Journal of Microbiological Methods 78 (2009) 344–350

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Application of a clustering-based peak alignment algorithm to analyze various DNA fingerprinting data

Satoshi Ishii [a,*], Koji Kadota [b], Keishi Senoo [a]

[a] Department of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Japan
[b] Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Japan

ARTICLE INFO

Article history:
Received 11 May 2009
Received in revised form 8 July 2009
Accepted 8 July 2009
Available online 17 July 2009

Keywords:
DNA fingerprinting
Complete linkage clustering
Peak alignment
Statistics
R program

ABSTRACT

DNA fingerprinting analysis such as amplified ribosomal DNA restriction analysis (ARDRA), repetitive extragenic palindromic PCR (rep-PCR), ribosomal intergenic spacer analysis (RISA), and denaturing gradient gel electrophoresis (DGGE) are frequently used in various fields of microbiology. The major difficulty in DNA fingerprinting data analysis is the alignment of multiple peak sets. We report here an R program for a clustering-based peak alignment algorithm, and its application to analyze various DNA fingerprinting data, such as ARDRA, rep-PCR, RISA, and DGGE data. The results obtained by our clustering algorithm and by BioNumerics software showed high similarity. Since several R packages have been established to statistically analyze various biological data, the distance matrix obtained by our R program can be used for subsequent statistical analyses, some of which were not previously performed but are useful in DNA fingerprinting studies.

1. Introduction

One-dimensional electrophoresis techniques such as polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP), repetitive extragenic palindromic PCR (rep-PCR), ribosomal intergenic spacer analysis (RISA), and denaturing gradient gel electrophoresis (DGGE), are powerful tools for use in molecular diagnostics, medical microbiology, epidemiological analyses, environmental microbiology, microbial ecology, and so forth. Amplified ribosomal DNA restriction analysis (ARDRA) is a form of PCR-RFLP analysis in which the ribosomal RNA gene is the target. While ARDRA and rep-PCR can characterize microbes at genus and subspecies/strain level, respectively (Chèneby et al., 2000; Ishii and Sadowsky, 2009; Lagacé et al., 2004), RISA and DGGE analyses enable us to compare complex microbial communities (Borneman and Triplett, 1997; Ikeda et al., 2007; Muyzer et al., 1993; Saito et al., 2008a). Since these techniques produce robust banding patterns of DNA, they are referred to as "DNA fingerprinting."

The major difficulty in one-dimensional electrophoretic data analysis is the alignment of multiple peak sets (i.e., peak matching). Although it may be possible to align multiple peaks by eye, it is subjective and labor intensive (Rademaker et al., 2008). In addition, between-gel comparison is almost impossible by visual evaluation of the DNA fingerprinting data. Several bioinformatics software applications are commercially available for gel analyses (e.g., BioNumerics and GelCompar, Applied Math, Sint-Martens-Latem, Belgium), but they are usually expensive and the types of analysis included are usually limited. Therefore, a free and flexible program is desirable for the analysis of DNA fingerprinting data.

Recently, complete linkage hierarchical clustering was successfully applied to analyze mass spectrometry data (Tibshirani et al., 2004; Yu et al., 2006) and differentially expressed fragments from cDNA-based amplified fragment length polymorphism (AFLP) (Kadota et al., 2007). This approach is simple and has a high throughput compared to other clustering algorithms (Kadota et al., 2007). In addition, visual evaluation of peak alignment can be easily performed (Kadota et al., 2007). However, the applicability of this algorithm to analyze other one-dimensional electrophoretic data (e.g., rep-PCR DNA fingerprinting data) has not been examined.

In this report, we developed an R program for a clustering-based peak alignment algorithm. Since several R packages have been established to statistically analyze various biological data, the distance matrix obtained by our R program can be used for subsequent statistical analyses. We applied this R program to analyze various DNA fingerprinting data, such as ARDRA, rep-PCR, RISA, and DGGE data, and compared the results obtained by our complete-linkage clustering algorithm with those obtained by BioNumerics software.

2. Materials and methods

2.1. Data sets

ARDRA analysis was performed to differentiate bacterial 16S rRNA gene fragments obtained from *Azospirillum brasilense* JCM1224^T,

* Corresponding author. Satoshi Ishii, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan. Tel.: +81 3 5841 5140; fax: +81 3 5841 8042.
E-mail address: anaerobe@mail.ecc.u-tokyo.ac.jp (S. Ishii).

0167-7012/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2009.07.005

Geobacillus kaustrophilus JCM12893, *Stenotrophomonas nitritireducens* JCM13311, *Ochrobactrum anthropi* JCM21032^T, and *Cupriavidus metallidurans* JCM21315^T. Conditions for DNA extraction and PCR amplification of the nearly full-length 16S rRNA gene were as described previously (Ashida et al., 2008). The PCR products were digested with restriction endonuclease *Afa* I (= *Rsa* I) (TaKaRa Bio, Otsu, Japan) according to the manufacturer's instructions. Digested DNA fragments were run on 2.5% agarose gel, stained with 0.5 μg ml^{-1} ethidium bromide, and visualized under UV light. The gel image was obtained using Printgraph GX CCD camera (ATTO, Tokyo, Japan).

ARDRA analysis was also performed *in silico*, using the 16 S rRNA gene sequences of the strains described above. The *Afa* I restriction sites were recognized and cut using NEBcutter ver. 2.0 (http://tools.neb.com/NEBcutter2/index.php). A pseudo-gel image was generated based on the restriction fragment lengths calculated.

Rep-PCR DNA fingerprinting, previously performed by Ishii et al. (2006) with ERIC1R and ERIC2 primers (Versalovic et al., 1991) (also called ERIC-PCR), was used in this study. The original data was used to characterize 38 *Campylobacter* strains isolated from *Cladophora* samples in Lake Michigan (Ishii et al., 2006). The DNA fingerprint patterns were previously analyzed using BioNumerics software ver. 2.5 (Ishii et al., 2006). By analyzing the same data, we could compare the results obtained by our clustering-based peak alignment algorithm with those obtained by BioNumerics software.

RISA experiment was previously conducted to compare microbial communities in soybean rhizospheres (Inaba et al., 2009). In their experiment, rhizosphere communities of the nodulating soybean cultivar Enrei and non-nodulating mutant En1282 were compared. Microbial communities around fresh and degraded nodules were also compared since degraded nodules produced significantly more N_2O than fresh nodules (Inaba et al., 2009). The intergenic spacer region of the bacterial 16S rRNA gene was used for RISA analysis as described previously (Ikeda et al., 2008; Inaba et al., 2009).

DGGE analysis was performed to compare bacterial communities in soil microcosms with or without denitrification stimulants. In sample TSNS, nitrate and succinate were added to soil obtained from a rice paddy field as an electron acceptor and donor, respectively, for denitrification. This condition, together with 24 h anoxic incubation at 30 °C, was previously shown to enhance soil-denitrifying activity (Saito et al., 2008b). Soil samples supplemented only with nitrate or

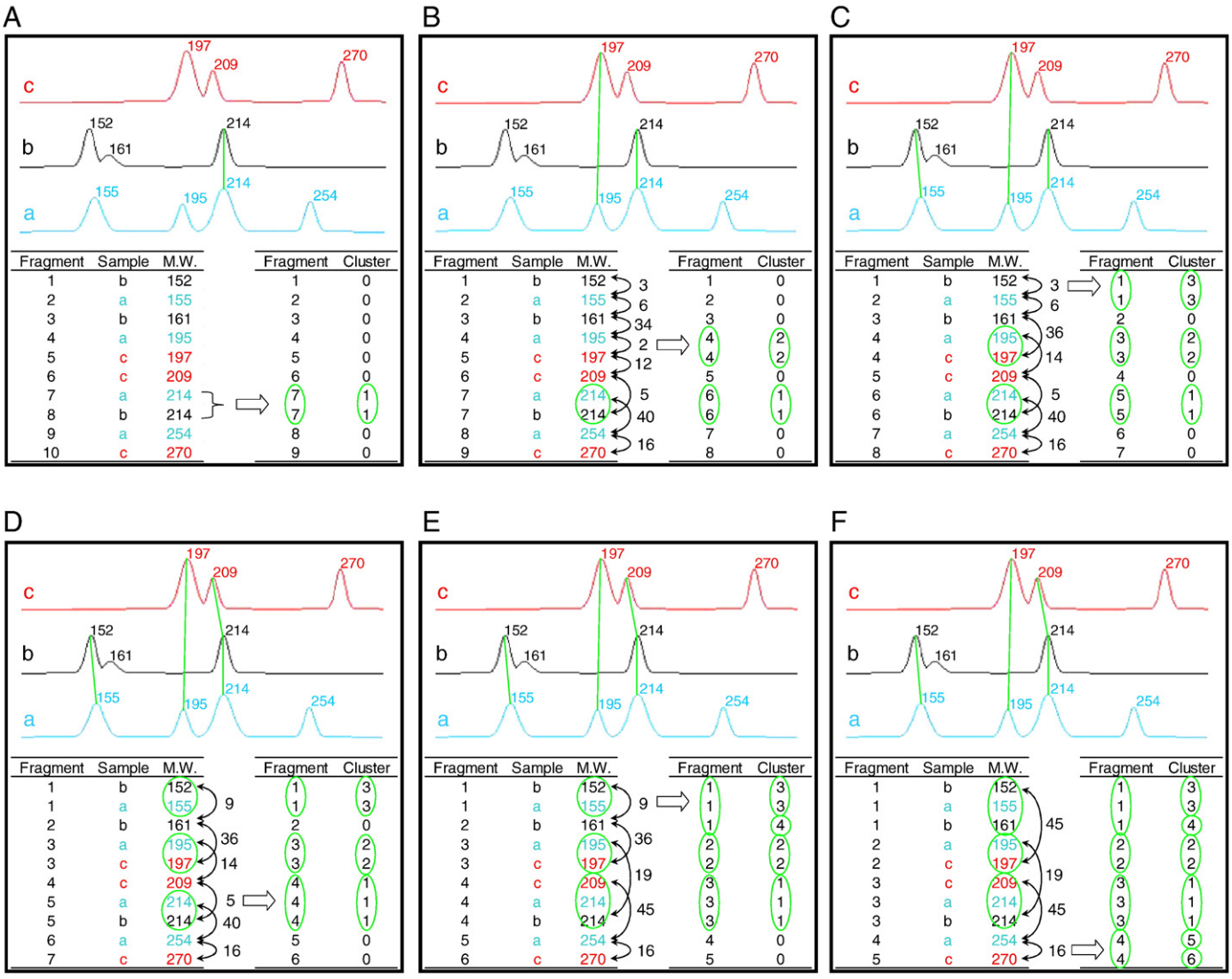

Fig. 1. Example of the clustering-based peak alignment algorithm used in this study. (A) Fragments with identical molecular weights (fragments 7 and 8) were merged, and a unique cluster number "1" was assigned. (B) Multiple fragments with the smallest distance (fragments 4 and 5) were merged. Since the merged fragments satisfy the two conditions described in the text (see Sections2.3)), they were assigned a unique cluster number "2". (C) Fragments 1 and 2 were merged, and a cluster number "3" was assigned. (D) Fragments 4 and 5 were merged, and a cluster number "1" was assigned. (E) Fragments 1 and 2 were merged, but different cluster numbers "3" and "4", respectively, were assigned (see text for detail). (F) Fragments 4 and 5 are merged, but different cluster numbers "5" and "6", respectively, were given (see text for detail). As a result, 10 fragments were aligned to form six clusters in this example.