Evolutionary Instability of Operon Structures Disclosed by Sequence Comparisons of Complete Microbial Genomes

Takeshi Itoh,†* *Keiko Takemoto,*‡* *Hirotada Mori,** *and Takashi Gojobori*†

*Research and Education Center for Genetic Information, Nara Institute of Science and Technology, Nara, Japan; †Center for Information Biology, National Institute of Genetics, Mishima, Japan; and ‡Institute for Virus Research, Kyoto University, Kyoto, Japan

Gene orders have been shown to be generally unstable by comprehensive analyses in several complete genomes. In this study, we examined instability of genome structures within operons, where functionally related genes are clustered. We compared gene orders of known operons obtained from *Escherichia coli* and *Bacillus subtilis* with corresponding those of operons in 11 complete genome sequences. We found that in many cases, gene orders within operons could be shuffled frequently during evolution, although several operon structures, such as ribosomal protein operons, were well conserved. This suggests that shuffling of a genome structure is virtually neutral in long-term evolution. Moreover, degrees of instability of the operon structures depended on the genomes examined. Variation in degrees of instability of the genome structures was likely to be related to differences in amounts of insertion sequences. Effects on transcription regulation are also discussed in association with operon destruction.

Introduction

Since the completion of the genome sequence of *Haemophilus influenzae* in 1995, more than a dozen genomes have been completed, and numerous genome projects are currently in progress. The advancement of genome research on various organisms gives us a unique opportunity for direct comparison of complete genome sequences to investigate the evolution of the genomes, particularly focusing on evolution of genome structures.

Structural changes in complete genome sequences have been examined among several eubacteria, and gene orders in bacterial genomes have been shown to be generally unstable (Mushegian and Koonin 1996; Tatusov et al. 1996; Himmelreich et al. 1997; Watanabe et al. 1997). In yeast, one whole-genome duplication and successive translocations were observed by the comparison of paralogous gene orders within the genome (Wolfe and Shields 1997), indicating instability of the genome structure in yeast. However, the causes of the genetic instability and its functional significance are still unknown. In order to elucidate them, we turned our attention to the structures of operons that are transcribed into polycistronic mRNAs.

Functionally related genes of eubacteria are often clustered on the genome and are organized into a transcriptional unit, termed an operon (for review, see Lewin 1997, p. 338). Similar gene organizations have been found in archaebacteria (Langer et al. 1995). Interestingly, several *Caenorhabditis elegans* genes appear to be cotranscribed polycistronically in clusters similar to bacterial operons (Spieth et al. 1993; Zorio et al. 1994). It is likely that they arose within the *Caenorhabditis* genus independently of bacteria (Spieth et al. 1993), although it was recently reported that the arrangement of eukaryotic operons seems to have predated the divergence of this genus (Evans et al. 1997).

Since the proximity of functionally related genes was proposed to possibly result in more efficient functioning (Demerec and Demerec 1956), coordinative expression of such genes is generally thought to be logical and economical. Therefore, the structures of operons should be important for efficient regulation by cotranscription, and they should be conserved in the course of evolution. Approximately 100 operons between *Bacillus subtilis* and *Escherichia coli* and only 14 operons between *B. subtilis* and *Synechocystis* sp. were reported to be conserved (Kunst et al. 1997), even though most operons used in the study were hypothetical transcription units. In this paper, we conducted the extensive and comparative analyses of the operon structures in 11 complete genomes that are currently available from public data banks: *H. influenzae* (*Hin*) (Fleischmann et al. 1995), *Mycoplasma genitalium* (*Mge*) (Fraser et al. 1995), *Synechocystis* sp. (*Syn*) (Kaneko et al. 1996), *Methanococcus jannaschii* (*Mja*) (Bult et al. 1996), *Mycoplasma pneumoniae* (*Mpn*) (Himmelreich et al. 1996), *Saccharomyces cerevisiae* (*Sce*) (Goffeau et al. 1997), *Helicobacter pylori* (*Hpy*) (Tomb et al. 1997), *E. coli* (*Eco*) (Blattner et al. 1997), *B. subtilis* (*Bsu*) (Kunst et al. 1997), *Methanobacterium thermoautotrophicum* (*Mth*) (Smith et al. 1997), and *Archaeoglobus fulgidus* (*Afu*) (Klenk et al. 1997). The genome sequences used were those of three proteobacteria, one cyanobacterium, three low-G+C gram-positive bacteria, three archaebacteria, and one eukaryote (fig. 1). For most organisms, few operon structures have been experimentally confirmed. Hence, we decided to compare the known operon structures of *E. coli* or *B. subtilis* with orthologous operons of other genomes, because a number of operons in *E. coli* and *B. subtilis* have been confirmed by experiments. Although cotranscription in protein-coding regions has not been found in *S. cerevisiae*, this genome

Abbreviations: *Afu, Archaeoglobus fulgidus*; *Bsu, Bacillus subtilis*; *Eco, Escherichia coli*; *Hin, Haemophilus influenzae*; *Hpy, Helicobacter pylori*; IS, insertion sequence; *Mge, Mycoplasma genitalium*; *Mja, Methanococcus jannaschii*; *Mpn, Mycoplasma pneumoniae*; *Mth, Methanobacterium thermoautotrophicum*; ORF, open reading frame; *Sce, Saccharomyces cerevisiae*; *Syn, Synechocystis* sp.

Key words: operon, complete genome, genome structure, insertion sequence.

Address for correspondence and reprints: Takashi Gojobori, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411–8540, Japan. E-mail: tgojobor@genes.nig.ac.jp.

Mol. Biol. Evol. 16(3):332–346. 1999
© 1999 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

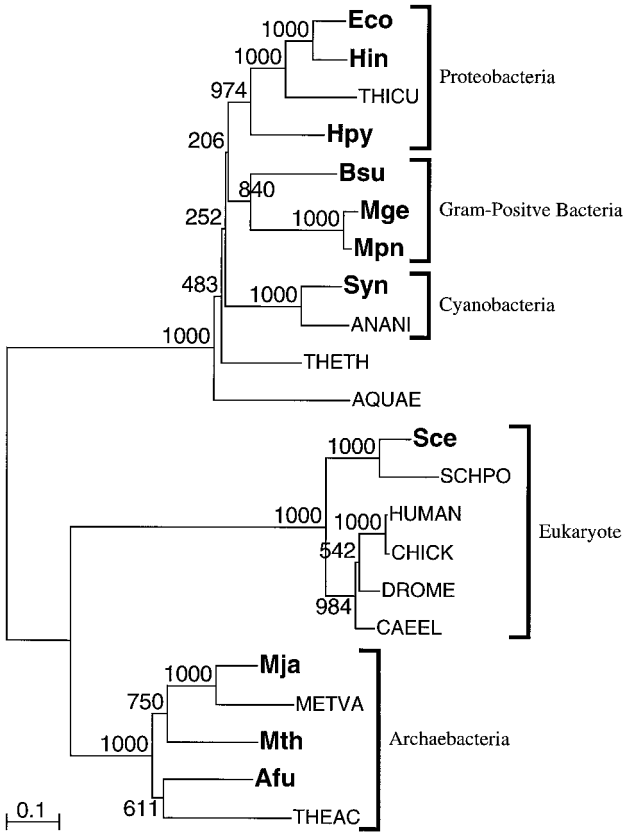

FIG. 1.—Unrooted phylogenetic tree of EF-2/G. Abbreviations: ANANI, *Anacystis nidulans*; AQUAE, *Aquifex aeolicus*; CAEEL, *Caenorhabditis elegans*; CHICK, *Gallus gallus*; DROME, *Drosophila melanogaster*; HUMAN, *Homo sapiens*; METVA, *Methanococcus vannielii*; SCHPO, *Schizosaccharomyces pombe*; THEAC, *Thermoplasma acidophilum*; THETH, *Thermus aquaticus* (subsp. *thermophilus*); THICU, *Thiobacillus cuprinus.* Numbers indicate bootstrap values for 1,000 replicates. The scale for branch lengths is shown below the figure.

was included in our analyses because some genes are known to be organized similarly to those in bacteria (St. John and Davis 1981). It is of particular interest to examine whether some genes have bacterial operon-like structures in *S. cerevisiae,* because their structures may be quite important in the course of evolution. We also estimated the relative stability of operon structures among the eubacterial genomes.

Materials and Methods

Genome Sequences

All of the complete genome sequences of *Hin, Mge, Syn, Mja, Mpn, Sce, Hpy, Eco, Bsu, Mth,* and *Afu* can be obtained from the ftp site at ftp://ncbi.nlm.nih.gov/genbank/genomes. The sequences are also available at http://mol.genes.nig.ac.jp/gib. ORFs were extracted from the feature tables of the GenBank files. The complete genome sequence of *E. coli* K-12 MG1655 was determined by Blattner et al. (1997). In addition, about 60% of the *E. coli* K-12 W3110 genome has been determined by the Japan *E. coli* genome sequencing team (Yura et al. 1992; Fujita et al. 1994; Aiba et al. 1996; Itoh et al. 1996; Oshima et al. 1996; Yamamoto et al. 1997), but the two strains of *E. coli* have essentially the same genome sequences, and the following study obtains the same results from either genome.

Phylogenetic Analysis

Amino acid sequences of EF-2/G were aligned with the CLUSTAL W program (Thompson, Higgins, and Gibson 1994). We discarded highly diverged regions of EF-2/G for the following phylogenetic analysis. The phylogenetic tree was reconstructed by the neighbor-joining (NJ) method (Saitou and Nei 1987). We used CLUSTAL W for the NJ tree reconstruction with corrections for multiple replacements, excluding sites with gaps.

Operon Data

We searched the databases and literature for any descriptions of the operons of *Eco* and *Bsu.* For *Eco,* we collected 256 operons consisting of two or more ORFs. One operon of *Eco* had, on average, 3.5 ORFs. For *Bsu,* 100 operons were collected, consisting, on average, of 4.1 ORFs per operon. The data sets for the operons are available at our WWW site, http://www.cib.nig.ac.jp/dda/taitoh/operondata.html.

Orthologous Pairs

Orthologous pairs were originally defined by Watanabe et al. (1997). In this study, we improved on this method. In particular, an orthologous pair was defined according to the following criteria: (1) orthologous open reading frames (ORFs) between two genomes compared must be the most similar ORF reciprocally (fig. 2*a*); (2) similarity of the pair has to show statistical significance; (3) if a particular ortholog shows more similarity to certain paralogs within the genome, all of the paralogs are regarded as being orthologous to the counterpart of the other genome (fig. 2*b*). The last criterion means that an orthologous pair should be represented as many-to-many relationships of paralogous groups between two species.

Similarity was calculated by the FASTA program (Pearson and Lipman 1988). We assumed that similarity was of statistical significance when the *z* value was larger than 6 in 300 random shuffles of a query sequence (Lipman and Pearson 1985).

Classification of Orthologous Operon Structures

When a structure similar to a known operon of *Eco* or *Bsu* was found in another genome sequence, such a gene cluster was regarded as a hypothetical operon, even if the operon was not confirmed experimentally in the genome.

Structures of orthologous operons were classified into four groups according to their conservation levels of gene orders (fig. 3): (1) an operon structure was ''identical'' if it was completely identical to that of *Eco* or *Bsu*; (2) we defined an operon structure as ''similar'' if a structure similar to the known operon was conserved in part, allowing translocations, deletions, and two insertions within an operon; (3) an operon structure was defined as ''destructed'' if two or more orthologs of an operon were found and the operon structure was not conserved; (4) and an operon structure was defined as

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

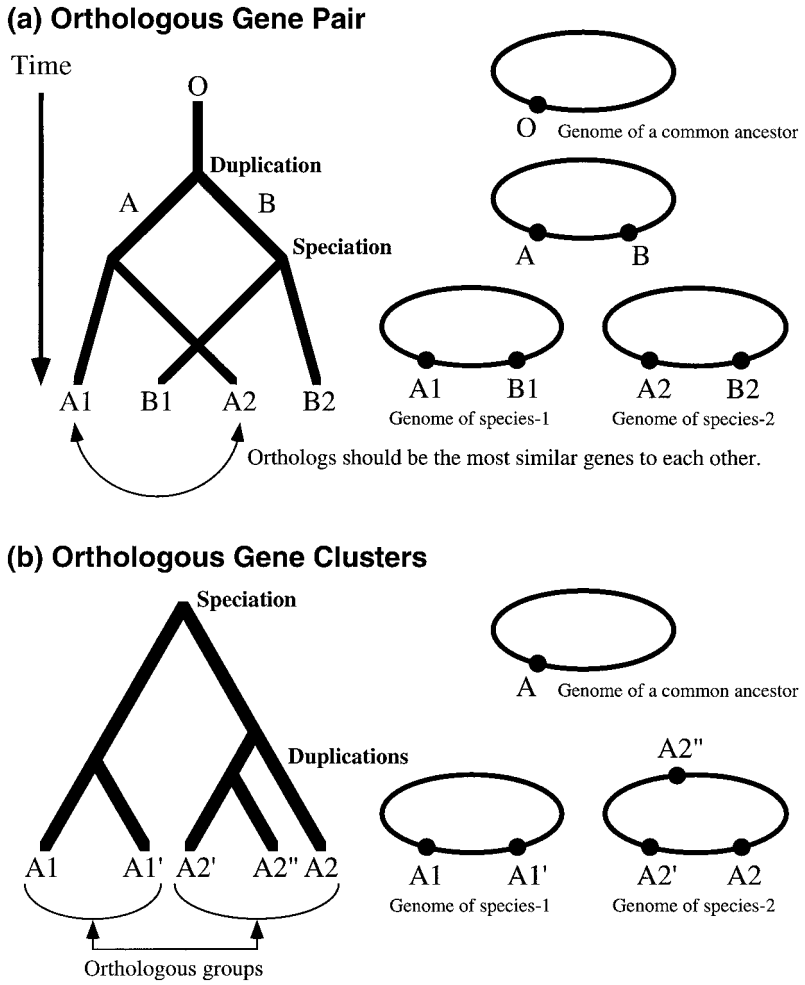

FIG. 2.—Definition of an orthologous gene pair when a duplication event has occurred (*a*) before and (*b*) after speciation. In *a*, A1 is orthologous to A2, and B1 is orthologous to B2 (Watanabe et al. 1997). In *b*, A1 and A1′ are orthologous to A2, A2′, and A2″.

‘‘unknown’’ if no orthologs or at most one ortholog within an operon was found, such that its structural conservation could not be estimated, because two or more orthologs were necessary for comparison of a gene order between genomes. Note that even though an operon appears to have been destructed in some genomes, it is also likely that the operon had been created in *Eco* or *Bsu* after speciation. However, we simply termed such cases ‘‘destructed.’’

Results

Instability of Operon Structures

Instability of operon structures between *Eco* and other genomes is summarized in table 1, and the degree of the instability for each genome is shown in figure 4*a*. Operons of unknown structures are not included in this figure. The operon structures were well-conserved in the most closely related organism, *Hin*, although the orders of other genes appeared to have been shuffled randomly (Watanabe et al. 1997). Nevertheless, most operons in other genomes were subjected to rearrangements in their structures. For example, *Syn* has only a few operon structures that are identical to those of *Eco*, as observed in archaebacteria. Moreover, no identical structures were found in the eukaryotic genome, *Sce*. Thus, the degree of instability of operon structures seems to be correlated with the degree of divergence between the genomes compared. However, this is not always the case, and the degree of instability depends on the evolutionary lineage. In fact, operon structures between *Eco* and each of three gram-positive bacteria (*Bsu*, *Mpn*, and *Mge*) were more conserved than were those between two gram-negative bacteria, *Eco* and *Hpy*. The degree of instability of operon structures between *Bsu* and other genomes is summarized in table 2 (see also fig. 4*b*). A similar tendency was also observed between *Bsu* and other genomes, as found between *Eco* and others.

These results indicate that a genome structure can be readily shuffled within operons in long-term evolution. For instance, although the *dnaK* operon consists of seven genes in *Bsu*, the operon in other genomes was destructed (fig. 5). In comparison with archaeal genomes, it is likely that the last common ancestor at least had the structure spanning *grpE-dnaJ*. Interestingly, the *dnaK* operons were destructed independently within the proteobacterium lineage, between *grpE* and *dnaK* in *Eco* and between *dnaK* and *dnaJ* in *Hpy* (fig. 5).

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

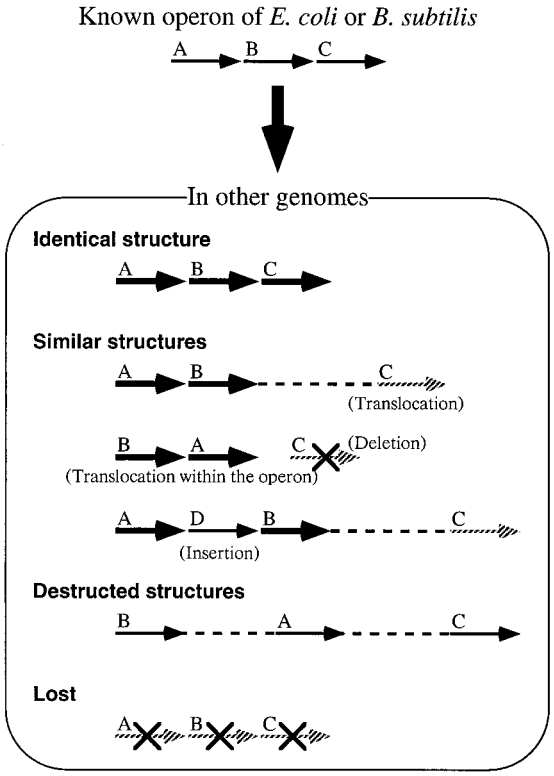

FIG. 3.—Several patterns of alteration of operon structures. Thick arrows are conserved parts of putative operons. Hatched arrows indicate destructed parts of operons.

Even though almost all of the operons were found to be rearranged, destructed, or lost, there were several exceptions. According to the functions of the listed operons, ribosomal protein operon structures such as S10-*spc*-α were well conserved among all genomes except *Sce*, as had been reported (Siefert et al. 1997; Watanabe et al. 1997). Several operons, *atp*, *groE*, *nusA-infB*, and *pheST*, were also well conserved within eubacterial genomes (tables 1 and 2).

Relative Instability of Operon Structure

As shown in figure 6, in order to qualify the degree of instability, let us consider orthologous operons among three species—1, 2, and 3. Suppose that an operon in species 1 is conserved in species 2. There can be two hypotheses for explaining an evolutionary history of alteration in operon structures: first, that their common ancestor had the same structure and that it was destructed only in the lineage of species 3 (fig. 6*a*), and second, that each operon was independently created. The former is more parsimonious than the latter, because at least two distinct translocations are needed for the latter (fig. 6*b*). Let us consider another example. Suppose that an operon is destructed in both species 2 and 3. In this case, it is possible that the operons have been destroyed independently in the lineages of species 2 and 3, although the common ancestor has the identical structure. It is equally possible that an operon was newly created in the lineage of species 1 when the common ancestor did not have any operon structure. Therefore, we estimated the number of destructed operons using the formula described below. Assume that operon structures had been destructed independently in species 2 and 3. In making a comparison of operon structures in species 1 with those in species 2 and 3, let N_2 be the number of conserved operons only in species 2 (i.e., destructed in species 3), let N_3 be the number of conserved operons only in species 3 (i.e., destructed in species 2), and let N_c be the number of commonly conserved operons in both genomes (fig. 6*c*). Under the assumption of independent destruction of operon structures in each species, according to the conditional probability of stochastic independence, we derived the following equation:

$$\frac{N_2 + N_c}{N_2 + N_3 + N_c + N_d} = \frac{N_c}{N_3 + N_c}.$$

Accordingly, the estimated number of destructed operons (N_d) in both genomes is

$$N_d = \frac{N_2 N_3}{N_c}.$$

Thus, the proportion of conserved operons in species 2 (R_2) is derived by the following formula:

$$R_2 = \frac{N_2 + N_c}{N_2 + N_3 + N_c + N_d}.$$

The proportion in species 3 can be calculated in the same way. Therefore, we can compare the relative stabilities of operon structures between species 2 and species 3 by using the estimated number of destructed operons in both species.

Among the proteobacteria, gram-positive bacteria, and cyanobacteria, speciation appears to be almost trifurcate (fig. 1). Thus, the above assumption of independent destruction of operon structures can be reasonably accepted. For the *Bsu* operon orthologs between each proteobacterium and cyanobacterium *Syn*, we counted the number of conserved operons (fig. 7*a–c*) only when operons were conserved or destructed in proteobacterium and *Syn*, and we computed the relative stability of the operon structures. In the same way, for the *Eco* operon orthologs between each gram-positive bacterium and the cyanobacterium, we compared the degree of conservation of the *Eco* operon structures between each gram-positive bacterium and *Syn* (fig. 7*d–f*).

The proportions in *Eco*, *Hin*, and *Bsu* were significantly larger ($P < 0.001$) than in *Syn*. Therefore, the conservation of operon structures are observed in the following order:

$$(Eco, Hin, Bsu) \gg Syn.$$

Similarly, in *Mpn* and *Mge*, the operon structures were much more conserved than in *Syn*, but less conserved than in *Eco*, *Hin*, or *Bsu*. *Hpy* contained only a few conserved operon structures, as observed for *Syn* (fig. 7*c*), although its proportion of conserved structures was larger than that of *Syn* ($P < 0.01$). We conclude that the relative stability of the operon structures among eubacteria is

$$(Eco, Hin, Bsu) > (Mpn \approx Mge) \gg Hpy > Syn.$$

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

Table 1
Conservation of Structures of 256 *Eco* Operons in Other Genomes

	Hin	*Hpy*	*Syn*	*Bsu*	*Mpn*	*Mge*	*Mja*	*Mth*	*Afu*	*Sce*	Function
accBC	I	I	D	I	—	—	x	x	x	x	AcetylCoA carboxylase
ace	—	—	—	—	—	—	—	—	x	D	Malate synthase
aceEF	I	—	D	S	S	S	—	x	x	D	Pyruvate dehydrogenase
ackA-pta	I	—	D	x	D	D	—	—	—	—	Activation of acetate to acetyl CoA
acrAB	x	—	x	x	—	—	x	—	x	x	Acriflavine resistance
acrEF	x	—	x	x	—	—	—	—	D	x	Acriflavine resistance
ada-alkB	—	—	—	x	—	—	x	x	x	x	DNA repair
aga	—	D	D	x	x	x	—	x	—	D	*N*-acetylgalactosamine uptake
α	I	I	S	S	S	S	S	S	S	D	Ribosomal protein
amiAhemF	—	x	D	x	—	—	—	—	—	x	Coproporphyrinogen oxidase
amiB	S	D	D	S	—	—	x	x	x	D	Complex operon
ampDE	x	—	—	x	—	—	—	—	x	x	Beta-lactamase regulation
appCBA	—	—	S	S	—	—	x	x	x	x	Cytochrome bd terminal oxidase
araBAD	x	—	—	S	x	—	—	—	—	D	Arabinose catabolic pathway
araFGH	—	—	D	—	x	x	x	x	x	x	Periplasmic binding protein
argECBH	x	—	D	S	—	—	D	D	D	D	Arginine biosynthesis
arsBC	—	—	D	S	—	—	D	D	x	—	Arsenical resistance
artPIQMJ	S	x	—	—	x	x	D	x	—	D	Arginine transport
ascFB	—	—	—	—	—	—	—	—	—	x	Cryptic sugar transport
atoDAB	I	I	—	D	—	—	—	—	D	—	Short-chain fatty acids metabolism
atoSC	—	D	x	D	—	—	—	S	D	—	Two-component system
atp	S	S	S	S	S	S	D	x	x	D	ATP synthase
baeSR	—	S	x	—	—	—	—	x	D	—	Two-component system
basRS	—	x	x	—	—	—	—	x	D	—	Two-component system
β	S	S	S	S	S	S	S	S	S	D	Ribosomal protein
bglGFB	—	—	—	S	—	—	—	—	—	D	β-glucoside transport
bioBFCD	S	D	D	S	—	—	S	x	x	D	Biotin biosynthesis
btuCED	x	x	x	x	—	—	x	—	x	x	Vitamin B12 transport
cad	—	—	x	x	—	—	—	x	—	D	Lysine decarboxylase
cai	—	—	x	x	—	—	x	x	D	D	Carnitine
carAB	—	D	D	I	—	—	D	S	I	x	Carbamoyl-phosphate synthetase
ccmA-H	S	x	x	x	x	—	—	—	D	x	c-type cytochrome biogenesis
celABCDF	—	—	—	S	—	—	—	—	—	—	Cellobiose uptake
cheRBYZ	—	—	x	D	—	x	D	—	D	—	Motility and chemotaxis
chpSB	—	—	x	x	—	—	—	—	—	—	Regulation of cell growth
clpPX	I	D	I	D	—	—	—	—	—	x	Clp protease
cobUST	—	—	D	x	—	—	D	x	D	x	Cobalamin biosynthesis
codBA	x	—	x	—	—	—	—	—	—	—	Cytosine transport
copRS	—	x	D	x	—	—	—	x	D	—	Two-component system
cpxAR	x	S	—	x	—	—	—	x	D	—	Two-component system
creABCD	—	x	—	—	—	—	x	x	D	—	Phosphate sensor
csgBA	—	—	—	—	—	—	—	—	—	—	Curlin
csgDEFG	—	—	x	—	—	—	—	—	—	D	Curli
cyd	I	—	I	I	—	—	x	—	x	—	Cytochrome *d* oxidase
cynTSX	—	x	D	x	—	—	—	x	x	—	Cyanate utilization
cyoABCDE	—	x	S	S	—	—	x	x	D	x	Cytochrome *o* ubiquinol oxidase
cysDNC	—	—	x	x	—	—	x	—	D	D	Cysteine biosynthesis
cysJIH	—	—	D	S	x	—	x	—	—	D	Cysteine biosynthesis
cysPTWAM	—	D	S	x	x	x	S	S	D	—	Thiosulfate transport
dadAX	-	S	x	D	—	—	x	x	—	x	D-amino acid dehydrogenase
dam	S	D	D	D	D	D	D	x	D	D	dam superoperon
dapAnlp	I	x	x	x	—	—	x	x	x	—	Lysine biosynthesis
dcd	I	x	D	x	—	—	x	x	—	—	Deoxycytidine triphosphate deaminase
dcm-vsr	x	x	x	x	—	—	x	x	—	—	DNA processing
dedCD	x	x	x	x	—	—	—	x	—	x	Folylpolyglutamate synthase
def-fmt	I	D	D	I	D	D	—	—	x	x	tRNA modification
deoAB	D	S	x	S	S	S	D	D	D	x	Nucleotide and deoxyribonucleotide catabolism
dfp-dut	I	D	D	D	—	—	x	D	x	D	DNA and pantothenate metabolism
dicB	—	—	—	—	—	—	—	—	—	x	Cell division control
dmsABC	I	—	x	x	—	—	—	x	D	x	Dimethyl sulfoxide reductase
dnaA	I	D	D	S	D	D	—	—	—	D	Replication
dnaK	I	D	D	I	D	D	x	I	x	D	Molecular chaperone
dnaTC	—	—	—	x	—	—	—	—	—	—	Replication
dpp	S	I	D	x	x	x	x	x	S	—	Dipeptide transport
dsdXA	x	—	—	x	—	—	—	—	—	—	D-serine uptake
ebgAC	—	—	—	—	—	—	—	—	x	—	β-galactoside utilization
edd-eda	x	I	D	x	—	—	x	—	—	x	Entner-Doudoroff pathway
entCEBA	—	—	—	S	—	—	—	—	x	x	Enterobactin biosynthesis
fabDGacpP	I	S	D	I	x	x	—	—	—	D	Fatty acid biosynthesis
fadBA	—	—	—	x	—	—	x	x	—	x	Fatty acid degradation

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

Table 1
Continued

	Hin	*Hpy*	*Syn*	*Bsu*	*Mpn*	*Mge*	*Mja*	*Mth*	*Afu*	*Sce*	Function
fdnGHI	S	—	—	—	—	—	x	—	x	—	Formate dehydrogenase
fecABCDE	D	S	S	S	—	—	x	—	S	—	Citrate-dependent iron(III) transport
feoAB	—	x	I	—	—	—	x	I	x	—	Ferrous iron transport
fepDGC	x	S	S	S	—	x	x	—	S	—	Ferric enterobactin uptake
fhuACDB	D	S	x	—	—	x	x	x	x	x	Ferrichrome-iron transport
fic-pabA	x	x	x	x	—	—	x	x	x	x	Para-aminobenzoate synthetase
fis	S	D	x	x	—	—	D	—	x	D	DNA inversion
fix	—	—	x	S	—	—	x	x	S	D	Redox process
flgAMN	—	—	—	x	—	—	—	—	—	—	Flagellar biosynthesis
flgBCDEFGHI	—	S	—	S	—	—	—	x	—	—	Flagellar biosynthesis
flgKL	—	D	—	I	—	—	—	—	—	x	Flagellar biosynthesis
flhBA	—	D	—	I	—	—	—	—	—	—	Flagellar biosynthesis
flhDC	—	—	—	—	—	—	—	—	—	—	Flagellar transcriptional activation
fliAZY	x	D	—	D	—	—	—	—	—	D	Flagellar biosynthesis
fliDST	—	S	—	S	—	—	x	—	x	—	Flagellar biosynthesis
fliFGHIJK	—	S	x	S	—	—	x	D	x	x	Flagellar biosynthesis
fliLMNOPQR	—	S	—	S	—	—	—	—	—	—	Flagellar biosynthesis
frd	I	S	—	S	—	—	D	D	S	—	Fumarate reductase
fru	I	—	—	S	S	S	—	—	—	x	Fructose uptake
frv	—	—	—	x	x	x	x	D	x	—	Fructose-specific enzymes
ftsYEX	I	D	x	S	x	x	x	D	x	—	Cell division
fucAO	x	—	—	x	—	—	—	D	D	x	Fucose metabolism
fucPIK	S	x	—	—	—	—	x	x	x	x	Fucose uptake metabolism
gal	S	x	x	D	x	x	x	D	D	S	Galactose metabolism
gat	—	x	x	x	D	x	—	—	x	x	Galactitol uptake
gcv	—	—	D	S	—	—	—	—	—	D	Glycine cleavage system
glc	—	x	D	x	—	—	x	—	x	x	Glycolate utilization
glgBXCAP	I	—	D	S	—	—	D	x	—	D	Glycogen synthesis
glmUS	D	D	D	D	—	—	x	D	x	x	Amino sugar biosynthesis
glnALG	x	D	D	D	—	—	x	D	D	x	Glutamine synthetase
glnHPQ	—	S	x	x	x	x	x	—	x	—	Glutamine transport
glpABC	I	x	—	—	—	—	D	—	—	D	Glycerol-3-phosphate dehydrogenase
glpFK	I	—	x	I	D	D	—	—	x	D	Glycerol uptake
glpTQ	S	x	—	x	—	—	—	—	—	D	sn-glycerol-3-phosphate uptake
gltBDF	—	x	D	S	—	—	x	x	x	x	Induction of the Ntr enzymes
gltJKL	—	x	x	x	x	x	x	—	x	—	Glutamate/aspartate transport
glvCBG	x	—	—	S	—	—	—	—	—	—	Glucoside uptake
glyQS	S	D	D	I	—	—	—	x	—	—	Glycine tRNA synthetase
gntKU	—	—	x	—	—	—	x	—	—	x	Gluconate uptake and catabolism
groE	I	I	I	I	I	I	x	x	x	D	Molecular chaperone
guaBA	I	D	D	D	—	—	D	D	x	D	Purine catabolite
hem	S	x	x	S	—	—	x	x	x	D	Porphyrin biosynthesis
hflA	S	x	x	x	—	—	x	—	x	x	High frequency of lysogenization
hip	x	—	—	—	—	—	x	—	x	—	High-frequency persistence to the lethal effects
hisGDCBHAFE	I	—	D	S	—	—	D	D	S	D	Histidine biosynthesis
hisMP	—	—	—	—	x	x	x	—	—	x	Histidine transport
hisT	D	D	D	D	x	x	D	D	D	D	Pseudouridine synthase
htrE	—	—	—	—	—	—	—	—	—	x	Periplasmic protein
hya	—	S	S	—	—	—	D	x	S	x	Hydrogenase-1
hyb	—	S	D	—	—	—	D	D	S	—	Hydrogenase-2
hyc	—	—	x	—	—	—	S	S	D	D	Hydrogenase-3
hydHG	—	D	x	x	—	—	—	S	D	—	Two-component system
hypABCDE	x	S	D	—	—	—	D	S	S	x	Hydrogenase
ileS-lsp	D	D	D	S	D	D	x	x	x	x	tRNA synthetase, peptidase
ilvBN	—	x	—	—	—	—	—	—	—	—	Acetolactate synthase I
ilvGMEDA	S	x	D	D	—	—	D	D	D	D	Valine and isoleucine biosynthesis
ilvIH	I	—	D	I	—	—	D	I	I	D	Acetolactate synthase III
kdpABC	—	x	S	—	—	—	—	D	D	—	Membrane K transport ATPases
kdsA	S	D	D	S	S	S	x	D	D	D	3-deoxy-D-manno-octulosonic acid metabolism and enterobacterial lipopolysaccharide biosynthesis
kdtAB	I	D	x	x	—	—	—	x	x	—	Lipopolysaccharide biosynthesis
L11	I	I	I	I	I	I	D	I	D	D	Ribosomal protein
lac	—	x	x	—	x	—	—	—	—	x	Galactoside utilization
lct	D	x	—	D	x	—	—	—	D	x	L-lactate dehydrogenase
lepAB	I	D	D	D	x	x	—	x	—	D	Secretory machinery
leu	I	—	D	I	—	—	D	S	S	D	Leucine biosynthesis
livKHMGF	—	—	D	x	—	—	S	—	S	x	Branched-chain amino acid transport
lpx	S	S	D	S	D	x	x	D	D	x	Lipid biosynthesis
malEFG	—	—	—	I	—	S	x	x	—	D	Maltodextrins and maltose transport

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

Table 1
Continued

	Hin	*Hpy*	*Syn*	*Bsu*	*Mpn*	*Mge*	*Mja*	*Mth*	*Afu*	*Sce*	Function
malK-lamB	—	x	—	—	x	x	x	x	—	—	Maltose, maltodextrins uptake
malPQ	x	—	x	x	—	—	x	—	—	x	Glycogen debranching
malXY	—	—	—	x	x	x	—	—	—	—	Maltose metabolism
manXYZ	—	—	—	I	—	—	—	—	—	—	Mannose uptake
marRAB	—	—	—	D	—	—	x	—	x	—	Antibiotic resistance
mdoGH	—	x	—	—	—	—	—	—	—	—	Oligosaccharide biosynthesis
melAB	—	—	x	x	—	—	—	—	—	—	Melibiose uptake
menFDBCE	S	x	D	S	x	—	—	—	x	D	Menaquinone biosynthesis
metBL	x	D	x	D	—	—	x	x	x	D	Methionine biosynthesis
mglBA	I	—	—	—	x	x	x	x	D	—	Galactose transport
mhp	—	—	x	D	x	x	x	x	x	D	Catechol dioxygenases
moaABCDE	S	S	S	S	—	—	D	D	D	—	Molybdenum cofactor biosynthesis
modABC	I	S	x	S	—	—	D	S	S	—	Molybdate transport
moeAB	I	D	D	S	—	—	x	D	D	x	Molybdopterin biosynthesis
motABcheAW	—	S	S	S	—	—	x	—	D	x	Chemotaxis
mrdAB	I	D	x	x	—	—	—	x	—	—	Rod shape
mtlAD	—	—	—	I	I	—	—	—	—	—	Mannitol uptake
murGC	I	D	D	D	—	—	—	D	—	—	Peptidoglycan synthesis
nagBACD	S	—	D	S	—	—	—	—	D	D	*N*-acetylglucosamine uptake
napAGHBC	S	x	—	—	—	—	D	—	D	x	Electron transfer
narGHJI	—	—	—	I	—	x	—	—	x	x	Nitrate reductase
narXL	—	—	I	I	—	—	—	x	—	—	Two-component system
narZYWV	—	—	—	I	x	—	x	—	x	—	Nitrate reductase
nikABCDE	—	—	D	x	S	S	x	x	—	x	Nickel transport
nirBDC	—	—	—	D	—	x	x	x	x	—	NADH-dependent nitrite reductase
nrdAB	I	x	x	—	—	—	—	x	x	D	Ribonucleoside diphosphate reductase
nrdEF	—	x	x	I	S	S	—	x	x	x	Class I ribonucleotide reductase
nrf	S	D	x	D	—	x	x	—	S	D	Nitrite reduction
nuo	—	S	S	D	—	—	D	S	S	D	NADH dehydrogenase
nusA	I	S	I	S	S	S	x	—	D	D	Transcription, translation
ompB	—	S	x	x	—	—	—	x	D	—	Two-component system
oppABCDF	I	S	D	I	S	S	x	D	S	—	Oligopeptide transport
otsBA	—	—	x	—	—	—	—	S	—	x	Trehalose synthesis
pdxH-tyrS	D	x	D	x	x	x	—	—	—	D	tRNA synthetase, oxidase
pdxJ	—	D	x	x	x	—	—	—	—	—	Pyridoxal 5′-phosphate biosynthesis
pheST	I	I	D	I	I	I	D	D	D	D	Phenylalanine tRNA synthetase
phn	—	—	x	D	x	S	—	D	D	D	Phosphonate utilization
phoBR	I	S	D	I	—	—	—	S	D	—	Phosphate regulon
phoPQ	—	x	—	—	—	—	—	x	D	x	Two-component system
potFGHI	—	x	—	—	S	S	D	x	S	—	Putrescine transport
ppx	x	D	D	—	—	—	—	—	—	D	Inorganic polyphosphate degradation
pqiAB	—	—	—	—	—	—	—	—	—	D	Paraquat-inducible protein
prfBlysS	x	D	D	D	x	x	x	—	x	x	Peptide, chain release factor
prmA	S	x	x	x	—	—	x	x	D	x	Ribosomal protein
proBA	D	—	D	I	—	—	—	x	—	D	Proline biosynthesis
proVWX	—	D	—	I	x	x	x	—	x	-	Proline transport
pspABCDE	—	—	D	D	—	—	—	—	—	x	Phage shock protein
ptr-recBD	S	—	x	x	—	—	x	—	x	D	Endopeptidase, exonuclease
pts	I		—	S	D	D	—	x	—	—	Sugar uptake
purEK	I	—	D	I	—	—	x	x	x	x	Phosphoribosylaminoimidazole carboxylase
purF	S	D	D	D	—	—	D	D	D	D	Purine biosynthesis
purHD	I	x	D	I	—	—	x	x	D	D	Purine biosynthesis
purMN	I	—	D	I	—	—	x	x	x	x	Purine biosynthesis
pyrBI	—	x	x	D	—	—	D	D	I	—	Pyrimidine biosynthesis
pyrF	I	x	D	x	—	—	x	D	x	x	OMP decarboxylase
rbfAtruB	I	x	D	I	x	x	x	x	x	x	RNA modification
rbs	I	—	x	S	S	S	D	D	x	D	Ribose transport
recR	I	D	D	I	x	x	x	—	—	—	DNA recombination
relA	D	x	D	D	x	x	—	—	x	—	Amino acid starvation
rfa	D	x	—	D	—	—	x	—	x	x	Lipopolysaccharide biosynthesis
rfb	x	—	D	S	x	x	D	S	S	x	Lipopolysaccharide biosynthesis
rhaBAD	—	—	—	S	x	—	—	—	—	—	L-rhamnose uptake
rnc	S	D	D	S	D	x	—	x	—	x	DNA/RNA processing
rph-pyrE	I	—	x	D	—	—	x	x	x	D	Ribonuclease, pyrimidine biosynthesis
rplMrpsI	I	I	I	I	I	I	I	x	I	D	Ribosomal protein
rplUrpmA	I	I	I	S	S	S	—	—	—	D	Ribosomal protein
rpmF	I	—	—	x	x	x	—	—	x	—	Ribosomal protein
rpmH	I	x	I	I	I	I	—	—	—	x	Ribosomal protein
rpoN	S	S	S	D	D	—	D	x	D	x	Organic nitrogen uptake and assimilation
rpsA	I	D	x	I	x	x	—	—	—	D	Ribosomal protein

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

Table 1
Continued

	Hin	*Hpy*	*Syn*	*Bsu*	*Mpn*	*Mge*	*Mja*	*Mth*	*Afu*	*Sce*	Function
rpsB	S	S	S	I	S	S	D	D	D	D	Ribosomal protein
rpsF-rplI	I	S	D	S	S	S	—	—	—	x	Ribosomal protein
rpsO-pnp	D	D	D	I	x	x	D	D	x	D	Ribosomal protein
rpsU-dnaG-rpoD	I	D	D	S	S	S	—	x	x	x	Ribosomal protein
rst	—	x	—	—	—	—	—	x	D	x	Two-component system
ruvAB	I	D	D	I	I	I	x	—	—	—	DNA repair and recombination
S10	I	I	S	I	S	I	S	S	S	D	Ribosomal protein
sap	S	x	D	—	S	S	D	x	—	—	Peptide uptake
sdaCB	I	I	—	x	—	—	—	—	—	—	Serine uptake
sdhCDAB	—	S	D	S	—	—	D	D	—	D	Succinate dehydrogenase
secD	S	S	S	S	—	D	S	x	—	D	tRNA modification, protein export
secE-nusG	I	x	I	x	I	x	x	—	—	—	Protein export, transcription
serA	I	x	D	x	—	—	D	D	D	D	L-serine biosynthesis
serBsms	D	D	D	x	—	—	x	x	x	x	Phosphatase, DNA repair
smtAmukFEB	S	x	—	—	—	—	—	—	—	D	Cell division
spc	I	S	S	S	S	S	S	S	S	D	Ribosomal protein
speED	—	x	—	x	—	—	S	—	x	x	Spermidine biosynthesis
speFpotE	I	—	x	x	—	—	—	x	—	x	Ornithine uptake
spoT	S	S	D	S	D	D	x	—	—	D	DNA/RNA processing
srlABD	x	x	x	D	—	—	—	x	x	—	Glucitol uptake
sspAB	I	—	—	x	—	—	—	x	—	—	Stringent starvation protein
str	I	S	I	I	S	S	S	I	S	D	Ribosomal protein
sucABCD	S	—	D	S	x	x	D	D	S	D	TCA cycle
surA-pdxA-ksgA-apaGH	S	D	D	D	D	x	x	D	D	D	RNA modification, chaperone
tdc	—	x	—	—	—	—	x	D	x	x	Threonine dehydratase
tdh	x	x	—	S	x	—	x	—	x	x	L-threonine metabolism
tehAB	D	—	—	—	—	—	D	—	x	—	Potassium tellurite resistance
thiCEFGH	D	D	D	S	—	x	x	D	D	x	Thiamin biosynthesis
thr	I	D	D	D	—	—	D	S	x	D	Threonine biosynthesis
tnaAB	—	—	—	x	—	—	x	—	—	x	Tryptophan transport
tolQRA	I	D	S	x	—	—	—	D	x	—	Colicin uptake
torCAD	—	x	—	D	—	—	x	—	D	—	Trimethylamine-N-oxide reductase
treBC	—	—	—	I	—	—	—	—	—	x	Trehalose uptake
trmD	I	S	D	S	S	S	—	—	x	x	Ribosomal protein
trp	S	I	D	S	x	—	S	S	S	D	Tryptophan biosynthesis
ttdAB	—	—	—	x	—	—	D	D	I	—	L-tartrate dehydratase
tyr	x	x	x	—	—	—	x	x	—	x	Aromatic amino acid biosynthesis
ubiCA	—	x	x	—	—	—	x	x	x	x	Ubiquinone biosynthesis
ugpBAECQ	—	—	D	D	S	S	x	x	—	—	sn-glycerol 3-phosphate transport
uhp	x	—	x	S	x	—	—	x	D	D	Hexose phosphate transport
uidABC	—	—	x	x	—	—	—	—	—	—	β-glucuronidase
umuDC	—	—	x	—	—	—	—	—	—	—	SOS response
upp-uraA	I	—	x	D	x	x	—	x	—	x	Uracil utilization
uvrYC	x	x	D	D	D	x	—	x	x	—	DNA repair
uxaCA	—	x	x	D	—	—	—	x	—	x	Pentose and glucuronate interconversion
uxuAB	x	—	—	x	—	—	—	—	—	D	Glucuronate pathway
xapAB	—	—	x	x	—	—	x	x	x	x	Xanthosine uptake
xylFGH	S	—	—	—	x	x	x	D	x	—	Xylose transport

NOTE.—Abbreviations: *Afu, Archaeoglobus fulgidus*; *Bsu, Bacillus subtilis*; *Eco, Escherichia coli*; *Hin, Haemophilus influenzae*; *Hpy, Helicobacter pylori*; *Mge, Mycoplasma genitalium*; *Mja, Methanococcus jannaschii*; *Mpn, Mycoplasma pneumoniae*; *Mth, Methanobacterium thermoautotrophicum*; *Sce, Saccharomyces cerevisiae*; *Syn, Synechocystis* sp.; I, identical; S, similar; D, destructed; x, unknown (only one ortholog found); —, unknown (no ortholog found).

Discussion

Detection of Orthologous ORF Pairs

Our method could detect orthologous pairs between not only closely but also distantly related organisms such as eubacteria and archaebacteria (tables 1 and 2). This indicates that the method is sufficiently effective in finding orthologs. However, the following three points should be carefully noted. First, when a pair of orthologs is highly diverged, it is difficult to detect the pair correctly by a sequence similarity search. Second, when orthologs consist of more than two domains, it is possible that orthologs may be identified by using only the domain with the highest similarity. Even if the remaining domains show significant similarity to different proteins, these domains are neglected. Finally, if some of the genes have been horizontally transferred, phylogenetic relationships among relevant species should be taken into account. Although these points should be improved for better detection of orthologs, the method is powerful and suitable for the present study, that is, an automatic mass analysis of orthologs between complete genomes.

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

Table 2
Conservation of Structures of 100 *Bsu* Operons in Other Genomes

	Eco	*Hin*	*Hpy*	*Syn*	*Mpn*	*Mge*	*Mja*	*Mth*	*Afu*	*Sce*	Function
acuABC	—	—	—	x	—	—	x	D	x	D	Acetoin utilization
ada	x	—	x	x	—	—	x	x	x	D	DNA alkyltransferase
als	—	—	x	x	—	—	—	—	—	x	Acetolactate
ans	D	x	x	—	—	—	x	x	x	—	L-asparaginase and L-aspartase
app	S	S	S	D	S	S	—	—	S	—	Oligopeptide permease
ara	S	S	—	D	S	S	D	D	D	D	L-arabinose
argC-F	S	D	D	D	S	—	D	S	D	D	Citrulline biosynthesis
atp	S	S	S	S	S	S	x	x	S	D	ATP synthase
bioWAFDBI	S	S	D	D	—	x	S	x	x	D	Biotin biosynthesis
cel	S	—	—	—	x	—	—	—	—	x	Cellobiose phosphotransferase
cgeAB	—	—	—	—	—	—	—	—	—	—	Sporulation
cgeCDE	—	—	—	x	—	—	—	—	—	—	Sporulation
codVWXY	S	S	S	x	—	—	x	D	D	—	Dipeptide permease operon regulation
comA	x	x	—	—	—	—	—	x	x	x	Late competence
comE	D	D	x	D	x	x	—	—	x	x	Late competence
comF	x	x	x	x	—	—	—	—	—	x	Late competence
comG	S	S	x	S	—	x	D	S	x	x	Late competence
cotJ	—	—	—	—	—	—	—	—	—	—	Spore coat composition
cotR	x	x	x	x	x	x	—	—	—	x	Spore coat
dap	D	D	D	S	—	—	D	S	D	D	Diaminopimelate
dhbACEBF	S	x	x	x	—	—	x	x	—	x	2,3-dihydroxybenzoate biosynthesis
dnaK	S	S	S	S	D	D	D	S	D	D	Chaperone
dpp	S	S	S	D	S	S	—	—	S	x	Dipeptide transport system
dra-nupC-pdp	S	x	—	x	S	S	D	x	x	—	Induction by deoxyribonucleosides
ecsABC	—	—	x	x	D	—	—	—	—	—	ABC transporter
feuABC	x	—	x	—	x	—	x	—	—	—	Iron-uptake system
flaA	S	D	S	x	D	x	D	x	D	x	Flagellar structure
fliDST	S	—	S	—	—	—	—	—	—	—	Flagellar protein
folic acid	D	S	S	D	—	—	D	S	S	D	Folic acid biosynthesis
ftsAZ	I	I	I	x	x	x	x	x	x	x	Developmental regulation
gbsAB	x	—	D	x	—	—	D	D	x	x	Glycine betaine synthesis
gcv	S	x	—	D	—	—	—	—	—	D	Glycine-cleavage system
gerA	x	x	—	—	—	—	—	—	—	x	Spore germination
gerB	x	x	x	—	—	—	—	—	—	x	Spore germination
glgBCDAP	S	S	—	D	—	x	D	x	x	D	Glycogen biosynthesis and degradation
glnQHMP	x	x	S	D	—	—	—	x	S	x	Glutamine ABC transporter
glnRA	x	x	D	x	—	—	x	x	x	D	Glutamine synthetase
glpFK	I	I	—	D	D	D	—	x	D	D	Glycerol catabolism
glpTQ	D	x	—	x	x	x	—	x	—	D	Glycerol catabolism
glv	S	—	—	x	x	x	—	x	x	x	Membrane transport
gnt	x	D	x	x	—	—	—	—	—	D	Gluconate
groESL	I	I	I	I	I	I	x	x	D	D	Chaperonin
hemAXCDBL	S	—	S	D	—	—	D	D	S	D	Haem synthesis
hemEHY	D	x	D	D	—	—	—	D	—	D	Protoheme IX biosynthesis
hut	D	—	x	D	D	x	x	x	D	D	Histidase
ilv-leu	S	S	D	D	—	—	D	S	S	D	Acetolactate synthesis
iol	D	D	D	D	D	x	D	x	x	D	Inositol dehydrogenase
kinB-kapB	—	—	x	—	—	—	—	x	x	x	Phosphorelay initiating sporulation
L21	S	S	S	S	I	I	—	—	—	D	Ribosomal protein
lepA-hemN	D	D	D	D	x	x	—	—	x	x	Coproporphyrinogen III oxidase
levanase	S	—	—	—	x	—	—	—	—	x	Fructose phosphotransferase
lytABC	x	D	D	D	—	—	x	x	x	—	Autolysin
menBE	x	x	x	D	—	—	—	—	x	D	Menaquinone
menDC	S	I	—	D	—	x	x	x	—	x	Menaquinone
mmgABCDE	S	x	D	x	x	—	x	x	S	D	Catabolite repression
motAB	I	x	I	x	—	—	—	x	—	—	Motility
mre-min	S	S	S	S	—	—	x	D	D	—	Rod shape, cell division
mtr	x	x	x	x	—	—	—	x	—	x	Transcription attenuation
mutSL	D	D	—	D	—	—	—	x	—	D	Mismatch repair
nasB	S	x	—	D	—	—	D	D	D	D	Nitrate/nitrite assimilation
nrd	S	S	D	D	S	S	—	x	x	D	Ribonucleotide reductase
nrgAB	S	x	—	D	—	—	S	I	I	x	Membrane-associated protein
nusA-infB	S	S	S	S	S	S	—	x	D	D	Translation, transcription
odhAB	I	I	—	x	x	x	—	—	—	D	2-oxoglutarate dehydrogenase
opp	I	S	S	D	S	S	—	x	S	x	Oligopeptide transport
pbpE	D	x	—	x	—	—	—	—	x	x	Penicillin-binding protein
PD-1	D	—	x	—	—	—	D	—	x	—	Flagellin synthesis
pho	I	I	S	D	—	—	x	S	D	x	Alkaline phosphatase regulation
ponA	x	x	x	—	—	x	—	—	—	—	Penicillin-binding protein

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

Table 2
Continued

	Eco	*Hin*	*Hpy*	*Syn*	*Mpn*	*Mge*	*Mja*	*Mth*	*Afu*	*Sce*	Function
pps	x	—	—	x	—	—	—	—	—	—	Peptide synthase
proU	S	—	D	x	x	x	x	x	S	D	Osmoprotection
pst	S	—	—	S	S	S	S	S	S	—	Phosphate-specific transport
ptsHI	I	I	—	—	D	D	—	x	—	—	Sugar phosphotransferase system
pur	S	S	D	S	—	—	S	S	S	D	Purine biosynthesis
pyr	S	D	D	S	D	—	D	S	S	D	Pyrimidine biosynthesis
qcr	x	—	S	S	—	—	—	—	—	x	Menaquinol : cytochrome *c* reductase
qox	I	—	x	S	—	—	—	D	x	x	Quinol oxidase
rbs	S	S	D	D	x	x	D	x	D	D	Ribose transport
rib	S	D	D	D	—	—	D	D	D	D	Riboflavin biosynthesis
rocDEF	—	—	D	x	x	—	—	x	—	D	Arginine catabolism
S10-*spc*-*α*	S	S	S	S	S	S	S	S	S	D	Ribosomal protein
sacPA	—	—	—	—	—	—	—	—	—	—	Sucrase
sacXY	—	x	x	—	—	—	—	—	—	D	Levanesucrase biosynthesis
sdh	S	S	S	D	—	—	—	—	S	D	Succinate dehydrogenase complex
sigA	I	I	D	D	I	I	—	x	D	—	RNA polymerase
sigB	D	x	—	D	—	—	—	—	—	—	Transcription factor, stress
spo0B	D	D	D	D	x	x	D	D	D	D	Sporulation, phenylalanine biosynthesis
spoIIIA	—	—	—	D	—	—	—	—	—	—	Sporulation
spoIVF	—	—	—	x	—	—	x	x	D	—	Sporulation
spoVA	—	—	—	—	—	—	—	—	—	x	Sporulation
srfA	—	—	x	D	x	—	x	x	—	—	Surfactin synthetase
tagABC	x	—	—	D	—	—	x	x	x	—	Polyglycerol phosphate biosynthesis
tagDEF	—	x	—	—	—	—	x	D	D	x	Polyglycerol phosphate biosynthesis
tagGH	—	—	—	I	—	—	—	—	I	—	Teichoic acid translocation
thyBdfrA	D	D	—	—	I	I	x	x	x	D	Thymidine biosynthesis, dihydrofolate reductase
trehalose	S	—	—	—	—	—	—	—	—	x	Trehalose
trp	S	S	S	D	—	—	S	S	S	D	Tryptophan
urease	—	I	S	D	—	—	—	—	—	—	Urease
xpt-pbuX	x	—	—	—	—	—	x	x	—	—	Xanthine
yllB-pbpB	S	S	D	D	S	S	—	—	—	x	Cell division

NOTE.—Abbreviations as in table 1.

Unstable Structures of Operons

As shown in comparisons of several genome sequences, the genes on the genomes seem to have been, in general, randomly shuffled (Watanabe et al. 1997). On the other hand, operon structures were expected to be more conserved than the outside regions of operons, because the polycistronic property of an operon structure is thought to be of functional importance. According to our results, however, only 56% of operons were identical between *Eco* and *Hin.* Moreover, the proportion of identical operon structures was 13% between *Eco* and *Hpy.* Consequently, the degree of conservation of operon structures was found to be generally quite low (fig. 4). These observations suggest that conservation of operon structures appears to be less important than expected, implying that destruction of operon structures is almost selectively neutral during long-term evolution. Functional constraint against coexpression of genes may be so weak that the organization of gene clusters in operon structures can be readily changed during evolution.

Genome rearrangement is thought to be caused by recombination between homologous DNA sequences (Himmelreich et al. 1997). Although in *Bsu,* the rate of inversions was estimated to be low (Toda, Tanaka, and Itaya 1996), duplications and deletions are known to frequently occur in *Eco* and *Salmonella* (Roth et al. 1996). This suggests that genome structures have undergone frequent alteration during long-term evolution, such that almost all of the operons could have been rearranged.

Gene orders of ribosomal operons are well conserved in both eubacteria and archaebacteria. In a study of ribosome assembly in vitro, ribosomal protein operons were suggested to correspond to assembly units for forming a ribosome (Herold and Nierhaus 1987). Therefore, we expect that the gene orders may correspond to the assembly order of ribosomal proteins. Nonetheless, we found that the orders of genes within the operons appeared to be irrelevant to the order of their assembly. Thus, the order of genes in a ribosomal protein operon may not be important for their assembly to form a ribosome and may not be crucial for their function. Rather, the conservation of ribosomal protein operons can be explained by the possibility that the genome rearrangement has been deleterious for high expression of ribosomal genes. When an operon is destructed and split into two units, it is quite possible that the transcription efficiency drastically decreases in the latter units. This is because the latter unit may lack transcription regulatory regions. Such decreases in amounts of transcripts may be seriously deleterious, because ribosomal proteins play an essential role in all organisms and should be highly expressed. Consequently, the operon structures should have been retained in the course of evolution.

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

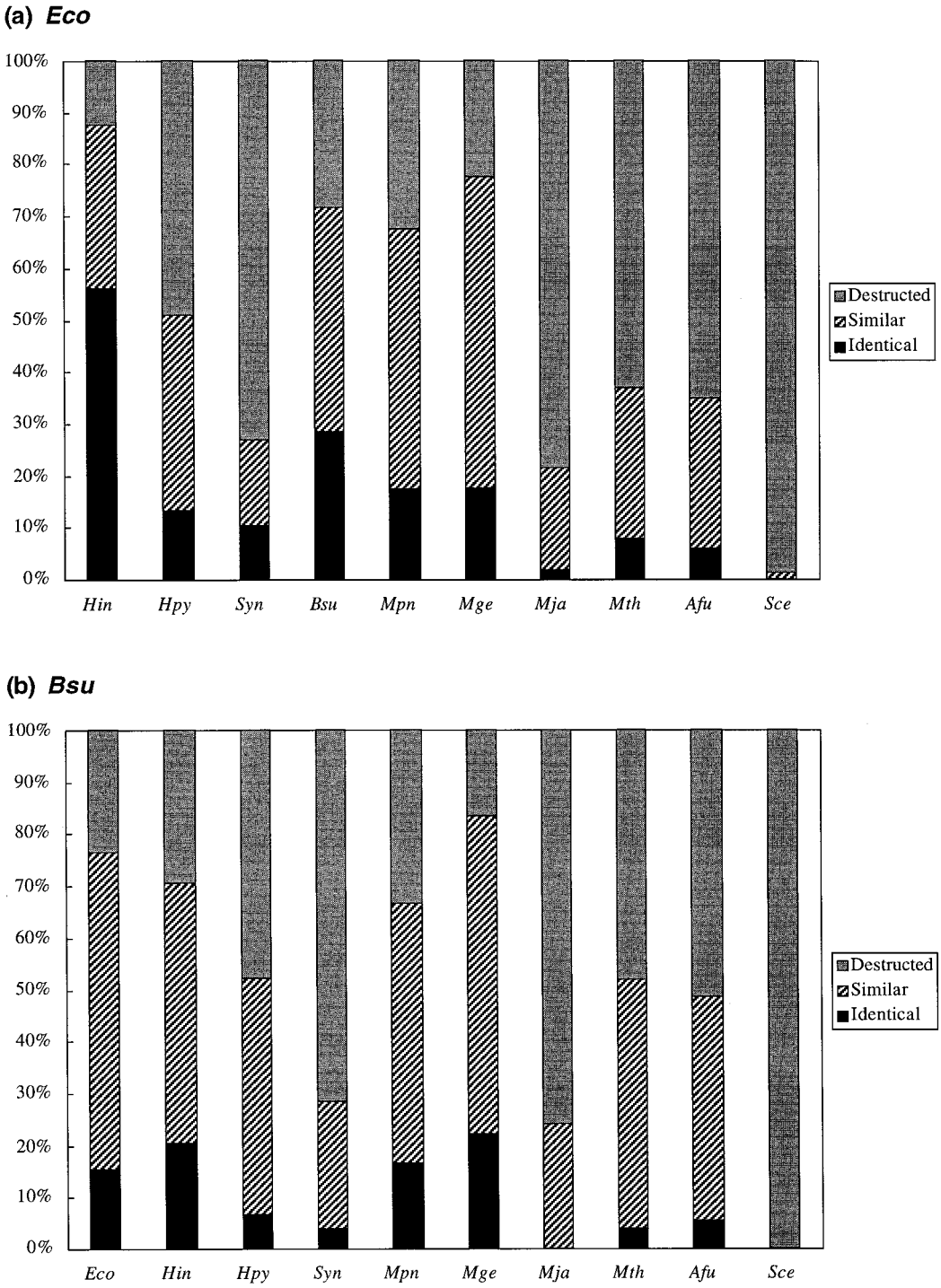

FIG. 4.—Conservation of structures of (*a*) 256 *Eco* and (*b*) 100 *Bsu* operons in other genomes. Black bars indicate identical portions, hatched bars indicate similar portions, and shaded bars indicate destructed portions. Unknown portions are not included.

Moreover, even operons which contain very important genes could sometimes be destructed in several species (e.g., fig. 5). In general, the conservation levels of operon structures appear irrelevant to the degree of their function (tables 1 and 2). This supports the above-mentioned notion that, in many cases, functional constraints against operon structures are weak and the structures can be frequently shuffled during long-term evolution.

Operon-like Structures in Archaebacteria and a Eukaryote

Although orthologous ORF pairs between an archaebacterium and a eubacterium are quite divergent,

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

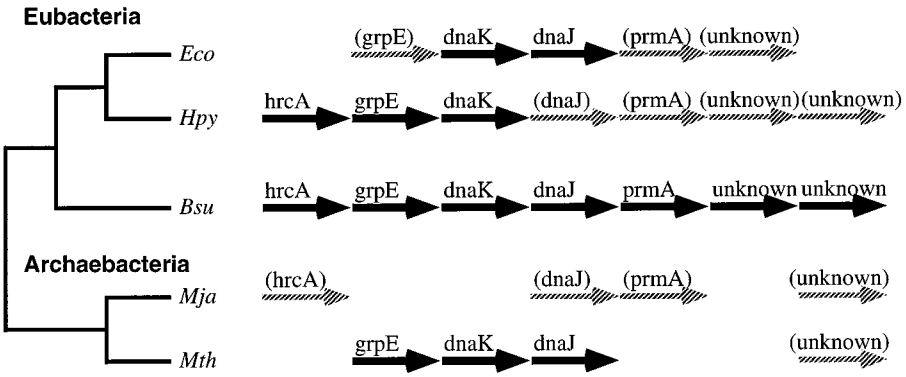

FIG. 5.—Conservation of the *dnaK* operons in five species. Thick arrows indicate known or putative cotranscribed gene clusters. Hatched arrows indicate translocated genes. An approximate phylogenetic relationship is shown on the left.

the relative degree of conservation of the operon structures can be evaluated by our analysis. In fact, *Mja* appears to be less conserved than *Mth* or *Afu* (fig. 4). There seems to be a difference in the instability of the operon structures among archaebacteria too.

There are more divergent relationships in operon structures between a eubacterium and a eukaryote *Sce*. Among the 256 *Eco* and 100 *Bsu* operons, only one gene cluster, the *gal* ortholog in the yeast genome, was found to retain an organization similar to that of the eubacterial operon. Since no other operon structures were found to be conserved, the genome structure of yeast seems to be

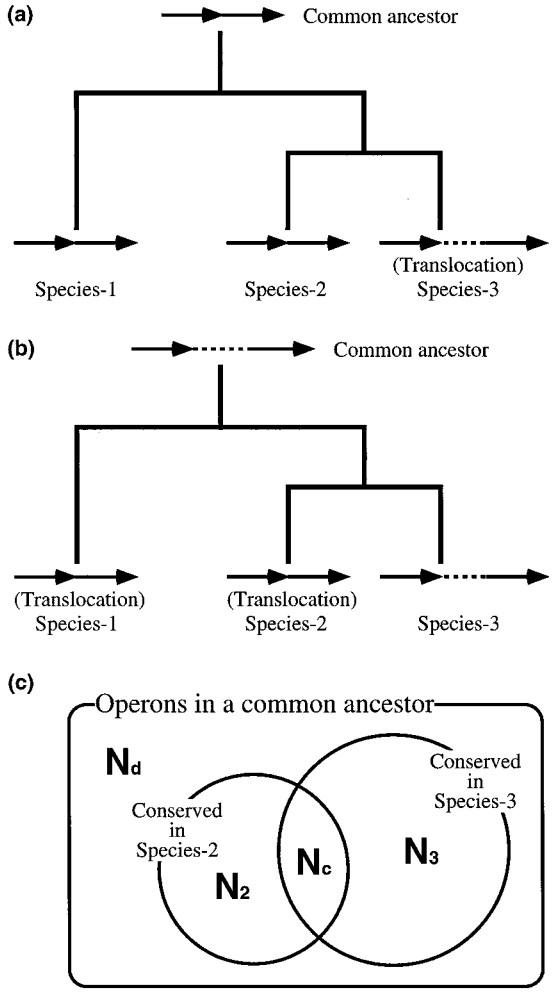

FIG. 6.—Parsimonious estimation of the ancestral operon structure.

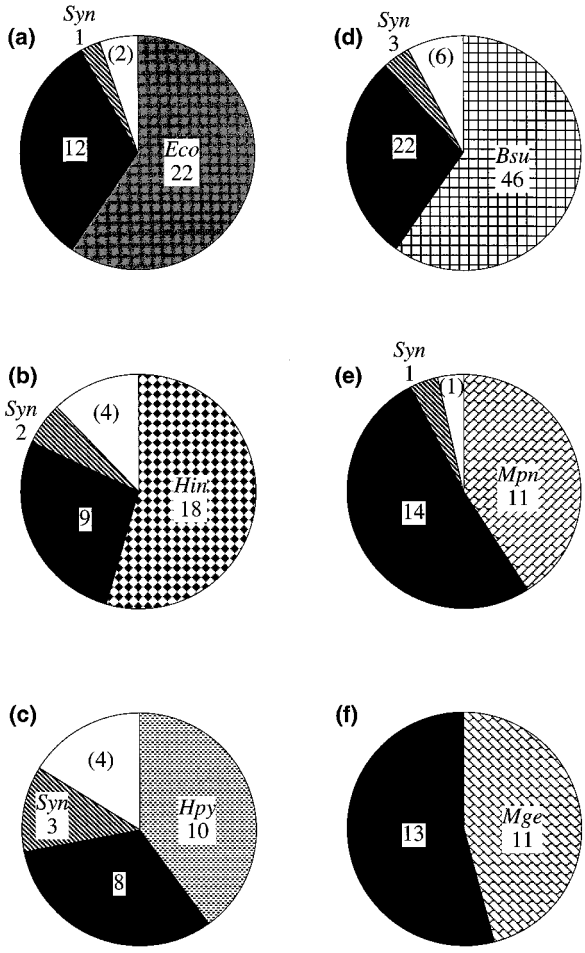

FIG. 7.—Relative stability of (*a–c*) the *Bsu* operon structures between a proteobacterium and *Syn* and (*d–f*) the *Eco* operon structures between a gram-positive bacterium and *Syn*. Black regions indicate conserved operons in both organisms. White regions indicate destructed operons in both organisms. The number of operons is shown in each region. Numbers in parentheses indicate estimated numbers of destructed operons.

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

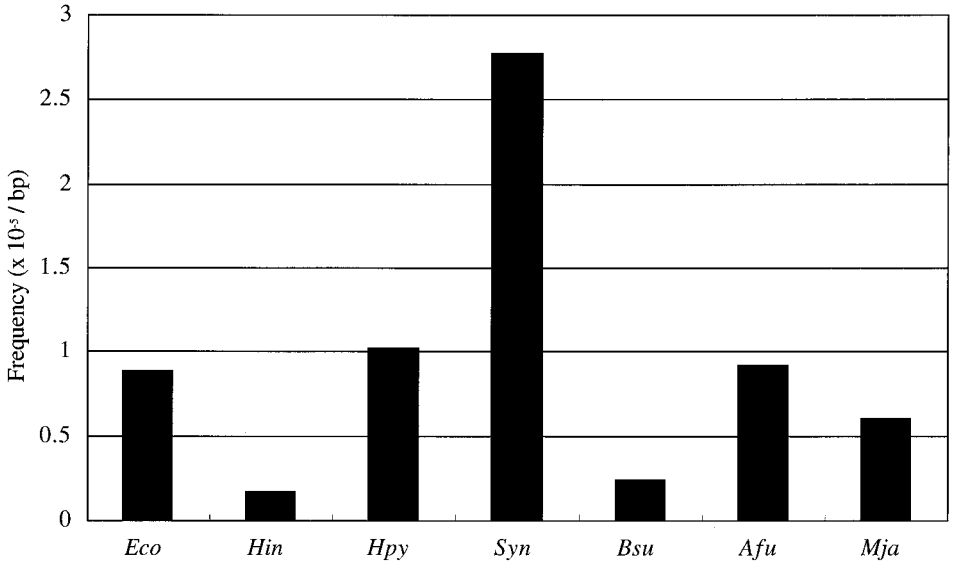

FIG. 8.—Frequency of occurrence of ISs per bp. The number of ISs in each genome was obtained from literature (Fleischmann et al. 1995; Fraser et al. 1995; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Blattner et al. 1997; Goffeau et al. 1997; Klenk et al. 1997; Kunst et al. 1997; Smith et al. 1997; Tomb et al. 1997). For *Syn,* the frequencies for transposases instead of ISs were calculated. Ten partial copies of ISs were included in *Hpy.* No IS-like element was reported in *Mge, Mpn,* or *Mth.*

completely different from that of eubacteria and archaebacteria. Its genome structure and transcription regulation may have evolved in a unique manner.

Important Role of Insertion Sequences in Instability of Operon Structures

Conservation levels of the operon structures were very different among the genome sequences studied here. The operons in *Bsu* were very well conserved. Indeed, its genome structure had been thought to be stable (Itaya 1993). In contrast, operons in *Syn* were drastically destructed. It is known that *Bsu* has few insertion sequence (IS)-like elements (Kasahara, Nakai, and Ogasawara 1997; Kunst et al. 1997), but *Syn* has about 100 transposases (Kaneko et al. 1996). If a genome sequence can be rearranged via homologous recombination, autonomously transposable elements like ISs are convincing candidates for a cause of genome instability (Naas et al. 1995; Deonier 1996). Although there exist a number of other repetitive DNA sequences in bacterial genomes (e.g., 314 BIME sequences in *E. coli*; Blattner et al. 1997), they are not large enough to frequently mediate homologous recombination, as a frequency of recombination increases with the size of the repetitive sequence (Bachellier et al. 1996). Moreover, ISs can be inserted and excised without regard to locations in a genome sequence, even within an operon. If ISs had been the main cause of genome rearrangements, positive

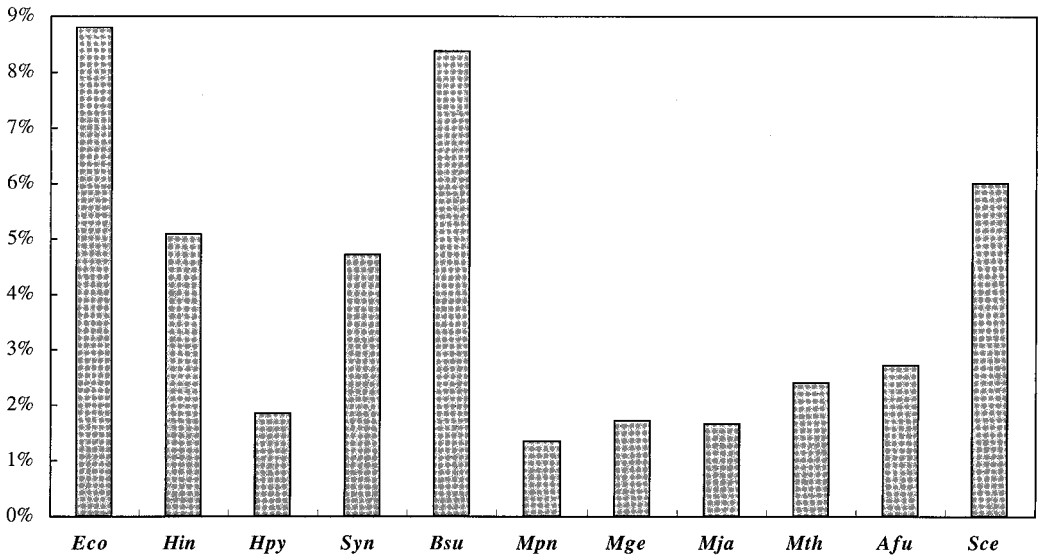

FIG. 9.—Proportions of possible transcription regulators. The number of predicted transcription regulators was divided by the number of all ORFs.

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

correlation would be observed between genome stability and the number of ISs. In fact, the unstable genomes of *Syn* and *Hpy* showed a relatively large number of ISs (fig. 8). No IS-like element was found in *Mge*, *Mpn*, or *Mth*, while approximately 8% of the *Mpn* genome is composed of repetitive DNA elements (Himmelreich et al. 1996). Note that the frequency of only reported ISs is shown in figure 8. *Eco* appears to possess many ISs (fig. 8), whereas *Bsu* contains few ISs. Accordingly, *Bsu* seems to have the most stable genome, and therefore it well retains the ancestral genome structure.

Moreover, *Synechococcus* sp., which is closely related to *Synechocystis* sp., contains multiple chromosome copies even at a slow growth rate (Binder and Chisholm 1990). Since this suggests that cyanobacteria have many chances for homologous recombination and their genome sequence can be shuffled more frequently than others, existence of multiple chromosomes may have accelerated genome rearrangements in cyanobacteria.

Effects on Transcription Regulation

When an operon is destructed and divided into two units, the latter unit requires new regulation for its transcription; otherwise, it will become a pseudogene(s). This may occur if the whole regulation of the operon becomes less important (i.e., alteration of the operon structure is almost selectively neutral) and if the members of the operon need not be highly expressed.

If an operon structure is destructed, the equivalent transcriptions of the destructed operons may be compensated by more complicated regulation. For all ORFs for each species, we computed the proportions of putative transcription regulators, which were identified by significant sequence similarity ($E < 10^{-4}$) to transcription regulators registered in SWISS-PROT 34 (fig. 9). The *Syn* and *Hpy* genomes, in which operon structures were drastically destructed, had fewer regulators than did *Eco*, *Hin*, and *Bsu*. Moreover, *Mpn* and *Mge* retained regulators at levels similar to that of *Hpy*. Although this observation is indirect evidence, it is suggested that the stability of a genome is irrelevant to the complexity of regulations if the number of regulators correlates simply to the degree of their complexity.

Unfortunately, much experimental information on operons and transcription regulation is unavailable for most species whose complete genomes have been sequenced, so it is difficult to predict transcription regulation in silico at present. Thus, more experiments for operons and their regulation are required in order to elucidate a relationship between operon destruction and its effect on regulation. Further systematic investigations of transcriptional regulation in these genomes are expected in the post-genome sequencing projects.

Acknowledgments

We appreciate Dr. S. Ohno for his helpful discussion and our colleagues for their comments. We would like to thank Drs. S. Gaudieri and M. Bellgard for proofreading the manuscript. This work was supported in part by a Grant-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan.

LITERATURE CITED

AIBA, H., T. BABA, K. HAYASHI et al. (33 co-authors). 1996. A 570-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 28.0–40.1 min region on the linkage map. DNA Res. **3**:363–377.

BACHELLIER, S., E. GILSON, M. HOFHUNG, and C. W. HILL. 1996. Repeated sequences. Pp. 2012–2040 *in* F. C. NEIDHARDT, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. B. LOW, B. MAGASANIK, W. S. REZNIKOFF, M. RILEY, M. SCHAECHTER, and H. E. UMBARGER, eds. *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.

BINDER, B. J., and S. W. CHISHOLM. 1990. Relationship between DNA cycle and growth rate in *Synechococcus* sp. strain PCC 6301. J. Bacteriol. **172**:2313–2319.

BLATTNER, F. R., G. PLUNKETT III, C. A. BLOCH et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1474.

BULT, C. J., O. WHITE, G. J. OLSEN et al. (40 co-authors). 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273**:1058–1073.

DEMEREC, M., and Z. E. DEMEREC. 1956. Analysis of linkage relationships in *Salmonella* by transduction techniques. Brookhaven Symp. Biol. **8**:75–84.

DEONIER, R. C. 1996. Native insertion sequence elements: locations, distributions, and sequence relationships. Pp. 2000–2011 *in* F. C. NEIDHARDT, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. B. LOW, B. MAGASANIK, W. S. REZNIKOFF, M. RILEY, M. SCHAECHTER, and H. E. UMBARGER, eds. *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.

EVANS, D., D. ZORIO, M. MACMORRIS, C. E. WINTER, K. LEA, and T. BLUMENTHAL. 1997. Operons and SL2 trans-splicing exist in nematodes outside the genus *Caenorhabditis.* Proc. Natl. Acad. Sci. USA **94**:9751–9756.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (40 co-authors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**:496–512.

FRASER, C. M., J. D. GOCAYNE, O. WHITE et al. (29 co-authors). 1995. The minimal gene complement of *Mycoplasma genitalium.* Science **270**:397–403.

FUJITA, N., H. MORI, T. YURA, and A. ISHIHAMA. 1994. Systematic sequencing of the *Escherichia coli* genome: analysis of the 2.4–4.1 min (110,917–193,643 bp) region. Nucleic Acids Res. **22**:1637–1639.

GOFFEAU, A., R. AERT, M. L. AGOSTINI-CARBONE et al. (633 co-authors). 1997. The yeast genome directory. Nature **387**(Suppl.):1–105.

HEROLD, M., and K. H. NIERHAUS. 1987. Incorporation of six additional proteins to complete the assembly map of the 50 S subunit from *Escherichia coli* ribosomes. J. Biol. Chem. **262**:8826–8833.

HIMMELREICH, R., H. HILBERT, H. PLAGENS, E. PIRKL, B.-C. LI, and R. HERRMANN. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res. **24**:4420–4449.

HIMMELREICH, R., H. PLAGENS, H. HILBERT, B. REINER, and R. HERRMANN. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium.* Nucleic Acids Res. **25**:701–712.

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022

ITAYA, M. 1993. Stability and asymmetric replication of the *Bacillus subtilis* 168 chromosome structure. J. Bacteriol. **175**:741–749.

ITOH, T., H. AIBA, T. BABA et al. (31 co-authors). 1996. A 460-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 40.1–50.0 min region on the linkage map. DNA Res. **3**:379–392.

KANEKO, T., S. SATO, H. KOTANI et al. (24 co-authors). 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3**:109–136.

KASAHARA, Y., S. NAKAI, and N. OGASAWARA. 1997. Sequence analysis of the 36-kb region between *gntZ* and *trnY* genes of *Bacillus subtilis* genome. DNA Res. **4**:155–159.

KLENK, H.-P., R. A. CLAYTON, J.-F. TOMB et al. (51 co-authors). 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature **390**:364–370.

KUNST, F., N. OGASAWARA, I. MOSZER et al. (151 co-authors). 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature **390**:249–256.

LANGER, D., J. HAIN, P. THURIAUX, and W. ZILLIG. 1995. Transcription in archaea: similarity to that in eucarya. Proc. Natl. Acad. Sci. USA **92**:5768–5772.

LEWIN, B. 1997. GENES VI. Oxford University Press, Oxford.

LIPMAN, D. J., and W. R. PEARSON. 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.

MUSHEGIAN, A. R., and E. V. KOONIN. 1996. Gene order is not conserved in bacterial evolution. Trends Genet. **12**:289–290.

NAAS, T., M. BLOT, W. M. FITCH, and W. ARBER. 1995. Dynamics of IS-related genetic rearrangements in resting *Escherichia coli* K-12. Mol. Biol. Evol. **12**:198–207.

OSHIMA, T., H. AIBA, T. BABA et al. (32 co-authors). 1996. A 718-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 12.7–28.0 min region on the linkage map. DNA Res. **3**:137–155.

PEARSON, W. R., and D. J. LIPMAN. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**:2444–2448.

ROTH, J. R., N. BENSON, T. GALITSKI, K. HAACK, J. G. LAWRENCE, and L. MIESEL. 1996. Rearrangements of the bacterial chromosome: formation and applications. Pp. 2256–2276 *in* F. C. NEIDHARDT, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. B. LOW, B. MAGASANIK, W. S. REZNIKOFF, M. RILEY, M. SCHAECHTER, and H. E. UMBARGER, eds. *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SIEFERT, J. L., K. A. MARTIN, F. ABDI, W. R. WIDGER, and G. E. FOX. 1997. Conserved gene clusters in bacterial genome provide further support for the primacy of RNA. J. Mol. Evol. **45**:467–472.

SMITH, D. R., L. A. DOUCETTE-STAMM, C. DELOUGHERY et al. (37 co-authors). 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. **179**:7135–7155.

SPIETH, J., G. BROOKE, S. KUERSTEN, K. LEA, and T. BLUMENTHAL. 1993. Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell **73**:521–532.

ST. JOHN, T. P., and R. W. DAVIS. 1981. The organization and transcription of the galactose gene cluster of *Saccharomyces.* J. Mol. Biol. **152**:285–315.

TATUSOV, R. L., A. R. MUSHEGIAN, P. BORK, N. P. BROWN, W. S. HAYSE, M. BORODOVSKY, K. E. RUDD, and E. V. KOONIN. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli.* Curr. Biol. **6**:279–291.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

TODA, T., T. TANAKA, and M. ITAYA. 1996. A method to invert DNA segments of the *Bacillus subtilis* 168 genome by recombination between two homologous sequences. Biosci. Biotechnol. Biochem. **60**:773–778.

TOMB, J.-F., O. WHITE, A. R. KERLAVAGE et al. (42 co-authors). 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature **388**:539–547.

WATANABE, H., H. MORI, T. ITOH, and T. GOJOBORI. 1997. Genome plasticity as a paradigm of eubacterial evolution. J. Mol. Evol. **44**(Suppl. 1):S57–S64.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

YAMAMOTO, Y., H. AIBA, T. BABA et al. (31 co-authors). 1997. Construction of a contiguous 874-kb sequence of the *Escherichia coli* K-12 genome corresponding to 50.0–68.8 min on the linkage map and analysis of its sequence features. DNA Res. **4**:91–113.

YURA, T., H. MORI, H. NAGAI, T. NAGATA, A. ISHIHAMA, N. FUJITA, K. ISONO, K. MIZOBUCHI, and A. NAKATA. 1992. Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region. Nucleic Acids Res. **20**:3305–3308.

ZORIO, D. A. R., N. N. CHENG, T. BLUMENTHAL, and J. SPIETH. 1994. Operons as a common form of chromosomal organization in *C. elegans.* Nature **372**:270–272.

MASAMI HASEGAWA, reviewing editor

Accepted October 26, 1998

Downloaded from https://academic.oup.com/mbe/article/16/3/332/2925393 by U.S. Department of Justice user on 16 August 2022