Environ. Sci. Technol. 2003, 37, 4554–4560

Statistical Evaluation of Bacterial Source Tracking Data Obtained by rep-PCR DNA Fingerprinting of *Escherichia coli*

JOHN M. ALBERT,[†]
JUNKO MUNAKATA-MARR,*[,†]
LUIS TENORIO,[‡] AND
ROBERT L. SIEGRIST[†]

Environmental Science & Engineering Division and Department of Mathematical & Computer Sciences, Colorado School of Mines, Golden, Colorado 80401

Pattern recognition has been applied to environmental systems for identification of numerous pollution sources including aerosolized lead and petroleum hydrocarbons. In recent years, DNA fingerprinting has gained widespread application as a means to characterize genetic variations for such purposes as microbial source tracking. This approach, however, is strongly dependent on the statistical and image analyses applied. Several statistical analyses of rep-PCR DNA fingerprints were assessed as a means to differentiate between potential sources of fecal contamination. GelCompar II and methods based on penalized discriminant analysis (PDA) and *k*-nearest neighbors (KNN) classification procedures were used to differentiate between 10 source groups within a library containing DNA fingerprints of 548 *Escherichia coli* isolates from known human and nonhuman sources. KNN performed significantly better than PDA in a jackknife analysis, though the library was not large enough to detect significant differences between GelCompar II and the other two methods. GelCompar II and KNN both attained ≥90% correct classification in a holdout procedure. In addition, interpoint distance analyses indicate coherency within source groups, while library randomization demonstrated that KNN does not create artificial groupings. This investigation stresses the need to understand limitations of statistical analyses used in pattern recognition of DNA fingerprints.

Introduction

Fecal contamination is a widespread problem throughout the United States that adversely affects surface and groundwaters and raises health concerns about their quality as drinking water sources. In a draft report, the U.S. EPA stated that 21 000 water bodies within the United States were impaired due to fecal contamination from both animal and human sources (*1*). In addition, the latest CDC waterborne disease outbreak report indicated 17.9% and 87% increases in outbreaks associated with drinking surface and groundwaters, respectively, since the previous report (*2*).

In the past 10 years, identifying sources of fecal contamination using bacterial source tracking (BST) techniques has become a large field of study, as evidenced by recent reviews (*3, 4*). The goal of all BST methods is to differentiate between sources of fecal contamination in order to effectively direct mitigation efforts. In general, BST methods rely on unique biomarkers from indicator organisms, biomarkers that are, ideally, specific to host population groups. Many BST methods generate phenotypic or genotypic profiles to characterize bacteria from known source groups. These profiles are then subjected to pattern recognition protocols, analogous to other pollution source-tracking applications (*5–10*). This approach, however, is strongly dependent on the statistical and, in the case of DNA fingerprinting, the image analyses applied (*11*).

Analytical methods used for the purposes of BST include antibiotic resistance assays, host-specific molecular markers, ribotyping, and repetitive element polymerase chain reaction (rep-PCR) (*12–19*). rep-PCR, a DNA amplification procedure that targets repetitive units within bacterial genomic DNA (*20*), has been used in BST because it is able to differentiate between organisms at the strain level (*21–24*). As an analytical tool, rep-PCR reproducibility has been demonstrated within single researcher studies (*25, 26*). However, slight changes in protocol have been shown to affect DNA fingerprints produced, making it difficult to reproduce results in different laboratories (*27*). DNA fingerprints can be analyzed using either banding patterns (band-based) or densitometric curves (curve-based). The latter method is preferred since manual band-based scoring methods introduce a bias (*28*); curved-based analysis should provide less biased analyses of DNA fingerprints (*26, 28, 29*).

Library-based BST methods involve the collection of indicator organisms from known host groups. Biomarkers obtained from the indicator organisms are stored in a database according to the host group of origin. Various classification methods such as discriminant analysis, cluster analysis, and principal component analysis have been used in library-based BST (*13, 14, 16–18, 30, 31*). The goals of these techniques are first, to describe the observations within the library and second, to assign new observations to the correct host. Many of these techniques are performed with commercial software packages that offer convenient templates. However, the limitations of the classification algorithms used in these closed-source software packages are not readily apparent. Such limitations include algorithm assumptions, library size, classification rate uncertainties, and input data range. Using the aforementioned techniques and/or commercial software, a wide range of average rates of correct classification (ARCC), from 67 to 87%, has been reported based on libraries ranging in size from 154 to several thousand indicator organisms (*13, 14, 17–19*).

Identification of contaminant sources using BST relies on the associations between indicator organisms and host groups. Evidence suggests that enteric bacteria may exhibit ecological structure due in part to host adaptation and geographic location (*32*). Some studies indicate that this ecological structure may be maintained in natural populations of *E. coli* (*15, 33–35*). Previous attempts to explore indicator-host correlations (i.e. source group coherency) include analysis of variance, randomization procedures, and G statistics (*32, 33, 36, 37*). New methods of exploring source group distributions are necessary to more directly assess specificity and identify strains of unknown origin. One such method is interpoint distances, a nonparametric procedure that compares multivariate probability distributions.

BST methods are promising, but several factors currently limit their effective application. Lack of both a uniform

* Corresponding author phone: (303)273-3421; fax: (303)273-3413; e-mail: jmmarr@mines.edu.
[†] Environmental Science & Engineering Division.
[‡] Department of Mathematical & Computer Sciences.

10.1021/es034211q CCC: $25.00
Published on Web 09/18/2003