Large-scale analysis of gene clustering in bacteria

Qingwu Yang and Sing-Hoi Sze

Genome Res. 2008 18: 949-956 originally published online April 4, 2008
Access the most recent version at doi:10.1101/gr.072322.107

Large-scale analysis of gene clustering in bacteria

Qingwu Yang[1] and Sing-Hoi Sze[1,2,3]

[1]*Department of Computer Science, Texas A&M University, College Station, Texas 77843, USA;* [2]*Department of Biochemistry & Biophysics, Texas A&M University, College Station, Texas 77843, USA*

An important strategy to study operons and their evolution is to investigate clustering of related genes across multiple bacterial genomes. Although existing algorithms are available that can identify gene clusters across two or more genomes, very few algorithms are efficient enough to study gene clusters across hundreds of genomes. We observe that a querying strategy can be used to analyze gene clusters across a large number of genomes and develop an efficient algorithm to identify all related clusters on a genome from a given query cluster. We use this algorithm to study gene clustering in 400 bacterial genomes by starting from a well-characterized list of operons in *Escherichia coli* K12 and perform comparative analysis of operon occurrences, gene orientations, and rearrangements both within and across clusters. We show that important biological insights can be obtained by comparing results across these categories. A software program implementing the algorithm (GCQuery) and supplementary data containing detailed results are available at http://faculty.cs.tamu.edu/shsze/gcquery.

[Supplemental material is available online at www.genome.org.]

In bacteria, one of the main mechanisms to facilitate control of gene expression is the organization of genes into operons, in which a number of algorithms are available for their predictions (Salgado et al. 2000; Price et al. 2005; Che et al. 2006). An important strategy to study operons and their evolution is to investigate clustering of related genes within localized regions across multiple bacterial genomes. Since operon structures can be altered by genome rearrangements (Coenye and Vandamme 2005), it is important to allow the investigation of unrestricted gene clusters that may not correspond to single operons across bacterial genomes. Although existing algorithms are available that can identify gene clusters across two or more genomes, including FISH (Calabrese et al. 2003), GeneTeams (Bergeron et al. 2002; Luc et al. 2003), HomologyTeams (He and Goldwasser 2005), and a generalized algorithm of GeneTeams and HomologyTeams in Kim et al. (2005), very few algorithms are efficient enough to study gene clusters across hundreds of genomes.

To overcome this difficulty, Lee and Sonnhammer (2003) analyzed each genome separately by identifying clusters of genes that belong to the same metabolic pathway and compared the results across a large number of genomes. One drawback of such a strategy is that it is not possible to utilize comparative data during the initial analysis. We observe that the following querying strategy can be used to analyze gene clusters across a large number of genomes. Suppose that a list of clusters is given on one of the genomes. For each given cluster Q, first find the locations of all the related genes on each chromosome c. By considering c as a sequence of genes, the distribution of related genes within any window on c can be modeled by the hypergeometric distribution. The list of windows on c with E-values below a cutoff then give rise to a list of clusters on c. An important advantage of such a querying strategy is that it is possible to obtain very high accuracy by choosing initial clusters that have been experimentally confirmed. We will show that this approach has higher accuracy than general purpose algorithms that do not assume that an initial list of clusters is given.

[3]Corresponding author.
E-mail shsze@cs.tamu.edu; fax (979) 847-8578.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.072322.107.

We study gene clustering in 400 bacterial genomes by starting from a well-characterized list of operons in *Escherichia coli* K12. We first validate our algorithm by performing queries on the well-studied *Bacillus subtilis* subsp. *subtilis* str. 168 genome and within the *E. coli* K12 genome itself, and compare the results to known experimentally verified operons. We then perform comparative analysis of operon occurrences among bacterial groups and study gene orientations within predicted clusters. We also study distributions of rearrangements, both within and across clusters. We show that important biological insights can be obtained by comparing results across these categories and our algorithm is well suited for analyzing gene clusters across a large number of genomes.

Methods

We represent each chromosome c by an ordered sequence of genes $(g_1,g_2,\ldots,g_n)$ while not imposing constraints on gene orientations. Given a query cluster Q, our algorithm GCQuery first identifies the set of all related genes on c. This defines a subsequence $c' = (g'_1,g'_2,\ldots,g'_{n'})$ of c such that each g'_i is related to at least one gene in Q. We think of a gene cluster as a set of genes on c' that are in close proximity to each other while excluding intervening genes from c. We consider each substring $(g'_j,g'_{j+1},\ldots,g'_{j+k'-1})$ on c' between the jth gene and the $(j + k' - 1)$th gene as a potential gene cluster that spans the window $(g_i,g_{i+1},\ldots,g_{i+k-1})$ on c between the ith gene and the $(i + k - 1)$th gene, where $g_i = g'_j$ and $g_{i+k-1} = g'_{j+k'-1}$ (see Figs. 1, 2), and estimate its E-value as follows. The probability of finding such a cluster of size at least k' is given by the hypergeometric distribution as

$$p(n,n',k,k') = \sum_{i=k'}^{k} \binom{n'}{i}\binom{n-n'}{k-i} \Big/ \binom{n}{k}.$$

The expected number of such clusters that span a window of length k on a linear chromosome c is given by the E-value

$$e(n,n',k,k') = (n - k + 1)\, p(n,n',k,k').$$

Algorithm GCQuery(Q,c) {
  $c' \leftarrow$ subsequence $(g'_1, g'_2, \ldots, g'_{n'})$ of $c = (g_1, g_2, \ldots, g_n)$ such that each g'_i is related to at least one gene in Q;
  for $k' \leftarrow 1$ to n' do {
    for $j \leftarrow 1$ to $n' - k' + 1$ do {
      compute $e(n, n', k, k')$ of the cluster $(g'_j, g'_{j+1}, \ldots, g'_{j+k'-1})$ on c' that spans the window $(g_i, g_{i+1}, \ldots, g_{i+k-1})$ on c, where $g_i = g'_j$ and $g_{i+k-1} = g'_{j+k'-1}$; } } }

Figure 1. Algorithm GCQuery to find all related gene clusters on a linear chromosome *c* from a query cluster *Q*. GCQuery is available at http://faculty.cs.tamu.edu/shsze/gcquery.

On a circular chromosome *c*, this *E*-value is given by

$$e(n,n',k,k') = n\, p(n,n',k,k').$$

Note that windows cannot extend beyond the right end on a linear chromosome (see Figs. 1, 2), but they can wrap around on a circular chromosome. Since n and n' are fixed, we precompute and store all the $O(n)$ binomial coefficients. For a fixed k', $p(n,n',k,k')$ can be obtained from $p(n,n',k,k'-1)$ in constant time, and it takes $O(n^2)$ time and space to obtain all the possible *E*-values. Since each cluster $(g'_j, g'_{j+1}, \ldots, g'_{j+k'-1})$ can be obtained from the previous one in constant time by removing g'_{j-1} and adding $g'_{j+k'-1}$, the time complexity of the algorithm is $O(n^2)$.

We use the above algorithm to study the organization of bacterial gene clusters by starting from a list of 123 *E. coli* K12 operons that are experimentally validated and contain at least four genes from the RegulonDB database (Huerta et al. 1998), with protein sequences from the MG1655 strain of *E. coli* K12 (Blattner et al. 1997). We analyze related clusters in 400 completely sequenced bacterial genomes with taxonomy information (Wheeler et al. 2000; see Supplemental Fig. S1 for all results). We follow the classification approach on the NCBI website (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and divide the genomes into 18 groups (Table 1). While *E. coli* K12 belongs to the Gammaproteobacteria class, the classes Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, and Epsilonproteobacteria belong to the Proteobacteria phylum. The largest group contains 98 bacterial genomes, while seven groups contain four or fewer genomes.

We consider a gene in a query cluster to be related to a gene in a genome if their protein–protein BLAST *E*-value is at most 10^{-7} when the set of all genes in the genome is used as a database (Altschul et al. 1990). We consider a predicted cluster to be significant if its *E*-value from the GCQuery algorithm is at most 10^{-5}. As seen below, these *E*-value cutoffs are chosen to obtain a suitable tradeoff between the number of predicted clusters and the accuracy of these clusters. Note that a predicted cluster does not necessarily correspond to a single operon since the orientations of genes within the cluster may not be the same and there may be intervening genes in the cluster.

Results

Gene clusters on *B. subtilis* subsp. *subtilis* str. 168

To validate our algorithm, we compare the results of querying each of the 123 *E. coli* K12 operons on the *B. subtilis* subsp. *subtilis* str. 168 genome to experimentally confirmed operons from the ODB database (Okuda et al. 2006). For each predicted cluster, we evaluate its accuracy with respect to a given operon from the database by computing the s_{min} score that divides the number of genes that appear in both the operon and the cluster by the minimum size of the operon and the cluster. A predicted cluster will have a high s_{min} score if it has significant overlaps with genes in the operon either with respect to the operon (in which case a large portion of the genes in the operon are included) or with respect to itself (in which case a large portion of the genes in the cluster belong to the operon). Since some of the predicted clusters can be extremely large, we also compute the s_{max} score that divides the number of genes that appear in both the operon and the cluster by the maximum size of the operon and the cluster, which provides a discrepancy estimate.

For each of the 123 *E. coli* K12 operons, we consider at most one predicted cluster in *B. subtilis* subsp. *subtilis* str. 168 from GCQuery with the lowest *E*-value that is below the given cutoff. Given a predicted cluster, we evaluate its accuracy with respect to the entire ODB database by finding the maximum values of s_{min} and s_{max} over all operons from the database. We then obtain the average values of s_{min} and s_{max} over the set of all predicted clusters. Note that these evaluations are only approximations since some clusters may correspond to real operons that are not in the ODB database or are not experimentally confirmed yet, and some clusters may correspond to more than one operon or to higher-level organizations such as superoperons.

Table 2 shows the average s_{min} and s_{max} scores with respect to the ODB database over different combinations of the two *E*-value cutoffs and their relationships to the number of predicted clusters. To choose appropriate *E*-value cutoffs, our goal is to ensure that the number of predicted clusters is not too low while having a high average s_{min} score and an acceptable average s_{max} score. Since the results may be different in other bacteria, the two *E*-value cutoffs should not be too relaxed so that related genes represent homology relationships accurately and predicted clusters represent biologically relevant clusters. Table 2 shows that the GCQuery *E*-value cutoff had a large effect on the number of predicted clusters, with a cutoff of 10^{-9} resulting in too few predicted clusters, a cutoff of 10^{-3} being too relaxed, and the results for the cutoff of 10^{-5} being always better than the corresponding results for the cutoff of 10^{-7}. On the other hand, the effect of the BLAST *E*-value cutoff on the number of predicted clusters was much smaller. When the GCQuery *E*-value cutoff is set to 10^{-5}, the average s_{min} score of 0.59 was the highest when the BLAST *E*-value cutoff is set to 10^{-7}, with an acceptable average s_{max} score of 0.31. We will use these *E*-value cutoffs to define

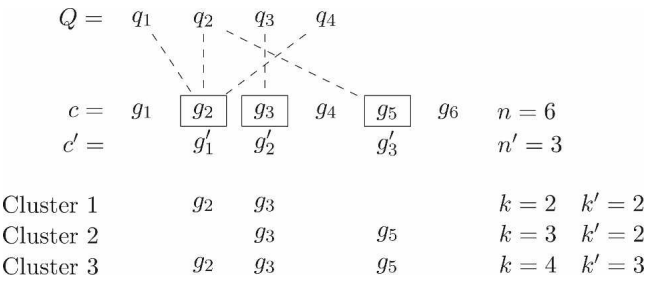

Figure 2. Illustration of all clusters of size >1 on a linear chromosome *c* from a query cluster *Q*. Dashed lines denote related genes. It is possible that each gene in *Q* can be related to more than one gene in *c* and vice versa.

Table 1. Number of genomes in each group and the minimum, maximum, and overall percentage under four categories in each group

		(A) Occurrence rate			(B) Same gene orientation			(C) Neighboring gene pairs			(D) Neighboring cluster pairs		
Group	No.	Min.	Max.	All	Min.	Max.	All	Min.	Max.	All	Min.	Max.	All
Acidobacteria	1	20.3	20.3	20.3	96.0	96.0	96.0	92.3	92.3	92.3	20.0	20.0	20.0
Actinobacteria	34	8.9	27.6	18.5	72.4	100.0	83.6	84.4	98.6	89.6	10.7	54.5	21.2
Alphaproteobacteria	52	4.1	39.8	19.4	67.9	100.0	85.2	80.0	100.0	89.9	7.9	80.0	20.6
Betaproteobacteria	36	17.1	48.8	33.5	82.1	100.0	91.2	91.0	96.9	94.1	7.8	33.3	16.4
Gammaproteobacteria	98	11.4	99.2	46.1	82.1	100.0	94.8	91.2	99.8	97.1	7.7	93.2	42.9
Deltaproteobacteria	14	11.4	29.3	21.9	74.3	100.0	90.2	88.6	100.0	93.0	8.3	42.9	16.4
Epsilonproteobacteria	11	8.9	17.1	13.0	80.0	100.0	94.9	87.1	95.8	90.6	9.5	40.0	23.3
Aquificae	1	7.3	7.3	7.3	100.0	100.0	100.0	82.6	82.6	82.6	11.1	11.1	11.1
Bacteroidetes/Chlorobi	10	8.1	17.9	14.9	94.1	100.0	98.4	80.9	94.4	89.1	17.6	52.9	31.1
Chlamydiae/Verrucomicrobia	11	7.3	15.4	9.2	88.9	100.0	95.2	91.2	100.0	96.5	15.8	41.7	31.5
Chloroflexi	2	13.0	15.4	14.2	84.2	93.8	88.6	86.0	88.5	87.2	12.5	15.8	14.3
Cyanobacteria	19	5.7	15.4	8.9	60.0	100.0	82.6	88.1	100.0	92.2	9.1	50.0	32.4
Deinococcus-Thermus	4	11.4	13.0	12.6	75.0	100.0	87.1	89.9	93.7	91.2	25.0	50.0	35.5
Firmicutes	95	2.4	30.9	14.5	70.8	100.0	93.0	82.5	100.0	92.3	8.3	85.7	24.0
Fusobacteria	1	11.4	11.4	11.4	85.7	85.7	85.7	91.2	91.2	91.2	14.3	14.3	14.3
Planctomycetes	1	11.4	11.4	11.4	100.0	100.0	100.0	90.0	90.0	90.0	14.3	14.3	14.3
Spirochaetes	9	5.7	12.2	9.4	76.9	100.0	93.3	92.2	100.0	95.7	15.4	62.5	30.6
Thermotogae	1	15.4	15.4	15.4	94.7	94.7	94.7	89.5	89.5	89.5	31.6	31.6	31.6
All	400	2.4	99.2	24.8	60.0	100.0	92.0	80.0	100.0	94.4	7.7	93.2	31.3

The minimum and maximum percentages are computed over all genomes in each group, while the overall percentage is computed by dividing the number of entries that satisfy the condition within a category by the total number of entries considered. For each pair of *E. coli* K12 operon and bacterial genome, only one significant cluster with the lowest *E*-value is considered (if it exists). The four categories are as follows (see text for details). (A) Percentage of occurrences of operons that are significant. (B) Percentage of significant clusters in which all genes share the same orientation. (C) Percentage of conserved neighboring gene pairs within significant clusters. (D) Percentage of conserved neighboring cluster pairs.

significant clusters in our later analysis. Note that by choosing *E*-value cutoffs in this manner, we do not distinguish between different types of genes that can have large variations on the number of homologs, which will result in large differences on the number of related genes n' relative to the total number of genes n on a chromosome and should be modeled reasonably well by the hypergeometric distribution.

To show that the above choice gives high accuracy, we compare our performance to GeneTeams (Bergeron et al. 2002; Luc et al. 2003) and HomologyTeams (He and Goldwasser, 2005), which obtain clusters across multiple genomes by restricting the number of intervening genes and the number of base pairs between adjacent genes in a cluster, respectively. We apply these algorithms on the two genomes *E. coli* K12 and *B. subtilis* subsp. *subtilis* str. 168 to obtain clusters that contain more than one gene while fixing the BLAST *E*-value cutoff at 10^{-7} for defining related genes. We use the same procedure as above to compute the average s_{min} and s_{max} scores with respect to the ODB database over the set of all predicted clusters.

Figure 3 shows that the average s_{min} score of GCQuery was always better than the average s_{min} score of GeneTeams or HomologyTeams over different distance cutoffs and was much better when the *E*-value cutoff is 10^{-5}, while the average s_{max} score of GCQuery was comparable to the average s_{max} score of GeneTeams or HomologyTeams when the *E*-value cutoff is around 10^{-5}. The average s_{min} and s_{max} scores of GCQuery both increased when the *E*-value cutoff becomes more stringent, and were higher than 0.9 and 0.4, respectively, when the *E*-value cutoff becomes very stringent, although the number of clusters predicted by GCQuery decreased rapidly as the *E*-value cutoff becomes more stringent (see also Table 2). A caution is that since our algorithm is based on 123 experimentally confirmed *E. coli* K12 operons that contain at least four genes, it can only find clusters that have counterparts in these operons. Thus the number of clusters found by GCQuery was much smaller than GeneTeams or HomologyTeams that do not impose such restrictions.

Gene clusters on *E. coli* K12

To further validate our algorithm, we consider each of the 123 *E. coli* K12 operons and apply GCQuery within its own genome (see Supplemental Fig. S2 for complete results). All but two operons

Table 2. Performance of GCQuery on *B. subtilis* subsp. *subtilis* str. 168 over different combinations of the two *E*-value cutoffs

	BLAST *E*-value cutoff			
GCQuery *E*-value cutoff	10^{-3}	10^{-5}	10^{-7}	10^{-9}
10^{-3}	(55, 0.55, 0.28)	(54, 0.52, 0.26)	(49, 0.56, 0.28)	(48, 0.54, 0.29)
10^{-5}	(34, 0.54, 0.28)	(36, 0.57, 0.31)	(38, 0.59, 0.31)	(35, 0.53, 0.33)
10^{-7}	(24, 0.43, 0.27)	(22, 0.47, 0.30)	(22, 0.47, 0.29)	(22, 0.48, 0.29)
10^{-9}	(14, 0.60, 0.38)	(14, 0.60, 0.38)	(13, 0.57, 0.32)	(11, 0.68, 0.38)

Each entry is a triple (n, s_{min}, s_{max}), where n is the number of predicted clusters, and s_{min} and s_{max} are two average overlap scores with respect to the ODB database (see text for details).

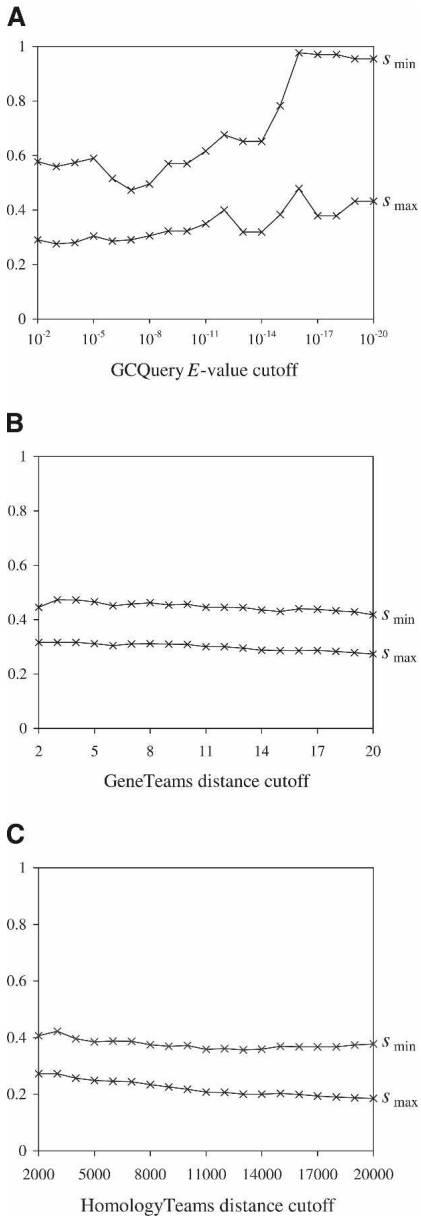

Figure 3. Performance of GCQuery (*A*), GeneTeams (*B*), and HomologyTeams (*C*) on *B. subtilis* subsp. *subtilis* str. 168 while fixing the BLAST *E*-value cutoff to 10^{-7} for defining related genes. For GeneTeams and HomologyTeams, the distance cutoff defines the maximum number of intervening genes and the maximum number of base pairs between adjacent genes in a predicted cluster, respectively.

were found in their entirety within significant clusters, with 11 clusters containing a few additional flanking genes, and four of these clusters containing flanking genes in both orientations. In the *thrLABC* operon, which is involved in threonine biosynthesis (Cossart et al. 1981), the gene *thrL* was missing since its protein contains only 21 amino acids and is difficult to detect with a low BLAST *E*-value. The five-component operon *argT-hisJQMP*, which encodes a transport system that is specific for the uptake of arginine (Wissenbach et al. 1995), was missed since the large number of homologs found for its genes lead to clusters that are not significant.

The operons *narGHJI* and *narZYWV* encode the alpha, beta, delta, and gamma units of nitrate reductase 1 and nitrate reductase 2, respectively (Blasco et al. 1990). Both operons were found when starting from either operon. The four-component operon *fixABCX* of the anaerobic carnitine metabolism consists of *fixA*, *fixB*, *fixC*, and *fixX* (Walt and Kahn, 2002). The putative five-component operon *ydiQRST-fadK* contains *ydiQ*, *ydiR*, *ydiS*, *ydiT*, and *ydiD* (*fadK*). The proteins *fixA*, *fixB*, *fixC*, and *fixX* have high sequence similarity with *ydiQ*, *ydiR*, *ydiS*, and *ydiT*, respectively (Campbell et al. 2003). When *fixABCX* is used as a query, both *fixABCX* and a cluster containing *ydiQ*, *ydiR*, *ydiS*, and *ydiT* were found. When *ydiQRST-fadK* is used as a query, both *ydiQRST-fadK* and *fixABCX* were found. These results show that GCQuery can be used to search for multiple occurrences of homologous gene clusters.

In the query results from the type 1 fimbrial operon *fimAI-CDFGH* (Lane et al. 2007), in addition to itself, an additional cluster *ycbQRSUVF* with *E*-value 1.1×10^{-8} was found. Although the genes within this cluster are of unknown function, the cluster strongly resembles the predicted transcription unit *ycbRSTUVF* in Moreno-Hagelsieb and Collado-Vides (2002). This shows that GCQuery can be used to search for potential new operons or gene clusters.

Comparative analysis of bacterial groups

We study the distribution of occurrences of the 123 *E. coli* K12 operons in 18 bacterial groups. We define the occurrence rate in a genome to be the percentage of *E. coli* K12 operons that have a significant predicted cluster in the genome. We define the overall occurrence rate in a group of genomes to be the percentage of pairs of operon and genome in the group that have significant occurrences. Table 1A shows the minimum, maximum, and overall occurrence rates within each group (see Supplemental Table S1 for detailed results). In *Mycoplasma mobile* 163K, only three operons (2.4%) had significant occurrences. In fact, these three operons are the only ones that occurred in all the 400 bacterial genomes (see below). In *E. coli* W3110, only one operon did not have significant occurrences, which is not surprising since it is also one of the *E. coli* K12 strains. In addition to *E. coli* K12 strains, six other *E. coli* strains are completely sequenced in our data set, including *E. coli* 536, *E. coli* APEC O1, *E. coli* CFT073, *E. coli* O157:H7 EDL933, *E. coli* O157:H7 str. Sakai, and *E. coli* UTI89. The occurrence rates for these six *E. coli* strains were 91.9%, 91.1%, 91.1%, 95.1%, 95.1%, and 91.9%, respectively, which indicate that only a small number of *E. coli* operons are not conserved during evolution. The overall occurrence rate in the 400 bacterial genomes was 24.8%. Only two groups had an overall occurrence rate over 24.8%. The highest was the group Gammaproteobacteria, to which *E. coli* K12 belongs, with an overall occurrence rate of 46.1%. The variations in occurrence rates within the group were very high, ranging from 11.4% to 99.2%. Another group was Betaproteobacteria, with an overall occurrence rate of 33.5%.

Comparative analysis of operon occurrences

We study occurrences of the 123 *E. coli* K12 operons in the 400 bacterial genomes. Overall, 106 operons (86.2%) had occurrences in less than half of the 400 bacterial genomes (Fig. 4; see Supplemental Table S2 for detailed results). Only three operons (2.4%)

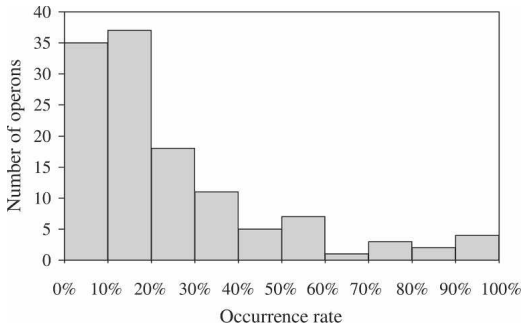

Figure 4. Distribution of the occurrence rate of the 123 *E. coli* K12 operons in the 400 bacterial genomes.

had occurrences in all the 400 bacterial genomes, including *S10*, *spc,* and *alpha*, which have high similarity among bacterial genomes (Watanabe et al. 1997). Coenye and Vandamme (2005) studied the organization of these operons in 99 bacterial genomes and found that many bacterial genomes miss ribosomal proteins that appear in *E. coli*. We investigate the occurrences of genes in these operons in detail.

In the *S10* operon, *rpsJ*, *rplC*, *rplD*, *rplW*, *rplB*, *rpsS*, *rplV*, *rpsC*, *rplP*, *rpmC*, and *rpsQ* encode ribosomal proteins. While *rpsC* and *rplP* occurred within significant clusters in all the 400 bacterial genomes, the occurrence rates of *rpsJ*, *rplC*, *rplD*, *rplB*, *rpsS*, *rplV*, and *rpsQ* were all over 97%. The genes *rplW* and *rpmC* were much less conservative with occurrence rates 54.5% and 42.2%, respectively (Fig. 5, left; see Supplemental Table S3 for detailed results). The smallest significant cluster found contains four consecutive genes in *Silicibacter pomeroyi* DSS-3 with an *E*-value of 1.4×10^{-7}, which are orthologs of *rpsJ*, *rplC*, *rplD*, and *rplW*. In *S. pomeroyi* DSS-3, a cluster of seven genes was also found with an *E*-value of 1.8×10^{-13}, which are orthologs of the rest of the operon. Interestingly, these two fragments together are similar to the operon, but there are 1077 intervening genes between them.

In the *spc* operon, *rplN*, *rplX*, *rplE*, *rpsN*, *rpsH*, *rplF*, *rplR*, *rpsE*, *rpmD*, *rplO*, and *rpmJ* encode ribosomal proteins, while *secY* encodes a preprotein translocase membrane subunit. Only *rpsH* occurred within significant clusters in all the 400 bacterial genomes, while the occurrence rates of *rplN*, *rplX*, *rplE*, *rplF*, *rplR*, *rpsE*, and *secY* were all over 95%. The genes *rpsN*, *rpmD*, *rpmJ*, and *rplO* were much less conservative with occurrence rates 87.5%, 69.2%, 65.0%, and 34.0%, respectively (Fig. 5, middle; see Supplemental Table S4 for detailed results).

The *alpha* operon consists of *rpsM*, *rpsK*, *rpsD*, *rpoA*, and *rplQ*. While the three genes *rpsK*, *rpoA*, and *rplQ* occurred within significant clusters in all the 400 bacterial genomes, the occurrence rates of *rpsM* and *rpsD* were 99.5% and 55.2%, respectively (Fig. 5, right; see Supplemental Table S5 for detailed results). The smallest significant cluster found contains three genes in *Magnetospirillum magneticum* AMB-1 with an *E*-value of 1.2×10^{-6}, which are orthologs of *rpsK*, *rpoA*, and *rplQ*. Significant clusters in other bacterial genomes contain at least four genes.

The operon with the lowest occurrence rate (2.2%) in the 400 bacterial genomes was the *frl* operon, whose genes are mainly involved in fructoselysine related metabolism (Wiame and Van Schaftingen, 2004). The operon occurred in nine bacterial genomes only: except for *C. acetobutylicum* ATCC 824 which belongs to Firmicutes, the other genomes belong to Gammaproteobacteria (see Supplemental Table S6 for detailed results).

The above results indicate that although very few *E. coli* K12 operons are shared by all the 400 bacterial genomes, the counterparts of most operons can be found in many bacteria. The GCQuery algorithm allows the evaluation of various hypotheses concerning evolution and conservation of gene clusters.

Gene orientations within clusters

One of the most important characteristics of an operon is that all genes are transcribed in the same direction. Since the GCQuery algorithm does not require that genes in a cluster must have the same orientation, the fact that all genes within a predicted cluster have the same orientation provides additional evidence that it is probably an operon.

We study the distribution of gene orientations within these clusters. With the requirement that at most one significant cluster with the lowest *E*-value is considered for each pair of *E. coli* K12 operon and bacterial genome, there were a total of 12,221 significant clusters. We define the percentage of clusters in which all genes share the same orientation in a genome only among these significant clusters. We define the overall percentage in a group of genomes to be the percentage over all significant clusters in the group of genomes. Within 92.0% of the 12,221 clusters, all genes share the same orientation, which indicates that gene orientations are highly preserved (Table 1B; see Supplemental Table S7 for detailed results). Within 14 groups (77.8%), one or more bacterial genomes had clusters in which all genes share the same orientation. In fact, in 103 bacterial genomes (25.8%), all clusters contain only one gene orientation.

Among the bacterial genomes, *Synechococcus sp*. CC9311 had the lowest percentage (60.0%) of clusters that contain only one gene orientation. While there were a total of nine predicted clusters in *Synechococcus sp*. CC9311, three of them had genes with different orientations. We investigate these orientation differences in detail. The operon *atpIBEFHAGDC* consists of nine genes that are subunits of ATP synthases (Kasimoglu et al. 1996).

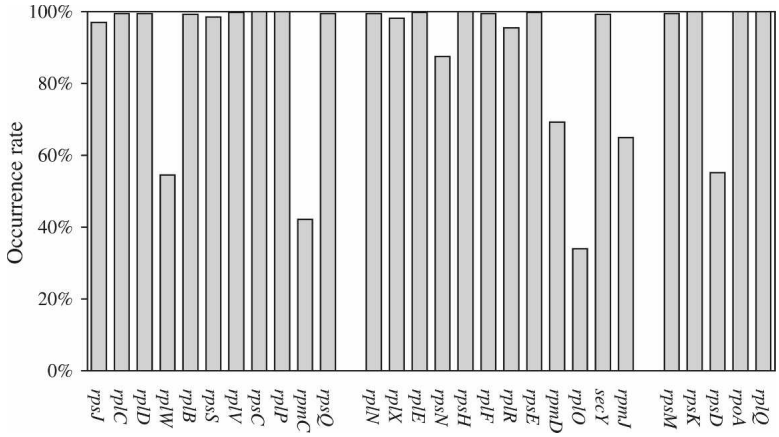

Figure 5. Distribution of the occurrence rate of genes within significant clusters in *S10*, *spc*, and *alpha* operons in the 400 bacterial genomes.

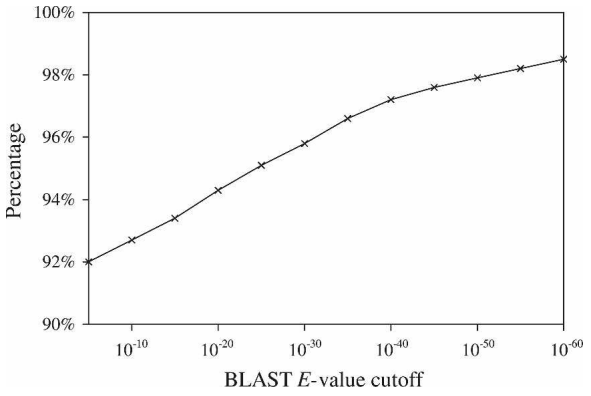

Figure 6. Percentage of clusters in which all genes share the same orientation for different BLAST *E*-value cutoffs.

In *Synechococcus sp*. CC9311, the predicted cluster consists of six genes that are separated into two parts with 26 intervening genes between them, with homologs of *atpD* and *atpC* in a different orientation from the other genes that are homologs of *atpB*, *atpH*, *atpA*, and *atpG*. The operon *cyoABCDE* encodes the cytochrome *o* oxidase complex (Cotter et al. 1990). In *Synechococcus sp*. CC9311, a cluster of four genes was found that include homologs of *cyoA*, *cyoB*, *cyoC*, and *cyoE*, with the homolog of *cyoE* in a different orientation from the other genes. The operon *moaABCDE* consists of genes that are involved in molybdopterin synthesis (Tao et al. 2005). In *Synechococcus sp*. CC9311, a cluster of four genes was found that include homologs of *moaA*, *moaC*, *moaE*, and *moaB* in order, with homologs of *moaE* and *moaA* in a different orientation from homologs of *moaC* and *moaB*.

We are also interested in the effect of varying the BLAST *E*-value cutoff that defines related genes on gene orientations. As the cutoff decreases from 10^{-5} to 10^{-60}, the total number of related genes decreases and the overall percentage of clusters that contain only one gene orientation increases from 92.0% to 98.5% (Fig. 6). This shows that a more stringent requirement can affect the results.

Rearrangements within clusters

In addition to common gene orientations, the spatial arrangement of genes within bacterial operons is important for function, expression, and regulation of these genes (Itoh et al. 1999; Tamames 2001). We study the distribution of gene order within the 12,221 clusters described above. For a given pair of *E. coli* K12 operon Q and predicted cluster C and a set of correspondences with each of them linking a gene in Q to a related gene in C, we obtain a subset of one-to-one corresponding pairs of links as follows: if there is more than one link for a gene in Q, only retain the one with the lowest BLAST *E*-value. After this step, if there is more than one link for a related gene in C, only retain the one with the lowest BLAST *E*-value. In the remaining set of k genes in Q and k related genes in C, assign a label from +1 to +k to each gene in Q according to the order of genes in *E. coli* K12, then assign the corresponding label for each related gene in C, while giving it a + direction if the related gene is on the forward strand and a − direction if the related gene is on the reverse strand. The sequence of k genes in C then corresponds to a signed permutation, in which each neighboring gene pair in C with labels l_1 and l_2 is considered to be a breakpoint if l_1 and l_2 are not consecutive, that is, $|l_1 - l_2| \neq 1$ (Kececioglu and Sankoff, 1995). We define the percentage of conserved neighboring gene pairs to be the total number of neighboring gene pairs that are not breakpoints divided by $k - 1$ (which is the total number of neighboring gene pairs), and use it to evaluate the degree of conservation of gene order. Among the 12,221 clusters, 84.8% of them had perfectly conserved neighboring gene pairs (Fig. 7), which means that the gene order within *E. coli* K12 and the gene order within each cluster are the same either in the forward or in the reverse direction.

When more than one pair of operon and predicted cluster are considered together, we define the overall percentage of conserved neighboring gene pairs to be the total number of neighboring gene pairs that are not breakpoints over all pairs of operon and predicted cluster divided by the total number of neighboring gene pairs over all pairs of operon and predicted cluster. Table 1C shows that the conservation of neighboring gene pairs was in general very high (see Supplemental Table S8 for detailed results). Among all the neighboring gene pairs, 94.4% of them were conserved. Only three groups had overall percentage over 94.4%, including Gammaproteobacteria (97.1%), Chlamydiae/Verrucomicrobia (96.5%), and Spirochaetes (95.7%). Among the bacterial genomes, *Sinorhizobium meliloti* 1021 had the lowest overall percentage (80.0%), with 13 clusters having non-conserved neighboring gene pairs among a total of 29 clusters.

Although gene order within operons can be unstable (Itoh et al. 1999), our results on gene orientation and gene order indicate that predicted clusters tend to contain only one gene orientation and the gene order tends to be conserved.

Rearrangements across clusters

While most previous approaches in analyzing genome rearrangements consider each gene as a basic unit (Kececioglu and Sankoff 1995), we study genome rearrangements also at the level of gene clusters. For each bacterial genome, we collect at most one significant cluster with the lowest *E*-value for each *E. coli* K12 operon. For each chromosome c, assign a label from 1 to k for each operon that has a significant cluster on c according to the order of operons in *E. coli* K12. Then assign the corresponding label to each significant cluster on c according to the starting position of the window that the cluster occupies. If the starting window positions of two clusters are the same, then order the clusters according to the ending window positions. Since our results indicated that none of the predicted clusters from different operons

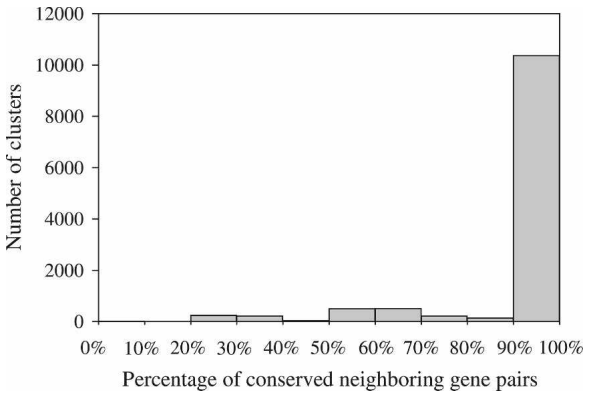

Figure 7. Distribution of the percentage of conserved neighboring gene pairs over all clusters.

occupy exactly the same window, this procedure always breaks a tie. The sequence of k clusters on c then corresponds to an unsigned permutation in which each neighboring cluster pair on c with labels l_1 and l_2 is considered to be a breakpoint if l_1 and l_2 are not consecutive, that is, $|l_1 - l_2| \neq 1$ (Kececioglu and Sankoff, 1995). We define the percentage of conserved neighboring cluster pairs to be the total number of neighboring cluster pairs that are not breakpoints divided by k (which is the total number of neighboring cluster pairs on a circular chromosome).

When more than one chromosome are considered together, we define the overall percentage of conserved neighboring cluster pairs to be the total number of neighboring cluster pairs that are not breakpoints over all chromosomes divided by the total number of neighboring cluster pairs over all chromosomes. Among all the neighboring cluster pairs, only 31.3% of them were conserved (Table 1D; see Supplemental Table S9 for detailed results). Among all groups, Gammaproteobacteria, to which *E. coli* K12 belongs, had the highest overall percentage (42.9%). The percentage was the highest in *E. coli* O157:H7 EDL933 (93.2%), which belongs to Gammaproteobacteria, with eight nonconserved neighboring cluster pairs among a total of 118 neighboring cluster pairs. Interestingly, *Psychrobacter cryohalolentis* K5, which had the lowest percentage (7.7%), is also in this group.

When compared to rearrangements within clusters, our results indicate that large-scale rearrangements at the level of clusters are much more pronounced, although the degree of conservation can still be very high among closely related bacterial genomes.

Comparisons across categories

To investigate whether there are special relationships across the above categories, we consider the set of 400 percentage values, one for each bacterial genome, for each category in Table 1, and compute the Pearson correlation coefficient between each pair of categories (Table 3). While there were significant positive correlations between almost all category pairs, there was no significant correlation between the frequency of operon occurrences in different bacterial genomes and the frequency of clusters in which all genes share the same orientation (with a P-value of 0.47). Since our algorithm does not impose constraints on gene orientations, we can conclude that there is a strong force to preserve gene orientations in bacterial clusters, which is a prerequisite for functioning as operons. Although operon structures can be easily destroyed by genome rearrangements (Coenye and Vandamme, 2005), the force to preserve the remaining operon structures as a unit at the operon level instead of at the gene level does not become weaker even when operon occurrences decrease with respect to *E. coli*, which is not expected in a random model in which operon structures are not involved.

We also observe that there were much stronger correlations between the frequency of clusters that contain only one gene orientation and the frequency of conserved neighboring gene pairs or between the conservation of the two neighboring relationships (with P-values $< 10^{-20}$) than between the rate of operon occurrences and the two neighboring relationships (with P-values $> 10^{-4}$). Thus the degree of preservation of operon structures is a much stronger factor in determining the conservation of neighboring gene pairs than the rate of operon occurrences with respect to *E. coli*.

Table 3. Pearson correlation coefficient of the 400 percentage values, one for each bacterial genome, between each pair of categories in Table 1 (with *P*-value in parentheses)

	(B)	(C)	(D)
(A)	0.036 (0.47)	0.18 (2×10^{-4})	0.16 (1×10^{-3})
(B)		0.45 (5×10^{-21})	0.20 (5×10^{-5})
(C)			0.47 (3×10^{-23})

The four categories are as follows (see text for details): (A) Percentage of occurrences of operons that are significant; (B) percentage of significant clusters in which all genes share the same orientation; (C) percentage of conserved neighboring gene pairs within significant clusters; (D) percentage of conserved neighboring cluster pairs.

Discussion

We have demonstrated that our querying strategy is well suited for analyzing gene clusters across a large number of genomes and important biological insights can be obtained from such analysis. Due to the speed of our algorithm, we were able to obtain all the results on 400 bacterial genomes in less than one day, which includes the time for performing BLAST from genes in the 123 *E. coli* K12 operons on all the 400 bacterial genomes. Since our algorithm does not make any assumptions on the orientation or the density of genes within clusters and only requires that genes in a significant cluster are closer together than expected by chance, it can also be used for analyzing gene clusters on higher organisms such as yeast. Other than using BLAST to establish relations between genes or proteins, it is also possible to use existing databases on homology relationships, such as COG (Tatusov et al. 1997) or INPARANOID (Remm et al. 2001), to identify related genes that are orthologs or paralogs. Since about two-thirds of the computation time of our algorithm was spent on BLAST, the use of such precomputed information can lead to a significant speedup.

Acknowledgments

We thank the anonymous reviewers for invaluable comments that significantly improved the paper. This work was supported by NSF grant DBI-0624077.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Bergeron, A., Corteel, S., and Raffinot, M. 2002. The algorithmic of gene teams. *Lect. Notes in Comput. Sci.* **2452:** 464–476.

Blasco, F., Iobbi, C., Ratouchniak, J., Bonnefoy, V., and Chippaux, M. 1990. Nitrate reductases of *Escherichia coli*: Sequence of the second nitrate reductase and comparison with that encoded by the *narGHJI* operon. *Mol. Gen. Genet.* **222:** 104–111.

Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Calabrese, P.P., Chakravarty, S., and Vision, T.J. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19:** SI74–SI80.

Campbell, J.W., Morgan-Kiss, R.M., and Cronan Jr., J.E. 2003. A new *Escherichia coli* metabolic competency: Growth on fatty acids by a novel anaerobic β-oxidation pathway. *Mol. Microbiol.* **47:** 793–805.

Che, D., Li, G., Mao, F., Wu, H., and Xu, Y. 2006. Detecting uber-operons in prokaryotic genomes. *Nucleic Acids Res.* **34:** 2418–2427.

Coenye, T. and Vandamme, P. 2005. Organisation of the *S10, spc* and *alpha* ribosomal protein gene clusters in prokaryotic genomes. *FEMS Microbiol. Lett.* **242:** 117–126.

Cossart, P., Katinka, M., and Yaniv, M. 1981. Nucleotide sequence of the *thrB* gene of *E. coli*, and its two adjacent regions; The *thrAB* and *thrBC* junctions. *Nucleic Acids Res.* **9:** 339–347.

Cotter, P.A., Chepuri, V., Gennis, R.B., and Gunsalus, R.P. 1990. Cytochrome *o* (*cyoABCDE*) and *d* (*cydAB*) oxidase gene expression in *Escherichia coli* is regulated by oxygen, pH, and the *fnr* gene product. *J. Bacteriol.* **172:** 6333–6338.

He, X. and Goldwasser, M.H. 2005. Identifying conserved gene clusters in the presence of homology families. *J. Comput. Biol.* **12:** 638–656.

Huerta, A.M., Salgado, H., Thieffry, D., and Collado-Vides, J. 1998. RegulonDB: A database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.* **26:** 55–59.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16:** 332–346.

Kasimoglu, E., Park, S.-J., Malek, J., Tseng, C.P., and Gunsalus, R.P. 1996. Transcriptional regulation of the proton-translocating ATPase (*atpIBEFHAGDC*) operon of *Escherichia coli*: Control by cell growth rate. *J. Bacteriol.* **178:** 5563–5567.

Kececioglu, J. and Sankoff, D. 1995. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica* **13:** 180–210.

Kim, S., Choi, J.H., and Yang, J. 2005. Gene teams with relaxed proximity constraint. *Proc. IEEE Comp. Sys. Bioinformatics Conf.* **2005:** 44–55.

Lane, M.C., Simms, A.N., and Mobley, H.L.T. 2007. Complex interplay between type 1 fimbrial expression and flagellum-mediated motility of uropathogenic *Escherichia coli*. *J. Bacteriol.* **189:** 5523–5533.

Lee, J.M. and Sonnhammer, E.L.L. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13:** 875–882.

Luc, N., Risler, J.-L., Bergeron, A., and Raffinot, M. 2003. Gene teams: A new formalization of gene clusters for comparative genomics. *Comput. Biol. Chem.* **27:** 59–67.

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

Okuda, S., Katayama, T., Kawashima, S., Goto, S., and Kanehisa, M. 2006. ODB: A database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res.* **34:** D358–D362.

Price, M.N., Huang, K.H., Alm, E.J., and Arkin, A.P. 2005. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* **33:** 880–892.

Remm, M., Storm, C.E.V., and Sonnhammer, E.L.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314:** 1041–1052.

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.

Tamames, J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* **2:** 1–11.

Tao, H., Hasona, A., Do, P.M., Ingram, L.O., and Shanmugam, K.T. 2005. Global gene expression analysis revealed an unsuspected *deo* operon under the control of molybdate sensor, ModE protein, in *Escherichia coli*. *Arch. Microbiol.* **184:** 225–233.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Walt, A. and Kahn, M.L. 2002. The *fixA* and *fixB* genes are necessary for anaerobic carnitine reduction in *Escherichia coli*. *J. Bacteriol.* **184:** 4044–4047.

Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. 1997. Genome plasticity as a paradigm of eubacteria evolution. *J. Mol. Evol.* **44:** S57–S64.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., and Rapp, B.A. 2000. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **28:** 10–14.

Wiame, E. and Van Schaftingen, E. 2004. Fructoselysine 3-epimerase, an enzyme involved in the metabolism of the unusual Amadori compound psicoselysine in *Escherichia coli*. *Biochem. J.* **378:** 1047–1052.

Wissenbach, U., Six, S., Bongaerts, J., Ternes, D., Steinwachs, S., and Unden, G. 1995. A third periplasmic transport system for L-arginine in *Escherichia coli*: Molecular characterization of the *artPIQMJ* genes, arginine binding and transport. *Mol. Microbiol.* **17:** 675–686.

Received October 13, 2007; accepted in revised form March 13, 2008.