J Mol Evol (1997) 44(Suppl 1):S57–S64

JOURNAL OF MOLECULAR EVOLUTION

Genome Plasticity as a Paradigm of Eubacteria Evolution

Hidemi Watanabe,[1,2] Hirotada Mori,[3] Takeshi Itoh,[3] Takashi Gojobori[1,*]

[1]National Institute of Genetics, Mishima 411, Japan
[2]New Energy and Industrial Technology Development Organization (NEDO), Tokyo 170, Japan
[3]Nara Institute of Science and Technology, Ikoma 630-01, Japan

Received: 8 July 1996 / Accepted: 20 August 1996

Abstract. To test the hypotheses that eubacterial genomes leave evolutionarily stable structures and that the variety of genome size is brought about through genome doubling during evolution, the genome structures of *Haemophilus influenzae, Mycoplasma genitalium, Escherichia coli,* and *Bacillus subtilis* were compared using the DNA sequences of the entire genome or substantial portions of genome. In these comparisons, the locations of orthologous genes were examined among different genomes. Using orthologous genes for the comparisons guaranteed that differences revealed in physical location would reflect changes in genome structure after speciation. We found that dynamic rearrangements have so frequently occurred in eubacterial genomes as to break operon structures during evolution, even after the relatively recent divergence between *E. coli* and *H. influenzae*. Interestingly, in such eubacterial genomes of high plasticity, we could find several highly conservative regions with the longest conserved region comprising the S10, *spc*, and α operons. This suggests that such exceptional conservative regions have undergone strong structural constraints during evolution.

Key words: Genome evolution — Eubacteria — Rearrangement — S10 region

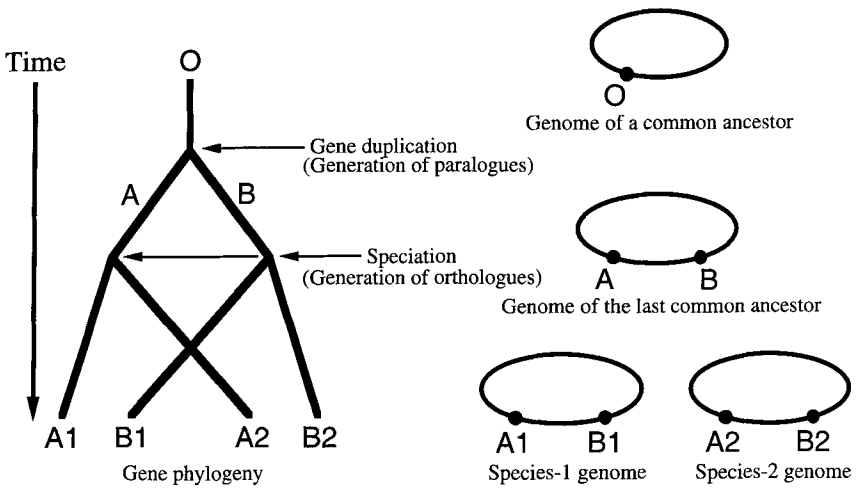
Introduction

The way genomes of different organisms are organized during species evolution has been an issue of great interest during the past few decades. As for eubacterial genomes, there have been various speculations about structural dynamics within individual genomes in terms of genome size and gene order and the transfer of genetic materials between species. Ochman and Wilson (1987) suggested that severely restricted recombination is the cause of temporal and spatial stability in a large number of independently evolving lineages in natural populations of *Escherichia coli* (*Ec*). On the other hand, Naas et al. (1994) showed that an amazing degree of genetic plasticity can be observed in the chromosome of the strain W3110 of *Ec* K-12 cultured in agar stabs, indicating that DNA rearrangements such as transpositions and inversions serve as the causes of a wide variety of genomic sequence variations (Arber 1995). Moreover, Herdman (1985) noted that the size variations in eubacterial genomes have been mainly brought about through genome doublings, whereas Labedan and Riley (1995) noted that full genome doubling in the evolution of the *Ec* genome is not supported by the test of regularity during physical intervals of evolutionarily related genes in the genome.

Recently, the complete genome sequences for *Haemophilus influenzae* Rd (*Hi*) and *Mycoplasma genitalium* (*Mg*) were disclosed (Fleischmann et al. 1995; Fraser et al. 1995). In addition to these eubacteria, partial, but fairly long sequence data is available for the genomes of *Ec* and *Bacillus subtilis* (*Bs*). These genome sequences

Abbreviations: Bs, Bacillus subtilis; Ec, Escherichia coli; Hi, Haemophilus influenzae; Mg, Mycoplasma genitalium; ORF, open-reading frame

* *Present address:* Center for Information Biology, National Institute of Genetics, Mishima 411, Japan; e-mail tgojobor@genes.nig.ac.jp
Correspondence to: T. Gojobori



Homologues: A1, B1, A2 and B2
Orthologous: A1 vs A2 and B1 vs B2
Paralogous: the other combinations

Fig. 1. Schematic representation of the difference between orthologue and paralogue.

are very useful for testing the contradictory genome evolution hypotheses.

To test those hypotheses, it is important to locate the orthologous genes in the genomes. This is because we are interested in tracing evolutionary changes in genome structures during species divergence. Thus, we first identified orthologous gene sets among *Hi, Mg, Ec,* and *Bs* and then compared the gene order between the genomes of these species by examining whether the orthologues of the genes comprising a region in a genome also compose a region in the same physical order in another genome.

Materials and Methods

Homologues can be classified into paralogues and orthologues (Fig. 1). Whether a pair of homologues are paralogous or orthologous depends on the phylogenetic relationship between the species carrying the homologues. Paralogues are defined as the genes generated in the same genome by gene duplication. On the other hand, orthologues are generated by speciation from the same gene in the last common ancestor of the species. As shown in Fig. 1, orthologues (A1 and A2 or B1 and B2) are most similar to each other among all the possible combinations of homologous genes between different species. In this context, for the identification of orthologous gene pairs between different species, the complete gene sets of the species are needed.

In this analysis, we used the complete gene sets of the *Hi* and *Mg* genomes, which consist of 1,727 and 468 genes or ORFs, respectively. As for *Ec* and *Bs,* 1,677 and 659 genes are available, respectively. These gene sets are not complete but provide much information. Orthologous gene pairs were identified between these species according to the following four criteria: (1) Each of an orthologous pair of genes is more similar to the other than to any other genes it can be paired with; (2) each orthologous gene pair shows similarity of statistical significance; (3) all the possible pairs of the genes of each orthologous gene group, which is constructed through linkage of orthologous gene pairing, should satisfy the previous two criteria; (4) gene phylogeny based on similarities of genes should be consistent with the phylogenetic relationships among the four species (Hori and Osawa 1987). In practice, if a pair of genes between *Hi* and *Ec* are in an orthologous gene group, the pair should be the most similar among all the pairs in the group. Similarity was computed with the FASTA program (version 2.0) (Pearson 1990) after each gene was translated into an amino acid sequence, and the statistical significance of each pair was examined using the PRDF program (Pearson 1990) to determine whether the statistical significance of each FASTA score was higher than 6.0 SD with 200 shuffles.

Results and Discussion

Locations of Orthologues

The correspondence of orthologous gene locations between different genomes is shown in Fig. 2. In this representation, we could find many adjoining parallel lines when gene order tended to be conserved in different genomes. However, it is obvious that parallel lines indicative of gene order conservation are very rare. This indicates that the genes in the genomes compared have been reshuffled so frequently as to appear to be randomly ordered. It is rather surprising that this situation is also found even in the comparison between the genomes of *Hi* and *Ec* (Fig. 2B), because these species are more closely related than between *Hi* and *Mg* at the primary sequence level (Fig. 3). Now, one may question whether those genomes have been reshuffled to the level of randomness. To answer such questions, we have to evaluate the difference level in genome structures quantitatively.

Quantitative Evaluation of Difference Level in Genome Structures

To evaluate the difference level of genome structures quantitatively, we invented two indices, *M* and *D,* for the evaluation of the conservation of gene order and the difference in gene number. These indices are defined with three types of parameters for the genomes com-

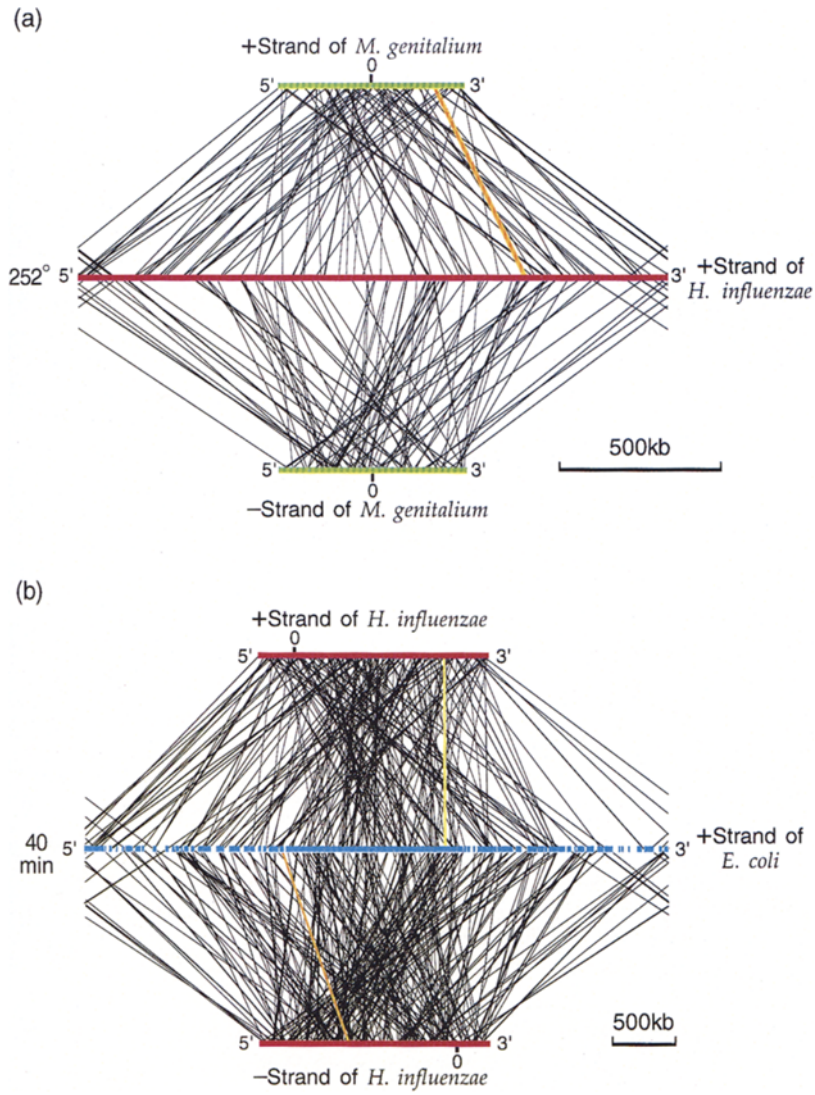


Fig. 2. Comparisons of physical positions of orthologous genes between two genomes. **a** *M. genitalium* (*green*) vs *H. influenzae* (*red*). **b** *H. influenzae* vs *E. coli* (*blue*). Orthologous genes are connected by *black lines*. The *orange and yellow lines* represent the correspondence between long conserved regions (see text). 0's indicate the origins of sequence coordinates specified in Fleischmann et al. (1995) and Fraser et al. (1995). The *blank portions* along the genomes represent regions in which physical map data and nucleotide sequences of genes are not available.

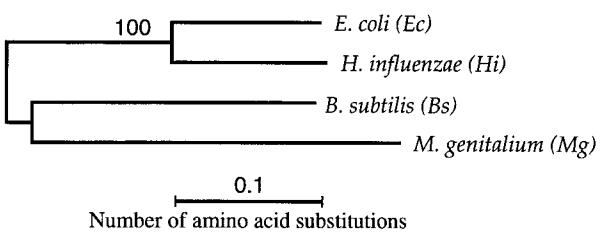


Fig. 3. Gene phylogeny constructed using the *dnaK* orthologous gene group. The phylogenetic tree was constructed by the neighbor-joining method (Saitou and Nei 1987). The *number* at the branching point indicates the percentage of bootstraps in which the cluster to the right was found.

pared—the numbers of constituent genes of genome i (N_i), the numbers of orthologous gene pairs ($N_{O(i,j)}$), and the numbers of contiguously conserved regions ($N_{C(i,j)}$) identified between genomes i and j. These parameters used for comparisons among the four genomes are shown in Table 1. In this study, ''conserved region'' means a region that is composed of one or more genes in a genome whose orthologues are located in the same order in a region of one other genome. According to this definition, a conserved region can be composed of only one gene.

M_{ij} is defined as the mean number of the orthologous genes per conserved region which are identified between genomes i and j, i.e.,

$$M_{ij} \equiv N_{O(i,j)}/N_{C(i,j)} \quad (1)$$

This index ranges from 1 to $N_{O(i,j)}$ in accordance with the conservation level in gene order. Generally speaking, if we compare two circular genomes one of which was derived by random shuffling of the other one, the expected value of M becomes $N_O/(N_O - 1)$. If identical genomes are compared, the M value becomes N_O or the number of their constituent genes. The M values calculated among the four species are shown in Table 2. The M values are all less than 2, meaning that the average length of conserved regions is only one or two genes. However, the values are not so low as $N_O/(N_O - 1)$, indicating that these genomes have not been completely shuffled yet. It is worth while noting that the M values

Table 1. Parameters for the genomes of *E. coli*, *H. influenzae*, *B. subtilis*, and *M. genitalium*[a]

	E. coli	*H. influenzae*	*B. subtilis*	*M. genitalium*
E. coli	1,677	384	141	97
H. influenzae	607	1,727	163	144
B. subtilis	202	215	659	56
M. genitalium	132	184	98	468

[a] Number of genes that have been mapped on each genome is shown on the diagonal. N_O is in the lower left, and N_C the upper right

Table 2. *M* and *D* values obtained in comparisons among genomes of different eubacteria[a]

	E. coli	*H. influenzae*	*B. subtilis*	*M. genitalium*
E. coli		0.783	0.905	0.934
H. influenzae	1.58		0.901	0.909
B. subtilis	1.43	1.32		0.905
M. genitalium	1.36	1.28	1.75	

[a] *M* is shown in the lower left, and *D* is shown in the upper right

between *Hi* and *Ec* and between *Mg* and *Bs* are relatively higher than those between the other pairs. This may suggest that divergence in gene order correlates with species divergence. Since *M* is defined using only orthologous genes, this index may be independent of species-specific genes and difference in genome size.

The other index *D* is a measure of evolutionary changes in the number of constituent genes of genomes (Fig. 4). *D* is defined between genomes *i* and *j* as

$$D_{ij} \equiv (N_i + N_j - 2N_{O(i,j)})/(N_i + N_j - N_{O(i,j)}) \quad (2)$$

If we assume that a virtual common ancestor of the two genomes consists of $(N_1 + N_2 - N_O)$ genes, the numbers of genes deleted during the evolution of genomes 1 and 2 are $\{(N_1 + N_2 - N_O) - N_1\}$ and $\{(N_1 + N_2 - N_O) - N_2\}$, respectively (Fig. 4, Shrinking model). Thus, the total number of genes deleted becomes $(N_1 + N_2 - 2N_O)$. Therefore, *D* represents the fraction of genes deleted during genome evolution from the total number of constituent genes in the common ancestor. On the other hand, if it is assumed that the common ancestor has a genome of only N_O genes as shown in Fig. 4 as the "Growing model," formula (2) may also be interpreted as the fraction of genes generated from the common ancestor during evolution.

Whichever interpretation of *D* is adopted, *D* has the following properties: *D* becomes 1 if the genomes compared are composed of completely different types of genes, whereas it becomes 0 if the sets of constituent genes are identical. This index is not dependent on gene order. However, we have to note that it depends on the ratio of sequenced regions when incomplete genome sequences are compared. The values of this index were calculated for the four species (Table 2). It is shown in Table 2 that the *D* value between *Hi* and *Ec* is the lowest, indicating that the genomes of these two species are composed of relatively similar sets of genes. However, the *D* values between other species pairs are almost equal, implying that *D*, i.e., gene number, is rather independent of phylogenetic divergence.

These mathematical approaches to divergence in genome structures after speciation will be useful in revealing difference and conservation between genome structures quantitatively.

Characteristic Conservations

We now know that genomes of eubacteria have been shuffled through genome rearrangements and the frequency of rearrangement seems to correlate with phylogenetic distances. Although the extensive rearrangement of genome structures makes it very difficult to trace genome evolution of those four species precisely, we found some conservation patterns in their genomes. The orthologous gene clusters conserved between different genomes are listed in Table 3. As expected from the low *M* values calculated, most of the conserved regions are "size 1 clusters" (data not shown) or "size 2 clusters," in which there are only one or two genes. This is consistent with the notion that eubacterial genomes have been reshuffled frequently. This implies that relative gene position in a genome is not substantially essential to gene function. Despite this general feature in genome evolution, however, we can see some exceptional characteristic conservations in Table 3.

The longest regions identified between the four species are all the same region—namely, the "S10 region." In the case of *Ec*, this region is comprised of three operons, the S10, *spc*, and α operons. Collectively, they encode 26 ribosomal proteins, the RNA polymerase α subunit, and the preprotein translocase SecY subunit. Conservation in this region is depicted in Fig. 5A. There

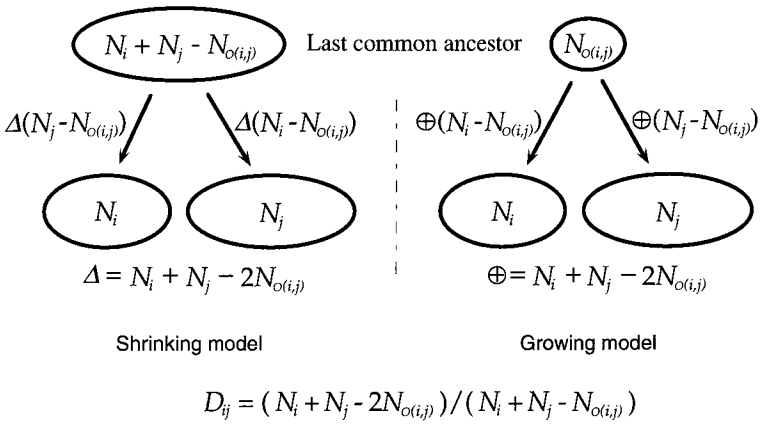


Fig. 4. Meanings of the *D* value. Both genome-shrinking and genome-growing models are considered.

are several characteristic features in the conservation in this region: All the genes in this region are in the same orientation; most of the genes in this region are related to translation; the 5′ halves are completely conserved; the 3′ halves contain divergences strongly correlating to the phylogeny of these species. The phylogeny dependencies found in the 3′ halves are as follows: The portion containing *adk, map,* and *infA* genes is located outside this region in gram-negatives while this portion exists in this region in gram-positives; the case of the *S4* gene is the converse of the previous case; the *L30* gene is absent only in *Mg,* implying that this gene has been depleted after speciation between *Bs* and *Mg*. The strong conservation of the S10 region indicates that there is strong functional and/or structural cohesion among the genes in this region.

The previous study (Lindahl et al. 1990) demonstrated the transcriptional organization of the S10, *spc* and α operons as follows: About 25% of the polymerases transcribing the end of the S10 operon continued transcription into the *spc* operon; most or all RNA polymerases transcribing the *spc* operon continued into the α operon; and about 30% of the transcripts of the α operon were initiated at the α operon promoter. The physical order of these operons must have been conserved to generate such a variety of transcriptional combinations of these operons. However, the extreme conservation of the gene order within these operons may not be explained in this way. Thus, it appears that there are other strong structural constraints on this region that have not been well characterized yet. We speculate that the gene order correlates to the assembly of ribosome subunits as well as the expression level of the genes.

Interestingly, a very similar situation can be also found in the region mainly comprised of genes participating in cell division and cell wall synthesis (Fig. 5B). All the genes in this region are in the same orientation, most of the genes are functionally related, several transcription initiation sites have been identified experimentally within this region in *Ec,* and there is some correlation in the conservation to the phylogeny. Most of the genes in this region are absent in *Mg,* which is a cell-wall–free organism. In addition to the features shared with the S10 region, there is a characteristic feature specific to this region in which paralogous genes exist, e.g., *murE, murF, murD,* and *murC*. The products of these four genes catalyze the sequential four chemical reactions converting UDP-*N*-acetylmuramic acid (UDP-MurNAc) to UDP-MurNAc-L-Ala-D-Glu-meso-diaminopimelic acid. The tendency for gene duplication is also found in a species-specific portion; for example, the *Bs*-specific portion that is composed of two genes homologous to *ftsI* of *Ec*. Thus, such gene duplications with no destruction of this region might be advantageous to the survival of eubacteria except for *Mg*. In addition, it may be of interest to note that the two hypothetical genes in the 5′ end of this region are completely conserved among the four species. The functions of these two genes have not been identified yet, but this conservation strongly suggests that these genes are important and may be functionally related to cell division.

The characteristic conservation in those two regions implies that they are conserved at least over various species of Bacteria. Thus, in a common ancestor of prokaryotes, the structures of these regions and the mechanisms for the regulation of the genes in these regions might have been completed, and the mechanisms are so complete that no species could not have invented any other mechanisms better than the existing one.

Although experimental approaches to the functions of the genes in these regions seem to have been carried out mainly by investigating the physiological properties of these individual genes, it would be also important to reveal the biological reason for the conservation of the structures in these regions.

Divergence in Genome Size

Alteration in genome size was addressed by counting the number of genes duplicated after speciation. If many genes were duplicated through genome doublings after speciation, many genes in a genome would show higher similarities to some genes within the same genome than to their orthologues. Thus we examined *Ec* and *Hi* be-

Table 3. Orthologous gene clusters conserved among different genomes[a]

Cluster size¶	Gene names	No. of clusters
(A) *E. coli—H. influenzae*		
2	*lepAB, dksAfolK*, thrAB*, yffBdapE*, hflKC*, yajGbolA, aroBdam*, guaBA*, bcpdapA*, rfaQI*$^{(1,0)}$†, *btuB*$^{(1,1)}$ *murB, fadRnhaB*, yjaDG*, ybaBrecR, rpiAserA*, hslVU*, menAG, deoDyjjl*†, aspAmopA, ksgAapaH*, asnCasnA, oxyRyijC*, gpsAcysE*, fucPfucA*†, deffmt, ycfCpurB*, kdtAB*, glpBC*, ytfMO*, yigZtrkH, hibPQ*, galTK*, slttrpR, vacByjfH*, purHD*, emrAB*, secAmutT, rpsBtsf, cdsAlpxD*, glyQS, ybaDribG, yaaCrpsT*, menBC*, accBpanF, yhdGfis, lspAlytB, yiaRS*$^{(1,0)}$**, cydAB*, ligcysK, yhbGptsN*, yjgApmbA, trxBcydC, sucCD*, mdhargR*†, cmk*$^{(0,1)}$*himD, pyrFyciH, aptdnaX*, aceElpdA*, dnaKJ, accDfolC*, yhbCA*, ribHnusB, infCrplT*, pntAB, pstCA, purMN*, trpBA*, xseBispA, moeAB*, sppAydjA*, bioF*$^{(1,0)}$*C*, ilvIH*, purEK*, sucAB, basRS, yjeQyjeR*	76
3	*holDrimIyjjT*, mreBCD*, yjeEmutLmiaA, focApflBA, rbsDCK, priBrpsRrplI*, hemXYcyaA, fucIKU*, tigclpPX, mglBAC*, yciBAI*, mobyihDdsbA, leuABD*, rnpAyidDC, tbpAyabKJ*, dmsABC*, pheSThimA*, hemMychBprsA*, yjgPQpepA*, gmkrpoZspoT**	20
4	*fdhDfdoH*$^{(1,0)}$*IfdhE*, yceDfabHDacpP*, queAyajCsecDF*, ygjDrpsUdnaGrpoD, yhhFftsYEX*, radCrpmBfpgdut, yiaMNOxylK*, fabZlpxABrnhB*, nrfABCD*, potABCD*, glgBCAP*, rplMrpsIsspAB*	12
5	*aspSyebDCruvCB, ybgCtolQRBpal*, frdABCDyjeA*	3
6	*xylFGHRxylAB**	1
7	*hisGDBHAFI, gidBatpBFAGDC*	2
8	*ybeBAmrdABrlpAybeDlipBA*	1
14	*yabBCmurEFmraYmurDftsWmurGCddlBftsQAZenvA*	1
23	*rpsJrplCDWBrpsSrplVrpsCrplPrpsQrplNErpsHrplFRrpsErplOprlArpsMKDrpoArplQ**	1
(B) *E. coli—E. subtilis*		
2	*rpsMK*, rpoArplQ*, purEK*, purMN, purHD*, treB*$^{(1,0)}$*C*, argCB, proBA, motAB*, ftsAZ, carA*$^{(0,1)}$*B, hslVU*, cheAW*, sucAB*, aroAcmk, ribGH, phoB*$^{(0,1)}$*R*$^{(0,1)}$*, dnaKJ, mreB*$^{(0,2)}$*C*, ilvCtrxA, frdA*$^{(1,0)}$*B*$^{(1,0)}$**, pheST*, fliCfliS*†, gidAB, dnaXrecR, prsApth*, rplAL*	27
3	*leuBilvIH*†, cyoABC*$^{(0,1)}$	2
4	*trpE*$^{(0,1)}$*CBA*, glgBCAP**	2
5	*murEmraYmurDftsWmurG, atpBAGDC**	2
17	*rpsJrplCDBrpsSrplVrpsCrplPrpsQrplNErpsHrplFRrpsErplOprlA**	1
(C) *E. coli—M. genitalium*		
2	*rpsRrplI, rpsMK*, rpoArplQ*, ugpA*$^{(1,0)}$*E*, pheST*, infCrplT*, tiglon, dnaGrpoD*, gidAB*$^{(0,1)}$***, pstA$^{(0,1)}$*B, rplM*$^{(0,1)}$*rpsl*	11
3	*potA*$^{(7,0)}$*B*$^{(1,0)}$*C*$^{(1,0)}$*, yabBC*$^{(0,1)}$*ftsZ, atpAGD, tsfpyrHfrr*$^{(0,1)}$	4
17	*rpsJrplCDBrpsSrplVrpsCrplPrpsQrplNErpsHrplFRrpsErplOprlA**	1
(D) *H. influenzae—B. subtilis*		
2	*rpS13S11, rpoArplQ, purEK, purMN*, purHD, ftsAZ, hslVU, nusAinfB*, sucAB*, trpBA, pilBxcpS*, dnaKJ, mreB*$^{(0,2)}$*C, rpL21L27*, ilvIH*, frdAB*, pheST, pyrGmurZ†, rpL11L1*, rpoBC**	20
4	*glgBCAP*	1
5	*murEmraYmurDftsWmurG, atpBAGDC*, rpL34HI1001thdFdnaArecF**	3
18	*rpS10L3L4L2S19L22S3L16S17L14L5S8L6L18S5L30L15secY*	1
(E) *H. influenzae—M. genitalium*		
2	*rpL11L1*, rpS18L9*, nusAinfB, rpS13S11, rpoArplQ, rpL21L27*, dnaGrpoD*, prfAhemK*$^{(1,0)}$**, rpoBC, rpL13*$^{(0,1)}$*S9, rpL34dnaA*†*	11
3	*potA*$^{(2,0)}$*BC*, strArpS7fusA**, HI1129HI1130$^{(0,1)}$*ftsZ, atpAGD, rpS16trmDrpL19*	5
4	*pheSTinfCrpL20*	1
17	*rpS10L3L4L2S19L22S3L16S17L14L5S8L6L18S5L15secY*	1
(F) *B. subtilis—M. genitalium*		
2	*rplKA, nusAinfB*, pheST, rpoBC*, gidAB*$^{(0,1)}$**, hprTftsH*, rpmHdnaA*	7
3	*gyrBAserS, rpsRrplIdnaC*$^{(0,1)}$**, rplUrpmAlon, pdhA*$^{(1,0)}$*B*$^{(1,0)}$*D*, atpAGD**	5
26	*rpsJrplCDWBrpsSrplVrpsCrplPrpsQrplNErpsHrplFRrpsErplOsecYadk map infArpmJrpsM KrpoArplQ*	1

[a] Key to symbols: ¶Indicating number of constituent orthologous genes of a conserved region; *Indicating that corresponding orthologous gene cluster is in a different strand; †Indicating that strand configuration in the cluster is not conserved; $^{(l,r)}$Indicating that the gene (its orthologous gene) shows greater similarity to $l(r)$ genes within the same genome than to its orthologous gene; $^{(0,0)}$is omitted

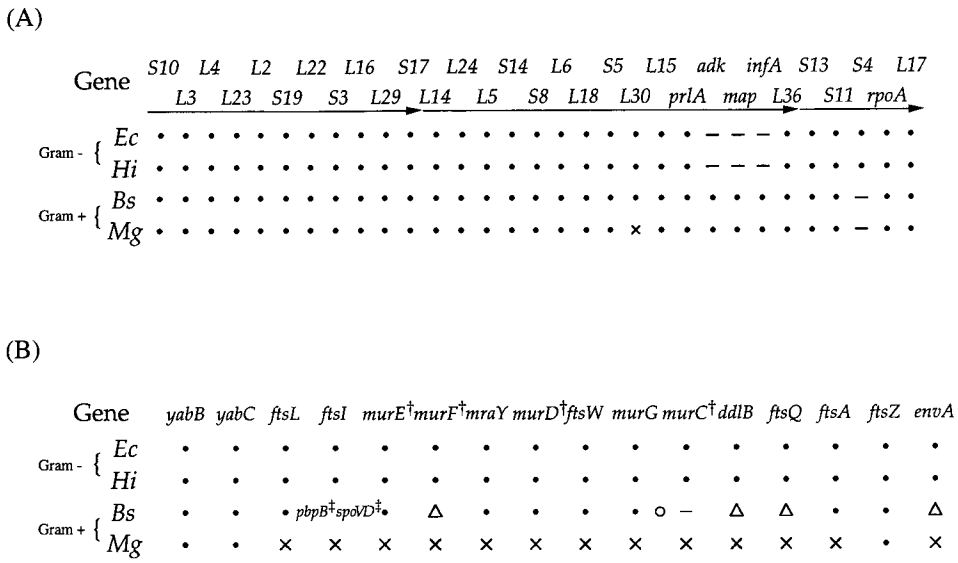


Fig. 5. Extensively conserved regions among *Ec, Hi, Bs,* and *Mg*. **A** S10 region. The *lateral three arrows* shown below gene names represent operons known in *Ec*. "•" indicates the existence of the gene at the site; "–" indicates that the gene is translocated elsewhere; "x" indicates that the gene is absent; *Gram–* and *Gram+* represent gram-negatives and gram-positives, respectively. **B** The region comprised of genes for cell-wall synthesis and cell division. †: paralogous genes; ‡: paralogues in *Bs* which are homologous to the *ftsI* gene in *Ec; open circle: Bs*-specific cluster of genes, *murB, divIB, ylxW, ylxX,* and *sbp;* a *triangle* indicates that the gene is absent in this region and has not been found in the other region.

cause they are relatively closely related, but their genomes are very different in size. Only 20 such cases were found in the *Ec* genome and only 10 cases in *Hi*, while 607 orthologous gene pairs were identified between these species, implying that only a low percentage of the *Ec* genome was generated after speciation. This strongly suggests that the main cause of the size difference between the genomes of these species is genome shrinking in *Hi* rather than genome doubling in *Ec* after speciation. This result alone cannot exclude the possibility that some genes, whose orthologues were deleted in *Hi,* have been duplicated so many times as to lead to growth in genome size in *Ec*. However, this hypothesis can be disregarded because there are not a sufficient number of genes whose orthologues were not found in *Hi* but for whom homologues were found in *Ec*.

Biological Significance of Genome Plasticity

The genome size divergence may have been brought about by genome rearrangement mechanisms such as homologous recombination. As for parasitic organisms existing in homeostatic environments provided by hosts, genome shrinking directly conduces to rapid replication and an important factor in their survival is rapid replication, rather than adaptability to many environments. Thus, genome shrinking may be advantageous to the survival of parasitic organisms such as *Hi* and *Mg* if the functions of deleted genes are complemented by the genes of hosts.

Recently, it was shown in *Ec* that transcriptional regulatory genes as well as transport genes have been extensively duplicated in comparison to the other genes (Watanabe and Otsuka 1995) and that the regulatory relationships between transcriptional regulators and their regulating genes have been generated by random combinations during evolution (Otsuka et al. 1996). Taking into account the instability of operon structures shown in this study, we would point out the possibility that transcriptional regulatory systems have been generated dynamically during eubacteria evolution. Thus, it is likely that genome plasticity has contributed to the dynamic generation of transcriptional regulatory systems.

The results in this study showing the evolutionary plasticity of genome structures are supported by the results provided by Tatusov et al. (1996). Our approach to the evolution of genome structures in eubacteria was brought about by identifying orthologous gene pairs among four different species. In this approach, it is assumed that sequence similarity between the genes determined to be orthologous correlates to species phylogeny. However, if lateral gene transfer between distantly related species has occurred frequently, this assumption becomes far from realistic. For example, if a gene in *Salmonella typhimurium,* which is much more closely related to *Ec* than *Hi,* was transferred to a locus in the genome of *Hi,* the gene transferred will be determined to be an orthologue of an *Ec* gene, although the real orthologue of the *Ec* gene may reside in another locus. Therefore, our approach must be carried out carefully if the frequency of lateral gene transfer is not sufficiently estimated.

Many genome projects have been started, and a large number of genome sequences have been exhaustively

determined. Sequence mass production was expected to provide us with unique opportunities to obtain much information relating to the structures and functions of many genes. So far, it has become common to infer the functions and structures of newly sequenced genes or their products by examining the similarities to "known" genes because of the effectiveness of this approach in identifying the features of new genes. However, recent studies on gene classification show that a large part of the open reading frames predicted in a genome have no counterpart of similarity in the pool of known genes. For example, it was estimated that 23% of the *Ec* genes show no similarity to any known genes while 39% of these genes are similar to some unknown genes (Koonin et al. 1995). This indicates that there is a limit to the effectiveness of the conventional approach in the case of its application to the genes obtained through comprehensive genome analyses and that new genes without similarities to known genes will become the majority as genome projects progress. Therefore, an approach such as the one we used in this study to identify sets of genes evolutionarily conserved in relative positions will be vital. In addition, it may be expected that unknown functions will be found on gene clusters evolutionarily conserved, and it might be concluded that, if such gene clusters are found in a wide spectrum of species, they were important parts of the common ancestral genome among the species.

Note Added in Proof

By applying the present method to the genomes of *Synechocystis* sp. strain PCC6803 (D63999-D64006, D90899-80917) and *Methanococcus jannaschii* (L77117) in addition to the four species used in this study, we could also find conservation of the S10 region in both of them, although the region corresponding to eubacterial S10 regions in the *M. jannaschii* genome has a few characteristic features (e.g., translocation of gene clusters and insertion of eukaryotic ribosomal protein genes). Thus, it seems to be certain that a prototype of the S10 region existed in the last common ancestor among Bacteria, Archaea, and probably Eucarya.

Acknowledgments. The authors would like to thank Dr. Hirohumi Yoshikawa of the University of Tokyo for helpful discussions. This study was supported, in part, by a research grant from the Ministry of Education, Science, Sports and Culture, and by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

References

Arber W (1995) The generation of variation in bacterial genomes. J Mol Evol 40:7–12

Fleischmann RD et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

Herdman M (1985) The evolution of bacterial genomes. In: Cavalier-Smith T (ed) The evolution of genome size. John Wiley, London, p 37

Hori H, Osawa S (1987) Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol Biol Evol 4:445–472

Koonin EV, Tatusov RL, Rudd KE (1995) Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. Proc Natl Acad Sci USA 92:11921–11925

Labedan B, Riley M (1995) Widespread protein sequence similarities: origins of *Escherichia coli* genes. J Bacteriol 177:1585–1588

Lindahl L, Sor F, Archer RH, Nomura M, Zengel JM (1990) Transcriptional organization of the S10, *spc* and α operons of *Escherichia coli*. Biochim Biophys Acta 1050:337–342

Naas T, Blot M, Fitch WM, Arber W (1994) Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. Genetics 136: 721–730

Ochman H, Wilson AC (1987) Evolutionary history of enteric bacteria. In: Neidhardt FC et al. (eds) *Escherichia coli* and *Salmonella typhimurium*. Cellular and molecular biology. American Society for Microbiology, Washington, DC, p 1034

Otsuka J, Watanabe H, Mori KT (1996) Evolution of transcriptional regulation system through promiscuous coupling of regulatory proteins with operons; suggestion from protein sequence similarities in *Escherichia coli*. J Theor Biol 178:183–204

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr Biol 6:279–291

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Watanabe H, Otsuka J (1995) A comprehensive representation of extensive similarity linkage between large numbers of proteins. Comput Appli Biosci 11:159–166