Elucidating Genome Structure Evolution by Analysis of Isoapostatic Gene Clusters using Statistics of Variance of Gene Distances

Naobumi V. Sasaki, and Naoki Sato*

Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, Tokyo, Japan

*Corresponding author: E-mail: naokisat@bio.c.u-tokyo.ac.jp.

Accepted: 15 December 2009 **Associate editor:** Dr. William Martin

Abstract

Identifying genomic regions that descended from a common ancestor is important for understanding the function and evolution of genomes. In related genomes, clusters of homologous gene pairs serve as evidence for candidate homologous regions, which make up genomic core. Previous studies on the structural organization of bacterial genomes revealed that basic backbone of genomic core is interrupted by genomic islands. Here, we applied statistics using variance of distances as a measure to classify conserved genes within a set of genomes according to their "isoapostatic" relationship, which keeps nearly identical distances of genes. The results of variance statistics analysis of cyanobacterial genomes including *Prochlorococcus*, *Synechococcus*, and *Anabaena* indicated that the conserved genes are classified into several groups called "virtual linkage groups (VLGs)" according to their positional conservation of orthologs over the genomes analyzed. The VLGs were used to define mosaic domain structure of the genomic core. The current model of mosaic genomic domains can explain global evolution of the genomic core of cyanobacteria. It also visualizes islands of lateral gene transfer. The stability and the robustness of the variance statistics are discussed. This method will also be useful in deciphering the structural organization of genomes in other groups of bacteria.

Key words: comparative genomics, cyanobacteria, gene distance profile, genome core, isoapostatic genes.

Introduction

In closely related genomes of bacteria, segments of genes with conserved gene order, which is referred to as synteny blocks, are found by mapping orthologous genes (or simply called orthologs) to each genome. In many cases, a stable genome structure consisting of many synteny blocks, called "genomic core," is shared by many genomes. Such structure reflects evolution of the genomes. There are also variable regions of "genomic islands" consisting of laterally transferred genes (Suyama and Bork 2001). Bacterial genomes are, therefore, envisaged as a mosaic of genomic core interspersed with genomic islands. Such mosaic structure of bacterial genome has been intensively studied in *Escherichia coli* (Dobrindt 2005; Rasko et al. 2008) and cyanobacterial genomes among others.

Cyanobacteria or photosynthetic prokaryotes with oxygen evolution exhibit ecological and morphological adaptation to wide ecological spectrum (Whitton and Potts 2000). In a classical review on the molecular evolution of cyanobacteria, Doolittle (1982) raised three questions, namely, 1) what is the proper phylogenetic position of the cyanobacteria within the prokaryotes? 2) what phylogenetic relationships exist within the cyanobacteria? and 3) what evolutionary relationship do cyanobacteria bear to eukaryotic photosynthesizers? As reviewed by Wilmotte (1994), botanical, bacteriological, and molecular approaches have contributed to respond to these questions. In particular, in marine species of unicellular cyanobacteria, various ecological variants called "ecotypes" are recognized. They are adapted to high light (upper layer of ocean) or low light (deep sea), with (coastal region) or without (open ocean) supply of rich nutrients. These ecotypes are phylogenetically closely related as analyzed by sequence conservation, such as the 16S–23S internal transcribed spacer sequences (Rocap et al. 2002; Johnson et al. 2006) or the 16S ribosomal RNA (rRNA) sequences that differ by at most 3% (Kettler et al. 2007). However, high genomic flexibility was reported among the *Prochlorococcus* ecotypes, namely,

only 40–67% of the genes are shared in all available *Prochlorococcus* genomes. Genomic comparison of these genomes revealed that the basic structures of the genome are identical to that in other bacterial genomes in that most of the shared orthologs are arranged in conserved order to form stable core, and additional genes are located within genomic islands. Coleman et al. (2006) suggested that the contextual flexibility is attained by mosaic structure of genomic islands and stable cores, whereas Dufresne et al. (2008) suggested that the core genome plays a constitutive function and the accessory genome is related to ecotype-specific functions. In spite of these findings, little has been argued about the evolution of fairly stable structure of the genome core. We aimed to analyze features of the genome core in cyanobacterial genomes.

To find synteny blocks, we need a multiple genome alignment in advance. Many software and algorithms have been developed such as LAMARCK (Wolf et al. 2001), Murasaki (http://murasaki.dna.bio.keio.ac.jp), MBGD (Uchiyama 2003), and LAGAN (Brudno et al. 2003) to obtain alignments. These methods involve many improvements from the basic idea of the alignment proposed by Sankoff et al. (1992), but the quality of alignment results by these algorithms depends on gap penalty during the process of alignment reduction. Even in the case of globally optimized alignments by maximum matching approach, correctness of local alignment is not always guaranteed (Brudno et al. 2004). In these algorithms, optimization strategy is combinatorial. In other words, the number of possible alignments will explode with increase in number and diversity of genomes.

We propose here an alternative approach, namely, a statistical one. Instead of using simple distances of orthologs, we use variance of ortholog distances as a measure of dissimilarity in multivariate analysis. Such analysis will detect groups of orthologs that keep constant distances over various genomes, which we call "isoapostatic" (similar distance in Greek) relationship. The method allowed us to analyze the mutual relationship of orthologs in a feature space. Clustering in the feature space resulted in groups of orthologs (virtual linkage groups [VLGs]) that keep isoapostatic relationship in real genomes. This method was successfully applied to detect synteny blocks over many marine species of cyanobacteria.

Materials and Methods

Principles of the Method The multiple genome alignment problem was originally described by Sankoff et al. (1996), which is intended to find a phylogenetic tree describing the most plausible rearrangement scenario for multiple genomes. We explain the method using an example (fig. 1) and then introduce our new method.

Suppose a set of genes *A, B, C, D,* and *E,* which are conserved over four genomes, 1, 2, 3, and 4. The order of these genes is different in the four genomes, and the genomes are supposed to be rearranged during the evolution but we do not know the evolutionary history. How can we reconstruct the evolutionary history of rearrangement of these five genes? Possible rearrangements and inversions are shown by lines (fig. 1*A*). A guide tree is also inferred using maximum parsimony strategy for rearrangements. In figure 1*A*, two blocks can be distinguished as shown by colors. Gene *A* and gene *B* are present within 1 or 2 distance units. Gene *C* and gene *E* also keep 1 or 2 units distance. The position of gene *D* changes with respect to these two blocks. Such erratic genes will make genome alignment difficult. Although introduction of gap penalty relaxes the problem, it is still compelling problem because no systematic method of estimating gap penalties for particular genomes is known. Even in the case of globally optimized alignments by maximum matching approach, validity of local alignment is not guaranteed (Brudno et al. 2004). In this approach, resulting alignment is affected by the topology of guide tree. That is why optimization of alignment and optimization of tree are inseparable. This is an example of small hypothetical genomes, but as the number of genomes and their diversity increase, construction of alignment and tree will be more difficult.

An alternative approach is a statistical method. Here, we describe a method using multivariate analysis. An advantage of this method is that we can obtain cluster of genes by the similarity of distance without considering hierarchical relationships of genomes. Let us consider a distance metric of a pair of orthologs. Figure 1*B* shows a matrix of "distance of orthologs," which is defined by mean distance of ortholog pair over all genomes. In this case, the number of intervening genes is taken as a measure of distance of orthologs. The data in figure 1*B* correspond exactly the situation in figure 1*A*. However, distance of orthologs within a syntenic cluster is continuous, and there is no clear-cut distinction between the distance values within a cluster and the distance values over different clusters. Then, what is an invariable inherent to a cluster? We propose to use "variance measure," which is defined as the variance of distance of orthologs over genomes. Figure 2 shows that the distance itself is not an invariant but the variance of distance is an invariant within a cluster if it is completely conserved over all genomes. In a more realistic case, the variance is a small positive real number within a cluster, whereas the variance is large for a pair of genes that do not belong to an identical cluster. Figure 1*C* shows a distance matrix using the variance measure for the example in figure 1*A*. The result of hierarchical clustering using the matrix of variance measure indicates that *AB* and *CE* are correctly clustered as in the result of combinatorial method (fig. 1*D*). Note that the gene *D* is associated with *CE* rather than *AB*. This is because the single long-distance transposition event made the variance larger. This shows that the variance measure is suited for detecting

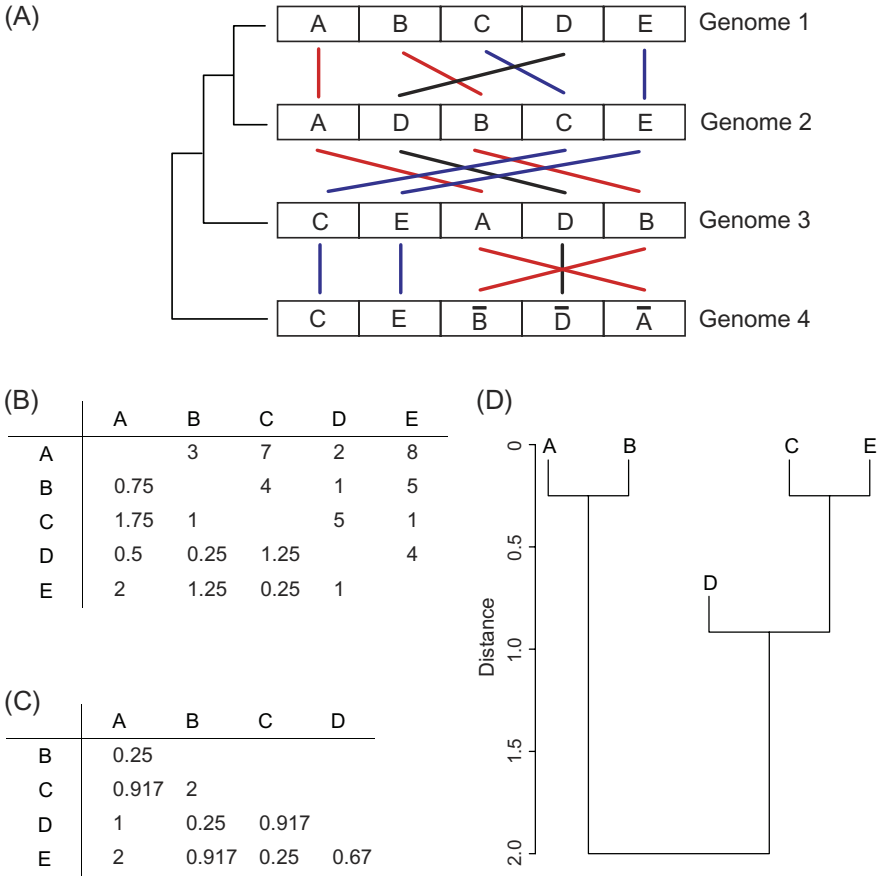

Fig. 1.—Schematic explanation of the method. (*A*) A model set of rearranged genomes. Each box indicates a gene. The genomes are ordered by the model evolutionally scenario as shown by the dendrogram on left. (*B*) Distance matrix of figure 1*A*. Upper triangle: total interval distance; lower triangle: mean interval distance. (*C*) Matrix of variance measure of figure 1*A*. (*D*) Dendrogram of the genomes in figure 1*A* calculated by hierarchical clustering using the variance matrix in (*C*).

conserved region. Hence, we call such relationship isoapostatic ("iso" = similar; "apostasis" = distance; in Greek). The final result might not be very different from the result of syntenic analysis in simple cases, but this method is more powerful if the genomes contain many small changes such as transposition, inversion, insertions/deletions (indels), and horizontal gene transfers. In conventional analysis of synteny block, only neighboring relationship is considered (such as A_1–A_2, A_2–A_3, ... A_{m-1}–A_m, etc: bold colored items in fig. 2*B*), but the clustering using variance measure considers all possible distance relationships (A_1-A_3, A_1-A_4, A_{m-2}-A_m, etc plus the above mentioned ones: colored area in fig. 2*C*). This is the basis of the robustness of the method using isoapostatic relationship.

Data table 1 lists the genomes used in the present study. We prepared two data sets. One includes 14 marine cyanobacterial genomes and the other consists of two *Anabaena* genomes. RefSeq (Pruitt et al. 2007) files of ten strains of *Prochlorococcus marinus* (MED4, MIT9313, CCMP1375, MIT9312, NATL2A, MIT9301, MIT9303, MIT9515, NATL1A, and AS9601), four strains of marine *Synechococcus* (WH8102, CC9902, CC9605, CC9311), and two *Anabaena* strains (PCC7120 and ATCC47912) were downloaded from the National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov) and gene order information was extracted by parsing the FEATURE field. The genes are indexed by their order in each genome. Next, the Cyano25 data set of homolog groups was obtained from the Gclust database (http://gclust.c.u-tokyo.ac.jp; Sato 2009). The Cyano25 data set contains all 75,709 proteins encoded in 25 cyanobacterial genomes. Then, we selected orthologous single-copy genes that are shared in all the genomes analyzed. Accordingly, we obtained 917 and 2778 orthologs in the two data sets, respectively.

Directional Circular Distance We aimed to detect isoapostatic genes that keep identical mutual spacing over various genomes. Because isoapostatic genes are expected to reside at any positions within a genome, multivariate analysis was employed rather than typical graph-searching algorithms that have been used for searching synteny or gene cluster. Let G^t be a circular genome having a set of k single-copy

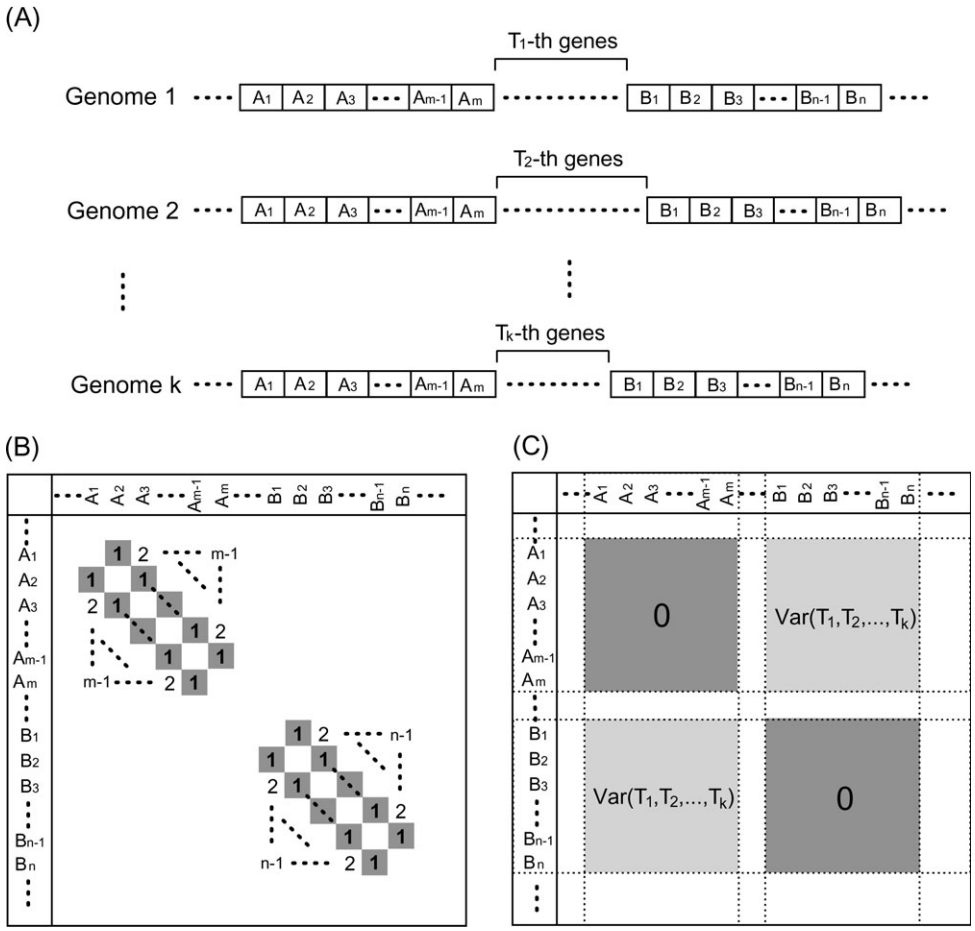

Fig. 2.—Illustration of the difference between the matrices with measures of mean distance and variance. (*A*) A model set of genomes with conserved synteny blocks. Each box indicates a gene and each alphabetical label corresponds to a synteny block. (*B*) Mean distance matrix of figure 2*A*. Elements representing relationship of neighboring genes are colored in gray. (*C*) Matrix of variance measure. Each gray-hatched domain has a unique value as shown. Note that each of the two blocks will be merged into a single domain in a feature space, when Var(T_1, T_2,…, T_k), the values corresponding to other regions are small enough.

orthologs shared by all the m genomes and n nonorthologous genes. Here, we do not use suffix for genome t in this paragraph for simplicity. Under the approximation that all genes are distributed over a unit circle evenly, the arc length $l(i, j)$ between the i-th and the j-th orthologs is defined by:

$$l(i,j) = \frac{2\pi * |i - j|}{n + k}. \quad (1)$$

Then absolute distance $f(i, j)$ is calculated by

$$f(i,j) = \pi - |\pi - l(i,j)|. \quad (2)$$

We then consider $d(i, j)$, a signed distance between orthologs i and j in the genome G^t. If gene j is located upstream of gene i, the sign is defined as negative. It can be formulated as:

$$d(i,j) = \begin{cases} |f(i,j)|, & \text{if gene } j \text{ is located downstream of gene } i, \\ -|f(i,j)|, & \text{otherwize.} \end{cases} \quad (3)$$

Definition of Distance Measures Consider positional relationships of n orthologs in m genomes and between each pair of orthologs (i, j), the distance between the two orthologs $d(i, j)$ can be measured by the count of intervening genes on each genome. By measuring all pairs of orthologs on each genome, we obtain the positional profile for n genes on m genomes. Here, we define intergenomic distance of the ortholog pair (i, j) such as

$$v_{ij} = \sum_{t=1}^{m} \{d^t(i,j) - \overline{d(i,j)}\}^2, \quad (4)$$

Table 1
Summary of the Genomes Used in This Study

Species or strain	Genes	Genome size (kb)	GC content (%)	Taxonomic group
Prochlorococcus marinus MED4	1717	1036	30.8	Low-B/A *Prochlorococcus* clade I
P. m. MIT9313	2269	2410	50.7	High-B/A *Prochlorococcus* clade IV
P. m. CCMP1375 (SS120)	1883	1751	36.4	High-B/A *Prochlorococcus* clade II
P. m. MIT9312	1810	1709	31.2	Low-B/A *Prochlorococcus* clade II
P. m. NATL2A	1892	1842	35.1	High-B/A *Prochlorococcus* clade I
P. m. MIT9301	1907	1641	31.3	Low-B/A *Prochlorococcus* clade II
P. m. MIT9303	2997	2682	50.0	High-B/A *Prochlorococcus* clade IV
P. m. MIT9515	1906	1704	30.8	Low-B/A *Prochlorococcus* clade I
P. m. NATL1A	2193	1864	34.0	High-B/A *Prochlorococcus* clade I
P. m. AS9601	1921	1669	31.3	Low-B/A *Prochlorococcus* clade II
Synechococcus sp. WH8102	2519	2434	59.4	Marine A *Synechococcus* clade III
S. CC9311	2892	2606	52.0	Marine A *Synechococcus* clade I
S. CC9605	2645	2510	52.9	Marine A *Synechococcus* clade II
S. CC9902	2307	2234	54.2	Marine A *Synechococcus* clade IV
Anabaena sp. PCC7120	5366	6413	41.3	*Anabaena/Nostoc*
A. variabilis ATCC29413	5043	6365	41.4	*Anabaena/Nostoc*

where $d^t(i,j)$ is the distance on the t-th genome, and $\overline{d(i,j)}$ is the mean distance of all genomes. The score v_{ij} is used as a variance measure. Then, we obtained the $n \times n$ matrix V consisting of variance scores was used to reconstruct the feature space by multidimensional scaling (Cox TF and Cox MAA 2001) using the "cmdscale" function in the stats package (version 2.3.1) on the R software platform (version 2.3.1; Becker et al. 1988), with the default parameter settings. This operation causes dimension contraction to satisfy the metric criteria that we call "isoapostasy" in the feature space.

Clustering in the Feature Space Clustering of isoapostatic genes that are mapped in the feature space was classified by the partitioning around medoids method, a variant of the "k-means" method (Kaufman and Rousseeuw 1990) using the "pam" function in the cluster package (version 1.11.0) in the R software platform, with the default parameter settings. The number of clusters was evaluated by the silhouette width (Rousseeuw 1987)

Phylogenetic Analysis Aligned sequences of 16S and 23S rRNA of cyanobacteria were obtained from the Ribosomal Database Project release 9 (http://rdp.cme.msu.edu, Cole et al. 2009) and European rRNA database (http://www.psb.ugent.be/rRNA/, Wuyts et al. 2004), respectively. All 16S and 23S rRNA sequences of the cyanobacterial genomes were obtained from the RefSeq database and aligned to the prealigned rRNA sequences by ClustalX software version 1.83 (Thompson et al. 1997). Subsequent sequence manipulation was performed by the SISEQ software version 1.59 (Sato 2000). The sites having gaps in more than 20% sequences were removed. Bayesian Interference tree was constructed by the MrBayes software version 3.1.2 (Ronquist and Huelsenbeck 2003), using the doublet model for base pairs and the 4 by 4 model for other sites, with nst = 6 and rates = invgamma.

Results

Isoapostatic Genes in Marine Cyanobacteria We analyzed distance statistics of 14 marine cyanobacterial genomes (*P. marinus* MED4, MIT9313, CCMP1375, MIT9312, NATL2A, MIT9301, MIT9303, MIT9515, NATL1A, AS9601, and *Synechococcus* sp. WH8102, CC9902, CC9605, CC9311) that had been sequenced before the start of the present study. They shared 917 unique orthologs. The result of simple distance statistics using multidimensional scaling by Euclidean distance (fig. 3) showed that most objects (orthologs) were arranged in a fragmented circular shape. It reflected roughly the circular structure of cyanobacterial genomes. Because the position of each object reflected its average position in the 14 genomes, the clusters of objects roughly corresponded to ortholog clusters shared by all the genomes.

Isoapostatic relationship of orthologs was analyzed by variance statistics using multidimensional scaling (fig. 4*A*). We found again clusters of orthologs, but this time, separation of clusters was clearer because the objects were distributed as patches rather than along a circle as in figure 3. Hence, we propose the name VLG for each cluster (the reason is explained in Discussion) because it represented a unit of gene assembly within a chromosome over various genomes. This situation is reminiscent of the linkage group in classical genetics, in which linkage group is a unit of gene assembly in genetic crosses. The VLG is, in contrast, a unit of gene assembly or rearrangement during genome evolution rather than genetic crosses. As a result, we obtained eight clusters by the clustering using the partitioning around medoids method. The statistics of silhouette widths indicated that the clustering into eight clusters was one of the bests (fig. 4*B* and *C*), and the remapping of this result onto figure 3 reconstructed sectors of orthologs. The two different

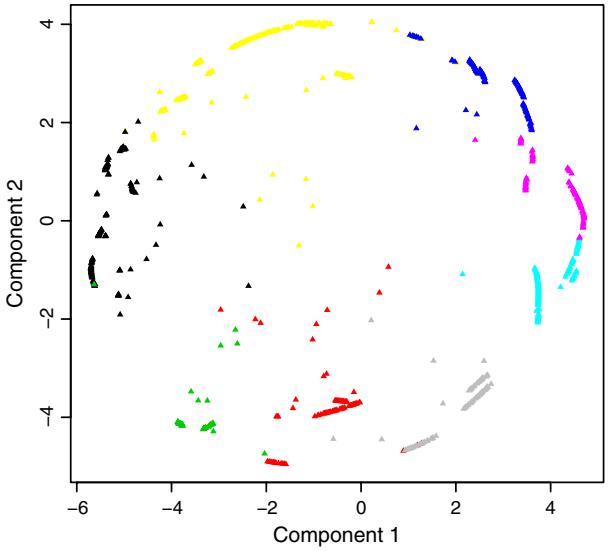

Fig. 3.—Embedding of orthologs in the 14 marine cyanobacterial genomes by multidimensional scaling using Euclidean distance. Data points indicate conserved orthologs and the color of each object indicates classification by VLG as later defined by the result of figure 4. Note that the data points are arranged on the periphery of a fragmented circular form and each VLG coincided with a sector of the circle.

methods of estimating silhouette widths did not affect much on the selection of the best clustering.

Close examination of figure 3 and figure 4 revealed, however, a curious similarity. The circular arrangement of the eight colors was similar in these two figures. Substructures such as the three subclusters in navy blue were apparent in both figures. However, figure 4*A* was better suited for clustering because the grouping was 2D rather than linear.

Isoapostatic Genes in *Anabaena* sp. PCC 7120 and *A. variabilis* ATCC 29413 We applied the same method to binary comparison of genomes of *Anabaena* (also called *Nostoc*) sp. PCC 7120 and *A. variabilis* ATCC 29413. These two filamentous cyanobacteria share 2,778 unique orthologs, which we used for the analysis. The result of multidimensional scaling by Euclidean distance (fig. 5) was again circular, with minor distribution in the midst of the circle, which represents genomic rearrangements. Such circular figure can be regarded as a consensus genome that represents the two component genomes.

The isoapostatic genes were clustered essentially as in the case of 14 marine cyanobacteria (fig. 6). We used the silhouette widths in selecting the best clustering (fig. 6*B* and *C*), and the VLGs were colored accordingly (fig. 6*A*), which were remapped onto the circular consensus genome (fig. 5). The fragmented pattern of figure 6 shows what isoapostatic genes are. Not all genes in magenta, for example, form a gene cluster in a strict sense of the words (fig. 1*A*), but they keep similar spacing over different genomes as depicted in figure 1*D*.

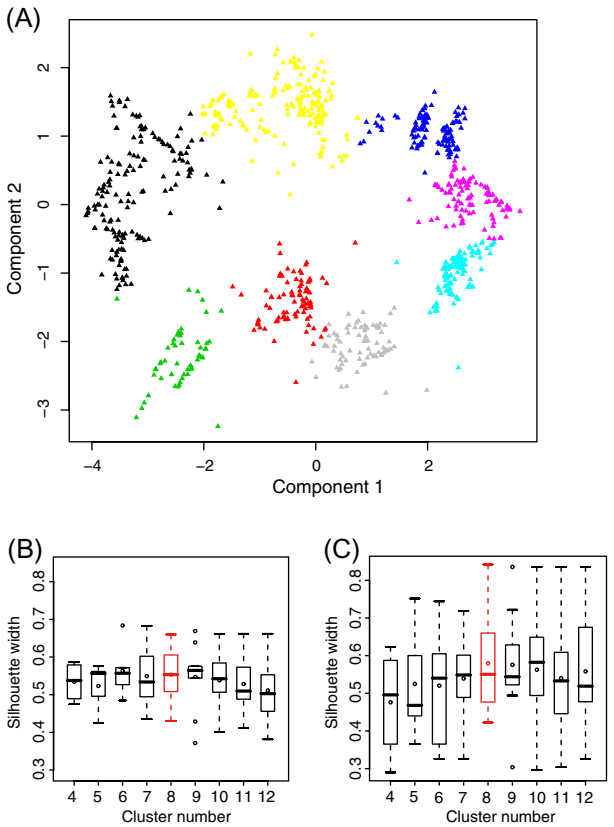

Fig. 4.—Clustering of orthologs into VLG in 14 marine cyanobacterial genomes. (*A*) Embedding of orthologs using a variance measure. The colors indicate VLG classification. Circular arrangement of objects is less clear but still recognizable in this figure. Each colored cluster corresponds to a sector of the circle in figure 3. (*B*) Box plot of average silhouette widths of clusters that were calculated for the results in figure 3. Notches on both sides indicate the maximum and minimum scores, respectively. Open boxes indicate interquartile ranges. A bold line in each box indicates the median. Each open circle indicates the mean. Open circles outside the notches indicate outliers. (*C*) Box plot of average silhouette widths that were calculated using the variance measure corresponding to the results in (*A*). The silhouette widths were calculated by the pam function of the R program.

In this case, the relationship of figure 5 and figure 6 was more complex than the relationship of figure 3 and figure 4. The general arrangement of the eight colors was similar in figure 5 and figure 6, but the arrangement of the data points looked quite different. Again, the result of isoapostatic analysis (fig. 6) was better suited for clustering. For example, the black and yellow points formed a single arc in figure 5, but they were distinctly separated in figure 6. This indicates that the isoapostatic relationship extends over wider regions of the genome in *Anabaena*.

Statistical Evaluation of Embedding by Multidimensional Scaling To evaluate the embedding of the objects, we checked the effect of dimension of multidimensional scaling on the clustering (table 2). In the Euclidean distance

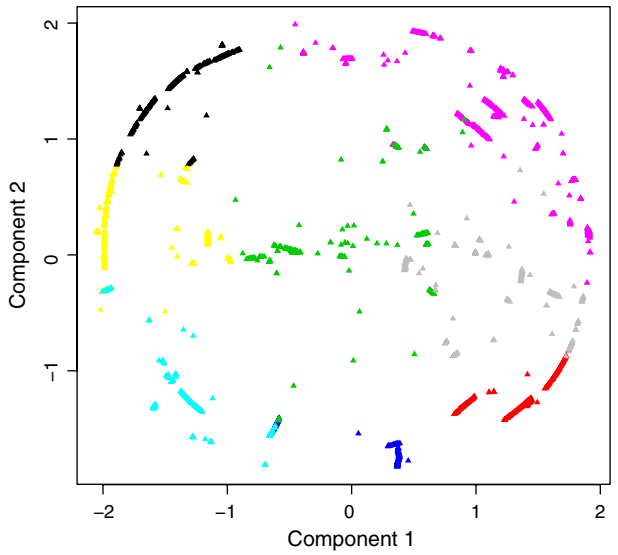

Fig. 5.—Embedding of orthologs in *Anabaena* PCC 7120 and ATCC 29413 genomes by multidimensional scaling using Euclidean distance. Points indicate conserved orthologs and the color of each object indicates classification of VLG.

data, the stress values of 1D and 2D embeddings were rather high but were reduced in three dimensions or above in both cases (table 2A). This suggests that the data structure of the distance relationships of the objects is best represented in three dimensions. table 2B shows the stress values of the variance data. Note that the stress value changed abruptly between dimensions 1 and 2 but gradually decreased at higher dimensions. This shows that reduction in dimensions to eliminate the noise is not necessary in the case of the variance data.

The contribution of genomic position of orthologs to the result of multidimensional clustering (fig. 4) was checked by linear multiple regression analysis using the lm function of the R program. table 3A shows the results of analysis with Euclidean distance. In table 3, 14 genomes were classified into high correlation group (MED4, MIT9312, MIT9301, NATL2A, MIT9515, NATL1A, and AS9601) and low correlation group (MIT9313, CCMP1375, MIT9303, WH8102, CC9311, CC9605, and CC9902) by combined correlation coefficient. The ecotypes in the high correlation group are close relatives within the genus *Prochlorococcus*, and the circular map (fig. 3) was interpreted to reflect mainly the gene positions in these cyanobacteria. Decomposition of the coefficient showed that the high correlation coefficient was mainly the result of high correlation coefficient of the argument component, whereas the features of low correlation group are represented in the radius component. table 3B shows the result of analysis using variance statistics. This showed a similar trend as in table 3A, though the combined correlation was slightly smaller. The outlier points in the midst of the circle were not those that lost their

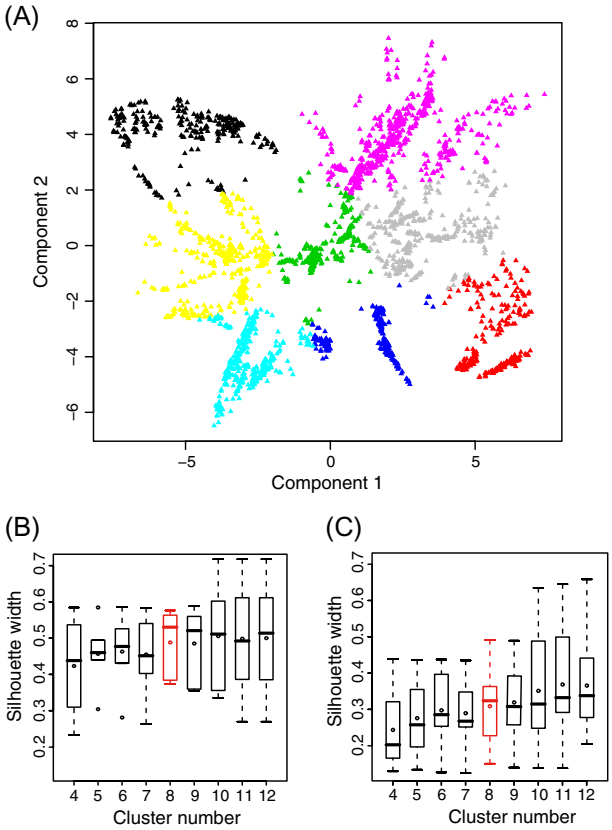

Fig. 6.—Clustering of orthologs into VLG in *Anabaena* PCC 7120 and ATCC 29413 genomes. (*A*) Embedding of ortholog genes using a variance measure. The colors indicate VLG classification. The VLG classification are almost corresponding to the wings of the plot shape. (*B*) and (*C*) are box plots of silhouette widths of clusters as shown in figure 4*B* and *C*. Scores in (*C*) are lower than (*B*), and it is difficult to determine the best number of clustering by the distribution. We chose the best number of clusters as four from the score distribution in (*B*) (indicated by red).

positional information but rather represented their conflicting information in radius direction, which could be represented by the third dimension.

These results justify that the arrangement of the data points can be considered to represent the consensus genome as stated above. The contribution of each genome to the consensus genome might reflect evolutionary relationship of the cyanobacteria. Figure 7 shows a phylogenetic tree of the 14 marine cyanobacteria based on the combined 16S and 23S rRNA sequences. *Synechococcus elongatus* PCC6301 was taken as an outgroup. The 14 species of marine cyanobacteria were split into the *Prochlorococcus* clade and the marine *Synechococcus* clade. These clades agreed with the grouping of high correlation group and low correlation group in table 3.

Position of the Isoapostatic Genes within the Genome To find the relationship of VLGs and the gene clusters in the real genomes, the VLGs were mapped back onto

Table 2
Stress Values in Different Dimensions

Dimensions	*Anabaena*	Marine
A. Euclidean distance data		
1	43.5	38.8
2	24.0	16.6
3	8.45	11.8
4	4.88	8.51
5	4.90	8.43
B. variance data		
1	55.0	43.9
2	42.5	26.4
3	33.0	25.9
4	28.5	20.6
5	26.6	18.4

NOTE.—Stress values in embedding into different dimensions by multidimensional scaling were calculated for each of Euclidean distance data (A) and variance data (B). *Anabaena*, comparison between *Anabaena* PCC 7120 and ATCC 29413; Marine, comparison within 14 marine cyanobacteria.

Table 3
Multi-regression Analyses on the Embedded Space for the Marine Cyanobacterial Data.

Species	Correlation (radius)	Correlation (argument)	Correlation (combined)	Lag
A. Results with Euclidean distance data				
Prochlorococcus marinus MIT9312	0.096	0.993	0.993	0.13
P. m. AS9601	0.094	0.993	0.993	0.12
P. m. MIT9515	0.095	0.993	0.993	0.12
P. m. MED4	0.111	0.993	0.993	0.13
P. m. MIT9301	0.096	0.990	0.990	0.12
P. m. NATL2A	0.113	0.975	0.975	2.04
P. m. CCMP1375 (SS120)	0.115	0.972	0.972	0.15
P. m. NATL1A	0.097	0.969	0.969	0.14
Synechococcus CC9605	0.391	0.403	0.493	1.04
S. CC9311	−0.654	0.017	0.468	4.39
P. m. MIT9303	−0.645	0.003	0.464	−3.96
S. WH8102	0.415	0.344	0.453	−0.86
P. m. MIT9313	−0.006	0.437	0.441	1.32
S. CC9902	0.370	0.356	0.438	−1.17
B. Results with variance data				
Prochlorococcus marinus MIT9312	−0.076	0.888	0.888	3.20
P. m AS9601	−0.072	0.882	0.882	3.18
P. m. MIT9301	−0.062	0.880	0.880	3.15
P. m. NATL2A	0.095	−0.874	0.875	−0.98
P. m. NATL1A	−0.107	0.860	0.861	3.52
P. m. MIT9515	−0.287	0.766	0.800	3.31
P. m. MED4	−0.266	0.771	0.799	3.26
P. m. CCMP1375 (SS120)	0.280	0.602	0.682	2.67
Synechococcus CC9605	0.661	−0.142	0.657	−2.66
S WH8102	−0.629	0.122	0.623	−3.37
S. CC9902	0.626	0.026	0.621	2.29
P. m. MIT9313	0.599	−0.072	0.590	3.09
S. CC9311	0.575	0.100	0.582	−1.97
P. m. MIT9303	0.512	0.179	0.548	2.30

NOTE.— Polar coordinates (radius and argument) were used in this calculation. Correlation coefficients for radius and argument components are shown separately and in combined form. Lag, difference of relative position between embedded space and genomic position.

the real genomes, and the orthologs were painted in corresponding colors (fig. 8). The three inner circles represent the genomes of *Prochlorococcus* MED4, AS9601, and CCMP1375 (high correlation group), respectively, and the outer two circles represent the genomes of *Synechococcus* WH8102 and CC9311 (low correlation group), respectively. The result indicated that the genes belonging to the same VLG were largely located in the neighborhood to form domains in the real genomes. A closer examination, however, revealed that the VLG pattern was different in different genomes, and the difference was obvious between the two cyanobacterial groups. The pattern in the high correlation group was more similar to the consensus genome shown in the center. The mosaic pattern might exhibit the history of genomic rearrangements, as it can be roughly traced in the figure. This is consistent with the previous report that traces of horizontal gene transfers were frequently found in variable regions in the genome (Coleman et al. 2006).

There were many cases in which a single cluster consisting of a single VLG in the genomes of the high correlation group was split into several different clusters in the genomes of the low correlation group. Detection of such variable clusters was a characteristic of our method. In this respect, isoapostasis is not complete for the genes in such clusters. However, we can also say that our method enabled successful detection of conservation of long distances between split clusters.

Discussion

Benefit of using Positional Profile Method for Cross-Genomic Analysis Because the scenario of genome rearrangement is closely related to phylogenetic relationship of genomes (Sankoff et al. 1992), conserved synteny has been used as a measure of genomic distance (Sankoff and Nadeau 1996). As first pointed out by Sankoff et al. (1992), detection of synteny blocks itself is part of the algorithm to calculate genomic distance. Kececioglu and Sankoff (1995) pointed out the importance of minimal reversal distance in assessing permutations of genome rearrangement. Although many improved algorithms using graph search (Tesler 2002) have been developed, the problem of combinatorial complexity still remain (Brudno et al. 2004), which is an obstacle for using synteny relationship as a phylogenetic marker in a large data set containing

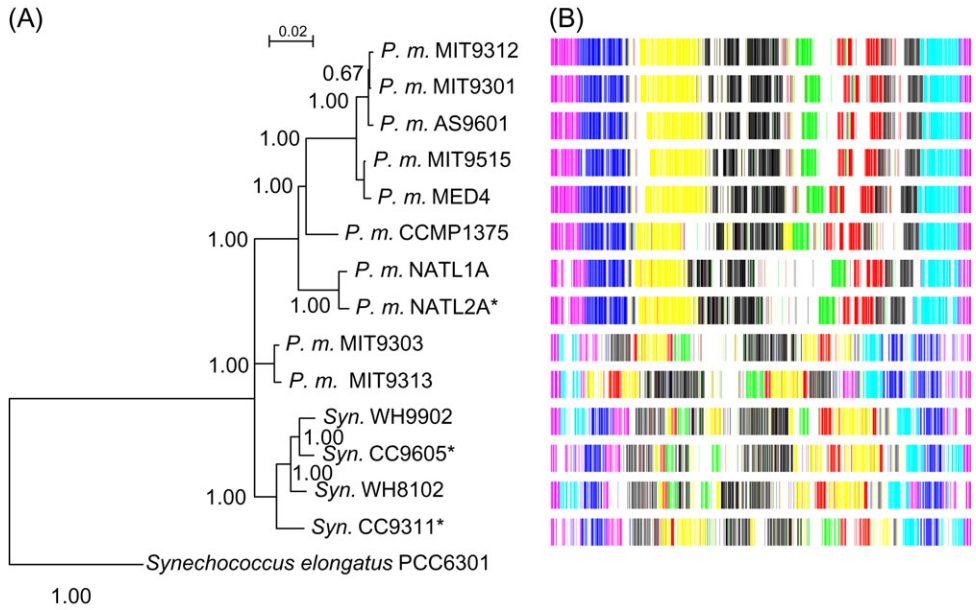

Fig. 7.—Phylogenetic tree in 14 marine cyanobacteria constructed by 16S and 23S rRNA. The tree was generated from Bayesian Inference method. *Synechococcus elongatus* PCC6301 was taken as an outgroup. The marine species are largely separated into the *Prochlorococcus* clade and the *Synechococcus* clade and MIT9312, MIT9301, AS9601, MIT 9515, and MED4 formed a rigid cluster. The posterior probability of each branch is shown. The branch of MIT9303 and MIT9313 was not resolved.

genomes of diverse organisms. Genome comparison within various strains of a bacterial species, such as *E. coli* (Ohnishi et al. 2002; Dobrindt 2005) and *Staphylococcus* (Takeuchi et al. 2005; Feng et al. 2008), revealed that a definite genomic backbone is present in these genomes, and a limited number of insertions of pathogenic islands and some deletions have been identified. Synteny relationship has also been used to estimate genomic core of marine ecotypes of a cyanobacterial species *P. marinus* (Kettler et al. 2007). But the genomic variation of various ecotypes of *P. marinus* is larger than their sequence-level diversity. This situation suggests that *P. marinus* is a good example of studying genomic rearrangement because identification of ortholog relationship is easy, but there are many rearrangements as well as indels due to horizontal gene transfer.

In the present study, we tested a statistical approach rather than combinatorial approach to analyze genomic rearrangements having many indels. The results showed that statistical method is capable of detecting global synteny relationship in genomes of related but significantly diversified organisms. Two types of multivariate analyses were tested: analysis using simple gene-to-gene distance and that using a variance measure. The latter was introduced as inspired by the leading notion of isoapostasy. First, multivariate analysis using gene-to-gene distances of orthologs reconstructed a consensus genome represented by a circular arrangement of orthologs or a virtual genome in a feature space. In this virtual genome, neighborhood relationships of orthologs are faithfully reconstructed (fig. 3). However, the border of syntenic blocks was not clear. Next, we tested multivariate analysis using a variance measure. The resultant virtual genome did not look like a simple circle, but units of rearrangement blocks were separated from each other (fig. 4). This comparison indicated that the variance measure is better suited to detect rearrangement blocks. The reason is simple: As seen in fig. 2, the variance measure is an invariant shared by an entire synteny block. In the real comparison of various genomes, indels and rearrangements occur in subgroups of genomes. Small indels, sporadic or systematic, have generally a small effect of increasing the variance value for the pairs of orthologs located on both sides of the indels. This might be the reason why syntenic genes are arranged in small ellipsoids in figures 4 and 6.

Rearrangements or large indels introduce large positive terms in the variance calculation of orthologs affected by the changes for the genomes belonging to different subgroups, but the contributions corresponding to the relationship within each subgroup remain zero. As a result, the synteny block is split into two parts in the feature space, but the two parts are still located in the vicinity. This is the situation seen in figures 4 and 6. It is possible to give different identifiers to the two parts, but it is convenient to classify them as a large group to emphasize such rearrangements in the real genomes. That is why we classify the syntenic clusters into a minimal number of VLGs. Evaluation of the clustering results (supplementary fig. S1 and supplementary table 1, Supplementary Material online) also suggested that the VLGs were appropriately clustered.

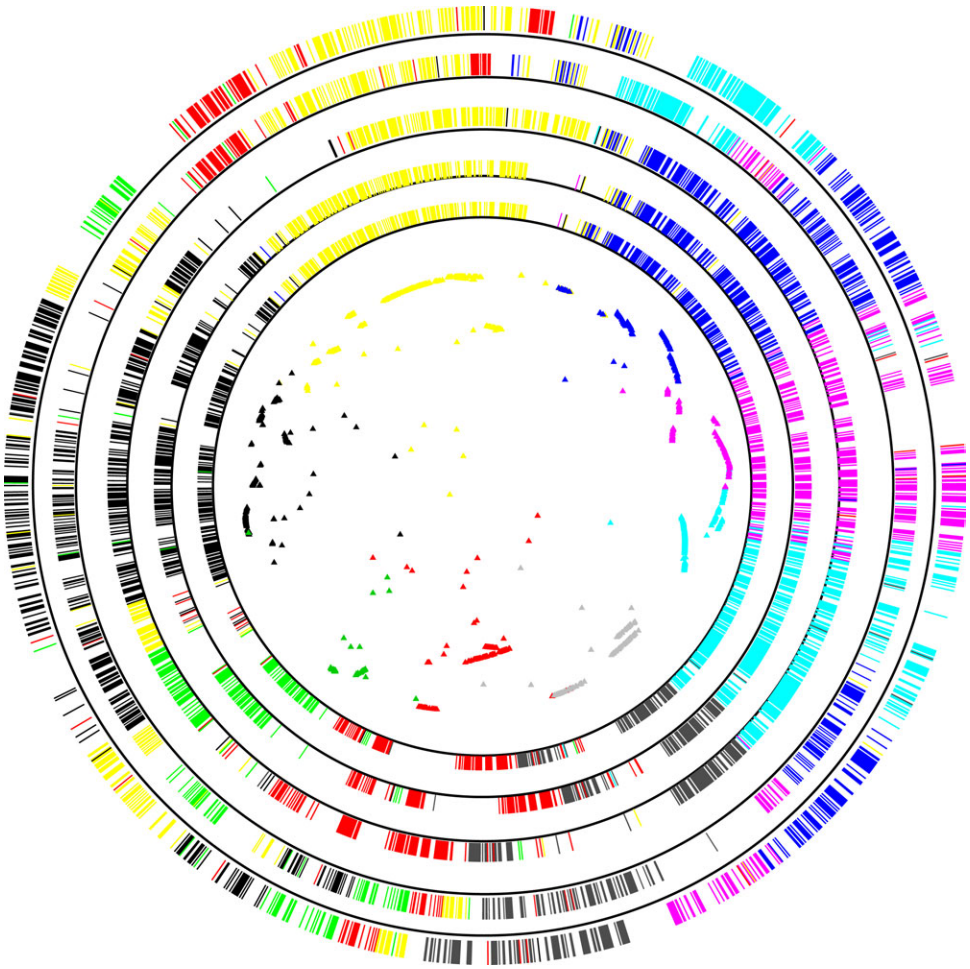

Fig. 8.—Localization of VLG on five selected marine cyanobacterial genomes. From the innermost ring to the outermost one, *Prochlorococcus* MED4, AS9601, CCMP1375 and *Synechococcus* WH8102 and CC9902 are shown. The eight VLGs are indicated by colors according to those in figure 3. The genomic circles are rotated so that the VLG positions roughly agree with those in the consensus genome (center of the figure). Most genomes of the *Prochlorococcus* clade are closely related and the domain patterns are similar. But the patterns are different between the *Prochlorococcus* and the *Synechococcus* clades.

That is the reason why the isoapostasy method can detect rearrangement blocks without definition of gap scores. This is a significant feature of our method involving statistics of long-distance terms over various graph-search methods, such as the alignment reduction (Sankoff et al. 1992) or anchor-base method such as GRIMM synteny method (Pevzner and Tesler 2003).

Mosaic Structure of Marine Cyanobacteria and Its Evolution Mosaic color patterns of the VLGs in marine cyanobacteria exhibit two prominent characteristics of the structure of these genomes, namely mosaic pattern of VLGs and genomic islands. In figure 8, we can see mosaic pattern of VLGs that constituted a stable common structure. The VLG pattern of each genome was generally consistent with the phylogenetic relationships obtained by molecular phylogenetic tree of rRNA (fig. 7). Because the VLG patterns were generated without guide tree, these patterns are not biased with the rearrangement scenario. VLG patterns may be a good marker of phylogenetic inference because Sankoff and Nadeau (1996) pointed out that rearrangement of conserved synteny should be used as a measure of genomic distance.

We also noted in figure 3 that blank regions interrupt continuation of a single VLG to produce fragmented synteny blocks. Blank regions also exist between different VLGs. These regions are regarded as genomic islands representing laterally transferred genes, which rarely contain highly conserved orthologs. Previous studies reported the presence of many insertion islands within the stable common structure formed by orthologs. The structure of genomic island is often linked to their functions such as the pathogenic islands in *E. coli* (Dobrindt 2005), *Staphylococcus aureus* (Feng et al. 2008) and *Streptococcus* (Brochet et al.

2008; Rasmussen et al. 2008), or symbiotic islands in nodulating α-proteobacteria (Kaneko et al. 2000; Sullivan et al. 2002). In cyanobacteria, expressed islands involved in heterocyst differentiation have been described (Ehira et al. 2003). Genomic islands of laterally acquired genes have also been reported in marine *Synechococcus* (Dufresne et al. 2008). Our results are, in general, consistent with their results, although the two methods are totally different. This does not mean that our method is useless. The reported results were obtained within the limit of graph searching algorithm, and it will be difficult to use the same methodology to more diversified genomes. The results of figure 6 indicated that our method is applicable to comparison of significantly diversified organisms, namely two different strains of *Anabaena*. They belong to the same genus, but the physiological properties are very different, and the genomes are highly rearranged with many large insertions. The isoapostasy method will be applicable to more diversified organisms because, as stated above, our method includes statistics of long-distance relationship and is more robust to rearrangements and indels.

Supplementary Material

Supplementary figure S1 and supplementary table 1 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

This work was partly supported by Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows (2011425 to N.V.S.), by Grant-in-Aid for Priority Areas (nos. 18017005 and 20017006 to N.S.) from the MEXT, Japan, and by Grant-in-Aid for Creative Scientific Research (16GS034 to N.S.) from JSPS. Computation was done in the Super Computer System, Human Genome Center, Institute of Medical Science, the University of Tokyo.

Literature Cited

Becker RA, Chambers JM, Wilks AR. 1988. The new S language: a programming environment for data analysis and graphics. Pacific Grove (CA): Wadsworth & Brooks/Cole Advanced Books & Software.

Brochet M, et al. 2008. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. Proc Natl Acad Sci U S A. 105:15961–15966.

Brudno M, et al. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721–731.

Brudno M, et al. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. Genome Res. 14:685–692.

Cole JR, et al. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37:D141–D145.

Coleman ML, et al. 2006. Genomic islands and the ecology and evolution of Prochlorococcus. Science. 311:1768–1770.

Cox TF, Cox MAA. 2001. Multidimensional scaling. Boca Raton (FL): Chapman & Hall/CRC.

Dobrindt U. 2005. (Patho-)genomics of *Escherichia coli*. Int J Med Microbiol. 295:357–371.

Doolittle WF. 1982. Molecular evolution. In: Carr NG, Whitton BA, editors. The biology of cyanobacteria. Berkeley (CA): University of California Press. p. 307–332.

Dufresne A, et al. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol. 9:R90.

Ehira S, Ohmori M, Sato N. 2003. Genome-wide expression analysis of the responses to nitrogen deprivation in the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 10: 97–113.

Feng Y, et al. 2008. Evolution and pathogenesis of *Staphylococcus aureus* and comparative genomics. FEMS Microbiol Rev. 32:23–37.

Johnson ZI, et al. 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science. 311:1737–1740.

Kaneko T, et al. 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res. 7: 331–338.

Kaufman L, Rousseeuw PJ. 1990. Finding groups in data: an introduction to cluster analysis. New York: Wiley.

Kececioglu J, Sankoff D. 1995. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. Algorithmica. 13:180–210.

Kettler GC, et al. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet. 3:2515–2528.

Ohnishi M, et al. 2002. Genomic diversity of enterohemorrhagic *Escherichia coli* O157 revealed by whole genome PCR scanning. Proc Natl Acad Sci U S A. 99:17043–17048.

Pevzner P, Tesler G. 2003. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res. 13:37–45.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.

Rasko DA, et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of E-coli commensal and pathogenic isolates. J Bacteriol. 190:6881–6893.

Rasmussen TB, et al. 2008. *Streptococcus thermophilus* core genome: comparative genome hybridization study of 47 strains. Appl Environ Microbiol. 74:4703–4710.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. 2002. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol. 68:1180–1191.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rousseeuw PJ. 1987. Silhouettes—a graphical aid to the interpretation and validation of cluster-analysis. J Comput Appl Math. 20: 53–65.

Sankoff D, et al. 1992. Gene order comparisons for phylogenetic inference—evolution of the mitochondrial genome. Proc Natl Acad Sci U S A. 89:6575–6579.

Sankoff D, Nadeau JH. 1996. Conserved synteny as a measure of genomic distance. Discrete Appl Math. 71:247–257.

Sankoff D, Sundaram G, Kececioglu J. 1996. Steiner points in the space of genome rearrangements. Int J Foundation Comput Sci. 7:1–9.

Sato N. 2000. SISEQ: manipulation of multiple sequence and large database files for common platforms. Bioinformatics. 16:180–181.

Sato N. 2009. Gclust: trans-kingdom classification of proteins using automatic individual threshold setting. Bioinformatics. 25:599–605.
Sullivan JT, et al. 2002. Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. J Bacteriol. 184: 3086–3095.
Suyama M, Bork P. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17:10–13.
Takeuchi F, et al. 2005. Whole-genome sequencing of *Staphylococcus haemolyticus* uncovers the extreme plasticity of its genome and evolution of human-colonizing staphylococcal species. J Bacteriol. 187:7292–7308.
Tesler G. 2002. Efficient algorithms for multichromosomal genome rearrangements. J Comput Syst Sci. 65:587–609.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.
Uchiyama I. 2003. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. 31:58–62.
Whitton BA, Potts M. 2000. The ecology of cyanobacteria: their diversity in time and space. Dordrecht: Kluwer Academic.
Wilmotte A. 1994. Molecular evolution and taxonomy of the cyanobacteria. In: Bryant DA, editor. The molecular biology of cyanobacteria. Dordrecht: Kluwer Academic Publishers. p. 1–25.
Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 11:356–372.
Wuyts J, Perriere G, Van de Peer Y. 2004. The European ribosomal RNA database. Nucleic Acids Res. 32:D101–D103.