288–298 *Nucleic Acids Research, 2007, Vol. 35, No. 1* *Published online 14 December 2006*
doi:10.1093/nar/gkl1018

Operon prediction using both genome-specific and general genomic information

Phuongan Dam[1], Victor Olman[1,2], Kyle Harris[1], Zhengchang Su[3] and Ying Xu[1,2,*]

[1]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, [2]Institute of Bioinformatics, University of Georgia, Athens, GA, USA and [3]Center for Bioinformatics Research, Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

Received September 7, 2006; Revised October 27, 2006; Accepted October 30, 2006

ABSTRACT

We have carried out a systematic analysis of the contribution of a set of selected features that include three new features to the accuracy of operon prediction. Our analyses have led to a number of new insights about operon prediction, including that (i) different features have different levels of discerning power when used on adjacent gene pairs with different ranges of intergenic distance, (ii) certain features are universally useful for operon prediction while others are more genome-specific and (iii) the prediction reliability of operons is dependent on intergenic distances. Based on these new insights, our newly developed operon-prediction program achieves more accurate operon prediction than the previous ones, and it uses features that are most readily available from genomic sequences. Our prediction results indicate that our (non-linear) decision tree-based classifier can predict operons in a prokaryotic genome very accurately when a substantial number of operons in the genome are already known. For example, the prediction accuracy of our program can reach 90.2 and 93.7% on *Bacillus subtilis* and *Escherichia coli* genomes, respectively. When no such information is available, our (linear) logistic function-based classifier can reach the prediction accuracy at 84.6 and 83.3% for *E.coli and B.subtilis*, respectively.

INTRODUCTION

In bacterium, an mRNA molecule (or a transcript) can contain one or multiple genes. In the case of multi-gene transcript, the set of genes found in the transcript is arranged in tandem in the chromosome, and named an operon. Although genes in an operon are in general found to be transcribed together, in some cases the same set of genes under different conditions may give rise to transcripts of different lengths. Functionally, genes (in an operon) transcribed into a single mRNA transcript are found to work in the same pathway or interact with each other, although examples of transcripts containing genes involved in different pathways, such as the *Escherichia coli rpsU-dnaG-rpoD* operon that encodes the 30S ribosomal protein S21, DNA primase and RNA polymerase have been documented (1). Biologically, organization of multiple genes into an operon serves as a transcription-regulation mechanism that subsequently regulates the activity of pathways and/or cellular responses. Therefore, successful prediction of operons can help to improve our ability in functional annotation of (conserved) hypothetical genes, a major challenge in functional annotation of genomes. Currently, the best prediction programs can reach a prediction accuracy level at 85–91% of specificity and sensitivity in terms of finding the correct operonic gene pairs in *Bacillus subtilis* and *E.coli*. In terms of correctly predicting the whole transcription unit (TU) that contains one or more genes, the sensitivity level varies from 50 to 79% in *E.coli* (2) due to the high false positive rates in classifying adjacent gene pairs.

It has been shown that a number of genomic features relevant to adjacent gene pairs (on the same strand) are useful for predicting whether the pairs belong to the same operons. These features include (i) the intergenic distance, (ii) the phylogenetic profiles of genes, (iii) the conservation of gene pairs (gene neighbourhood) across multiple genomes, (iv) the functional annotation, such as COG or Riley's classification, (v) the involvement of a gene pair in the same biological pathway, protein complexes or physical interactions and (vi) the correlation of their gene expression patterns. Using these features, a number of computational methods have been developed for operon prediction, including (i) hidden Markov model-based method (3), (ii) machine learning-based technique (4), (iii) simple statistical methods (5), (iv) Bayesian methods (6–9), (v) graph-theoretic approaches (10–12), (vi) neural networks (13), (vii) logistic regression(14), (viii) support vector machine (15) and (ix) a few other methods (2,16,17).

In most of these methods, the predictors were trained and tested on the known operon data of either *E.coli* or *B.subtilis* because only these two organisms have substantial numbers of experimentally verified operons. As a result, a general

*To whom correspondence should be addressed. Tel: +1 706 542 9779; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu

problem with current methods is that they do not seem to generalize well from one genome to another. For example, an operon-prediction program, trained on *E.coli* data, could have 91% prediction accuracy on (other) *E.coli* operonic gene pairs but have its accuracy dropped to 64% when tested on *B.subtilis* operonic gene pairs (2). Among factors that could have affected their generalization ability is the (possibly unintentional) use of genome-specific features, leading to performance deduction of these methods when applied to a new genome. Though numerous studies have been carried out to combine different genomic features in various ways for operon prediction, very little has been done to examine the contribution of these features, individually and in combination, for operon prediction in genomes other than the genome(s) on which a prediction program is trained. Another factor that could have affected the generalization ability of the methods is the choice of classification functions. For example, the use of non-linear classification methods, such as Bayesian classification schemes or support vector machines could lead to an over-trained predictor and performance deduction when applied to a new genome. To address these issues, we have evaluated the performance of a number of classification methods for operon prediction on *E.coli* and *B.subtilis* when using a training dataset from the same genome versus a different genome. We have also assessed the usefulness of several information sources that were used in previous operon prediction programs, such as conserved gene neighborhood, phylogenetic profiles and intergenic distances, through comparing the classification errors when these features were used and not used. In addition to the previously used features, we have applied in our prediction program three new features, namely the length ratio between a pair of adjacent genes, Gene Ontology (GO)-based functional similarity between adjacent genes and the frequency of a specific DNA motif in the intergenic region, which seem to be generally useful for operon prediction. Together, we found that a flexible framework of linear and non-linear classification methods and a combination of the aforementioned features have substantially improved the accuracy of prediction and the generalization ability of an operon predictor.

MATERIALS AND METHODS

Data preparation

All genome sequences and their annotated genes were downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html). In this study, we used 258 bacterial and archaeal genomes, and the genome list can be found at http://csbl.bmb.uga.edu/~phd/.

From the RegulonDB database (version 4.0) (18), we collected 690 TUs, including 545 unique TUs and 145 overlapped TUs that share common genes. Among the unique TUs, 256 TUs are single-gene TUs and 289 are multi-gene TUs. After removing genes that were deleted from the new version of the *E.coli* files at NCBI, we obtained 707 operonic gene pairs (i.e. adjacent genes within an operon) and 497 boundary pairs. A boundary pair consists of either the upstream adjacent gene and the first gene of the TU, or the last gene of the TU and the adjacent downstream gene, providing that these genes are transcribed in the same direction. From the *B.subtilis* TU database (8), we collected 992 TUs including 338 unique multi-gene TUs and 645 unique single-gene TUs that yielded 850 operonic gene pairs and 775 boundary pairs.

Feature scores

To evaluate the contribution of selected features in operon prediction, we have calculated the numerical values of the features, and then used these values individually and in combination to train a classifier. The features used in our study are (i) the intergenic distance, (ii) the conserved gene neighborhood, (iii) distances between adjacent genes' phylogenetic profiles, (iv) the ratio between the lengths of two adjacent genes and (v) frequencies of specific DNA motifs in the intergenic regions.

Intergenic distance. Because the intergenic distance has been shown to be a critical attribute in predicting operon/boundary pairs (19), we have calculated the intergenic distance (D_I) between each adjacent gene pair using a previously described formula [D_I = downstream_gene_start − (upstream_gene_end + 1)] (19). Furthermore, by observing the distributions of D_I in *E.coli* and *B.subtilis*, we have found that there are very small number of D_I values that is lower than −50 (i.e. two genes whose sequences are overlapped by 50 nt) whereas most of known gene pairs with $D_I > 250$ are found to be boundary pairs. Therefore, we have used −50 and 250 as the lowest and highest cutoff values, respectively, in our study.

Neighborhood conservation. We have used a score for measuring the neighborhood conservation of two genes with respect to a group of reference genomes (in our case, all the involved genomes in our study), defined in an earlier work by our laboratory (20). In brief, the score is defined as: $S = -\sum_{k=1}^{K} L(g_i, g_j, G^k)$, where $L(g_i, g_j, G^k)$ is the log-likelihood of a gene pair to be neighbors in the kth genome G^k. The log-likelihood score is computed as the probability that g_i and g_j are neighbors within a distance $d_{k(i,j)}$ in G^k, or $L(g_i, g_j, G^k) = \log P_{ij}$; P_{ij} is defined as follows:

(i) $P_{ij} = (1 - p_{ik})(1 - p_{jk})$, if both genes are absent from genome G^k,
(ii) $P_{ij} = (1 - p_{ik})\ p_{jk}$, if only gene j is present in genome G^k,
(iii) $P_{ij} = p_{ik}\ (1 - p_{jk})$, if only gene i is present in genome G^k,
(iv) $P_{ij} = (p_{ik}p_{jk}d_{k(i,j)}\ (2N_k - d_{k(i,j)} - 1))/(N_k(N_k - 1))$, if genes i and j are present in genome G^k.

$d_{k(ij)}$ is the number of genes between g_i and g_j; N_k is the number of genes in genome G^k; and p_{ik} is the probability that gene g_i is present in genome G^k. To compute p_{ik}, all reference genomes were divided into 13 groups based on their affiliated phylum, and then p_{ik} was calculated as the frequency of gene g_i present in the phylum that G^k belongs to. Our study has showed that small S values are generally associated with gene pairs that are functionally related (20).

Phylogenetic distance. Phylogenetic profiles, which measure the co-presence or co-absence of a pair of genes in reference to a group of genomes, have been previously used to

predict the functional relatedness of genes (21). Previous studies have shown that genes with highly similar phylogenetic profiles (i.e. with short distance of the phylogenetic profiles) are often functionally related. To calculate the distance between the phylogenetic profiles of two genes, we used two approaches, namely the Hamming distance and the Shannon entropy distance. For the Hamming distance between two genes A and B, we sum the number of times that only A or B is found in the genome, $D_H = \sum_{i=1}^{n} d_i$, where n is the number of genomes, $d_i = 0$ if the orthologs of A and B are both present or both absent in genome i, and $d_i = 1$ otherwise. The Shannon entropy distance is calculated as $D_E = n - (n - D_H)\sqrt{E(p)}$, where p is the proportion of 0 identities among all identities in the phylogenetic profiles of gene A and B, and $E(p) = -p \log(p) - (1-p) \log(1-p)$.

To be thorough, we have calculated this score using, as reference genomes, (i) 258 archaeal and bacterial genomes found at NCBI, (ii) a set of 121 archaeal and bacterial genomes comprised of the largest genome from each genus, (iii) a set of 198 archaeal and bacterial genomes comprised of the largest genome from each species. Based on our preliminary study, we found that the scores obtained from (i) gave the best prediction results. Hence (i) is used in our study described in the Results section.

Inclusion of short DNA motifs for operon prediction. Based on our preliminary study, we found that some DNA motifs seem to be often associated with the inter-operonic regions. To select the DNA motifs with the most discerning power (between operon pairs and boundary pairs), we have counted the number of occurrences for each DNA motif in the intergenic region of each gene pair. For each promoter of length L, the normalized frequency of occurrences of a DNA motif $M = (t_1 \ldots t_d)$, $t_i \in \{A, C, G, T\}$, is calculated as the ratio of observed occurrences and the expected number of occurrences, or

$$F_m = \frac{X}{(L - d + 1) * p},$$

whereas X is the number of observed occurrences of the motif in our target genome, d is the length of the motif and p is the expected frequency of this motif in the genome, assuming independence of nucleotides in the sequence, i.e. $p = \prod_{i=1}^{d} p(t_i)$, where $p(t_i)$ is the frequency of nucleotide t_i in the promoter set. For each motif with length ranging from 3 to 5 nt, we considered all possible combination of A, C, G and T, yielding 64 three-letter motifs, 256 four-letter motifs and 1024 five-letter motifs.

To calculate the motif frequencies, for each gene pair we extracted 100 nt upstream of the translational start site of the downstream gene. The number of times a motif is present in this sequence is counted and then normalized. To select DNA motifs with possibly discerning power between operonic and boundary gene pairs, we have used K-mean method to cluster the normalized frequencies of each motif into two groups. After clustering, each cluster is assigned to one of the two classes, operon or boundary, based on the composition of gene pairs in the cluster, and the classification error rate was calculated. After all motifs of length ranging from 3 to 5 letters are considered, we rank the motifs according to their classification errors. We found that the top motifs are consistent in both *E.coli* and *B.subtilis* datasets. Hence we have included the scores of six top-ranked DNA motifs, TTT, ATA, TTTT, TATA, TTTTT and TTTTC, in the feature list for further evaluation. Later on, we used a formula described in the Supplementary Data to confirm that the motif is significantly over-represented in the genome.

Similarity score between GO terms of gene pairs. Each gene can have a GO number (http://www.geneontology.org/) indicating its biological function. We have previously developed a scoring scheme for measuring the GO-based functional similarity between a pair of genes (20). In brief, the GO similarity score $s(V_i,V_j)$ of a gene pair V_i and V_j is the number of common terms between paths in the two GO graphs V_i, V_j induced by the GO terms of each gene. The GO graph induced by V_i is a direct acyclic graph that includes V_i at the bottom most level of the graph and its ancestor GO terms at the upper levels. Then, the $S_{GO}(g_i,g_j) = \max s(V_i,V_j)$, where maximum is taken over all graphs V_i, V_j induced from GO of V_i and V_j. We have previously shown that the larger the score, the more likely that two genes are functionally related (20).

Length ratio between a pair of genes. The score is calculated as the natural log of the length ratio of upstream gene and downstream gene, or $L = \ln(l_i/l_j)$, $j = i + 1$, whereas l_i and l_j are the length of the genes. The data are shown in the Supplementary Table 4.

Classification of gene pairs

We have utilized an existing Matlab toolbox freely available named PRTools and tried a number of different classification functions in this toolbox. Detailed discussions and implemented algorithms used to create these classifiers can be found at http://130.161.42.18/prtools/prtools.html.

We have tested 11 classification functions, both linear and non-linear, provided in toolbox, on our training data. For each classification function, we have carried out the following procedure to assess the performance in separating operonic gene pairs from boundary gene pairs: (i) we randomly partition the training dataset into two subsets, named training set and testing set; (ii) we train a classifier using a provided classification function on the training set and (iii) we validate the trained classifier using the testing set and record the classification error rate. We run this procedure 100 times for each classification function, and then select the function with the lowest classification error rate as the optimal classifier for this particular classification function. In addition, we tested each trained classifier on another genome to assess the classifier's generalization ability.

Performance measurement

We have used the following measures to assess the performance of our classifiers: (i) sensitivity, (ii) specificity, (iii) the accuracy and (iv) classification error rate, which are calculated as follows:

$$\text{Sensitivity}\,(S_T) = \frac{TP}{WO}$$

$$\text{Specificity}\,(P_T) = \frac{TP}{TP + FP}$$

$$\text{Accuracy(A)} = \frac{\text{TP} + \text{TN}}{\text{WO} + \text{TUB}}$$

$$\text{Error rate} = 1 - \text{A},$$

where TP is the number of true positive (operonic) pairs being predicted correctly, TN is the number of true negative (boundary) pairs being predicted correctly, FP is the number of false positive pairs (i.e. boundary pairs predicted to be operonic pairs), FN is the number of false negative pairs (i.e. operonic pairs predicted to be boundary pairs), WO is the total number of operonic pairs in the dataset, and TUB is the total number of boundary pairs in the dataset. We understand that our accuracy measure is calculated differently from that used by others, which is defined as the average of the sensitivity S_T and the specificity P_T. To calculate the sensitivity and specificity for prediction of boundary pairs, we substituted TN and FN for TP and FP from the above formula.

RESULTS

Operon prediction using different classification functions

Among features that are often used for operon prediction, the length of the intergenic region between a pair of adjacent genes was reported to be one of the most reliable indicators of whether it is an operon pair or a boundary pair (19). To examine the relationship between the intergenic distance and a gene pair's being an operonic or a boundary pair, we have calculated the frequency distributions of intergenic distances at intervals of 20 bases for a set of operonic gene pairs and a set of boundary gene pairs found in the known *E.coli* and *B.subtilis* operon sets, and then compared the frequencies of operonic pairs and the frequencies of boundary pairs at each interval of the intergenic distance. We found that when the intergenic distances between gene pairs are <40 nt in *E.coli* and *B.subtilis*, ~85–92% of the gene pairs are operonic pairs, whereas when the intergenic distances between gene pairs are >200 nt in *E.coli* and *B.subtilis*, ~85–95% of the gene pairs are boundary pairs. When the intergenic distances between gene pairs are >40 but <200 nt, the ratio of the frequency of operonic gene pairs and the frequency of boundary gene pairs reduces as the intergenic distance increases (Supplementary Table 1). These observations led us to closely examine the relationship between the performance of our classifiers and the intergenic distances between adjacent gene pairs. We have employed two approaches to examine this. Our first approach, which is referred as 'whole data-based training', was to use either *E.coli* or *B.subtilis* data as a whole. In the second approach, which is referred as 'subgroup-based training', we divided the known dataset of either *E.coli* or *B.subtilis* into three groups based on the intergenic distance, and then selected the optimum classifier for each group. The three groups are (i) U40 for gene pairs whose intergenic distances are <40 nt, (ii) U200 for gene pairs whose intergenic distances are <200 nt but ⩾40 nt and (iii) O200 for gene pairs whose intergenic distances are ⩾200 nt. The total classification error of the whole dataset is calculated based on the classification errors of these subgroups. The examination results are given later in this section.

To derive a set of useful features for operon-boundary prediction, we have computed scores for all features listed in the Feature scores, including three new features: the frequencies of six DNA motifs, the GO-based functional similarity between adjacent genes, and the ratio between the lengths of two consecutive genes (The data are shown in the Supplementary Table 4). These features in different combinations were used to train a linear classifier, and the features that give low classification error rates for both *B.subtilis* and *E.coli* datasets were identified as the core features for further studies in this section and following sections. This set of core features include the distances between adjacent genes' phylogenetic profiles, the frequencies of the DNA motif (TTTTT) found in the intergenic regions, the intergenic distances and the GO-based functional similarity scores. The contribution by each of these as well as other features to the prediction accuracy is addressed later in Contribution of selected features to operon prediction.

The average classification errors for 100 classification trials using the optimal classifiers (see Classification of gene pairs) are shown in Tables 1 and 2. In each trial, we used the default setting of the classifiers, and calculated the average classification errors for the testing data. The optimal classifier was selected based on the procedure outlined in Classification of gene pairs. Although we have tested classifiers using all classification methods available in the PRTools package, we only report the top four methods that give smallest average classification errors in Tables 1 and 2, and the remaining results can be found in Supplementary Table 2. From these results, we have the following key observations (A) and (B).

(A) When training and testing data are from the same genome, non-linear classifiers yield lower classification

Table 1. The average classification errors of various linear classifiers and non-linear classifiers (*) when using whole data-based training approach (All), or subgroup-based training approach (Subgroup): the training and the testing sets are from the same genome

Classifier	*E.coli*		*B.subtilis*	
	All	Subgroup	All	Subgroup
Loglc	15.80	14.29	17.16	16.83
Fisherc	16.29	14.62	17.16	16.87
Naivebc*	13.98	12.88	17.17	16.03
Treec*	20.72	9.91	31.84	15.16

Table 2. The average classification errors of various linear classifiers and non-linear classifiers (*) when using whole data-based training approach (All), or subgroup-based training approach (Subgroup): the training set is from *B.subtilis* and the testing set is from *E.coli* (column 2–3) and the other training set is from *E.coli* and testing set is from *B.subtilis* (column 4–5)

Classifier	*E.coli*		*B.subtilis*	
	All	Subgroup	All	Subgroup
Loglc	16.95	15.56	17.60	18.32
Fisherc	16.92	16.32	17.57	18.22
Naivebc*	17.63	18.35	20.27	18.25
Treec*	41.44	21.59	38.55	23.32

errors. In cases like *E.coli* or *B.subtilis*, a substantial number of operons have been obtained from previous experiments (8,19). For such cases, we found that very low classification errors can be achieved by dividing the dataset into groups according to the intergenic distance and by using a non-linear classifier, at 9.9% for *E.coli* and 15.2% for *B.subtilis*, as shown in Table 1. By comparing the strategy of whole data-based training and the strategy of subgroup-based training, we found the latter in general yields better results, as shown in Table 1. By comparing the prediction results between non-linear and linear classifiers, we found that the Naive Bayesian classifiers consistently perform well on several datasets, although decision tree-based classifiers give the most dramatic improvement in subgroup-based training. Our data suggests a new strategy for operon prediction, i.e. with a well-studied genome, such as *E.coli* or *B.subtilis*, non-linear classifiers, such as decision tree-based or Naive Bayesian method, when used in conjunction with subgroup-based training, will lead to an improvement in the prediction accuracy.

(B) Linear classifiers give the most generalization ability. An important application of operon prediction is to predict operons in new genomes using known operon data from well-studied genomes, such as *B.subtilis* or *E.coli*. Our evaluation shows that this approach increases the classification errors. In detail, we found that when using *E.coli*-trained classifiers to predict operonic pairs in *B. subtilis*, the smallest average classification errors is 17.57%, compared to 15.16% if the classifiers were trained with *B.subtilis* data, as shown in Tables 1 and 2. When using the *B.subtilis*-trained classifiers to predict operonic pairs in *E.coli*, the smallest average classification errors is 15.56%, compared with 9.91% if the classifiers were trained with *E.coli* data, as shown in Tables 1 and 2. Therefore, the increases of classification errors are from 2.4 to 5.8%. We also found that linear classifiers whose methods are based on the maximization of the likelihood criterion using the logistic function or the minimization of the errors in the least square sense consistently perform well across all datasets. By comparing the strategy of using whole data-based training and the strategy of using subgroup-based training, we found that the latter do not substantially reduce the classification errors, as shown in Table 2. Among the top two linear classifiers, the differences in classification errors, when changing from using the whole data-based training approach to using subgroup-based training approach, range from −0.7 to 1.4%, much smaller than the improvement we found when using decision tree-based classifiers and the subgroup-based training approach, as shown in the above section (A). Therefore, for generalization purpose, dividing the dataset into subgroups based on the intergenic distance does not give a clear advantage over using the whole dataset.

Together, our study suggests a new protocol for predicting operons in prokaryotic genomes that includes (i) with a well-studied genome, such as *E.coli* or *B.subtilis*, the dataset should be divided into subgroups based on the intergenic distance and trained with a non-linear classifier using the decision tree-based method or the Naïve Bayesian method and (ii) when the goal of training a classifier is to apply to other genomes, a linear classifier should be used, and the best methods are the maximization of the likelihood criterion using the logistic function and the minimization of the errors in the least square sense.

Contribution of selected features to operon prediction

Among information sources having been used for operon prediction, the most reliable features are derived directly from genomic sequences, such as the intergenic distance, the ratio between the lengths of an adjacent pair of genes, and frequencies of specific DNA motifs in the intergenic region. Features computed through mapping orthologous genes across genomes, such as neighborhood conservation scores or similarities of phylogenetic profiles can also be readily obtained, although they are not as reliable. Furthermore, experimentally derived data, such as GO annotation, the Riley's classification, knowledge of a gene pair being in the same biological pathway or complex, and correlation of gene expression patterns measured using microarray experiments may not be readily available for most gene pairs because very few organisms have genome-scale experimental data as outlined above. Because it is desirable to use features that are readily obtainable for operon prediction, we have avoided using many experimentally-derived features in our study.

In Operon prediction using different classification functions, we have used a set of core features found through our preliminary study that give good prediction results. These features include (i) the distance of the phylogenetic profiles, (ii) the frequency of motif TTTTT in the intergenic region, (iii) the intergenic distance and (iv) the GO-based functional similarity score. We now examine each feature in the core set and expand this core set to include a few other features. To evaluate the usefulness of these features in improving the performance of our classifiers, we trained our classifiers with or without these groups of features and compared the results.

(A) Prediction accuracy is affected by the length of the intergenic distance. As discussed in Operon prediction using different classification functions, the ratios of the frequency of operon pairs and the frequency of boundary pairs changes as the intergenic distance changes, which led us to closely examine the dependency of the classification errors and the intergenic distance, while previous studies often examine the prediction accuracy on the whole dataset. To address the issue, we report the average classification errors when the training and testing sets are from the same genome, and the dataset was divided into three groups based on the intergenic distance, resulting in three classifiers per classification method per dataset. This approach was shown in Operon prediction using different classification functions to be useful for operon prediction when the training and testing data are from the same genome. Our results in Table 3 suggest that (i) the U200 groups in both *E.coli* and *B.subtilis* yield the highest classification errors, averaging from 15 to 24%. (ii) Classifiers using decision tree based-method greatly reduce the classification errors in the O200 groups, resulting in almost 42% reduction in error rates, as shown in Table 3. We also found when the *E.coli*-trained classifiers were used to predict operonic pairs from *B.subtilis* genome, or when *B.subtilis*-trained classifiers were used to predict operonic pairs from *E.coli* genome, the U200 groups also yield the

Table 3. The dependency of the classification errors on the intergenic distances of gene pairs

Classifier	*E.coli*			*B.subtilis*		
	U40	U200	O200	U40	U200	O200
Naivebc*	8.75 ± 0.62	21.14 ± 2.13	7.56 ± 1.18	12.16 ± 0.38	22.16 ± 1.48	8.63 ± 0.97
Treec*	8.16 ± 0.95	15.06 ± 2.32	4.35 ± 1.31	9.41 ± 1.46	23.95 ± 1.85	5.17 ± 1.23

The dataset was divided into three subgroups including U40, U200 and O200 based on the intergenic distance of the gene pairs. The non-linear classifiers were trained using the subgroup-based training approach. The training and testing sets are from the same genome.

Table 4. The contribution of features in improving the classification errors of the decision tree-based classifier

Phylo	Length	IG	TTTTT	Neighbor	GO	*E.coli*			*B.subtilis*		
						U40	U200	O200	U40	U200	O200
+	−	−	−	−	−	8.08 ± 0.50	19.91 ± 2.49	5.26 ± 0.53	10.03 ± 0.6	26.09 ± 0.90	5.39 ± 0.54
−	+	−	−	−	−	6.44 ± 0.26	15.41 ± 0.53	5.13 ± 0.49	8.28 ± 0.34	18.02 ± 0.37	5.57 ± 0.40
−	−	+	−	−	−	9.47 ± 0.02	35.69 ± 0.85	9.56 ± 0.36	12.37 ± 0.03	28.16 ± 0.39	9.86 ± 0.04
−	−	−	+	−	−	9.47 ± 0.00	40.28 ± 0.00	9.91 ± 0.00	12.37 ± 0.00	30.98 ± 0.00	9.87 ± 0.00
−	−	−	−	+	−	7.61 ± 0.52	23.74 ± 1.13	5.19 ± 0.49	10.05 ± 0.38	22.23 ± 0.61	7.22 ± 0.72
−	−	−	−	−	+	9.47 ± 0.00	34.2 ± 0.56	7.35 ± 1.13	12.30 ± 0.07	29.72 ± 0.27	9.62 ± 0.22
+	+	−	−	−	−	6.07 ± 0.34	14.81 ± 0.70	5.24 ± 0.52	8.47 ± 0.38	17.62 ± 0.44	4.71 ± 0.24
−	+	−	−	+	−	5.87 ± 0.30	15.59 ± 0.67	5.05 ± 0.52	8.45 ± 0.35	17.46 ± 0.46	5.21 ± 0.42
+	+	−	−	+	−	5.82 ± 0.41	14.29 ± 0.79	4.53 ± 0.64	7.96 ± 0.33	17.60 ± 0.62	4.53 ± 0.39
+	+	+	+	+	−	5.79 ± 0.29	13.60 ± 0.36	4.81 ± 0.62	8.35 ± 0.33	15.97 ± 0.45	5.16 ± 0.32
+	+	+	+	+	+	5.72 ± 0.19	12.82 ± 0.63	4.43 ± 0.51	8.29 ± 0.33	16.21 ± 0.41	5.07 ± 0.30

The testing and training sets are from the same genome, and features are present (+) or absent (−) from the combination. Besides shown features, no other feature is used. The dataset was divided into three subgroups including U40, U200 and O200 based on the intergenic distance of the gene pairs.

highest average classification errors, and the result are reported in the Supplementary Table 3. Similar to our results, Zhang *et al*. (15) reported that most of the miss-assigned operonic pairs have the intergenic distance between 50 and 200 bases. Our results suggest that operon-pair prediction for adjacent gene pairs with intergenic distance between 40 and 200 bases is less reliable than other situations.

(B) In subgroup-based training, the distance of the phylogenetic profiles, the neighborhood conservation score and length ratio are critical features contributing to operon-prediction accuracy. We have first trained decision tree-based classifiers with various combinations of the core features, using half of the data from a genome as a training set, and the other half as the testing set. To test whether other features may also contribute to the improvement in the classification of gene pairs, we added new features including conserved neighborhood information, the remaining five DNA motifs discussed in the Materials and Methods, and the ratio of the lengths of a gene pair to our core feature set. As shown in Table 4, when comparing the performance, we found that among the tested features, the length ratios between the gene pairs give the most discerning power in all groups of both *E.coli and B.subtilis* genomes. In addition, both distances of the phylogenetic profiles and the neighborhood conservation scores also give good improvement to the prediction accuracy. However, the information provided by the phylogenetic profiles and the neighborhood conservation scores appears to overlap each other because addition of both features to the length ratio feature does not significantly improve the prediction accuracy when comparing to cases using only one of the two features, as shown in Table 4. Although the distance of phylogenetic profiles between a gene pair was previously used as a feature in operon prediction, it was not shown to be a main contributor to the predictors by any previous studies (13,15). Our results suggest that small phylogenetic distance is a very strong indicator of a gene pair belonging to the same operon when the intergenic distance of the gene pair is >200 nt.

Our data are the first to show that the ratio of the lengths of a gene pair is a powerful discerning feature for operon prediction. When inspecting the natural log of the length ratios, we found that boundary gene pairs are often associated with small values of the natural log of the length ratio, as shown in Supplementary Table 4. Because the length ratio is calculated as the ratio of the upstream-gene length to the downstream-gene length, our results suggest that the length of the downstream gene in proportion to the upstream gene affects the chance of the two genes being transcribed together. Besides the consistent improvement in classification errors found when the length ratio feature is used, the inclusion of other features including the intergenic distance, the GO-based functional similarity score and other DNA motifs also led to a consistent but moderate reduction in classification errors in the U200 groups as shown in Table 4.

When multiple features were combined, two of combinations yielded best results are shown on the last rows of Table 4. The small reduction in classification errors by the inclusion of GO-based functional similarity scores, the intergenic distance and the DNA motif is encouraging to us because it suggests that in this case operon prediction can work well even in the absence of these features. Although other DNA motifs can also be used, their contributions to error reduction seem to be genome-dependent (data not shown). Furthermore, we have confirmed that the TTTTT motif is significantly over-represented in both *E.coli* and *B.subtilis* genomes, using the formula discussed in the Supplementary Data.

(C) For generalization purpose, a combination of multiple features substantially improves the prediction accuracy in the

U200 groups. In section (B) Linear classifiers give the most generalization ability, we have found that linear classifiers generalize well across genomes. To further understand the contribution of the features used to prediction accuracy, the classifier whose method is based on the maximization of the likelihood criterion using the logistic function (logistic function-based) was trained on various combinations of selected features using the whole data-based training approach. The trained logistic function-based classifier was then used to predict if an adjacent gene pair in another genome is an operon pair or not. The prediction results were divided into three groups, namely U40, U200 and O200, and the classification error was calculated for each group. Table 5 shows a portion of the prediction results on both *E.coli* and *B.subtilis*. Our results indicate that the U200 groups have the best improvement in prediction accuracy when additional features were used. In the U200 groups, we see 12–21% improvement when comparing the results from using all features against the results from using the intergenic distances alone, as shown in Table 5. For the two other groups, inclusion of other features yields very little reduction in prediction errors. Among all tested features, the GO-based functional similarity score accounts for 5–7% of the 12–21% improvement in prediction accuracy for the U200 groups, suggesting that inclusion of GO-based functional similarity scores is necessary. Comparing the list of features that we found useful in this section and above section, we found that both phylogenetic distances and length ratio, while they are critical for improving the prediction accuracy when the testing and training data are from the same genome, do not yield a similar level of improvement when the training and testing data are from different genomes. This suggests that these features are genome-specific. In addition, we did not find any individual feature that gives a large improvement in classification errors as found in the case of using training and testing data from the same genome. Rather, the analysis results suggest that the combination of all features give better results than using only one feature, as shown in Table 5 (and data not shown). However, inclusion of other features to the last combination set of features shown in Table 5 did not further improve the prediction accuracy (data not shown).

Prediction results

(A) Our classifiers have lower classification errors than previously published results. Based on results shown in the previous sections, we trained several classifiers using either the logistic function-based method or the decision tree-based method, and used the trained classifiers to predict if an adjacent gene pair is an operon or a boundary pair, for all gene pairs in *E.coli* and *B.subtilis*. Then we calculated the specificity, sensitivity and the accuracy of each trained classifier on the known datasets from these two genomes, and reported the results in Table 6. Because in many previous reports, the accuracy was computed as an average of specificity and sensitivity of the positive set (the set of operonic gene pairs), we also included this score in the 'Average' column of Table 6 for the ease of comparison. Our results show that when training a classifier using operon data from the same genome as the testing genome, the non-linear classifier using decision tree-based method gives the prediction accuracy at 90.2 and 93.7% in *B.subtilis* and *E.coli*, respectively, or 90.7 and 94.7% if the accuracy is calculated as the average of prediction sensitivity and specificity. Hence our prediction program performs substantially better than the best previously reported studies which gave prediction accuracy, when using training and testing data from the same genome, at 88% in *B. subtilis* to 91% in *E.coli* (2,8,15). It is worth noting that the only piece of experimental data that we used is GO-based functional annotation, whereas to achieve the 91% prediction accuracy in *E.coli*, other authors have used multiple sources of data that are either derived directly from experiments or indirectly through mapping of experimental data from other organisms. These data include protein complex, genes in the same pathway, and microarray data in the case of *B.subtilis*.

If we do not use training data from the same genome as the target genome, our study suggests that the linear classifier using logistic function-based method is among methods that give the best prediction results. For example, we have achieved 84.6% prediction accuracy on *E.coli* when using known operons from *B.subtilis* as the training set; similarly we have achieved 83.2% prediction accuracy on *B.subtilis*, when the classifier is trained on *E.coli*. If the accuracy is calculated as the average of sensitivity and specificity of the

Table 5. The contribution of features in improving the classification errors of the logistic function-based classifier

IG	Neighbor	GO	Length	Phylo	TTTTT	ATA	*E.coli*			*B.subtilis*		
							U40	U200	O200	U40	U200	O200
+	−	−	−	−	−	−	9.47 ± 0.00	33.77 ± 0.53	9.91 ± 0.00	12.37 ± 0.00	28.95 ± 0.12	9.87 ± 0.00
+	+	−	−	−	−	−	9.47 ± 0.00	32.16 ± 0.57	9.91 ± 0.00	12.37 ± 0.00	27.98 ± 0.86	9.64 ± 0.24
+	−	+	−	−	−	−	9.47 ± 0.00	31.75 ± 0.93	9.91 ± 0.00	12.36 ± 0.08	27.38 ± 0.97	9.82 ± 0.14
+	−	−	+	−	−	−	9.47 ± 0.00	33.01 ± 0.63	9.91 ± 0.00	12.37 ± 0.00	28.93 ± 0.27	9.87 ± 0.00
+	−	−	−	+	−	−	9.51 ± 0.07	33.29 ± 0.65	9.91 ± 0.00	13.11 ± 0.57	29.07 ± 0.78	9.87 ± 0.00
+	−	−	−	−	+	−	9.47 ± 0.00	34.03 ± 0.65	9.91 ± 0.00	12.68 ± 0.24	26.46 ± 0.51	9.87 ± 0.00
+	−	−	−	−	−	+	9.53 ± 0.08	33.01 ± 0.68	9.91 ± 0.00	12.37 ± 0.00	29.24 ± 0.53	9.87 ± 0.00
+	+	+	−	−	−	−	9.47 ± 0.00	31.45 ± 0.32	9.91 ± 0.00	12.36 ± 0.05	26.72 ± 0.45	9.64 ± 0.32
+	+	+	+	−	−	−	9.47 ± 0.00	30.52 ± 1.06	9.91 ± 0.00	12.36 ± 0.05	26.62 ± 0.82	9.51 ± 0.35
+	+	+	+	+	−	−	9.49 ± 0.06	30.21 ± 0.78	9.91 ± 0.00	13.48 ± 0.79	26.97 ± 0.81	9.60 ± 0.31
+	+	+	+	+	+	−	9.42 ± 0.12	29.48 ± 1.07	9.91 ± 0.00	13.54 ± 0.68	24.58 ± 0.47	9.60 ± 0.38
+	+	−	+	+	+	+	9.28 ± 0.21	31.42 ± 1.37	9.91 ± 0.00	13.61 ± 0.48	25.36 ± 0.83	9.78 ± 0.28
+	+	+	+	+	+	+	9.25 ± 0.14	29.57 ± 0.82	9.86 ± 0.15	13.45 ± 0.50	23.92 ± 0.70	9.42 ± 0.47

The testing and training sets are from different genomes, and features are present (+) or absent (−) from the combination. Besides shown features, no other feature is used. The whole data-based training approach was used. After prediction, the classification errors were calculated for each subgroup.

Table 6. Sensitivity, specificity, the average of sensitivity and specificity and accuracy of operon prediction

Train/Test	Boundary gene pairs				Operonic gene pairs			Accuracy (%)
	Genome	Sensitivity (%)	Specificity (%)	Average (%)	Sensitivity (%)	Specificity (%)	Average (%)	
Same	*E.coli*	90.54	93.95	92.24	95.90	93.52	94.71	93.69
Genome	*B.subtilis*	89.55	89.90	89.72	90.82	90.50	90.66	90.22
Different	*E.coli*	81.09	81.58	81.33	87.13	86.76	86.94	84.63
Genomes	*B.subtilis*	81.03	83.40	82.22	85.29	83.14	84.22	83.26

When the testing and training data are from the same genome, the decision tree-based classifier was used, whereas when the testing and training sets are from different genomes, the logistic function-based classifier was used.

positive set, these numbers are equivalent to 84.9s and 86.2% in *B.subtilis* and *E.coli*, respectively. These numbers compare favorably to the previously best programs, which only gave 64% prediction accuracy on *B.subtilis* when trained on *E.coli* (2).

To confirm that the final classifiers used in this step are not over-trained, we randomly split the *E.coli* known dataset into two groups, and then calculated the prediction accuracy for each group using the same classifier. After 100 trials, we found that using decision tree-based classifiers, the average difference between the classification error rates of the two groups is <0.57%, and the standard deviation of the differences is <3%. Using the logistic function-based classifiers, the average difference between the classification error rates of the two groups is <0.84%, and the standard deviation of the differences is <2.8%. Similar results were observed with the classifiers trained on the *B.subtilis* known dataset. The small average differences and standard deviations suggest that the optimum classifiers chosen are not over-trained.

(B) We have better accuracy in predicting operon boundaries. The ultimate goal of operon prediction is to define the boundaries of operons in a genome. From the RegulonDB database, we have collected 289 unique TUs that contain two or more genes, and 256 unique single-gene TUs in *E.coli*. In addition, there are additional 145 TUs that are partially overlapped with each other. Previous study reported 69% accuracy in predicting unique multi-gene TUs, and an overall of 79% accuracy in predicting unique TUs (containing both single and multi-gene) in *E.coli* (2). For these two measures, our program has achieved 80.6 and 92.2% prediction accuracy on the same dataset. In *B.subtilis*, our program has achieved 73% accuracy in predicting unique multi-gene TUs, and 88.2% accuracy in predicting all unique TUs.

Overall, we have predicted 2586 TUs (single-gene and multi-gene) in *E.coli* and 2540 TUs in *B.subtilis*. In *E.coli*, we predicted 1680 single-gene TUs, and 906 multi-gene operons. In *B.subtilis*, we predicted 1734 single-gene TUs and 806 multi-gene operons. The total number of *E.coli* TUs predicted by our program is similar to others such as 2381 TUs (3) and 2646-2796 TUs (19). However, the number of multi-gene operons was previously predicted to be 717-831(19) or 837 (15) in *E.coli*, but it is slightly higher (906) as predicted by our program, and the reason will be addressed in below section. The list of these operons is available at http://csbl.bmb.uga.edu/~phd/.

(C) Implications of our prediction results on the overlapped transcripts. The genes belonging to 145 overlapping TUs in *E.coli* are found to cluster in 57 regions of the *E. coli* chromosome, suggesting that an average of 2.54 TU can be transcribed from each of these DNA regions. When using the operon prediction program, each DNA region should be predicted to belong to at least one TU (or more), producing at least 57 predicted operons from these regions. Among the 145 overlapped TUs, we found that 32 are exactly matched to our predicted operons. By closely examining these matched overlapped-TUs, we found that 18 predicted operons were matched to the longest overlapped-transcripts, and 11 were matched to the shortest overlapped-transcripts among all possible transcripts, suggesting that our decision tree-based predictor tends to predict either the longest or the shortest forms when multiple transcripts containing the same gene are observed. Because our protocol tends to predict the longest form of the overlapped TUs, it is not surprising to find that the number of multi-gene operons in the *E.coli* genome predicted by our program is slightly higher than the previously published results, as noted above.

(D) Our results correlate well with gene expression data. Our prediction results can also be evaluated using the gene expression data. If a gene pair is correctly predicted to be an operon pair, its Pearson correlation coefficient should be high as this gene pair would be transcribed together in one mRNA. Using the microarray gene expression data obtained from 180 experiments in *E.coli* (http://genome-www5.stanford.edu/ #35), we computed the distribution of the Pearson correlation coefficient of the set of operonic gene pairs and the set of boundary gene pairs, and then compared the results to the distribution of the Pearson correlation coefficient of the set of randomly chosen gene pairs. The results showed in Figure 1 suggest that the distribution of the Pearson correlation coefficient of 1000 randomly chosen gene pairs broadly follows a normal distribution, whereas the distribution of the known operonic gene pairs is skewed to the right as expected because genes in the same operons are usually transcribed together. Interestingly, the distribution of the Pearson correlation coefficients of the known boundary pairs is also slightly skewed to the right, suggesting that although an adjacent gene pair is not from the same operon, their gene expression could still be slightly correlated. Our observation makes biological sense as the RNA polymerase machinery probably has better chance to continue transcribing the downstream adjacent gene after it has finished transcribing the upstream gene or operon. Furthermore, we observed that 50% of the operon pairs and 18% of the boundary pairs have the Pearson correlation coefficient of larger than or equal to 0.78, confirming that adjacent pairs belonging to the same operons are in general transcribed together. Similarly, we have computed the Pearson correlation coefficient of unknown pairs that are predicted to be operonic gene pairs or boundary gene pairs, and calculated the distribution of Pearson correlation coefficient of these

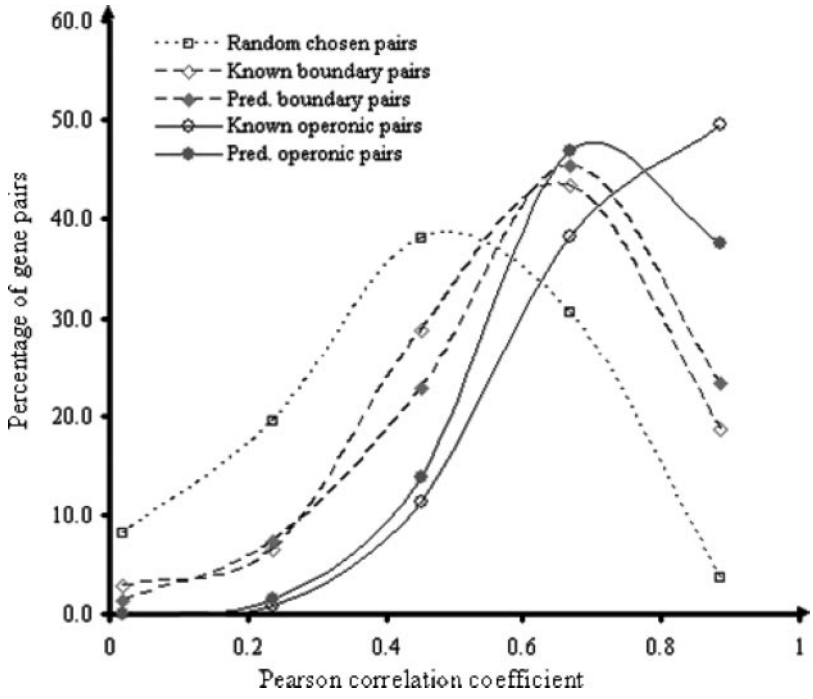

Figure 1. The distribution of Pearson correlation coefficients of *E.coli* gene pairs calculated from the gene expression data. The X axis indicates the Pearson correlation coefficients. The Y axis is the density function for each of the following five sets of gene pairs: the randomly chosen pairs (square), the known boundary pairs (diamond), the unknown pairs predicted to be boundary pairs (filled diamond), the known operonic pairs (circle) and the unknown pairs predicted to be operonic pairs (filled circle).

sets of gene pairs. We found that the distribution of the Pearson correlation coefficient of the unknown pairs that are predicted to be operonic gene pairs are closely resemble that of known operonic gene pairs, and the distribution of the Pearson correlation coefficient of the unknown pairs predicted to be boundary pairs are closely resemble that of known boundary pairs. These results indicate that at the gene expression level, our prediction results of the unknown gene pairs are closely resemble the known gene pairs, suggesting that the classification errors calculated from the known gene pairs probably correspond well with the classification errors of the unknown gene pairs.

DISCUSSION

In this study, we have reported a new operon prediction method that utilizes multiple sources of genomic information. To obtain the best possible results, we have tested a number of classification methods, and found that the decision tree-based classifier gives the best prediction performance when the classifier is trained using data from the same genome as the target genome, and a linear classifier using the logistic function to maximize the likelihood criterion is one of the top classifiers that gives the best prediction performance when the classifier is trained on a genome other than the target genome. By using experimentally verified data from two genomes and exploring different classification functions, we have gained new insights about what type of classification function is most effective for operon prediction under different conditions. We believe that this study has led to a better prediction program with better generalization ability. Compared to existing operon prediction programs, our program performs better in predicting operonic gene pairs as well as in recognizing operon boundary. We believe that the improvement in predicting operon boundaries is critical for the successful application of other genome analysis tools such as computational prediction (and experimental confirmation) of transcriptional factor binding sites or functional annotation of unknown genes.

Our study is the first to show the discerning power of the length ratio between an adjacent gene pair; and this feature is most valuable when the training and testing data are from the same genome. Very little improvement is gained when including this feature in a linear classifier, implicating that these scores are not compatible between *E.coli* and *B.subtilis*. This observation suggests that there may be differences in the activity of *B.subtilis* RNA polymerase and *E.coli* RNA polymerase. Although we were not able to find any experimental study that addressed this issue directly, we found a previous report suggesting that *E.coli* and *B.subtilis* RNA polymerases could exhibit differences at the level of catalysis and signal recognition (22). Further experiments are clearly needed to confirm or reject this hypothesis. In addition, our study also indicates that for the U200 and O200 groups, the distance of the phylogenetic profiles is a major contributor to the reduction in error rates in our prediction, confirming that in bacterium there is a pressure to keep genes working in the same biological process to be transcribed together.

The improvement in the operon prediction accuracy achieved in our study might be due to the following factors: (i) a larger dataset compared to previous similar studies, (ii) one new feature used and (iii) a new classification

method. Certainly, our datasets are larger than most of the datasets used in the previous studies. For example, for the *E.coli* dataset, we obtained 289 operons with 707 operonic gene pairs. Previous works often used 237 operons, and the operonic gene pairs range from 641 to 807. For the *B.subtilis* data, previous studies used 100–635 known operons with 582–703 operonic gene pairs. However, in Table 1, we have shown that when training on all data, using the naïve Bayesian network method and five features (the distances of the phylogenetic profiles, the intergenic distances, the frequency of the TTTTT motif and the GO-based functional similarity score), we only achieved 14 and 17% misclassification error rates on the *E.coli* and *B.subtilis* data, respectively. These results are not better than the previous results, suggesting that the larger datasets do not help in improving the prediction accuracy. However, when using the decision-tree based method and using the strategy of splitting the data to three subgroups based on the intergenic distances, we achieved 10 and 15% misclassification errors on the same datasets, respectively, suggesting that it is the decision-tree based method that helps to improve the prediction accuracy. Furthermore, in Table 4, we show that the length ratio feature alone gives the best prediction accuracy when using the decision-tree based method. Therefore, we believe that the improvement in the prediction accuracy is due to the use of the decision-tree based method and the new feature.

Based on our prediction results, we predicted 218 multigene transcripts that contain both hypothetical proteins and annotated genes, and another 35 operons containing only hypothetical proteins in *E.coli*. In *B.subtilis*, we predicted 172 transcripts that contain both hypothetical protein and genes with annotated functions, and 393 operons containing only hypothetical proteins. The list of these operons is available at http://csbl.bmb.uga.edu/~phd/. As shown in Figure 1, many of these unknown genes predicted to be in the same operons also have high Pearson correlation coefficient scores of their expression patterns, suggesting that they are possibly transcriptionally co-regulated. Although the high Pearson correlation coefficient score is an indication of co-transcription, further experiments are needed to confirm our prediction.

From the results of our study, it is apparent that for a genome with a large number of known operons such as *B.subtilis* or *E.coli* we can achieve 90 and 94% accuracy in predicting whether a gene pair belongs to the same operon. In theory, the results will correspond to a maximum of 81 and 88% accuracy in predicting the boundary of an operon in *B.subtilis* and *E.coli,* respectively, although we only achieved 73 and 81% in our study. When applying a trained classifier to another genome at the gene pair level, prediction accuracy drops at least 8–10%, suggesting that the corresponding accuracy in predicting the operon boundary will be maximally at the level of 69–72%. Similar approach to ours can be used to further improve the generalization of the classifiers if other large datasets are available. In fact, Okuda and colleagues have already begun their effort to collect all known operons in other genomes (23).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Hongwei Wu, Dr Fenglou Mao, Dr Xin Chen and other members of the Computational Systems Biology Laboratory for their helpful discussions. This work was supported in part by the National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204, NSF/DBI-0542119), and also by a 'Distinguished Scholar' grant from the Georgia Cancer Coalition. Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation (NSF/DBI-0354771).

Conflict of interest statement. None declared.

REFERENCES

1. Burton,Z.F., Gross,C.A., Watanabe,K.K. and Burgess,R.R. (1983) The operon that encodes the sigma subunit of RNA polymerase also encodes ribosomal protein S21 and DNA primase in *E.coli* K12. *Cell*, **32**, 335–349.
2. Romero,P.R. and Karp,P.D. (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*, **20**, 709–717.
3. 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.
4. 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.
5. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
6. 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.
7. Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227–1235.
8. De Hoon,M.J., 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.*, 276–287.
9. Westover,B.P., Buhler,J.D., Sonnenburg,J.L. and Gordon,J.I. (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880–888.
10. 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.
11. Chen,X., Su,Z., Dam,P., Palenik,B., Xu,Y. and Jiang,T. (2004) Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. *Nucleic Acids Res.*, **32**, 2147–2157.
12. Edwards,M.T., Rison,S.C.G., Stoker,N.G. and Wernisch,L. (2005) A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic Acids Res.*, **33**, 3253–3262.
13. Chen,X., Su,Z., Xu,Y. and Jiang,T. (2004) Computational prediction of operons in Synechococcus sp. WH8102. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 211–222.
14. 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.
15. 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.
16. 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.
17. Jacob,E., Sasikumar,R. and Nair,K.N. (2005) A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics*, **21**, 1403–1407.

18. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303–D306.
19. 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.
20. Wu,H., Su,Z., Mao,F., Olman,V. and Xu,Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.*, **33**, 2822–2837.
21. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
22. Artsimovitch,I., Svetlov,V., Anthony,L., Burgess,R.R. and Landick,,R. (2000) RNA Polymerases from Bacillus subtilis and *Escherichia coli* Differ in Recognition of Regulatory Signals. *In Vitro. J. Bacteriol.*, **182**, 6027–6035.
23. Okuda,S., Katayama,T., Kawashima,S., Goto,S. and Kanehisa,M. (2006) ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res.*, **34**, D358–D362.