The Operonic Location of Auto-transcriptional Repressors Is Highly Conserved in Bacteria

Nimrod D. Rubinstein,[†,1,2] David Zeevi,[†,1] Yaara Oren,[1] Gil Segal,[3] and Tal Pupko*[,1,2]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel

[2]National Evolutionary Synthesis Center, Durham, North Carolina

[3]Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel

†These authors contributed equally to this work.

***Corresponding author:** E-mail: talp@post.tau.ac.il.

Associate editor: Helen Piontkivska

Research article

Abstract

Bacterial genes are commonly encoded in clusters, known as operons, which share transcriptional regulatory control and often encode functionally related proteins that take part in certain biological pathways. Operons that are coregulated are known to colocalize in the genome, suggesting that their spatial organization is under selection for efficient expression regulation. However, the internal order of genes within operons is believed to be poorly conserved, and hence expression requirements are claimed to be too weak to oppose gene rearrangements. In light of these opposing views, we set out to investigate whether the internal location of the regulatory genes within operons is under selection. Our analysis shows that transcription factors (TFs) are preferentially encoded as either first or last in their operons, in the two diverged model bacteria *Escherichia coli* and *Bacillus subtilis*. In a higher resolution, we find that TFs that repress transcription of the operon in which they are encoded (autorepressors), contribute most of this signal by specific preference of the first operon position. We show that this trend is strikingly conserved throughout highly diverged bacterial phyla. Moreover, these autorepressors regulate operons that carry out highly diverse biological functions. We propose a model according to which autorepressors are selected to be located first in their operons in order to optimize transcription regulation. Specifically, the first operon position helps autorepressors to minimize leaky transcription of the operon structural genes, thus minimizing energy waste. Our analysis provides statistically robust evidence for a paradigm of bacterial autorepressor preferential operonic location. Corroborated with our suggested model, an additional layer of operon expression control that is common throughout the bacterial domain is revealed.

Key words: activator, alloregulator, autoactivator, autoregulator, autorepressor, conservation, gene expression regulation, genome organization, genomic structure, molecular evolution, operon, repressor, transcription factor.

Introduction

Bacterial genes are typically organized in clusters, which are cotranscribed as polycistrons and subject to common regulatory control. This model was originally termed an operon by Jacob and Monod (1961) and was thought to encode proteins that carry out sequential steps in a certain metabolic pathway and/or are functionally related (Blumenthal 1998; Lawrence 2002, 2003), where even the order of the genes within the operon tends to follow the order of the biochemical process (Rocha 2008). However, many examples were found to deviate from this classical definition (Lawrence 2003), and hence the term transcriptional unit is often used instead or interchangeably (e.g., Salgado et al. 2000). In this work, we only use the term operon.

Several hypotheses have been proposed to explain why genes are organized in operons: What are the selection advantages, if any, for the existence of operons, and how are they created and maintained (reviewed in Lawrence 2003; Rocha 2008). It was first proposed that operons allow efficient regulation of gene expression providing a selective explanation for the operonic structure. Coregulation may explain why operons are maintained, however, it does not provide a good explanation for their formation (Lawrence 1997). To name a few caveats, precise juxtapositioning of previously unlinked genes is a rare event, and there are many examples of genes whose coregulation would be highly beneficial, yet are not found in operons (Lawrence 1997). Hence, an alternative hypothesis was suggested, according to which operons are selfish entities whose structure does not provide a selective benefit to the individual organism but is vital for efficient propagation and integration via horizontal gene transfer (HGT) within bacterial genomes (Lawrence 1997). In contrast, Price et al. (2005) provided evidence that HGT is not the cause of operon formation but rather promotes the prevalence of preexisting ones and suggested that operons are efficient structures for complex coregulation of expressionally related genes. In support of this, Davids and Zhang (2008) showed that horizontally transferred genes have a significantly weaker tendency to be encoded in defined operons compared with core genes (genes that appear in multiple related bacterial strains) which presumably have not been horizontally transferred. Furthermore, the work by Perez and Groisman (2009), studying the evolution of bacterial regulons among two closely related species, suggests that horizontally transferred genes are often transcribed by existing transcription factors (TFs) which also regulate core genes. This implies that genes may be transferred as separate units rather than as entire operons with their regulatory mechanisms.

Analyzing their spatial organization in the genome, operons which are coregulated and even operons which regulate each other were found to colocalize, which is in line with the view that regulatory control is a major evolutionary driving force in determining the spatial organization of genomes (Warren and ten Wolde 2004). Adding to that, the hierarchy of the regulatory network was also found to be correlated with spatial localization (Seshasayee et al. 2009). Specifically, TFs regulating a limited number of operons (lower in the hierarchy) strongly colocalize with their binding sites whereas global, master, TFs, which regulate a large number of operons (at the top of the hierarchy) and are expressed in higher copy numbers, do not show such a spatial pattern. This further supports claims for selection for efficient regulatory control and also suggests an economical tradeoff between copy number and distance of TF-binding site search since TFs move along the genome via diffusion (Kolesov et al. 2007; Wunderlich and Mirny 2008).

Notwithstanding, the internal order of genes within operons was claimed to be poorly conserved over long periods of evolutionary times, leave several exceptions such as ribosomal operon structures and other operons whose products physically interact (Dandekar et al. 1998; Itoh et al. 1999). Disruption of the operonic structure due to molecular mechanisms such as rearrangements, deletions and HGT insertions, and gene displacements (Fondi et al. 2009) was thus invoked to be nearly selectively neutral, possibly since coexpression constraints are too weak to oppose these processes (Mushegian and Koonin 1996; Itoh et al. 1999).

The observation that operon structures tend to undergo disruption during evolution seems to be in partial contrast with the view that regulatory control is a strong evolutionary driving force shaping the structure of bacterial genomes. In this work, we show that some features of the operon structure are highly conserved throughout evolution. Specifically, we show that the operonic locations of TFs that regulate their own operon, relative to the locations of other genes in their operon, have been subject to strong purifying selection and hence remained conserved even among highly diverged bacterial phyla. Our investigation reveals that TFs have a significant tendency to be located at the terminal operon positions, that is, first and last. Moreover, we find that this trend is mainly attributed to autoregulatory TFs. Furthermore, we show that autorepressors, regulating functionally diverse operons, are preferentially located at the first position within their operons and hence contribute mostly to this trend. We then use a data set that spans 13 bacterial phyla to show that this trend is widely common and hence strongly evolutionary conserved. We conclude by presenting a model that explains the biochemical mechanism that drives selection for preferential location of functionally diverse autorepressors at the start of their operons.

Materials and Methods

TF, Operon, and Regulation Data

Escherichia coli and *Bacillus subtilis* were used as reference organisms as they are the most comprehensively annotated bacteria, especially with regards to transcription regulation. All experimentally validated TFs, operon structures, and modes of regulation were derived from the BioCyc database (Karp et al. 2005; Keseler et al. 2009), where for *E. coli*, RegulonDB database (Gama-Castro et al. 2011) was additionally used. Whenever a certain TF is indicated to regulate several transcriptional units derived from the same operon, the longest transcriptional unit was selected as the representative operon and the number of genes it encodes was defined as its size. Each TF that is indicated to regulate an operon in which it is encoded was classified as an autoregulator, regardless if it is additionally indicated to regulate operons in which it is not encoded (for illustration, see supplementary fig. S1*A*, Supplementary Material online). Otherwise, the TF was classified as an alloregulator (for illustration, see supplementary fig. S1*B*, Supplementary Material online). Each autoregulator that is indicated to repress the operon in which it is encoded was classified as an autorepressor, regardless of the regulatory effect it is indicated to exert on other operons it regulates. The same definitions were used to classify TFs as autoactivators, where the regulatory effect exerted by the TF on the operon in which it is encoded is activation. An autoregulator that is indicated to both repress and activate the operon in which it is encoded was classified as a dual-autoregulator. Alloregulators that are indicated to repress all their target operons were classified as allorepressors. Similarly, alloregulators that are indicated to activate all their target operons were classified as alloactivators. Finally, alloregulators that are indicated both to repress and to activate their target operons were classified as dual-alloregulators.

Supplementary tables S1 and S2 (Supplementary Material online) provide information regarding the mode of regulation, operonic location, and DNA-binding domain (see below) for each of the *E. coli* and *B. subtilis* TFs that were used in this work, respectively (which was derived from the Swiss-Prot Protein knowledgebase (Gasteiger et al. 2003) and from the BioCyc database (Karp et al. 2005; Keseler et al. 2009), unless stated differently).

The predicted operon structures of all other bacterial organisms used in this work were retrieved from the BioCyc database (Karp et al. 2005). Notably, for these bacteria no knowledge regarding the mode of transcription regulation is available (i.e., whether they are autorepressors, autoactivators, etc.).

TF DNA–Binding Domain Data

TF DNA–binding domains were derived from the proTF database (Bai et al. 2010), which holds an extensive collection of predicted TFs and their DNA-binding domain, based on the methodology that was used to construct the DBD database (Wilson et al. 2008). Each chromosomally encoded TF (as opposed to a plasmid-encoded TF), which could be matched to a gene in any of the BioCyc predicted operons for the relevant organism, was used in our work. As a result, 573 bacterial organisms (listed in supplementary table S3, Supplementary Material online) spanning the following 13 phyla: Deinococcus-Thermus, Actinobacteria, Chloroflexi, Bacteroidetes, Chlorobi, Spirochaetes, Acidobacteria, Firmicutes, Tenericutes, Thermotogae, Cyanobacteria, Proteobacteria, and Verrucomicrobia were retained for our analysis. The phylogenetic lineage information (i.e., phylum) for each of these organisms was derived from the NCBI taxonomy database (Benson et al. 2011; Sayers et al. 2011).

The phylogenetic tree topologies presented in figures 1–3 were derived from the prokaryotic species tree available at the MicrobesOnline database (Dehal et al. 2010).

Statistical Inference

In order to test whether an observed number of TFs encoded at a certain operon position: first, middle, last, or terminal (first or last), is significantly greater from what is expected by chance we carried out the following randomization procedure. The genes in each of the operons were shuffled 100,000 times. For each such shuffle repetition, the number of TFs observed at each operon position was counted. The reported *P* value of the statistical significance for an observed number of TFs encoded at a certain operon position was thus defined as the fraction of shuffle repetitions which showed an equal or higher number of TFs at that certain operon position. We note that this procedure ensures that the computed *P* value accounts for the size of the operon in which a TF is encoded.

Homology Analysis of the *mar*R Gene of *E. coli*

Orthologs of the *mar*R gene of *E. coli* were derived from the BioCyc database (Karp et al. 2005; Keseler et al. 2009), which uses bidrectional best BLAST hits. Other genes that reside in the operons encoding *E. coli mar*R orthologs were similarly identified: The *mar*A and *mar*B genes, which are encoded in the *E. coli mar* operon, downstream to *mar*R and the *emr*A and *emr*B genes, which are encoded downstream to the *E. coli mpr*A autorepressor (also of the MarR DNA–binding domain family) and appear in several *mar* orthologous operons.

Results

Transcription Factors Are Preferentially Located at Terminal Operon Positions

Our initial null hypothesis is that TFs are randomly distributed within their encoding operons, that is, there is no preference to reside at any specific operonic location such as the first gene in an operon. To test this hypothesis, we classified the operon positions of all the annotated TFs of the Gram-negative and Gram-positive model bacteria *E. coli* and *B. subtilis*, respectively, residing in operons of three or more genes as either 1) terminal (i.e., first or last positions) or 2) middle (otherwise) (see supplementary tables S1 and S2, Supplementary Material online for detailed information on the TFs and their encoding operons in *E. coli* and *B. subtilis*, respectively). For the 35 TFs of *E. coli* that meet the above criterion, we observed a strong and statistically significant tendency to be located at terminal positions (26 of the 35 TFs; $P = 0.0014$; randomization test, see Materials and Methods; table 1 and supplementary fig. S2*A*, Supplementary Material online). The same analysis for *B. subtilis* revealed a similar trend, where 19 of the 30 relevant TFs are located at terminal operon positions ($P = 0.02$; table 1 and supplementary fig. S2*B*, Supplementary Material online). Interestingly, in both *E. coli* and *B. subtilis*, TFs located at the first operon position mainly contribute to this signal (16 of 26 and 14 of 19 for *E. coli* and *B. subtilis*, respectively).

Autorepressors Are Preferentially Located at the First Operon Position

Aiming to further refine the trend of TF operonic location, we divided TFs into the following two categories: 1) autoregulators—TFs that regulate their encoding operon (but may also regulate other operons, for illustration, see supplementary fig. S1*A*, Supplementary Material online) and 2) alloregulators—TFs that do not regulate their encoding operon (for illustration, see supplementary fig. S1*B*, Supplementary Material online). We thus repeated the previous analysis separately for autoregulators and alloregulators. In the previous analysis, we tested whether TFs show preference for terminal operon positions, that is, first or last versus middle positions. Hence, operons with two genes were excluded since they do not include middle positions. In this analysis, operons with only two genes were also considered since we differentiated between first and last operon positions. Specifically, here, we classified each TF as either first, middle, or last in its operon. This analysis revealed that *E. coli* autoregulators show a strong and statistically significant tendency to be located at the first operon position (27 first, 7 middle, and 17 last; respective *P* values = 0.005, 0.999, and 0.686; table 1 and supplementary fig. S3*A*, Supplementary Material online). This trend was also found for *B. subtilis* autoregulators (17 first, 7 middle, and 7 last; respective *P* values = 0.006, 0.982, and 0.933; table 1 and supplementary fig. S3*B*, Supplementary Material online). In contrast, the same could not be concluded about alloregulators, neither of *E. coli* (12 first, 2 middle, and 13 last; respective *P* values = 0.577, 0.9, and 0.418; table 1 and supplementary fig. S3*A*, Supplementary Material online) nor of *B. subtilis* (13 first, 4 middle, and 10 last; respective *P* values = 0.2, 0.949, and 0.652; table 1 and supplementary fig. S3*B*, Supplementary Material online). This refinement thus establishes that the trend for TF localization at terminal operon positions mainly stems from autoregulators located first in their operons.

TFs can be further classified according to the regulation effect they exert on transcription of a specific target (reviewed in van Hijum et al. 2009): 1) repressors—TFs that work to prevent or decrease transcription of their targets by any of several physical mechanisms that hinder RNA polymerase activity; 2) activators—TFs that work to allow or increase transcription of their targets by any of several physical mechanisms that facilitate RNA polymerase activity; and 3) dual-regulators—TFs that can act both as repressors and as activators (Perez-Rueda and Collado-Vides 2000). We used this classification to further refine the operonic localization signal we observed for TFs, classifying autoregulators as: 1) autorepressors—TFs that repress transcription of their encoding operon; 2) autoactivators—TFs that activate transcription of their encoding operon; and 3) dual-autoregulators—TFs that can both repress and activate transcription of their encoding operon (for illustration, see supplementary fig. S1A, Supplementary Material online). Repeating the localization analysis, this time only for autoregulators (autorepressors, autoactivators, and dual-autoregulators) revealed that autorepressors show a strong and statistically significant tendency to be located at the first operon position both in *E. coli* (16 first, 2 middle, and 6 last; respective *P* values = 0.0009, 0.999, and 0.91; table 1 and supplementary fig. S4A, Supplementary Material online) and in *B. subtilis* (16 first, 5 middle, and 3 last; respective *P* values = 0.0002, 0.991, and 0.994; table 1 and supplementary fig. S4B, Supplementary Material online). In contrast, no such signal was observed for autoactivators of *E. coli* (9 first, 4 middle, and 10 last; respective *P* values = 0.399, 0.984, and 0.239; table 1 and supplementary fig. S4A, Supplementary Material online). Lack of preference for operonic location seems also to be the case for autoactivators of *B. subtilis* (1 first, 2 middle, and 4 last; respective *P* values = 0.962, 0.674, and 0.212; table 1 and supplementary fig. S4B, Supplementary Material online). Dual-autoregulators were only observed in *E. coli*, and similar to autoactivators, no preferential location was detected, yet these are insufficient data for a robust statistical conclusion (2 first, 1 middle, and 1 last; table 1).

The two-component system operon type, which prototypically consists of a histidine protein kinase that senses extracellular stimuli and a response regulator protein (which is a TF) that is activated by the histidine kinase to elicit the specific response (Stock et al. 2000), is considerably common in our data (18 two-component system TFs in *E. coli* and 12 in *B. subtilis*, supplementary tables S1 and S2, Supplementary Material online, respectively). Nevertheless, as none of the two-component system TFs are autorepressors, neither in *E. coli* nor in *B. subtilis*, they do not artificially skew the analysis. We note, however, that *E. coli* autoactivators, other than two-component system TFs, show a preference for the last operon position (3 first, 4 middle, and 7 last; respective *P* values = 0.835, 0.946, and 0.07; table 1). The few autoactivators, which are not two-component system TFs, in *B. subtilis*, do not seem to show preference for any operon position (1 first, 1 middle, and 2 last; table 1), yet the data are too scant to decisively determine this. Finally, *E. coli* dual-autoregulators, which are not two-component system TFs, show no preference for any operon position (1 first and 1 middle; table 1).

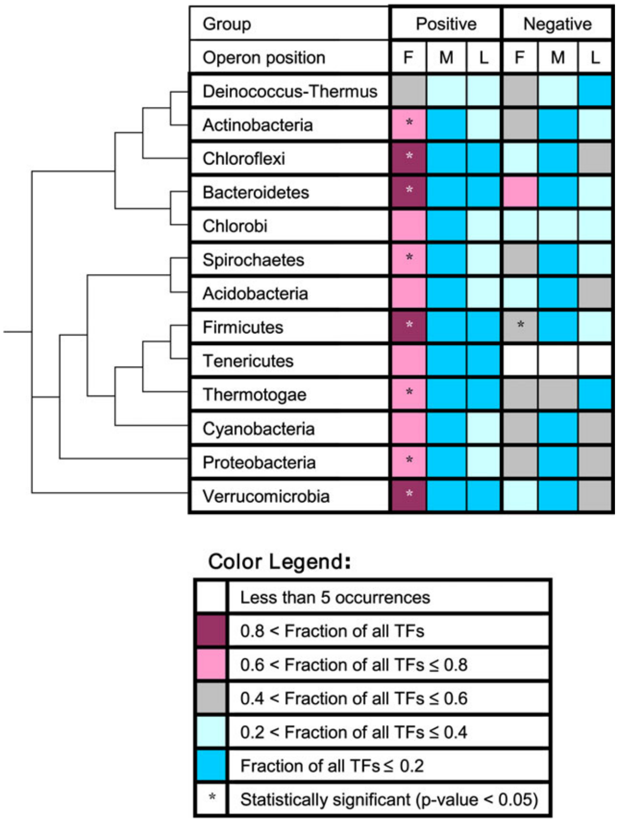

FIG. 1. A heat map describing the phylogenetic signal of TF preferential operonic location. The positive group consists of DNA-binding domains that occur only as an autorepressor in either *Escherichia coli* or *Bacillus subtilis*. The negative group consists of DNA-binding domains that do not occur as an autorepressor neither in *E. coli* nor in *B. subtilis*. Each phylum is represented by a specific organism (see supplementary table S4, Supplementary Material online). The fraction of the TFs in each operon position (F, first; M, middle; L, last) is color coded according to the color legend, and an asterisk marks a statistically significant ($P < 0.05$) preference for the relevant operon position.

We thus conclude that the trend for TF localization at terminal operon positions mainly stems from autorepressors located first in their operons. The fact that autorepressors both in *E. coli* and in *B. subtilis*, which are considerably diverged bacteria, show a similar and striking localization trend raises the question whether there is a common biological mechanism selecting for preferential operonic localization of autorepressors.

The Operonic Location of Autorepressors Is Highly Conserved among Diverse Bacterial Phyla

Autorepressors of both *E. coli* and *B. subtilis*, which are highly evolutionary diverged, show a similar strong and statistically significant preference to be located first in their respective operons. We thus hypothesized that the

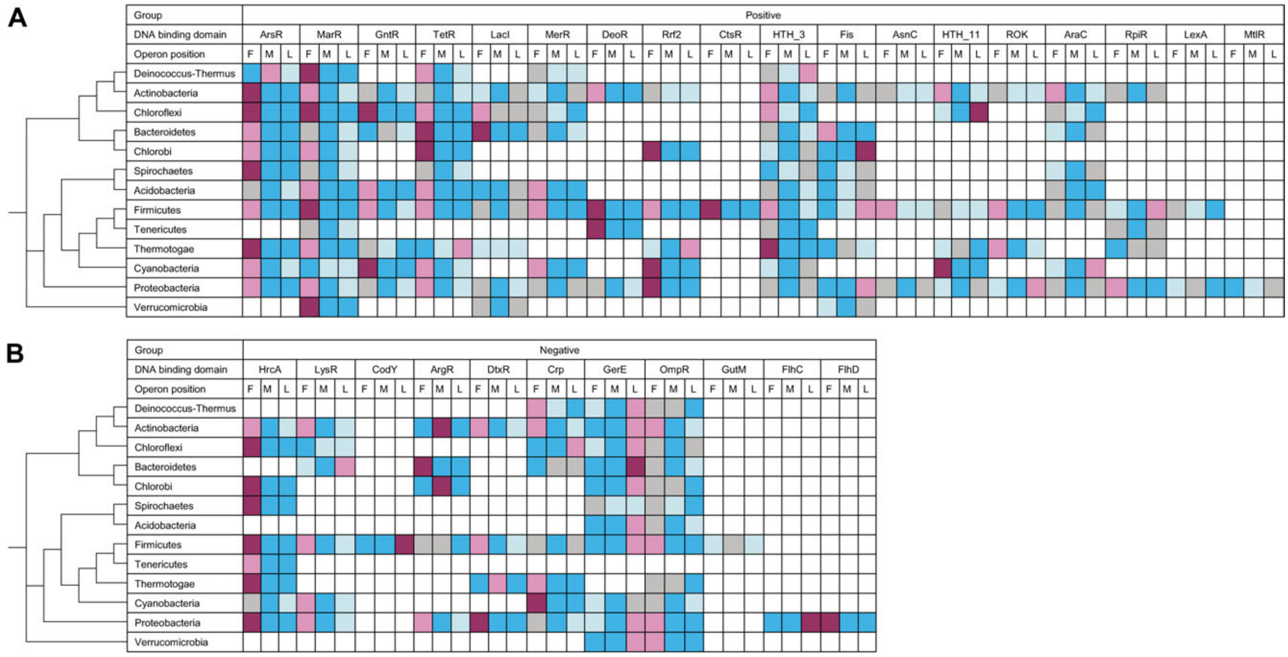

FIG. 2. A heat map describing the phylogenetic signal of TF preferential operonic location. The positive group (*A*) consists of DNA-binding domains that occur as an autorepressor at least once in *Escherichia coli* or *Bacillus subtilis*. The negative group (*B*) consists of DNA-binding domains that do not occur as an autorepressor neither in *E. coli* nor in *B. subtilis*. Each phylum which has at least five occurrences of a certain DNA-binding domain was considered for this analysis. The fraction of TFs of a certain DNA-binding domain in each operon position (F, first; M, middle; L, last) is color coded according to the color legend in figure 1.

operonic location of bacterial autorepressors is maintained due to strong purifying selection. To this end, we tested whether the signal for autorepressor preference of the first operon position is recapitulated on a wider evolutionary span of bacteria. Although hundreds of bacterial genomes are completely sequenced, high-resolution genome-wide annotation regarding transcriptional regulation exists only for *E. coli* and *B. subtilis*. Nevertheless, TFs can be detected due to the strong sequence signatures of their DNA-binding domains (Kummerfeld and Teichmann 2006). In addition, the distance between genes integrated with functional information, when available, allows to predict operons and transcriptional units (Romero and Karp 2004). We used these predictions for detecting TFs and their respective operon positions in 573 bacterial genomes spanning 13 phyla (for details, see Materials and Methods and supplementary table S4, Supplementary Material online). We used the regulatory roles annotated for the *E. coli* and *B. subtilis* TFs (e.g., autorepressors and autoactivators) as proxies for the regulatory roles of TFs in other bacteria according to the DNA-binding domain families.

To test whether the preference of the first operon position of autorepressors is under strong purifying selection throughout these bacteria, we divided any of the DNA-binding domains that occur either in *E. coli* or in *B. subtilis* into two groups: 1) positive group—any of the DNA-binding domains that occur only as autorepressors in either *E. coli* or *B. subtilis* (ArsR, ROK, AsnC, MarR, CtsR, LexA, and MtlR) and 2) negative group—any of the DNA-binding domains that do not occur as autorepressors neither in *E. coli* nor in *B. subtilis* (GerE, Crp, FlhC, LysR/TdcR, OmpR, HrcA, FlhD, CodY, ArgR, DtxR, GutM, and HTH_10). To avoid analyzing organisms that are too evolutionarily related, we selected a single organism from each of the 13 phyla for each of the groups. Specifically, the organism with the highest number of TFs corresponding to the DNA-binding domains of that group was selected (with at least five occurrences). Thus, the representative organisms in each group (provided in supplementary table S4, Supplementary Material online) are not necessarily identical. We thus compared the operon localization of the positive group with that of the negative group, for each of the 13 phyla representatives. The heat map in figure 1 presents the results of this comparison, where 8 of the 13 (~61.5%) phyla show a statistically significant preference for the first operon position in the positive group compared with only 1 of the 12 (~8.3%) relevant phyla in the negative group (only 12 phyla were used for the negative group since we require at least five TF occurrences in the representative organism). This striking trend thus indicates that autorepressors, which regulate diverse biological functions, have a strong tendency to be located first in their respective operons, throughout nearly the entire bacterial domain.

We performed an additional phylogenetic analysis, in which we did not pool together different DNA-binding domains but rather considered them separately. We also did not sample a single organism from each of the 13 phyla in this analysis but used all 573 organisms. The positive group here was defined to include any DNA-binding domain that occurs at least once as an autorepressor in *E. coli* or in *B. subtilis* (but may also occur as other types of regulators) and the negative group was defined to include

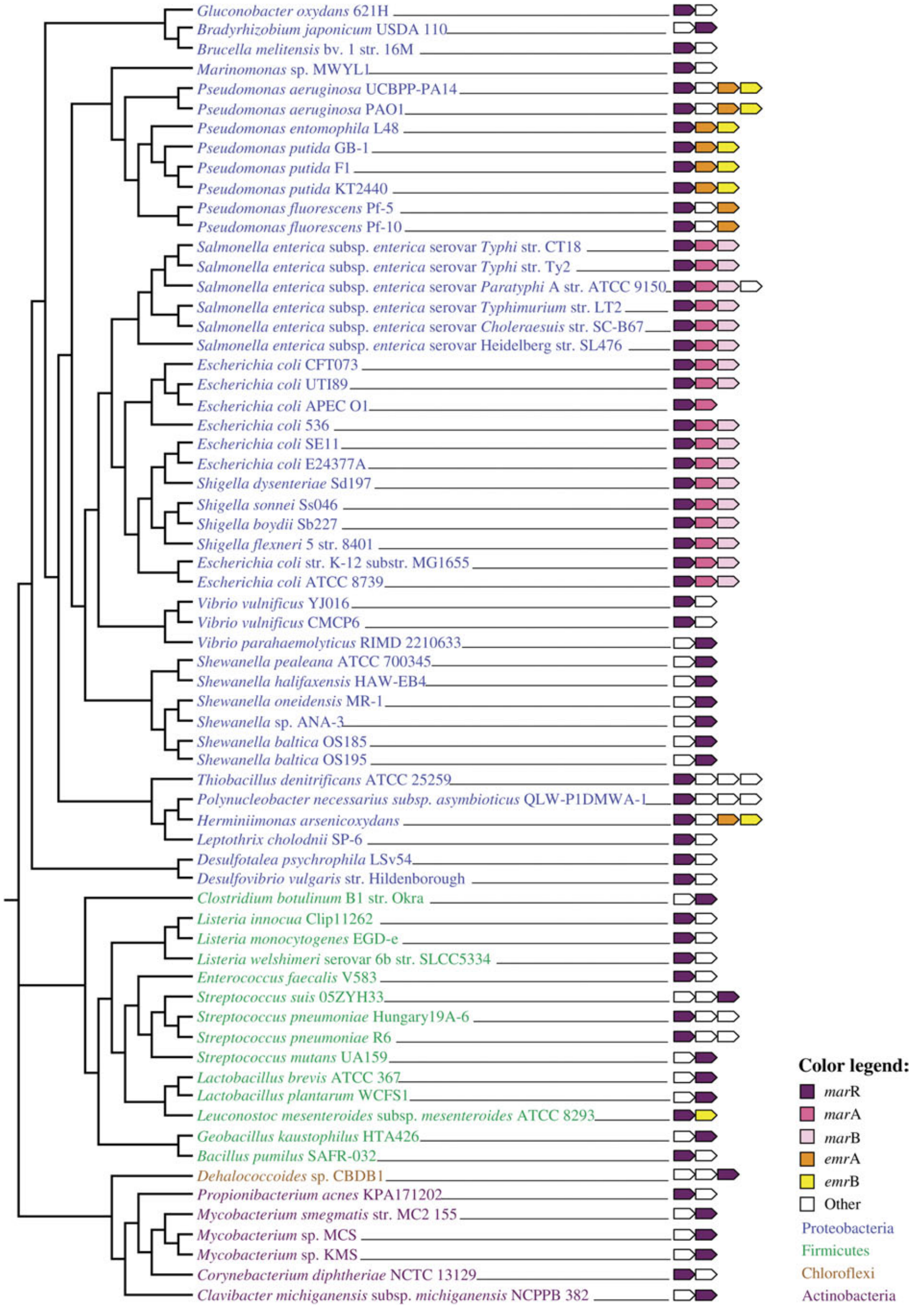
Gluconobacter oxydans 621H
Bradyrhizobium japonicum USDA 110
Brucella melitensis bv. 1 str. 16M
Marinomonas sp. MWYL1
Pseudomonas aeruginosa UCBPP-PA14
Pseudomonas aeruginosa PAO1
Pseudomonas entomophila L48
Pseudomonas putida GB-1
Pseudomonas putida F1
Pseudomonas putida KT2440
Pseudomonas fluorescens Pf-5
Pseudomonas fluorescens Pf-10
Salmonella enterica subsp. enterica serovar Typhi str. CT18
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476
Escherichia coli CFT073
Escherichia coli UTI89
Escherichia coli APEC O1
Escherichia coli 536
Escherichia coli SE11
Escherichia coli E24377A
Shigella dysenteriae Sd197
Shigella sonnei Ss046
Shigella boydii Sb227
Shigella flexneri 5 str. 8401
Escherichia coli str. K-12 substr. MG1655
Escherichia coli ATCC 8739
Vibrio vulnificus YJ016
Vibrio vulnificus CMCP6
Vibrio parahaemolyticus RIMD 2210633
Shewanella pealeana ATCC 700345
Shewanella halifaxensis HAW-EB4
Shewanella oneidensis MR-1
Shewanella sp. ANA-3
Shewanella baltica OS185
Shewanella baltica OS195
Thiobacillus denitrificans ATCC 25259
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1
Herminiimonas arsenicoxydans
Leptothrix cholodnii SP-6
Desulfotalea psychrophila LSv54
Desulfovibrio vulgaris str. Hildenborough
Clostridium botulinum B1 str. Okra
Listeria innocua Clip11262
Listeria monocytogenes EGD-e
Listeria welshimeri serovar 6b str. SLCC5334
Enterococcus faecalis V583
Streptococcus suis 05ZYH33
Streptococcus pneumoniae Hungary19A-6
Streptococcus pneumoniae R6
Streptococcus mutans UA159
Lactobacillus brevis ATCC 367
Lactobacillus plantarum WCFS1
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293
Geobacillus kaustophilus HTA426
Bacillus pumilus SAFR-032
Dehalococcoides sp. CBDB1
Propionibacterium acnes KPA171202
Mycobacterium smegmatis str. MC2 155
Mycobacterium sp. MCS
Mycobacterium sp. KMS
Corynebacterium diphtheriae NCTC 13129
Clavibacter michiganensis subsp. michiganensis NCPPB 382
Color legend:
marR
marA
marB
emrA
emrB
Other
Proteobacteria
Firmicutes
Chloroflexi
Actinobacteria

Fig. 3. The preferential operonic location the *mar*R autorepressor. Homology analysis of the operonic location of *Escherichia coli*'s *mar*R autorepressor. The identity of other operon genes and the phyla to which the organisms belong are described in the color legend.

all other DNA-binding domains that occur in *E. coli* or in *B. subtilis*. The heat map in figure 2 largely recapitulates the results presented in figure 1. Namely, the positive group shows marked preference for the first operon position relative to the negative group. The statistical significance of the operonic location of TFs in the above analysis (randomly shuffling each operonic location within each genome independently) is justified if the representative bacteria from each phylum are distant enough from each other to avoid confounding by phylogenetic relatedness and conservation. In this analysis, this assumption may not hold since closely related bacteria are included. It is thus possible that close phylogenetic relatedness contributes to the conservation signal observed in figure 2. Therefore, we did not compute statistical significance of operonic location in this analysis. Nevertheless, the fact that the negative group still does not show as striking preferential operonic location as the positive group, suggests that the bias due to phylogenetic relatedness is limited and at least part of the signal of the positive group stems from evolutionary conservation.

Notwithstanding, some DNA-binding domains of the positive group deviate in their preferential operonic location. In addition to autorepressors, many of such

Table 1. The Observed Number of TFs in the Different Operon Positions and Its Statistical Significance.

Organism	*Escherichia coli*				*Bacillus subtilis*			
Operon Position	**Terminal**	**First**	**Middle**	**Last**	**Terminal**	**First**	**Middle**	**Last**
All TFs[a]	26 (*P* = 0.0014)		9		19 (*P* = 0.02)		11	
Autoregulators[b]		27 (*P* = 0.005)	7 (*P* = 0.999)	17 (*P* = 0.686)		17 (*P* = 0.006)	7 (*P* = 0.982)	7 (*P* = 0.933)
Alloregulators[b]		12 (*P* = 0.577)	2 (*P* = 0.9)	13 (*P* = 0.418)		13 (*P* = 0.2)	4 (*P* = 0.949)	10 (*P* = 0.652)
Autorepressors[b]		16 (*P* = 0.0009)	2 (*P* = 0.999)	6 (*P* = 0.91)		16 (*P* = 0.0002)	5 (*P* = 0.991)	3 (*P* = 0.994)
Autoactivators[b]		9 (*P* = 0.399)	4 (*P* = 0.984)	10 (*P* = 0.239)		1 (*P* = 0.962)	2 (*P* = 0.674)	4 (*P* = 0.212)
Dual-autoregulators[b]		2	1	1				
Non-TCS autoactivators[b]		3 (*P* = 0.835)	4 (*P* = 0.946)	7 (*P* = 0.07)		1	1	2
Non-TCS dual-autoregulators[b]		1	1					

[a] Derived from operons with three or more genes.

[b] Derived from operons with two or more genes.

DNA-binding domains occur also as other types of regulators either in the same organism or in different ones. For example, the RpiR DNA-binding domain occurs as an autorepressor in *E. coli* (repressing the operon involved in transport and catabolism of D-allose and low-affinity transport of D-ribose) but as an autoactivator in *B. subtilis* (activating the maltose-metabolism operon). Accordingly, in Proteobacteria it shows a significant preference for the first operon position, whereas in Firmicutes (and in Thermotogae) it shows, instead, a significant preference for the last operon position. Similarly, the negative group includes several DNA-binding domains that show marked preference for the first operon position. Perhaps the most striking one is the HrcA DNA–binding domain. This DNA-binding domain occurs only once in *B. subtilis* (the *hrc*A gene), where it is an allorepressor of the *gro*E operon, which encodes ATP-dependent class I heat shock genes. The HrcA TF is encoded within the *dna*K operon, which encodes other ATP-dependent class I heat shock genes. Although no binding site has been found for the *hrc*A product near the *dna*K promoter, it has been experimentally shown to repress the *dna*K operon (Schulz and Schumann 1996).

Autorepressors Located at the First Operon Position Regulate Diverse Biological Functions

In our analyses, we have pooled together TFs according to the regulatory effect they carry out on the transcription of their target genes. We present six examples of autorepressors encoded first in the operons they regulate, three of *E. coli* and three of *B. subtilis*. These examples illustrate that despite their common regulatory role, these specific TFs are involved in regulating considerably diverse biological functions. 1) The *ars*R gene of *E. coli* (ArsR DNA–binding domain) encodes a transcriptional repressor of the energy-dependent arsenate reductase and an inner membrane–associated arsenite export system gene (Carlin et al. 1995); 2) the *pdh*R gene (GntR DNA–binding domain) encodes the transcriptional repressor of the genes that constitute the *E. coli* pyruvate dehydrogenase complex and additionally represses other operons, which encode downstream elements of the energy production pathway (Quail and Guest 1995; Ogasawara et al. 2007). The GntR DNA–binding domain was not included in the positive group in the phylogenetic analysis of pooled DNA-binding domains (fig. 1) since it occurs once as a dual-autoregulator in *E. coli*. However, the specific DNA-binding domain phylogenetic analysis (fig. 2) reveals a marked preference for the first operon position of GntR TFs; 3) the *mar*R gene (MarR DNA–binding domain) encodes the transcriptional repressor of the *E. coli mar* operon, which encodes the master transcriptional activator of the multidrug resistance phenotype (the *mar*A gene, Alekshun and Levy 1997). Figure 3 presents the operons in which *mar*R orthologs are encoded (for details, see Materials and Methods). This strict phylogenetic analysis shows that 47 of the 66 *mar*R orthologs (71.2%) are encoded in the first operon position, which further illustrates the preferential operonic location of autorepressors; 4) the *azl*B gene product (AsnC DNA–binding domain) represses the transcription of the *B. subtilis azl*BCDEF operon encoding branched amino-acid transport proteins (Belitsky et al. 1997); 5) the *cts*R gene product (CtsR DNA–binding domain) represses the *B. subtilis clp*C operon, where *clp*C along with *clp*E and *clp*P which are encoded in different operons and are also repressed by the *cts*R gene product, encode ATP-dependent class III heat shock proteins (Miethke et al. 2006); and 6) the *gln*R gene product (MerR DNA–binding domain) represses the *B. subtilis gln*RA operon, where *gln*A encodes a glutamine synthetase that plays a central role in nitrogen metabolism (Schreier et al. 1989; Fisher 1999). The MerR DNA–binding domain was also not included in the positive group in the pooled DNA-binding domain phylogenetic analysis (fig. 1) since it occurs once as an alloactivator in *E. coli*. However, the DNA-binding domain specific phylogenetic analysis (fig. 2) reveals a preference for the first operon position of MerR TFs.

A Model for the Preferential Operonic Localization of Autorepressors

All of the above are examples of operons that encode a specific biological mechanism, which is maintained repressed as long as the inducer (that leads to derepression) is absent. Once the inducer is present and derepression is established, all cistrons within the operon are expected to be expressed. The amounts of protein produced from each of the cistrons in an operon is expected to follow a decreasing gradient according to their distance from the promoter due to abortive transcription termination (Ullmann et al. 1979; Danchin and Ullmann 1980; Li and Altman 2004; Lee et al. 2008), abortive translation (Menninger 1976), or even posttranslational effects (Murakawa et al. 1991); a phenomenon known as natural polarity (Kurland 1992) (for a supporting analysis, see supplementary material and fig. S5, Supplementary Material online). Autorepressors located first in their operons are thus expected to be produced to the largest amounts relative to the other proteins encoded within their operons during induction, despite the fact that repression is not required at all during that time. This begs the question, why would such a seemingly wasteful operon organization be selected for?

While the operon is effectively maintained repressed, the autorepressor is not produced, and thus its concentration within the cell is expected to gradually decrease with time due to protein degradation. Conditions that cause operon induction may be encountered seldom enough to allow the autorepressor concentration to fall beneath a level that may no longer maintain effective repression. In this case, leaky transcription is expected to occur, which would then restore the autorepressor concentration required for effective repression. Thus, a steady state is expected, in which the concentration of the autorepressor, and hence effective repression, is maintained via protein degradation and leaky transcription.

While repression is by and large energy independent and achieved via a single protein, other functions encoded by the operon cistrons such as carbon source utilization, efflux of noxious agents, stress induced proteolysis, folding, and assembly require a far larger array of energy-dependent proteins. An apparent wasteful expression of a repressor during operon induction would thus be less wasteful than expression of energy-consuming proteins due to leaky transcription. An autorepressor encoded first in its operon thus seems to provide the most economical solution to the potential wastefulness of leaky transcription. This rational also extends to maximizing autorepressor production in times of operon induction, in order to optimize effective repression once the inducer is no longer present.

Discussion

Transcriptional regulation in bacteria has been studied for more than half a century. Starting from pioneering works of Jacob and Monod and their colleagues describing regulation of the *lac* operon to nowadays descriptions of hierarchical and modular regulatory networks (reviewed in McAdams et al. 2004) and recurring regulation patterns (network motifs, reviewed in Alon 2007). Clearly, this field remains in the focus of intense research as more and more of its intricacies are revealed. Notwithstanding, these novel insights continue to support the recurrent view that gene control is mainly selected for facilitating bacteria to adjust to changes in their environments in order to optimize their growth. Transcriptional regulation, namely repression, plays a key role in these processes.

Gene organization is clearly an important aspect of transcriptional regulation (reviewed in Lawrence 2003), yet current knowledge concerning gene organization at the operon level is relatively limited. Our initial observation was that TFs localize preferentially at terminal operon positions in *E. coli* and in the evolutionary remote bacteria *B. subtilis*. To better understand this phenomenon, we refined our analysis by distinguishing between autoregulators and alloregulators and then between autorepressors and autoactivators. In each of these steps, the signal for selection toward nonrandom TF localization was evident and statistically supported. Our phylogenetic analysis was further able to show that this signal is not limited to these two bacteria, but rather, is common to nearly the entire bacterial domain. Our results thus reveal that regulation is a major determinant in operon structure, despite evolutionary flexibility in the operon structure (Mushegian and Koonin 1996; Itoh et al. 1999; Fondi et al. 2009).

Our phylogenetic analysis necessitated a main premise that the type of DNA-binding domain of a TF, which is based on sequence information, is a good proxy for the type of the regulatory role that the TF performs. There are several examples of TFs with the same type of DNA-binding domain that have close functional relatedness (e.g., Weickert and Adhya 1992; Alekshun and Levy 1999; Busenlehner et al. 2003; Ramos et al. 2005; Elgrably-Weiss et al. 2006; Elsholz et al. 2010), rendering our premise valid and our conclusions robust. In fact, this premise only limits the resolution of our analysis since several of the DNA-binding domains include TFs with considerably diverse regulatory roles and accordingly show a mixed phylogenetic signal of preferred operonic location (e.g., the RpiR DNA–binding domain, fig. 2). Given finer annotation regarding the modes of regulation of the TFs analyzed in this work, we can expect the phylogenetic signal we have detected to be even better resolved.

The strong operonic location preference of autorepressors spans both highly diverged phyla as well as biologically diverse functions. This striking observation suggests that operonic localization of autorepressors is a paradigm and invokes a generalizing explanation of the underlying mechanism responsible for this trend. The model suggested in this work is consistent with preestablished notions regarding expression regulation and transcription dynamics. Namely, it emphasizes the importance of avoiding costly leaky transcription, which strongly supports the economical mode of expression of genes whose induction strongly relies on specific and infrequent environmental conditions. An alternative model that explains autoregulator preference for the first operon position could have been selection for proximity of the TF to its binding site. This explanation is in general agreement with the observed colocalization of alloregulators with their targets (Warren and ten Wolde 2004; Seshasayee et al. 2009); however, it should predict preference for the first operon position of both autorepressors and autoactivators. Perhaps the fact that autoactivators do not show preference for the first operon position lends support to our model since leaky transcription in this case would be highly wasteful, which is in contradiction to what our model predicts. This also corroborates with the marked preference that *E. coli* autoactivators show for the last operon position. Nevertheless, there may well be other factors that contribute to the preferential localization of autorepressors not accounted for by our model. For example, the dependence of some autorepressors to work in combination with other regulators in order to achieve their regulatory effect (e.g., *pdh*R in *E. coli* and *fru*R in *B. subtilis*). Furthermore, whether or not an autorepressor regulates divergent operons in addition to the operon in which it is encoded (e.g., *mpr*A in *E. coli* and *iol*R in *B. subtilis*) may also determine its operonic location.

To conclude, our analysis has revealed a striking feature of genomic organization, which is highly conserved in evolution. Our model suggests that biochemical energetic considerations related to efficient transcription regulation are the evolutionary driving force for this trend. This work thus enhances our understanding of the evolutionary importance of TF operonic location.

Supplementary Material

Supplementary material, figures S1–S5, and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

D.Z. is supported by the Safra Bioinformatics Foundation at Tel Aviv University. N.D.R. and T.P. are supported by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606. This study was supported by an Israel Science Foundation grant 878/09 to T.P.

References

Alekshun MN, Levy SB. 1997. Regulation of chromosomally mediated multiple antibiotic resistance: the *mar* regulon. *Antimicrob Agents Chemother*. 41:2067–2075.

Alekshun MN, Levy SB. 1999. The *mar* regulon: multiple resistance to antibiotics and other toxic chemicals. *Trends Microbiol*. 7:410–413.

Alon U. 2007. Network motifs: theory and experimental approaches. *Nat Rev Genet*. 8:450–461.

Bai J, Wang J, Xue F, et al. (12 co-authors). 2010. proTF: a comprehensive data and phylogenomics resource for prokaryotic transcription factors. *Bioinformatics* 26:2493–2495.

Belitsky BR, Gustafsson MC, Sonenshein AL, Von Wachenfeldt C. 1997. An *lrp*-like gene of *Bacillus subtilis* involved in branched-chain amino acid transport. *J Bacteriol*. 179:5448–5457.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2011. GenBank. *Nucleic Acids Res*. 39:D32–D37.

Blumenthal T. 1998. Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* 20:480–487.

Busenlehner LS, Pennella MA, Giedroc DP. 2003. The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. *FEMS Microbiol Rev*. 27:131–143.

Carlin A, Shi W, Dey S, Rosen BP. 1995. The *ars* operon of *Escherichia coli* confers arsenical and antimonial resistance. *J Bacteriol*. 177:981–986.

Danchin A, Ullmann A. 1980. The coordinate expression of polycistronic operons in bacteria. *Trends Biochem Sci*. 5:51–52.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*. 23:324–328.

Davids W, Zhang Z. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. *BMC Evol Biol*. 8:23.

Dehal PS, Joachimiak MP, Price MN, et al. (13 co-authors). 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res*. 38:D396–D400.

Elgrably-Weiss M, Schlosser-Silverman E, Rosenshine I, Altuvia S. 2006. DeoT, a DeoR-type transcriptional regulator of multiple target genes. *FEMS Microbiol Lett*. 254:141–148.

Elsholz AK, Michalik S, Zuhlke D, Hecker M, Gerth U. 2010. CtsR, the Gram-positive master regulator of protein quality control, feels the heat. *EMBO J*. 29:3621–3629.

Fisher SH. 1999. Regulation of nitrogen metabolism in *Bacillus subtilis*: vive la difference! *Mol Microbiol*. 32:223–232.

Fondi M, Emiliani G, Fani R. 2009. Origin and evolution of operons and metabolic pathways. *Res Microbiol*. 160:502–512.

Gama-Castro S, Salgado H, Peralta-Gil M, et al. (28 co-authors). 2011. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*. 39:D98–D105.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 31:3784–3788.

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

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*. 3:318–356.

Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N. 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*. 33:6083–6089.

Keseler IM, Bonavides-Martinez C, Collado-Vides J, et al. (14 co-authors). 2009. EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res*. 37:D464–D470.

Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA. 2007. How gene order is influenced by the biophysics of transcription regulation. *Proc Natl Acad Sci U S A*. 104:13948–13953.

Kummerfeld SK, Teichmann SA. 2006. DBD: a transcription factor prediction database. *Nucleic Acids Res*. 34:D74–D81.

Kurland CG. 1992. Translational accuracy and the fitness of bacteria. *Annu Rev Genet*. 26:29–50.

Lawrence JG. 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol*. 5:355–359.

Lawrence JG. 2002. Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell* 110:407–413.

Lawrence JG. 2003. Gene organization: selection, selfishness, and serendipity. *Annu Rev Microbiol*. 57:419–440.

Lee HJ, Jeon HJ, Ji SC, Yun SH, Lim HM. 2008. Establishment of an mRNA gradient depends on the promoter: an investigation of polarity in gene expression. *J Mol Biol*. 378:318–327.

Li Y, Altman S. 2004. Polarity effects in the lactose operon of *Escherichia coli*. *J Mol Biol*. 339:31–39.

McAdams HH, Srinivasan B, Arkin AP. 2004. The evolution of genetic regulatory systems in bacteria. *Nat Rev Genet*. 5:169–178.

Menninger JR. 1976. Peptidyl transfer RNA dissociates during protein synthesis from ribosomes of *Escherichia coli*. *J Biol Chem*. 251:3392–3398.

Miethke M, Hecker M, Gerth U. 2006. Involvement of *Bacillus subtilis* ClpE in CtsR degradation and protein quality control. *J Bacteriol*. 188:4610–4619.

Murakawa GJ, Kwan C, Yamashita J, Nierlich DP. 1991. Transcription and decay of the *lac* messenger: role of an intergenic terminator. *J Bacteriol*. 173:28–36.

Mushegian AR, Koonin EV. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet*. 12:289–290.

Ogasawara H, Ishida Y, Yamada K, Yamamoto K, Ishihama A. 2007. PdhR (pyruvate dehydrogenase complex regulator) controls the respiratory electron transport system in *Escherichia coli*. *J Bacteriol*. 189:5534–5541.

Perez JC, Groisman EA. 2009. Transcription factor function and promoter architecture govern the evolution of bacterial regulons. *Proc Natl Acad Sci U S A*. 106:4319–4324.

Perez-Rueda E, Collado-Vides J. 2000. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res*. 28:1838–1847.

Price MN, Huang KH, Arkin AP, Alm EJ. 2005. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res*. 15:809–819.

Quail MA, Guest JR. 1995. Purification, characterization and mode of action of PdhR, the transcriptional repressor of the *pdh*R-*ace*EF-*lpd* operon of *Escherichia coli*. *Mol Microbiol*. 15:519–529.

Ramos JL, Martinez-Bueno M, Molina-Henares AJ, Teran W, Watanabe K, Zhang X, Gallegos MT, Brennan R, Tobes R. 2005. The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev*. 69:326–356.

Rocha EP. 2008. The organization of the bacterial genome. *Annu Rev Genet*. 42:211–233.

Romero PR, Karp PD. 2004. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics* 20:709–717.

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. 2000. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A*. 97:6652–6657.

Sayers EW, Barrett T, Benson DA, et al. (42 co-authors). 2011. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 39:D38–D51.

Schreier HJ, Brown SW, Hirschi KD, Nomellini JF, Sonenshein AL. 1989. Regulation of *Bacillus subtilis* glutamine synthetase gene expression by the product of the *gln*R gene. *J Mol Biol*. 210:51–63.

Schulz A, Schumann W. 1996. *hrc*A, the first gene of the *Bacillus subtilis dna*K operon encodes a negative regulator of class I heat shock genes. *J Bacteriol*. 178:1088–1093.

Seshasayee AS, Fraser GM, Babu MM, Luscombe NM. 2009. Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*. *Genome Res*. 19:79–91.

Stock AM, Robinson VL, Goudreau PN. 2000. Two-component signal transduction. *Annu Rev Biochem*. 69:183–215.

Ullmann A, Joseph E, Danchin A. 1979. Cyclic AMP as a modulator of polarity in polycistronic transcriptional units. *Proc Natl Acad Sci U S A*. 76:3194–3197.

van Hijum SA, Medema MH, Kuipers OP. 2009. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol Mol Biol Rev*. 73:481–509.

Warren PB, ten Wolde PR. 2004. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. *J Mol Biol*. 342:1379–1390.

Weickert MJ, Adhya S. 1992. A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem*. 267:15869–15874.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*. 36:D88–D92.

Wunderlich Z, Mirny LA. 2008. Spatial effects on the speed and reliability of protein-DNA search. *Nucleic Acids Res*. 36: 3570–3578.