Genetica (2005) 124: 145–166
DOI 10.1007/s10709-005-0950-5

Evolutionary conservation of bacterial operons: does transcriptional connectivity matter?

Einat Hazkani-Covo[1] & Dan Graur[1,2,*]
[1]*Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel;* [2]*Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA;* **Author for correspondence (E-mail: dgraur@uh.edu)*

Received 31 October 2004 Accepted 7 January 2005

Key words: bacterial operons, evolutionary rates, transcriptional connectivity

Abstract

In the literature, it has been frequently suggested that the connectivity of a protein, i.e., the number of proteins with which it interacts, is inversely correlated with the rate of evolution. We attempted to extrapolate from proteins to operons by testing the hypothesis that operons with high transcriptional connectivity, i.e., operons that are controlled through interactions with many transcription factors, are evolutionarily more conserved at the structure and sequence levels than low-connectivity operons. With *Escherichia coli* used as reference, two structural- and two sequence-conservation measures were determined for 82 groups of homologous operons from 30 completely-sequenced bacterial genomes. In *E. coli*, large operons tend to be regulated by more transcription factors than either smaller operons or single genes. Large *E. coli* operons that are regulated by single transcription factors were found to be regulated by activators more frequently than by repressors. Levels of sequence conservation and structural conservation of operons were found to be independent of each other, i.e., structurally conserved operons may be divergent in sequence, and vice versa. Transcriptional connectivity was found to influence neither sequence nor structural conservation of operons. Although this finding seems to contradict the situation in genes, a critical review of the literature indicates that although gene connectivity is frequently touted as a factor in determining rates of evolution, only a very small fraction of the variability in degrees of evolutionary conservation is explainable by this factor.

Introduction

The relationship between connectivity and evolutionary conservation has been a focus of large number of studies (e.g., Clarke, Mittenthal & Senn, 1993; Thattai & van Oudenaarden, 2001; Fraser et al., 2002; Bastolla et al., 2003; Hahn, Conant & Wagner, 2004). Connectivity of a protein is usually measured by the number of proteins with which it interacts, i.e., the number of edges associated with a protein node within a protein–protein interaction network. There are several measures of conservation in proteins, the most common of which is the rate of amino-acid replacement. We note that connectivity is usually defined for a single model organism, almost invariably a well studied simple unicellular organism. By necessity, therefore, in comparative studies it is implicitly assumed that the connectivity in the organisms for which such knowledge is lacking is the same as that in the model organism. Degrees of evolutionary conservation, on the other hand, can only be defined comparatively, i.e., data on two or more organisms are needed.

Several authors have studied the relationship between protein connectivity and evolutionary rates. From comparisons involving *Saccharomyces*

cerevisiae, Schizosaccharomyces pombe, Candida albicans, and *Caenorhabditis elegans*, it was concluded that highly connected proteins evolve slower than sparsely connected ones (Fraser et al., 2002; Fraser, Wall & Hirsh, 2003). This correlation was said to be independent of protein essentiality. The notion of a negative correlation between protein connectivity and evolutionary rate was challenged by Jordan et al. (2003a, b), who claimed that the correlation appears to be due to a few highly interactive proteins that evolve exceptionally slow. Moreover, in *Escherichia coli*, Hahn et al. (2004) reported no effect of metabolic connectivity on evolutionary rate.

Many factors beside protein connectivity affect the rates of molecular evolution of a gene. The most important factor may be functional constraint (Graur & Li, 2000), with genes having different functions and, hence, different selective constraints, evolving at different rates. Other factors that may determine the evolutionary rate of a gene are gene essentiality (Wilson, Carlson, & White, 1977; Hirsh & Fraser, 2001), level of gene expression (Pal, Papp, & Hurst, 2001), and gene duplication (Yang, Gu, & Li, 2003). The physical position within the genome might also affect the evolutionary rate of a gene (Williams & Hurst, 2000). The relative contributions of each of these factors on the molecular evolution of a gene are not known at present.

It was recently suggested that sequence conservation should not be used exclusively to measure evolutionary conservation. Krylov et al. (2003) used a gene's propensity to be lost during evolution as a measure of evolutionary conservation. With this measure, they found that proteins involved in many interactions are lost less frequently during evolution than proteins with fewer interactions.

In this study we extrapolate from genes to higher order genetic structures and study the relationship between connectivity and evolutionary conservation in operons. By analogy to the rules said to govern the evolution of single genes, we hypothesize that operons with high degrees of connectivity will be more conserved than operons with low degrees of connectivity. We note, however, that the extrapolation is not straightforward, since the measures of connectivity and evolutionary conservation that are useful as far as genes are concerned are not directly applicable to operons. We, therefore, modified the definitions and adjusted them to operons (see below).

The architecture of bacterial operons is seldom conserved during evolution (Mushegian & Koonin, 1996; Siefert et al., 1997; Watanabe et al., 1997; Itoh et al., 1999). Bacterial genomes can be characterized by three properties: gene makeup, gene order, and sequence. Studies comparing these three properties indicate that gene order is the least conserved property (Huynen & Bork, 1998), although its degree of conservation is positively correlated with those of the other two properties (Wolf et al., 2001).

In order to ascertain whether operon connectivity is related to operon conservation, we used the RegulonDB database (Salgado et al., 2000, 2004) as a source of information on operons and transcription factors in *E. coli*. We, then, compared the *E. coli* operons to complete or partial counterparts in 30 completely sequenced bacterial genomes.

Definitions

Operon connectivity was defined as the number of transcription factors that bind to the promoter of an operon and regulate its transcription. Operon connectivity data is only available for well-studied model organisms. In this study, we use data from *E. coli*.

Operon conservation can be defined at the sequence level, as is the practice in genes, or at the structural level. We used two *sequence conservation measures* for operons. The first was the mean evolutionary rate for all the genes within the operon. This measure may be problematic because of the variation in rates among the genes within the operon. To sidestep this difficulty, we devised a second measure, the evolutionary rate of the most conserved gene within the operon. In both cases, we used the number of amino-acid replacements as calculated from sequence comparisons with the homologous protein products from *E. coli*. Two *structural conservation measures* were used: *operon identity*, defined as the existence of an operon with the same gene makeup and geneorder as its homolog in *E. coli*, and *operon similarity*, defined as the existence of an operon whose gene complement resembles but is not identical to that of its homolog in *E. coli*. For a

detailed description of the manner in which these variables were calculated, see Data and Methods.

We tested the *operon-connectivity/operon-conservation hypothesis* that predicts that operons that are regulated by several transcription factors should be more conserved during evolution than those that bind a single transcription factor.

Data and methods

Operons, transcription units, and transcription factors in Escherichia coli

We used the 1/2003 version of RegulonDB in XML format (Salgado et al., 2000, 2004) for identifying and linking operons, transcription units and transcription factors in *E. coli K-12*. RegulonDB includes both operon and transcription unit data. A transcription unit differs from an operon because by definition an operon must contain two or more genes, whereas a transcription unit may also contain a single gene. An operon may also include several promoters to which transcription factors may bind, whereas a one-to-one relationship exists between transcription units and promoters (Karp et al., 2002).

We first compiled a dataset of 237 entries consisting of single-transcription-unit operons and of transcribed single genes that do not belong to operons, for which the number of transcription factors is known. The category of 'transcribed single genes that do not belong to operons' may be thought of as single-transcription-unit operons of length 1. We note that all the entries in this dataset are independent of one another and, hence, no complicated statistical precautions should be taken in subsequent analyses. We shall, henceforth, refer to this compilation as dataset I.

We next compiled a reduced dataset for which not only the number of transcription factors, but also the type of regulation (activation or repression) are known. After omitting all entries with either unknown or dual (activation and repression) regulation, we are left with 214 entries in dataset II.

Dataset III was derived from dataset I by removing all single gene transcripts. The number of operons in this dataset is 140.

Orthology search

The COG (Clusters of Orthologous Groups) database (Tatusov, Koonin & Lipman, 1997; Tatusov et al., 2001) was used for identification of orthologous proteins from different bacterial species. Since the COG accession is not part of RegulonDB, for each gene in database III, we first translated the Blattner number (*b*-number) in RegulonDB into its corresponding COG number. This translation was performed using version 7/2003 of the *E. coli* K-12 protein table file (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K12/NC_000913_ptt). *E. coli* operons for which not all b numbers had corresponding COG numbers were removed from further analysis. The total number of *E. coli* operons used in the comparative part of the study (dataset IV) was 82. The COG accession number of each protein in dataset IV was, then, used to locate its orthologs in the other bacterial genomes. The bacterial protein table files including this information were downloaded from ftp://ncbi.nih.gov/genomes/Bacteria/.

Bacterial genomes

The COG database consists of proteins from 50 completely sequenced genomes. Taxonomic exclusions from our study were based on several criteria: (a) In cases in which several strains from the same species were listed, we selected the one with the highest number of genes associated with COG accession numbers, and discarded the other strains. (b) Strains of *E. coli* other than the reference strain were excluded. (c) Bacterial with either unspecified taxonomic affiliation or defined in the COG database as 'Bacteria,' were omitted. In addition, we decided not to include members of higher taxa of dubious monophyly, such as members of the beta and delta/epsilon proteobacteria, which for unknown reasons are clumped under the unspecific name 'Proteobacteria.' After the 'triage,' our study comprised of 30 bacterial genomes (other than *E. coli*), classified into four taxa (see Table 1). In cases in which the genomic sequence was compartmentalized into several fragments, chromosomes or plasmids, the sequence was linearized and concatenated.

Table 1. Bacterial taxa used in the comparative analysis. Each entry is listed as in the COG database. Species abbreviations are listed in parentheses. Taxa above the species level are in boldface with numbers of constituent species listed in parentheses

Taxon	GenBank Accession Number
Alpha proteobacteria (7)	
Agrobacterium tumefaciens C58 Cereon (Atu)	NC_003062, NC_003063, NC_003064, NC_003065
Brucella melitensis (Bme)	NC_003317, NC_003318
Caulobacter crescentus (Ccr)	NC_002696
Mesorhizobium loti (Mlo)	NC_002678, NC_002679, NC_002682
Rickettsia conorii (Rco)	NC_003103
Rickettsia prowazekii (Rpr)	NC_000963
Sinorhizobium meliloti (Sme)	NC_003047, NC_003078, NC_003037
Gamma proteobacteria (8)	
Buchnera sp. (Buc)	NC_002528
Haemophilus influenzae (Hin)	NC_000907
Pasteurella multocida (Pmu)	NC_002663
Pseudomonas aeruginosa (Pae)	NC_002516
Salmonella typhimurium LT2 (Sty)	NC_003197, NC_003277
Vibrio cholerae (Vch)	NC_002505, NC_002506
Xylella fastidiosa (Xfa)	NC_002488, NC_002490
Yersinia pestis CO92 (Ype)	NC_003143, NC_003131, NC_003134
Gram plus bacteria (12)	
Bacillus halodurans (Bha)	NC_002570
Bacillus subtilis (Bsu)	NC_000964
Clostridium acetobutylicum (Cac)	NC_003030, NC_001988
Lactococcus lactis (Lia)	NC_002662
Listeria innocua (Lin)	NC_003212, NC_003383
Mycoplasma genitalium (Mge)	NC_000908
Mycoplasma pneumoniae (Mpn)	NC_000912
Mycoplasma pulmonis (Mpu)	NC_002771
Staphylococcus aureus N315 (Sau)	NC_002745, NC_003140
Streptococcus pneumoniae TIGR4 (Spn)	NC_003028
Streptococcus pyogenes (Spy)	NC_002737
Ureaplasma urealyticum (Uur)	NC_002162
Actinobacteria (3)	
Corynebacterium glutamicum (Cgl)	NC_003450
Mycobacterium leprae (Mle)	NC_002677
Mycobacterium tuberculosis H37Rv (Mtu)	NC_000962

Identification of homologous operons

We looked for the appearance of orthologous genes from each *E. coli* operon in the other bacterial genomes. A homologous operon was defined as a collection of orthologous genes to those within an *E. coli* operon that appeared in close proximity to one another in the other genome. A homologous operon was defined as such even if its operonic identity was not confirmed experimentally. Structural similarities between an operon in *E. coli* and its homologous counterpart in another bacterium were classified into four groups according to gene order conservation, by

using slightly modified definitions from Itoh et al. (1999): (a) The highest degree of structural conservation is 'identity,' whereby both gene makeup and gene order are the same as in the *E. coli* counterpart. Identity, as far as gene order is concerned, means that the relative positions of an operon's constituent genes relative to the mRNA transcript have been strictly preserved. In our analysis, operons with internal gene duplications are also considered as identical. For example, operons ABC and ABBC are considered identical. (b) An operon structure is defined as 'similar' if the two operons differ from each other by internal translocations, deletions, and at most two insertions. (c) An operon structure was defined as 'destructed' if two or more orthologs of the genes within an *E. coli* operon were found in the other genome, but the operon itself was not. (d) An operon structure was defined as 'unknown' if no orthologous genes or at most one orthologous gene from the *E. coli* operon was found in the other genome. We divided the 'unknown' category into two subcategories depending on whether one or no orthologous genes were found.

Structural conservation scores

All structural conservation scores were defined per taxonomic group, rather than for pairs of taxa. We used three conservation scores (expressed as percentages): (a) *Unweighted score for identical operons* was defined as the number of bacterial species within a taxonomic group in which the operon under study is identical, divided by the size of the taxonomic group. (b) *Weighted score for identical operons* was defined as the number of bacteria in which the operon is identical, divided by the number of bacteria whose genomes contain all the genes in the operon. (c) *Weighted score for similar operons* was defined as the number of bacteria with a similar or an identical operon in their genome, divided by the number of bacteria whose genome contains at least two genes from this operon. All three scores were calculated for each operon in each of four taxonomic groups. We used Spearman non-parametric correlation test and Mann-Whitney non-parametric test to test the difference between any two groups.

Sequence conservation scores

Proteins derived from *E. coli* operons that had an identical counterpart in at least one other species were extracted from the *E. coli* K-12 protein FASTA file (ftp://ftp.ncbi.nih.gov/genomes/ Bacteria/Escherichia_coli_K12/NC_000913.faa). Names of genes from the other bacteria were extracted from protein table files according to their COG numbers. Protein sequences were extracted from the protein FASTA files (ftp://ncbi.nih.gov/genomes/Bacteria/).

Orthologous proteins were aligned using ClustalW (Higgins, Thompson & Gibson, 1996). Pairwise protein distances were only calculated for proteins derived from identical operons. The distances were calculated by using the PROTDIST program from the PHYLIP package (Felsenstein, 1993) with the PAM matrix. In cases in which more than one conserved operon appeared in a bacterial genome, the operon with the average shortest distance to the proteins from the *E. coli* operon was chosen.

Two sequence conservation scores were calculated per operon for each of the four taxonomic groups: (a) mean distance over all the proteins in the operons from all bacteria in a taxonomic group, and (b) mean distance for the most conserved gene within the operon.

Results

Characterization of transcription units in Escherichia coli

We first study the internal relationships between the number of transcription factors and type of regulation in *E. coli*, on the one hand, and transcription-unit size (defined as the number of genes within it), on the other (for definitions, see Data and methods). We used dataset I, which includes 127 single genes and 42, 33, 14, 11, 2, 5, 1, and 2 operons of sizes of 2, 3, 4, 5, 6, 7, 8, and 15, respectively. A significant positive association ($r = 0.15$, $P = 0.015$) between transcription-unit size and number of transcription factors was obtained with the Spearman's rank correlation text. This indicates that, larger transcription units tend to be regulated by more transcription factors.

We used dataset II to test for a possible relationship between the type of regulation (activation

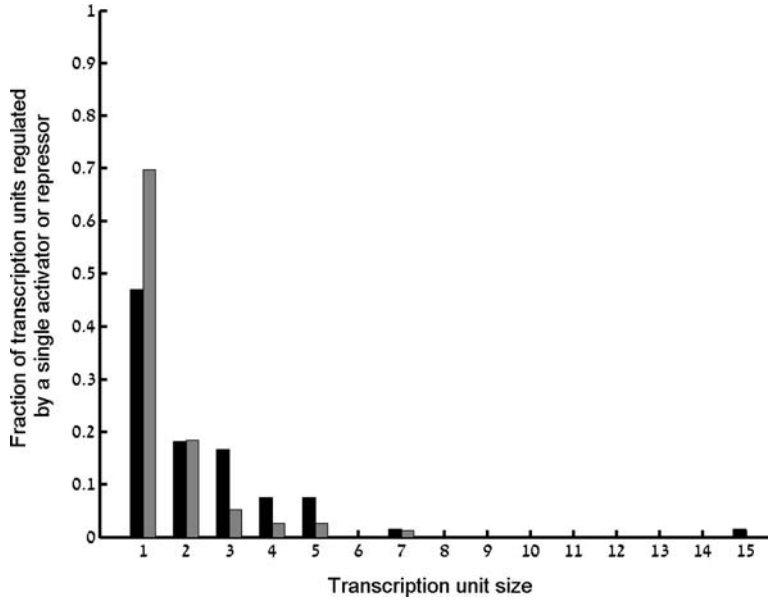

Figure 1. Fraction of transcription units regulated by a single activator (black) or a single repressor (grey).

or repression) and transcription-unit size. For transcription units regulated by a single transcription factor, we found a significant association with type of regulation. That is, larger transcription units were found to be regulated by activators more often than shorter transcription units, which were mainly regulated by repressors (Mann–Whitney test, $n = 142$, $P = 0.002$, see Figure 1). This result was also obtained when self-regulated transcription units, suspected to be mainly regulated by repression (Martinez-Antonio & Collado-Vides, 2003) were removed from the analysis. No such correlation was found between the type of regulation for transcription units regulated by more than a single transcription factor (Kruskal–Wallis nonparametric ANOVA for two transcription factors, $n = 56$; $P = 0.768$; Kruskal–Wallis nonparametric ANOVA for three transcription factors, $n = 12$; $P = 0.800$).

Structural evolutionary conservation of operons

In the comparative part of our study, we used the 82 operons for which all the constituent genes in *E. coli* had COG accessions (dataset IV). Four structural conservation measures were studied: identity, similarity, destruction, and unknown (see Data and methods). Degrees of structural conservation for each of the 82 operons in each of 30 bacterial genomes are listed in Table 2.

Sixty-four of the *E. coli* operons were found in identical gene makeup and gene order in at least one other bacterium. The numbers of *E. coli* operons found in at least one other bacterium were 61, 22, 22, and 7 for gamma proteobacteria, alpha proteobacteria, Gram plus bacteria, and actinobacteria, respectively. As far as similar and identical operons are concerned, 78 operons appeared in at least one genome, while 4 operons did not appear in any of the genomes included in our study. The numbers of similar plus identical operons found in each taxonomic group were 77, 52, 55 and 36 for gamma proteobacteria, alpha proteobacteria, Gram plus bacteria, and actinobacteria, respectively. As expected, the number of conserved operons in the different bacterial taxa increased with phylogenetic relatedness to *E. coli*. For example, the number of conserved operons was the highest in gamma proteobacteria, to which *E. coli* also belongs. The group with the least structural similarity was actinobacteria. These results are in agreement with Itoh et al. (1999), however, due to a different definition of orthologs, our results are not identical to those in their Table 1.

The most conserved operon was *pho*BR (Makino et al., 1986; Wanner & Chang, 1987),

Table 2. Degrees of structural conservation of 82 *Escherichia coli* operons in 30 other bacterial genomes. *L* = number of genes in the *E. coli* operon. *TF* = number of transc- ription factors *in E. coli*. Abbreviations of bacterial species names are as in Table 1. Degrees of structural conservation: I, identical; S similar, D, destructed; X1, only one ortholog was found; X0, no orthologs were found. The numerals following D, S and X denote the number of COG orthologs found in this genome. For example, D2 indicates the operon is destructed, but that two of the constituent genes are found in the genome

Operon	*L*	*TF*	Alpha proteobacteria							Gamma proteobacteria								Gram plus bacteria												Actinobacteria		
			Atu	Bme	Ccr	Mlo	Rco	Rpr	Sme	Buc	Hin	Pmu	Pae	Sty	Vch	Xfa	Ype	Bha	Bsu	Cac	Lla	Lin	Mge	Mpn	Mpu	Sau	Spn	Spy	Uur	Cgl	Mle	Mtu
ahpCF	2	1	X0	X0	I	X1	X0	X0	X0	X0	X0	X0	I	I	X0	I	X0	X0	I	X0	I	X0	X0	X0	X0	I	X0	I	X0	X0	X0	X0
araBAD	3	2	X1	D2	X0	D2	X0	X0	D3	X0	X1	X1	X1	I	X0	X0	S3	S3	S3	S2	X0	X1	X0	X0	X0	X0	X1	X0	X0	X0	X0	X0
argCBH	3	1	D3	D3	D3	D3	X0	X0	D3	S3	X0	I	D3	I	S3	S3	I	S3	S3	D3	S3	S3	X0	X0	X0	S3	X0	X0	X0	S3	S3	S3
aroF-tyrA	2	1	X0	X0	D2	D2	X0	X0	X0	X0	D2	X1	D2	I	I	D2	X1	X0	X0	X0	D2	X0	X0	X0	X0	X0	D2	X0	X0	D2	X0	X0
atoDAE	3	1	S2	S2	S2	S2	S2	X0	S2	X0	I	X0	S3	X0	X0	X0	X0	S3	S2	S2	X0	X0	X0	X0	X0	X0	X0	S3	X0	S3	X0	S2
betIBA	3	2	I	S3	D3	S3	X0	X0	S3	X0	X1	D2	I	D2	S2	X1	I	D2	D2	D2	X1	D2	X0	X0	X0	S3	D2	D2	D2	D2	D3	S3
bioBFCD	4	1	S4	S4	S4	S4	D2	D2	D2	S3	S4	S4	S4	I	I	D4	I	S4	S4	D3	X1	X1	X0	X0	X0	S4	X1	X1	X0	D3	S4	S4
cadBA	2	1	X0	X1	X1	X1	X1	X1	S2	X0	S2	S2	S2	I	I	X1	I	D2	D2	D2	X1	D2	X1	X1	X1	D2	D2	X1	X1	X0	D2	D2
caiTABCDE	6	2	S5	S5	S5	S5	D2	X0	S6	X0	X1	D2	S6	I	D5	D3	D4	S5	S5	S3	X0	D2	X0	X0	X0	S4	X0	D2	X0	D5	D3	S6
codBA	2	1	X1	X1	X1	X1	X0	X0	X1	X0	X0	I	I	I	I	X0	X0	X1	D2	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0
cydAB	2	2	I	I	I	I	I	I	I	X0	I	I	I	I	I	X0	I	I	I	X0	I	I	X0	X0	X0	I	X0	X0	X0	I	X0	I
cynTSX	3	1	X0	D2	X1	X0	X0	X0	X0	X0	X0	X0	S3	D2	X0	X0	X1	X0	D2	D2	X0	X0	X0	X0	X0	X0	D2	X0	X0	D2	X0	X1
cyoABCDE	5	2	S5	S5	S5	S5	S4	S4	S4	I	X0	X0	I	I	X0	S5	I	S5	S5	X0	X0	S5	X0	X0	X0	S5	X0	X0	X0	D2	D4	D4
cysDNC	3	1	S2	S2	S2	S2	X0	X0	S3	I	X0	X0	S3	I	I	S2	I	S2	S2	S3	X0	X0	X0	X0	X0	X0	X0	X0	X0	S2	X0	S2
cysJIH	3	1	D3	S3	S3	S2	X0	X0	D2	I	X0	X0	D3	I	I	I	S3	S3	S3	D2	X0	X0	X0	X0	X0	X0	X0	X0	X0	S2	X0	S2
cysPUWAM	5	1	S5	S5	S4	S4	X0	X0	S4	X0	D2	D2	S4	S5	S3	S4	S4	S4	X1	D2	X1	X0	X0	X0	X0	X1	X0	X0	X0	X1	D2	S4
dadAX	2	1	S2	S2	D2	S2	X0	X0	D2	X0	D2	D2	S2	I	D2	I	D2	D2	D2	D2	X0	D2	X0	X0	X0	X1	D2	D2	X0	D2	D2	D2
dmsABC	3	2	X1	X1	X1	I	X0	X0	S2	X0	I	I	D2	I	D2	X0	I	X1	X1	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X1
dppABCDF	5	1	I	I	X0	I	X0	X0	I	X0	S4	S4	S4	S5	S4	X0	S5	I	S4	S4	S4	S4	S3	S3	S4	I	S3	I	S4	S3	S3	S3
ebgAC	2	2	X1	X0	X0	X0	X0	X0	X1	X0	X0	X1	X0	X1	X0	X0	D2	X0	X0	X0	X0	X0	X0	X0	X0	X0	D2	X0	X0	X0	X0	X0
edd-eda	2	1	D2	S2	D2	D2	X0	X0	S2	X0	D2	D2	D2	I	D2	I	D2	D2	D2	D2	D2	X0	X0	X0	X0	X0	D2	X0	X0	X0	X0	X0
entCEBA	5	1	D2	S4	D2	D2	X1	X1	D2	X0	D3	D3	S4	I	S4	X1	D5	S2	S5	X1	D3	S3	X0	X0	X0	S3	D2	D2	X0	D3	D2	D5
epd-pgk	2	1	S2	I	I	S2	X0	X0	S2	D2	D2	D2	I	I	I	D2	I	I	I	I	D2	I	I	I	D2	I	D2	D2	X0	I	I	I
fadBA	2	2	D2	D2	D2	D2	D2	D2	S2	X0	X0	X0	I	S2	X1	X0	I	I	S2	D2	X1	X0	X0	X0	X0	X1	X0	X1	X0	X0	S2	S2
fdnGHI	3	2	X1	X1	X1	S2	X0	X0	I	X0	S3	I	S3	I	S3	X0	S3	X1	X1	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X1
fecABCDE	5	1	S5	S5	S4	S5	X0	X0	S5	X0	S4	S5	S5	S5	S5	X1	S5	S4	S4	S4	S3	S4	X0	X0	X0	S4	S4	S4	S3	S4	X0	X1
fecIR	2	2	I	X1	I	X1	X0	X0	I	X0	X0	X1	I	X0	X1	X0	X0	X1	X1	X1	X0	X1	X0	X0	X0	X1	X0	X0	X0	X1	X1	X1
fepA-entD	2	1	D2	D2	X1	D2	X0	X0	X1	X0	X1	X1	D2	D2	D2	X1	D2	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0

Table 2. (Continued)

Operon	*L*	*TF*	Alpha proteobacteria							Gamma proteobacteria								Gram plus bacteria												Actinobacteria		
			Atu	Bme	Ccr	Mlo	Rco	Rpr	Sme	Buc	Hin	Pmu	Pae	Sty	Vch	Xfa	Ype	Bha	Bsu	Cac	Lla	Lin	Mge	Mpn	Mpu	Sau	Spn	Spy	Uur	Cgl	Mle	Mtu
fepDGC	3	1	S3	S3	S2	S3	X0	X0	S3	X0	S2	S3	S3	S3	S3	X0	S3	S3	S3	S3	D2	S3	X0	X0	X0	S3	S3	S3	S3	S3	X0	X0
fhuACDB	4	1	S4	S4	S4	S4	X0	X0	S4	X0	S4	S4	I	I	I	X1	S4	S3	S3	S3	S2	S3	X0	X0	X0	S3	S3	S3	S2	S3	X0	X1
fixABCX	4	1	S3	S3	I	X0	X0	I	X0	X0	X0	S3	I	X0	S3	X0	S2	S2	S3	X0	X0	X0	X0	X0	X0	X0	X0	X0	S3	S3	S3	S3
frdABCD	4	2	S2	S2	S2	S2	D2	D2	S2	X0	I	I	S2	I	I	S2	I	S2	S2	X0	X0	X0	X0	X0	X0	S2	X0	X0	X0	S2	S2	I
fruBKA	3	1	X0	X0	X0	X1	X0	X0	X0	X0	I	S3	D2	S2	S2	X0	D2	D2	D2	D2	D2	D2	D2	D2	X0	D2	D2	D2	X0	S2	X0	X0
ftsQAZ	3	1	I	I	I	I	S3	S3	I	S2	I	I	I	I	I	I	I	S3	S3	S3	S3	I	X0	X0	X0	I	S3	I	X0	S2	S2	S2
fucAO	2	2	X1	D2	X0	D2	X0	X0	D2	X0	X1	X1	D2	I	D2	X0	I	D2	D2	D2	X0	D2	X0	X0	X0	X0	D2	X0	X0	X1	X0	X0
fucPIKUR	5	2	D2	D3	X1	S2	X0	X0	S2	X0	S4	D3	D2	S5	D2	X1	D4	S3	S4	S3	D3	D2	X0	X0	X0	D3	D3	X1	X0	S3	X0	X0
gcvTHP	3	3	S2	I	S3	I	X0	X0	I	X0	X0	X0	S2	I	X0	S2	S2	S3	S3	X0	X0	S3	X0	X0	X0	S3	X0	X0	X0	X0	D2	D2
glnHPQ	3	2	I	I	I	I	D3	D3	I	X0	S2	X0	I	I	I	X0	I	I	I	I	I	S3	X0	X0	X0	I	3	S3	X0	I	X0	X0
glpACB	3	4	D2	D2	X0	D2	X0	X0	D2	X0	I	I	D2	I	I	X0	I	D2	D2	X0	D2	X0	X0	X0	X0	X0	X0	X0	X0	X0	D2	D2
glpFK	2	2	D2	D2	D2	D2	X0	X0	D2	X0	I	I	I	I	X0	I	D2	I	I	S2	S2	I	D2	D2	S2	I	S2	S2	X0	X0	X0	X0
glpTQ	2	3	X1	X0	X1	X1	X0	X0	X1	X0	S2	I	D2	I	I	X0	D2	X1	I	X1	D2	X1	X1	X1	X0	D2	X0	D2	X0	X0	X0	X1
guaBA	2	1	X1	X1	X1	X1	X0	X0	X1	X0	X0	X1	X1	X1	X1	I	D2	D2	D2	X1	X1	X1	X0	X0	X0	X0	D2	D2	X0	X0	X1	D2
ilvIH	2	1	I	I	I	I	X0	X0	I	I	I	I	I	I	I	X0	I	I	I	D2	I	I	X0	X0	X0	I	I	X0	X0	I	I	I
kbl-tdh	2	1	D2	X1	D2	I	X0	X0	I	X0	D2	X0	D2	I	I	D2	I	D2	S2	X0	X0	X1	X0	X0	X0	D2	X1	X0	X0	X0	X0	D2
kdpABC	3	1	I	X0	I	I	X0	X0	I	X0	X0	X0	I	I	X0	X0	I	X0	X0	I	X0	I	X0	X0	X0	I	X0	X0	X0	X0	X0	I
lacZYA	3	2	D3	D2	D2	D2	D2	X1	D3	X1	X1	D2	S2	D2	D2	D2	D2	D3	D2	D2	S2	D2	X1	X1	X1	D2	D3	D3	X0	D2	D2	D2
lexA-dinF	2	1	D2	D2	D2	D2	X0	X0	D2	X0	D2	D2	D2	I	S2	D2	D2	D2	D2	D2	D2	D2	X0	X0	X1	D2	X1	X0	X1	D2	X0	D2
malEFG	3	2	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	S2	S2	X0	S2	S2	S2	X0	S2	S2	X0	X0	X0	S2	S2	S2	X0	X0	X0	X0
malPQ	2	2	X0	X0	X0	X0	X0	X0	X0	X0	D2	D2	D2	I	I	X0	D2	X0	X0	X1	D2	X0	X0	X0	X0	X0	S2	S2	X0	D2	X0	D2
malXY	2	2	X0	X0	X0	X0	X0	X0	X0	X0	X0	X1	X0	D2	X1	X0	X0	D2	D2	D2	D2	D2	X0	X0	X1	X1	D2	X0	X0	D2	X0	X1
manXYZ	3	3	X0	X0	X0	X0	X0	X0	X0	X0	X0	S2	X0	I	X0	X0	I	X0	I	I	S2	I	X0	X0	X0	X0	I	I	X0	X0	X0	X0
MelAB	2	2	X1	X0	X1	X1	X0	X0	X1	X0	X0	X0	X0	I	X0	X0	D2	X1	D2	D2	X1	X1	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0
mglBAC	3	2	I	S3	I	I	X0	X0	I	X0	S3	S3	I	S3	S3	X0	I	I	S3	X1	S3	X0	X0	X0	X0	X0	X0	X0	X0	S3	X0	X0
mtlADR	3	2	X1	X0	X0	X1	X0	X0	X1	X0	X0	I	X0	I	I	X0	I	S2	S2	S2	X1	X0	X0	S2	S2	S2	S2	X0	X0	X0	X0	X0
nagBACD	4	2	S4	D4	D4	D4	X0	X0	D4	X0	S3	S3	D3	I	S3	D2	I	S4	S4	S3	D4	S4	X0	X0	D3	D4	D4	D4	X0	D3	D2	D4
operon[1]	15	2	S13	S9	S9	S11	D5	D5	S14	X1	S13	S15	S15	S15	S15	S8	S14	S5	S6	S6	D3	D3	X0	D2	D2	D4	D3	D3	D2	S3	S3	S4
narGHJI	4	3	X1	S4	X1	X1	X0	X0	X1	X0	X1	X1	S4	S4	X1	X0	X1	X1	S4	X0	X0	X0	X0	X0	X0	S3	X0	X0	X0	S3	X0	S4

Table 2. (Continued)

narXL	2	1	X1	X1	X0	X1	X0	X0	X1	X0	D2	D2	I	I	D2	X1	D2	X1	X1	X1	X0	X1	X0	X0	X0	X1	X1	X1	X0	X1	X0	X1
nirBDC-cysG	4	2	S3	X0	S2	S3	X0	X0	S2	X0	X0	X0	S4	I	X1	X0	S3	S3	S4	D3	X0	D2	X0	X0	X0	S4	X0	X0	X0	X0	X0	S2
nrdAB	2	2	I	I	D2	X0	D2	I	X0	I	X0	S2	I	I	I	I	I	I	I	I	I	I	S2	S2	S2	I	I	I	X0	D2	S2	S2
nrfABCDEFG	7	2	S3	S3	S3	S4	D2	D2	S4	X0	S4	S6	S4	S6	D3	S3	S4	X1	X1	X1	X0	X1	X0	X0	X0	X1	X0	X0	X0	X0	X0	X0
oppABCDF	5	1	I	I	X0	I	X0	X0	I	X0	S4	S4	S4	S5	S4	X0	S5	I	S4	S4	S4	S4	S3	S3	S4	I	S3	I	S4	S3	S3	S3
operon[2]	15	1	S15	D7	S7	S15	D4	D3	S15	X0	D2	D2	S15	D3	D5	D2	S14	S8	D4	S4	S8	S4	D2	D2	S5	S8	D5	S5	D3	S8	X0	D4
phoBR	2	1	I	I	I	I	I	I	I	X0	I	I	I	I	I	I	I	I	I	I	I	I	X0	X0	X0	I	I	I	X0	I	I	I
operon[3]	5	1	I	I	S5	I	X0	X0	S5	X0	S3	S4	S3	I	S5	I	I	I	S4	I	S5	S5	S4	S4	X0	I	I	S5	S4	S3	S5	S5
purEK	2	1	I	I	I	I	X0	X0	I	X0	I	I	I	I	I	I	I	I	I	X0	I	I	X0	X0	X0	I	I	I	X0	S2	S2	S2
purHD	2	1	D2	D2	D2	D2	X0	X0	D2	X0	I	S2	I	I	I	I	I	I	I	I	D2	I	X0	X0	X0	I	I	X0	X0	D2	D2	D2
purMN	2	1	I	I	I	I	X0	X0	I	X0	I	I	I	I	I	S2	I	I	I	I	I	I	X0	X0	X0	I	I	I	X0	D2	D2	D2
rbsDACBK	5	2	S4	S4	S4	S4	X0	X0	S4	X0	I	I	S4	I	I	X1	S5	S5	S5	D2	S5	X0	X0	X0	D2	S2	X0	X0	X0	S5	D2	X1
rhaSR	2	1	D2	D2	D2	D2	X0	X0	D2	X0	D2	X0	D2	D2	D2	X0	D2	D2	D2	D2	D2	D2	X0	X0	X0	D2	D2	X0	X0	D2	X0	D2
speA	2	1	X1	X1	X0	X1	X0	X0	X1	X0	X0	I	D2	D2	I	D2	X0	X1	X1	X0	X0	X0	X0	X0	X0	X0	X1	X0	X0	X0	X0	X0
tauABCD	4	2	S3	S3	S3	S4	X0	X0	S3	X0	S3	S3	S4	X0	D2	X0	S4	S3	S3	S3	X0	X0	X0	X0	X0	S3	S3	X0	X0	S3	X0	X1
tdcABCDEFG	7	6	D4	D5	S4	D5	D2	D2	D5	D2	S7	S7	S6	S7	S7	D3	S7	D5	D5	D5	S6	D6	X0	X0	X0	D6	D6	D5	X0	D6	D4	D5
torCAD	3	2	S2	X1	X1	X1	X0	X0	S2	X0	S3	I	2	I	S3	X0	S3	X1	X1	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X1
treBC	2	1	X1	X0	D2	X1	X0	X0	X1	X0	X0	X1	D2	D2	D2	X0	X1	D2	D2	D2	D2	D2	X0	X0	D2	I	I	I	X0	D2	X0	X1
umuDC	2	1	D2	D2	D2	D2	X0	X0	D2	X0	X0	D2	D2	I	D2	X0	D2	D2	D2	D2	D2	D2	X0	X0	X0	D2	X0	X0	X0	D2	X0	D2
uxuAB	2	1	S2	D2	X0	D2	X0	X0	D2	X0	X0	X0	X0	I	X0	X0	D2	D2	D2	D2	S2	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0
operon[4]	5	1	S5	D5	D5	S5	D2	D2	D5	X0	D2	S5	S5	S5	S5	D4	D3	D4	S4	D4	D3	D4	X1	X1	X0	D4	D4	D3	X0	S4	D3	D4
XapAB	2	1	D2	X1	X1	D2	X1	X1	D2	X1	X1	X1	D2	I	X1	D2	D2	D2	D2	D2	X1	D2	X1	X1	X1	X1	D2	D2	X0	X1	D2	D2
XylAB	2	1	S2	S2	X0	S2	X0	X0	S2	X0	I	X1	X1	I	X0	X0	I	I	I	X1	I	X1	X0	X0	X0	X0	X0	X0	X0	X1	X0	X0
ycfC-purB	2	1	X1	X1	X1	X1	X0	X0	X1	I	I	I	S2	I	I	D2	I	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0	X0
yhdG-fis	2	1	X1	X1	X1	X1	X0	X0	X1	X0	I	I	I	I	I	D2	I	X0	X1	X1	X0	X0	X0	X0	X0	X1	X0	X0	X0	X0	X0	X0

operon[1] = napFDAGHBC-ccmABCDEF-dsbE-ccmH.
operon[2] = phnCDE-f73-phnFGHIJKLMNOP.
operon[3] = pstSCAB-phoU.
operon[4] = wza-wzb-b2060-wcaAB.

Table 3. Three structural conservation scores (expressed as percentages) calculated separately for each of the 82 operons in each of the four higher taxa. X – At the most one ortholog was found in each genome of the higher taxa, therefore a score was not calculated

Operon	Unweighted score for identical operons				Weighted score for identical operons				Weighted score for similar operons			
	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria
ahpCF	14.3	37.5	33.3	0	100	100	100	X	100	100	100	X
araBAD	0	12.5	0	0	0	50	0	X	0	100	100	X
argCBH	0	37.5	0	0	0	42.9	0	0	0	85.7	83.3	100
aroF-tyrA	0	25	0	0	0	40	0	0	0	40	0	0
atoDAE	0	12.5	0	0	X	50	0	0	100	100	100	100
betIBA	14.3	25	0	0	20	100	0	0	80	60	12.5	33.3
bioBFCD	0	37.5	0	0	0	42.9	0	0	57.1	87.5	75	66.7
cadBA	0	37.5	0	0	0	50	0	0	100	100	0	0
caiTABCDE	0	12.5	0	0	0	50	X	0	83.3	33.3	66.7	33.3
codBA	0	37.5	0	0	X	100	0	X	X	100	0	X
cydAB	100	75	41.7	66.7	100	100	100	100	100	100	100	100
cynTSX	0	0	0	0	X	0	X	X	0	50	0	0
cyoABCDE	0	50	0	0	0	80	0	X	100	100	100	0
cysDNC	0	50	0	0	0	80	0	X	100	100	100	100
cysJIH	0	50	0	0	0	66.7	0	X	60	83.3	66.7	100
cysPUWAM	0	0	0	0	0	0	X	X	100	71.4	50	50
dadAX	0	25	0	0	0	28.6	0	0	60	42.9	0	0
dmsABC	14.3	50	0	0	100	100	X	X	100	66.7	X	X
dppABCDF	57.1	0	25	0	100	0	100	X	100	100	100	100
ebgAC	0	0	0	0	X	0	0	X	X	0	0	X
edd-eda	0	25	0	0	0	28.6	0	X	40	28.6	0	X
entCEBA	0	12.5	0	0	X	50	0	0	20	50	57.1	0
epd-pgk	28.6	50	58.3	100	40	50	63.6	100	100	50	63.6	100
fadBA	0	25	8.3	0	0	66.7	33.3	0	14.3	100	66.7	100
fdnGHI	14.3	25	0	0	100	33.3	X	X	100	100	X	X
fecABCDE	0	0	0	0	0	0	X	X	100	100	100	100
FecIR	42.9	12.5	0	0	100	100	X	X	100	100	X	X
fepA-entD	0	0	0	0	0	0	X	X	0	0	X	X
fepDGC	0	0	0	0	0	0	0	0	100	100	88.9	100

fhuACDB	0	37.5	0	0	0	50	X	X	100	100	100	100
fixABCX	28.6	12.5	0	0	100	100	X	X	100	100	100	100
frdABCD	0	62.5	0	33.3	X	100	X	100	71.4	100	100	100
fruBKA	0	12.5	0	0	X	50	X	X	X	66.7	0	100
ftsQAZ	71.4	87.5	25	0	71.4	100	37.5	X	100	100	100	100
fucAO	0	25	0	0	0	50	0	X	0	50	0	X
fucPIKUR	0	0	0	0	X	0	X	X	50	33.3	42.9	100
gcvTHP	42.9	12.5	0	0	75	100	0	X	100	100	100	0
glnHPQ	71.4	50	41.7	33.3	71.4	100	62.5	100	71.4	100	100	100
glpACB	0	62.5	0	0	X	100	X	X	0	83.3	0	0
glpFK	0	62.5	33.3	0	0	83.3	36.4	X	0	83.3	81.8	X
glpTQ	0	37.5	8.3	0	X	50	25	X	X	66.7	25	X
guaBA	0	12.5	0	0	X	50	0	0	X	50	0	0
ilvIH	71.4	87.5	50	100	100	100	85.7	100	100	100	85.7	100
kbl-tdh	28.6	37.5	0	0	50	50	0	0	50	50	33.3	0
kdpABC	57.1	37.5	25	33.3	100	100	100	100	100	100	100	100
lacZYA	0	0	0	0	0	X	0	X	0	16.7	12.5	0
lexA-dinF	0	12.5	0	0	0	14.3	0	0	0	28.6	0	0
malEFG	0	0	0	0	X	X	X	X	X	100	100	X
malPQ	0	25	0	0	X	33.3	0	0	X	33.3	66.7	0
malXY	0	0	0	0	X	0	0	0	X	0	0	0
manXYZ	0	25	41.7	0	X	100	100	X	X	100	100	X
melAB	0	12.5	0	0	X	50	0	X	X	50	0	X
mglBAC	57.1	25	8.3	0	80	33.3	33.3	0	100	100	100	100
mtlADR	0	50	0	0	X	100	X	X	X	100	100	X
nagBACD	0	25	0	0	0	100	0	0	20	71.4	44.4	0
operon[1]	0	0	0	0	X	0	X	X	71.4	100	27.3	100
narGHJI	0	0	0	0	0	0	0	0	100	100	100	100
narXL	0	25	0	0	X	33.3	X	X	X	33.3	X	X
nirBDC-cysG	0	12.5	0	0	X	50	0	X	100	100	60	100
nrdAB	42.9	75	66.7	0	60	85.7	72.7	0	60	100	100	66.7
nrfABCDEFG	0	0	0	0	X	X	X	X	71.4	85.7	X	X
oppABCDF	57.1	0	25	0	100	0	100	X	100	100	100	100

Table 3. (Continued)

Operon	Unweighted score for identical operons				Weighted score for identical operons				Weighted score for similar operons			
	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria
operon[2]	0	0	0	0	0	0	X	X	57.1	28.6	58.3	50
phoBR	100	87.5	66.7	100	100	100	100	100	100	100	100	100
operon[3]	42.9	37.5	33.3	0	60	75	57.1	0	100	100	100	100
purEK	71.4	87.5	58.3	0	100	100	100	0	100	100	100	100
purHD	0	75	50	0	0	85.7	85.7	0	0	100	85.7	0
purMN	71.4	75	66.7	0	100	85.7	100	0	100	100	100	0
rbsDACBK	0	50	0	0	X	80	0	0	100	100	66.7	50
rhaSR	0	0	0	0	0	0	0	0	0	0	0	0
speA	0	25	0	0	X	40	X	X	X	40	X	X
tauABCD	0	0	0	0	0	0	X	X	100	80	100	100
tdcABCDEFG	0	0	0	0	X	0	X	X	14.3	75	12.5	0
torCAD	0	25	0	0	X	40	X	X	100	100	X	X
treBC	0	0	25	0	0	0	33.3	0	0	0	33.3	0
umuDC	0	12.5	0	0	0	20	0	0	0	20	0	0
uxuAB	0	12.5	0	0	0	50	0	X	25	50	25	X
operon[4]	0	0	0	0	0	0	X	X	28.6	57.1	12.5	33.3
xapAB	0	12.5	0	0	0	25	0	0	0	25	0	0
xylAB	0	37.5	25	0	0	100	100	X	100	100	100	X
ycfC-purB	0	75	0	0	X	75	X	X	X	87.5	X	X
yhdG-fis	0	75	0	0	X	85.7	X	X	X	85.7	X	X

operon[1] = napFDAGHBC-ccmABCDEF-dsbE-ccmH.
operon[2] = phnCDE-f73-phnFGHIJKLMNOP.
operon[3] = pstSCAB-phoU.
operon[4] = wza-wzb-b2060-wcaAB.

which was found intact in 83% of the genomes in our study.

Effect of transcriptional regulation on structural conservation in identical operons

The distribution of unweighted and weighted conservation scores (Table 3) in the four higher taxa are shown in Figures 2a, b. While only up to 5% of *E. coli* operons had unweighted conservation scores larger than 80% in gamma proteobacteria (Figure 2a), 32% of the operons in *E. coli* had weighted scores larger than 80% (Figure 2b). Unsurprisingly, the highest weighted and unweighted scores are found in gamma proteobacteria.

First, we attempted to rule out the possibility that operon size may influence our results. We found no correlation between the weighted score and operon size. In contrast, a negative Spearman's rank correlation coefficient was found for the relation between the unweighted score and operon size in two out of the four taxonomic groups (Table 4a). This difference seems to result from the definitions of these two variables. In the unweighted score, the absence of genes from the genome scores as 0. Thus, *E. coli* operons with a relatively high number of genes have a high probability of lacking some homologous genes in other bacteria. These operons

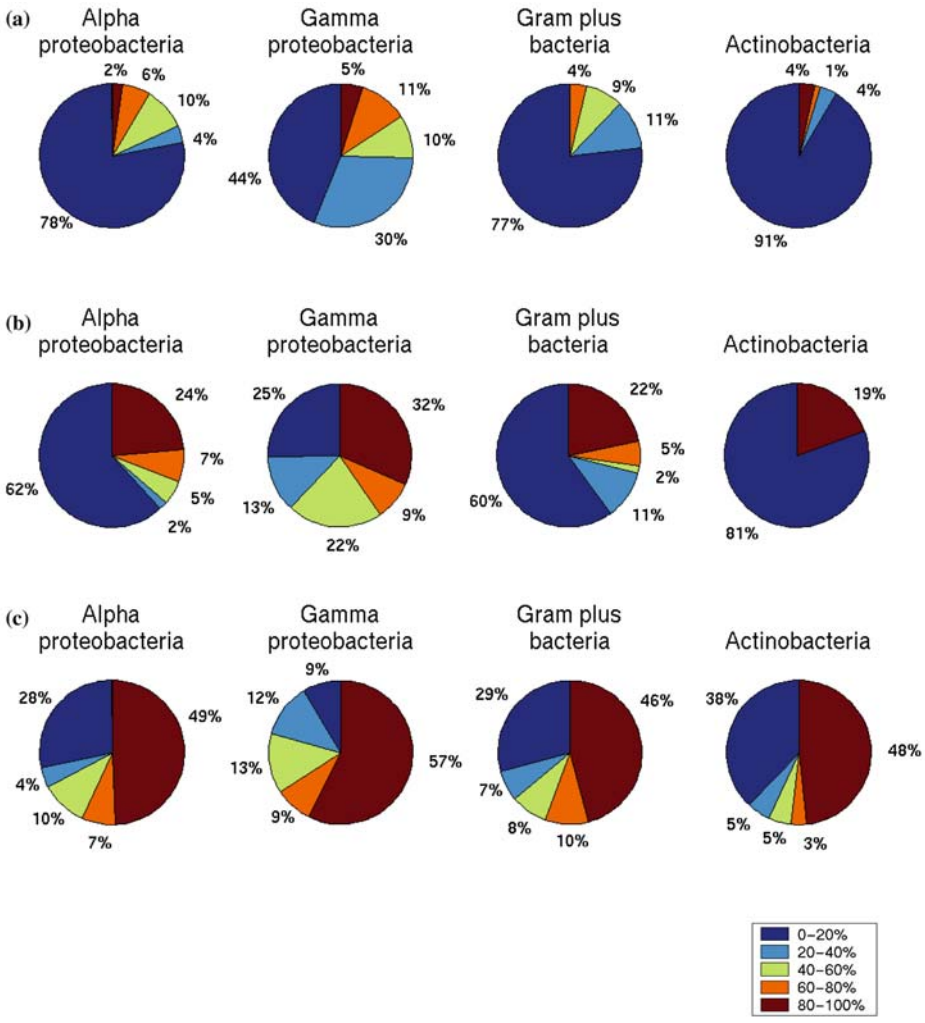

Figure 2. Conservation scores in various taxa. (a) Distribution of unweighted scores for identical operons. (b) Distribution of weighted scores for identical operons. (c) Distribution of weighted scores for similar operons.

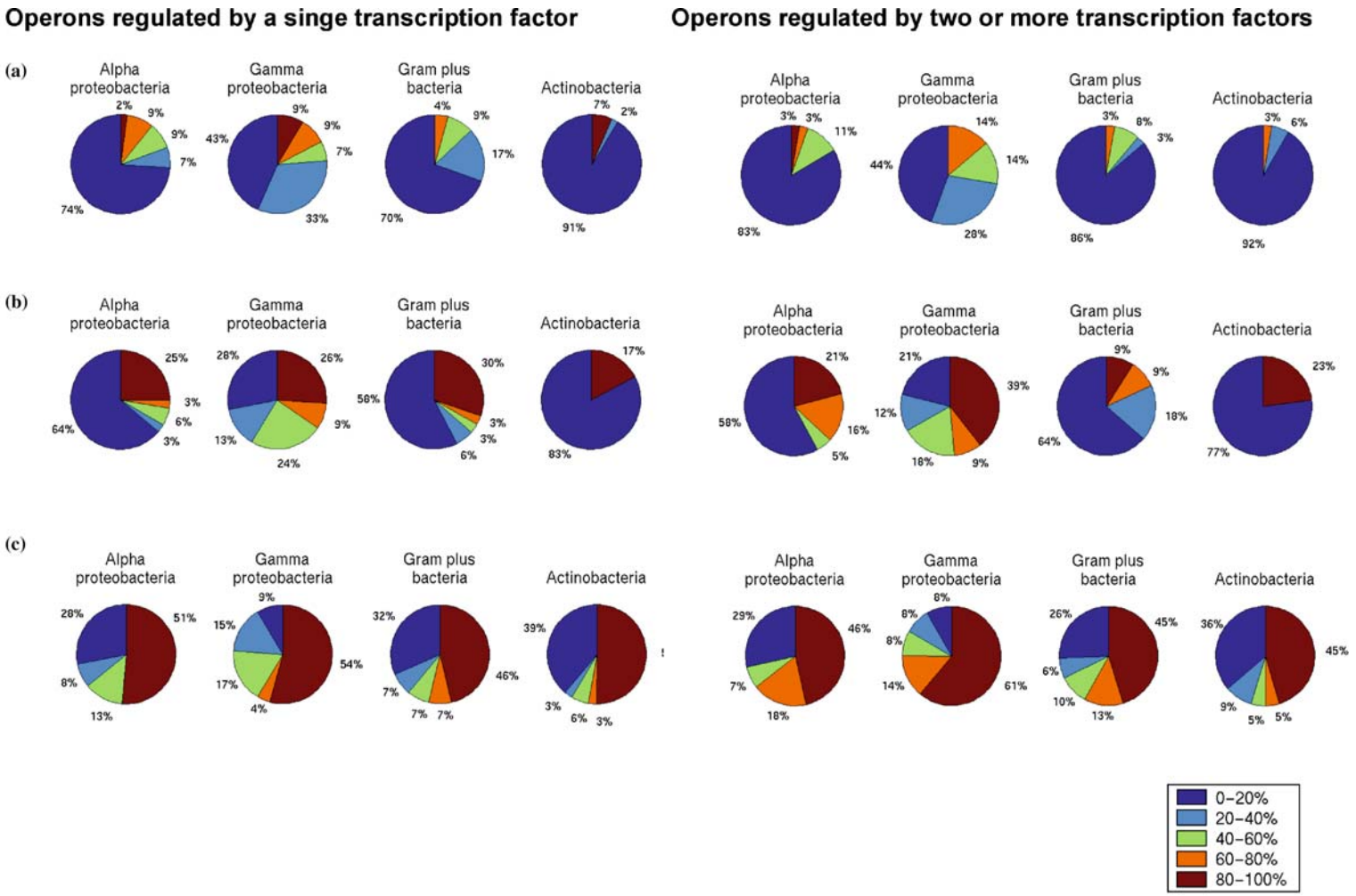

Figure 3. Distribution of conservation scores in operons regulated by either a single transcription factor (left-hand side) or by two or more transcription factors (right-hand side). (a) Distribution of unweighted scores for identical operons. (b) Distribution of weighted scores for identical operons. (c) Distribution of weighted scores for similar operons.

Table 4. Factors affecting the conservation scores for identical operons. In sections a–c we used Spearman correlation. In section d we used Mann–Whitney test

		Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actinobacteria
a. Operon size vs.	Unweighted scores	$n = 82$	$n = 82$	$n = 82$	$n = 82$
		$r = -0.08$	$r = -0.35$	$r = -0.26$	$r = -0.12$
		$p = 0.474$	$p = 0.001$	$p = 0.020$	$p = 0.301$
	Weighted scores	$n = 55$	$n = 79$	$n = 55$	$n = 36$
		$r = -0.05$	$r = -0.19$	$r = -0.09$	$r = -0.04$
		$p = 0.708$	$p = 0.099$	$p = 0.505$	$p = 0.826$
b. Number of binding transcription factors vs.	Unweighted scores	$n = 82$	$n = 82$	$n = 82$	$n = 82$
		$r = -0.07$	$r = -0.08$	$r = -0.09$	$r = -0.03$
		$p = 0.552$	$p = 0.473$	$p = 0.402$	$p = 0.769$
	Weighted scores	$n = 55$	$n = 79$	$n = 55$	$n = 36$
		$r = 0.08$	$r = 0.14$	$r = -0.11$	$r = 0.06$
		$p = 0.585$	$p = 0.222$	$p = 0.418$	$p = 0.744$
c. Number of binding transcription facros vs.	Unweighted scores for Size 2	$n = 37$	$n = 37$	$n = 37$	$n = 37$
		$r = 0.01$	$r = -0.11$	$r = 0.05$	$r = -0.06$
		$p = 0.969$	$p = 0.530$	$p = 0.767$	$p = 0.743$
	Unweighted scores for Size 3	$n = 22$	$n = 22$	$n = 22$	$n = 22$
		$r = 0.11$	$r = 0.03$	$r = 0.07$	$r = -0.10$
		$p = 0.610$	$p = 0.894$	p = 0.763	$p = 0.673$
d. Type of single transcription factor of operon in size 2 vs.	Unweighted scores	$p = 0.257$	$p = 0.801$	$p = 0.244$	$p = 0.186$
	Weighted scores	$p = 0.186$	$p = 0.345$	$p = 0.361$	$p = 0.841$

tend to lower the correlation coefficient. This phenomenon disappears when the score is weighted.

Next, we ask whether the number of transcription factors affects the probability that an operon will remain identical in gene makeup and gene order in the four higher taxa. No correlation was found between either of the two conservation scores and the number of transcription factors (Table 4b). We note that since the absence of genes from a genome disqualifies an operon from being used in the weighted scheme, the numbers of operons within a taxon are different from each other depending on weighting. The results are presented as pie distributions of scores for one and two or more transcription factors (Fig 3a, b). There seems to be no relationship between structural conservation and number of transcription factors.

The relationship between the unweighted conservation scores and the number of transcription factors was tested separately for operons of different sizes (Table 4c). Out of the 82 tested operons, 37 had two genes and 22 had three genes. The number of operons with four or more genes was too small to be included in a meaningful statistical analysis. Again, no correlation was found between the number of transcription factors and the unweighed scores for both size 2 and size 3 operons. Thus, the conservation of operons seems to be independent of the number of transcription factors. In a Mann-Whitney test, no dependence was found between conservation and type of transcription regulation in a group of operons of size 2 that are controlled by a single transcription factor ($n = 23$) (Table 4d).

Effect of transcriptional regulation on structural conservation of similar operons

Since the number of fully conserved operons is quite small in bacteria that are only distantly related to *E. coli*, we attempted to test the influence of the number of transcription factor with a less strict conservation score. Thus, we combined the weighted scores for identical and similar operons. The weighted conservation scores are listed in Table 3

Table 5. Factors affecting the conservation scores for similar operons. Rows A–C were detected by Spearman correlation. Row D was detected using Mann–Whitney test

		Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actinobacteria
a. Operon size vs. scores		$n = 67$ $r = 0.22$ $p = 0.067$	$n = 82$ $r = 0.25$ $p = 0.022$	$n = 72$ $r = 0.30$ $p = 0.010$	$n = 58$ $r = 0.28$ $p = 0.031$
b. Number of binding transcription factors vs. scores		$n = 67$ $r = -0.003$ $p = 0.979$	$n = 82$ $r = 0.12$ $p = 0.280$	$n = 72$ $r = 0.06$ $p = 0.605$	$n = 58$ $r = -0.05$ $p = 0.696$
c. Number of transcription factors after cleaning size effect for operon scores	Size 2	$n = 26$ $r = -0.01$ $p = 0.975$	$n = 37$ $r = 0.06$ $p = 0.744$	$n = 31$ $r = 0.03$ $p = 0.856$	$n = 21$ $r = 0.27$ $p = 0.240$
	Size 3	$n = 18$ $r = -0.12$ $p = 0.648$	$n = 22$ $r = 0.08$ $p = 0.713$	$n = 19$ $r = 0.13$ $p = 0.602$	$n = 15$ $r = -0.603$ $p = 0.013$
d. Type of single transcription factor of operon in size 2 vs. scores		$p = 0.488$	$p = 0.361$	$p = 0.879$	$p = 0.850$

and their distribution in the four taxa are shown in Figure 2c. About 50% of the operons in the different taxonomic groups had weighted conservation scores larger than 80%. The results obtained for the combined group of identical and similar operons are comparable to those obtained for the group of identical operons only (Table 5). Operon size was found to have an effect on the degree of structural conservation, however, this effect was in the opposite direction to that observed for the group of identical operons. Hence, longer operons seem to be more conserved. Notwithstanding this difference, the correlation may be a combinatorial artifact resulting from the over-sampling of similar fragments from long operons. When dividing the operons into different operon sizes, a dependence between conservation and number of transcription factors was only found in one data set (operons of length 3 in actinobacteria; $n = 15$. We can, therefore, conclude that the regulation of an operon is unrelated to its structural conservation.

Effect of transcriptional regulation on sequence conservation of operons

Possible effects of the number of transcription factors on sequence conservation were tested for 64 *E. coli* operons, for which structurally identical counterparts were found in at least one other bacterium. The number of amino acid replacements was calculated in a pairwise manner between each *E. coli* gene and its homolog. Two measures of aminoacid sequence distance were used, mean distance and least distance (Table 6). As was the case with the structural conservation scores, the highest sequence conservation was found in gamma proteobacteria and the lowest in actinobacteria. With Spearman nonparametric test, we found no correlation between the number of transcription factors and either distances (Table 7).

Lack of relationship between structural conservation and sequence conservation

The relationship between structural and sequence conservation scores was tested by using the weighted scores for identical and similar operons and the two sequence scores. With Spearman nonparametric test, we found no correlation between structural conservation and sequence conservation in any of the tests (Table 8).

Discussion

Transcription regulation has been extensively studied in *E. coli* (e.g. Thieffry et al., 1998; Babu &

Table 6. Two sequence conservation measures calculated separately for each of the 64 operons for which structurally identical counterparts were found. Scores are based on protein distances and are calculated for each of the four higher taxa

Operon	Operon mean distance				Operon least distance			
	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria
ahpCF	0.43	0.35	0.60	nd	0.33	0.28	0.51	nd
araBAD	nd	0.06	nd	nd	nd	0.03	nd	nd
argCBH	nd	0.23	nd	nd	nd	0.15	nd	nd
aroF-tyrA	nd	0.24	nd	nd	nd	0.22	nd	nd
atoDAE	nd	0.41	nd	nd	nd	0.33	nd	nd
betIBA	0.93	0.35	nd	nd	0.68	0.24	nd	nd
bioBFCD	nd	0.50	nd	nd	nd	0.20	nd	nd
cadBA	nd	0.56	nd	nd	nd	0.51	nd	nd
caiTABCDE	nd	0.09	nd	nd	nd	0.02	nd	nd
codBA	nd	0.26	nd	nd	nd	0.25	nd	nd
cydAB	1.25	0.54	1.77	1.43	1.13	0.47	1.43	1.38
cyoABCDE	nd	0.45	nd	nd	nd	0.25	nd	nd
cysDNC	nd	0.32	nd	nd	nd	0.18	nd	nd
cysJIH	nd	0.52	nd	nd	nd	0.38	nd	nd
dadAX	nd	0.55	nd	nd	nd	0.44	nd	nd
dmsABC	2.20	0.37	nd	nd	1.53	0.21	nd	nd
dppABCDF	1.30	Nd	1.56	nd	1.02	nd	1.31	nd
edd-eda	nd	0.47	nd	nd	nd	0.35	nd	nd
entCEBA	nd	0.13	nd	nd	nd	0.09	nd	nd
epd-pgk	0.92	0.27	0.99	1.01	0.88	0.19	0.95	0.97
fadBA	nd	0.42	1.19	nd	nd	0.35	0.93	nd
fdnGHI	0.91	0.29	nd	nd	0.60	0.19	nd	nd
fecIR	1.88	1.14	nd	nd	1.75	0.93	nd	nd
fhuACDB	nd	1.26	nd	nd	nd	0.84	nd	nd
fixABCX	1.50	0.13	nd	nd	1.29	0.12	nd	nd
frdABCD	nd	0.42	nd	1.18	nd	0.22	nd	0.73
fruBKA	nd	0.71	nd	nd	nd	0.65	nd	nd
ftsQAZ	1.60	0.67	1.69	nd	0.87	0.36	0.82	nd
fucAO	nd	0.82	nd	nd	nd	0.59	nd	nd
gcvTHP	0.94	0.03	nd	nd	0.62	0.01	nd	nd
glnHPQ	1.29	0.66	1.20	1.47	0.82	0.42	0.71	0.79
glpACB	nd	0.56	nd	nd	nd	0.33	nd	nd

Table 6. (Continued)

Operon	Operon mean distance				Operon least distance			
	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actino-bacteria
glpFK	nd	0.40	0.98	nd	nd	0.30	0.53	nd
glpTQ	nd	0.36	1.02	nd	nd	0.31	0.58	nd
guaBA	nd	0.48	nd	nd	nd	0.45	nd	nd
ilvIH	0.84	0.31	1.01	0.96	0.78	0.28	0.93	0.92
kbl-tdh	0.37	0.13	nd	nd	0.36	0.08	nd	nd
kdpABC	0.72	0.43	0.94	0.75	0.49	0.28	0.58	0.48
lexA-dinF	nd	0.08	nd	nd	nd	0.03	nd	nd
malPQ	nd	0.45	nd	nd	nd	0.40	nd	nd
manXYZ	nd	0.75	0.70	nd	nd	0.67	0.60	nd
melAB	nd	0.11	nd	nd	nd	0.06	nd	nd
mglBAC	1.24	1.51	1.00	nd	0.97	1.17	0.78	nd
mtlADR	nd	0.43	nd	nd	nd	0.27	nd	nd
nagBACD	nd	0.13	nd	nd	nd	0.09	nd	nd
narXL	nd	0.60	nd	nd	nd	0.32	nd	nd
nirBDC-cysG	nd	0.07	nd	nd	nd	0.05	nd	nd
nrdAB	2.21	0.75	2.07	nd	2.11	0.69	2.00	nd
oppABCDF	1.51	nd	1.68	nd	1.09	nd	1.30	nd
phoBR	1.75	0.65	1.61	1.59	1.31	0.41	1.23	1.30
pstSCAB-phoU	1.35	0.34	1.67	nd	0.58	0.19	0.83	nd
purEK	0.92	0.58	1.15	nd	0.52	0.34	0.66	nd
purHD	nd	0.28	0.81	nd	nd	0.25	0.74	nd
purMN	0.86	0.38	1.05	nd	0.71	0.26	0.79	nd
rbsDACBK	nd	0.35	nd	nd	nd	0.23	nd	nd
speA	nd	0.52	nd	nd	nd	0.45	nd	nd
torCAD	nd	0.71	nd	nd	nd	0.40	nd	nd
treBC	nd	nd	0.92	nd	nd	Nd	0.80	nd
umuDC	nd	0.24	nd	nd	nd	0.18	nd	nd
uxuAB	nd	0.16	nd	nd	nd	0.09	nd	nd
xapAB	nd	0.13	nd	nd	nd	0.12	nd	nd
xylAB	nd	0.44	1.09	nd	nd	0.20	0.82	nd
ycfC-purB	nd	0.41	nd	nd	nd	0.24	nd	nd
yhdG-fis	nd	0.30	nd	nd	nd	0.20	nd	nd

nd = not detected.

Table 7. Spearman correlation tests for relationship between sequence conservation scores and the number of transcription factors

	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actinobacteria
Operon mean distance	$n = 22$	$n = 61$	$n = 22$	$n = 7$
	$r = 0.27$	$r = 0.09$	$r = 0.02$	nd
	$p = 0.22$	$p = 0.51$	$p = 0.93$	
Operon least distance	$n = 22$	$n = 61$	$n = 22$	$n = 7$
	$r = 0.28$	$r = 0.07$	$r = -0.19$	nd
	$p = 0.20$	$p = 0.59$	$p = 0.41$	

nd – not detected.

Teichmann, 2003; Martinez-Antonio & Collado-Vides, 2003). In our analysis, large transcription units from *E. coli* were found to be regulated by more transcription factors than smaller ones. Large transcription units, thus, appear to require tighter regulation. We note, however, that the correlation coefficient was small ($r = 0.15$, $P = 0.015$). In addition, we found that when regulated by a single transcription factor, large *E. coli* operons have a higher chance of being regulated by activators than either small operons or single genes (Figure 1). Since promoters regulated by repressors are known to be stronger, i.e., to produce higher quantities of transcripts than those regulated by activators (Choy & Adhya, 1996), it is possible that the preponderance of activators in large operons, as well as the large number of transcription factors associated with large operons, may have evolved to prevent unnecessary and energetically costly transcription and translation of these operons.

In agreement with previous reports (Mushegian and Koonin, 1996; Siefert et al., 1997; Watanabe et al., 1997; Itoh et al., 1999), we find that operons tend to evolve rapidly. A minor exception to this rule seems to be the two-gene *pho*BR operon

Table 8. Spearman correlation tests for relationship between structural conservation score and sequence conservation scores

	Alpha proteobacteria	Gamma proteobacteria	Gram plus bacteria	Actinobacteria
Weighted score for identical operons vs. mean distance	$n = 22$	$n = 61$	$n = 22$	$n = 7$
	$r = 0.02$	$r = 0.11$	$r = 0.05$	nd
	$p = 0.93$	$p = 0.39$	$p = 0.82$	
Weighted score for identical operons vs. least distance	$n = 22$	$n = 61$	$n = 22$	$n = 7$
	$r = 0.16$	$r = 0.09$	$r = 0.14$	nd
	$p = 0.47$	$p = 0.50$	$p = 0.53$	
Weighted score for similar operons vs. mean distance	$n = 22$	$n = 61$	$n = 22$	$n = 7$
	$r = 0.03$	$r = 0.24$	$r = 0.40$	nd
	$p = 0.89$	$p = 0.07$	$p = 0.06$	
Weighted score for similar operons vs. least distance	$n = 22$	$n = 61$	$n = 22$	$n = 7$
	$r = 0.06$	$r = 0.21$	$r = 0.15$	nd
	$p = 0.79$	$p = 0.11$	$p = 0.51$	

nd – not detected.

(Makino et al., 1986; Wanner & Chang, 1987; Anba et al., 1990; von Kruger, Humphreys & Ketley, 1999; Pragai et al., 2004). One of the genes in this operon is *pho*B, which belongs to a very small group of positively autoregulated genes (Thieffry et al., 1998). This transcription factor controls the phosphate regulon, which is composed of at least 31 genes located in eight operons (Wanner, 1993). The second gene in this operon is *pho*R, which is a sensory protein for the same *pho*BR-controlled regulon (Wanner & Chang, 1987).

How can we explain this lack of conservation? Itoh et al. (1999) suggested that selection against the destruction of genes may be weak, such that changes in operon structure and composition may be selectively neutral during long-term evolution. This hypothesis implies that the main attributes of operons, i.e., co-regulation of transcription and cotranslation, are unimportant. If, however, transcription co-regulation and co-translation *are* important, then the question arises as to how are these two processes maintained during evolution in the absence of operon-structure conservation. As far as co-regulation of transcription is concerned, one promising answer involves the concept of *regulon* (Maas & Clark, 1964) i.e., the possibility that co-regulation of two or more genes may be maintained even in the absence of cohabitation within the same operon. A second possible answer involves the concept of *uber-operon* (Lathe, Snel & Bork, 2000). According to this hypothesis, what is conserved are sets of operons, whereas the composition of each individual operon within the set is not. As far as co-translation is concerned, the current consensus is that co-translation cannot be maintained in the absence of physical proximity. Indeed, gene pairs whose physical proximity is maintained during evolution appear to produce proteins that co-interact physically (Dandekar et al., 1998) or belong to the same biochemical pathway (Yanai, Mellor & DeLisi, 2002).

Our findings suggest that the number of transcription factors has no influence on either the structural or sequence conservation of operons. These conclusions were independent of the measures of conservation used in our analyses. These findings stand in contrast to results pertaining to single genes in which it had been shown that gene loss occurs less frequently in highly connected genes (Krylov et al., 2003). The question now is: Why is the loss of a single gene dependent on its connectivity, whereas the loss of an operon is not?

We propose several explanations for this disparity. First, the difference may be due to the use of different network types. Studies dealing with the influence of connectivity on single genes used networks of protein-protein interactions, whereas our analysis used a transcriptional regulatory network.

Second, it is possible that our results are affected by an assumption concerning the conservation of connectivity during evolution. In a manner similar to the usual practice in the literature (Fraser et al., 2002; Fraser, Wall & Hirsh, 2003; Jordan, Wolf & Koonin, 2003a, b; Krylov et al., 2003; Hahn, Conant & Wagner, 2004), we too assumed identical connectivities among all the genomes under comparison. Such an assumption is necessary due to data insufficiency. However, this assumption may not apply or may only partially apply in nature, and one may obtain different outcomes when data on protein and regulatory networks become available for bacteria other than *E. coli*.

Third, the difference may be due to the fact that the correlation between protein connectivity and evolutionary conservation is mostly due to the existence of highly connected protein 'hubs' (e.g., Krolov et al., 2003). Note, that operons regulated by more than two transcription factors constituted only a minute fraction ($\sim$7%) in our dataset. Although very complex network motifs were reported for the *E. coli* regulation network (Shen-Orr et al., 2002), it appears that compensating for protein hubs in protein networks is more difficult than in transcrition-regulation networks. Such compensatory effects could be achieved by regulons or *uber operons* as was previously suggested (Lathe, Snel & Bork, 2000).

Finally, we might wish to consider the possibility that the difference between gene connectivity and operon connectivity is only apparent. A critical review of the literature indicates that although gene connectivity is frequently touted as an important factor in determining rates of evolution, only a very small fraction of the variability in degrees of evolutionary conservation is explainable by this factor (Fraser et al., 2002; Fraser, Wall & Hirsh, 2003; Jordan, Wolf & Koonin, 2003a, b; Krylov et al., 2003; Hahn, Conant & Wagner, 2004). The explainable fraction ranges from 0%

(Hahn, Conant & Wagner, 2004) to 12% (Krylov et al., 2003). It is, therefore, possible that detecting such a weak effect in a complex genetic entity, such as the operon, is extremely difficult, especially with small sample sizes.

Acknowledgements

We thank David Wool and Tal Dagan for their help. E. H.-C. thanks Shay Covo for critical review of the manuscript.

References

Anba, J., M. Bidaud, M.L. Vasil & A. Lazdunski, 1990. Nucleotide sequence of the *Pseudomonas aeruginosa pho*B gene, the regulatory gene for the phosphate regulon. J. Bacteriol. 172: 4685–4689.

Babu, M.M & S.A Teichmann, 2003. Functional determinants of transcription factors in *Escherichia coli:* protein families and binding sites. Trends Genet. 19: 75–79.

Bastolla, U., M. Porto, M.H. Eduardo Roman & M.H. Vendruscolo, 2003. Connectivity of neutral networks, overdispersion & structural conservation in protein evolution. J. Mol. Evol. 56: 243–254.

Choy, H. & S. Adhya, 1996. Negative Control in *Escherichia coli* and *Salmonella*, Cellular and Molecular Biology, pp. 1287–1299 in American Society for Microbiology, edited by F.C. Neidhardt, R. Curtiss, J.L. Ingraham, E.C.C. Lin & K.B. Low. Washington, DC.

Clarke, B., J.E. Mittenthal & M. Senn, 1993. A model for the evolution of networks of genes. J Theor. Biol. 165: 269–289.

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

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics. University of Washington, Seattle.

Fraser, H.B., A.E. Hirsh, L.M. Steinmetz, C. Scharfe & M.W. Feldman, 2002. Evolutionary rate in the protein interaction network. Science 296: 750–752.

Fraser, H.B., D.P. Wall & A.E. Hirsh, 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol. Biol. 3: 11.

Graur D, Li WH, (2000). Fundamentals of Molecular Evolution. Second Edition, Sinauer Associates, Sunderland, MA, 481 pp.

Hahn, M.W., G.C. Conant & A. Wagner, 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J. Mol. Evol. 58: 203–211.

Higgins, D.G., J.D. Thompson & T.J. Gibson, 1996. Using CLUSTAL for multiple sequence alignments. Meth. Enzymol. 266: 383–402.

Hirsh, A.E. & H.B. Fraser, 2001. Protein dispensability and rate of evolution. Nature 411: 1046–1049.

Huynen, M.A. & P. Bork, 1998. Measuring genome evolution. Proc. Natl. Acad. Sci. USA 95: 5849–5856.

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

Jordan, I.K., Y.I. Wolf & E.V. Koonin, 2003a. Correction. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. 3: 5.

Jordan, I.K., Y.I. Wolf & E.V. Koonin, 2003b. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. 3: 1.

Karp, P.D., M. Riley, M. Saier, I.T. Paulsen & J. Collado-Vides, 2002. The EcoCyc Database. Nucleic Acids Res. 30: 56–58.

Krylov, D.M., Y.I. Wolf, I.B. Rogozin & E.V. Koonin, 2003. Gene loss, protein sequence divergence, gene dispensability, expression level & interactivity are correlated in eukaryotic evolution. Genome Res. 13: 2229–2235.

LatheIII, W.C., B. Snel & P. Bork, 2000. Gene context conservation of a higher order than operons. Trends Biochem. Sci. 25: 474–479.

Maas, W.K. & A.J. Clark, 1964. Studies on the mechanism of repression of arginine biosynthesis in *E. coli*. II. Dominance of repressibility in diploids. J. Mol. Biol. 8: 365–370.

Makino, K., H. Shinagawa, M. Amemura & A. Nakata, 1986. Nucleotide sequence of the pho B gene, the positive regulatory gene for the phosphate regulon of *Escherichia coli* K-12. J. Mol. Biol. 190: 37–44.

Martinez-Antonio, A. & J. Collado-Vides, 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol. 6: 482–489.

Mushegian, A.R. & E.V. Koonin, 1996. Gene order is not conserved in bacterial evolution. Trends Genet. 12: 289–290.

Pal, C., B. Papp & L.D. Hurst, 2001. Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.

Pragai, Z., N.E. Allenby, N. O'Connor, S. Dubrac & G. Rapoport, 2004. Transcriptional regulation of the *pho*PR operon in *Bacillus subtilis*. J. Bacteriol. 186: 1182–1190.

Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate & F.R. Blattner, 2000. RegulonDB (version 3.0): Transcriptional regulation and operon organization in *Escherichia coli* K-12. Nucleic Acids Res. 28: 65–67.

Salgado, H., S. Gama-Castro, A. Martinez-Antonio, E. Diaz-Peredo & F. Sanchez-Solano, 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. Nucleic Acids Res. 32: D303–D306.

Shen-Orr, S.S., R. Milo, S. Mangan & U. Alon, 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31: 64–68.

Siefert, J.L., K.A. Martin, F. Abdi, W.R. Widger & G.E. Fox, 1997. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. J Mol. Evol. 45: 467–472.

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

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova & U.T. Shankavaram, 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29: 22–28.

Thattai, M. & A. van Oudenaarden, 2001. Intrinsic noise in gene regulatory networks. Proc. Natl. Acad. Sci. USA 98: 8614–8619.

Thieffry, D., A. M. Huerta, E. Perez-Rueda & J. Collado-Vides, 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays 20: 433–440.

von Kruger, W.M.A., S. Humphreys & J.M. Ketley, 1999. A role for the *Pho*BR regulatory system homologue in the *Vibrio cholerae* phosphate-limitation response and intestinal colonization. Microbiology 145: 2463–2475.

Wanner, B.L., 1993. Gene regulation by phosphate in enteric bacteria. J. Cell Biochem. 51: 47–54.

Wanner, B.L. & B.D. Chang, 1987. *The pho*BR operon in *Escherichia coli* K-12. J. Bacteriol. 169: 5569–5574.

Watanabe, H., H. Mori, T. Itoh & T. Gojobori, 1997. Genome plasticity as a paradigm of eubacteria evolution. J. Mol. Evol. 44: S57–64.

Williams, E.J. & L.D. Hurst, 2000. The proteins of linked genes evolve at similar rates. Nature 407: 900–903.

Wilson, A.C., S.S. Carlson & T.J. White, 1977. Biochemical evolution. Annu. Rev. Biochem. 46: 573–639.

Wolf, Y.I., I.B. Rogozin, A.S. Kondrashov & E.V. Koonin, 2001. Genome alignment, evolution of prokaryotic genome organization & prediction of gene function using genomic context. Genome Res. 11: 356–372.

Wolf, Y.I., G. Karev & E.V. Koonin, 2002. Scale-free networks in biology: new insights into the fundamentals of evolution? Bioessays 24: 105–109.

Yanai, I., J.C. Mellor & C. DeLisi, 2002. Identifying functional links between genes using conserved chromosomal proximity. Trends Genet. 18: 176–179.

Yang, J., Z. Gu & W.H. Li, 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol. Biol. Evol. 20: 772–774.