Title

A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes

Authors

Price, Morgan N.
Huang, Katherine H.
Alm, Eric J.
et al.

Publication Date

2004-12-01

Peer reviewed

Title: A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes

Authors: Morgan N. Price, Katherine H. Huang, Eric J. Alm, and Adam P. Arkin

Author affiliation: Lawrence Berkeley Lab, Berkeley CA, USA. A.P.A. is also affiliated with the Howard Hughes Medical Institute and the UC Berkeley Dept. of Bioengineering.

Corresponding author: Eric J. Alm, ejalm@lbl.gov, phone 510-843-1794, fax 510-486-6059, address Lawrence Berkeley National Lab, 1 Cyclotron Road, Mailstop 939R704, Berkeley, CA 94720

Abstract:

We combine comparative genomic measures and the distance separating adjacent genes to predict operons in 124 completely sequenced prokaryotic genomes. Our method automatically tailors itself to each genome using sequence information alone, and thus can be applied to any prokaryote. For *Escherichia coli K12* and *Bacillus subtilis*, our method is 85% and 83% accurate, respectively, which is similar to the accuracy of methods that use the same features but are trained on experimentally characterized transcripts. In *Halobacterium NRC-1* and in *Helicobacter pylori*, our method correctly infers that genes in operons are separated by shorter distances than they are in *E. coli*, and its predictions using distance alone are more accurate than distance-only predictions trained on a database of *E. coli* transcripts. We use microarray data from six phylogenetically diverse prokaryotes to show that combining intergenic distance with comparative genomic measures further improves accuracy and that our method is broadly effective. Finally, we survey operon structure across 124 genomes, and find several surprises: *H. pylori* has many operons, contrary to previous reports; *Bacillus anthracis* has an unusual number of pseudogenes within conserved operons; and *Synechocystis PCC 6803* has many operons even though it has unusually wide spacings between conserved adjacent genes.

Introduction

As the gap grows between the sequencing of complete microbial genomes and the characterization of transcriptional regulation in those organisms, automated methods for predicting regulatory interactions are becoming a high priority. Automated prediction of operon structure in prokaryotic genomes is particularly important because it provides the most confident predictions that two genes are co-regulated and because other computational analyses, such as prediction of cis-regulatory elements, often rely on operon predictions.

Most previous efforts to predict operons focused on *Escherichia coli* and *Bacillus subtilis*, and relied on databases of experimentally identified transcripts for training and for validation [1, 2, 3, 4, 5, 6]. Unfortunately, databases of characterized transcripts are available for only a few organisms, making it difficult to judge the accuracy of current operon prediction methods on new genome sequences. Thus, unsupervised methods for operon prediction – methods that do not require large databases of known operons – are needed, along with new methods for validation of those predictions.

We present a statistical framework for estimating the likelihood that two adjacent genes are contained within the same transcriptional unit (TU). Our method is based on genome sequences only, and is free from parameters optimized to reproduce experimentally characterized operons. Nevertheless, our method's predictions correspond well with databases of experimentally determined operons in *E. coli* and *B. subtilis*. To show that our method is effective across the prokaryotes, we use the observation that genes in the same operon usually have similar expression profiles, whereas other adjacent genes do not [3]. We demonstrate qualitative agreement between our method's predictions and microarray data from six phylogenetically diverse prokaryotes, and introduce a procedure to estimate the quantitative accuracy of operon predictions from microarray data.

Two approaches have previously been proposed to predict operons in uncharacterized species. The first relies on identifying operons that are conserved in multiple species, as genes that remain adjacent across long stretches of evolutionary time are likely to be in the same operon [7]. This method allows highly confident prediction of many operons, but the majority of the operons in *E. coli* cannot be predicted this way [7]. We suspect that this is because many operons are evolutionarily new ([8]; M.N.P, K.H.H, E.J.A, A.P.A., submitted), and neutral conservation of gene order within the closely related genomes that do contain these new operons makes it hard to distinguish new operons from non-operons by conserved gene order alone.

The second method relies on the fact that genes in the same operon tend to be separated by fewer base pairs of DNA. In *E. coli* and *B. subtilis*, this tendency can be quantified from known transcripts to give the probability that two adjacent genes are on the same operon as a function of their intergenic distance. It has been proposed that these probabilistic "distance models" can be transferred from one species to another unrelated species, but this *ad hoc* approach has only been validated for *E. coli* and *B. subtilis* [9]. A subsequent study indicated that, in general, intergenic distances within conserved operons vary across species [10]. Thus, *E. coli*'s distance model may not always be effective – indeed, we use microarray data to show that it is less effective for *Halobacterium NRC-1* or for *Helicobacter pylori*.

An Unsupervised Approach to Predicting Operons

Principles

The key elements of our approach are (i) to use both comparative and distance information and (ii) to infer a genome-specific distance model from preliminary comparative-only predictions. The method relies on a key assumption that the greater conservation of adjacency for genes on the same strand of DNA, compared to opposite-strand pairs, is entirely due to operons. This assumption has previously been used to identify conserved operons [7]. In practice, this assumption implies that most adjacent pairs that are conserved across significant evolutionary distances (e.g., across the γ-proteobacteria) are operon pairs, with a probability increasing with the extent of conservation.

Although in some cases, pairs which are clearly not in operons (opposite-strand pairs) are conserved across significant evolutionary distances [11, 12], we do not know of any process that would produce conserved not-operon (same-strand) pairs but not conserved opposite-strand pairs. We also make the analogous assumption for the greater functional relatedness of same-strand versus opposing-strand pairs. In the Results, we validate the combined assumption directly by analyzing known transcripts in *E. coli* and *B. subtilis*. We do *not* make any assumption about the intergenic distances between genes on the same or different strands, because there are biological reasons for these to be different for convergent, divergent, and not-operon (same-strand) gene pairs [10].

As with most previous approaches to operon prediction, we focus on pairs of adjacent genes, and estimate the probability that each pair is in the same operon. We do not attempt to predict alternative transcripts due to internal promoters, terminator read-through, etc., as this remains a challenging problem even in *E. coli*, where transcriptional control features are relatively well characterized [4]. Instead, we define two adjacent genes to be on the same operon if a transcript that contains both genes exists, even if alternative transcripts exist that contain only one of the genes.

Features

For each pair of adjacent genes on the same strand, we consider:

- *distance* – the number of base pairs separating the two genes,
- *comparative features* – how often their orthologs are near each other (within 5 kb) in other genomes,
- *functional similarity* – whether their predicted functions are in the same category (from COG [13]), and
- *similarity of CAI* – the similarity of their codon adaptation index (CAI), a measure of synonymous codon usage [14].

Both distances between genes and comparative features have previously been used in unsupervised operon predictions and are the most informative features for predicting operons [1, 9, 7]. To increase the sensitivity of the comparative approach, we computed separate features for the closely related and distantly related genomes (see Methods). We used features that reflect similarity of function and similarity of codon usage because such features have been reported to improve prediction accuracy in *E. coli* [1, 2, 5].

Statistical Inference

The key challenge for an unsupervised approach is to estimate, from sequence alone, the probability that two adjacent genes are in the same operon given the values of the features. We first infer the distribution of the comparative and functional features for operon and not-operon pairs by using the assumption described above, as shown in Figure 1. The observed distribution of values for same-strand pairs is a mixture of the distribution for operon pairs, which is unknown, and for not-operon pairs, which by assumption is approximated by the observed distribution for opposite-strand pairs. If we know the relative fraction of operon and not-operon pairs in the same-strand set, then we can estimate the unknown distribution $P(Value|Operon)$ for operon pairs by "subtracting" out the contribution from not-operon pairs. This proportion of operon pairs in the same-strand set ($P(Operon|Same)$) can be estimated from the number of runs of same-strand pairs in the genome [7, 15]. In the Methods, we extend this approach to estimate $P(Operon|Same)$ to genomes with coding strand bias.

To perform the "subtraction," we use likelihood ratios rather than probabilities. Specifically, we estimate the likelihood ratio $P(Same|Values)/P(NotSame|Values)$, where "Values" refers to the comparative/functional features and "Same" refers to same-strand vs. opposing-strand pairs, from the observed distributions (see Methods). We then use the following formula:

$$\frac{P(Values|Operon)}{P(Values|NotOperon)} \approx \frac{\frac{P(NotSame)}{P(Same)} \cdot \frac{P(Same|Values)}{P(NotSame|Values)} - P(NotOperon|Same)}{P(Operon|Same)} \qquad \text{(Eq. 1)}$$

which can be derived from our assumption

$$P(Values|NotOperon) \approx P(Values|NotSame) \qquad \text{(Eq. 2)}$$

by treating $P(Values|Same)$ as a mixture of $P(Values|Operon)$ and $P(Values|NotOperon)$.

We then produce a genome-specific distance model from these likelihood ratios. This follows the same approach of considering distributions as mixtures, but is slightly more complicated because we do not have a "true negative" set of gene pairs to train from (we consider only distances between genes on the same strand). Instead, we split the pairs into those with high and low comparative/functional likelihood ratios, and treat these as preliminary operon predictions. By once again invoking the key assumption that not-operon pairs resemble opposite-strand pairs with respect to the comparative/functional features, we estimate that the false positive error rate of these predictions equals the fraction of opposite-strand gene pairs "predicted" to be on the same operon. We make these predictions for the opposite-strand pairs, even though we already know that they can never be co-transcribed, only so that we can estimate the false positive error rate $P(NotOperon|High)$. Thus we have

$$P(High|NotOperon) \approx P(High|NotSame) \qquad \text{(Eq. 3)}$$

$$P(NotOperon|High) \approx P(High|NotSame) \cdot \frac{P(NotOperon|Same)}{P(High|Same)} \qquad \text{(Eq. 4)}$$

where “High” refers to pairs with high comparative/functional likelihood ratios, which are more likely to be in the same operon. The false negative error rate $P(Operon|Low)$ can be derived from the number of “missing” predictions:

$$P(Operon|Same) = P(Operon|High) \cdot P(High|Same) + P(Operon|Low) \cdot P(Low|Same) \qquad \text{(Eq. 5)}$$

We estimate the likelihood ratio $P(Distance|Operon)$ / $P(Distance|NotOperon)$ from these error rates and the observed distributions $P(Distance|High)$ and $P(Distance|Low)$ for the two sets of same-strand pairs (see Methods). At this point, we have likelihood ratios from the comparative/functional features and from the genome-specific distance model. We use these preliminary predictions to estimate likelihood ratios for the remaining feature, the similarity of CAI, but without the error estimation step. Finally, we multiply the likelihood ratios for all the features with the *a priori* likelihood ratio to give the overall prediction:

$$\frac{P(Operon|AllFeatures)}{P(NotOperon|AllFeatures)} = \frac{P(Operon|Same)}{P(NotOperon|Same)} \cdot \frac{P(Values|Operon)}{P(Values|NotOperon)} \cdot \frac{P(Distance|Operon)}{P(Distance|NotOperon)} \cdot \frac{P(CAI|Operon)}{P(CAI|NotOperon)} \qquad \text{(Eq. 6)}$$

This “naive Bayes” approach makes the assumption that distance, the comparative/functional features, and the similarity of CAI are conditionally independent, which is approximately true (data not shown).

Results

Test of key assumption

We tested the key assumption – that not-operon pairs and opposite-strand pairs will have the same distributions of values for the comparative and functional features – against databases of characterized transcripts for *E. coli* and *B. subtilis* [16, 17]. Specifically, we compared the preliminary comparative/functional predictions for "known" not-operon pairs to the corresponding "predictions" for opposite-strand pairs. We defined known not-operon pairs as those same-strand pairs that straddle the boundaries of a known TU and are not in any known alternative transcript (following [1]). As shown in Figure 2A & 2B, the distribution for the known not-operon pairs is similar to that for opposite-strand pairs in both organisms.

Interestingly, in *B. subtilis*, some of the not-operon pairs have unusually low probabilities of being in an operon, highlighting a potential caveat of using these primarily literature-culled databases: there is a predominance of highly conserved genes (present in many other genomes) in this small data set. Because the comparative/functional predictions will only conclude that two genes are very *un*likely to be in the same operon if both genes are conserved but present in different regions of the genome in several other genomes, genes that are conserved in more genomes will tend to be more confidently predicted to fall in different operons (left-shifted in Figure 2B).

In addition, the not-operon set contains too many genes strongly predicted to occur in the same operon, particularly for *B. subtilis.* A previous investigation of conserved "known" not-operon pairs in *E. coli* found evidence in the literature that many of them are in fact co-transcribed [7]. In *B. subtilis*, we checked the 19 known not-operon pairs that we predicted to be $> 90\%$ likely to fall in the same operon (based on the comparative and functional features) against TU diagrams and Northern hybridizations at BSORF (http://bacillus.genome.ad.jp/bsorf.html). Northerns were only available for three pairs (*sul*/*folA*, *mmgE*/*yqiQ*, and *deoR*/*dra*), but in all three cases, there was a transcript containing both genes that was not present in the original database. Furthermore, in both *E. coli* and *B. subtilis*, the conserved and/or functionally related not-operons (those with comparative/functional $P(Operon) > 0.9$) are significantly more co-expressed than other not-operon pairs (Figure 2C & 2D: both $p < 0.01$, Kolmogorov-Smirnov test). Based on these results, we conclude that the modest deviations from the assumption are due to co-transcription of the "known" not-operon pairs, perhaps reflecting alternative transcripts. In the next section, we demonstrate that the assumption ultimately leads to accurate operon predictions.

Accuracy of operon predictions

We tested the accuracy of our unsupervised method in three ways. First, for *E. coli* and *B. subtilis*, we compared our predictions to known operons. We also compared the performance of the unsupervised method to that of a similar supervised method that we optimized using the known operons. Second, we defined a procedure to estimate prediction accuracy from microarray expression data, and measured our performance this way across six phylogenetically diverse prokaryotes. Finally, we established that our internal confidence values approximate the observed accuracy of individual predictions and then used this internal estimate of accuracy as an indicator of performance in genomes for which no additional data is available to test against.

Accuracy against known transcripts

The simplest metrics to describe the effectiveness of an operon prediction method are sensitivity – the proportion of true operon pairs that are correctly predicted – and specificity – the proportion of true not-operon pairs that are correctly predicted. These metrics require binary predictions (a pair of genes is either in an operon or not) – we used a threshold of $P(Operon|AllFeatures) > 0.5$, or more likely to be in an operon than not. Other thresholds can be used if higher sensitivity or specificity is preferred. With this threshold, the unsupervised method has sensitivity and specificity of 88.3% and 79.9% respectively in *E. coli* and 90.9% and 71.0% respectively in *B. subtilis*. For a threshold-independent measure of accuracy, we used the area under the receiver operating characteristic curve (AOC, [18]) shown in Figure 3A and 3B. AOC is equal to the probability that a randomly selected known operon pair will have a higher score than a randomly selected known not-operon pair. Thus, an AOC of 0.5 reflects an uninformative (random) predictor, and an AOC of 1.0 corresponds to perfect predictions. In *E. coli*, the AOC is 0.920 for the unsupervised approach, versus 0.919 for the supervised method, and in *B.subtilis*, the AOCs are 0.888 and 0.907 respectively. (To measure the accuracy of the supervised method, we used 100-fold cross-validation.) Furthermore, the distance models inferred by our unsupervised approach are similar to the supervised models in both organisms (Figures 3C and 3D). We also compared our unsupervised results to several previously published supervised methods, and found that its accuracy was comparable except when the supervised methods used significant additional information, such as microarray data (Supplementary Table 1). Overall, the unsupervised method is quite accurate at predicting known TU boundaries, even though known transcripts are not used to optimize any part of the method.

Accuracy against microarray data

To test operon predictions more broadly, we compared the unsupervised predictions to microarray data from six species. We found that the microarray data correlates with predictions and obtained quantitative estimates of prediction accuracy from the microarray data. To measure whether genes predicted to be in the same operon have similar expression patterns, we used a standard metric: the Pearson (linear) correlation between the normalized log-ratios of the two genes (r).

In all six species, predicted operon pairs are strongly coexpressed relative to other adjacent pairs on the same strand (Figure 4). Predicted not-operon pairs show little coexpression, similar to opposite-strand pairs, which we used as negative controls. Moreover, as shown in Figure 5A, the average strength of the correlations increases with the estimated probability $P(Operon|AllFeatures)$ that the genes are in the same operon.

We also used agreement with gene expression data to test whether the method was using informative features, and whether it was combining those features effectively. The distance models are responsible for a majority of the agreement with microarrays, which strongly suggests that the method is predicting operons rather than identifying functionally related pairs of adjacent genes (Table 1). Combining comparative genomics with intergenic distance improves agreement greatly over using either measure alone (Table 1), and in five of the six species, the combined comparative/functional predictions outperform the best single comparative feature (not shown). In contrast, similarity of CAI has little effect on the final predictions and does not give a consistent improvement (not shown). The greater agreement with microarrays of distance-only predictions, relative to the comparative/functional predictions, is consistent with the hypothesis that many operons are too new to be identified by comparative genomics alone [8].

Finally, we used microarray data to estimate the absolute accuracy of the predictions. To do this, we modeled the observed distributions of correlations for predicted operon and not-operon pairs as mixtures of the distributions for true positives and false positives. We approximated the latter with the observed distribution for opposite-strand pairs, following the assumption that not-operon pairs resemble opposite-strand pairs. To estimate the distribution for true positives, we used those gene pairs that were strongly predicted to be on the same operon ($P(Operon|AllFeatures) > 0.95$). These genes comprise a set of high-quality predictions that have very low intergenic separations and/or conserved gene order in distantly-related species, and display high specificity when compared to known operons in *E. coli* and *B. subtilis* (see Figure 5A). For further information about this accuracy estimation procedure, see Methods.

The microarray-based estimates of accuracy are consistent with the accuracy expected from the predicted probabilities, and, in *E. coli* and *B. subtilis*, with the observed accuracy on known operons (Table 2). We observe good agreement for the larger data sets (*E. coli*, *B. subtilis*, and *Chlamydia trachomatis*), while in *Helicobacter pylori* and *Halobacterium NRC-1* there is insufficient data for reliable estimates (not shown). Although overall accuracy in *Synechocystis PCC 6803* according to the microarrays is 72% ± 5%, consistent with the method's internal estimate of 73%, this reflects the combination of a high false positive rate and a low false negative rate, due to an overly high *a priori* estimate of $P(Operon|Same)$. The unusual operon structure observed in *Synechocystis* is discussed in a later section.

Accuracy in other genomes

To test the predictions for 124 genomes, where neither databases of known transcripts nor microarray data are generally available, we used the $P(Operon|AllFeatures)$ values themselves as an internal estimate of prediction accuracy. Several lines of evidence suggest that these internal estimates may be a good indicator of performance. First, in all six species, the average microarray similarity (r) rises sharply as $P(Operon|AllFeatures)$ approaches one, and falls to nearly zero as $P(Operon|AllFeatures)$ approaches zero (Figure 5A). Second, unsupervised estimates of $P(Operon|AllFeatures)$ agree with the accuracy of predictions for known operons in *E. coli* and *B. subtilis* (Figure 5B). Finally, as shown in the previous section, predicted accuracies are in quantitative agreement with estimates from gene expression data.

We calculated the estimated accuracy of the predictions in 124 genomes from the average over all pairs of each prediction's confidence, which equals $P(Operon|AllFeatures)$ for predicted operon pairs (those with $P(Operon|AllFeatures) > 0.5$) and $1 - P(Operon|AllFeatures)$ for predicted not-operon pairs. These predicted accuracies range from 71% to 96%, with half of the genomes lying between 82% and 87%. Accuracy is independently correlated with the excess conservation of same-strand pairs and with the strength of the relationship between close spacing and conservation (Spearman $r = 0.47$ and 0.64, respectively; both $p < 10^{-7}$). Accuracy is below 75% in three genomes which have unusually weak relationships between conservation and close spacing: *Methanocaldococcus jannaschii* (formerly *Methanococcus*), *Synechocystis* (discussed below), and *Desulfovibrio vulgaris*, which improves to 79% when recently sequenced relatives are added (not shown). The only other genome with such low accuracy is *Rickettsia prowazekii*, probably because of large numbers of pseudogenes and "split" genes [19]. Overall, we predict that the accuracy of the method is $\geq 82\%$ for most genomes.

Operon structure across 124 genomes

Having validated our predictions in a number of genomes, we investigated whether these predictions could highlight biological differences among genomes when applied to a large set of diverse prokaryotes. We first turned to the genome-specific "distance models", which are the estimates of log-likelihood ratios for operon and not-operon pairs given the intergenic distance between them ($ln(P(Distance|Operon)/P(Distance|NotOperon))$). Most genome-specific distance models have the shape expected from *E. coli* and *B. subtilis*, but *E. coli* has particularly extreme values at very short and very high separations (Figure 6). *E. coli* may have an unusually strong correlation between intergenic distance and conserved proximity, or gene starts in other genomes may simply be less accurate (e.g., [9]).

These variations in distance models support our motivation for developing an unsupervised method. To determine whether the observed differences among species reflect actual biological variation, or are simply an artifact of our method, we examined two genomes with significant differences to the *E. coli* model for which we also had gene expression data: *Halobacterium* and *H. pylori.*

Distance models vary

As shown in Figure 7, in *E. coli*, microarray similarity decays gradually with increasing distance, but both *Halobacterium* and *H. pylori* show sharp and significant drop-offs – *Halobacterium* around 20 bp and *H. pylori* around 50 bp – as predicted by the genome-specific distance models. These differences in the distance models arise from statistically significant differences in how likely these pairs at intermediate distances are to be conserved in a distant genome (Supplementary Table 2).

For both genomes, predictions made using the genome-specific distance model show significantly better agreement with microarray data than predictions from a model trained on known *E. coli* operons (the method of [9]). In *Halobacterium*, the Spearman correlation of binary distance-only predictions with microarray similarity (Pearson r) is 0.210 for the genome-specific model vs. 0.127 for the *E.coli* distance model ($p = 0.04$, two-sided t-test of correlation of rank(r) vs. difference in predictions). The corresponding test in *H. pylori* gives 0.328 vs. 0.307 ($p = 0.008$). For the four other genomes, the two levels of agreement are almost identical (not shown; all $p > 0.05$). This latter result explains why a previous study focusing only on *E. coli* and *B. subtilis* reached the conclusion that distance models can be applied across species [9]; however, our results suggest that this is not true in general.

Pseudogenes in ancestral operons

The correlation between intergenic distance and conserved proximity might be weakened in some genomes by the disruption of genes within ancestral operons. For example, *Bacillus anthracis str. Ames* has an unusual distance model, while its relative *B. subtilis* has a typical model (Figure 6). *B. anthracis* has 12 apparent pseudogenes (BLASTn hits to an annotated ORF of over 200 bases in length) within operons conserved in a distant genome, whereas *B. subtilis* has none. We examined two of these pseudogenes, and found that those open reading frames were also disrupted in another sequenced strain, so these pseudogenes are unlikely to be sequencing errors. Over all same-strand pairs in *B. anthracis*, we found that 166 were separated by candidate pseudogenes that were syntenic in *B. cereus* (a close relative), so that pseudogenes may be a sufficient explanation for the unusual distance model of *B. anthracis*..

Operons in the ϵ-Proteobacteria

It has been suggested that *Helicobacter pylori* and its relative *Campylobacter jejuni* have few operons [20, 21]. However, we observe a clear excess of same-strand pairs, which indicates organization of genes into operons [15]. Indeed, from the number of same-strand pairs, we estimate that most such pairs in these genomes are in operons – 71% in *H. pylori* and 72% in *C. jejuni*, higher rates than observed for *E. coli* or *B. subtilis*. In addition, 20.5% of these same-strand gene pairs in *H. pylori* are conserved within 5 kb in *C.jejuni*, versus only 3.4% of opposite-strand pairs ($p < 10^{-13}$, χ^2 test). These conserved pairs are separated by smaller distances than other pairs in both genomes (not shown). Finally, and most significantly, microarray data for *H. pylori* indicates that predicted operon pairs have much greater similarity in expression profiles than do predicted not-operon pairs (Figure 4), and this is largely due to the distance model (Table 1). Thus, both comparative genomics and microarray data confirm the existence of many operons in these genomes.

Unusual operons in Synechocystis

From the number of same-strand pairs, we estimate that 48% of same-strand pairs in *Synechocystis* are in operons. The microarray-based estimate, however, is significantly lower and suggests that only 34%±6% of same-strand pairs are co-transcribed in operons (Table 2). Furthermore, our results and those of others suggest that many conserved operons in *Synechocystis* have large distances between genes (see [9, 10] and Figure 6). We investigated a number of possible reasons for these discrepancies. First, it has been suggested that the gene models may be inaccurate because of the absence of TTG initiation codons [9]. To rule out this explanation, we analyzed alternative gene models from CyanoBase (http://www.kazusa.or.jp/cyano/) or produced by CRITICA [22] as well as the standard set from NCBI. Both alternative sets of gene models included TTG start codons and produced the same anomalous distance model (not shown). Thus, the unusual distribution of intergenic distances for genes within operons in *Synechocystis* is not an artifact and reflects a biological difference in the structure of this genome. Second, we ruled out strong strand bias or unusual numbers of pseudogenes, either of which might affect our method for estimating $P(Operon|Same)$. Thus, assuming that the microarray-based estimates are more accurate than the sequence-based estimates of the total number of operons, it is a mystery why genes that are not co-transcribed would tend to occur on the same strand of DNA.

Discussion

Interpreting the wealth of microbial sequence data requires unsupervised methods for statistical inference and careful validation against experiment across as many phylogenetically diverse species as possible. We have demonstrated accurate unsupervised prediction of operons by combining comparative genomics and genome-specific distance models. Our method relies on the assumption, first introduced by [7] and which we have validated against known operons and against microarray data, that not-operon pairs resemble opposite-strand pairs with respect to conservation and functional similarity.

We used microarray data to estimate the accuracy of our operon predictions and to show that the unsupervised predictions are effective in six phylogenetically diverse prokaryotes, including the archaeon *Halobacterium NRC-1*, a Gram-positive bacterium (*B. subtilis*) with strong coding strand bias, a member (*Helicobacter pylori*) of the ϵ-proteobacteria, which have been described as having few operons [20, 21], the cyanobacterium *Synechocystis PCC 6803*, which has unusual operon structure [9, 10], and the intracellular parasite *Chlamydia trachomatis*. Furthermore, in *E. coli* and *B. subtilis*, unsupervised predictions are about as accurate as supervised predictions that are optimized using known operon structure. Because the predictions for other genomes were not validated against known operons, it is conceivable that the method is predicting some other kind of functional relationship between adjacent genes, rather than operons. However, most of the power of this method to predict coexpression comes from the genome-specific distance models, and the extent of agreement with both microarrays and known operons is quantitatively consistent with the method's internal estimate of its accuracy, so we argue that the method must be predicting pairs of genes that are co-transcribed.

It has been proposed that intergenic distances between genes in operons are similar in all prokaryotes. Moreover, it has been suggested that the distance distribution from *E. coli* can be used to predict operons in unrelated prokaryotes and to estimate the total number of TUs in their genomes [9]. However, we found that many genome-specific distance models are different from *E. coli.* Using comparative genomics and gene expression data, we confirmed that genes in operons in both *Halobacterium* and *H. pylori* are closer together than genes in *E. coli* operons, and that our genome-specific approach improved prediction accuracy in these genomes. In contrast, operons in *B. anthracis* appear to be widely spaced due to large numbers of pseudogenes within ancestral operons. We do not know whether such operons that have been disrupted by pseudogenes are still functional. In *Synechocystis*, the unusually wide spacing within conserved operons [9, 10] seems not to be due to errors in gene start predictions [9] or pseudogenes, and might be related to the apparent surplus of same-strand not-operon gene pairs.

We further improved our predictions by combining genome-specific distance models with comparative features (conserved proximity) and a functional feature (matching COG function codes). We also improved the accuracy of comparative operon prediction by handling distantly and closely related species differently. As more genomes are sequenced, these comparative features should become more powerful. We considered using patterns of gene co-occurrence ("phylogenetic profiles," [23]), but this did not provide statistically significant additional information (not shown). The similarity of textual annotations has been used to select a genome-specific distance threshold, but this threshold and the underlying feature were used to aid functional annotation, and their effectiveness for operon prediction was not directly tested [24]. This feature and other other precise measures of functional similarity (e.g., from metabolism [2]) might improve accuracy. Finally, we suspect that transcription intiation or rho-independent termination sites that are conserved across species might aid prediction.

These operon predictions will be useful for analyses of gene regulation, for example, to focus the search for new transcription factor binding motifs to those upstream regions which are likely to contain promoters [25, 26]. They should also aid in analyzing microarray data – averaging expression profiles over several genes in a predicted operon can reduce noise and improve the effectiveness of clustering algorithms (R.P. Koche and E.J. Alm, unpublished observations). As both conserved gene order [11, 27] and distances between genes [24] have shown promise in the assignment of gene function, our results may also aid the annotation of uncharacterized genes. Predictions for over 120 genomes, as well as source code, are freely available from the VIMSS website (http://vimss.org/operons).

Methods

Data sources

Sequences. We downloaded the complete annotated genomes of 124 prokaryotes from NCBI complete microbial genomes (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html), TIGR (http://www.tigr.org), and DOE's JGI (http://www.jgi.doe.gov/), and excluded plasmids and non-protein-coding genes from our analyses.

Microarrays. We obtained data for *E. coli*, *B. subtilis*, and *H. pylori* from the Stanford Microarray Database (74, 78, and 31 arrays, respectively, from http://genome-www.stanford.edu/microarray, [28]), for *Synechocystis* from the Kyoto Encyclopedia of Genes and Genomes (49 arrays from http://www.genome.ad.jp/kegg/expression/), for *C. trachomatis* from T. Nicholson and R. Stephens (12 experiments times 2-3 replicates), and for *Halobacterium* from R. Bonneau and N. Baliga (44 arrays).

Features

The comparative or "gene neighbor" scores measure how often two genes are near each other across many genomes [11, 27]. We used putative orthologs from bidirectional best BLASTp hits with 75% coverage both ways, and asked how often the genes have orthologs that are within 5 kb (this cutoff was determined empirically). Previous workers threw out closely related genomes [27, 10] or reduced sensitivity when they were present [7], as these genomes show conserved pairs of unrelated genes because of insufficient evolutionary time to shuffle them apart. Instead, we computed separate scores for distantly and closely related genomes. To distinguish closely related genomes, we clustered all genomes by the extent of conserved gene order, placing in the same cluster any pair of genomes for which 5% or more of opposite-strand pairs were conserved within 5 kb. To get useful information from these closely related genomes, we introduced a penalty if both orthologs exist but are not within 5 kb. Specifically, for each same-strand pair, this "within-cluster" score was the sum, across the closely related genomes that also contained orthologs for both genes, of a positive term if the pair was within 5 kb, or a negative term if the pair was not within 5 kb. We used "pseudo-log-likelihood" scoring, so that the magnitudes of these two terms were the logarithms of the proportions of all opposing-strand pairs that were conserved within 5 kb or not, respectively. We computed a second feature from the distantly related genomes by summing, across clusters, the maximum term within each cluster (excluding the cluster containing the genome itself, and using only positive terms). We also computed the sum of terms, including penalties, over all genomes irregardless of clustering, giving a third comparative feature.

To determine COG function codes, we assigned genes to COGs [13] via reverse position-specific BLAST [29] against CDD [30], or by using COG membership from NCBI. Pairs of genes were assigned to three categories: matching, not matching, or one or both genes are uncharacterized (function codes “R” or “S,” or not in COG).

We used similarity of CAI, a measure of synonymous codon usage [14], instead of a related feature proposed by [5] because similarity of CAI shows better agreement with operons in *E. coli* and *B. subtilis* (data not shown). The reference set for CAI in each genome was identified by choosing the most 100 biased genes with at least 300 amino acids among a set of 500 COGs which show bias across many genomes. Our similarity measure was defined as $s_{CAI} = ((rank(CAI_1) - mean(rank)) * (rank(CAI_2) - mean(rank)) - (rank(CAI_1) - rank(CAI_2))^2$. Both terms showed modest but statistically significant agreement with operons (not shown).

Estimating likelihood ratios

We begin with values for a feature d, such as the distance between two genes, for each pair. The values are split into two sets, such as the same-strand pairs with high and low comparative/functional likelihood ratios. When inferring the genome-specific distance models, we also have error rates in the training data, $P(NotOperon|High)$ and $P(Operon|Low)$. We wish to estimate the likelihood ratio $P(d|Operon)/P(d|NotOperon)$, which corresponds to the probability

$$p_d \equiv \frac{P(d|Operon)}{P(d|Operon) + P(d|NotOperon)} \qquad \text{(Eq. 7)}$$

which can be thought of as the probability of a pair separated by distance d being an operon pair if operons and not-operons were equally likely. The likelihood ratio is equal to $p_d/(1 - p_d)$.

We first grouped the values into overlapping bins of 100–200 items and estimated the likelihood ratio within each bin. We obtained a likelihood ratio for each specific value by interpolating and then smoothing (via local regression). We used ranks rather than raw values.

To estimate likelihood ratios within each bin, we used a maximum likelihood approach. We solved numerically for the p_d that maximized the joint probability of p_d and the data – the counts of high and low pairs within bin d (n_{Hd} and n_{Ld}, respectively) – given a prior distribution $\pi(p_d)$:

$$P(n_{Hd}, n_{Ld}, p_d) = P(n_{Hd}, n_{Ld}|p_d) \cdot \pi(p_d) \qquad \text{(Eq. 8)}$$

$$P(n_{Hd}, n_{Ld}|p_d) \propto P(High|d)^{n_{Hd}} \cdot P(Low|d)^{n_{Ld}} \qquad \text{(Eq. 9)}$$

$$\pi(p_d) \equiv p_d \cdot (1 - p_d) \qquad \text{(Eq. 10)}$$

where $P(High|d)$ is an unknown probability, not the observed proportion, and is related to p_d by

$$\frac{P(High|d)}{P(Low|d)} = \frac{P(d|High)}{P(d|Low)} \cdot \frac{P(High)}{P(Low)}$$

$$= \frac{p_d \cdot P(Operon|High) + (1 - p_d) \cdot P(NotOperon|High)}{p_d \cdot P(Operon|Low) + (1 - p_d) \cdot P(NotOperon|Low)} \cdot \frac{P(High)}{P(Low)} \qquad \text{(Eq. 11)}$$

where $P(High)$ and $P(Low)$ can be estimated from the observed proportions over the entire data set.

Because of our choice of prior, our maximum likelihood estimator is a generalization of pseudocounts, or adding counts to the observations to avoid overfitting. In the absence of errors, the maximum likelihood estimate with this prior is given by adding a total of two pseudocounts to each bin [31].

Combining the comparative log-likelihood ratios

To combine the comparative/functional log-likelihood ratios – from three raw "gene neighbor" scores and the COG similarity score – into a combined log likelihood $ln(P(Values|Same) \ / \ P(Values|NotSame))$, we did not use the naive Bayesian method. These variables are not conditionally independent, so multiplying likelihood ratios or, equivalently, summing log-likelihood ratios, would overstate the confidence of predictions. Instead, we found the best-fitting linear combination of log-likelihood ratios using logistic regression (*glm* in the R statistics package, http://www.r-project.org/). All four features contained statistically significant additional information for discriminating same-strand from opposing-strand pairs in the majority of genomes (generalized ANOVA, data not shown).

Prior estimate of P(Operon|Same)

The proportion of same-strand pairs that are in operons can be estimated by observing the proportion of adjacent pairs of genes that are same-strand pairs [7, 15]. If independent transcripts are equally likely to occur on the same or different strands, then $1 - P(Operon|Same) \cdot P(Same) = 2 \cdot P(NotSame)$, which gives $P(Operon|Same) = 2 - 1/P(Same)$. This method agrees with other estimates for *E. coli*, but is not accurate for genomes with an excess of genes on the leading strand of DNA replication [15].

To account for these strand biased genomes, such as *B. subtilis*, we use our rather surprising observation that adjacent pairs of genes on either the leading or lagging strand of DNA are equally likely to be co-transcribed in an operon (M.N.P., E.J.A., A.P.A., submitted). Based on this observation, we assume that $P(Operon|Leading_1, Leading_2) = P(Operon|Lagging_1, Lagging_2)$, where "1" refers to a first gene that might be in the same operon or on the same strand as the next gene downstream ("2"). From this we derive:

$$P(Operon|Same) = \frac{P(Operon|Lagging_1)}{P(Lagging_2|Lagging_1)} = \frac{P(Operon|Leading_1)}{P(Leading_2|Leading_1)}$$

$$a \cdot P(Operon|Lagging_1)^2 + b \cdot P(Operon|Lagging_1) + c = 0$$

$$a = \frac{P(Leading_2|Leading_1)}{P(Lagging_2|Lagging_1)}$$

$$b = -2 \cdot P(Leading_2|Leading_1)$$

$$c = P(Leading_2|Leading_1) + P(Lagging_2|Lagging_1) - 1 \qquad \text{(Eq. 12)}$$

We also tried a simpler approach based on the plausible but unsupported hypothesis that TUs assort to the leading and lagging strands independent of their length. Compared to this "strand-naive" approach, the "strand-wise" formula gave better prediction accuracy on known operons, better agreement with microarray data, and better agreement with an independent estimate of $P(Operon|Same)$ based on *E. coli* distance models [9] (Supplementary Table 3).

Estimating accuracy from microarray data

Given the "true positive" and "true negative" pairs described in the Results, as well as the predicted operon and not-operon pairs, we modeled these four distributions of microarray similarites with a Gaussian kernel. We then used linear regression on the densities to estimate the proportion of true operon pairs in each set of predictions. We also corrected for the expression levels of the different sets of genes – the high-confidence predictions are more highly expressed and have higher microarray similarity than other operon pairs, probably due to reduced noise (not shown). Specifically, we divided each distribution into four quartiles by their expression level and reweighted these fractions before the regression. To put confidence intervals around these estimates of accuracy, we used a jackknife approach: we reran the estimation procedure with individual conditions (manually identified groups of similar experiments, such as "heat shock") removed from the data set. We multiplied the variance of these leave-1-out estimates by $(m-1) \cdot m/(m+1)$, where m is the number of conditions, to account for the fact that the jackknife estimates are correlated as they mostly use the same data, and used a t test to give 95% confidence intervals.

Acknowledgments

We thank Nitin Baliga, Richard Bonneau, Tracy Nicholson, and Richard Stephens for providing unpublished microarray data, Inna Dubchak for insightful discussions and sharing resources, and the rest of the Arkin lab for comments on the manuscript. This work was supported by a grant from the DOE Genomics:GTL program (DE-AC03-76SF00098).

References

[1] Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., & Collado-Vides, J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* **97**:6652–7.

[2] Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., & Kasif, S. (2002) Computational identification of operons in microbial genomes. *Genome Res.* **12**:1221–30.

[3] Sabatti, C., Rohlin, L., Oh, M.K., & Liao, J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* **30**:2886–93.

[4] Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F., & Craven, M. (2003a) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics* **19 Suppl. 1**:I34–I43.

[5] Bockhorst, J., Craven, M., Page, D., Shavlik, J., & Glasner, J. (2003b) A Bayesian network approach to operon prediction. *Bioinformatics* **19**:1227–35.

[6] de Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., & Miyano, S. (2004) Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomputing* 2004.

[7] Ermolaeva, M.D., White, O., & Salzberg, S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29**:1216–21.

[8] Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., Koonin, E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11:**356–72**.

[9] Moreno-Hagelsieb, G. & Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18 Suppl. 1**:S329–36.

[10] Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., & Koonin, E.V. (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res.* **30**:4264–71.

[11] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., & Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**:2896–901.

[12] Korbel, J.O., Jensen, L.J., von Mering, C., & Bork, P. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22**:911–7.

[13] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., & Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**:22–8.

[14] Sharp, P.M. & Li, W.-H. (1987) Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. *Nucl. Acids Res.* **15**:1281–1295.

[15] Cherry, J.L. (2003) Genome size and operon content. *J. Theor. Biol.* **221**:401–10.

[16] Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., & Gama-Castro, S. (2002) The EcoCyc Database. *Nucleic Acids Res.* **30**:56-8.

[17] Itoh, T., Takemoto, K., Mori, H., & Gojobori, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16**:332–46.

[18] DeLong, E.R, DeLong, D.M., & Clarke-Pearson, D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**:837–45.

[19] Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P.E., Barbe, V., Samson, D., Roux, V., Cossart, P., Weissenbach, J., Claverie, J.M., & Raoult, D. (2001) Mechanisms of evolution in Rickettsia conorii and R. prowazekii. Science **293**:2093-8.

[20] Thompson, L.J., Merrell, D.S., Neilan, B.A., Mitchell, H., Lee, A., & Falkow, S. (2003) Gene expression profiling of Helicobacter pylori reveals a growth-phase-dependent switch in virulence gene expression. *Infect Immun.* **71**:2643–55.

[21] Parkhill, J., Wren B,W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S. & others. (2000) The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature* **403**:665–8.

[22] Badger, J. H. & Olsen, G. J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**: 512-524.

[23] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., & Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**:4285-8.

[24] Strong, M., Mallick, P., Pellegrini, M, Thompson, M.J., & Eisenberg, D. (2003) Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biol.* **4**:R59.

[25] McGuire, A.M., Hughes, J.D., & Church, G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10**:744–57.

[26] McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., & Lawrence, C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29**:774–82.

[27] Huynen M., Snel, B., Lathe 3rd, W., & Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**:1204–10.

[28] Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., & others. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* **31**:94–6.

[29] Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., & others. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**:2994–3005.

[30] Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., & others. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31**:383–7.

[31] Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler,D. (1996) Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology. *Comput. Appl. Biosci.* 12(4): 327–45.

Web References

http://bacillus.genome.ad.jp/bsorf.html – BSORF (*B. subtilis* Northerns)

http://www.kazusa.or.jp/cyano/ – CyanoBase (*Synechocystis* gene models)

http://vimss.org/operons – VIMSS comparative genomics browser

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html – NCBI's complete microbial genomes

http://www.tigr.org – The Institute for Genomic Research

http://www.jgi.doe.gov/ – DOE Joint Genome Institute

http://genome-www.stanford.edu/microarray – Stanford Microarray Database

http://www.genome.ad.jp/kegg/expression/ – KEGG Microarray Database

http://www.r-project.org/ – the R statistics package

Genome	Distance	Comparative	All features
E. coli K12	0.406	0.401	0.494
B. subtilis	0.420	0.335	0.461
Helicobacter pylori	0.275	0.231	0.343
Chlamydia trachomatis	0.260	0.167	0.303
Synechocystis	0.159	0.222	0.268
Halobacterium	0.198	0.159	0.215

Table 1: The majority of the agreement between predictions and microarrays is due to the genome-specific distance models. For each genome we show the Spearman correlation between the microarray similarity (the Pearson correlation of the normalized log-ratios for two adjacent genes) and the predicted probability that the two genes are in the same operon ($P(Operon)$) using just intergenic distance, using just the comparative/functional features, or using all features.

Attribute	*E. coli*	*B. subtilis*	*C. trachomatis*	*Synechocystis*
% accuracy of predicted operon pairs				
from sequence	89.4	84.2	86.2	76.5
from microarrays	88.6 ± 1.3	76.7 ± 3.5	94.8 ± 5.7	58.2 ± 10.9
from known operons	85.4	77.0		
% accuracy of predicted not-operon pairs				
from sequence	85.4	83.5	82.4	70.3
from microarrays	85.7 ± 2.3	80.6 ± 1.1	91.6 ± 14.0	86.5 ± 7.1
from known operons	83.7	88.0		
a priori % in operons ($P(Operon\|Same)$)				
from sequence	57.0	51.7	59.7	48.5
from microarrays	56.0 ± 1.6	49.7 ± 2.1	59.9 ± 9.7	32.1 ± 5.6

Table 2: Estimates of prediction accuracy from the method itself agree with estimates from microarrays or from known operons. Ranges are 95% confidence intervals.

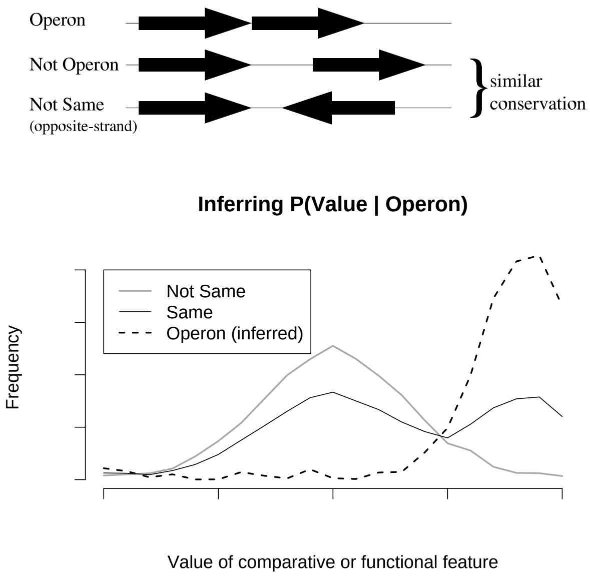

Figure 1: Building a model of operons without training data. Above, we show the three types of pairs of adjacent genes, and the key assumption. Below, we use this assumption to infer $P(Value|Operon)$, the distribution of a comparative or functional feature for operon pairs, from the observed distributions $P(Value|Same)$ and $P(Value|NotSame)$. The graph is purely schematic.

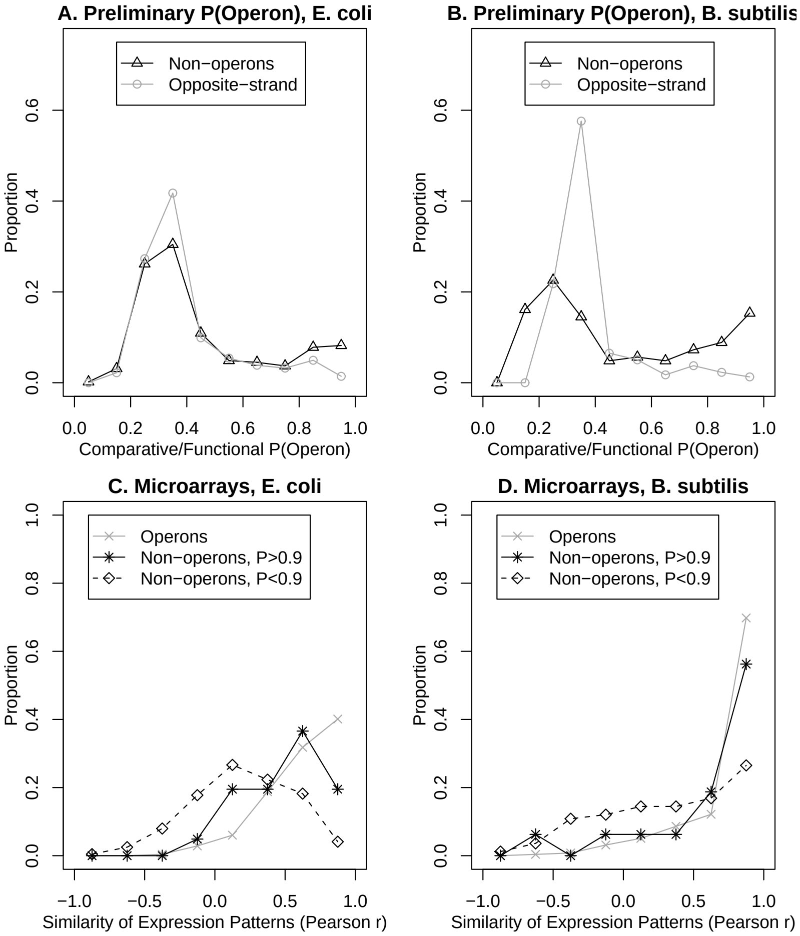

Figure 2: Conservation and co-expression of "known" not-operon pairs. (A & B) The distribution of preliminary estimates of $P(Operon)$, using only the comparative and functional features, for opposite-strand pairs and "known" not-operon pairs in (A) *E. coli K12* and (B) *B. subtilis*. (C & D) Histograms of microarray similarity (Pearson correlation, x axis) for known operon pairs, for known not-operon pairs strongly predicted to be in an operon by the comparative/function features ($P(Operon) > 0.9$), and for other known not-operon pairs, in (C) *E. coli* and (D) *B. subtilis*.

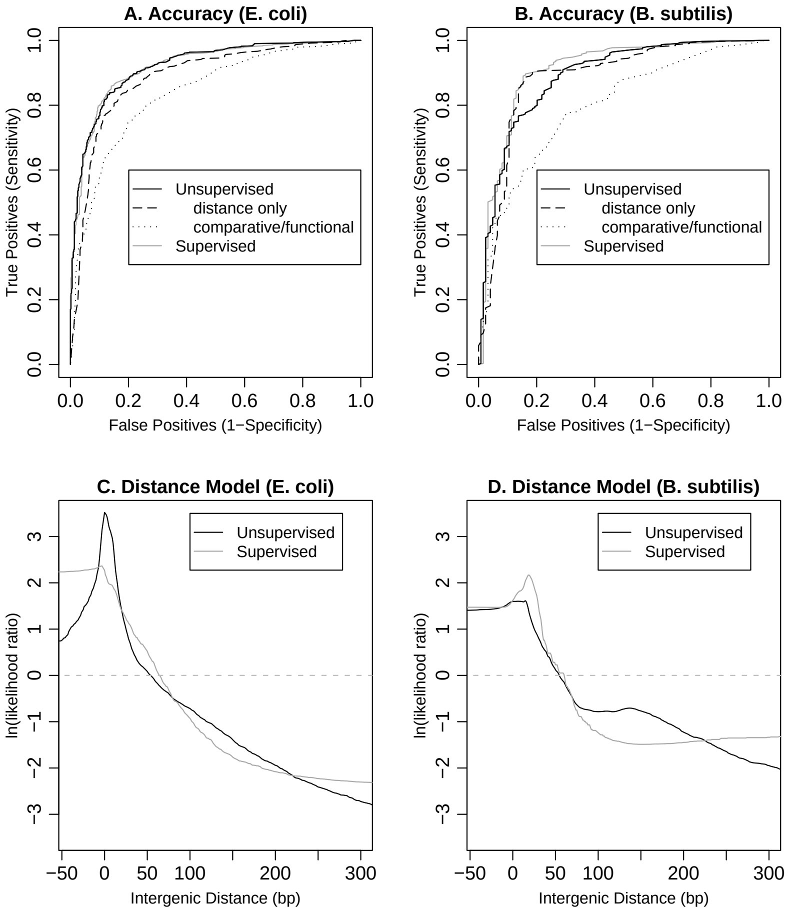

Figure 3: Unsupervised predictions are accurate and similar to supervised predictions in *E. coli K12* **(left) and** *B. subtilis* **(right).** (A & B) Accuracy on known operon and not-operon pairs as the prediction threshold varies, also known as the Receiver Operating Characteristic (ROC) curve, for (A) *E. coli* and (B) *B. subtilis.* We show the ROC curves for unsupervised predictions using all features, for unsupervised predictions using only distance or only the comparative/functional features, and also for supervised predictions (using all features and 100-fold cross-validation). (C & D) Distance models, with intergenic distance in base pairs on the x axis and log likelihood ratios ($ln(P(Operon|Distance)/P(NotOperon|Distance))$) on the y axis, for (C) *E. coli* and (D) *B. subtilis.* A log likelihood ratio of zero (dashed line) indicates pairs that are equally likely to be in an operon or not.

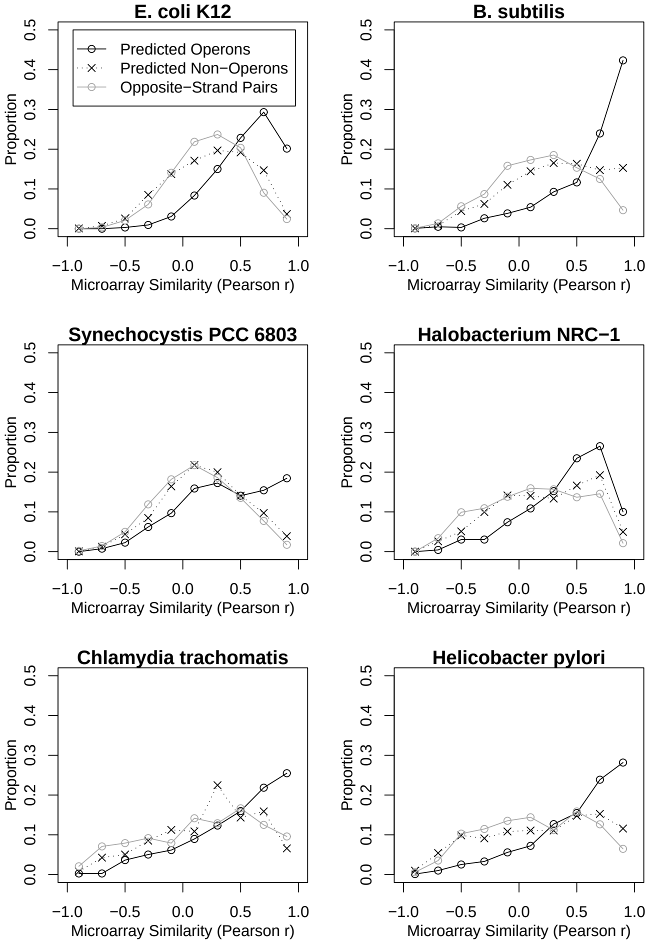

Figure 4: Unsupervised predictions agree with microarray data from six species. For each species, we show histograms of microarray similarity (Pearson r, x axes) for predicted operon pairs, for predicted not-operon pairs on the same strand, and for opposite-strand pairs. Predicted not-operon pairs show a similar distribution as the opposite-strand pairs and are significantly less correlated than predicted operon pairs ($p < 10^{-7}$ for all genomes, Kolmogorov-Smirnov test, D-statistic=0.22–0.37).

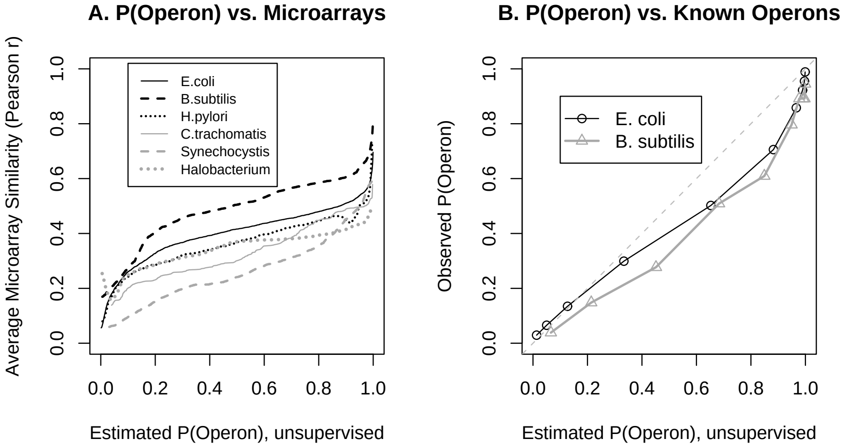

Figure 5: $P(Operon|AllFeatures)$ **is consistent with known operons and with microarray data.** (A) The smoothed average of the similarity of gene expression profiles (Pearson r) as a function of $P(Operon|AllFeatures)$, computed by local regression (loess) on r vs. $rank(P(Operon|AllFeatures))$. (B) Accuracy of unsupervised estimates of $P(Operon|AllFeatures)$ for known operons from *E. coli* and *B. subtilis*. For both genomes, we grouped the known operon or not-operon pairs together into 10 bins of equal size based on $P(Operon|AllFeatures)$. For each bin, we show $P(Operon|AllFeatures)$ versus the actual proportion of operon pairs, after correcting for the greater number of known operon pairs in the test set. The straight line shows ideal performance ($x = y$).

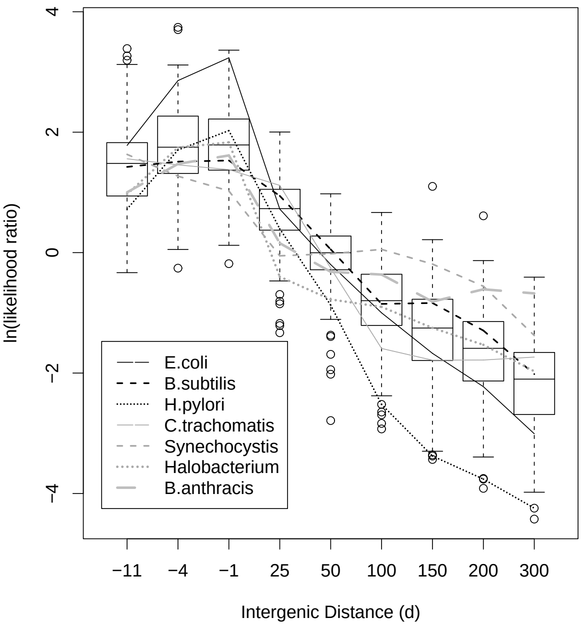

Figure 6: Unsupervised distance models across 124 genomes. We show boxplots, across 124 genomes, of the genome-specific log-likelihoods $ln(P(Distance|Operon)/P(Distance|NotOperon))$ at the indicated distances. Where the log likelihood is zero, operon and not-operon pairs are predicted to be equally likely to have that distance. The boxes show quartiles and medians, whiskers extend up to 1.5x the interquartile range from the box, and dots show outlying genomes. The non-linear x-axis highlights the sharp peak around the common separations of -1 and -4. Distance models for a few specific genomes are shown with lines. Although most genomes follow the same trend of more operons at lower separations, significant differences are seen in the shape and magnitude of their distance models.

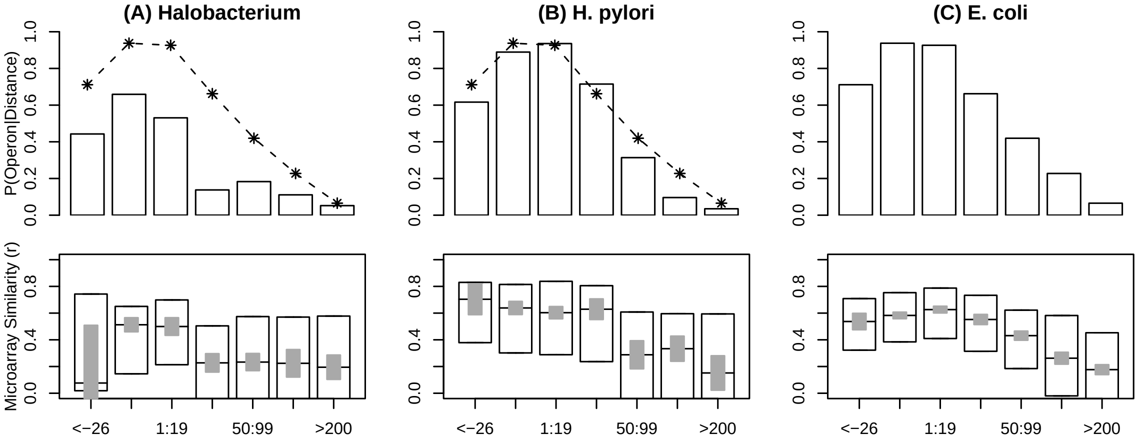

Figure 7: Microarrays confirm genome-specific differences in distance models. The panels show the genome-specific distance models (top) and boxplots of microarray similarity (bottom) for same-strand pairs separated by various intergenic distances in (A) *Halobacterium NRC-1*, (B) *H. pylori*, and (C) *E. coli.* The ranges of distances were selected to make the number of pairs within each range more uniform. In the top panels, the bars show the average of the genome-specific distance model $P(Operon|Distance)$ within each range, and the stars with lines show the corresponding value from *E. coli.* In the bottom panels, the boxes show quartiles and medians of microarray similarity (Pearson r) within each range, and the grey bars show 90% confidence intervals around the median. If two bars do not overlap then the medians are significantly different ($p < 0.05$).

Supplementary Table 1: Agreement of our unsupervised and supervised predictions with experimentally identified operon and not-operon pairs in *E. coli* and *B. subtilis*. AOC is the area under the operating curve (e.g., Figure 3A), or the probability that an operon pair will have a better score than a not-operon pair if both pairs are chosen at random. Default sensitivity (fraction of known operon pairs which are correctly predicted) and specificity (fraction of known not-operon pairs which are correctly predicted) are computed with a threshold of predicted $p > 0.5$, and maximum accuracy is the maximum over all possible thresholds of the average of sensitivity and specificity. The unsupervised microarray-based predictions, which are shown only in this table, use a logistic regression of the microarray data (rank of Pearson r, total intensity, and total absolute change of log-levels for the pair, with pairwise interactions) versus the usual unsupervised predictions (thresholded at 0.5).

For comparison, we show results from our supervised predictions, from Salgado *et al.* 2000 for *E. coli* (using distance and Monica Riley's functional classification, or just distance), from Sabatti *et al.* 2002 for *E. coli* (using correlation in microarray data and/or distance as features, on a somewhat different training set), from Bockhorst *et al.* 2003b for *E. coli* (distance-only or distance plus microarrays and further sequence-based features), from Moreno-Hagelsieb and Collado-Vides 2002 for *B. subtilis* (using a distance model trained in *E. coli*), and from De Hoon *et al.* 2004 for *B. subtilis* (using distance and/or microarray correlation, and a much larger unpublished training set). We do not show the results of Bockhorst *et al.* 2003a because they report accuracy for predicting transcripts, not individual pairs of genes.

Measure	**AOC**	**Max. Acc.**	**Def. Sens.**	**Def. Spec.**
E. coli				
Unsupervised (Sequence-only)	0.920	0.852	0.883	0.799
Distance-only	0.886	0.829	0.794	0.857
Unsupervised with microarrays	0.925	0.863	0.890	0.817
Microarray-only	0.820	0.750	0.834	0.660
Supervised (Sequence-only)	0.919	0.859	0.865	0.850
Salgado *et al.* 2000	–	0.87	–	–
Distance-only	–	0.82	–	–
Sabatti *et al.* 2002	–	0.88	0.88	0.88
Distance-only	–	0.83	0.84	0.82
Microarray-only	–	0.76	0.82	0.70
Bockhorst *et al.* 2003b	0.929	–	0.78	0.90
Distance-only	0.915	–	–	–
B. subtilis				
Unsupervised (Sequence-only)	0.888	0.815	0.909	0.710
Distance-only	0.882	0.863	0.825	0.863
Unsupervised with microarrays	0.885	0.844	0.922	0.727
Microarray-only	0.748	0.692	0.804	0.545
Supervised (Sequence-only)	0.907	0.868	0.877	0.847
Moreno-Hagelsieb & Collado-Vides 2002	–	0.82	–	–
De Hoon *et al.* 2004	–	0.884	0.888	0.879
Distance-only	–	0.856	0.821	0.890
Microarray-only	–	0.796	0.801	0.791

Supplementary Table 2: Statistical tests of differences between *E. coli*'s distance model and those of *Halobacterium NRC-1* and *Helicobacter pylori*. To confirm differences in distance models, we tested same-strand pairs separated by 20-49 base pairs (*E. coli* vs. *Halobacterium*) or by 50-99 base pairs (*E. coli* vs. *H. pylori*). We compared how often these pairs were conserved within 5 kb in a distant genome, relative to other pairs in the same genome. We show the 90% confidence intervals of the odds ratios from the Fisher exact test. In both cases the odds ratio in *E. coli* is higher, indicating significantly greater conservation at these separations ($p < 0.05$).

		Conserved within 5 kb		
Genome	Range (bp)	In-range pairs	Other pairs	Odds Ratio
Halobacterium	20–49	12/194 (6.2%)	173/1017 (17.0%)	0.18–0.55
E. coli	20–49	127/324 (39.4%)	956/2681 (35.7%)	0.95–1.4
H. pylori	50–99	15/143 (10.5%)	314/1083 (29.0%)	0.17–0.46
E. coli	50–99	117/426 (27.5%)	966/2,579 (37.5%)	0.52–0.77

Supplementary Table 3: Comparison of "strand-wise" and "strand-naive" models for estimating P(Operon|Same). The strand-wise estimate leads to significantly more accurate unsupervised predictions in *B. subtilis.* The poor agreement between both estimates and the *E. coli* distance model-based method of Moreno-Hagelsieb and Collado-Vides (2002) probably reflects the biologically meaningful variation in the distance distributions of different genomes (Rogozin *et al.* 2002).

Issue	Measure	Strand-wise	Strand-naive	p
# Operons in *B. subtilis*	% same-strand pairs that are within operons	51.7%	41.3%	–
Accuracy on known operons in *B. subtilis*	Area under the operating curve	0.888	0.864	$< 10^{-5}$, test of DeLong *et al.* 1988
Agreement with microarray data for *B. subtilis*	Spearman correlation of $P(Operon\|AllFeatures)$ with microarray similarity r	0.461	0.433	$< 10^{-10}$, two-sided t-test of correlation between $rank(r)$ and differences in $rank(p)$
Agreement of estimated # operons with *E. coli*-based estimates	Spearman correlation, 124 genomes	0.363	0.223	0.04, correlation test of ranked differences