The Journal of Microbiology, March 2004, p.9-13 Vol. 42, No. 1


# A Method for Comparing Multiple Bacterial Community Structures from 16S rDNA Clone Library Sequences

**Inae Hur[1] and Jongsik Chun[1,2]***

*[1]Interdisciplinary Program in Bioinformatics and*
*[2]School of Biological Sciences,Seoul National University, 56-1 Shillim-dong, Kwanak-gu, Seoul 151-742, Republic of Korea*

(Received November 17, 2003 / Accepted January 30, 2004)

**Culture-independent approaches, based on 16S rDNA sequences, are extensively used in modern microbial ecology. Sequencing of the clone library generated from environmental DNA has advantages over fingerprint-based methods, such as denaturing gradient gel electrophoresis, as it provides precise identification and quantification of the phylotypes present in samples. However, to date, no method exists for comparing multiple bacterial community structures using clone library sequences. In this study, an automated method to achieve this has been developed, by applying pair wise alignment, hierarchical clustering and principle component analysis. The method has been demonstrated to be successful in comparing samples from various environments. The program, named CommCluster, was written in JAVA, and is now freely available, at http://chunlab.snu.ac.kr/commcluster/.**

***Key words*:** bioinformatics, microbial community, 16S rRNA, JAVA, hierarchical clustering

The use of 16S rDNA has been the cornerstone of microbial ecology, and has led to a wealth of information concerning prokaryotic diversity (Amann, 2000; Theron and Cloete, 2000). Microbial community structure analysis, based on 16S rDNA, has given an understanding of functional and biogeographical relationships, and such data vital for an improved understanding of ecosystem processes and the role bacteria play in various environments. Several methods have been developed for determining bacterial community structures. Among these, denaturing gradient gel electrophoresis (DGGE) and terminal restriction fragment length polymorphisms (TRFLP) provide a rapid and overall view of community structures (Muyzer, 1999; Osborn *et al.*, 2000). However, these are basically fingerprint methods, and generate only limited and incomplete information, and may be problematic, especially when a sample with a highly diverse community structure is considered. In addition, an interlaboratory comparison is not possible. In contrast, sequencing of 16S rDNA from clone libraries, of DNAs from environmental samples, provides an alternative means of elucidating bacterial community structures, as each phylotype can be precisely identified and quantified.

Analysis based on the 16S rDNA clone library can be laborious in producing a number of sequences large enough to cover a whole community. The technique is also limited by the difficulty to compare libraries, and in determining if they are significantly different. Hierarchical clustering and ordination methods have been used to compare gel electrophoretic fingerprints, and exhibit community structures of bacteria and viruses (Amp and Miambi, 2000; el Fantroussi *et al.*, 1999; Wommack *et al.*, 1999). However, no method or software tool is available for comparing multiple bacterial community structures based on the 16S rDNA clone library. In this study, an automated software tool for comparing multiple samples was developed, using sequences in clone libraries. The effectiveness of this method was demonstrated with an analysis of various bacterial communities collected from the Genbank database.

## Materials and Methods

### *Algorithm*

The scheme of data analysis used in this study is summarized in Fig. 1, and was comprised of the following steps: (i) A sequence similarity matrix was generated from sequences: Each operational taxonomic units (OTU), represented as sequences, were pair wise aligned, using a modified dynamic programming algorithm (Wheeler and Hughey, 2000). Multiple sequence alignment of the OTUs was not possible, due to the extensive computing costs. The resultant sequence similarity matrix was saved for the next step. (ii) Hierarchical clustering analysis of the OTUs: hierarchical clustering was achieved using the unweighted pair group method, with the arithmetic mean (UPGMA; Sneath and Sokal, 1973).

* To whom correspondence should be addressed.
(Tel) 82-2-880-8153; (Fax) 82-2-888-4911
(E-mail) jchun@snu.ac.kr

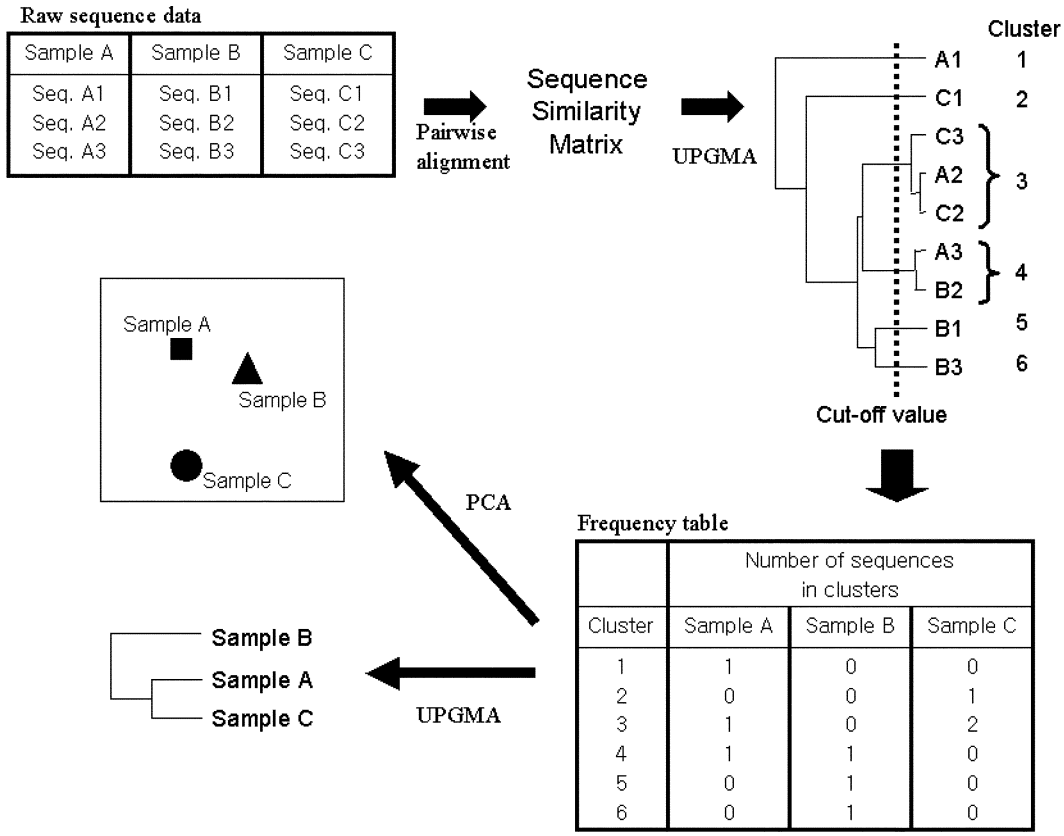


**Fig. 1.** Major steps in the CummCluster program.

**Table 1.** Dataset 1 used in this study

| Sample | Description | Geographical origin | Number of sequences | Reference |
|---|---|---|---|---|
| Getbol | Intertidal flat sediment (0~10cm) | Ganghwa, Korea | 103 | Kim *et al.*, 2004 |
| Savannah-RCP1 | Surface soil in contaminated region close to the coal pile | Savannah River Site, USA | 29 | Brofft *et al.*, 2002 |
| Savannah-RCP2 | Surface soil 60m from Savannah-RCP1 | Savannah River Site, USA | 35 | Brofft *et al.*, 2002 |
| Savannah-FW | Surface soil from an unaffected region by coal pile | Savannah River Site, USA | 85 | Brofft *et al.*, 2002 |
| Georgia-Soil | Agricultural Soil | Georgia, USA | 95 | Furlong *et al.*, 2002 |
| Georgia-Earthworm | Earthworm casts | Georgia, USA | 102 | Furlong *et al.*, 2002 |
| Tokoy-02 | Marine sediment (0~2cm) | Tokyo Bay, Japan | 37 | Urakawa *et al.*, 1999 |
| Tokoy-68 | Marine sediment (6~8cm) | Tokyo Bay, Japan | 36 | Urakawa *et al.*, 1999 |

(iii) A cutoff similarity value, provided by the user, was used to define the clusters in the UPMGA dendrogram. The OTUs belonging to each cluster were identified, and used to generate frequency tables containing the numbers of OTUs in each cluster, including single-membered clusters, for each sample. Each cluster is considered as a variable. (iv) Comparing samples using the hierarchical clustering and ordination methods: The frequency table was used to generate a distance or similarity matrix, using Pearsons correlation or Dice coefficients (Sneath and Sokal, 1973). The resultant matrix was subsequently used for clustering samples (sequence sets), using the UPMGA. An ordination analysis was performed, using principle component analysis (PCA; Sneath and Sokal, 1973), and the resultant principle components plotted in two or three dimensional spaces.

***Test datasets***

Two different datasets were used in this study. Dataset 1 consisted of sequences from various samples (Table 1), whereas dataset 2 contained sequences from samples collected from the Changjiang River, and associated lakes in China (Table 2).

***Program implementation***

A computer program, named CommCluster, was written

**Table 2.** Dataset 2 used in this study. All samples were collected from the Changjiang River, and associated lakes in China (Sekiguchi *et al.*, 2002)

| Sample | Description | Number of sequences |
|---|---|---|
| CR99-2 | Sampled at site 2 in 1999 | 76 |
| CR98-5 | Sampled at site 5 in 1998 | 76 |
| CR99-7 | Sampled at site 7 in 1999 | 76 |
| CR99-D | Sampled at Lake Dongting 24 in 1999 | 76 |
| CR99-P | Sampled at Lake Poyang 24 in 1999 | 76 |
| CR98-24 | Sampled at site 24 in 1998 | 76 |
| CR99-24 | Sampled at site 24 in 1999 | 76 |
| CR98-35 | Sampled at site 35 in 1998 | 76 |
| CR99-35 | Sampled at site 35 in 1999 | 76 |

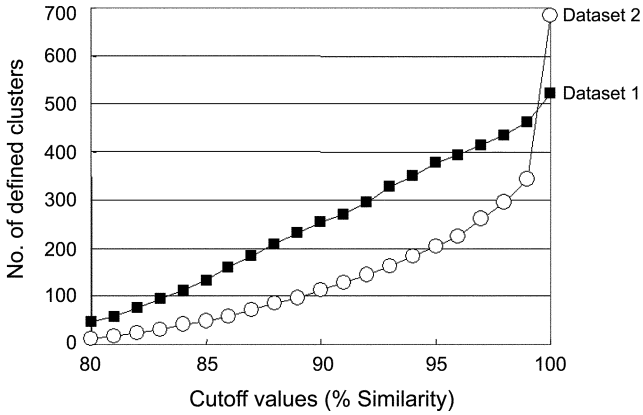


**Fig. 2.** Relationship between the cutoff values and the numbers of clusters defined.

in JAVA, which allowed portability across multiple computer platforms. The Java implementation also provided the opportunity to quickly create a web-based interface, through the use of Java servlets and jsp pages. CommCluster required nucleotide sequences, from multiple sets, as input data, and a cutoff similarity value for defining clusters from the UPGMA dendrogram. An output can also be generated for other computer programs, such as the NTSYS program (Exeter Software).

***Program availability***

The CommCluster program and test datasets are freely available from the web site: http://chunlab.snu.ac.kr/commcluster/.

## Results and Discussion

In this study, two datasets were used to demonstrate and validate the CommCluster method. Dataset 1 consisted of 522 16S rDNA sequences from 8 samples, collected from soils and sediments in the different geographical locations. Due to the extensive computing costs, multiple sequence alignment could not be applied. To assign an adequate cutoff similarity value for the cluster definition, the relationship between the cutoff values and the number of defined clusters was analyzed (Fig. 2).

The number of clusters defined by the cutoff similarity values in dataset 1 gradually decreased, whereas a sharp decrease, between 99 and 100%, in the cutoff values was noticed for dataset 2. This result implies there were more sequences with similarity to each other, with a difference of less than 1%, in dataset 2 than dataset 1. This was to be expected, as the sequences in dataset 2 were from one geographical area, that being the Changjiang River and associated lakes, whereas those in dataset 1 originated from worldwide sources.

Pearsons correlation and Dice coefficients were used to perform cluster analyses on the resultant frequency tables. The former utilizes the quantity information of each variable (i.e. cluster), but the latter does not.

The cluster analyses of the samples in dataset 1 are summarized in Fig. 3. Three different cutoff values were used, with the 97, 90 and 80% cutoff values roughly defining the clusters corresponding species, family and division, respectively. Using Pearsons correlation coefficients, different dendrograms were recovered from the different cutoff values. Three marine sediment samples; Getbol, Tokyo-02, Tokyo-68), as well as three soil samples from the Savannah River sites, were clustered in the 90 and 80% dendrograms, respectively. The soil sample from Georgia (Georgia-Soil) was only clustered with the earthworm casts sample from the same site (Georgia-Earthworm) in the 90% cutoff dendrogram. Overall, the 90% cutoff value, which represents the family-level grouping, was successful in clustering the samples from the same geographical origins. Dendrograms based on the Dice coefficients showed similar groupings, and the 90% cutoff dendrogram showed identical topology to that based on the Pearsons correlation coefficient. The 90% cutoff value may be a good choice for comparing the physiological components in a bacterial community, as the organisms belonging to the same family, or sharing 90%, or higher 16S rDNA similarity, are likely to have the same metabolism and physiology (e.g. sulfate-reducing bacteria).

Dataset 2 was examined using PCA. The sampling sites for the Changjiang River, were located as an inland-2-5-7-D-P-24-35-sea sequence (Sekiguchi *et al.*, 2002). The

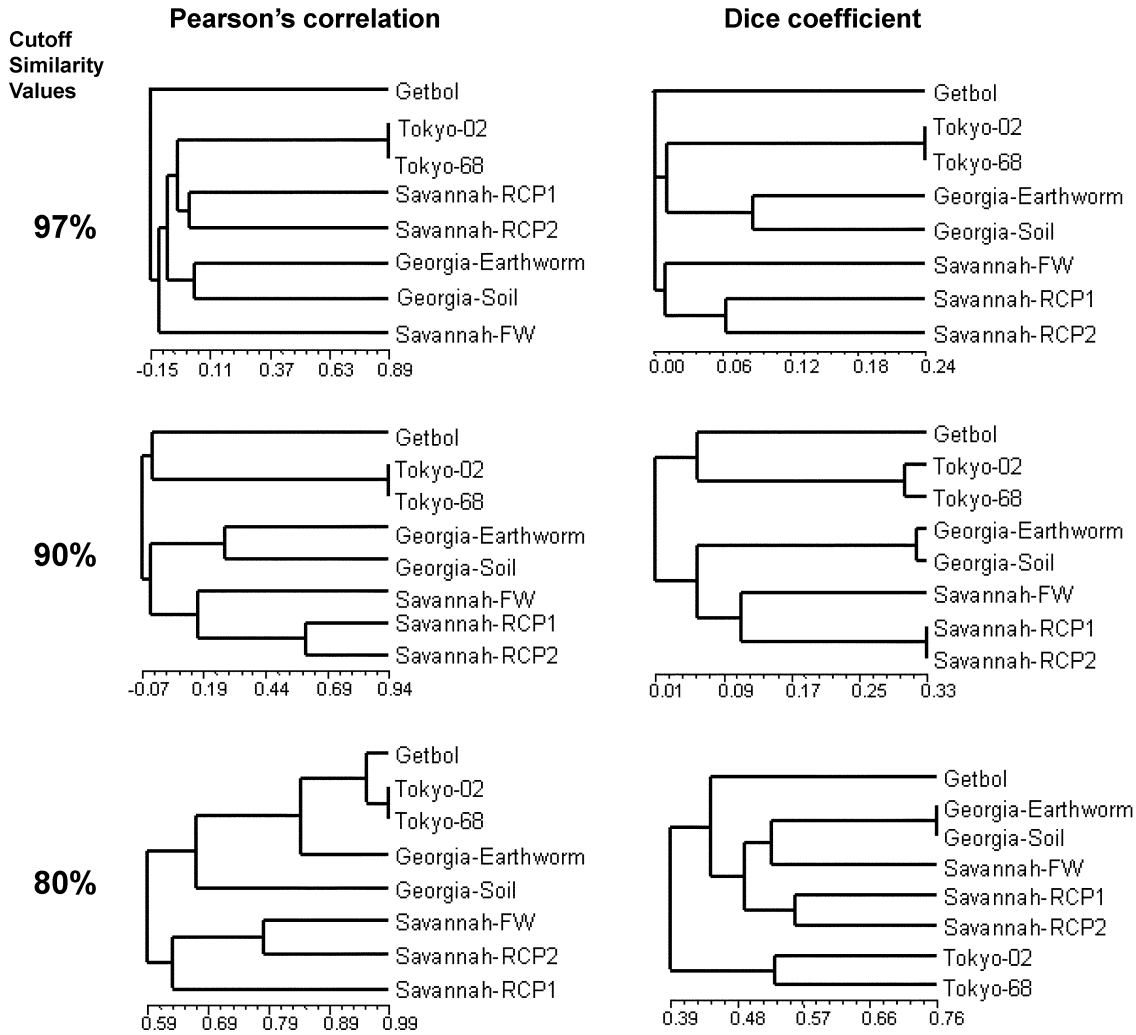


**Fig. 3.** Hierarchical clustering analysis of dataset 1. Samples were clustered using different cutoff similarity values, Pearsons correlation and Dice coefficients.

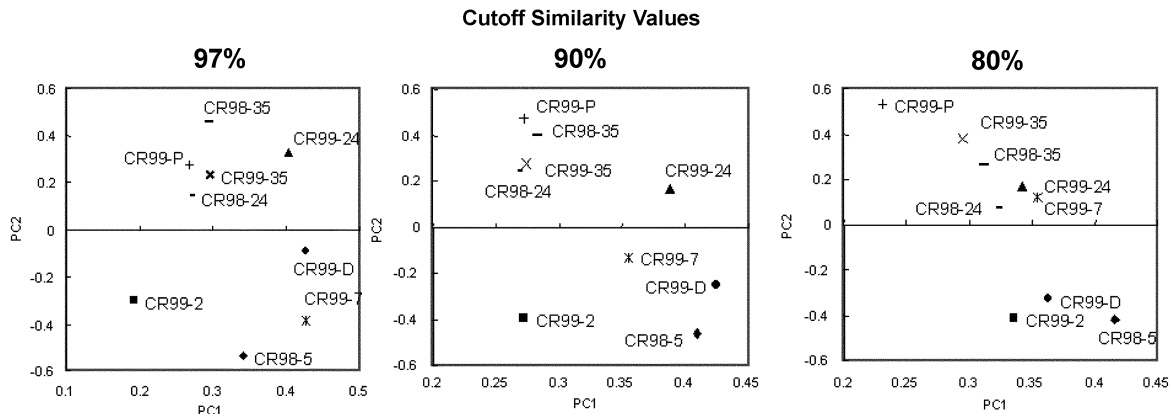


**Fig. 4.** Ordination diagrams, based on PCA analysis of dataset 2. The PCA analysis was performed using frequency tables defined at different cutoff similarity values.

sites, which were close to each other, were also found to be closely associated in the ordination diagram based on the first two principal components (Fig. 4). It was evident from the clone library sequences that the samples from the same sites were always clustered together, even though they had been sampled in different years. In general, the samples from the upper reaches of the Changjiang River, namely sites 2, 5 and 7, and the associated lake, Dongting,

were clustered together in the PCA ordination diagram. Similarly, samples collected from the lower river, namely sites 24 and 35, and the associated lake, Poyang, were found to be closely related in the PCA plot. The PCA result obtained from this study was similar to that based on the PCA analysis of the DGGE band patterns (Sekiguchi *et al.*, 2002). It would be fair to say that the method presented in this study was effective and comparable to those based on fingerprinting techniques, such as DGGE.

A comparison of the two datasets employed in this study was not possible, as they represent different regions in the 16S rDNA. Even though the method proposed in this study was useful when comparing multiple bacterial community structures, there were a few limitations. First, the parts sequenced were not consistent throughout the published studies. Some workers sequenced the 5'-end, whereas others sequenced the 3'-end or middle part. Secondly, only representative sequences have been deposited in the public database from most of the microbial ecological studies. This was particularly problematic, as our CommCluster approach requires all available sequences for the Pearsons correlation and PCA analyses.

The microbial community analysis has been extended to include eukaryotic microorganisms (Dawson and Pace, 2002), which will stay as the framework for the microbial ecology. Of the many methods available, clone library analysis has great potential as the cost of sequencing has been, and will be, continuously going down. It is clear from this study that the CommCluster approach provides a novel means of comparing multiple bacterial community structures. The program is freely available to the public, and should serve as an essential tool for understanding and elucidating the roles of microbes in various environmental niches.

## Acknowledgment

This study was supported by a grant (01-PJ11-PG9-01BT00B-0003) from the International Mobile Telecommunications 2000 R&D Project, the Ministry of Information & Communication, Republic of Korea.

## References

Amann, R. 2000. Who is out there? Microbial aspects of biodiversity. *Syst. Appl. Microbiol.* 23, 1-8.

Amp, F. and E. Miambi. 2000. Cluster analysis, richness and biodiversity indexes derived from denaturing gradient gel electrophoresis fingerprints of bacterial communities demonstrate that traditional maize fermentations are driven by the transformation process. *Int. J. Food. Microbiol.* 60, 91-97.

Brofft, J.E., J.V. McArthur, and L.J. Shimkets. 2002. Recovery of novel bacterial diversity from a forested wetland impacted by reject coal. *Environ. Microbiol.* 4, 764-769.

Dawson, S.C. and N.R. Pace. 2002. Novel kingdom-level eukaryotic diversity in anoxic environments. *Proc. Natl. Acad. Sci. USA* 99, 8324-8329.

el Fantroussi, S., L. Verschuere, W. Verstraete, and E.M. Top. 1999. Effect of phenyl urea herbicides on soil microbial communities estimated by analysis of 16S rRNA gene fingerprints and community-level physiological profiles. *Appl. Environ. Microbiol.* 65, 982-988.

Furlong, M.A., D.R. Singleton, D.C. Coleman, and W.B. Whitman. 2002. Molecular and culture-based analyses of prokaryotic communities from an agricultural soil and the burrows and casts of the earthworm *Lumbricus rubellus*. *Appl. Environ. Microbiol.* 68, 1265-1279.

Kim, B.S., H.M. Oh, H. Kang, S.S. Park, and J. Chun. 2004. Remarkable bacterial diversity in the tidal flat sediment as revealed by 16S rDNA analysis. *J. Microbiol. Biotechnol.* (14, 205-211.

Muyzer, G. 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Curr. Opin. Microbiol.* 2, 317-322.

Osborn, A.M., E.R. Moore, and K.N. Timmis. 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ. Microbiol.* 2, 39-50.

Sekiguchi, H., M. Watanabe, T. Nakahara, B. Xu, and H. Uchiyama. 2002. Succession of bacterial community structure along the Changjiang River determined by denaturing gradient gel electrophoresis and clone library analysis. *Appl. Environ. Microbiol.* 68, 5142-5150.

Sneath, P.H.A. and R.R. Sokal. 1973. *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco, CA: W. H. Freeman.

Theron, J. and T.E. Cloete. 2000. Molecular techniques for determining microbial diversity and community structure in natural environments. *Crit. Rev. Microbiol.* 26, 37-57.

Urakawa, H., K. Kita-Tsukamoto, and K. Ohwada. 1999. Microbial diversity in marine sediments from Sagami Bay and Tokyo Bay, Japan, as determined by 16S rRNA gene analysis. *Microbiology* 145, 3305-3315.

Wheeler, R. and R. Hughey. 2000. Optimizing reduced-space sequence analysis. *Bioinformatics* 16, 1082-1090.

Wommack, K.E., J. Ravel, R.T. Hill, J. Chun, and R.R. Colwell. 1999. Population dynamics of Chesapeake Bay virioplankton: Total-community analysis by pulsed-field gel electrophoresis. *Appl. Environ. Microbiol.* 65, 231-240.