Comprehensive comparison between locations of orthologous genes on archaeal and bacterial genomes

Katsuhisa Horimoto1,∗, Satoshi Fukuchi2 and Kentaro Mori3

1Laboratory of Mathematics, Saga Medical School, 5-1-1 Nabeshima, Saga 849-8501, Japan, 2Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan and 3Japan Network Information Center, 1-2 Kanda-Ogawa, Chiyoda, Tokyo 101-0052, Japan

Received on December 1, 2000; revised on April 4, 2001; accepted on May 7, 2001

ABSTRACT

Motivation: Following an extensive search for orthologous genes between the complete genomes from archaea and bacteria, the spatial association of the orthologs has been investigated in terms of synteny, the conservation of the order of neighboring genes. However, the relationships between the relative locations of remote orthologs over entire genomes have not been shown.

Results: Comprehensive comparisons between the locations of orthologs on nineteen archaeal and bacterial genomes are presented by the location to location correspondence based on the gene–location distance. When the two genomes are rotated such that a pair of orthologs with the shortest distance is set in the same angle, a statistically significant number of orthologs maintain their relative locations between the genomes. Even by the short distances at the 5% significance level, the rotations are restricted within a narrow range, suggesting an intrinsic angle for realizing similar locations between the orthologs in each genome pair. Furthermore, the rotations in the restricted range agree with the replication origin and terminus sites for the analyzed genomes where such sites are known. The relationship between location-maintained orthologs and gene function is also discussed.

Contact: horimoto@post.saga-med.ac.jp

INTRODUCTION

Complete sequence data of various archaeal and bacterial genomes have rapidly accumulated. Many possibilities emerged with the determination of the complete genomic structures. One of them is that the gene organization can be studied at a comprehensive level (Doolittle, 1998). In investigations of the spatial association, the orders of the orthologous genes are extensively investigated. Two aspects of the order of orthologs have been described by synteny analyses. Some conservation of the gene orders in a small fraction of all orthologs between genomes has been elucidated in terms of gene function (Siefert et al., 1997; Tamames et al., 1997; Dandekar et al., 1998). On the other hand, drastic rearrangements of gene order in a large fraction have been emphasized in terms of statistics (Mushegian and Koonin, 1996; Kolsto, 1997; Koonin and Galperin, 1997; Watanabe et al., 1997; Casjens, 1998; Huynen and Bork, 1998).

In contrast to the extensive analyses of near or neighboring genes, the relationships between remote genes have only been narrowly investigated. This is partly because the wide variety of gene contents between bacterial and archaeal genomes (Doolittle, 1998) has easily led to the hypothetical conclusion that the relative locations of remote genes may not be maintained, and partly because some difficulties in the methodological aspects have emerged in comparisons between remote genes on circular genomes. As for the methodology, simple procedures have been adopted for synteny analyses of gene locations. Following transformation of the circular data of gene locations into linear data, the colinearity of gene orders and the co-occurrence of neighboring genes between genomes have been evaluated by visual or counting procedures (Mushegian and Koonin, 1996; Huynen and Bork, 1998). In contrast, the direct analysis of circular data requires statistical techniques (Fisher, 1993). In previous reports (Horimoto et al., 1998a,b), a quantitative measure, the gene–location distance, was proposed by the application of a technique (Fisher and Lee, 1983) to investigate the spatial association between remote genes on circular genomes. The performance of the distance for investigating the relationship between the locations of remote genes was validated by its application to forty-three circular mitochondrial genomes.

In this paper, we investigate the relationship between the locations of the orthologs, in nineteen microbial
genomes from archaea and bacteria, by the location to location correspondence based on the gene–location distance. The correspondence revealed that there are many orthologs that are separately located on each genome, but maintain their relative locations between the two genomes. The nature of the correspondence between orthologs is discussed in terms of the correlation with replication sites and the gene function.

**MATERIALS AND METHODS**

**Genomic data**


The definition of an orthologous relationship between the genes in the analyzed genomes is adopted from the Clusters of Orthologous Groups (COGs) of proteins (Tatusov et al., 1997, 2000; Koonin et al., 1998) site (released in October 1999, http://www.ncbi.nlm.nih.gov/COG/). In this version, 2097 groups of orthologous genes between twenty-one genomes are compiled in terms of sequence similarity and phylogenetic relationship. Since orthologs on circular genomes are analyzed in this study, the linear genomes from two organisms, *Saccharomyces cerevisiae* and *Borrelia burgdorferi*, are excluded from the original data in the COGs.

**Procedure for the location to location correspondence**

The procedure for the correspondence between the locations of orthologous genes on two genomes is composed of three steps. At the first step, the gene–location distances (Horimoto et al., 1998a) are calculated for each of the orthologs between the two genomes. At the second step, a pair of orthologs with the shortest gene–location distance is found among all orthologs. Finally, the two genomes are rotated so that the pair with the shortest distance is located in the same angle.

The gene–location distance for the *i*th ortholog pair is defined by the following equation:

\[
D_{i}^{s(A, B)} = 0.5 \left[ 1 - \sum_{j, j \neq i}^{N_0} \sin^{2}(\theta_{i}^{A} - \theta_{j}^{A}) \cdot \sin^{2}(\theta_{i}^{B} - \theta_{j}^{B}) \right] + 1.0
\]

where \(N_0\) and \(\theta_{i(j, B)}\) are, respectively, the number of orthologous genes between the compared genomes, and the angle of the *i*th (or *j*th) ortholog on genomes *A* and *B*. The gene–location distance in equation (1) is based on the correlation coefficient for circular data (Fisher and Lee, 1983), and the coefficient is expressed as follows:

\[
\rho = \frac{\sum_{1 \leq i < j \leq N} \sin(\Theta_i - \Theta_j) \cdot \sin(\Phi_i - \Phi_j)}{\sqrt{\sum_{1 \leq i < j \leq N} \sin^{2}(\Theta_i - \Theta_j) \cdot \sum_{1 \leq i < j \leq N} \sin^{2}(\Phi_i - \Phi_j)}}
\]

where \(N, \Theta, \) and \(\Phi\) are the number of samples and the two angular variables on the two circles, respectively. The most distinctive point of the gene–location distance from the correlation coefficient is the sample handling. The gene–location distance handles the samples of orthologs that are related between two genomes. By contrast, the correlation coefficient handles the samples that are distinctly unrelated, such as window directions that are measured at different times in the same place (Fisher and Lee, 1983). The gene–location distances in equation (1), therefore, are calculated for each pair of orthologs, while a correlation coefficient in equation (2) is calculated for all samples. Additionally, the distance is designed to increase along with the dissimilarity and to range from 0.0 to 1.0, against the increase of the coefficient along with the similarity in the range from −1.0 to 1.0; a term in the brace of equation (1) is multiplied by −1 for the increase along with dissimilarity, and then the range is modified to be from 0.0 to 1.0 by two values, 0.5 and 1.0.

The essential properties of gene–location distances for location to location correspondence are illustrated by some hypothetical genomes in Figure 1. On the three pairs of genomes at the top of the figure, a gene of one orthologous pair on one genome transposes by 0, 90, and 180°, while the corresponding gene on the counter genome stays at the same location. As easily seen in the figure, the distance is proportional to the degree of gene transposition. The proportionality of a gene–location distance for one pair of genes is not exactly held when many orthologs concomitantly transpose on practical genomes. Additionally, the distance is sensitive when the genes change their orders on the genomes encoding a few orthologs (Horimoto et al., 1998a). However, the gene–location distance is effective to totally describe the difference between all gene locations over entire genomes. This property is illustrated on the three pairs of genomes at the bottom of the figure. The three orthologs that show the shortest, the middle, and the longest distances are set in the same angle by the rotations of entire genomes, respectively.
Orthologous gene location on archaeal and bacterial genomes

Fig. 1. Illustration of the properties of gene–location distance. Two hypothetical genomes are shown by outer and inner circles. In the three pairs of genomes at the top of the figure, three orthologous genes (named a, b, and c, as denoted on the larger circle) are encoded, and gene a on the inner genome is transposed by 0, 90, and 180°, respectively. The gene–location distances of gene a are indicated at the top of the larger circle, and are proportional to the degree of gene a transposition. In the remaining three pairs at the lower side, ten orthologs (named a–j) are encoded. Ten gene–location distances for each pair of orthologs in the two genomes are calculated, and the two genomes are rotated so that the a, f, and c orthologs, with the smallest, the middle, and the longest gene–location distances, are located in the same angle at the top of the figure. Each ortholog location is bound with a line, and in particular, orthologs that are mutually located within the range of 20° between two genomes are bound with bold lines.

Notes on the present data analysis
The gene–location distance in equation (1) is invariant, regardless of the selection of the point for measuring the angles. In the present analysis, the angles were measured from the starting nucleotide of each gene. In general, the data on orthologs contain some information, such as the sequence similarity, as well as the angle. Since the gene–location distance is principally defined by the angle of the gene, for simplicity, the additional information was excluded in the calculation of the distance in this study. Although the gene-coding orientation is an important factor for the genome comparison, some complexity arises in the calculation of gene–location distance. Two distances for the ortholog locations with each orientation are calculated, and it is unclear which orientation corresponds to each other in the comparison of the two genomes, due to the variety of the fraction of orthologs of the respective strands to all orthologs in the comparison between the three pairs of genomes, the gene–location distance shows larger values (0.07, 0.43, and 0.88), while the numbers of orthologs within the range of 20° decrease (6, 3, and 2). In general, the gene–location distances shorten as the numbers of orthologs that are located in the similar angle between two genomes increase, when the orthologs are set in the same angle. When one pair of orthologs with the shortest distance is set in the same angle by the rotations of two genomes, therefore, one can discern a gene configuration in which the locations of the orthologs correspond most effectively.
Fig. 2. Location to location correspondence between six pairs of microbial genomes. The corresponded genomes are denoted by two circles, in which the larger genome is on the outside and the smaller genome is on the inside. The orthologous genes with the shortest gene–location distances in each location to location correspondence are fixed in the same angle at the top of each figure; the fixed orthologs and their angles are MTH314 (51.1°) and MJ0709 (137.3°) between *M. thermoautotrophicum* and *M. jannaschii*, Rv1283c (117.2°) and BS_appB (103.8°) between *M. tuberculosis* and *B. subtilis*, evgS (192.6°) and HI0232 (51.9°) between *E. coli* and *H. influenzae*, BS_purE (59.6°) and MJ0616 (118.1°) between *B. subtilis* and *M. jannaschii*, argI (347.2°) and MJ0881 (174.4°) between *E. coli* and *M. jannaschii*, and argI (347.2°) and BS_argF (102.7°) between *E. coli* and *B. subtilis*. The orthologs within the range of 20° are bound by solid lines, and the start site in the Entrez Genome files is indicated on each circle.

The location to location correspondence was performed between nineteen archaeal and bacterial genomes. Since the correspondences amounted to one hundred and seventy-one, we will focus on six correspondences calculated for all combinations of orthologs within one group, and the ortholog with the shortest gene–location distance was selected from the orthologs in each group. By the selection, the effect of tandem duplication of genes is eliminated from the calculation of gene–location distance. In addition, statistically short distances as well as the shortest distance are further considered to eliminate any bias by the selection procedure in the evaluation of the nature of location to location correspondence.

**RESULTS**

The location to location correspondence was performed between nineteen archaeal and bacterial genomes. Since the correspondences amounted to one hundred and seventy-one, we will focus on six correspondences...
Orthologous gene location on archaeal and bacterial genomes

Fig. 3. Distribution of the numbers of orthologous genes within the range of 20°, when each ortholog between the two genomes is fixed in the same angle. The abscissa represents the numbers of orthologs within the range of 20° that are counted when each ortholog is set in the same angle. In each distribution, the observed number of orthologs by the shortest gene–location distance and the expected number of orthologs within 20° (about 11% = 40°/360° of all orthologs) are indicated with a solid arrow and a broken arrow, respectively, at the upper side of the figure. Additionally, the two values on the horizontal axis in each figure denote the maximum and minimum numbers of orthologs among all numbers of orthologs that are obtained by setting each ortholog in the same angle.

between genomes. The corresponded genomes are *M.thermoautotrophicum* and *M.jannaschii* in archaea, and *B.subtilis*, *M.tuberculosis*, *E.coli*, and *H.influenzae* in bacteria. Since the complete sequences of these genomes were determined at the early stages of complete genome sequencing, the orthologous relationships have been refined and extensive biological information has been accumulated. Correspondences between the other genomes not shown here are available at http://www.ged.saga-med.ac.jp/horimoto/L2LC/.

Visualization of location to location correspondence

According to the procedure for location to location correspondence, the orthologous genes with the shortest gene–location distance are set in the same angle by rotating genomes. In the correspondences, the orthologs whose locations are mutually ranged within 20° are visualized. The visualization of orthologs that are similarly located serves to facilitate an intuitive understanding of the features of the correspondence.

The six location to location correspondences are depicted in Figure 2. It is clearly seen that there are many orthologs that are mutually located within the range of 20° between the two genomes, but are separately located on each genome. The total numbers of orthologs within 20° were 114 of 792 orthologous genes between *M.thermoautotrophicum* and *M.jannaschii*, 142 of 918 between *M.tuberculosis* and *B.subtilis*, 133 of 881 between *E.coli* and *H.influenzae*, 63 of 463 between *B.subtilis* and *M.jannaschii*, 78 of 485 between *E.coli* and *M.jannaschii*, and 153 of 1080 between *E.coli* and *B.subtilis*. About 15% of the orthologs were mutually located within 20° in every correspondence. Although a few gene clusters, which are seen as a bundle of lines in the figure, were also found, one of the interesting features is that the orthologs within 20° were found over the entire regions
of the two genomes. This indicates that the rotation by the correspondence is attributed to the orthologs maintaining their relative locations over entire genomes, as well as the local gene clusters. Other correspondences not shown here also share similar features with the six correspondences (see our web site).

Notably, the number of orthologs within a range varies in the width of the range. When a range for location similarity is set to be larger than 20°, the number of orthologs within the range is more than that within 20°. In the following sections, therefore, the statistical inference of the correspondence is described in various ranges.

### Statistical evaluation of similarity between the locations of orthologous genes in the correspondence

As a preliminary investigation of the similarity between the ortholog locations in the location to location correspondence, the similarity is visually illustrated for the six correspondences. In one pair of genomes, each ortholog correspondence realizes a similar configuration of orthologs between genome pairs.

To statistically infer the similarity between the ortholog locations in the location to location correspondence, a simulation procedure was designed. Firstly, a set of hypothetical locations of genes with the same number of original orthologs was sampled from all locations of orthologs in one of the genomes, with replacement. Then, the rotation procedure was applied to obtain the null distribution of the number of orthologs within a range, like the distributions of the numbers of orthologs in Figure 3. According to the above rule, one hundred sets of null distributions were independently generated for each genome pair. From the one hundred sets, the average and the standard deviation were calculated. Using the average and the standard deviation in each genome pair, the probability for the number of orthologs in the correspondence was calculated on the assumption that the distribution approximated a normal distribution. Thus, the real locations of orthologs in the two genomes can be tested against the random locations.

According to the above procedure, the similarity between the locations of orthologs within the range of 20° was statistically estimated to be less than the 5% significance level in the six correspondences. The averages and the standard deviations in one hundred sets of sampling data were calculated in the six correspondences: 88.0 and 9.25, 101.9 and 9.69, 97.9 and 9.60, 51.4 and 6.74, 53.8 and 6.85.
Fig. 4. Allocation of the angles of orthologous genes with the short gene–location distances at a 5% significance level into those of all orthologs. The axes indicate the angles of the orthologs on each genome, which are cited from the Entrez Genome files (Benson et al., 1999). The orthologs with short gene–location distances, as estimated by assuming that the distribution of the gene–location distances approximates the normal distribution, are denoted by black circles, and the remaining pairs are indicated by open circles.

and 7.04, and 119.9 and 10.44. The averages are almost equal to the expected values (40°/360° ≈ 11% of total orthologs), suggesting that the present bootstrap procedure generated the random distribution well. (The generated distributions are available at our web site.) Using the above values, the probabilities in each correspondence were estimated: 0.0025 between M.thermoautotrophicum and M.jannaschii, 0.0011 between M.tuberculosis and B.subtilis, 0.0001 between E.coli and H.influenzae, 0.0425 between B.subtilis and M.jannaschii, 0.0003 between E.coli and M.jannaschii, and 0.0008 between E.coli and B.subtilis.

Since the number of orthologs within a range depends on the width of the range, the probabilities are further calculated by the bootstrap procedure in the ranges varying from 1 to 60° by 1° increments. Thus, the numbers of orthologs in the correspondences are completely tested in sixty ranges for the similar location. The statistical evaluation of the correspondence for all pairs of genomes is shown in Table 1. As seen in Table 1, the numbers of orthologs that maintain their relative locations are statistically significant in most of the genomes. When the range for location similarity is set to 20°, the probabilities are less than 0.05 in 169 of 171 correspondences. In the sixty ranges from 1 to 60°, less than 0.05 of the probabilities frequently emerged in all correspondences. Even in the three correspondences with more than 0.05 of probability in the range of 20°, some ranges with less than 0.05 of probability are found. In conclusion, the correspondence by the shortest gene–location distance promises to find a rotation angle, where statistically significant numbers of orthologs are similarly located between the genomes.

Stability of location to location correspondence
The robustness of the location to location correspondence for realizing a similar configuration between the orthologous genes is estimated from another aspect. We examined whether a rotation angle by the shortest gene–location distance is similar to the rotation angles by the short distances.

In the six correspondences, at first, the short gene–
A characteristic distribution of the orthologs with short location distances at a 5% significance level is calculated on the assumption that the distribution of the distances approximates a normal distribution, and then the angles of orthologous genes with the distances at a 5% significance level are allocated into those of all orthologs, as shown in Figure 4. From the scattered distribution of all orthologs, a characteristic distribution of the orthologs with short location distances at a 5% significance level is distributed around the angles of the origin site in M. pneumoniae, according to the previous report (Himmelreich et al., 1997). An open circle indicates a successful case when the allocated region included known replication sites, and a cross indicates a false case. Note that the angles of the origin site in M. pneumoniae were reversibly measured from the first site of the original genomic data, i.e. $360^\circ - \theta_M^{Mpn}$, according to the previous report (Himmelreich et al., 1997).

### Table 2. Agreement of the allocated regions of replication sites in the location to location correspondences with experimental data

<table>
<thead>
<tr>
<th>Genome pair</th>
<th>Allocated region</th>
<th>Experimental data</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eco vs Bsu</td>
<td>$-11^\circ - 64^\circ$ 168$^\circ - 243^\circ$</td>
<td>0$^\circ$ 172.0$^\circ$</td>
<td>○</td>
</tr>
<tr>
<td>Eco vs Mtu</td>
<td>$-5^\circ - 60^\circ$</td>
<td>0$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Eco vs Hin</td>
<td>100$^\circ - 175^\circ$ 280$^\circ - 355^\circ$ 118.4$^\circ$ 288.4$^\circ$</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Eco vs Mge</td>
<td>10$^\circ - 155^\circ$ -</td>
<td>0.0$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Eco vs Mpn</td>
<td>$-20^\circ - 60^\circ$ -</td>
<td>269.6$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Bsu vs Mtu</td>
<td>$-50^\circ - 70^\circ$ -</td>
<td>0.0$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Bsu vs Hin</td>
<td>100$^\circ - 160^\circ$ 280$^\circ - 340^\circ$ 118.4$^\circ$ 288.4$^\circ$</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Bsu vs Mge</td>
<td>$-20^\circ - 40^\circ$ -</td>
<td>0.0$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Bsu vs Mpn</td>
<td>250$^\circ - 330^\circ$ -</td>
<td>269.6$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Mtu vs Hin</td>
<td>90$^\circ - 140^\circ$ 260$^\circ - 310^\circ$ 118.4$^\circ$ 288.4$^\circ$</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>Mtu vs Mge</td>
<td>$-10^\circ - 60^\circ$ -</td>
<td>0.0$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Mth vs Mpn</td>
<td>260$^\circ - 350^\circ$ -</td>
<td>269.6$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Hin vs Mge</td>
<td>$-5^\circ - 80^\circ$ -</td>
<td>0.0$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Hin vs Mpn</td>
<td>260$^\circ - 350^\circ$ -</td>
<td>269.6$^\circ$ n/a</td>
<td>○</td>
</tr>
<tr>
<td>Mge vs Mpn</td>
<td>271.4$^\circ$ -</td>
<td>269.6$^\circ$ n/a</td>
<td>○</td>
</tr>
</tbody>
</table>

The replication sites of the former genome are allocated to the region of the latter genome by the rotation range that is restricted by the short gene–location distances at the 5% significance level. On the latter genome, the allocated region is compared with the known replication sites (Kunst et al., 1997; Fraser et al., 1995; Himmelreich et al., 1996; Cole et al., 1998; Blattner et al., 1997; Fleischmann et al., 1995). An open circle indicates a successful case when the allocated region included known replication sites, and a cross indicates a false case. Note that the angles of the origin site in M. pneumoniae were reversibly measured from the first site of the original genomic data, i.e. $360^\circ - \theta_M^{Mpn}$, according to the previous report (Himmelreich et al., 1997).

![Number of genome pairs vs Rotation width](image)

**Fig. 5.** Rotation widths by short gene–location distances at a 5% significant level in the location to location correspondences between nineteen microbial genomes. The differences between the angles of orthologous genes with short distances are measured, and the maximum difference is regarded as the rotation width of each correspondence. The rotation widths are counted by $10^\circ$ intervals.
the short distances at the 5% level are similar to the direction of the shortest distance for the best correspondence characteristic of each genome pair. In other words, the restriction of the rotations indicates that each genome pair tends to have an intrinsic angle for realizing the similar locations of orthologs between two genomes.

**DISCUSSION**

**Correlation of location to location correspondences with DNA replication sites**

One of the striking features of the restricted ranges of rotations in the location to location correspondences is that the restricted rotations appear to be correlated with the origins and the termini of DNA replication in the compared genomes. Between the six genomes with known replication sites, the correlation between the rotation angles and the replication sites is summarized in Table 2. For example, a correlation between the rotation angles and the replication sites is illustrated in the correspondence between *E.coli* and *B.subtilis*. As seen in Figure 4, the rotation of the first site of the *B.subtilis* genome data is restricted in the range from −120 to −45°. By the restricted rotation, the origin on *E.coli* (304.4°) (Blattner *et al.*, 1997) is allocated to the range from −11 to 64° on *B.subtilis*. The allocated range of the origin of *E.coli* on *B.subtilis* includes the origin site of *B.subtilis* (0°) (Kunst *et al.*, 1997). Concomitantly, the terminus site of *E.coli* (122.8°) is allocated to the range from 168 to 243° on *B.subtilis*, where the terminus site of *B.subtilis* (172.0°) is included. Although both replication sites are allocated at the edges of ranges, the rotation range agrees with the replication sites between *E.coli* and *B.subtilis*. In 12 of 14 remaining genome pairs, the agreement of the replication sites by the rotation is also seen; the allocated regions of the replication sites of one genome on the counter genome include the replication sites of the counter genome. Between *M.genitalium* and *M.pneumoniae*, the allocation site of the *M.genitalium* origin site on *M.pneumoniae* is positioned quite near the *M.pneumoniae* origin site. In this case, a pair of orthologs with the shortest distance, instead of the range by the short distances, is selected for the allocation, since the gene order is well conserved and is mutually reversed between the two genomes (Himmelreich *et al.*, 1997). The exceptional cases are seen between *E.coli* and *M.pneumoniae* and between *E.coli* and *M.genitalium*. In the latter case, however, the *E.coli* origin site is allocated in the range from 10 to 155° on *M.genitalium*, and the edge of this region is slid from the origin site of *M.genitalium* by only 10°. Consequently, the rotations in which the orthologs share a similar configuration result in the agreement of the replication sites between the genomes.

Recently, some observations between the ortholog locations on the genomes of the closely related bacteria revealed that a substantial proportion of gene rearrangements results from recombination sites that are determined by the positions of the replication forks, suggesting that replication sites have a major role in directing genome evolution (Anderson, 2000; Read *et al.*, 2000; Tillier and Collins, 2000). The gene rearrangement by replication-directed translocation supports the present agreement of the restricted ranges with the replication sites. If most of the gene translocations are attributed to the replication-dependent rearrangements, then the most similar configuration between the two genomes may emerge when the replication sites are set in the same angle. Thus, the agreement even in the distantly related genomes implies that the replication-directed translocation plays an important role for the gene rearrangement together with the insertion or deletion, in a wide spectrum of bacteria with various genome sizes.

**Classification of location-maintained orthologs in terms of gene function**

To further explore the nature of location to location correspondence, the location-maintained orthologs are grouped into fifteen classes, according to the gene function classification scheme in COGs (Tatusov *et al.*, 1997, 2000; Koonin *et al.*, 1998). Then, the bias of the numbers of location-maintained orthologs against those of all orthologs in the fifteen classes is statistically estimated by the $\chi^2$-test. By this test, one can assess the issue of whether the orthologs in the class are more or less maintained. If the observed value is larger than the expected one in a class, then the class is regarded as the well-maintained class, and if the observed value is less than the expected one, then the class is regarded as the less-maintained class.

Among the distributions of location-maintained orthologs with the bias at a 5% significance level, the classes are counted in terms of the difference between the observed and expected values in Table 3. One of the remarkable features is that a similar pattern of more or less location-maintained classes is seen in the three groups, although the degree of difference between the two values depends on the classes. In the three classes (classes J, K and L) of the information storage and processing category, the observed number of RNA related proteins (K) is more than the expected one, and by contrast, the observed number is less than the expected one in DNA related proteins (J and L). In six classes (D, O, M, N, P and T) of the cellular processes category, the well-maintained classes (P and T) appear to be proteins involved in small-molecule transport, in contrast to the intra- and intercellular communication proteins in the less-maintained classes (D, O and N). Although an evolutionary investigation is be-
from the structural and evolutionary points of view. theolog locations may be associated with the gene function more or less maintenance of ortholog locations, the or-
further analyses are needed to resolve the causality of biosynthesis maintain their relative locations. Although in their relative locations, proteins involved in amino acid terms of the maintenance. While ribosomal proteins vary frequently form an operon structure are distinctive in the two classes (J and E) that contain the genes that (Kyrpides et al. The genome content of the last common ancestor of life E.coli clustering of the sulfur metabolism-related genes on the classes (E and H) might be involved in the non-random an unclear pattern (C). Furthermore, the well-maintained on the living environment and the structure of microbes (G, E and H) appear to contain the genes less dependent on the comparison groups. The well-maintained classes of the metabolism category, four classes (G, E, H and I) show a clear feature, while two classes (C and F) depend on the comparison groups. The well-maintained classes (G, E and H) appear to contain the genes less dependent on the living environment and the structure of microbes than the less-maintained class (I) and the one class with an unclear pattern (C). Furthermore, the well-maintained classes (E and H) might be involved in the non-random clustering of the sulfur metabolism-related genes on the E.coli chromosome (Rocha et al., 2000). Interestingly, the two classes (J and E) that contain the genes that frequently form an operon scheme will be presented in the future. At any rate, the present approach can be applied to any relationship in the poorly characterized category in COGs were excluded in this classification.

Concluding remarks

The location to location correspondence realizes a comprehensive comparison between circular genomes in terms of gene location. The present analysis by the procedure reveals that statistically significant numbers of remote orthologs maintain their relative locations over the entire genomes of archaea and bacteria, against the scattered orders of orthologs in the narrow regions of genomes. The agreement is found between the rotations in the effective correspondences and the known replication sites, consistent with a mode for the gene translocation between closely related bacteria. The classification of the location-maintained orthologs suggests the relationship between locations and functions in orthologs.

<table>
<thead>
<tr>
<th>Gene class</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>D</th>
<th>O</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>E</th>
<th>F</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within archaea</td>
<td>+</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>2</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>29</td>
<td>29</td>
<td>16</td>
<td>29</td>
<td>29</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>29</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>27</td>
<td>19</td>
<td>22</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>Within bacteria</td>
<td>+</td>
<td>79</td>
<td>737</td>
<td>79</td>
<td>215</td>
<td>51</td>
<td>137</td>
<td>26</td>
<td>707</td>
<td>737</td>
<td>361</td>
<td>683</td>
<td>700</td>
<td>651</td>
<td>724</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>688</td>
<td>0</td>
<td>522</td>
<td>522</td>
<td>686</td>
<td>600</td>
<td>711</td>
<td>30</td>
<td>0</td>
<td>376</td>
<td>54</td>
<td>37</td>
<td>86</td>
<td>13</td>
</tr>
<tr>
<td>Archaea vs bacteria</td>
<td>+</td>
<td>1</td>
<td>120</td>
<td>21</td>
<td>19</td>
<td>41</td>
<td>76</td>
<td>1</td>
<td>107</td>
<td>120</td>
<td>45</td>
<td>120</td>
<td>115</td>
<td>88</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>119</td>
<td>0</td>
<td>101</td>
<td>101</td>
<td>79</td>
<td>44</td>
<td>119</td>
<td>13</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td>5</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>

At first, the distribution of the location-maintained orthologs in fifteen classes, which shows significant bias ($P < 0.05$, by the $\chi^2$-test), is detected in the location to location correspondences by the shortest gene location distances. The ranges for similar locations vary from 5 to 30° by 1° increments. Then, in each distribution with the bias at a 5% significant level, the class in which the observed number of the location-maintained orthologs is more than the expected one is regarded as the well-maintained class (denoted by ‘+’), and the class in which the observed number is less than the expected one is regarded as the less maintained class (‘−’). The classes are counted in three groups: within archaea, within bacteria, and between archaea and bacteria. The total numbers of classification tests in the three groups were 156 ($4C_2$ [combination of genomes] $\times 26$ [similarity ranges 5 to 30°]), 2730 ($15C_2 \times 26$), and 1560 ($4 \times 15 \times 26$), respectively. Among the tests, the numbers of distributions with a 5% significance level were 29, 737, and 120, respectively. (All distributions with $P < 0.05$ are available at our web site.) Gene classes are adopted from COGs (Tatusov et al., 1997, 2000; Koonin et al., 1998). The abbreviations of the gene classes are adopted from COGs: J, Translation, ribosomal structure and biogenesis; K, Transcription; L, DNA replication, recombination and repair; D, Cell division and chromosome partitioning; O, Posttranslational modification, protein turnover, chaperonins; M, Cell envelope biogenesis, outer membrane; N, Cell motility and secretion; P, Inorganic ion transport and metabolism; T, Signal transduction mechanisms; C, Energy production and conversion; G, Carbohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme metabolism; I, Lipid metabolism. For clarity, two classes in the poorly characterized category in COGs were excluded in this classification.

Table 3. Pattern of more or less location-maintained orthologs in the classification of gene function
ACKNOWLEDGEMENTS

We would like to thank Dr Hiroyuki Toh, of the Biomolecular Engineering Research Institute, and Professor Ken Nishikawa, of the National Institute of Genetics, for valuable discussions. Also, we would like to thank two anonymous referees for their instructive comments in improving the earlier version of the manuscript. K.H. was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) ‘Genome Information Science’ from the Ministry of Education, Science, Sports, and Culture of Japan (grant 12208038 and 1308028), and by the NIG Cooperative Research Program (grant 2000–43). S.F. was supported by Japan Science and Technology Corporation.

REFERENCES

the Escherichia coli genome: markers of the cell’s architecture? FEBS Lett., 476, 8–11.