR based verification of co-transcription of gene pairs

Score	Number of gene pairs tested	Number of gene pairs verified	Range of intergenic distance (bp)
$s \geqslant 1.2$	163	122 (75%)	−32 to 131
$1 \leqslant s<1.2$	91	61 (67%)	−8 to 178
$0 \leqslant s<1$	114	67 (59%)	−13 to 178

It is possible that some of the tested gene pairs that were not positively verified to be on the same operon are false positive predictions. However, a conclusion on those gene pairs cannot be drawn easily. The sample RNA used for RT-PCR was pooled from wild-type M145 cells at different culture stages. In contrast, the transcript profiles were obtained also from different mutants across a range of culture conditions. It is also possible that some of the tested operons were not transcribed in any of the cell samples collected. A complete list of all the gene pairs tested is available in Supplementary Table S5.

Verification of false negative predictions. From the results of binary classification, 6% of the KOPs have a negative score suggesting that the SVM model did not accurately classify them to be on the same operon. Also the 92% recall of SVM classifier X (Table 1) indicates that the model will misclassify 8% of the KOPs. Therefore, it is likely that some of the 2804 same-strand gene pairs with $s<0$ are in fact co-transcribed. To identify some of these gene pairs, we compared the scores of these 2804 gene pairs with the operon predictions of Price *et al.* (4). The authors used a Bayesian approach with features derived from the genome sequence to predict the probability (pOp) that two adjacent genes in *S. coelicolor* are in the same operon. We performed RT-PCR on a restricted subset of 60 gene pairs which have $s<0$ and pOp > 0.6. A PCR product was observed in 16 of these gene pairs. A closer examination of their transcript profiles revealed that 10 of these 16 pairs have correlation $r<0.7$ in all the nine sets of transcriptome data and only two pairs have $r > 0.7$ in more than one set. The weak correlation between transcript levels in these pairs is a potential cause for their misclassification by our model.

Extension of boundaries of known operons. In our analysis, we sought to identify groups of adjacent genes that are expressed as a single transcription unit. When we grouped consecutive gene pairs in the training set with score greater than zero, we observed in four cases that the

Table 6. Extension of cistron boundary of known operons

No.	Known operon	Known size	Predicted operon	Gene pairs verified by RT-PCR	Reference
1	*SCO5736-5737* (*rspO-pnp*) (Protein synthesis)	2	*SCO5737-5740*	*SCO5737-5738, SCO5739-5740*	(39,40)
2	*SCO2050-2054* (*hisAHBCD*) (Histidine biosynthesis)	5	*SCO2048-2054*	*SCO2049-2050*	(67,68)
3	*SCO5583-5585* (*amtB-glnK-glnD*) (Nitrogen metabolism)	3	*SCO5583-5586*	*SCO5585-5586*	(69)

identified polycistron size was greater than the size of the known operon. This indicates that additional gene pairs at cistron boundaries with positive score were predicted to be co-transcribed. One of these is the *rspO-pnp* (*SCO5736–5737*) dicistron (39,40). The transcript profile of *SCO5738*, encoding a putative protease downstream of this operon is strongly correlated with *SCO5737*. Moreover, the gene pair, *SCO5737–5738*, has a high score of 1.1. We verified the presence of a transcript across their intergenic region indicating that the operon has more than two genes (Figure 5). Further, the two downstream genes *SCO5739* and *SCO5740* encoding putative dihydrodipicolinate reductase and putative membrane protein, respectively are strongly correlated with *SCO5738* and both the genes are predicted to be part of the same operon. We have also verified the co-transcription of the genes *SCO5739* and *SCO5740* by RT-PCR.

Table 6 lists examples of two other operons for which pairs of adjacent genes at cistron boundaries with positive score were verified to be on the same operon. It is important to note that the reports characterizing these operons did not exclude the possibility of additional adjacent genes being part of the same transcription unit. By definition, the gene pairs at these cistron boundaries were used as NOPs in the training set and they were consistently classified as false positives due to their positive scores. We have shown that the genes in these pairs are indeed co-transcribed in agreement with our predictions. Hence the leave-one-out false positive rate (FPR) from our predictions is likely to be lower than 14%.

DISCUSSION

Operon is the unit of transcriptional regulation in prokaryotes. Identifying operon structure in a genome is important to the study of gene expression regulation. The estimation of transcript level of a gene using microarrays can also be improved by using information about the transcriptional activity of the genes that are co-transcribed with it. This can also translate into an improvement in identification of differentially expressed genes (50). In our study of *S. coelicolor* A3(2) and disruption mutants of regulatory genes, we have compiled a series of time profiles of transcriptome data. Such dynamic transcriptome profiles can be valuable in elucidating operon structure. Combining with transcriptome data on several *S. coelicolor* strains and culture conditions from two other studies (18,21), the dynamic behavior of adjacent genes in the genome were used to assess the likelihood of their being in the same operon. In principle, the expression profiles of genes on the same operon should be well correlated, at least under conditions of no alternative regulation. However, in reality transcriptome data are often riddled with noise especially when the transcription level is low, rendering microarray assay insensitive to dynamic changes. In our analysis, we thus incorporated other features characteristic of genes in the same operon.

Dependence of transcript dynamics for operon prediction

An essential condition for accurate calculation of correlation between genes is that they are expressed above noise level and exhibit sufficient dynamics across different experiments (5,51). Among the KOPs in the training set, gene pairs which have a higher score tend to exhibit more dynamics in their temporal transcript profiles. Consistent with this notion the Shannon entropy, a measure of transcript variation, of same-strand pairs with high score ($s \geqslant 1.2$) was found to be greater than that of same-strand pairs with score in the range $0 \leqslant s < 1$ (P-value $= 1.3 \times 10^{-73}$, Kolmogorov–Smirnov test).

The operon predictions improve significantly when transcriptome data from diverse experimental conditions are incorporated in the model. Using transcript profiles from only M145 strain in modified R5 medium, merely 60% of the KOPs could be accurately classified with a high FPR of 35%. This study thus included temporal transcriptome data from several *S. coelicolor* strains under very different culture conditions and from different sources including ours. Using all those transcriptome data (Supplementary Table S1), 80% of the KOPs could be classified at a considerably lower FPR of 21%. Moreover, the performance of the transcriptome-based classifier was comparable to the intergenic distance-based classifier (Figure 4). As more transcriptome data become available in the future, especially when conditions under which data are acquired increase, the classification framework can be used to further expand the repertoire of operons identified.

Other features

Intergenic distance feature. The intergenic distance between two adjacent genes in an operon is shorter on an average compared to that of same-strand pairs, which are not in the same operon. This feature was first used for operon prediction in *E. coli* (2,52). Several studies have subsequently showed that intergenic distance can be effectively used for operon prediction in other prokaryotes (4,8), for which the genome sequence is available. Using a log-likelihood function based on intergenic distance distribution, 75% of transcription units in *E. coli* could be predicted (2). It has been reported that operon predictions using intergenic distance has the highest

Table 7. Size of the predicted transcription units

Cistron size	Number of cistrons	Number of cistrons with $s > 1$ in all gene pairs	Number of cistrons with $s > 1.2$ in all gene pairs
1	4386	–	–
2	839	203	85
3	235	33	13
4	111	17	6
5	46	5	3
>5	47	2	0

recall and lowest FPR among the various features derived from genome sequence including codon usage, promoter predictions and terminator predictions (14). In this study using intergenic distance a FPR of 20% was seen at a recall of 80% (Figure 4). However, the FPR increased sharply to 35% as recall increased to 85%, indicating that intergenic distance cannot alone be used to achieve high recall at acceptable error rates.

Transcription terminator feature. Transcripts of operons, particularly those without any internal regulation, are likely to terminate at a single transcription terminator. Therefore, the likelihood of a terminator in the intergenic region of intra-operonic genes is low. Several studies have used this feature for operon prediction in prokaryotes (7,10,14,31). Due to the highly degenerate nature of the binding site of Rho-factor (called 'rut' site) (53), identification of rho-dependent terminator site is difficult. Most terminator prediction algorithms identify rho-independent transcription terminators, which have a characteristic hairpin structure (22,54–56). Among the same-strand pairs in the *S. coelicolor* genome, < 5% have high confidence (probability > 0.9) terminator predictions in their intergenic region. Hence this feature cannot be used to infer the operon status of a large fraction of the same-strand pairs.

Prediction of transcription units

In this study, every same-strand gene pair was assigned a score, and a score threshold of zero was used to group consecutive gene pairs into operons. A total of 5664 transcription units were predicted of which 1278 are polycistronic with two or more genes. The predicted operon assignment of every gene is given in Supplementary Table S6. The distribution of cistrons of different sizes is summarized in Table 7.

Large operons. Among the polycistrons, 47 (3.7%) have more than five genes of which 11 cistrons have 10 or more genes. This includes a large 27.5 kb, 21 gene operon *SCO0381–SCO0401* comprising a secondary metabolite gene cluster. The genes in this operon encode deoxysugar synthases involved in the synthesis of an unknown secondary metabolite. Eight of the 20 gene pairs in this operon have $s \geqslant 1.2$ implying strong predictions of co-transcription of these genes. Another 16.5 kb long, 15 gene operon *SCO3235–SCO3249* encodes for genes involved in the synthesis of calcium-dependent antibiotic (CDA)—a peptide antibiotic synthesized by non-ribosomal peptide synthases. Due to strong correlation between their transcript profiles, 6 of the 14 gene pairs in this operon have scores greater than 1.2.

Operons with internal regulation. Organization of genes into operons allows for an efficient way of coordinated response by the organism to environmental changes. However, there might also be circumstances in which an organism may need the product of the genes on an operon differently, either stoichiometrically or temporally, than the way they are normally prescribed. Thus, some flexibility to allow for an escape from the coordinated expression in an operon is necessary in some cases. A promoter or transcription terminator within an operon allows for differential expression of genes in the same operon. Among the known operons in the training set, many have an internal regulation site in the intergenic region of intra-operonic gene pairs. Several of these gene pairs have wide intergenic spacing. Interestingly, wide spacing between adjacent genes in operons has been reported to indicate complex regulation (1). Based on transcriptome data and prediction of transcription factor binding sites and terminator sites, nearly 20% of operonic genes in *S. coelicolor* are thought likely to be internally regulated (45). The existence of internal regulation in a same-strand gene pair may reduce the degree of correlation of their transcript profiles. A better prediction of internal regulation will certainly improve the operon predictions.

Further evidence of robustness—comparison of recently identified operons

Since we conducted the operon predictions there have been reports of polycistrons that were not included in our training set. We compared our prediction with the reports on those operons. This includes a five gene operon *nikABCDE* involved in nickel transport (57), a dicistron *devA-devB* (58), and class Ia and class II ribonucleotide reductases (RNR) encoded by *nrdABS* and *nrdRJ* operons (59). Seven of the eight gene pairs in these operons have a positive score, of which six pairs have transcript correction $r > 0.7$ in at least two of the nine sets. Only one gene pair—*nrdB-nrdS* has a negative score. Therefore, the genes in these operons were nearly all correctly predicted to be co-transcribed according to our methodology including one pair *nrdR-nrdJ* that is separated by a large distance (168 bp).

Using operon predictions for functional annotations

Chromosomal proximity of adjacent genes in multiple prokaryotic genomes and their co-transcription is often an indication of their functional relatedness. The information can be used to infer their functional annotation (47,48,60). In this study, we observed an increasing trend of functional similarity between pairs of adjacent genes with increasing scores. A significant number of pairs are comprised of genes that do not share the same functional class, many of which have a high score ($s \geqslant 1.2$) and/or have been verified by RT-PCR as being on the same

operon. Of these, 113 pairs have only one gene functionally annotated, while the other gene is either hypothetical or unclassified. The transcript profiles in almost all these gene pairs are strongly correlated ($r > 0.7$) in at least one set of microarray data. Potential relatedness of the physiological function may be inferred from the well annotated neighboring gene. Details of these genes are provided in Supplementary Table S7.

Comparison of operon predictions with earlier report

In this study we employed SVM as a classification tool for operon prediction. SVMs have been recently used to study several classification problems in bioinformatics (61–65). These studies have demonstrated that SVM-based classifiers produce results that are better than or at least as good as those obtained by other supervised methods, as the classification models that they generate tend to better generalize on unseen instances (i.e. instances that were not used during training).

Other methods have also been used successfully for operon prediction. A Bayesian approach was used to predict operons in all sequenced prokaryotes including *S. coelicolor* (4). The method relies on features based on genome sequence alone and uses intergenic distance, conservation of gene clusters across different organisms, codon usage and functional similarity to predict a probability (pOp) that two adjacent genes with the same orientation are co-transcribed. The authors used *S. coelicolor* genome annotation from The Institute of Genomic Research (TIGR), whereas the primary annotation from Sanger Institute (http://streptomyces.org.uk/) was used for this study. Out of 4965 same-strand pairs, 4549 pairs can be compared with their predictions.

A 5-fold cross-validation of the training set was used to compare the two methods. A reduced training set of 139 KOPs and 58 NOPs for which operon predictions were available from both studies was used for cross-validation. A comparison of the vertically averaged ROC curves is shown in Supplementary Figure S1. The difference between the predictions from the two methods is most evident at recall $>70\%$, where the FPR from our method is significantly less than the FPR from the predictions of Price and co-workers. Consequently, the AUC for our SVM classifier is greater than the latter study (P-value $= 5.2 \times 10^{-4}$, Wilcoxon signed rank test). However, it is important to note that the size of training set used in this comparison is smaller than the total number of KOPs and NOPs used for comparison of different SVM classifiers described earlier. For small sample sizes, the estimate of error obtained from cross-validation (leave-one-out as well as k-fold) can be highly variable. Classifiers based on small sample sizes are particularly vulnerable to situations where the perturbation introduced by k-fold partitioning results in an unstable classifier—a classifier with unreliable accuracy estimates (29,66).

We also compared the predictions of the two methods on all the same-strand gene pairs in the genome. Using a threshold of 0 and 0.5 for score and pOp, respectively, 3730 of the 4410 same-strand pairs with unknown operon status have the same predictions. The gene pairs for which the predictions of the two methods do not match include 125 pairs, which have $s > 0.5$ and pOp < 0.4. Adjacent genes in these pairs are predicted to be co-transcribed by our SVM model but not the Bayesian approach. More than 70% of these gene pairs have a transcript correlation $r > 0.7$ in at least two of the nine sets of microarray data suggesting that our method emphasizes expression correlation to predict the likelihood of co-transcription. On the other hand, 151 gene pairs have $s < -0.5$ and pOp > 0.6 indicating that these pairs were predicted to be co-transcribed by the Bayesian model but not by our SVM model. Among these pairs, 68% do not have a transcript correlation $r > 0.7$ in any of the nine sets and only 7% have correlation $r > 0.7$ in two or more sets. These results suggest that the information based on similarity of transcript profiles of adjacent genes plays an important role in our model predictions and thereby enhances the performance of operon prediction models that rely on genome sequence-based features alone.

CONCLUSIONS

Streptomyces coelicolor has a rich genome that encodes more genes than even the eukaryote, *Saccharomyces cerevisiae*. Its versatility to undergo differentiation with a complex life cycle and produce secondary metabolites after the cessation of growth is reflected in its highly dynamic temporal transcriptome profile. A large number of genes on the genome are not well annotated, with many annotated as involved in, or hypothesized to be involved in, regulation. A better understanding of its operon structure will be valuable in gene annotation and large-scale gene expression studies for elucidating its regulatory networks that control differentiation and antibiotics production. In this study, we used a SVM-based supervised classification approach to predict operon structure for this organism. In the past few years the transcriptome data of this organism has become a valuable resource for gaining physiological insights. The use of time series transcriptome data enhanced the predictive capability of the classifier that employed genome sequence-based features including intergenic distance and transcription terminator predictions. The experimental verification of a large set of those predicted by the classifier further demonstrates the utility of the method. As more transcriptome data become available and the conditions under which they are obtained diversify, the framework established in this study will also become more versatile in further enhancing the operon predictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported in part by grants from National Institutes of Health GM55850 to WSH and

NSF EIA-9986042, ACI-0133464, IIS-0431135, NIH RLM008713A to GK. Computational support was provided by University of Minnesota Supercomputing Institute. We thank Marlene Castro for her assistance in compiling the list of known operons and Lewis Marshall for his help in verification of operons. Funding to pay the Open Access publication charges for this article was provided by WSH at University of Minnesota.

Conflict of interest statement. None declared.

REFERENCES

1. Price,M.N., Arkin,A.P. and Alm,E.J. (2006) The life-cycle of operons. *PLoS Genet*., **2**, e96.
2. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
3. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res*., **29**, 1216–1221.
4. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*., **33**, 880–892.
5. Sabatti,C., Rohlin,L., Oh,M.K. and Liao,J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res*., **30**, 2886–2893.
6. Westover,B.P., Buhler,J.D., Sonnenburg,J.L. and Gordon,J.I. (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880–888.
7. Wang,L., Trawick,J.D., Yamamoto,R. and Zamudio,C. (2004) Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Res*., **32**, 3689–3702.
8. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18** (Suppl. 1), S329–S336.
9. Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) Using multiple levels of learning and diverse evidence sources to uncover coordinately controlled genes. *Proceedings of 17th International Conference on Machine Learning*. Morgan Kaufmann, Stanford, CA, USA.
10. Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 116–127.
11. Zhang,G.Q., Cao,Z.W., Luo,Q.M., Cai,Y.D. and Li,Y.X. (2006) Operon prediction based on SVM. *Comput. Biol. Chem*., **30**, 233–240.
12. Tjaden,B., Haynor,D.R., Stolyar,S., Rosenow,C. and Kolker,E. (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, **18**, S337–S344.
13. De Hoon,M.J.L., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2004) Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput*., **9**, 276–287.
14. Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227–1235.
15. Bentley,S.D., Chater,K.F., Cerdeno-Tarraga,A.M., Challis,G.L., Thomson,N.R., James,K.D., Harris,D.E., Quail,M.A., Kieser,H. *et al*. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.
16. Baltz,R.H. (1998) Genetic manipulation of antibiotic-producing *Streptomyces*. *Trends Microbiol*., **6**, 76–83.
17. Huang,J., Lih,C.J., Pan,K.H. and Cohen,S.N. (2001) Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays. *Genes Dev*., **15**, 3183–3192.
18. Huang,J., Shi,J., Molle,V., Sohlberg,B., Weaver,D., Bibb,M.J., Karoonuthaisiri,N., Lih,C.J., Kao,C.M. *et al*. (2005) Cross-regulation among disparate antibiotic biosynthetic pathways of *Streptomyces coelicolor*. *Mol. Microbiol*., **58**, 1276–1287.
19. Mehra,S., Lian,W., Jayapal,K., Charaniya,S., Sherman,D. and Hu,WS. (2006) A framework to analyze multiple time series data – a case study with *Streptomyces coelicolor*. *J. Ind. Microbiol. Biotechnol*., **33**, 159–172.
20. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
21. Karoonuthaisiri,N., Weaver,D., Huang,J., Cohen,S.N. and Kao,C.M. (2005) Regional organization of gene expression in *Streptomyces coelicolor*. *Gene*, **353**, 53–66.
22. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol*., **301**, 27–33.
23. Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory* Springer, New York.
24. Vapnik,V.N. (1998) *Statistical Learning Theory* Wiley, New York.
25. Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkofp,B., Burges,C.J.C. and Smola,A.J. (eds), *Advances in Kernel Methods – Support Vector Learning*, M.I.T. Press, pp. 169–184.
26. Lee,J.H. and Lin,C.J. (2000). Automatic model selection for support vector machines. *Technical report*. Department of Computer Science and Information Engineering, National Taiwan University, Taipei.
27. Takeuchi,T., Sawada,H., Tanaka,F. and Matsuda,I. (1996) Phylogenetic analysis of *Streptomyces* spp. causing potato scab based on 16S rRNA sequences. *Int. J. Syst. Bacteriol*., **46**, 476–479.
28. Leblond,P., Redenbach,M. and Cullum,J. (1993) Physical map of the Streptomyces lividans 66 genome and comparison with that of the related strain *Streptomyces coelicolor* A3(2). *J. Bacteriol*., **175**, 3422–3429.
29. Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of Fourteenth International Conference on Artificial Intelligence*, Montreal, CA, pp. 1137–1143.
30. Fawcett,T. (2004). ROC graphs: Notes and practical considerations for researchers. *Technical report*. HP Laboratories, Palo Alto, pp. 38.
31. Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F. and Craven,M. (2003) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19** (Suppl. 1), i34–i43.
32. Takano,H., Obitsu,S., Beppu,T. and Ueda,K. (2005) Light-induced carotenogenesis in *Streptomyces coelicolor* A3(2): identification of an extracytoplasmic function sigma factor that directs photodependent transcription of the carotenoid biosynthesis gene cluster. *J. Bacteriol*., **187**, 1825–1832.
33. Lee,E.J., Cho,Y.H., Kim,H.S. and Roe,J.H. (2004) Identification of sigmaB-dependent promoters using consensus-directed search of *Streptomyces coelicolor* genome. *J. Microbiol*., **42**, 147–151.
34. Cho,Y.H., Lee,E.J., Ahn,B.E. and Roe,J.H. (2001) SigB, an RNA polymerase sigma factor required for osmoprotection and proper differentiation of *Streptomyces coelicolor*. *Mol. Microbiol*., **42**, 205–214.
35. Hu,D.S., Hood,D.W., Heidstra,R. and Hodgson,D.A. (1999) The expression of the trpD, trpC and trpBA genes of *Streptomyces coelicolor* A3(2) is regulated by growth rate and growth phase but not by feedback repression. *Mol. Microbiol*., **32**, 869–880.
36. Sevcikova,B. and Kormanec,J. (2002) Activity of the *Streptomyces coelicolor* stress-response sigma factor sigmaH is regulated by an anti-sigma factor. *FEMS Microbiol. Lett*., **209**, 229–235.
37. Sevcikova,B., Benada,O., Kofronova,O. and Kormanec,J. (2001) Stress-response sigma factor sigma(H) is essential for morphological differentiation of *Streptomyces coelicolor* A3(2). *Arch. Microbiol*., **177**, 98–106.
38. Kormanec,J., Sevcikova,B., Halgasova,N., Knirschova,R. and Rezuchova,B. (2000) Identification and transcriptional characterization of the gene encoding the stress-response sigma factor sigma(H) in *Streptomyces coelicolor* A3(2). *FEMS Microbiol. Lett*., **189**, 31–38.

39. Bralley,P. and Jones,G.H. (2004) Organization and expression of the polynucleotide phosphorylase gene (pnp) of Streptomyces: processing of pnp transcripts in *Streptomyces antibioticus*. *J. Bacteriol.*, **186**, 3160–3172.
40. Chang,S.A., Bralley,P. and Jones,G.H. (2005) The absB gene encodes a double strand-specific endoribonuclease that cleaves the read-through transcript of the rpsO-pnp operon in *Streptomyces coelicolor*. *J. Biol. Chem.*, **280**, 33213–33219.
41. Bucca,G., Brassington,A.M., Hotchkiss,G., Mersinias,V. and Smith,C.P. (2003) Negative feedback regulation of dnaK, clpB and lon expression by the DnaK chaperone machine in *Streptomyces coelicolor*, identified by transcriptome and in vivo DnaK-depletion analysis. *Mol. Microbiol.*, **50**, 153–166.
42. Fornwald,J.A., Schmidt,F.J., Adams,C.W., Rosenberg,M. and Brawner,M.E. (1987) Two promoters, one inducible and one constitutive, control transcription of the *Streptomyces lividans* galactose operon. *Proc. Natl Acad. Sci. USA*, **84**, 2130–2134.
43. Vierling,S., Weber,T., Wohlleben,W. and Muth,G. (2000) Transcriptional and mutational analyses of the *Streptomyces lividans* recX gene and its interference with RecA activity. *J. Bacteriol.*, **182**, 4005–4011.
44. Tieleman,L.N., van Wezel,G.P., Bibb,M.J. and Kraal,B. (1997) Growth phase-dependent transcription of the *Streptomyces ramocissimus* tuf1 gene occurs from two promoters. *J. Bacteriol.*, **179**, 3619–3624.
45. Laing,E., Mersinias,V., Smith,C.P. and Hubbard,S.J. (2006) Analysis of gene expression in operons of *Streptomyces coelicolor*. *Genome Biol.*, **7**, R46.
46. Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.
47. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
48. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.*, **1**, 93–108.
49. Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.
50. Xiao,G., Martinez-Vaz,B., Pan,W. and Khodursky,A.B. (2006) Operon information improves gene expression estimation for cDNA microarrays. *BMC Genomics*, **7**, 87.
51. Kuramochi,M. and Karypis,G. (2001) Gene classification using expression profiles: a feasibility study. *Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering Conference*,Bethesda,MD, pp. 191–200.
52. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Blattner,F.R. and Collado-Vides,J. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 65–67.
53. Banerjee,S., Chalissery,J., Bandey,I. and Sen,R. (2006) Rho-dependent transcription termination: more questions than answers. *J. Microbiol.*, **44**, 11–22.
54. Brendel,V. and Trifonov,E.N. (1984) Computer-aided mapping of DNA-protein interaction sites. *CODATA Bulletin,* , **56**, 17–20.
55. Brendel,V. and Trifonov,E.N. (1984) A computer algorithm for testing potential prokaryotic terminators. *Nucleic Acids Res.*, **12**, 4411–4427.
56. Unniraman,S., Prakash,R. and Nagaraja,V. (2002) Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res.*, **30**, 675–684.
57. Ahn,B.E., Cha,J., Lee,E.J., Han,A.R., Thompson,C.J. and Roe,J.H. (2006) Nur, a nickel-responsive regulator of the Fur family, regulates superoxide dismutases and nickel transport in *Streptomyces coelicolor*. *Mol. Microbiol.*, **59**, 1848–1858.
58. Hoskisson,P.A., Rigali,S., Fowler,K., Findlay,K.C. and Buttner,M.J. (2006) DevA, a GntR-like transcriptional regulator required for development in *Streptomyces coelicolor*. *J. Bacteriol.*, **188**, 5014–5023.
59. Borovok,I., Gorovitz,B., Schreiber,R., Aharonowitz,Y. and Cohen,G. (2006) Coenzyme B12 controls transcription of the *Streptomyces* class Ia ribonucleotide reductase nrdABS operon via a riboswitch mechanism. *J. Bacteriol.*, **188**, 2512–2520.
60. Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res.*, **12**, 1221–1230.
61. Noble,W.S. (2004) Support vector machine applications in computational biology. In Scholkofp,B., Tsuda,K. and Vert,J. (eds), *Kernel methods in computational biology*, M.I.T. Press, pp. 71–92.
62. Zien,A., Rätsch,G., Mika,S., Schölkopf,B., Lengauer,T. and Müller,K.R. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.
63. Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.
64. Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
65. Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.Jr. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
66. Braga-Neto,U.M. and Dougherty,E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.
67. Limauro,D., Avitabile,A., Cappellano,C., Puglia,A.M. and Bruni,C.B. (1990) Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2). *Gene*, **90**, 31–41.
68. Carere,A., Russi,S., Bignami,M. and Sermonti,G. (1973) An operon for histidine biosynthesis in *Streptomyces coelicolor*. I. Genetic evidence. *Mol. Gen. Genet.*, **123**, 219–224.
69. Fink,D., Weissschuh,N., Reuther,J., Wohlleben,W. and Engels,A. (2002) Two transcriptional regulators GlnR and GlnRII are involved in regulation of nitrogen metabolism in *Streptomyces coelicolor* A3(2). *Mol. Microbiol.*, **46**, 331–347.