Nucleic Acids Research, 2005, Vol. 33, No. 10 ***3253–3262***
doi:10.1093/nar/gki634

A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context

Martin T. Edwards*, Stuart C. G. Rison[1], Neil G. Stoker[1] and Lorenz Wernisch

School of Crystallography, Birkbeck College, London WC1E 7HX, UK and [1]Department of Pathology and Infectious Diseases, Royal Veterinary College, London NW1 0TU, UK

Received August 10, 2004; Revised September 27, 2004; Accepted May 16, 2005

ABSTRACT

An important step in understanding the regulation of a prokaryotic genome is the generation of its transcription unit map. The current strongest operon predictor depends on the distributions of intergenic distances (IGD) separating adjacent genes within and between operons. Unfortunately, experimental data on these distance distributions are limited to *Escherichia coli* and *Bacillus subtilis*. We suggest a new graph algorithmic approach based on comparative genomics to identify clusters of conserved genes independent of IGD and conservation of gene order. As a consequence, distance distributions of operon pairs for any arbitrary prokaryotic genome can be inferred. For *E.coli*, the algorithm predicts 854 conserved adjacent pairs with a precision of 85%. The IGD distribution for these pairs is virtually identical to the *E.coli* operon pair distribution. Statistical analysis of the predicted pair IGD distribution allows estimation of a genome-specific operon IGD cut-off, obviating the requirement for a training set in IGD-based operon prediction. We apply the method to a representative set of eight genomes, and show that these genome-specific IGD distributions differ considerably from each other and from the distribution in *E.coli*.

INTRODUCTION

Following determination of the genome sequence and gene identification, a major goal in understanding an organism's biology is to define the transcriptional regulatory networks. In prokaryotes, a key regulatory feature is the operon—a collection of genes transcribed into a polycistronic mRNA. Thus, determining the transcription unit (TU) or operon map indirectly locates promoters, and assists in regulon definition, identification of regulatory networks, regulatory motif detection and interpretation of microarray expression data. As co-transcription is an efficient regulation of genes that have a related biological function, determination of operons is also useful in assignment of gene function and metabolic reconstruction.

Intergenic distance

The most generally applicable and successful pairwise operon prediction method is the intergenic distance method developed by Collado-Vides and co-workers (1,2). A within-operon likelihood for a gene pair is derived by comparing the frequencies of within-operon and between-operon gene pairs at the given intergenic distance. The procedure has also been incorporated into several combination prediction methods (3,4). Using only intergenic distance, this method has a pairwise accuracy (mean of sensitivity and specificity) of 81% in the case of *Escherichia coli*, producing a map of TUs recovering 65% of known operons from the RegulonDB dataset (5) (http://www.cifn.unam.mx/Computational_Genomics/regulondb/). Besides distance, other predictors such as expression level, predicted promoter/terminator and operon size usually add comparatively little to the accuracy of predictions (4,6). Despite the importance of the distribution of inter-operon distances for operon prediction, such distribution is well established only for *E.coli* and *Bacillus subtilis*, the only organisms with enough experimentally verified operons to enable a reasonable density estimate. Due to this limitation, the distributions from *E.coli* or *B.subtilis* are usually copied over to other organisms for the purpose of operon prediction (7). This re-use of the *E.coli* operon set assumes that all prokaryotic genomes share a common within-operon intergenic distance distribution.

Conservation methods

Complementary to the intra-genome property of intergenic distance is the inter-genome property of contextual

*To whom correspondence should be addressed. Tel: +44 20 7631 6831; Fax: +44 20 7631 6803; Email: m.edwards@mail.cryst.bbk.ac.uk

conservation of genes. Evolution in prokaryotes has occurred in a variety of niches. Operons coding for ubiquitous core processes, and to a lesser extent more specialized processes, will have evolved and been conserved across several organisms, in groups of varied phylogenetic composition. Occurrence of operons conserved in these groups has three possible explanations: the shared operon belongs to evolutionarily related organisms; there has been a horizontal transfer of the genes; or unrelated organisms have converged on a solution to a common problem using similar gene activities. While the first two cases initially impose a gene order, genome rearrangement is common (8), and convergent evolution imposes no gene order. Therefore, the content of the operon should remain broadly the same and observation of a conserved cluster of genes in proximity suggests an operon.

This idea was used by Ermolaeva *et al*. (9) who developed a method to assess the probability of an adjacent gene pair being within an operon by observing the frequency with which their homologues occur in other genomes, imposing the constraints of immediate adjacency and a 200 bp intergenic distance cut-off on the candidate gene pairs and their homologues.

Gene pair adjacency may not be conserved between species, even when genes stay within the same operon. There might be insertions, deletions or reshufflings within an operon. These problems have been addressed in several works. Overbeek *et al*. (10) allowed insertions in their assessment of conserved gene pairs. Wolf *et al*. (11) developed a method that allows for the insertion of genes into an instance of a conserved gene cluster, but the gene order must be at least partially conserved. The lack of conserved gene order is also partly addressed in later work on conserved gene neighbourhoods (12) and über-operons (13). The method of Zheng *et al*. (14) also relaxes the immediate adjacency constraint in their gapped phylogenetic profile method, but it remains limited to seeking gene pairs in a second genome separated by at most two genes. This gapped comparison increased the number and accuracy of their predictions of functional dependency.

Here, we describe a novel graph algorithmic method that identifies conserved clusters of functionally related genes. These clusters are predominantly operons or fragments of operons. This is not surprising as co-transcription is probably the strongest constraint on conserving a gene's genomic context. The method is flexible enough to identify an instance of a conserved cluster that has been completely shuffled relative to the canonical operon and has any number of insertions. The method seeks to identify conserved clusters of genes in a query genome by finding reoccurring unordered groups of homologues from several genomes under the assumption that functionally related genes co-occur in the same directon. We call this process a directon versus directon analysis (DVDA). A directon is a set of consecutive genes on one strand of DNA which are not interrupted by RNA genes or genes on the opposite strand. Each directon is composed of one or more TUs. Conversely, all the genes in an operon exist in the same directon because operons do not normally cross strands.

Using the intergenic distance distribution of pairs identified by the DVDA algorithm, we take an empirical Bayes approach to infer the probability that a pair of adjacent genes with a certain intergenic distance belongs to a common operon. This amounts to establishing an intergenic distance cut-off beyond which it becomes increasingly unlikely that the gene pair belongs to the same operon. A complete genome operon map can be created by joining gene pairs with a high probability of belonging to the same operon. By predicting an operon pair set and using this to establish a genome-specific intergenic distance distribution, we surmount the need for a known operon training set and can create a TU map of any completely sequenced genome with only a minimal annotation, that is, an open reading frame (ORF) prediction. The results of this analysis indicate that there is no universally applicable within-operon intergenic distance distribution, and that for intergenic distance-based operon prediction, each genome must be considered as a distinct case.

In this paper, we suggest a series of progressively stronger assumptions on functionally related gene pairs and present the results that can be derived by adopting the assumptions. First, a graph algorithm exploits the effect of keeping genes which are functionally related in close vicinity throughout evolution. This stage results in a list of functionally related gene pairs. Such functional pairs mostly comprise gene pairs in a common operon but also a few nonoperon pairs (e.g. repressors or inducers near the controlled operon). For the next stage, we assume that the distance distribution within the functional pairs is representative of the distance distribution within operon pairs. Finally, we provide probability estimates for operon pairs in case one is prepared to accept all pairs identified by the graph algorithm as operon pairs. Although this introduces false positives (FPs), their number is quite limited and might be acceptable in view of the gain in accuracy in the probability estimations.

MATERIALS AND METHODS

Our approach to generating a genome-specific intergenic distance distribution for genes in operons consists of three steps. In the first step, conserved clusters of genes are identified by finding pairs of directons from different genomes with similar gene composition in a DVDA. Next, a graph algorithmic analysis of the directon comparisons results in a set of candidate operon gene pairs. Finally, a statistical analysis of the distance distribution of these gene pairs provides probabilities that a pair belongs to a common operon based on its intergenic distance.

Genomes and operon resources

A set of 74 genomes/chromosomes was obtained from the European Bioinformatics Institute ftp site (ftp://ftp.ebi.ac.uk/pub/databases/genomes/). We ensured that no two genomes of the set are too similar using the criteria set down in Ref. (2), assessing the evolutionary separation of two genomes by the degree of sequence divergence of a limited gene set common to both organisms. Such non-redundant set reduces the inclusion of gene pairs belonging to different operons but retaining adjacency due to evolutionary proximity. The set of non-redundant genomes comprised 1 Aquificae, 2 Crenarchaeota, 7 Euryarchaeota, 20 Proteobacteria, 3 Spirochaetes, 21 Firmicutes, 3 Actinobacteria, 2 Bacteroides, 3 Chlamydiae, 1 Chlorobi, 1 Deinococci, 1 Fusobacteria, 1 Cyanobacteria and 1 Thermotogae.

The *E.coli*-verified operon set was provided by RegulonDB and 100 *B.subtilis* operons are available online (http://www.

cib.nig.ac.jp/dda/taitoh/bsub.operon.html). The generation of a putative non-operon pair set was achieved by pairing the first and last gene of an operon with its neighbour in the same directon, if such gene exists. This ensures that for each non-operon pair a known transcriptional boundary is crossed. This is not an ideal solution as the extra gene could very well belong to the operon when expressed under different conditions or as a readthrough. However, using such a set does allow a rough estimate of prediction precision.

Identification of maximum matchings between directons

Each of the gene sets of the target genomes was compared to that of the query genome, for example, *E.coli*, using BLASTP (15), with the BLOSUM62 matrix and an *E*-value cut-off of 10^{-5}. Assignment of homologous pairs, as detailed below, was not dependent on a high degree of sequence similarity. For each query directon with homologues in at least three organisms (suggesting conservation), the scored (BLASTP bitscore) homologues from a single target directon were retrieved. In general, there are instances where a query gene has more than one homologue in the target directon, and two query genes may exhibit homology to a single target gene. While the naïve approach would be to choose the homologue with the highest bitscore, this may force another gene in the query directon to align with a weaker homologue or not at all. To solve this problem, the following graph algorithmic approach was developed [for graph theoretic concepts and algorithms, see (16)].

The optimal assignment of query genes to their homologues in a single directon is achieved using a maximum weight maximum cardinality bipartite matching algorithm (graph algorithms were applied using LEDA—a Library of Efficient Datatypes and Algorithms; http://www.algorithmic-solutions.com/enleda.htm). A graph $G(V, E)$ is bipartite if its vertex set V can be divided into two sets V_1 and V_2, such that all edges are between V_1 and V_2 and there are no edges within V_1 or V_2. In our application, V_1 represents the genes in a directon of *E.coli* and V_2, the genes in a directon of a target genome. An edge $e \in E$ connects a vertex in V_1 with one in V_2 and represents a similarity above the cut-off; it is weighted with the alignment's bitscore. A matching is a set of edges that do not share any common vertices. A maximum weight maximum cardinality matching is a matching where the number of vertices matched is maximal and, in case there are several such matchings, one among them with the largest cumulative edge weight. Such a matching is more likely to yield the correct pair of orthologues as context and sequence similarity is a stronger indication of orthology than sequence similarity alone. Of all possible matchings for a query directon with all directons of a target organism, only the highest cardinality matching is retained. For some directons of *E.coli* and some target genomes, no matchings could be constructed due to the lack of homologies above the similarity cut-off.

Figure 1 shows examples of directon versus directon comparison, while Figure 2 clarifies the graph matching process and underlines how the process achieves a directon alignment independent of gene order. In Figure 2, we see that there is contention in the assignment of orthologues when searching for the 'best' alignment between the directons. A gene (circle) may have a similarity relationship (line) with more than one gene in a directon. In this example and its matching solution (lower half), gene I has been matched to A even though it has greater similarity to B. This is the effect of a search for a maximal weighted matching. The final matching is shown in the lower half of Figure 2. This has the highest cardinality possible and a larger score of 25 compared to the other three gene matchings. In the example shown, there is only one 'gap' in the alignment (gene II) and only one rearrangement (genes III and IV are inverted relative to genes B and C). However, there is no constraint on how many genes can be inserted or how shuffled these genes are with conserved functional core.

The assignment of orthology between a single gene from one directon to a single gene in another directon is complicated by gene duplication, that is, the presence of paralogues. If a gene in the query directon has more than one potential match, and either paralogue can legitimately be chosen without detriment to the overall matching, then the paralogue with the highest BLAST score will be chosen. Similarly for the case where there are two genes in the query directon that can match to a single gene in the target directon, the highest scoring pair

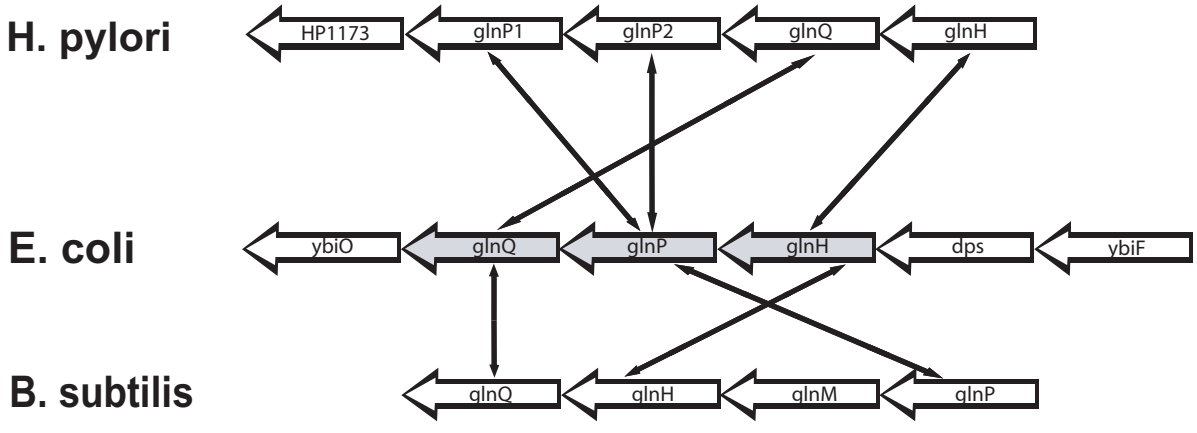

Figure 1. Example directon alignment. Presented is the *E.coli* directon that contains the glnHPQ operon (shaded). Large arrows represent genes, each row of genes is an observed directon. Black arrows indicate sequence similarity. Clearly, rearrangement of genes is common, with insertion of extra genes and paralogues. Use of the matching algorithm ensures an appropriate assignment of homologous pairs without reliance on conservation of order.

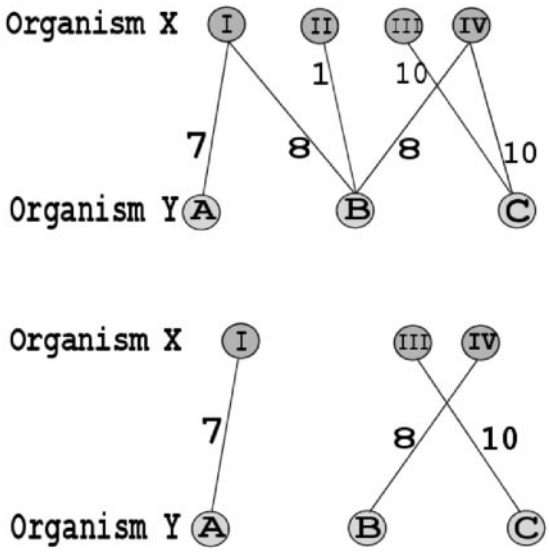

Figure 2. An example matching. Two artificial directons from two organisms are matched using a maximum weight maximum cardinality bipartite matching algorithm. See main text for description.

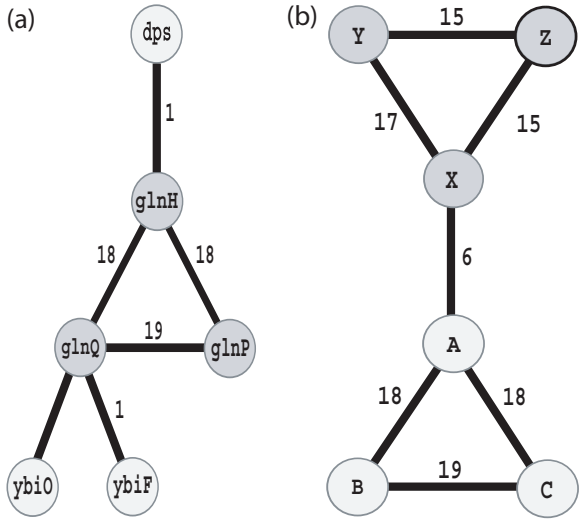

Figure 3. (**A**) Merged graph optimization. Matchings for the *E.coli* directon that contains the glnHPQ operon are merged into an undirected graph with edge weights of the number of target genomes with homologous genes in the same directon. The trimming procedure, utilizing a minimum cut algorithm, removes the dps-glnH, ybiO-glnQ and ybiF-glnQ edges to reveal the core glnHPQ operon. Note the smaller edge weights of gene pairs outside the core cluster. Further trimming of the graph occurs until all possible remaining cuts have cut weights above the threshold. (**B**) Removal of whole subclusters. In this artificial example, edge XA would be removed by the minimum cut, producing a pair of disconnected graphs. With its lower cumulative edge weight, the XYZ graph would be discarded, to be recovered in further iterations.

will be matched. The other paralogue can be in either of two general positions relative to the matched paralogue. It can be inside, in the sense that it is between the matched paralogue and another gene matched to its orthologue; or it can be outside in the sense that it is not bound by other genes of the matching. If the unmatched paralogue is inside, then it will be recovered anyway (described in the next section). If it is outside, then it will be lost for the comparison of the query organism against that single target organism. Integrating such paralogues in an operon is beyond the reach of an approach based on comparing genomes. The final case is where both the query and target have a pair of paralogues, A and its paralogue B in the query and A′ and its paralogue B′ in the target. The assignment of the most likely orthologue match A with A′ and B with B′.

Merging matchings for a single query directon

For each directon of the query organism, all matchings are merged to create a graph $G(V_d, E_d)$ as follows. The vertices $v \in V_d$ are genes from the query directon. All vertices are connected to each other by edges of weight 0. The maximum matching of the first target genome is retrieved. For any two genes $v_1, v_2 \in V_d$ in the query directon, it is established whether they are both connected to homologues from the target genome in the maximum matching. By construction of maximum matchings, the two homologues are different and located in the same directon of the target genome. If v_1 and v_2 both have homologues, the weight of the edge connecting them is incremented by one, otherwise the weight remains unchanged. The procedure is repeated for each target genome. Finally, all edges with weight zero are removed. This construction ensures that the integer weight of an edge $e \in E_d$ connecting two genes in a directon of the query organism reflects the co-occurrences of homologues in a single directon across all the other target genomes.

To maximize the use of the context data, the resulting merged graph is pruned using a minimum cut algorithm. A subset of edges of a connected graph is separating if their removal leaves a disconnected graph. A minimum cut algorithm returns a separating set of edges that has minimal total weight among all possible separating sets. If the weight of a minimal cut is below a fraction of 0.9 of the total edge weight of the graph, the edges of the cut are removed. This cut-off fraction was chosen after varying the threshold and assessing the specificity of the predictions. Decreasing this threshold would result in the inclusion of less well-conserved gene clusters, but at the expense of including more FP operon pairs.

Furthermore, only the component with the largest total edge weight is retained while vertices in the other components are discarded (to be reclaimed in later iterations). At this stage of the process, the graph is trimmed down to a core of highly conserved genes within the query directon with all low frequency associations removed (see Figure 3A and B).

The genes that remain in the core cluster are taken as the boundary of a set of genes contained in the same operon. The boundary genes as well as genes between them on the chromosome are collected into a conserved cluster. We include these intervening genes so as to account for organism-specific 'hitch-hiking' genes (12). These genes may be inserted and retained in an operon purely on the basis of a fortuitous coincidence of beneficial expression levels between the two TUs. For ease of discussion, the whole cluster will be referred to as 'conserved', though in general not all genes in the cluster will be found in the other genomes. The genes in the conserved cluster are then removed from all graph data structures and the above process is iterated for the remaining genes in the directon. In this way, all conserved clusters from all directons are gathered. The final output of the graph algorithmic stage of the

analysis is a list of all adjacent pairs of genes which occur in a common conserved cluster.

Estimating the probability of an operon pair

In this section, we discuss an empirical Bayes approach (17) to the estimation of the probability that a pair of adjacent genes is in the same operon given the number, d, of base pairs between them. The DVDA algorithm as described above produces a list of DVDA pairs, pairs of adjacent genes that are functionally related. In the following, we denote the event that a pair is a DVDA pair by G ($\bar{G}$ indicates that a pair is not a DVDA pair). If a pair of adjacent genes is in the same operon, it is an operon pair. The event of a pair being in the same operon is denoted by O ($\bar{O}$ indicates that a pair is not in a common operon). In this section, we propose a few simplifying assumptions that will allow us to estimate probabilities for gene pairs to be in the same operon depending on their distance.

Comparing the densities of distances in DVDA (dashed line in Figure 4a) and operon pairs in *E.coli* as obtained from RegulonDB (solid line) suggests that these distributions are actually very similar. Our first assumption thus is

$$P(d \mid O) = P(d \mid G). \quad \mathbf{1}$$

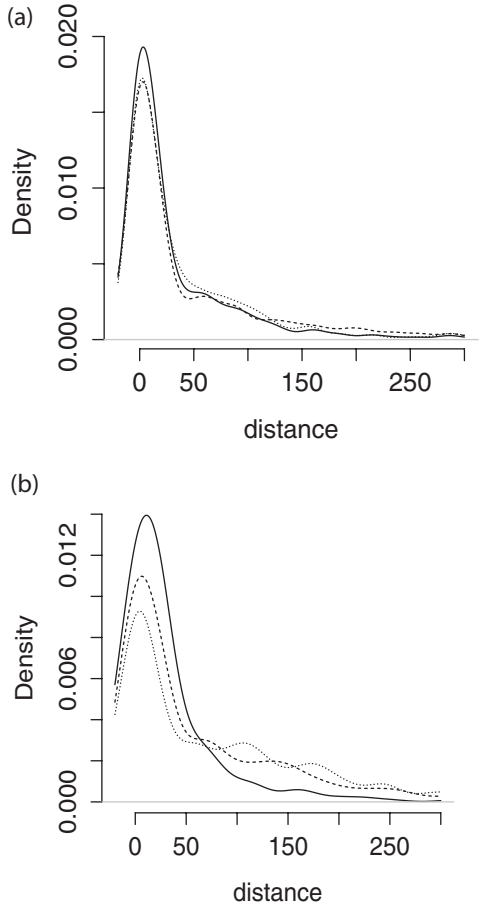

Figure 4. Comparison of kernel density estimates for (**a**) *E.coli* and (**b**) *B.subtilis* of $P(d \mid G)$ (dashed), $P(d \mid O)$ (solid) and $P(d \mid O, \bar{G})$ (dotted) (Gaussian kernel density, with bandwidth (a) 10 and (b) 15. The densities are almost identical for (a) and similar in (b).

A second approximating, although not strictly true, assumption is that all genes pairs with no base pairs between them belong to the same operon:

$$P(O \mid d = 0) = 1. \quad \mathbf{2}$$

Using the assumption from Equation 1 and Bayes theorem, the posterior probability that a gene pair is an operon pair conditioned on its distance d is

$$P(O \mid d) = \frac{P(d \mid O)P(O)}{P(d)} = P(G \mid d)\frac{P(O)}{P(G)}. \quad \mathbf{3}$$

We estimate $P(G \mid d)$ by a nonparametric logistic regression of a variable indicating whether a particular gene pair is a DVDA pair or not depending on its distance d. The regression was performed using the R statistical language (version 2.0.1) (http://cran.r-project.org) and the *mgcv* package (version 1.1-8) (18). A penalized smoothing regression spline is fitted in a logistic model by the procedure *gam*, which automatically determines the amount of smoothing by generalized cross-validation. The unknown proportion $P(O)$ of operon pairs among all gene pairs has the role of a scaling factor in Equation 3, and is set so that the assumption in Equation 2 is fulfilled. An alternative estimate of $P(O)$ can be obtained from

$$\frac{P(d=0)}{P(d=0 \mid G)} = \frac{P(O \mid d=0)P(d=0)}{P(d=0 \mid O)} = P(O), \quad \mathbf{4}$$

using assumptions from Equations 1 and 2. We estimate the two densities around 0 by the relative frequency of gene pairs with their distance d within the range from -30 to 30. As seen in Table 1, both estimates of $P(O)$ are very similar.

Equation 3 is completely general and does not depend on knowledge whether a gene pair is a DVDA pair or not. It rests on the assumption that the intergenic distance distributions of operon pairs and DVDA pairs are the same. A third assumption allows us to take the information about the outcome of a DVDA analysis into account as well. Looking at Figure 4a and noting the similarity of density plots of distances of DVDA pairs, $P(d \mid G)$ (dashed line), and of gene pairs which are in operons (according to RegulonDB) but not captured by DVDA, $P(d \mid O, \bar{G})$(dotted line), suggests the following simplifying assumption:

$$P(d \mid O, \bar{G}) = P(d \mid G) = f_1(d). \quad \mathbf{5}$$

We define $f_0(d) = P(d \mid \bar{O}, \bar{G})$, $\pi_1 = P(O \mid \bar{G})$ and $\pi_0 = P(\bar{O} \mid \bar{G})$. Furthermore, let $g(d)$ be the result of a weighted logistic regression of samples with density $f_1(d) = P(d \mid G)$ and samples with density $f(d) = \pi_0 f_0(d) + \pi_1 f_1(d) = P(d \mid \bar{G})$ on d, for example, the samples from G and $\bar{G}$. The samples need to be weighted so that the weighted total sample size of G is equal to that of $\bar{G}$. After obtaining $g(d)$, we calculate

$$\frac{\pi_1 g(d)}{1-g(d)} = \frac{\pi_1 f_1(d)/(f_1(d)+f(d))}{1-f_1(d)/(f_1(d)+f(d))} = \frac{\pi_1 f_1(d)}{f(d)} = \frac{P(O \mid \bar{G})P(d \mid O, \bar{G})}{P(d \mid \bar{G})} = P(O \mid d, \bar{G}). \quad \mathbf{6}$$

Table 1. Estimated proportions of operon pairs and distance cut-offs for operon pairs for eight genomes

Organism	$P_1(O)$	$P_2(O)$	Cut-off $\bar{G}$	Cut-off all	$P(G)$
Escherichia coli	0.69 [0.65,0.73]	0.71 [0.68,0.74]	57 [40,78]	89 [61,206]	0.29
Escherichia coli nh	0.72 [0.66,0.77]	0.79 [0.75,0.83]	80 [49,176]	122 [67,242]	0.2
Bacillus subtilis	0.79 [0.71,0.81]	0.74 [0.71,0.77]	124 [62,170]	327 [101,497]	0.35
Mycobacterium tuberculosis	0.72 [0.68,0.76]	0.72 [0.69,0.76]	69 [51,89]	89 [66,112]	0.21
Mycobacterium tuberculosis nh	0.7 [0.63,0.76]	0.75 [0.71,0.8]	60 [37,87]	71 [44,102]	0.15
Bacillus cereus	0.68 [0.64,0.7]	0.66 [0.63,0.69]	93 [66,111]	145 [107,174]	0.29
Clostridium perfringens	0.74 [0.68,0.78]	0.77 [0.72,0.82]	105 [45,215]	320 [207,406]	0.39
Aeropyrum pernix	0.51 [0.14,0.56]	0.44 [0.38,0.51]	106 [68,151]	123 [82,171]	0.14
Pseudomonas aeruginosa	0.73 [0.67,0.76]	0.75 [0.72,0.78]	76 [55,108]	129 [92,234]	0.3
Pseudomonas aeruginosa nh	0.72 [0.68,0.77]	0.81 [0.77,0.84]	78 [57,116]	112 [82,251]	0.23
Xylella fastidiosa	0.82 [0.74,0.85]	0.82 [0.76,0.9]	188 [145,230]	211 [167,269]	0.23
Xylella fastidiosa nh	0.77 [0.65,0.84]	0.87 [0.8,0.95]	167 [93,236]	180 [115,257]	0.14

$P_1(O)$, an estimation of the proportion of operon pairs, is obtained by choosing a proper scaling factor in Equation 3, $P_2(O)$ is obtained from Equation 4. Cut-offs $d: P(O \mid d, \bar{G}) = 0.5$ for non-DVDA pairs and a universal cut-off d: $P(O \mid d) = 0.5$ for all gene pairs are obtained from Equations 3 and 6. The 95% confidence intervals in brackets are obtained from 2000 bootstrap replications. $P(G)$ is the proportion of DVDA pairs among all gene pairs. For genomes with more than 50% non-hypothetical genes, the result of the analysis restricted to nonhypothetical genes is provided as well (nh).

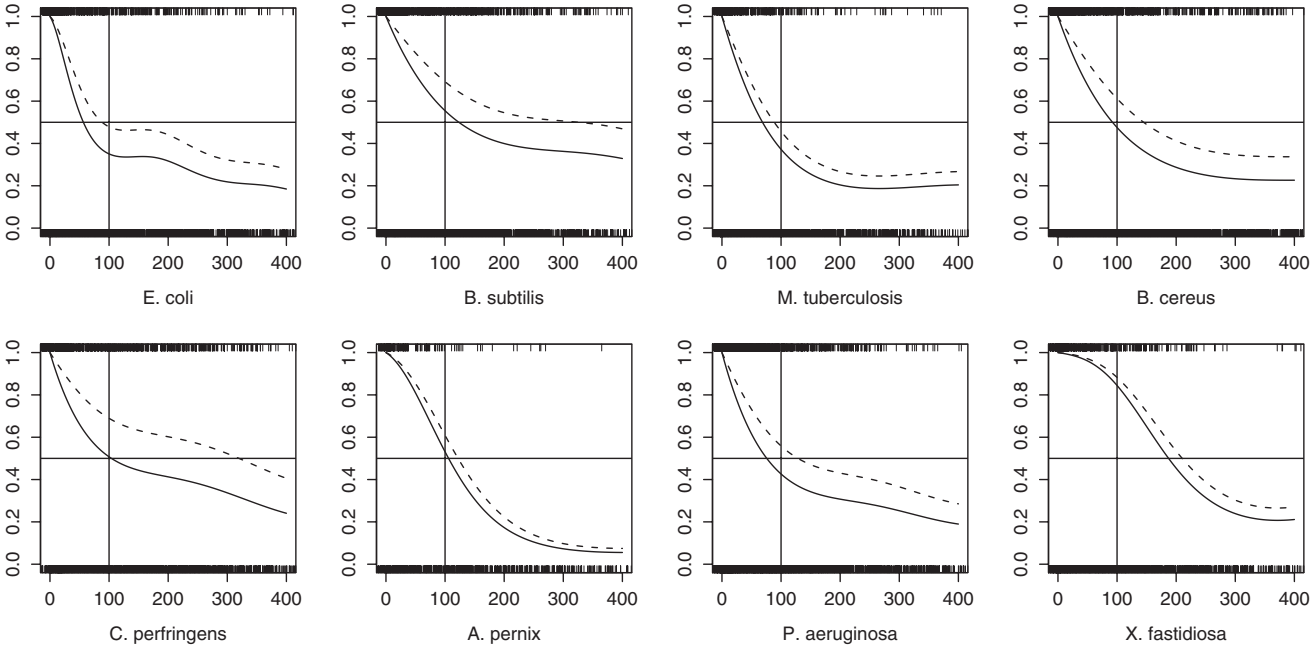

Figure 5. Estimates of the probability $P(O \mid d)$ (dashed) and $P(O \mid d, \bar{G})$ (solid) for several genomes. Shown are also rug plots of distances of graph pairs G (top) and non-graph pairs $\bar{G}$ (bottom).

Again, the scaling factor $\pi_1 = P(O \mid \bar{G})$ is chosen so that $P(O \mid d = 0, \bar{G}) = 1$, which follows from the assumption in Equation 2.

To summarize, Equation 3 rests on the assumption from Equation 1 only, and Equation 6 on the assumption from Equation 5 only, where both assumptions are suggested by Figure 4a. Assumption from Equation 5 alone does not determine $P(O \mid d, G)$, but together with the assumption from Equation 1, it implies $P(d \mid O, G) = P(d \mid G)$ and $P(O \mid d, G) = P(O \mid G)$. Adding assumption from Equation 2 results in $P(O \mid G) = 1$. If these implications are not fully plausible on their own, one has to keep in mind that they are the result of a combination of simplifications and approximations.

Figure 5 shows examples of estimated densities $P(O \mid d)$ and $P(O \mid d, \bar{G})$ for several genomes. Cut-off points for distances d can be derived for a decision whether to count a gene pair as operon pair or not. Assuming a cost of 1 for each misclassified gene pair and a cost of 0 for each correct classification, the expected loss is minimized when exactly the pairs with $P(O \mid d) > 0.5$—or $P(O \mid d, \bar{G}) > 0.5$ if information on DVDA pairs is used—are classified as operon pairs. Table 1 shows estimates of $P(O)$ and distance cut-offs based on this decision rule. It also shows estimates of the 95% confidence interval for all features as derived from the lower and upper 2.5% quantiles of the features for 2000 bootstrap samples. The bootstrap sets of distances were obtained from the original sets by randomly drawing, with replacement, the same number of distances from them. In the Supplementary Material, plots of $P(O \mid d)$ and $P(O \mid d, \bar{G})$ as obtained from DVDA pairs are compared with plots obtained from distances in RegulonDB.

RESULTS

We assess the precision of our DVDA predictions by comparing all the adjacent gene pairs found in DVDA conserved gene clusters from *E.coli* to the RegulonDB known operon list and a putative non-operon set (defined in Materials and Methods). Of the 854 gene pairs in *E.coli* DVDA clusters (of size 2–18 genes), 334 are TP (true positive, found in RegulonDB) and 58 are FP (false positive, found in the putative non-operon pair set). This amounts to a recovery (proportion of TP among true pairs) of 49.1% of the RegulonDB known operon pairs and a precision (proportion of TP among positives) of 85.2%. Importantly, DVDA does not constrain intergenic distance and is able to predict conserved gene pairs with large intergenic separation, in contrast to the method of Salgado *et al*. (1) which cannot predict beyond 100 bp separation due to the paucity of known operon pairs at these distances. DVDA achieves a precision of (39%) beyond 100 bp separation, with 26 TP and 41 FP. This reduction in precision may be due in part to reduced coverage by RegulonDB at these large distances. Table 2 contains instances of possible operon pairs that are classified as FP pairs due to our definition. The set of DVDA predicted pairs separated by <100 bp has a precision of 95%. The *B.subtilis* 1050 predicted pairs set has 158 TP and 38 FP, a precision of 80%.

These high precision DVDA predictions are suitable to infer an intergenic distance distribution for operon pairs that is almost identical to the true (RegulonDB) distribution in the case of *E.coli*. Figure 4 illustrates the similarity of the distributions of distances for DVDA pairs and the known operon set. The agreement is slightly less impressive for *B.subtilis* than *E.coli*, this may be due to a distance bias in the smaller *B.subtilis* operon set. We make the assumption from Equation 1 in Materials and Methods on the basis of the similarity of these densities. This assumption states that the probability of seeing a gene pair separated by distance d, given that the genes belong to an operon, is the same as the probability given that they are a DVDA pair.

Our approach of predicting adjacent operon pairs and estimating an intergenic distance distribution from them was applied to a representative set of eight query genomes (3 Proteobacteria, 4 Firmicutes and 1 Archaea). These results suggest that there are differences between genomes in the distribution of intergenic distances (IGDs) of predicted operon pairs. Figure 5 shows probabilities that a gene pair belongs to a common operon depending on their IGD, as inferred from the DVDA predictions, for several genomes. More specifically, the overall probability $P(O \mid d)$ (dashed) and the probability$P(O \mid d, \bar{G})$ as restricted to non-DVDA pairs (solid) are shown. There is some difference noticeable between these two probabilities. $P(O \mid d)$ represents the probability of being an operon pair for all gene pairs while $P(O \mid d, \bar{G})$ is the probability of being an operon pair for those gene pairs which are not identified by the DVDA method. Since the latter set is enriched in non-operon pairs, a non-DVDA pair is a priori more likely to be a non-operon pair, which shifts the cut-off for an operon pair towards smaller distances.

Table 1 describes the intersection points of the two probability estimates with a 0.5 cut-off, with 95% confidence intervals shown in brackets. It is evident that there are considerable differences in intra-operon distances between genomes. There is no obvious correlation of intra-operon distance with parameters such as number and length of directons, median IGD or size of genome (see the Supplementary Material for some examples of such parameters).

The predicted *E.coli* cut-off values are very close to the values generated when the RegulonDB (known operon set) IGD distribution is used instead of the DVDA (predicted pair set) IGD distribution in the evaluation [$P(O \mid d) = 0.5$ cut-off is 76 bp for RegulonDB compared to 89 bp for DVDA, see Supplementary Material for figures].

Table 1 provides two different estimates $P_1(O)$ and $P_2(O)$ of the proportion of operon pairs in directons (adjacent gene pairs across directon boundaries are not counted). On average about two-thirds of adjacent gene pairs in directons are in a common operon. Again there is some variability for this characteristic between genomes that seems uncorrelated to any obvious genome statistic and suggests significant operon structure differences amongst prokaryotic genomes.

Table 2. Conserved pairs of unknown operon status

Gene pair	IGD (bp)	Comment
queA tgt	56	Possible functional relationship (28)
tolA tolB	133	Operon in *Pseudomonas putida* (29,30)
pflA pflB	192	Part of anaerobically-induced operon (26)
flgL flgK	12	Operon pair in *Borrelia burgdorferi* (31)
plsX fabH	68	Fatty acid biosynthesis in *Escherichia coli* (32)
edd zwf	235	Close in Entner–Doudoroff pathway (26)
tar chew	145	Tar binds CheW (33)
fliA fliC	321	Part of the flagellar functional class (26)
flik fliL	105	Part of the flagellar functional class (26)
rnc lepB	272	Operon in *Rickettsia rickettsii* (34)
hsdM hsdR	201	Host modification–restriction system (26)
lon hupB	209	Lon protease degrades HupB in the absence of HupA
hisJ argT	221	Substrate binding proteins for HisJQMP transporter
lpxD fabZ	105	Pathways linked by common substrate (35–37)
serA rpiA	256	Part of a putative purine regulon (38)

DVDA detects conserved functionally related gene clusters, many of which are known to be co-transcribed. Presented are instances of conserved gene pairs which the literature suggests may belong to common operons or are at least functionally related. IGD refers to the intergenic distance between the pair.

DISCUSSION

Operon prediction with DVDA

The generation of operon maps is becoming an increasingly vital part of prokaryotic genome annotation, and IGD is the most informative predictor for this task. The inclusion of other data improves the accuracy of these predictions only marginally and in any case many of these data are available for only a limited number of genomes. Therefore, the generation of a reliable operon IGD distribution is crucial. We present a flexible operon pair prediction method able to detect permuted gene clusters containing conserved gene pairs separated by large IGD. Further, we show that predicted pairs can be used to generate genome-specific within-operon IGD distributions. From these, whole genome operon maps can be created. While the *E.coli* genome has become the *de facto* source of IGD distribution for prokaryotic operon prediction due to its experimentally validated operon set, our data suggest that transferring the distance distribution from *E.coli* to other genomes for operon prediction is questionable, as indicated

by the variability in genome-specific values seen in Table 1 and Figure 5. The application, for instance, of the *E.coli* IGD cut-off to the *Xylella fastidiosa* genome would result in a large number of false negatives (FNs) (a true operon pair labelled as a TU boundary), as the *E.coli* cut-off is approximately a third of the length of the *X.fastidiosa* cut-off. Given the range of within-operon intergenic distance cut-off values in our small test set of genomes, any operon predictions made under the assumption that all genomes have a similar operon IGD distribution should be treated with caution.

The DVDA protocol differs from other methods in the field of gene context conservation in two important regards. First, in assignment of orthology, sequence similarity and context are balanced. A strong sequence similarity between two genes is not considered enough evidence to assign orthology, unless there is evidence of a conserved functional context. The whole genome is scanned for these contexts and even a bi-directional best hit could be discarded as not being the orthologue if a better context is found for that gene. Second, the conservation methods described in Introduction rely at least in part on conservation of gene order, even though this conservation is known to be very limited. DVDA is completely independent of gene order as the graph representation of the directon to directon comparison contains no information on the order of the genes on the chromosome. DVDA is also independent of IGD and so is able to predict operon pairs at relatively large separation. This property is particularly advantageous since there are verified instances of operon pairs with large intergenic separation but, due to the relative scarcity of these pairs in the RegulonDB training set, these are missed by the original IGD predictor (1). Given the occasionally very large $P(O \mid d)$ and $P(O \mid d, \bar{G})$ distance cut-offs, freedom from any IGD constraint is essential for the prediction of the maximum number of conserved operon pairs.

B.subtilis shows a very large cut-off for operon pairs. This large cut-off is only partly reflected in the known operon set of *B.subtilis* (see section 2 in the Supplementary Material). In fact, this is one of the reasons why we think the *B.subtilis* known operon pairs might not be representative. The prediction method is independent of any training data, excepting ORF prediction.

Two of the genomes examined have large $P(O \mid d, \bar{G})$ distance cut-offs. There are a few verified examples of large separation between *B.subtilis* operon pairs, notably *infA map* (312 bp), *flgE fliL* (252 bp), *yrbA yrbB* (226 bp) and *rocD rocE* (223 bp). Another genome with a large predicted cut-off is *X.fastidiosa*. Comparison with the *Xanthomonas campestis* xanthan gum operon has revealed a syntenic and homologous region in *X.fastidiosa* termed the fastidian gum operon (19). Within this operon, gene pairs are separated by large IGDs: *gumC gumD* (212 bp); *gumF gumH* (411 bp); *gumH gumJ* (377 bp); and *gumJ gumK*(409 bp).

The DVDA method also allows us to estimate the proportion of within-operon to within-directon pairs for the tested organism. There is notable variability for this proportion between genomes (from ~50 to 82%) that seems uncorrelated to any obvious genome statistic. At one extreme, the plant pathogen *X.fastidiosa* is estimated to have ~4 out of 5 of its adjacent within-directon gene pairs belonging to operons. Again, this diversity suggests there is no universal operon distribution among prokaryotes.

Contextual conservation of functionally interacting genes

A conserved operon is not the only mechanism which can result in a persistent gene pair and here we describe some interesting predictions of conservation made by DVDA. These predictions further support DVDA as an accurate predictor of conserved functionally related genes, of which the operon set is the major component. Of the 58 predictions not in the RegulonDB set, 42 have an intergenic separation greater than 100 bp. Eight of these 42 pairs (*lacZ lacI*, 123 bp separation; *pheS pheM* 284 bp; *atoC atoD* 196 bp; *ebgR ebgA* 148 bp; *mglB galS* 280 bp; *uhpT uhpC* 138 bp; *treB treR* 119 bp; and *uxuB uxuR* 215 bp) consist of the leader peptide/inducer/repressor of an operon and the first gene of that operon. Retention of an operon repressor or inducer protein in the vicinity of its cognate binding site makes intuitive sense; the induction or repression mechanism would be more sensitive. Small changes in expression of the regulatory protein would result in a large change in the effective local concentration of the protein. When present, the leader peptide of an operon is necessarily immediately upstream of the operon it regulates. It is the conformation of the transcript of the leader peptide that determines whether transcription continues and the downstream operon is expressed. Leader peptides are usually associated with amino acid synthetic processes; *pheM* is the leader peptide for the *pheST* genes which code for phenylalanyl-tRNA synthetase.

An interesting example of a predicted pair with ambiguous operon status is *hisJ argT* (221 bp separation). Active transport of metabolites in Gram-negative bacteria is facilitated by periplasmic substrate binding proteins. Both products of the conserved *hisJ* and *argT* pair are substrate binding proteins for the His permease (*hisJQMP*); HisJ is specific for L-Histidine while ArgT binds L-Lysine, L-Arginine and L-Ornithine (20,21). Microarray analysis (22) has suggested that both these sets of genes (*argT* and *hisJQMP*) are under the control of the nitrogen regulation system (*ntrBC*), and that an NtrC controlled promoter of the *hisJQMP* operon may be upstream of *argT* (23), i.e. under some conditions, these genes form an operon.

An example of a conserved pair not associated with a known operon, but linked by some regulatory process, is the *lon hupB* predicted pair (209 bp separation). The *hupB* gene product (HU1) forms a histone like protein (HU) in complex with HU2 (product of the *hupA* gene). HU1 is degraded by the Lon protease in the absence of HU2 (24). During the log growth phase of *E.coli*, HU is a HU2 homodimer, whilst during stationary phase HU is a HU1–HU2 heterodimer. Claret and Rouviere-Yaniv (25) hypothesized that Lon inhibits the formation of HU1 homodimers in the absence of HU2. The HU1 homodimer lacks the necessary activity observed in the two other dimer species. This interaction is a clear case for the retention of *lon* in the vicinity of *hupB*. The intergenic region between *lon* and *hupB* in *E.coli* encompasses four promoters, three FIS binding sites (which extend more than 200 bp upstream of the *hupB* start codon) and a CRP binding site [EcoCyc, (26)], suggesting an explanation of the large separation.

Further examples of DVDA predicted pairs that are not found in the RegulonDB set are presented in Table 2. Although

these exceptions slightly distort the prediction of distances in operon pairs, they seem to be rare enough not to invalidate our approach and support DVDA as an accurate predictor of conserved functional interaction.

It is difficult to make direct comparisons to existing gene conservation methods since databases are continuously growing and methods improved. For example, the method published by Ermolaeva *et al*. (9) was originally based on only 34 genomes. While the results currently available from the corresponding website (http://www.tigr.org/tigr-scripts/operons/operons.cgi) are based on a larger set of genomes, these predictions have been produced by a modified, but as yet unpublished, method. One way to compare the results of the current (February 2005) version of the Ermolaeva method to our method is to choose a confidence threshold for the Ermolaeva predictions so that each method produces approximately the same precision. Choosing a *P*-value of 0.6 results in 670 putative operon pairs of which 301 are TP and 49 are FN, i.e. a precision of 86.0% and a recovery of 44.3% based on the RegulonDB dataset of 680 operon pairs. This might be compared with a precision of 85.2% and a recovery of 49.1% of the 854 pairs predicted by the DVDA method on the same set.

Assumptions made in the method

We assume in general that a recent gene duplication (i.e. where the paralogous genes are still adjacent or at least in the same directon) are likely to remain within the same TU. There is no obvious way to assess at what point a paralogue has diverged sufficiently, in its regulatory or catalytic properties, to represent a new gene function with an optimal transcription level different enough to its sibling to justify, in terms of efficiency, the creation or modification of a separate operon.

As with all distance-based operon prediction methods, precise IGDs are dependent on the accuracy of annotation of translational start sites. It could be argued that annotation of hypothetical genes is less certain than of those with defined function. In order to test the robustness of DVDA, we re-analysed data for genomes with more than 50% of non-hypothetical genes (*E.coli*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *X.fastidiosa*), removing all clusters containing a hypothetical gene. As shown in Table 1, the estimates of cut-offs for operon pairs are very similar and well within the sampling error as indicated by the bootstrap confidence intervals (see the Supplementary Material for plots of logistic regressions comparable to the ones in Figure 5). Though the genome-specific quality of ORF predictions makes comparison of genomic characteristics difficult, the examples presented are in some cases different enough that the effect is unlikely to be entirely due to methodological artifact.

The detection of non-operon pairs that are still functionally related and conserved in proximity is a problem affecting all prediction methods based on context. The two most obvious causes of such conservation are über-operons (13) (collections of operons retaining proximity in unrelated organisms) and operon regulatory genes. Our results show instances of the second case, and these may push the 0.5 confidence threshold to larger IGDs. Presumably, this factor affects all genomes, and will not unduly impact cross-genome comparison.

Further application of DVDA

DVDA uses bipartite matching to decide which single region from a genome *A* is most similar to a single region in a genome *B* on the basis of overall homology rather than a few strong sequence similarities between the regions. The matched pairs of genes are probably true orthologues, an assignment supported not only by their sequence similarity but also by their conserved functional context. A family of orthologues could be created by collecting all the matchings for a single directon and clustering the genes by which (for example) *E.coli* gene they match to. It would be interesting to compare these families of orthologous proteins to the COGs database (27) and investigate any inconsistency.

CONCLUSION

In this paper, we describe a comparative genomics method for predicting gene pairs whose functional relationship depends on spatial vicinity. The majority of these pairs belong to operons. The method is based on the comparison of homologous genes across genomes and uses advanced graph algorithmic and statistical methods. No training data are required so the method can be applied to any prokaryotic genome almost immediately after it has been sequenced. In contrast to the modest input requirements (an ORF prediction), the output from applying the method is extensive: a complete genome operon map; a collection of context-assisted orthologue assignments from a broad range of unrelated organisms; and instances where functional assignments can be made on the basis of this orthology with already annotated genes. The method is independent of any conservation of gene order and will accept any number of gaps in the alignment between query and target gene clusters. Prediction of conservation is independent of IGD, allowing DVDA to identify conserved gene pairs at arbitrarily large separation. Due to the comparative nature of DVDA, as more genomes become available prediction quantity and quality will increase. DVDA predictions for the eight genomes presented and supplementary information are available online (http://dvda.cryst.bbk.ac.uk).

ACKNOWLEDGEMENTS

Thanks are due to the reviewers whose comments were appreciated. M.T.E. acknowledges a studentship by BBSRC and dedicates this work to the memory of Kenneth Peter Packwood. S.C.G.R. was funded by the Wellcome Trust (Grant 062508). L.W. acknowledges JISC and that funding to pay the Open Access publication charges for this article was provided by a Wellcome Trust Functional Genomics grant.

Conflict of interest statement. None declared.

REFERENCES

1. Salgado,H., Moreno-Hagelsieb,G., Smith,T. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
2. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329–S336.
3. Strong,M., Mallick,P., Pellegrini,M., Thompson,M. and 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**, R59.1–R59.16.
4. Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, 20–23 August, La Jolla, California, USA, Vol. 8, pp. 116–127.
5. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Daz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimnez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martnez,C. and Collado-Vides,J. (2004) Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* k-12. *Nucleic Acids Res*., **32**, D307–D310.
6. Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F. and Craven,M. (2003) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19**, i34–i43.
7. Paredes,C., Rigoutsos,I. and Papoutsakis,E. (2004) Transcriptional organization of the *Clostridium acetobutylicum* genome. *Nucleic Acids Res*., **32**, 1973–1981.
8. Mushegian,A. and Koonin,E. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet*., **12**, 289–290.
9. Ermolaeva,M., White,O. and Salzberg,S. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res*., **29**, 1216–1221.
10. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
11. Wolf,Y., Rogozin,I., Kondrashov,A. and Koonin,E. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res*., **11**, 356–372.
12. Rogozin,I., Makarova,K., Murvai,J., Czabarka,E., Wolf,Y., Tatusov,R., Szekely,L. and Koonin,E. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*., **30**, 2212–2223.
13. Lathe,W.C.,III, Snel,B. and Bork,P. (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci*., **25**, 474–479.
14. Zheng,Y., Roberts,R.J. and Kasif,S. (2003) Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol*., **3**, RESEARCH0060.1–0060.9.
15. Altschul,S.F., Madden,T.L., Schaffer,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.
16. Gibbons,A. (1985) *Algorithmic Graph Theory*. Cambridge University Press, Cambridge, UK.
17. Efron,B., Tibshirani,R., Storey,J.D. and Tusher,V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*., **96**, 1151–1160.
18. Wood,S.N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. B*, **62**, 413–428.
19. da Silva,F.R., Vettore,A.L., Kemper,E.L., Leite,A. and Arruda,P. (2001) Fastidian gum: the *Xylella fastidiosa* exopolysaccharide possibly involved in bacterial pathogenicity. *FEMS Microbiol. Lett*., **203**, 165–171.
20. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
21. Lodish,H., Berk,A., Zipursky,S., Matsudaira,P., Baltimore,P. and Darnell,J. (2000) *Molecular Cell Biology. 4th edn* W.H.Freeman and Company, NY.
22. Zimmer,D., Soupene,E., Lee,H., Wendisch,V., Khodursky,A., Peter,B., Bender,R. and Kustu,S. (2000) Nitrogen regulatory protein c-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. *Proc. Natl Acad. Sci. USA*, **97**, 14674–14679.
23. Schmitz,G., Durre,P., Mullenbach,G. and Ames,G.F. (1987) Nitrogen regulation of transport operons—analysis of promoters argTr and dhuA. *Mol. Gen. Genet*., **209**, 403–407.
24. Bonnefoy,E., Almeida,A. and Rouviere-Yaniv,J. (1989) Lon-dependent regulation of the DNA binding protein HU in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **86**, 7691–7695.
25. Claret,L. and Rouviere-Yaniv,J. (1997) Variation in HU composition during growth of *Escherichia coli*: the heterodimer is required for -long term survival. *J. Mol. Biol*., **273**, 93–104.
26. Karp,P., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res*., **30**, 56–58.
27. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
28. Iwata-Reuyl,D. (2003) Biosynthesis of the 7-deazaguanosine hypermodified nucleosides of transfer RNA. *Bioorg. Chem*., **31**, 24–43.
29. Llamas,M.A., Ramos,J.L. and Rodriguez-Herva,J.J. (2003) Transcriptional organization of the *Pseudomonas putida* tol-oprL genes. *J. Bacteriol*., **185**, 184–195.
30. Dubuisson,J., Vianney,A. and Lazzaroni,J. (2002) Mutational analysis of the TolA C-terminal domain of *Escherichia coli* and genetic evidence for an interaction between TolA and TolB. *J. Bacteriol*., **184**, 4620–4625.
31. Ge,Y., Old,I., Girons,I. and Charon,N. (1997) The flgK motility operon of *Borrelia burgdorferi* is initiated by a sigma 70-like promoter. *Microbiology*, **143**, 1681–1690.
32. Zhang,Y. and Cronan,J. (1998) Transcriptional analysis of essential genes of the *Escherichia coli* fatty acid biosynthesis gene cluster by functional replacement with the analogous *Salmonella typhimurium* gene cluster. *J. Bacteriol*., **180**, 3295–3303.
33. Boukhvalova,M.S., Dahlquist,F.W. and Stewart,R.C. (2002) CheW binding interactions with CheA and Tar. Importance for chemotaxis signaling in *Escherichia coli*. *J. Biol. Chem*., **277**, 22251–22259.
34. Rahman,M.S., Simser,J.A., Macaluso,K.R. and Azad,A.F. (2003) Molecular and functional analysis of the *lepB* gene, encoding a type I signal peptidase from *Rickettsia rickettsii* and *Rickettsia typhi*. *J. Bacteriol*., **185**, 4578–4584.
35. Clements,J.M., Coignard,F., Johnson,I., Chandler,S., Palan,S., Waller,A., Wijkmans,J. and Hunter,M.G. (2002) Antibacterial activities and characterization of novel inhibitors of LpxC. *Antimicrob. Agents Chemother*., **46**, 1793–1799.
36. de Cock,H., Pasveer,M., Tommassen,J. and Bouveret,E. (2001) Identification of phospholipids as new components that assist in the *in vitro* trimerization of a bacterial pore protein. *Eur. J. Biochem*., **268**, 865–875.
37. Kloser,A., Laird,M., Deng,M. and Misra,R. (1998) Modulations in lipid A and phospholipid biosynthesis pathways influence outer membrane protein assembly in *Escherichia coli* K-12. *Mol. Microb*., **27**, 1003–1008.
38. Ravcheev,D., Gelfand,M., Mironov,A. and Rakhmaninova,A. (2002) Purine regulon of gamma-proteobacteria: a detailed description. *Genetika*, **38**, 1203–1214.