Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome

Yanbin Yin[a,b,1], Han Zhang[a,d,1], Victor Olman[a], and Ying Xu[a,b,c,2]

[a]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30605; [b]Department of Energy BioEnergy Science Center (BESC), Oak Ridge, TN 37831; [c]College of Computer Science and Technology, Jilin University, Changchun, Jilin, China; and [d]Department of Automation, Nankai University, Tianjin, China

Edited* by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved March 2, 2010 (received for review September 29, 2009)

It is generally known that bacterial genes working in the same biological pathways tend to group into operons, possibly to facilitate cotranscription and to provide stoichiometry. However, very little is understood about what may determine the global arrangement of bacterial genes in a genome beyond the operon level. Here we present evidence that the global arrangement of operons in a bacterial genome is largely influenced by the tendency that a bacterium keeps its operons encoding the same biological pathway in nearby genomic locations, and by the tendency to keep operons involved in multiple pathways in locations close to the other members of their participating pathways. We also observed that the activation frequencies of pathways also influence the genomic locations of their encoding operons, tending to have operons of the more frequently activated pathways more tightly clustered together. We have quantitatively assessed the influences on the global genomic arrangement of operons by different factors. We found that the current arrangements of operons in most of the bacterial genomes we studied tend to minimize the overall distance between consecutive operons of a same pathway across all pathways encoded in the genome.

bacterial genome | bioinformatics | genome organization | nucleoid | neighboring genes

A fundamental question in studying bacterial genomes is why genes in a genome are sequentially arranged the way they are. Currently we understand that genes encoding the same biological pathways tend to group into operons, possibly to facilitate cotranscription (1–3) and to provide stoichiometry, thanks to the discovery of operons 50 years ago (2, 4). In addition to this important understanding about the local arrangement of genes in a genome, we began to understand some global properties of bacterial genomes. For example, it has been observed that essential genes tend to locate on the leading strand of a bacterial genome (5), one of the two DNA strands going through replication in parallel using different replication mechanisms; and genes of certain functions such as those encoding rRNAs and ribosomal proteins tend to be located close to the origin of replication on the leading genomic strand (6, 7). Some other efforts were made to study gene clustering in a large scale or on particular types of genes (8, 9). It has also been observed that the bacterial chromosomes exhibit periodicities in terms of both gene coexpression (10, 11) and gene coevolution (12) patterns; and this periodicity may be related to the supercoiled domains in the folded structures (i.e., the nucleoid) of a bacterial chromosome (10, 12–15). It was recently speculated that the genomic organization of bacterial genes may be affected and constrained by multiple cellular processes, specifically gene transcription, genome replication, and nucleoid compaction, at both the local and the global levels (16). Still, our overall understanding about the global arrangement of operons in a bacterial genome is very limited and fragmented. Basically we do not yet know what may influence the genomic locations of operons at a genome scale.

Results and Discussion

We have carried out a computational study aiming to reveal factors that may influence the global arrangement of operons in a bacterial genome. Our analysis suggests two possible dominating factors in influencing the global arrangement of operons in a bacterial genome: (*i*) Biological pathways may have constrained where their encoding operons are located in a genome; and (*ii*) the multiple functional roles of individual operons in different pathways also influence where the operons are located. In this study, a pathway refers to a collection of chemical reactions in sequence or in parallel enabled or participated by proteins, which collectively implement a specific biological process, as defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (17). To derive exactly how these factors may have influenced the global arrangement of operons, we have carried out our study on *Escherichia coli* K-12 and *Bacillus subtilis* strain 168, which are the best studied bacteria and have the most amount of experimental data in the public domain, e.g., microarray gene expression data.

We have retrieved all the 317 and 263 well-characterized biological pathways of *E. coli* K-12 and of *B. subtilis* str. 168 from the SEED database (http://www.theseed.org/) (each called a *subsystem* in SEED) (18), respectively, which are encoded by 1,057 and 915 operons, accounting for 41% and 35% of all the known operons [including both experimentally characterized (19, 20) and computationally predicted operons (21)] in the two organisms, respectively (Table S1). We assume that operons in each genome are ordered clockwise starting from the origin of replication in the circular genome (Fig. S1). So the meaning of the *i*th operon in a genome or in a pathway is well defined.

Operons Participating in More Pathways Are Under Stronger Constraints in Their Genomic Locations

We found that at least 40% of the operons in each of the two genomes participate in multiple SEED pathways (Table S2), and the actual number could be substantially higher as more pathways encoded in these genomes are elucidated and considered in our study; hence we reason that the genomic locations of such operons may be constrained by multiple pathways. Our data show that the more pathways in which an operon participates, the more distant the operon's closest (neighboring) operon is, averaged over all the pathways of which it is part, as shown in Fig. 1 *A* and *B*. In addition, Spearman's rank correlation tests were

Author contributions: Y.X. conceived the research; Y.Y., H.Z., and Y.X. designed research; Y.Y. and H.Z. performed research; Y.Y. and H.Z. contributed new reagents/analytic tools; Y.Y., H.Z., and V.O. analyzed data; and Y.Y. and Y.X. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]Y.Y. and H.Z. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: xyn@bmb.uga.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0911237107/DCSupplemental.

performed, which show that in both organisms, there are significant positive correlations (correlation coefficients rho > 0.26, P values <5.0e-16) between the x axis and the y axis in Fig. 1 [i.e., the number of pathways that operons participate and the distance (number of operons)]. This suggests that, as an operon gets involved in more pathways, it gets "pulled" away from its closest operon by the other pathways it participates, indicating that *the genomic location of an operon is indeed influenced by all the pathways relevant to the operon*.

We have also conducted similar analyses on all the well-characterized pathways of *E. coli* K-12 and of *B. subtilis* str. 168 from the KEGG database (17) and the BioCyc database (22, 23), two other popular pathway databases, respectively, totaling 123 KEGG and 280 BioCyc pathways for *E. coli* K-12, and 117 and 335 pathways for *B. subtilis* str. 168, respectively. Highly similar results (Fig. 1 *C–F*) are obtained to those on the SEED pathways. We noted that the sizes of the pathways from these three databases vary substantially, ranging from 1.73 operons per pathway in BioCyc to 3.34 operons in SEED and 6.49 operons in KEGG (see Table S1), suggesting that our observation is a general property of the genomic arrangement of operons imposed by their relevant pathways regardless of the functional scope (small or large) of the pathways.

More Frequently Activated Pathways Have Their Operons More Clustered Together in Genome

Intuitively we would expect that the more frequently activated pathways (transcriptionally) will have their operons more compactly arranged in the genome, based on the observation made above. To check for this quantitatively, we have used the following formula to measure the compactness c_i of each pathway in terms of how spread out its operons are in the genome. Consider a genome that encodes N pathways and the ith pathway is encoded by M_i operons; d_{ij} is the distance between the jth operon and the $[M_i/2]$th operon (i.e., the median operon) in the ith pathway, measured in terms of the number of operons between the two operons (exclusive) in the genome plus one:

$$c_i = \sum_{j=1}^{M_i} d_{ij} \qquad [1]$$

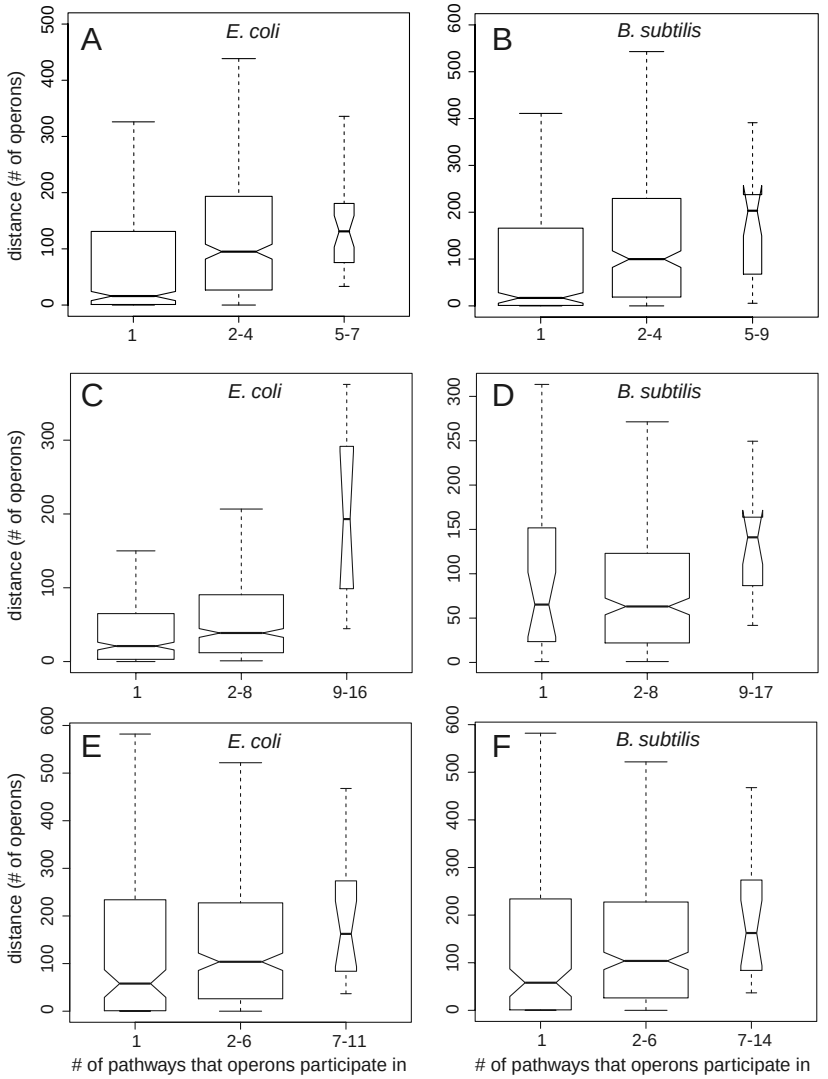

Fig. 1. Box plot of the number of pathways that an operon participates in (x axis) versus the distance between the operon and its closest operon averaged over all the pathways the operon participates in (y axis), measured in terms of the number of operons between the operons. Throughout this paper, the distance between two operons is defined as the number of all operons (not just the ones covered by the pathways under study) between the two operons (see Fig. S1). (*A* and *B*) SEED pathways; (*C* and *D*) KEGG pathways; (*E* and *F*) BioCyc pathways. The boxes are drawn with widths proportional to the square roots of the number of operons in the groups. A notch is drawn in each side of the box toward the median.

Clearly a pathway with a higher c_i value indicates that its operons are more spread out, i.e., less compact.

We have estimated the activation frequency of each pathway, based on the available microarray gene expression data. Specifically, we have used the microarray data for *E. coli* K-12 collected under 380 conditions from the M3D database (24), and the microarray data for *B. subtilis* str. 168 collected under 86 conditions from the KEGG database (17). All data are normalized across different experimental conditions so that the expression data for each gene collected under different conditions can be compared directly (24). We consider a pathway is *activated* if and only if at least $X\%$ of its operons are activated (different columns in Table 1), where an operon is considered activated if and only if its (average) expression value is higher than $Y\%$ of its expression values across all (available) conditions (different rows in Table 1), for parameters X and Y. Then we counted under how many experimental conditions the pathways are activated. Thus the activation frequency is between 0 and 380 for *E. coli* K-12 pathways and between 0 and 86 for *B. subtilis* str. 168. We have tried different X and Y values at 60, 70, and 80, respectively, and calculated the relationships between the activation frequencies of pathways and the compactness of their encoding operons for different X and Y values.

Table 1 summarizes the calculation results, from which we can see that *there is a strong negative correlation between the compactness of pathways and their activation frequencies* for each definition of a pathway being activated (all with P values $<1e\text{-}10$) for the SEED pathways of *E. coli* K-12 and *B. subtilis* str. 168, respectively.

Similar analyses were conducted on the KEGG and the BioCyc pathways of the two organisms. Highly similar results are obtained (Tables S3 and S4), suggesting that this observed relationship is true regardless at what (complexity) level pathways are defined (in a sense, the existing definitions of a pathway, as part of a large cellular network, are somewhat arbitrary), considering that the sizes of the pathways from the three databases span a large spectrum in terms of the number of operons they each cover, with the largest pathway having 76 operons and the smallest having one. Although all the well-characterized pathways cover no more than half of the (known) operons in both *E. coli* K-12 and *B. subtilis* str. 168 (see Table S1), we believe that our observation will continue to hold as more pathways are elucidated for these two organisms.

Table 1. Negative correlation between compactness (*c* values) and activation frequencies of pathways

	60%	70%	80%
	Spearman's rank correlation coefficient rho (*E. coli*)		
60%	−0.47	−0.54	−0.60
70%	−0.52	−0.59	−0.59
80%	−0.67	−0.68	−0.67
	Spearman's rank correlation coefficient rho (*B. subtilis*)		
60%	−0.44	−0.55	−0.58
70%	−0.64	−0.65	−0.61
80%	−0.71	−0.68	−0.63

The first row defines a pathway that is activated if and only if at least $X\%$ of its operons are activated. The first column defines that an operon is considered activated if and only if its expression value is higher than the $Y\%$ quantile value of its expression distribution across all experimental conditions. For all X and Y combinations, the pathway activation frequencies and their c values are analyzed to check if there is a statistically significant linear correlation. The Spearman's rank correlation coefficient rho is reported; all have P values <1e-10. Only pathways with at least two operons are considered.

Biological Pathways Constrain the Global Arrangement of Operons in a Genome

The two observations made above indicate that the global arrangement of operons in a genome is influenced by some global forces. Our additional analysis suggests that a bacterium tends to keep its operons encoding the same biological pathway as tightly clustered together as possible and, at the same time, tends to keep operons involved in multiple pathways in locations as close as possible to the other members of their participating pathways. To make these observations more quantitative, we define the following quantity over all the N (known) pathways encoded in a bacterial genome,

$$C = \sum_{i=1}^{N} c_i, \qquad [2]$$

with c_i representing the compactness of the ith pathway encoded in a genome, as defined in formula **1**. We hypothesize that the current arrangement of operons in a bacterial genome tends to minimize this quantity compared to alternative genomic arrangements of operons in the genome.

To check if this is indeed the case, we have created one-million permutations of the *E. coli* K-12 genome by randomly shuffling $X\%$ of operons encoding the SEED pathways, and then calculating the C value, defined above, for each reshuffled genome, and do this for $X = 10, 20, \ldots, 100$. We use the following two-step procedure to randomly shuffle a specified fraction ($X\%$) of operons. We first randomly select operons among all operons of a bacterial genome for 10,000 times and then randomly permute their locations 100 times for each specific selection of the 10,000. So we do a total of one-million permutations and calculate the C-value distribution over the one-million rearranged genomes. We did the same calculation on *B. subtilis* str. 168. Fig. 2 shows the C-value distributions for different percentages of reshuffled operons for both genomes. We can clearly see that the current genomic arrangement of operons in both genomes have lower C values (the vertical dashed lines) than the vast majority of the C values of the reshuffled genomes (i.e., alternative arrangements of operons in the genomes), respectively. In addition, statistical tests also confirmed that the permutated genomes have significant larger C values than the actual genomes (all P values <0.02, see Table S5). This strongly supports our speculation that bacterial genomes have evolved to minimize the C value (or some variation of the C function).

Fig. 2 also shows that, as a higher percentage of operons have their locations randomly reshuffled, the C value of the resulting rearranged genome goes up more substantially (see Fig. 2 *A* and *B*). Similar observation was made when studying the KEGG and the BioCyc pathways of the two organisms (Figs. S2 and S3). It is even more interesting to note that reshuffling the arrangement of operons that participate in more pathways tends to give rise to more substantial increases in the C value compared to the ones involved in fewer pathways (Fig. 2 *C* and *D*), which is in agreement with our first observation, suggesting that operons participating in more pathways are under stronger selective constraints.

We have also compared the C values of the known pathways of *E. coli* K12 and *B. subtilis* str. 168 with artificial pathways generated through arbitrarily grouping operons (into pathways) for each organism, respectively. Specifically, for each known pathway in *E. coli* K12, we arbitrarily selected the same number of operons from the pool of all operons covered by the known pathways to form an artificial pathway, and do this for every known pathway of *E. coli* K12. We created one-million sets of such artificial (size-matched) pathways and plotted the C-value distributions. We did the similar thing for *B. subtilis* str. 168. The C-value distributions for the one-million sets of artificial pathways are shown in Fig. 2 *C* and *D* as dotted curves, respectively. Again, the C values for the

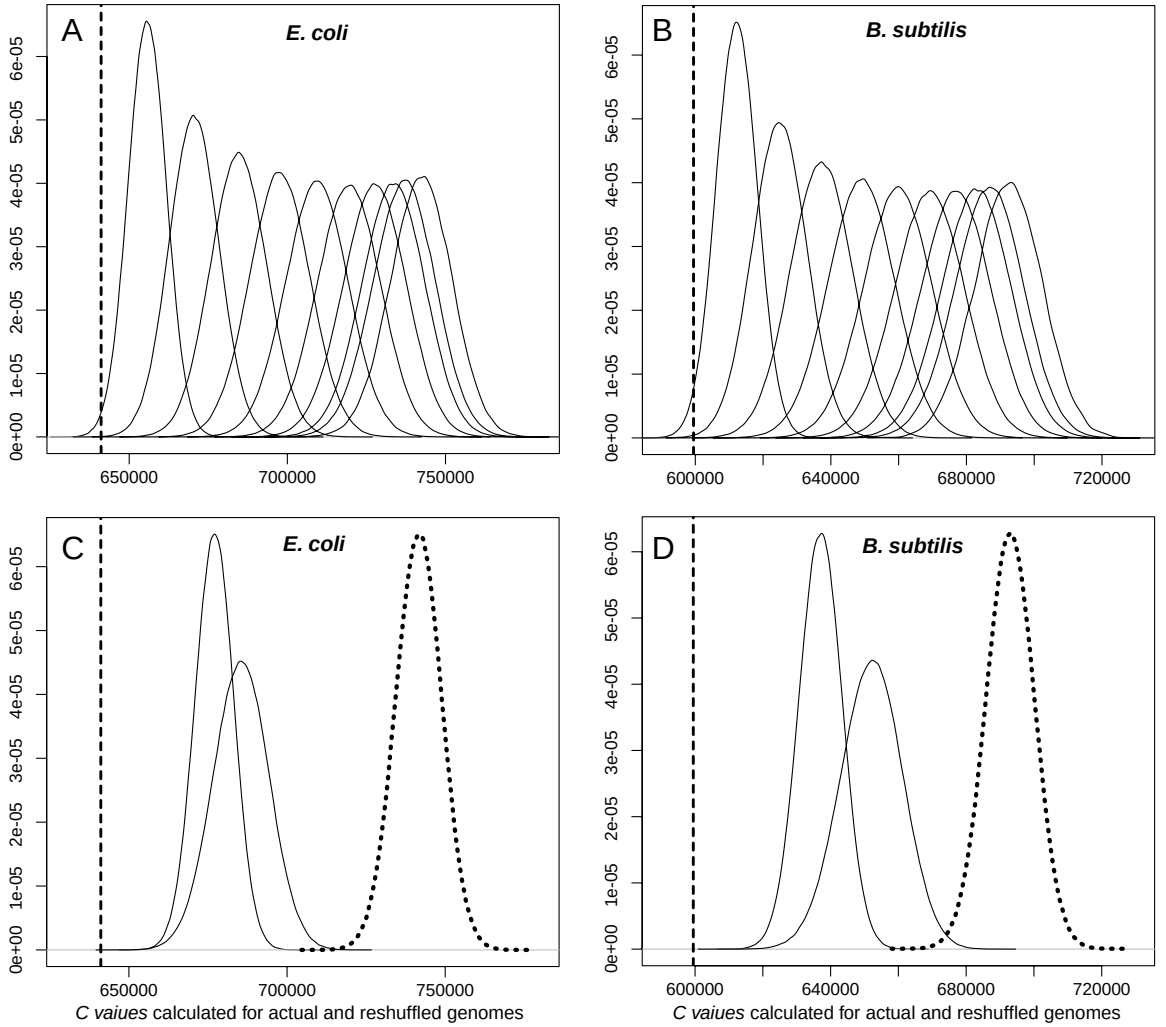

Fig. 2. Distributions of *C* values calculated for actual and reshuffled genomes. In each panel, the *x* axis represents the *C* values (the unit is the number of operons), and the *y* axis is the frequency (density). Each curve is calculated using one-million permutations of the current arrangement of the operons in a genome under a specified constraint. Ten *C* distributions are calculated in (*A*) *E. coli* K-12 and in (*B*) *B. subtilis* str. 168, respectively, with each distribution calculated allowing *X*% of operons randomly selected among all the operons under consideration and being randomly permutated, with $X = 10, 20, \ldots, 100$, respectively, where the 10 curves from left to right in *A* or *B* are consistent with the order of *X*. The vertical dash line shows the *C* value for the current arrangement of the operons in a target genome. (*C*) A comparison between the *C* distributions when randomly permuting all 300 operons participating only in one pathway (curve on the left) versus randomly permuting 300 operons participating in more than one pathway (curve on the right) in the genome of *E. coli* K-12. (*D*) The same as *C* but for *B. subtilis* str. 168. The dotted curves represent the distributions of *C* values when using artificially composed pathway models (one-million times).

real pathways in both organisms are significantly smaller than the corresponding *C* values of the artificial pathways.

Possible Interpretations for Our Observations

Our main observations, i.e., (*i*) operons encoding the same pathway tend to cluster together to facilitate cotranscription (Table 1), but generally do not minimize their genomic dispersion due to the constraints of other pathways involving some of these operons (Fig. 1); and (*ii*) the current arrangement of operons in a bacterial genome tends to minimize the *C* value (Fig. 2), i.e., the overall dispersion of pathways in terms of their operons' genomic locations, could be possibly interpreted as follows. For (*i*), it could be possibly (partially) explained in the same spirit of the selfish operon model (25), proposed to explain the formation of operons. This model states that having functionally related genes grouped into operons could reduce the probability of losing the entire functionality of this group of genes because such a genomic arrangement facilitates the restoration of the functionality of the entire operon via one horizontal gene transfer. A similar argument could be made about operons versus their participating pathways, although more careful analyses will be needed. For (*ii*), we speculate that the genomic arrangement of operons has evolved to minimize the total effort in locating and activating all the pathways during the lifecycle of an organism. The assumption we used here is that the shorter the genomic region covering all the operons of a pathway, the less effort it takes to locate and activate the whole pathway, which may involve remodeling/unfolding of the (dynamic) relevant supercoiled domains to make the targeted operons exposed on the surface to facilitate cotranscription (12, 13, 26). This explanation is consistent with our data that more frequently activated pathways tend to have their operons more tightly clustered together (Table 1). In spirit, this is in agreement with the coregulation model (4) which was also proposed to explain the formation of operons. It should be noted that (*i*) is essentially a prerequisite of (*ii*); hence (*i*) could also be possibly explained in terms of transcription coactivation. Further studies are clearly needed to make our explanation less speculative.

Extension to Other Prokaryotic Genomes

To check for the generality of the above observation, we have also performed similar studies on seven other bacteria: *Synechocystis* sp. PCC6803 (phylum: *Cyanobacteria*), *Mycobacterium leprae* TN (*Actinobacteria*), *Thermotoga maritima* MSB8 (*Thermotogae*), *Cytophaga hutchinsonii* ATCC 33406 (*Bacteroidetes*), *Acinetobacter* sp. ADP1 (*Proteobacteria*), *Chlamydophila abortus* S26/3 (*Chlamydiae*), and *Mycoplasma genitalium* G-37 (*Firmicutes* and one of the smallest bacterial genomes), which were chosen using two criteria: (*i*) they are from a set of bacterial phyla covering the majority of the sequenced bacterial genomes, and (*ii*) each has a relatively high coverage by SEED pathways. In addition, we have also selected two archaeal genomes: *Pyrococcus furiosus* DSM 3638 and *Methanococcus maripaludis* S2, which are selected because our lab's current research involves these two organisms. Out of these nine selected bacteria and archaea, seven have similar results (Fig. S4) to the ones shown in Fig. 2 on *E. coli* and *B. subtilis*. On the other two bacteria (*Acinetobacter* sp. ADP1 and *Chlamydophila abortus* S26/3), the *C* value calculated for the actual genome is not significantly lower than those calculated from reshuffled genomes. This may be due to two possible reasons: (*i*) the quality and the coverage of pathway annotation and operon prediction are not nearly as good as those for *E. coli* and *B. subtilis*, on which our analyses rely, and (*ii*) other forces could also play a role in determining the global arrangement of operons in a genome. We anticipate that, with continued improvement in pathway annotation and operon prediction, we should be able to better differentiate the two possible reasons. Considering the diversity of the genomes used in this study and the consistency of the derived results, we believe that our observations may apply to the majority, if not all, of the prokaryotic genomes.

It should be noted that all of the analyses conducted in this work are based on operons annotated by the current pathway databases (SEED, KEGG, and BioCyc). Our separate analyses on a wide range of prokaryotic genomes from different phyla using the pathway databases suggest that all of the above observations will remain to be true as more operons will be annotated to be involved in the to-be-identified pathways.

Concluding Remarks

In summary, we presented strong evidence that the global arrangement of operons in a bacterial genome is strongly influenced by the biological pathways the genome encodes, specifically by the (frequency of the) activation of the pathways. We believe that this could be a primary principle, probably among a few others, that tightly governs the genomic arrangement of operons (see ref. 16 for more information). For example, the regulatory model (4) and the gene transfer model (25) have been proposed to explain the existence of operons and might also apply to the higher level interoperons organization studied in this paper. We anticipate that numerous fundamental questions could be effectively addressed through the application of this principle of genomic organization of operons, such as how flexible each operon is in terms of its genomic location in a genome. Note that our discovery concerns only the relative locations among operons without referring to any landmarks in a genome. We believe that overall it could be the joint force of this discovered principle and others, such as the preference of certain operons to specific genomic landmarks such as the origin of replication, which determines the genomic locations of all the operons in a bacterial genome.

Methods and Methods

Genomes and Operons. The genomes of *E. coli* K-12 MG1655 and *B. subtilis* str. 168 as well as of *Synechocystis* sp. PCC6803, *Mycobacterium leprae* TN, *Thermotoga maritima* MSB8, *Cytophaga hutchinsonii* ATCC 33406, *Mycoplasma genitalium* G-37, *Acinetobacter* sp. ADP1, *Chlamydophila abortus* S26/3, *Pyrococcus furiosus* DSM 3638, and *Methanococcus maripaludis* S2 were downloaded from ftp://ftp.ncbi.nih.gov as of January 14, 2009. All the predicted operons for these organisms were downloaded from the DOOR (Database of prOkaryotic OpeRons) (21) database at http://csbl1.bmb.uga.edu/OperonDB. The reported operon prediction accuracies on *E. coli* K-12 MG1655 and *B. subtilis* str. 168 are both better than 90%, and the prediction accuracy on prokaryotic genomes in general is about 80% (21). We have checked that the predicted operons are all consistent with the experimentally verified operons for both the genomes of *E. coli* and *B. subtilis* str. 168 retrieved from the RegulonDB (19) and the DBTBS (database of *Bacillus subtilis* transcription factors and promoters) (20) databases.

Pathways. All pathways used in this study were downloaded from three pathway databases. All subsystems for the nine organisms were downloaded from http://seed-viewer.theseed.org/ as of August 2009. All relevant KEGG pathways were downloaded from ftp://ftp.genome.jp/pub/kegg/ as of March 2009. All relevant BioCyc pathways were obtained from http://biocyc.org/ as of August 2009.

Distance Between Operons of a Same Pathway. Briefly, we consider all operons in a specified pathway arranged in a list as follows. We do not distinguish between operons on different strands but only consider their locations in the (circular) genome so adjacency relationship among operons in the pathway is uniquely defined. We remove the longest interoperonic distance from this circular list so the two relevant operons are viewed as the two ends of the pathway, which gives a unique list of operons. Note that the interoperonic distance is the number of operons inserted between two considered operons, so it does not have unit. See Fig. S1 for details.

Microarray Data. The microarray data for *E. coli* K-12 were downloaded from the M3D database (24) at http://m3d.bu.edu/. These data were collected under 380 experimental conditions and have been normalized across all the experiments so that the expressions of one gene can be compared directly across different experiments (24). The microarray data collected under 86 experimental conditions for *B. subtilis* str. 168 were downloaded from ftp://ftp.genome.jp/pub/db/community/expression/bsu/. Unlike the array data of *E. coli* K-12, these data are not normalized across different experiments. The Xpander v5 software (27) is used to perform a quantile normalization (28) on these microarray data so the resulting expression data can be compared across different experiments.

ACKNOWLEDGMENTS. We are grateful to the editor and the two anonymous reviewers for their insightful and invaluable comments and suggestions, which have helped to improve the overall quality of the paper. We thank Drs. Juan Cui and Xizeng Mao of the Computational Systems Biology Laboratory for their helpful discussion throughout this project. This work was supported by National Science Foundation (DEB-0830024 and DBI-0542119) and the BioEnergy Science Center grant (DE-PS02-06ER64304), which is supported by the Office of Biological and Environmental Research in the Department of Energy Office of Science.

1. Demerec M, Hartman PE (1959) Complex loci in microorganisms. *Ann Rev Microbiol* 13:377–406.
2. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356.
3. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S (2002) Computational identification of operons in microbial genomes. *Genome Res* 12(8):1221–1230.
4. Jacob F, Perrin D, Sanchez C, Monod J (1960) Operon: A group of genes with the expression coordinated by an operator (Translated from French). *C R Hebd Seances Acad Sci* 250:1727–1729.
5. Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 34(4):377–378.
6. Jin DJ, Cabrera JE (2006) Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in Escherichia coli. *J Struct Biol* 156(2):284–291.
7. Karlin S, Mrazek J (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* 182(18):5238–5250.
8. Fang G, Rocha EP, Danchin A (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics* 9:4.
9. Yang Q, Sze SH (2008) Large-scale analysis of gene clustering in bacteria. *Genome Res* 18(6):949–956.
10. Jeong KS, Ahn J, Khodursky AB (2004) Spatial patterns of transcriptional activity in the chromosome of Escherichia coli. *Genome Biol* 5(11):R86.
11. Kepes F (2004) Periodic transcriptional organization of the E. coli genome. *J Mol Biol* 340(5):957–964.
12. Wright MA, Kharchenko P, Church GM, Segre D (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc Natl Acad Sci USA* 104(25):10559–10564.

13. Carpentier AS, Torresani B, Grossmann A, Henaut A (2005) Decoding the nucleoid organisation of Bacillus subtilis and Escherichia coli through gene expression data. *BMC Genomics* 6(1):84.
14. Postow L, Hardy CD, Arsuaga J, Cozzarelli NR (2004) Topological domain structure of the Escherichia coli chromosome. *Gene Dev* 18(14):1766–1779.
15. Allen TE, Price ND, Joyce AR, Palsson BO (2006) Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Comput Biol* 2(1):e2.
16. Rocha EP (2008) The organization of the bacterial genome. *Ann Rev Genet* 42:211–233.
17. Kanehisa M, et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36(Database issue):D480–484.
18. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33(17):5691–5702.
19. Salgado H, et al. (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34(Database issue):D394–397.
20. Sierro N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: A database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 36(Database issue):D93–96.
21. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: A database for prokaryotic operons. *Nucleic Acids Res* 37(Database issue):D459–463.
22. Keseler IM, et al. (2009) EcoCyc: A comprehensive view of Escherichia coli biology. *Nucleic Acids Res* 37(Database issue):D464–470.
23. Karp PD, et al. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 33(19):6083–6089.
24. Faith JJ, et al. (2008) Many microbe microarrays database: Uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res* 36(Database issue):D866–870.
25. Lawrence JG, Roth JR (1996) Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* 143(4):1843–1860.
26. Audit B, Ouzounis CA (2003) From genes to genomes: Universal scale-invariant properties of microbial chromosome organisation. *J Mol Biol* 332(3):617–633.
27. Shamir R, et al. (2005) EXPANDER—An integrative program suite for microarray data analysis. *BMC Bioinformatics* 6:232.
28. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, (2) pp:185–193.