BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 7 2005, pages 880–888
doi:10.1093/bioinformatics/bti123

Genome analysis

Operon prediction without a training set

B. P. Westover[1], J. D. Buhler[1,*], J. L. Sonnenburg[2] and J. I. Gordon[2]

[1]Department of Computer Science and Engineering, Washington University, St. Louis, MO 63130, USA and [2]Department of Molecular Biology and Pharmacology and Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108, USA

Received on May 17, 2004; revised and accepted on October 29, 2004
Advance Access publication November 11, 2004

ABSTRACT

Motivation: Annotation of operons in a bacterial genome is an important step in determining an organism's transcriptional regulatory program. While extensive studies of operon structure have been carried out in a few species such as *Escherichia coli*, fewer resources exist to inform operon prediction in newly sequenced genomes. In particular, many extant operon finders require a large body of training examples to learn the properties of operons in the target organism. For newly sequenced genomes, such examples are generally not available; moreover, a model of operons trained on one species may not reflect the properties of other, distantly related organisms. We encountered these issues in the course of predicting operons in the genome of *Bacteroides thetaiotaomicron* (*B.theta*), a common anaerobe that is a prominent component of the normal adult human intestinal microbial community.

Results: We describe an operon predictor designed to work without extensive training data. We rely on a small set of *a priori* assumptions about the properties of the genome being annotated that permit estimation of the probability that two adjacent genes lie in a common operon. Predictions integrate several sources of information, including intergenic distance, common functional annotation and a novel formulation of conserved gene order. We validate our predictor both on the known operons of *E.coli* and on the genome of *B.theta*, using expression data to evaluate our predictions in the latter.

Availability: The software is available online at http://www.cse.wustl.edu/~jbuhler/research/operons

Contact: jbuhler@cse.wustl.edu

1 INTRODUCTION

Operons—sets of genes that are co-transcribed into a single polycistronic mRNA sequence—are a fundamental mechanism by which bacteria implement co-expression of related genes. Identifying putative operons is therefore a key step in characterizing gene regulation in newly sequenced bacteria. A number of computational methods (Salgado *et al*., 2000; Ermolaeva *et al*., 2001; Bockhorst *et al*., 2003a; De Hoon *et al*., 2004) exist for operon prediction. The majority of these methods identify operons using a model inferred from a training set of known operons in the organism of interest.

The need for training data limits the utility of most operon predictors to organisms with many experimentally characterized operons, such as *Escherichia coli* (Salgado *et al*., 2000; Bockhorst *et al*., 2003a) and *Bacillus subtilis* (De Hoon *et al*., 2004). In contrast, new genomes lack a set of validated operons, so predictors that train organism-specific models cannot easily be applied to them. In such cases, one may hope that training sets of operons from one species might be useful in another. For example, Moreno-Hagelsieb and Collado-Vides (2002) found that an operon predictor based on intergenic distances in *E.coli* worked equally well when applied to the known operons of *B.subtilis*. However, other operon finders trained in one organism have proved less portable when the target species is not closely related to that used for training. For example, Romero and Karp (2004) found that a predictor trained on *E.coli* performed relatively well for that organism (69% of known operons correctly predicted) but performed substantially less well when applied to *B.subtilis* (46% of known operons correctly predicted). Portability of an operon finder may depend on the types and amount of information used for training, particularly if no explicit effort is made to produce a predictor that is portable across genomes.

In this work, we seek to perform operon prediction in *Bacteroides thetaiotaomicron* (*B.theta*), a prominent yet relatively uncharacterized member of the human intestinal microbiota (Xu *et al*., 2003). *Bacteroides* is the predominant genus in the normal adult human distal intestine. *B.theta* was the first member of this genus to have its genome completely sequenced and assembled, so there were as yet few well-characterized operons on which to train; moreover, *B.theta* is not closely related either to the *Gammaproteobacteria* or to the *Bacillaceae*, so we were wary of applying an operon finder trained on *E.coli* or *B.subtilis*. Instead, we developed a system for operon prediction that does not demand a training set of operons for the genomes of interest. Our predictor relies on *a priori* assumptions about the properties of operons to convert several types of genomic evidence—intergenic distance, common functional annotation and conservation of gene order—into a single probabilistic prediction. Our system is designed to enable reliable operon prediction not just in *B.theta* but generally in bacteria containing few well-characterized operons.

The remainder of the paper is organized as follows. After reviewing related work on operon finding, Section 2 describes the methods used to implement our predictor. Section 3 assesses the performance of our predictor both on its intended target, *B.theta*, and on the better-characterized *E.coli* strain K12. For *E.coli*, we can assess the performance of our software against the large set of known operons in the RegulonDB database (Salgado *et al*., 2004). For *B.theta*, we validate our predictor by comparing its predictions to observations

*To whom correspondence should be addressed.

from DNA microarray-based expression studies. Finally, Section 4 concludes and identifies opportunities for future work.

1.1 Relation to previous work

The design of our software is informed by previous operon finders by Salgado *et al*. (2000); Moreno-Hagelsieb and Collado-Vides (2002); Bockhorst *et al*. (2003a); Ermolaeva *et al*. (2001) and TIGR (2004). Salgado's predictor, trained on *E.coli*, predicts operons on the basis of intergenic distance and manual functional classification of genes. The measured distributions of intergenic distance for pairs of adjacent genes known to be in operons and for pairs known *not* to be in operons are used to produce a log-likelihood ratio test for whether two adjacent genes are in a common operon. Genes with both the right intergenic distance and a common functional classification are considered most likely to be in a common operon. Once trained, this predictor correctly classified 88% of pairs of adjacent *E.coli* genes from its training set. Our predictor, in common with Salgado's, makes use of intergenic distance and common functional information, but its *a priori* approach works around both the lack of training examples and the lack of an equivalent manual functional classification for the genes of *B.theta*.

The software of Bockhorst *et al*. makes predictions based on a number of features of adjacent genes, including codon usage statistics, gene expression data, intergenic distance and regulatory features. The authors report a predictive accuracy of 78% true positives for 10% false positives in *E.coli*. A major contribution of this predictor is its rigorous procedure for combining multiple types of information using a Bayesian network. We use a simplified version of this Bayesian strategy for combining information in our predictor. Other predictors that combine information sources, such as those of Yada *et al*. (1999) and Tjaden *et al*. (2002), use a more elaborate HMM-based gene model that predicts operons as one of its features. In contrast, we assume that genes have already been annotated in the target species and focus purely on operon prediction.

The predictor of Ermolaeva *et al*. is perhaps closest in spirit to our own. It predicts operons with a high degree of confidence based on the notion that pairs of genes that occur adjacent to one another in multiple organisms are likely to be members of the same operon. The method has a specificity of 92% on known pairs of adjacent *E.coli* genes in a common operon and an estimated sensitivity of 30–50% on all gene pairs in that genome. We have adapted this predictor's probability model, including *a priori* assumptions about the distribution of operons, to provide a rigorous probabilistic basis for our use of intergenic distance and predicted gene function. However, we generalize its strategy for using conserved gene order to consider not only pairs of adjacent genes but clusters of several nearby genes.

Finally, we note that not all groups use the same definition of correct or 'true positive' detection of a known operon, so that the above quoted sensitivity and specificity values are not necessarily comparable across studies. We follow the convention of reporting as true positives those pairs of adjacent genes that are correctly identified as being in a common operon.

2 METHODS

2.1 Problem formulation

We define an operon to be a set of one or more genes, occurring contiguously in a genome, that are transcribed into a single mRNA molecule. For mathematical convenience, we consider a single gene transcribed by itself to be a special case of an operon. Although most definitions of operons include regulatory elements, our predictions focus only on determining which sets of genes are co-transcribed.

The inputs to our operon predictor are a *target* genome, with annotated gene locations and putative functional annotations for each gene, along with a set of *informant* genomes (e.g. those available from GenBank) and their respective gene locations and functional annotations. Given this information, we construct an *operon map* for the target genome that predicts, for each pair of adjacent genes in the target, whether the pair belongs to the same operon.

Because genes on opposite strands of a genome cannot belong to the same operon, we limit prediction to contiguous genes on the same strand with no intervening gene on the opposite strand. We will refer to such runs of same-stranded genes as *directons* (Salgado *et al*., 2000).

2.2 Data sources and the *a priori* model

We use three data sources to predict whether a pair of adjacent genes in a directon belong to the same operon: distance between their open reading frames, similarity of their functional annotations and interspecies conservation of gene clusters containing the two genes. All of these sources of information have proven previously to be useful predictors of operon structure in *E.coli*.

Let S be the event that two adjacent genes in a genome occur on the same strand, and let O be the event that the genes are in the same operon. Moreover, let X be a random variable observed for the two genes (e.g. their intergenic distance). We wish to estimate $\Pr(O \mid S, X = x)$, the probability that two same-stranded genes belong to the same operon given an observed value x for X. Estimating this probability requires either training examples, in the form of gene pairs labeled as to whether or not they are in the same operon, or *a priori* assumptions.

We apply two key assumptions formulated by Ermolaeva *et al*. (2001). The first assumption regards the prior $\Pr(O \mid S)$ that two same-stranded genes are in a common operon. We assume that the strandedness of each operon is chosen uniformly at random, that is, that two adjacent genes not in the same operon are equally likely to be on the same strand or on opposite strands. Under this assumption, the probability that a directon contains exactly n operons is $(1/2)^n$, and the expected number of operons per directon (including single-gene 'operons') is $\sum_{i=1}^{\infty} i(1/2)^i = 2$. Hence, if there are on average two operons per directon, then there is on average one operon boundary per directon, and we estimate $\Pr(O \mid S)$ by $1 -$ (# of directons/# of S pairs).

The second assumption, which we apply to some but not all data sources, permits computation of the posterior $\Pr(O \mid S, X = x)$. We assume that the distribution of the observed statistic X is the same for all non-operon gene pairs, *whether or not they are on the same strand*. Using this assumption, we may estimate $\Pr(X = x \mid \overline{O})$ as $\Pr(X = x \mid \overline{S})$, the corresponding probability for adjacent gene pairs that occur on opposite strands, since such pairs are known not to be in a common operon. Given both this and the preceding assumption, it follows (see Section A.1 in the Appendix) that

$$\Pr(O \mid S, X = x) = 1 - \frac{\Pr(X = x \mid \overline{S})}{\Pr(X = x \mid S)} \Pr(\overline{O} \mid S). \quad (1)$$

All of the terms on the right-hand side can be estimated from the annotated genes of the genome or obtained from the prior.

2.3 Use of intergenic distance

The distance between adjacent genes is a powerful signal for operon prediction (Salgado *et al*., 2000; Moreno-Hagelsieb and Collado-Vides, 2002). Among the many features used to predict operons, Bockhorst *et al*. (2003a) found intergenic distance to be the best single predictor of operons in *E.coli*. Genes belonging to the same operon tend to exhibit small intergenic distance; indeed, it is not uncommon for the distances between these genes to be *negative*; i.e., the end of one gene overlaps the start of the next. In contrast, genes not in the same operon have a more uniform distribution of intergenic distance.

We assume that intergenic distance in non-operon gene pairs is not strongly biased between genes on the same strand and genes on opposite strands. Using this assumption, we apply Equation (1) to estimate for each gene

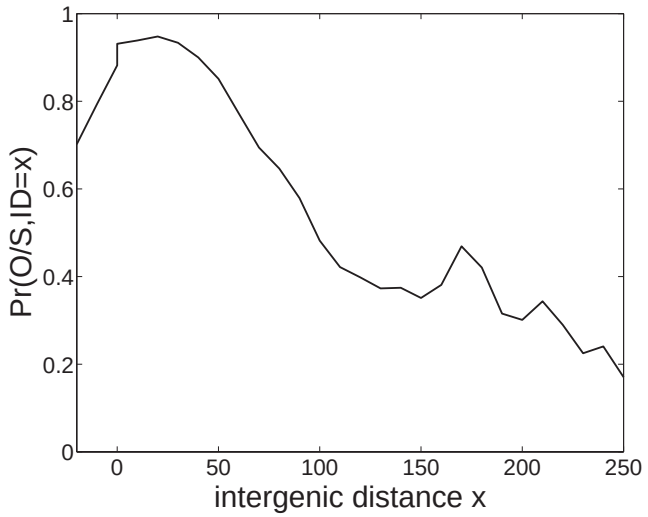

Fig. 1. Estimated probability in *B.theta* that a gene pair belongs to a common operon as a function of its intergenic distance, computed as described in Section 2.3.

pair the probability $\Pr(O \mid S, ID = x)$, where *ID* is intergenic distance. This probability is plotted as a function of the distance in Figure 1. To reduce noise in the observed distribution of intergenic distance, we smooth the distribution by assigning each observation of a distance between $10x - 4$ and $10x + 5$ to a bin centered at $10x$ and storing only the size of each bin.

2.4 Use of functional relatedness

Genes within operons tend to have related functions. Hence, the functional relatedness of two adjacent genes suggests that they may belong to the same operon. However, quantifying the function of a gene is challenging, particularly in a newly sequenced genome. The best available evidence in such a genome typically comes from comparative annotation, in particular strong protein-level similarity between a new gene and another gene of known function. These annotations are summarized for microbial genomes in GenBank by one-line textual descriptions in the file of predicted genes included with each genome. A more controlled classification, such as Gene Ontology (GO) terms or Enzyme Commission (EC) numbers, or a manual classification like that in Neidhardt *et al*. (1996), might be more informative but requires substantial labor and has not been done for many genomes of interest.

As a surrogate statistic for functional relatedness of two genes, we use the length of the longest common substring in their one-line annotations. Once again, we apply the *a priori* model to estimate $\Pr(O \mid S, CL = y)$, where *CL* is the common substring length. To limit the number of matches between unrelated genes, we remove common stop words, such as pronouns, and generic terms such as 'protein,' 'conserved,' or 'hypothetical,' from the descriptions before comparing them.

2.5 Conserved gene clusters

In the absence of selective pressure, genes in bacteria tend to become rearranged with respect to each other (Mushegian and Koonin, 1996). Genes that cluster together in multiple organisms are therefore more likely to be members of the same operon. Ermolaeva *et al*. (2001) found that a pair of adjacent, same-stranded genes A and B whose corresponding orthologs A' and B' are adjacent in another genome are likely to belong to the same operon. Genes A and A' in two organisms were judged orthologous if each was the other's highest-scoring BLASTP match in the other organism.

Our predictor relaxes both the adjacency and orthology criteria for declaring that a pair of genes belong to a common cluster, in order to account for some common features of operons. We relax the criterion of adjacency by allowing a cluster in two genomes to differ by one or more missing or inserted genes. Instances of gene insertion and deletion in actual operons are well known in *E.coli* and *B.subtilis*; for example, the glycogen biosynthesis and degradation operon glgBCDAP in *B.subtilis* (Itoh, 2004, http://www.cib.nig.ac.jp/dda/taitoh/bsub.operon.html) is represented as glgCAP in *E.coli* (Salgado *et al*., 2004). Moreover, we relax the criterion of orthology by considering matches between genes in which each gene is not necessarily the other's best BLAST hit. This criterion allows for uncertainty about which gene in one genome is orthologous to a given gene in the other.

Our algorithm for detecting gene clusters compares the *target genome*, i.e. the one whose operons are being annotated, to a second *informant genome*. We first divide the target genome into directons, since clusters corresponding to operons cannot cross directon boundaries. We compare all genes in each directon to the target genome using NCBI BLASTP (Altschul *et al*., 1997) and keep all matches with E-values less than or equal to some threshold τ. These matches indicate *possible* (though by no means definite) orthologous gene pairs between target and informant; in cases of ambiguity, multiple matches are retained for a single target gene. From these matches, we create a directed graph containing one node for each BLAST hit and an edge between each pair of nodes representing two genes on the same strand separated by at most ψ genes in the informant genome. In the current implementation, we follow the practice of Ermolaeva *et al*. (2001) and the TIGR Operon Finder (TIGR, 2004, http://www.tigr.org/tigr-scripts/operons/operons.cgi) by setting $\tau = 10^{-5}$ and $\psi = 4$. Finally, we enumerate every chain of at least two genes in this graph; every subset of such a chain is a candidate cluster. Figure 2 illustrates a chain of four genes representing a potential conserved cluster.

We enumerate chains in the graph for a directon by depth-first search, starting from every node without an incoming edge. In the extreme case where each of m genes in the target directon matches every one of n genes in the informant genome, the time complexity for building and searching the graph is $O(\psi n^2 m)$. In practice, however, most genes in the target match few or no genes in the informant, so that the total time complexity is closer to $O(\psi m)$.

2.6 Significance of clusters

Once we have identified all candidate clusters for a directon, we want to use for operon prediction only those clusters that are unlikely to have arisen by chance alone. We therefore assess the significance of each candidate cluster and weight its contribution to operon prediction by its significance. Again, we consider any subset of genes in a chain through the graph to be a candidate cluster. The size of a typical chain is less than eight genes, so the number of subsets considered per chain remains computationally feasible.

We test cluster significance against the null hypothesis that the genes in the informant genome are randomly ordered relative to the genes in the target genome. In other words, we assume that the informant genome is a uniformly chosen random permutation of the target. Under this null hypothesis we ask, what is the chance that a cluster from the target would also have occurred by chance in the informant?

Our null hypothesis is not unreasonable provided the target and informant genomes are sufficiently diverged. However, it is likely to be grossly violated for closely related genomes. To avoid using informant genomes that are closely related to the target, we compute for each informant genome the ratio $\rho = (n - b)/n$, where n is the total number of orthologous genes between the target and informant (defined by bidirectional best BLASTP hits) and b is the *breakpoint distance* (Watterson *et al*., 1982) between the two genomes, defined as the number of pairs of adjacent genes in the informant whose orthologs are *not* adjacent, or whose relative orientation is not preserved, in the target. We keep only those informants with ρ at most at some threshold β.

The parameter β, which fixes the set of informants used, should be chosen for each target genome so as to exclude informants that are likely to have many non-operon genes in conserved order and orientation relative to the target,

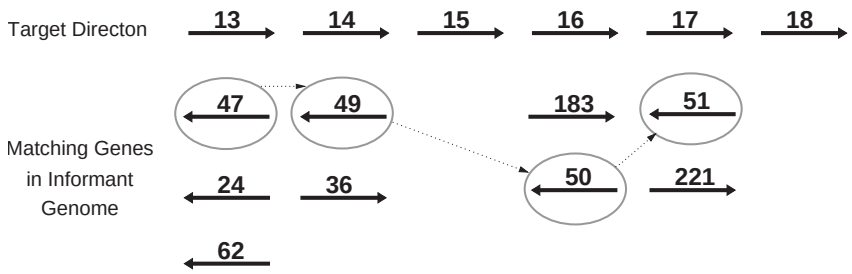

Fig. 2. Example of a chain of conserved genes. Genes are labeled with numbers corresponding to their order in the respective genomes. Circled genes form a chain in the order indicated by the connecting edges.

while still including informants that are likely to share common operons with the target. This problem is similar in spirit to that of choosing informant genomes for conservation-based gene prediction (Korf *et al*., 2001). We choose β based on *a priori* biological considerations; for example, for *E.coli*, we set β so as to exclude other gammaproteobacteria from the set of informants while including more distantly related organisms.

We now describe the significance test used to score clusters. The statistical framework for this test was described by Durand and Sankoff (2003). If the target and the informant have n genes in common, then we assess the significance of a cluster of size k as follows. Given a fixed window w_1 of size m within a directon of the target and a fixed window w_2 of r contiguous same-stranded genes in the informant, the probability $P_e(n,k,m,r)$ that w_2 contains *exactly* k genes from w_1 in the same relative order is given by the hypergeometric distribution,

$$P_e(n,k,m,r) = \frac{\binom{m}{k}\binom{n-m}{r-k}}{\binom{n}{r}k!}.$$

The numerator counts the number of ways to divide the r genes of w_2 into k genes with matches in w_1 and $r-k$ non-cluster-associated genes not from w_1. The denominator counts the total number of ways to choose the r genes in w_2, while $k!$ is the number of possible ways to order the k cluster-associated genes of w_2.

We now wish to know the probability $P_u(n,k,m,r)$ that w_2 contains a cluster of *at least* k genes from w_1. This probability is obtained by summing $P_e(n,k,m,r)$ over all cluster sizes $i \geq k$:

$$P_u(n,k,m,r) = \sum_{i=k}^{\min(r,m)} P_e(n,i,m,r).$$

The window sizes r and m and the length of the shared chain k are properties of a given cluster, so $P_u(n,k,m,r)$ is the chance of seeing a cluster as good as that actually observed purely by chance, at a *fixed position* in the target and informant genomes.

Let w_{2j} be the window of size r starting at position j in the target genome. Define the indicator function

$$X_j = \begin{cases} 1, & \text{if } w_{2j} \text{ has } \geq k \text{ genes from } w_1, \\ 0, & \text{otherwise.} \end{cases}$$

For all j, the expected value of X_j is given by

$$E(X_j) = \Pr(X_j = 1) = P_u(n,k,m,r).$$

To calculate the expected number of windows of size r in the informant genome with $\geq k$ genes from a single window w_1 in the target genome, we sum over the total number N_r of windows of size r in the informant genome:

$$E\left(\sum_{j=1}^{N_r} X_j\right) = \sum_{j=1}^{N_r} E(X_j) = N_r P_u(n,k,m,r).$$

Finally, the total E-value of a cluster, i.e. the expected number $S(n,k,m,r)$ of times such a cluster is expected to be observed between the target and informant genomes by chance, is obtained by summing over all possible windows of size m in the target genome:

$$S(n,k,m,r) = \sum_{i=1}^{D} \sum_{j=1}^{\max(0,d_i-m+1)} N_r P_u(n,k,m,r),$$

where D is the total number of directons in the target genome and d_i is the number of genes in its ith directon.

The E-value $S(n,k,m,r)$ assumes that each gene in the target has only one possible match in the informant. However, our algorithm permits clusters to contain informant genes that are not the best BLAST hits to their target genes. To adjust our E-values for such suboptimal hits, we use the notion of *gene families*, in which one gene from the target may match any of several genes in the informant. For example, if for a given cluster, a gene A' in the informant is the third best match to gene A in the target, A is considered to match a family of three genes.

If the ℓ^{th} informant gene in a cluster is the Φ_ℓ^{th} best match to its target gene, then a cluster containing any of the Φ_ℓ genes in the same family would be at least as interesting as the one actually observed. We must therefore multiply our E-value $S(n,k,m,r)$ by the number of ways of picking informant genes from the families matching each target gene of the cluster:

$$S_f(n,k,m,r) = \left(\prod_{\ell=1}^{k} \Phi_\ell\right) S(n,k,m,r).$$

2.7 Use of clusters in operon prediction

Suppose two adjacent genes in the target genome are found to be in a cluster with E-value $e < 1$. By Markov's inequality, the value e is an upper bound on the p-value p, which is the probability that the cluster occurred purely by chance under the null hypothesis.[1] We therefore let $p = e$, discarding clusters for which e approaches or exceeds 1 ($e > 0.9$).

Because our clustering procedure provides p-values for clusters, we use these values directly for prediction rather than inspecting opposite-stranded gene pairs as for the other data sources. Let $C(p)$ be the event that a pair of adjacent genes occurs in a cluster with p-value p. Let F be a 0-1 indicator for the event that the null hypothesis is false, i.e. that the observed cluster is not a chance event. Then

$$\begin{aligned}\Pr(O \mid S, C(p)) = {} & \Pr(O \mid S, F=1)\Pr(F=1 \mid S, C(p)) \\ & + \Pr(O \mid S, F=0)[1 - \Pr(F=1 \mid S, C(p))].\end{aligned}$$

To incorporate the p-value, we take $\Pr(F=1 \mid S, C(p)) = 1-p$. We set $\Pr(O \mid S, F=1) = 1$; although there might be other reasons for same-stranded genes to cluster besides being in a common operon, we take the view that such clusters are still of biological interest, since our informant

[1] In principle, we must correct for testing multiple clusters for a given pair of genes; however, the large overlap among the clusters covering a given gene pair typically makes their occurrence highly correlated.

genomes are chosen to be highly rearranged versus the target. Finally, we set $\Pr(O \mid S, F = 0) = \Pr(O \mid S)$. The use of the prior here is *not* conservative, since genes lacking a conserved cluster are less likely than average to be in a common operon. However, over 80% of same-stranded gene pairs in *B.theta* exhibit no cluster with $p < 1$, so a large majority of the instances that go into estimating the prior are from the $F = 0$ case. In summary, we estimate

$$\Pr(O \mid S, C(p)) = (1 - p) + p \Pr(O \mid S).$$

Given the large number of available bacterial genomes, it is desirable to find a way to use multiple informant genomes for a given target. When using multiple informants, we consider a pair of adjacent genes in the target to be in a cluster if *at least one* informant genome yields a cluster containing both genes. We retain the p-value for the most significant cluster in any informant but correct it for tests against multiple informants by multiplying by g, the number of informant genomes for which a cluster spanning the two genes *could* occur given the observed BLAST hits between target and informant. Experimentation revealed that cluster scores were being overweighted in the final analysis, possibly because the multiple test correction is insufficient. In order to compensate, $\Pr(O \mid S, C(p))$ was limited to be ≤ 0.95.

2.8 Combining information

Each of our three sources of information—intergenic distance, common annotation length and inclusion in a common cluster—assigns an *attribute* to a pair of adjacent genes. For each attribute X, we have estimated $\Pr(O \mid S, X = x)$ individually. We must now combine this information into a single probability $\Pr(O \mid S, X_1 = x_1 \ldots X_3 = x_3)$, which is our final estimate of whether a gene pair is likely to belong to a common operon.

We use a naive Bayesian combining strategy (Mitchell, 1997, Chapter 6), which assumes that the values of the various attributes are independent given that we know whether or not a gene pair is in a common operon. Under this assumption,

$$\Pr(O \mid S, X_1 = x_1 \ldots X_3 = x_3) = \frac{\prod_{i=1}^{3} \Pr(X_i = x_i \mid S, O) \Pr(O \mid S)}{\Pr(X_1 = x_1 \ldots X_3 = x_3 | S)}.$$

Moreover, we have by Bayes' rule that

$$\Pr(X_i = x_i \mid S, O) = \frac{\Pr(O \mid S, X_i = x_i) \Pr(X_i = x_i \mid S)}{\Pr(O \mid S)}.$$

Combining these observations, we conclude that

$$\Pr(O \mid S, X_1 = x_1 \ldots X_3 = x_3) = \prod_{i=1}^{3} \left[\frac{\Pr(O \mid S, X_i = x_i)}{\Pr(O \mid S)} \right] \Pr(O \mid S) \cdot \gamma,$$

where γ is a constant independent of O, and all the remaining terms either reflect the prior or are computable by the methods of the previous sections.

We make our final operon prediction for a pair of genes by computing each of the two probabilities $v_O = \Pr(O \mid S, X_1 = x_1 \ldots X_3 = x_3)$ and $v_{\overline{O}} = \Pr(\overline{O} \mid S, X_1 = x_1 \ldots X_3 = x_3)$, according to the method above. These two probabilities must sum to 1, which allows us to infer the normalizing factor γ. We may then set a probability cutoff θ between 0 and 1; gene pairs with $v_O \geq \theta$ are considered to be in a common operon, while pairs with $v_O < \theta$ are not.

3 EXPERIMENTAL VALIDATION

We validated our operon predictor on two bacterial genomes, one well-studied and one novel. To enable performance comparison with existing operon finders, we first tested our predictor on the genome of *E.coli*. We then applied the predictor to its intended target, the genome of *B.theta.*, using gene expression measurements as our best available surrogate for ground truth about the genome's operon structure.

3.1 Validation in *E.coli*

We first tested our predictor on the K12 strain of *E.coli* (GenBank accession NC_0009131). The *E.coli* K12 genome has been extensively annotated for both known operons and non-operon gene pairs; these annotations are available through the RegulonDB database (Salgado *et al*., 2004). From this database, we obtained 797 pairs of adjacent, same-stranded genes known to belong to a common operon and 294 such pairs known *not* to belong to a common operon. Non-operon pairs occurred at boundaries between two annotated operons. For each gene pair, we used our predictor to infer whether the pair was in a common operon and compared our result to the known annotation.

The informant genomes used for *E.coli* were derived from a set of 181 bacterial and archeal genomes in GenBank. A complete list of these informants is given in our Supplemental Data (http://www.cse.wustl.edu/∼jbuhler/research/operons). Setting a threshold $\beta = 0.35$ sufficed to exclude other gammaproteobacteria from the set of informants.

Figure 3A shows a receiver operating characteristic (ROC) curve describing the performance of our operon finder. The curve was obtained by varying the threshold θ for the overall probability that an adjacent gene pair is in an operon, as computed in Section 2.8. Pairs scoring above θ were labeled 'operon,' while the remaining pairs were labeled 'non-operon.' Pairs of genes labeled 'operon' and known to belong to a common operon were considered true positives, while pairs labeled 'operon' but known *not* to belong to a common operon were considered false positives.

We achieved a true positive rate of 88% at 20% false positives. Although this true positive rate is slightly lower than that reported for more highly tuned operon finders such as that of Salgado *et al*. (2000), it was obtained with no prior training of parameters on known *E.coli* operons and non-operons.

Overall, our predictor's output is highly enriched for true operons and hence is of value in choosing putative operons for experimental validation in a new genome.

3.2 Validation in *B.theta*

We next applied our operon predictor to the genome of *B.theta* strain VPI-5482 (GenBank accession NC_004663.1). Unlike *E.coli*, *B.theta* does not have a large number of closely related bacterial genomes in GenBank; hence, we set the threshold $\beta = 0.4$, which utilized all of our informant genomes (except *B.theta* itself). In particular, we included as an informant *Porphyromonas gingivalis*, the closest relative of *B.theta* present in GenBank at the time of writing.

Because it has not been extensively annotated, *B.theta* lacks a large database of experimentally confirmed operons that could be used as ground truth for validation. We therefore devised a scheme by which gene expression data acted as a surrogate for knowledge of whether a pair of genes belong to a common operon.

3.2.1 Use of expression data We performed comprehensive transcriptional profiling of *B.theta* using custom Affymetrix GeneChips to obtain growth-phase-associated expression measurements (see Section A.2 in the Appendix for details). Pairs of adjacent, same-stranded genes were hypothesized to belong to the same operon if their expression displayed significant covariation over multiple experimental time points.

More precisely, for each gene in the genome of *B.theta*, we obtained an expression level along with its estimated standard deviation using the dChip analysis software (Li and Wong, 2001). We treated the measurement $\tilde{E}(t)$ at each time point as a Gaussian random variable centered about the true expression $E(t)$ at that time, with the observed standard deviation. We limited attention

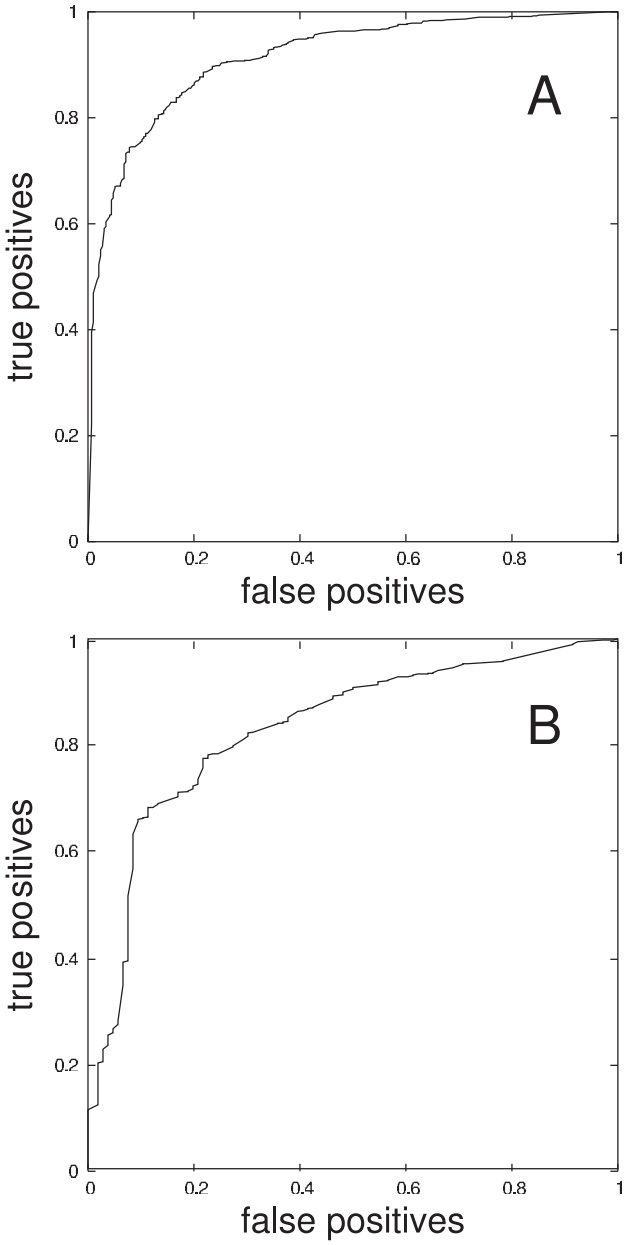

Fig. 3. (**A**) ROC plot for operon prediction in *E.coli*, using RegulonDB as ground truth as described in Section 3.1; (**B**) ROC plot for operon prediction in *B.theta*, using concordant gene expression as ground truth as described in Section 3.2.

to intervals of time during which both genes exhibited a significant change in expression, as follows. For measured expression values $\tilde{E}(t)$ and $\tilde{E}(t+1)$ of a gene at time points t and $t+1$, we computed the probability $\Pr(E(t) < E(t+1))$ that the true expression value $E(t+1)$ exceeds the true value $E(t)$. If this probability exceeded a high threshold τ_{i}, then expression was held to increase significantly; conversely, if it exceeded a low threshold τ_{d}, then expression was held to decrease significantly. Otherwise, no significant change was recorded. In the current implementation, $\tau_{\mathrm{i}} = 0.8$ and $\tau_{\mathrm{d}} = 0.2$.

When a pair of adjacent, same-stranded genes consistently increased or decreased significantly across two or more pairs of consecutive time points, we called those genes *concordant* and labeled them as being putatively in a common operon (a positive example). If the genes exhibited significant changes in opposite directions across two or more pairs of consecutive time points, we called those genes *discordant* and labeled them as being putatively *not* in a common operon (a negative example). Gene pairs that were neither concordant nor discordant were not used in validation. In *B.theta*, this labeling procedure produced 936 positive and 106 negative gene pairs.

Identifying putative operons from expression data can be error-prone, since adjacent, same-stranded genes can display covariant expression due to common regulation without being part of the same polycistronic transcript. As a measure of the accuracy of our surrogate for ground truth, we applied the above labeling procedure to expression measurements from a comparable Affymetrix GeneChip experiment performed with *E.coli* K12. Our expression-based labeling of gene pairs as operon or non-operon agreed with that given by RegulonDB 84% of the time.

3.2.2 Results Figure 3B gives the performance of our predictor relative to expression-based labeling in *B.theta*. True positives represent concordant gene pairs labeled 'operon,' while false positives represent discordant gene pairs labeled 'operon.' We obtained a true positive rate of 73% with 20% false positives.

3.3 Sensitivity analysis

To assess the utility of different data sources in our predictor, we measured the predictive performance when each data source in turn is removed from the predictor. Figure 4 shows the results of these experiments. Sensitivity analysis to changes in parameter values are described in our Supplementary Data.

For both target organisms, each data source by itself gave significantly better predictions than chance alone (data not shown). However, some data sources proved redundant when other high-quality information was available. For *E.coli*, shown in Figure 4A, the best single source of information was intergenic distance, as may be expected given the results of (Salgado *et al*., 2000; Moreno-Hagelsieb and Collado-Vides, 2002). Because *E.coli* has been extensively annotated, functional relatedness also proved a useful source of information. Information from clustering in this case proved redundant to the other two data sources combined, though it is not redundant to either source alone.

Operon prediction for *B.theta* is more challenging than for *E.coli* because the former's genome has been less extensively studied. The available genomes yielded fewer common annotations and fewer conserved clusters, leading to a greater reliance on intergenic distance. The relative dearth of common annotations compared to *E.coli* can be traced to the fact that a larger fraction of genes in *B.theta* (41% versus 34% in *E.coli*) are still labeled only as 'hypothetical.' Moreover, the typical evolutionary distance between *B.theta* and other bacteria in GenBank is much greater than that for *E.coli*, resulting in fewer opportunities to discover conserved clusters.

Sensitivity analysis of the *B.theta* results, shown in Figure 4B, is consistent with our observations about data availability for this organism. The negative effect of removing intergenic distance is considerably more dramatic, while functional relatedness information was typically of little benefit. In contrast to *E.coli*, significant utility was obtained from clustering even given the other data sources,

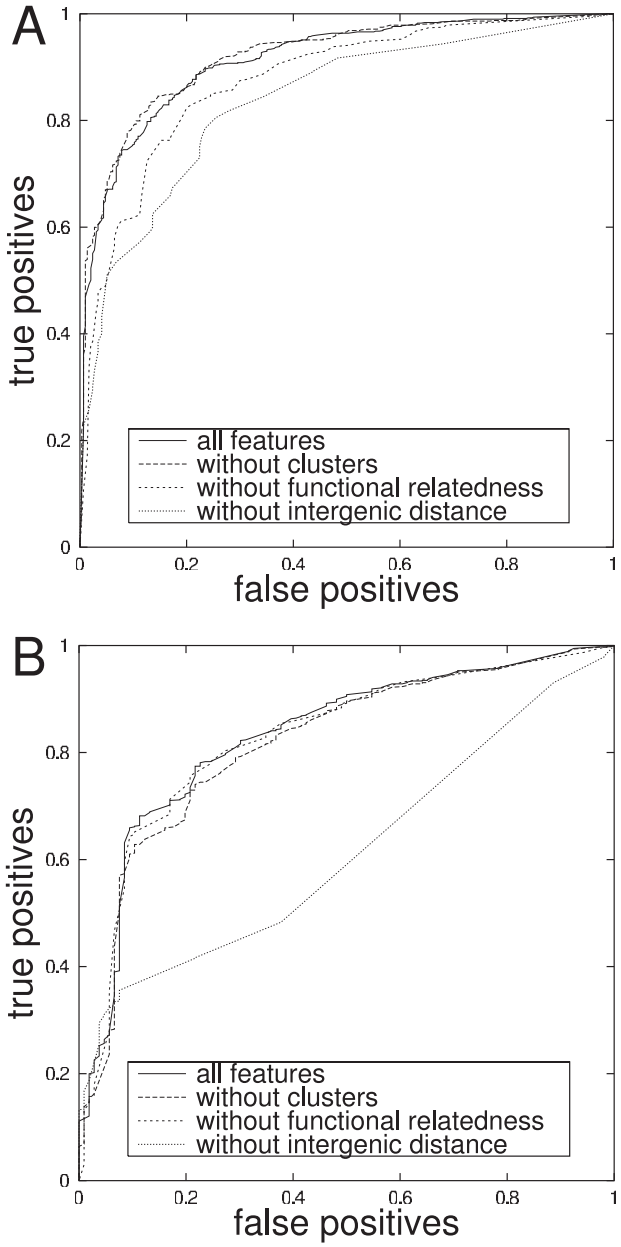

Fig. 4. Sensitivity analysis of operon predictor. (**A**) Effects on *E.coli* predictions of removing each individual data source. Curves for all features and those for all features except clusters are nearly coincident. (**B**) Effects on *B.theta* predictions of removing each individual data source.

particularly for false positive rates around 20%. Most of the benefit of clustering derived from conserved clusters in *P.gingivalis*, the closest relative of *B.theta* in GenBank.

Overall, we conclude that, while clustering and functional relatedness were not always useful, neither one can consistently be eliminated from the predictor without impacting on performance. We would expect that annotations are generally more useful for well-studied genomes, while clusters are more useful for recently sequenced organisms. As the quality of annotation and availability of informant genomes in *Bacteroides* and related groups improve, we will be able to exploit this information to improve our predictions for *B.theta*.

4 DISCUSSION

We have presented a procedure for operon finding designed to work in genomes where operons have not been previously identified. Our predictor combines information from intergenic distance, functional relatedness of genes and conserved gene clusters. Validation with the known operons of *E.coli*, along with corroboration from gene expression measurements in *B.theta*, suggest that our operon finder is robust enough to yield reasonable predictive performance across widely divergent species. While a predictor trained on many known operons of a given organism remains the most accurate available option, our approach provides useful results even in the absence of such training data. Indeed, our method needs only a target genome, a set of gene predictions and minimal functional annotations for it, and one or more informant genomes. Complete sets of predictions for *B.theta* and *E.coli*, along with the source code of the operon finder and other supplemental information, may be obtained online at http://www.cse.wustl.edu/∼jbuhler/research/operons.

A number of potential opportunities exist for improving our predictor while preserving its *a priori* nature. One limitation of our approach to estimating $\Pr(X = x \mid S, \overline{O})$ as described in Section 2.2 is its assumption that intergenic regions between genes on opposite strands have properties similar to regions between same-stranded genes not in a common operon. In observing opposite-stranded genes, we combine statistics from both *convergent* gene pairs (those whose 3′ ends face each other) and *divergent* pairs (those whose 5′ ends face each other). For certain attributes, these two types of gene pairs may look different. For example, we expect divergent pairs to have a larger intergenic distance than convergent pairs, in order to accommodate promoter sites. It may therefore be helpful to consider convergent and divergent pairs separately for parameterizing our *a priori* models.

In using gene clusters, it may be desirable to permit local gene order rearrangements within a cluster. Such changes have been observed in, e.g., the L-arabinose operon, whose genes in *B.subtilis* appear as araA-araB-araD (Itoh, 2004) but in *E.coli* appear as araB-araA-araD. It should be straightforward to extend our E-value estimates to accommodate this change to the cluster model, but such a change will tend to lower the significance of any clusters observed and so must be evaluated for potential loss of sensitivity. More generally, we wish to extend the estimation of significance for clustering to provide more accurate accounting for multiple clusters and multiple informants. However, such an extension is challenging because it must account for the fact that overlapping clusters from one or several related species are not independent events.

Our criteria for choosing informant genomes for clustering are biologically rather than statistically motivated. While biological knowledge was sufficient to make reasonable prior choices of organisms in this work, one might wish for a cluster scoring system that uses *all* genomes as informants while appropriately discounting those that prove too closely related to the the target. The scoring system of Ermolaeva *et al*. (2001) has this property for gene pairs; for larger clusters, a modification of our system may be possible. The principal barrier is not statistical but computational: the cost of enumerating and scoring all clusters shared by two closely related genomes is quite high. Future work may be able to reduce this cost.

Finally, while this work reserved gene expression measurements to validate our predictor, our results in *E.coli* suggest that integrating this data into the operon finder, as has been done by, e.g., Bockhorst *et al*. (2003b), would be of considerable value. Our measure of concordant versus discordant expression could be used as an attribute of gene pairs for prediction.

ACKNOWLEDGEMENTS

The authors wish to thank Jeremy Weatherford for invaluable assistance in revising the manuscript and preparing the software for distribution. This work was supported by NSF awards DBI-0237902 and EF-0333284, and by NIH award CDK30292.

REFERENCES

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003a) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227–1235.

Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F. and Craven,M. (2003b) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19**, i34–i43.

De Hoon,M.J.L., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2004) Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. In *Proceedings of the 2004 Pacific Symposium on Biocomputing*, World Scientific, Singapore, pp. 276–287.

Durand,D. and Sankoff,D. (2003) Tests for gene clustering. *J. Comput. Biol.*, **10**, 453–482.

Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.

Itoh,T. (2004) *Bacillus subtilis* operon predictions, http://www.cib.nig.ac.jp/dda/taitoh/bsub.operon.html

Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**(Suppl. 1), S140–S148.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci.*, **98**, 31–36.

Mitchell,T. (1997) *Machine Learning*. McGraw Hill, New York.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in eukaryotes. *Bioinformatics*, **18**(S1), S329–S336.

Mushegian,A.R. and Koonin,E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet.*, **12**, 289–290.

Neidhardt,F., Curtiss,R., Ingraham,J., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W., Riley,M., Schaechter,M. and Umbarger,H.E. (eds) (1996) *E.coli Gene Products: Physiological Functions and Common Ancestries*, 2nd edition. American Society for Microbiology, Washington, DC, pp. 2118–2202.

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.

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.*, **97**, 6652–6657.

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.

TIGR (2004) The Institute for Genome Research. TIGR operon finder.

Tjaden,B., Haynor,D.R., Stolyar,S., Rosenow,C. and Kolker,E. (2002) Identifying operons and untranslated regions of transcripts using Escherichia coli RNA expression analysis. *Bioinformatics*, **18**, S337–S344.

Watterson,W.A., Ewens,W.J., Hall,T.E. and Morgan,A. (1982) The chromosome inversion problem. *J. Theoret. Biol.*, **99**, 1–7.

Xu,J., Bjursell,M.K., Himrod,J., Deng,S., Carmichael,L.K., Chiang,H.C., Hooper,L.V. and Gordon,J.I. (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science*, **299**, 2074–2076.

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.

Table A1. Formulation of TYG-rich media for *B.theta*

Component	Concentration
Tryptone	1%
Yeast extract	0.5%
Glucose	0.2%
Potassium phosphate buffer, pH 7.2	100 mM
Cysteine	4.1 mM
Histidine	200 μM
$CaCl_2$	6.8 μM
$FeSO_4$	140 nM
$MgSO_4$	81 μM
$NaHCO_3$	4.8 mM
NaCl	1.4 mM
Hematin	1.9 μM
Vitamin K_3	5.8 μM

APPENDIX

A.1 Derivation of Equation (1)

We wish to estimate $\Pr(O \mid S, X = x)$. We have

$$\begin{aligned}\Pr(O \mid S, X = x) &= 1 - \Pr(\overline{O} \mid S, X = x) \\ &= 1 - \Pr(\overline{O}, S \mid S, X = x).\end{aligned}$$

Applying Bayes' rule gives

$$\begin{aligned}\Pr(\overline{O}, S \mid S, X = x) &= \frac{\Pr(S, X = x \mid \overline{O}, S)\Pr(\overline{O}, S)}{\Pr(S, X = x)} \\ &= \frac{\Pr(X = x \mid \overline{O}, S)\Pr(\overline{O}, S)}{\Pr(S, X = x)}.\end{aligned}$$

Our second assumption permits us to make the substitution

$$\Pr(X = x \mid \overline{O}, S) = \Pr(X = x \mid \overline{S}),$$

giving

$$\Pr(\overline{O}, S \mid S, X = x) = \frac{\Pr(X = x \mid \overline{S})\Pr(\overline{O}, S)}{\Pr(S, X = x)}.$$

Finally, applying the chain rule for probabilities, we have that $Pr(\overline{O}, S) = Pr(\overline{O} \mid S)Pr(S)$, and that $Pr(S, X = x) = Pr(X = x \mid S)Pr(S)$. Equation (1) follows immediately.

A.2 Experimental protocol for expression microarrays

We performed comprehensive transcriptional profiling of *B.theta* using custom Affymetrix GeneChips including probe sets for protein-coding genes and tRNA species on the chromosome and plasmid (p5482) of strain VPI-5482. Probe sets were created for 4719 *B.theta* genes, with 13 probe pairs per gene.

B.theta was grown on TYG rich media (tryptone, yeast extract and glucose; see Table A1 for exact formula) in a BioFlo-110 chemostat (New Brunswick Scientific, Edison, NJ) equipped with twin 1.3 L fermentation vessels. Growth was allowed to proceed in an atmosphere of 80% N_2:20% CO_2.

Samples of *B.theta* were collected at five time points—3.5, 4.5, 5.5, 6.5 and 8.83 h after inoculation—during growth from mid-log to stationary phase. At each time point, aliquots were removed from each vessel and placed in RNAProtect (Qiagen, Valencia, CA), and RNA was isolated (RNeasy; Qiagen). Genomic DNA contamination was minimized by treatment with DNAfree (Ambion, Austin, TX). cDNA targets were prepared using methods described in the *E.coli* Antisense Genome Array Manual (Affymetrix, Santa Clara, CA). Scanned images of hybridized arrays were quantified and interpreted using the dChip software (Li and Wong, 2001).

For comparison, we performed a time course experiment similar to the above using *E.coli* strain MG1655 grown in standard Luria-Bertani media. Gene expression was measured using Affymetrix *E.coli* ASV2 GeneChips at 2.5, 3.8, 4.5, 5.3 and 7.8 h post-inoculation.