Oligonucleotide Profiling for Discriminating Bacteria in Bacterial Communities

Ping-An He[*,1,2] and Li Xia[3]

[1]*College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, P.R. China*

[2]*College of Life Science, Zhejiang Sci-Tech University, Hangzhou 310018, P.R. China*

[3]*T-Life Research Center, Fudan University, Shanghai 200433, P.R. China*

Abstract: Based on the relative ratios of di- and tri-nucleotides in the DNA sequences, the profiles of 164 genome sequences from 152 representative microbial organisms were computed. By comparing the profiles of the genomes and their substrings with length 500 bps, the fluctuations of the relative abundances of di- and tri-nucleotides of these genomic sequences were analyzed. A new method to discriminate the origins of orphan DNA sequences was proposed, and the origins of 17 uncultured bacterium sequences from a bacterial community in the human gut were postulated and discussed.

Keywords: Bacterial community, DNA sequences, genome, profile.

INTRODUCTION

Since the development of shotgun sequencing technology, many bacterial genomic sequences have been sequenced. However, the bacteria represented in GenBank are still only a small number of the bacteria species on earth [1, 3, 18]. The traditional sequencing techniques for bacterial DNA have been restricted by the prerequisite of obtaining pure cultures [15, 25].

Recently, a new sequencing approach of environmental sampling has been developed [21, 22]. Using whole genome shotgun (WGS) sequencing, Venter *et al*. [22] sampled water from the Sargasso Sea which is one of the most well characterized ocean regions [22]. This genomic DNA based analysis overcomes limitations of conventional culture based technology and provides a powerful tool for researching structure-function relationships of microbial communities. However, several methodological problems remain unsolved, despite the proven potential of the metagenomic approaches to broaden our knowledge about the composition and function of natural microbial communities. For example, identification of the short fragments' organismal origin remains one of the major challenges in bioinformatics.

Many orphan DNA sequences in GenBank have been cloned and sequenced, but their origins have not been identified. These sequences are unique and have shown no homology in BLAST search or other conventional sequence alignment with any known sequences. Thus it is very difficult to postulate their phylogenetic positions. For example, Wei *et al.* [24] obtained 17 uncultured bacterium DNA fragments with a length ranging from 500-1000 bases from a sampling of a bacterial community in the human gut. However, only three sequences showed high similarity with some regions of the *Bacteroides thetaiotaomicron* genome according to a BLAST search, while the origins of the others remained unknown.

Recently, many alignment-free methods have been proposed for the comparison of biological sequences [2, 7-14, 17, 19, 20, 23]. Karlin *et al.* [9, 8, 12] selected the relative abundances of oligonucleotides as a genomic signature to analyze compositional biases in the whole genome. Sandberg *et al.* [17] investigated the possibility of predicting the genome of origin for a specific genomic sequence, and developed a naïve Bayesian classifier. Qi *et al.* [13] proposed a Markov model to subtract the random background from biological sequences in order to get a better evolutionary tree. Applying tetranucleotide frequencies biases, Pride *et al.* [13] compared nucleotide usage pattern conservation for related prokaryotes by examining the representation of DNA tetranucleotide combinations in 27 representative microbial genomes, and Teeling *et al.* [20] discriminated longer genomic fragments (40 kb) based on the idea of z-score. Also, we introduced an approach for comparison of DNA sequences based on the sieve ratio of a trinucleotide [6]. All these results indicate the great potential of utilizing linguistic approaches to solve the fragment identification problem.

In our work, one practical question was considered, namely, which species corresponds to each sequence among the 17 uncultured bacterium sequences described by Wei *et al.* [24]. Following our previous approach [6], we provided a method to identify the homology of these 17 unknown sequences based on the relative abundance of their di- and tri-nucleotide of DNA sequences. The results are organized as follows: First, based on the relative abundances of di- and tri-nucleotides, the profile of representative microbial genomes in Genbank were calculated. Second, for each of the genomes, the fluctuation of the k-profile-distance was examined. Finally, the method of discriminating bacteria for unknown sequences is proposed. In all representative genomes, we evaluated the reliability index of the method by inspecting the identifying index. According to the evaluation index 5, the five known similar species (Genbank accession number) of the 17 unknown sequences were obtained.

*Address correspondence to this author at the College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, P.R. China; Tel: +86 571 86843192; E-mail: pinganhe@yahoo.com.cn and pinganhe@zstu.edu.cn

1386-2073/07 $50.00+.00

MATERIALS AND METHODS

Data

The 17 uncultured bacterium sequences coming from bacterial community in human gut were downloaded from Genbank. They were grouped according to the experiment order, by which sequences coming from the same experiment were put together. The 7 sequences in group 1 were labeled as 1a, 1b, 1c, 1d, 1e, 1f, and 1g; the 6 sequences in group 2 were labeled as 2a, 2b, 2c, 2d, 2e, and 2f; the 4 sequences in Group 3 were labeled as 3a, 3b, 3c, and 3d. Their Genbank accession numbers were AY350322— AY350338.

Collection of Complete Genome of Bacteria in Genbank

The 152 representative microbial organisms which were also downloaded from the NCBI website are listed in Table **1** (ftp://ftp.ncbi.nih.gov/genbank/Bacteria).

The Definition of Profile

In the section, we define the profile of a DNA sequence as follows: for any given DNA sequence, $p\left(a_1a_2\cdots a_k\right)$ denotes the **frequency** of a k-string $a_1a_2\cdots a_k$ (where $a_i \in \{a,c,g,t\}$) occurrence in the sequence, so that

$$p\left(a_1a_2\cdots a_k\right)=\frac{m_k}{L-k+1} \qquad (1)$$

where k is the length of the oligonucleotide and m_k is the count of occurrence in the sequence. Since the occurrence of each oligonucleotide $a_1a_2\cdots a_k$ is determined by occurrence of $a_1a_2\cdots a_{k-1}$, $a_2a_3\cdots a_k$ and $a_2a_3\cdots a_{k-1}$, respectively, the **expected frequency** of the oligonucleotide $a_1a_2\cdots a_k$ can be represented as

$$e\left(a_1a_2\cdots a_k\right)=\frac{p(a_1a_2\cdots a_{k-1})p(a_2a_3\cdots a_k)}{p(a_2a_3\cdots a_{k-1})} \qquad (2)$$

then, we have

$$f\left(a_1a_2\cdots a_k\right)=\frac{p\left(a_1a_2\cdots a_k\right)}{e\left(a_1a_2\cdots a_k\right)}=\frac{p(a_1a_2\cdots a_k)p(a_2a_3\cdots a_{k-1})}{p(a_1a_2\cdots a_{k-1})p(a_2a_3\cdots a_k)} \qquad (3)$$

where $f\left(a_1a_2\cdots a_k\right)$ is called the **relative abundance ratio** of oligonucleotide $a_1a_2\cdots a_k$.

There are 4^k possible k-strings that can occur in a DNA sequence. Thus, for a given DNA sequence, we can write a 4^k-dimension vector to denote a DNA sequence of length n, each component of which is the relative abundance ratio of oligonucleotide with length k in the sequence. We call the k-dimension vector the **k-profile** of the sequence. That is, the k-profile of a DNA sequence is a vector constructed by the relative abundances ratio of oligonucleotide with length k of the sequence. When k=2, we take $p\left(a_2a_3\cdots a_{k-1}\right)=1$, then, the formula (3) is the definition of odds ratio of a dinucleotide (Karlin *et al.*, 1995). Therefore, the profile here is a generalization for the odds ratio of a dinucleotide.

According to the profiles of DNA sequences, we define the **k-profile-distance** of two DNA sequences as follows:

Table 1. 152 Representative Microbial Organisms in GenBank

Aeropyrum pernix K1, *Agrobacterium tumefaciens* strain C58 Cereon, *Agrobacterium tumefaciens* strain C58 UWash, *Aquifex aeolicus* VF5, *Archaeoglobus fulgidus* DSM 4304, *Bacillus anthracis str.* Ames, *Bacillus cereus* ATCC.14579, *Bacillus halodurans* C-125, *Bacillus subtilis* ssp. *Subtilis* 168, *Bacteroides thetaiotaomicron* VPI-5482, *Bdellovibrio bacteriovorus*, *Bifidobacterium longum* NCC2705, *Blochmannia floridanus*, *Bordetella bronchiseptica strain* RB50, *Bordetella parapertussis strain* 12822, *Bordetella pertussis strain Tohama* I, *Borrelia burgdorferi* B31, *Bradyrhizobium japonicum* USDA110, *Brucella melitensis*16M, *Brucella suis* 1330, *Buchnera aphidicola* APS, *Buchnera aphidicola* Sg, *Campylobacter jejuni* ssp. *jejuni* NCTC 11168, *Caulobacter crescentus* CB15, *Chlamydia muridarum* strain Nigg, *Chlamydia trachomatis* serovar D, *Chlamydophila caviae* GPIC, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* CWL029, *Chlamydophila pneumoniae* J138, *Chlamydophila pneumoniae* Tw183, *Chlorobium tepidum* TLS, *Chromobacterium violaceum* ATCC 12472, *Clostridium acetobutylicum* ATCC 824, *Clostridium perfringens* 13, *Clostridium tetani* E88, *Corynebacterium diphtheriae gravis* NCTC13129, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum*ATCC 13032, *Coxiella burnetii* RSA 493, *Deinococcus radiodurans*R1, *Enterococcus faecalis* V583, *Escherichia coli* CFT073, *Escherichia coli* K12-MG1655, *Escherichia coli* O157:H7 VT2-Sakai, *Escherichia coli* O157:H7 EDL933, *Fusobacterium nucleatum* subsp *nucleatum* ATCC 25586, *Geobacter sulfurreducens*, *Gloeobacter violaceus* PCC 7421, *Haemophilus ducreyi strain* 35000HP, *Haemophilus influenzae* Rd, *Halobacterium* sp. NRC-1, *Helicobacter hepaticus ATCC* 51449, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Lactobacillus johnsonii* NCC 533, *Lactobacillus plantarum* WCFS1, *Lactococcus lactis* subsp *lactis* IL1403, *Leptospira interrogans* serovar lai strain 56601, *Listeria innocua* CLIP 11262, *Listeria monocytogenes* EGD-e, *Mesorhizobium loti* MAFF303099, Methanobacterium *thermoautotrophicus* delta H, *Methanocaldococcus jannaschii* DSM2671, *Methanopyrus kandleri* AV19, *Methanosarcina acetivorans* C2A, *Methanosarcina mazei* Goe1, *Mycobacterium avium subsp. paratuberculosis str.* k10, *Mycobacterium bovis subsp. bovis* AF2122 97, *Mycobacterium leprae* TN, *Mycobacterium tuberculosis*CDC1551, *Mycobacterium tuberculosis* H37Rv, *Mycoplasma gallisepticum strain* R, *Mycoplasma genitalium* G37, *Mycoplasma mycoides subsp. mycoides* SC, *Mycoplasma penetrans* HF-2, *Mycoplasma pneumoniae* M129, *Mycoplasma pulmonis* UAB CTIP, *Neisseria meningitidis* serogroup A Z2491, *Neisseria meningitidis*MC58, *Nostoc* sp. PCC 7120, *Nitrosomonas europaea* ATCC 19718, *Oceanobacillus iheyensis*HTE831, *Onion yellows phytoplasma*, *Pasteurella multocida* PM70, *Photorhabdus luminescens subsp. laumondii* TTO1, *Pirellula* sp. strain 1, *Porphyromonas gingivalis W83*, *Prochlorococcus marinus* MED4, *Prochlorococcus marinus* MIT9313, *Prochlorococcus marinus subsp. marinus str.* CCMP1375, *Pseudomonas aeruginosa* PA01, *Pseudomonas putida* KT2440, *Pseudomonas syringae pv. tomato str.* DC3000, *Pyrobaculum aerophilum* IM2, *Pyrococcus abyssi* GE5, *Pyrococcus furiosus* DSM 3638, *Pyrococcus horikoshii*shinkaj OT3, *Ralstonia solanacearum* GMI1000, *Rhodopseudomonas palustris* CGA009, *Rickettsia conorii* Malish 7, *Rickettsia prowazekii* Madrid E, *Salmonella enterica* ssp. *enterica serovar* Typhi, *Salmonella typhi* CT18, *Salmonella typhimurium* LT2 SGSC1412, *Shewanella oneidensis* MR1, *Shigella flexneri* 2a strain 301, *Shigella flexneri* 2a 2457T, *Sinorhizobium meliloti* 1021, *Staphylococcus aureus* Mu50, *Staphylococcus aureus* ssp. aureus MW2, *Staphylococcus aureus* ssp. aureus N315, *Staphylococcus epidermidis* ATCC 12228, *Streptococcus agalactiae* 2603 V/R, *Streptococcus agalactiae*NEM316, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* R6, *Streptococcus pneumoniae* TIGR4, *Streptococcus pyogenes* MGAS315, *Streptococcus pyogenes*MGAS8232, *Streptococcus pyogenes* SSI-1, *Streptococcus pyogenes* SF370 serotype M1, *Streptomyces avermitilis*, *Streptomyces coelicolor* A3(2), *Sulfolobus solfataricus* P2, *Sulfolobus tokodaii* strain 7, *Synechococcus sp.* WH8102, *Synechocystis* sp. PCC 6803, *Thermoanaerobacter tengcongensis* MB4T, *Thermoplasma acidophilum* DSM 1728, *Thermoplasma volcanium*GSS1, *Thermosynechococcus elongatus* BP-1, *Thermotoga maritima* MSB8, *Treponema denticola* ATCC 35405, *Treponema pallidum* Nichols, *Tropheryma whipplei* TW0827, *Tropheryma whipplei* strain Twist, *Ureaplasma urealyticum* parvum biovar serovar 3, *Vibrio cholerae* El Tor N16961, *Vibrio parahaemolyticus* RIMD 2210633, *Vibrio vulnificus* CMCP6, *Vibrio vulnificus* YJ016, *Wigglesworthia brevipalpis*, *Wolbachia endosymbiont of Drosophila melanogaster*, *Wolinella succinogenes*, *Xanthomonas axonopodis* pv. citri 306, *Xanthomonas campestris* pv. campestris ATCC 33913, *Xylella fastidiosa*, *Xylella fastidiosa* Temecula1, *Yersinia pestis* CO92, *Yersinia pestis* KIM.

$$D(P,Q)=\sum_{i=1}^{4^k}\frac{2(p_i-q_i)^2}{p_i+q_i} \quad (4)$$

where P, Q are the k-profiles of two sequences and p_i, q_i are the components of the two k-profile, respectively.

Similarities of Short DNA Sequence Fragments and Genomic Sequences

In the section, a method is provided to compute similarities between short DNA sequence fragments and genomic sequences. We first calculated the k-profile of the microbial genome based on the relative abundance ratio of k-string for the two strands of each genome. In this work, we only take k=2 and 3. When k≧4, some values of $p(a_1a_2\cdots a_k)$ are 0 in 17 unknown sequences. For example, tetranucleotides AAAA and AAAG do not occur in sequence 1a.

For two strands of each genomic sequence, we chose a 500 bp wide sliding window that slides in the complete genome with a stride of 20 bp, and computed the 2-profile and 3-profile for each window sequence. Thus, we obtained two profile sets for each complete genome, and named the 2-profile set of all window sequences $\mathbf{G_2}$ and the 3-profile set of all window sequences $\mathbf{G_3}$, respectively. The 2- and 3-profile-distances were calculated between the profile of the genomic sequence and all elements in $\mathbf{G_2}$ and $\mathbf{G_3}$ so that we could obtain some information regarding the oligonucleotide composition of the genomic sequence.

Furthermore, for a given short DNA sequence (500-1000 bp) and genomic sequence, we defined the **2-distance** and **3-distance** between the short sequences and the genomic sequence. Recalling the set $\mathbf{G_2}$ and $\mathbf{G_3}$ for the genomic sequence, the 2- and 3-profile-distances were computed between the profile of the short DNA sequence and all elements in $\mathbf{G_2}$ and $\mathbf{G_3}$. In all calculated 2- and 3-profile-distances, the minimum was selected as the distance between the short DNA sequences and the genomic sequence, and these two distances were called their **2-distance** and **3-distance**, respectively. In addition, the similarities were computed between the short DNA sequence fragment and the genomic sequence based on the 2-distance and 3-distance. Normally, the smaller the distance the more similar the sequences.

RESULTS AND DISCUSSION

The Profile of 152 Microbial Organisms

Observing the 2- and 3-profiles of all genomic sequences, we found that the relative abundance ratio of an oligonucleotide X_1X_2 for a dinucleotide or $X_1X_2X_3$ for a trinucleotide was close to the profile of its reversed complement sequence, $X'_2X'_1$ or $X'_3X'_2X'_1$, respectively, where X'_i was D-W complement letter of X_i (results not shown). These results implied that the 2- and 3-profiles of one strand of the genomic sequences for all bacteria approach the 2- and 3-profiles of the other strand. In other words, the component of profile of genomic sequences could be reduced from 16 to 10 for 2-profile and from 64 to 32 for 3-profile. Furthermore, another phenomenon was observed: over-represention of dinucleotides (mainly occurring in AA, TT and GC), where their relative abundance ratio was more than 1.20. Under-represented di-nucleotides with a relative abundance ratio of less than 0.8, mainly occurred in AC, GT, CG, and TA for all microbial genomic sequences. Similarly, for all microbial genomic sequences, over-represented tri-nucleotides ($f(X_1X_2X_3)>1.2$) mainly occurred in ACC, GGT, ACG, CGT, AGC, GCT, ATA, TAT, ATC, GAT, CAG, CTG, CCA, TGG, GAA, TTC, GTA, and TAC, respectively, while underrepresented ($f(X_1X_2X_3)<0.8$) mainly occurred in AAT, ATT, ACT, AGT, CCC, GGG, CTA, TAG, CTC, GAG, GAC, GTC, GCA, TGC, GGA, TCC, TAA and TTA, respectively.

Comparing the 2- and 3-profiles of complete genomic sequences, we found that the 2-profile and 3-profile of each genus had its common identity, but the 2- and 3-profiles of bacteria from two different genuses were more obviously different. That is, the 2- and 3-profiles contained genus-specific signatures. For example, there were seven species belonging to the *Chlamydiae* genus among the 152 microbial organisms studied. The seven species were *C. muridarum*

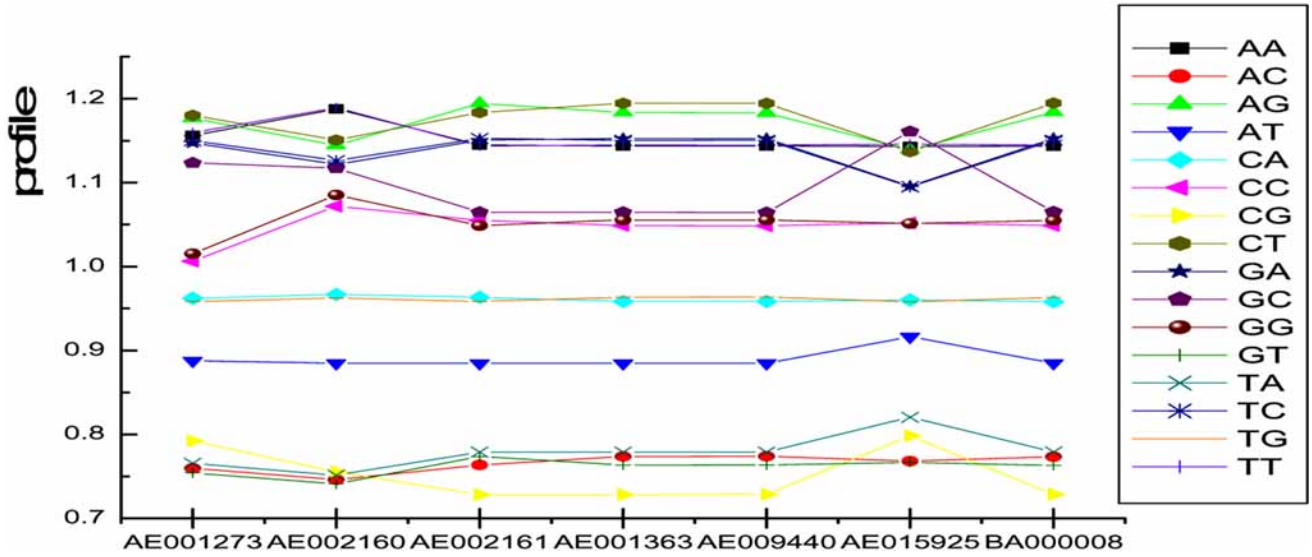

Fig. (1). The 2-profiles of the *Chlamydiae* genus in GenBank.

strain Nigg, *C. trachomatis* serovar D, *C. caviae* GPIC, *C. pneumoniae* AR39, *C. pneumoniae* CWL029, *C. pneumoniae* J138, and *C. pneumoniae* Tw183, respectively. Their 2-profiles are plotted in Fig. **1**, in which these values showed only small variations.

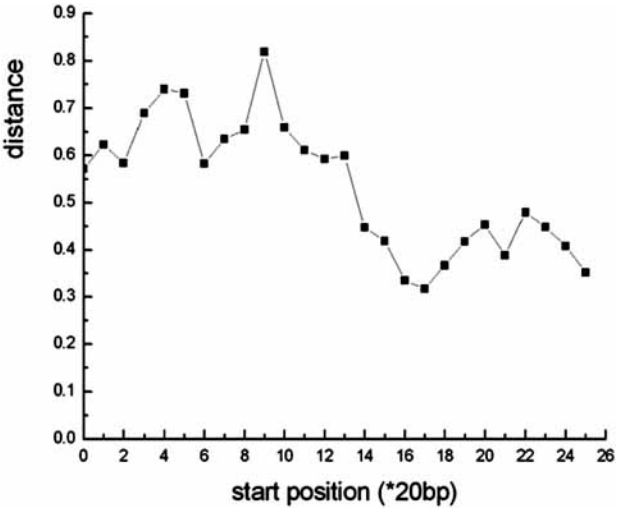

Fig. (2). The 2-profile-distance fluctuation curve for the first 1000 bp of *E. coli* K-12 MG1655.

The Fluctuation of K-Profile-Distance of Genomic Sequences

First, a 1000 bp sequence of the *Escherichia coli* K-12 MG1655 genome was used to illustrate how to compute the fluctuation of k-profile-distance of genomic sequences. First, we computed the profile set $\mathbf{G_2}$ and the 2-profile of *E. coli* K-12 MG1655 genome, and the set $\mathbf{G_2}$ contained 26 elements. Then, the 2-profile-distance between *E. coli* K-12 MG1655 and all 26 subsequences with length 500bps were calculated according to the formula (4). The results of the fluctuation curve of 2-profile-distance are plotted in Fig. **2**. Some information regarding oligonucleotide composition was obtained for each microbial genome *via* the observation of fluctuation curve of the 2- and 3-profile-distances. The larger the distance the more different was the oligonucleotide composition between two sequences.

The fluctuation curves of 2-profile-distance and 3-profile-distance are shown in Figs. **3-6** for the four species, *E. coli* CFT073, *E. coli* K12-MG1655, *E. coli* O157:H7 VT2-Sakai, and *E. coli* O157:H7 EDL933. From these figures, we found that the compositions of oligonucleotides were very stable in the majority of regions of the genomic sequences. In more than 95% of the regions, the 2-profile-distances between the genome and its substrings were less than 1.2, and the 3-profile-distances were less than 16. For the 2-profile-distances and 3-profile-distances of other bacterial genomes, similar results were obtained.

On the other hand, large 2- and 3-profile-distances (very large dissimilarity) were found between the whole genome and its substrings in Figs. **3-6**, and large 2- and 3-profile-distances corresponding to the genome's substring fragments were mainly repeated sequences. In these repeated sequences, some sequences were encoding sequences in genomic sequences. By comparing these coding sequences, we found that there were great similarities within the same species, even sometimes in the same genus. For instance, we checked the oligonucleotides fragments in the four complete genomes, *E. coli* CFT073, *E. coli* K12-MG1655, *E. coli* O157:H7 VT2-Sakai, and *E. coli* O157:H7 EDL933, which corresponded to the largest 2- and 3-profile-distances, and we found that these fragments all belonged to the tolA gene, which corresponded to the cell envelope integrity inner membrane protein. The tolA sequence of *E. coli* CFT073 and *E. coli* K12-MG1655 are identical (Figs. **3** and **4**), while the tolA sequence of *E. coli* O157:H7 VT2-Sakai and *E. coli* O157:H7 EDL933 share the same gene encoding sequence (Figs. **5** and **6**). The tolA sequence of *E. coli* CFT073 is similar the tolA sequence of *E. coli* O157:H7 VT2-Sakai (Figs. **3** and **5**).

Furthermore, the tolA gene was investigated in other genomes. We found that some 12 other organism belonging to 6 different genuses also contain the tolA gene. These were

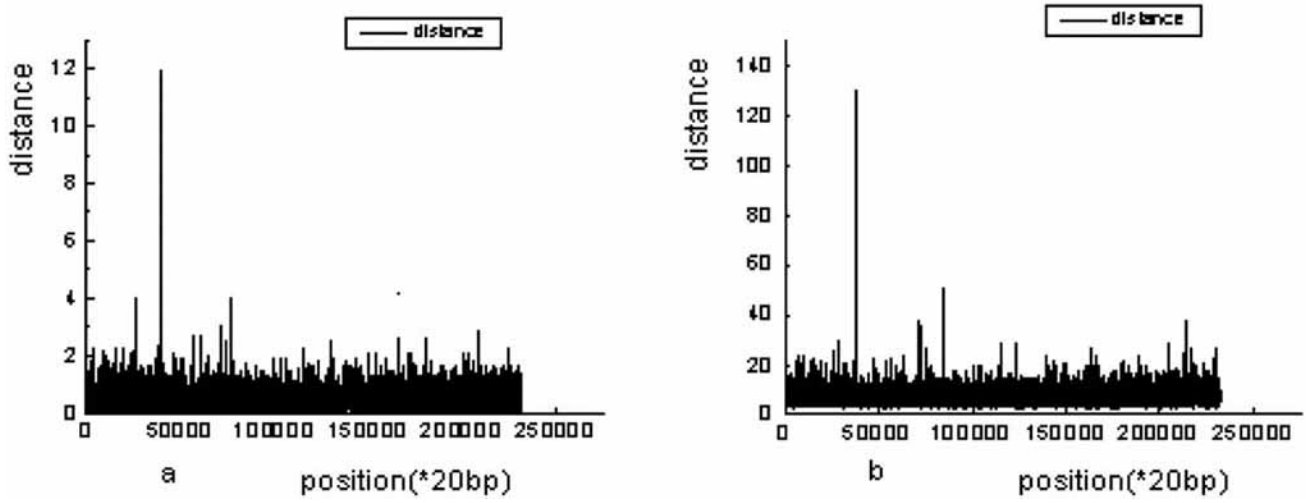

Fig. (3). The k-profile-distance fluctuation curves in the *E. coli* K-12 MG1655 genome. a) 2-Profile-distance fluctuation curves; b) 3-profile-distance fluctuation curves. The location of the sequence fragment of the maximal distance occurring in the genome is from 775565 bp to 776830 bp, which is the sequences fragment that codes gene tolA.

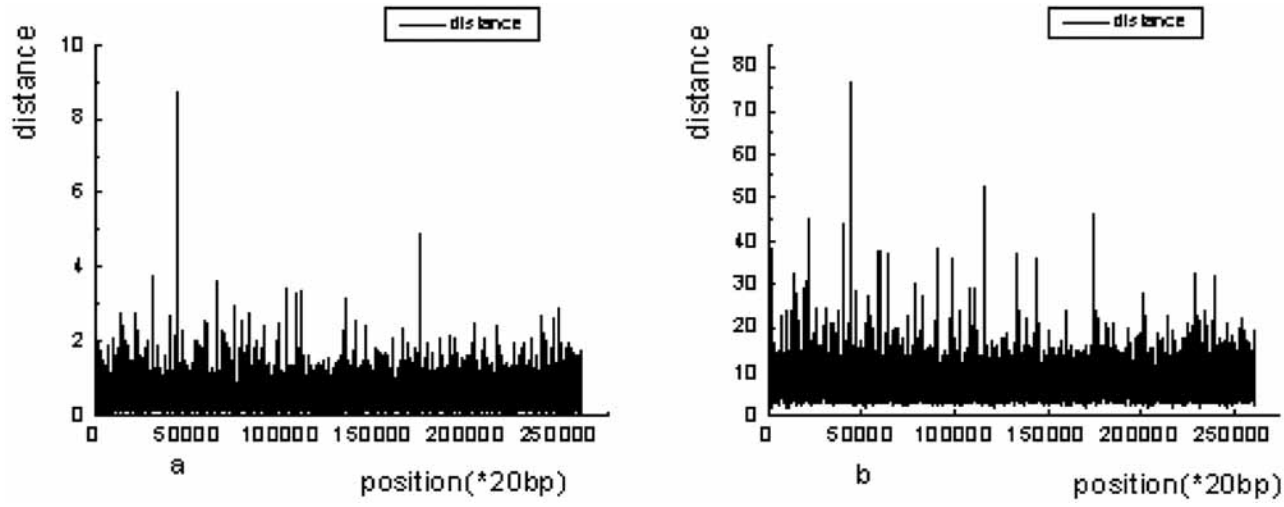

Fig. (4). The k-profile-distance fluctuation curves in the *E. coli* CFT073 genome. a) 2-Profile-distance fluctuation curves; b) 3-profile-distance fluctuation curves. The location of the sequence fragment of the maximal distance occurring in the genome is from 799012 bp to 800277 bp, which is the sequences fragment that codes gene tolA.

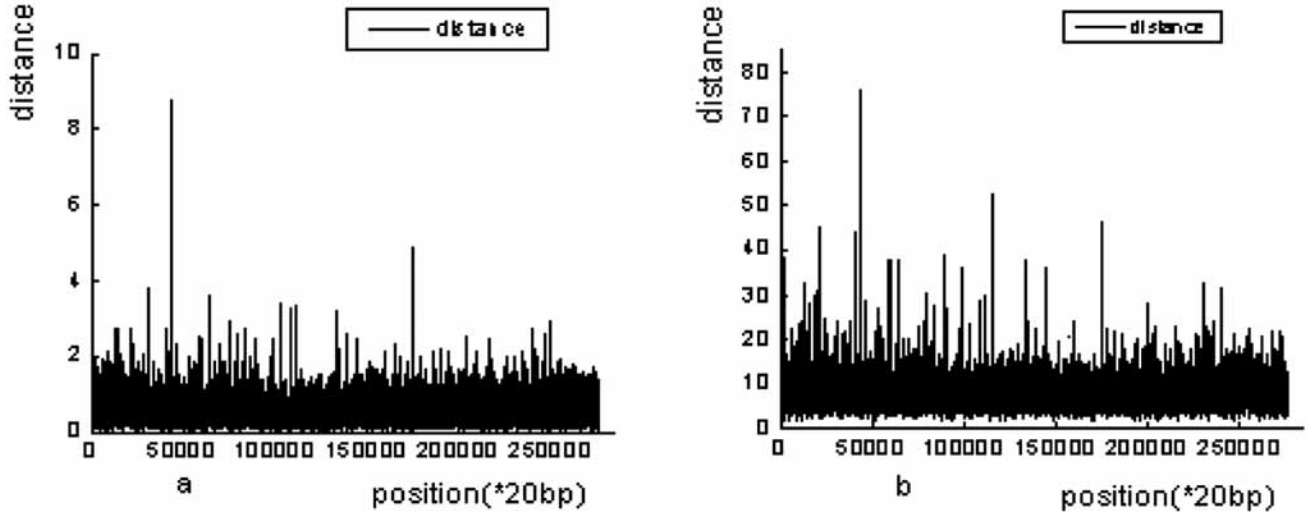

Fig. (5). The k-profile-distance fluctuation curves in the *E. coli* O157H7 genomic sequence. a) 2-Profile-distance fluctuation curves; b) 3-profile-distance fluctuation curves. The location of sequence fragment of the maximal distance occurring in the genome is from 860848 bp to 862032 bp, which is the sequences fragment that codes gene tolA.

Photorhabdus luminescens subsp. laumondii TTO1, *Pseudomonas putida* KT2440, *Salmonella enterica* ssp. *enterica serovar* Typhi, *Salmonella typhi* T18, *Salmonella typhimurium* LT2 SGSC1412, *Shigella flexneri* 2a strain 301, *Shigella flexneri* 2a 2457T, *Vibrio cholerae* El Tor N16961, *Yersinia pestis*CO92, and *Yersinia pestis* KIM.

Again, through observing the fluctuation curves of 2-profile-distance and 3-profile-distance of other bacterial genomes, we found the same repeated sequences occurring in *Staphylococcus aureus* Mu50, *Staphylococcus aureus* ssp. aureus MW2, and *Staphylococcus aureus* ssp. aureus N315. Each of these bacterial genomes contained the genes clfB, sdrC and sdrD. Moreover, the genomes of *Listeria innocua* CLIP 11262 and *Listeria monocytogenes* EGD-e contained the repeated sequences LPXTG motif.

Discriminating Method and its Evaluation Index

The 152 microbial organisms have a total of 164 genomic sequences, because some species have two chromosomes. In these 164 genomic sequences, there are just 120 genomic sequences with more than two genomic sequences in the same genus. So, we selected the 120 genomic sequences from 164 genomic sequences, and divided them into two parts and named them, "training set" and "testing set," respectively. The training set contained 68 genomes sequences, and the testing set contained 52 genomes sequences. Their contents are listed in Tables **2** and **3**. For each genomic sequence in the testing set, there is more than one genomic sequence in the same genus in the training set.

In [24], the 17 unknown sequences were randomly chosen from the bacterial community sampling of the human gut. For each of the genomic sequences in the testing set, we randomly selected 26 subsequences with length 501+20i (i=0, 1,…, 25) to model experimentation. We computed the distances between each of the 26 subsequences and 68 genomes in the training set, based on the definition of 2-distance and 3-distance between the short sequence fragment

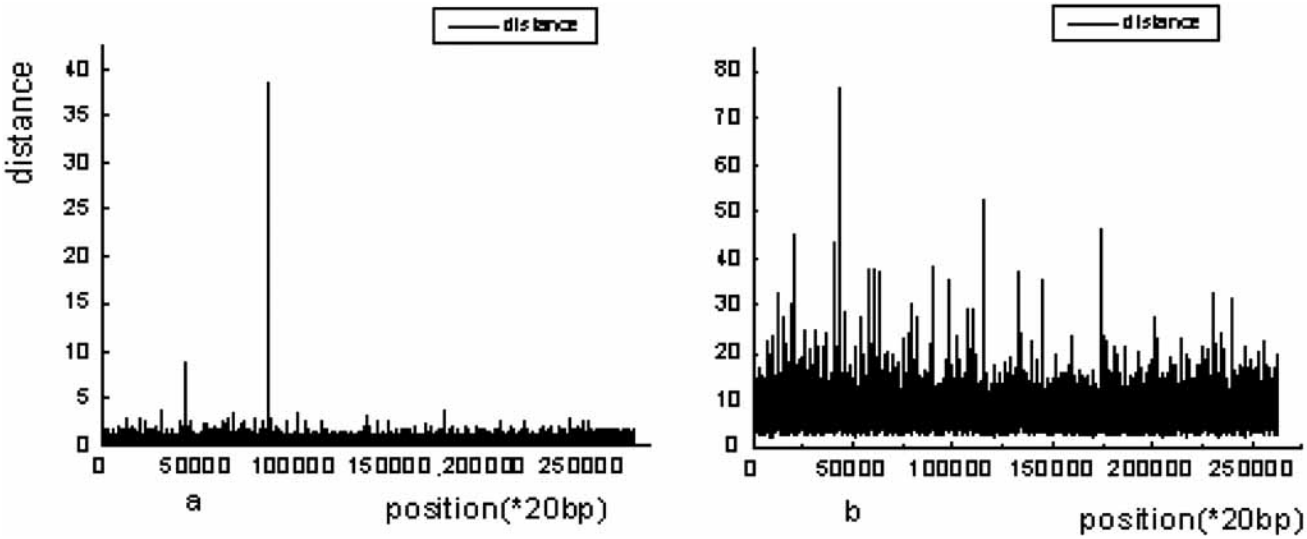

Fig. (6). The k-profile-distance fluctuation curves in *E. coli* O157H7_EDL933 genome. a) 2-Profile-distance fluctuation curves; b) 3-profile-distance fluctuation curves. The location of the sequence fragment of the maximal distance occurring in the genome is from 862502 bp to 863686 bp, which is the sequence fragment that codes gene tolA. (The characters in the sequences fragment from 1725749 bp to 1729749 bp in the genome are all N, so we ignore the sequence fragment).

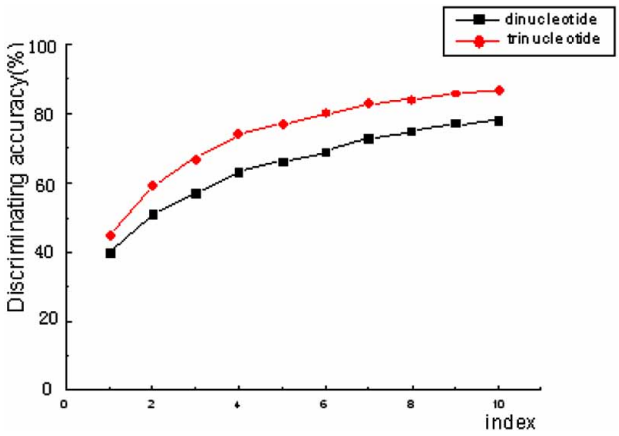

Fig. (7). Dependence discriminating accuracy against index.

and the genomic sequence, and then listed all 68 values in ascending order. In the 68 ordinal values, the ordinal number of the same genus that first occurred was defined as the index of the sequence, and named *index i*. Thus, there were 26 indexes, which corresponded to 26 short subsequences for each genomic sequence, and the average of the 26 indexes was considered as the index of the genus. For example, we randomly selected 26 subsequences from the *Streptomyces coelicolor* A3(2) genomic sequence (AL645882). In all indexes corresponding to the 26 subsequences, the indexes of 22 subsequences had the value 1, and the indexes of 2 subsequences had the value 2, and the indexes of the 2 sequences had the value 3. So, the average 1/26(1*22+2*2+3*2) = 2 was considered to be the index of the *Streptomyces* genus. Finally, the average of the indexes of 52 genomic sequences in the testing set referred to the index of the method that would be used in the part of "Discriminating unknown sequences". Normally, the smaller the index and the higher the accuracy mean greater reliability. The results of 2- and 3-distances are plotted in Fig. **5**, where the discriminating accuracy depended on the index from 1 to 10. When the index was 5, we found that the reliability of the criterion should be 77% and 67% for tri- and dinucleotides, respectively. Thus, we chose the average of all 52*26 indexes, 5, as the discriminating criterion index for the method.

Table 2. The Genomic Sequences in the Training Set (Only GenBank Accession Number)

AE007869, AE007870, AE016879, BX248353, AE000516, AE000520, AE001273, AE001437, AE001439, AE002160, AE003849, AE003852, AE003853, AE004091, AE004437, AE005672, AE005674, AE006468, AE006470, AE006641, AE006914, AE008917, AE008923, AE009440, AE009948, AE009949, AE009952, AE010299, AE010300, AE014074, AE014075, AE014184, AE014291, AE015450, AE016795, AE016796, AE016826, AE016853, AE016877, AE016958, AE017126, AE017143, AE017196, AE017198, AE017226, AF222894, AL096836, AL139299, AL157959, AL450380, AL513382, AL592022, AL646052, AL732656, BA000001, BA000003, BA000007, BA000008, BA000018, BA000030, BA000033, BA000035, BX293980, BX470249, BX470250, BX548174, BX842601, U00089.

Table 3. The Genomic Sequences in the Testing Set (Only GenBank Accession Number)

AE008689, AE017226, BX842601, AE000511, AE002098, AE002161, AE004092, AE005174, AE007317, AE008384, AE008918, AE008922, AE009442, AE009950, AE010301, AE013218, AE014073, AE014133, AE014292, AE014613, AE015451, AE015925, AE015927, AE015929, AE017125, AJ235269, AL009126, AL123456, AL445566, AL590842, AL591824, AL645882, AL935263, BA000004, BA000011, BA000016, BA000017, BA000023, BA000026, BA000031, BA000032, BA000034, BA000036, BA000037, BA000038, BX072543, BX248333, BX470248, BX548175, U00096, L42023, L43967.

Discriminating Unknown Sequences

In this section, we took all 164 microbial genomic sequences as the training set, and 17 uncultured bacterium se-

Table 4. Similarities Table for 17 Uncultured Bacterium Sequences Based on the Profiles of Di- and Trinucleotides

Unknown Sequence	Genomic Sequence	2-Distance	Genomic Sequence	3-Distance
1a	**AE015451**	0.023484	AE014075	2.908087
	BX548175	0.024802	U00096	2.953511
	BA000037	0.024929	AE005174	3.124335
	BX119912	0.029530	AE017180	3.136746
	AE016853	0.032505	BA000007	3.149035
1b	AE009952	0.041787	BX842601	5.702246
	AL590842	0.041787	AE014075	6.187696
	AE015928	0.053199	**AE015928**	6.201896
	AE017198	0.057366	AE010299	6.205095
	AL935263	0.058521	AE002098	6.310886
1c	AE005174	0.014330	AL009126	4.007225
	AL935263	0.014970	BX470251	4.161818
	BA000007	0.017543	BA000038	4.419935
	AL590842	0.018805	AB001339	4.430572
	AE004439	0.019666	AE016853	4.459725
1d	BA000028	0.026113	AE004439	3.788560
	BA000011	0.030802	AE008691	3.878428
	AE000666	0.034429	BX842601	3.898293
	AE015928	0.035054	NC003450	3.937844
	AL139299	0.038389	AE008384	3.960859
1e	BA000002	0.045653	BA000035	6.526810
	AE010299	0.070637	BX248353	7.122858
	BA000019	0.075084	BA000017	7.255012
	AE015928	0.076419	BA000018	7.259932
	BA000011	0.077633	AL096836	7.318569
1f	AB001339	0.077405	**AE015928**	5.074988
	U00089	0.125180	AE010299	5.538611
	BA000019	0.132248	AL591688	5.595195
	AL591824	0.155209	AE008922	5.770153
	AE010300	0.166006	AE015924	5.816370
1g	BA000035	0.035198	AE008923	5.413208
	AE017180	0.043188	BA000045	5.557992
	AE006470	0.047948	AL591688	5.627130
	AE014295	0.053425	BX119912	5.648832
	AL954747	0.059045	AE016853	5.879656
2a	**AE015928**	0.044871	**AE015451**	3.525918
	BA000011	0.046573	BA000039	3.689116
	AE016796	0.046995	AL646053	3.871787
	AE015924	0.047202	**AE015928**	4.022744
	AL096836	0.047285	BX470251	4.040708
2b	AE004437	0.020041	**AE015924**	2.989321
	BX248353	0.049689	AE003849	3.152230
	AE009439	0.053531	AE014073	3.173817
	AE015928	0.055292	AE008917	3.255469
	BA000045	0.057757	BX842601	3.275928
2c	AL954747	0.027716	**AE015451**	3.369904
	AE009948	0.030661	BX470248	3.397388
	BA000011	0.033040	AL450380	3.571710
	BX548175	0.039769	AL591824	3.639649
	AE002161	0.043601	AE000512	3.661922

(Table 4) contd…..

Unknown Sequence	Genomic Sequence	2-Distance	Genomic Sequence	3-Distance
2d	AE004437	0.020041	AE004091	2.682966
	BX248353	0.049689	AE009442	2.778905
	AE009439	0.053531	***AL139299***	2.891064
	AE015928	0.055292	AE014073	2.960381
	BA000045	0.057757	BX470248	2.976455
2e	***AL139299***	0.022449	AE016877	3.293186
	AL590842	0.022989	AJ235269	3.873049
	BA000028	0.023085	**AE015928**	3.922774
	AE009952	0.023672	***AL139299***	3.954790
	AE017125	0.024402	BA000007	3.959378
2f	BX548175	0.023389	AE003853	3.300902
	NC003450	0.037711	AE014073	3.328468
	AE015451	0.039026	AE016796	3.331219
	AE017180	0.039870	AE005674	3.449416
	BA000001	0.042271	AE016877	3.511325
3a	BX548175	0.030087	AE008917	2.931090
	BA000011	0.040887	AE014291	2.931090
	AL954747	0.044347	AE006470	3.048056
	AE000520	0.045624	NC003450	3.109679
	AE015924	0.045984	AE016853	3.197760
3b	**AE015451**	0.020849	BX470251	2.706370
	AL139299	0.026943	***BA000011***	2.722835
	AE014299	0.032843	AE014292	2.725614
	BA000019	0.034353	AE008918	2.741842
	AE015924	0.035948	AL732656	2.790280
3c	BA000019	0.015054	AE003852	2.434237
	AE014299	0.021602	BA000037	2.512702
	AE017196	0.024546	BA000028	2.578067
	AE005672	0.028877	AL591824	2.630593
	AE015929	0.029628	AE003849	2.664877
3d	**AE015924**	0.013998	AE006470	2.856421
	AL139299	0.015865	AE005174	3.022110
	AE014295	0.017673	BA000007	3.173329
	AE017180	0.021083	AE007869	3.193097
	BX470251	0.022682	AE014075	3.245225

quences as the testing set to discriminate between bacteria in bacterial communities. We calculated the 2- and 3-distances between each uncultured bacterium sequence and 164 complete genomic sequences to compare their similarities, so that we discriminated the homology of the sequence and 152 microbial organisms. The results are listed in Table **4** based on the discriminating criterion index 5.

Observing Table **4**, we found some information about the homology of the 17 uncultured bacterium sequences and representative microbial genomes. For example, *Bacteroides thetaiottaomicron VPI-5482* (AE015928) occurred most frequently in Table **4**, and it occurred 10 times. This implied that *B. thetaiottaomicron VPI-5482* was the most similar of the representative microbial genomes to the 17 uncultured bacterium sequences. We searched the species according to frequency of occurrence in Table **4**. The five most frequently occurring genomes were *Bacteroides thetaiottaomicron VPI-5482* (10 times), *Porphyromonas gingivalis* W83(AE015924) (6 times), *Thermoplasma volcanium* GSS1(BA000011) (6 times), *Thermoplasma acidophilum* DSM1728(AL139299) (6 times) and *Pseudomonas putida* KT2440 (AE015451) (5 times), respectively. We found the organisms *B. thetaiottaomicron VPI-5482* and *P. gingivalis* W83 belong to the same phylum, class and order, classified as B20.1.1.1.1 and B20.1.1.3.1, respectively, which showed that two organisms should be very similar species.

In 17 uncultured bacterium sequences, 1b, 1d, 1e, 1f, 2a, 2b, 2d 2e, 3a, 3b, and 3d, were similar to *B. thetaiottaomicron VPI-5482* and *P. gingivalis* W83. In particular, the sequences, 1b and 2a are strongly similar to the *B. thetaiottaomicron* VPI-5482 di- and tri-nucleotides. These results imply that the origins of the two sequences were homologous species of *B. thetaiottaomicron* VPI-5482. In Table **4**, *B. thetaiottaomicron* VPI-5482 and *P. gingivalis* W83 occurred

simultaneously in the similar species of 1f, 2a and 2b. The results suggested similarities between these sequences and B20.1.1. *. *. Recalling the results of Wei *et al.* [24], 1e, 2b and 3a at the nucleic acid level, and 2f, 3c and 3d at the amino acid level shared homology with regions of *B. thetaiottaomicron* by using the BLAST search in Genbank. Synthesizing the results of Wei and our results, we conclude that the sequences 1b, 1e, 1f, 2a, 2b, and 3d are homologous to B20.1.1. *. *.

Except for the two organisms above, *T. volcanium* GSS1 and *T. acidophilum DSM1728* should also be of concern, because the count of their occurrence in Table **4** was only less than *B. thetaiottaomicron.* Another reason was that they belonged to Archaea, which were all A2.5.1.1.1. It is well known that Archaea and Bacteria belong to different kingdoms. In Table **4**, the sequences 1d, 1e, 2a, 2c, 2d, 3a, 3b and 3d are similar to the two organisms. Especially, the sequences, 1d, 2e and 3b are very similar to the two organisms. It needs to proven whether or not there exist similarities between 1d, 2e and 3b and the two organisms.

Among the species containing sequences similar to 1a, 1c, 1g, 2f, and 3c, most were classified as *Proteobacteria* of the *Gammaproteobacteria* class (B12.3.*.*.*), which contains *Pseudomonas putida KT2440* (AE015451). Since many *Proteobacteria* species exist in the human gut, we conclude that the species corresponding to the five sequences possibly came from B12.3.

CONCLUSIONS

When using the sequencing approach of environmental sampling, over one million nucleic acid sequences are obtained which are usually short oligonucleotide fragments. How to classify them is a very difficult problem at present, since few complete DNA databases exist for comparison or alignment. Therefore, the identification of the organisms of origin of these fragments is a limiting factor of the environmental sampling approach.

To complement this alignment issue, the linguistic approaches have many advantages. In this paper, we propose a method to discriminate bacteria in bacterial communities based on their oligonucleotide profiles. In this method, the minimal distance between short sequences and all subsequences of genomic sequence denotes the similarity between short sequences and genomic sequence. We choose a 500 bp wide sliding window that slides in the complete genome with a stride of 20 bp, and compute the 2- and 3-profile for each window sequence. In fact, for the size of sliding window, we can select different length based on the length of the unknown sequence.

Based on the 2- and 3-profiles, genomic fragments derived from the bacteria community can be discriminated. Thus, together with such widely used identification approaches such as G+C content, BLAST etc., the use of our method will enhance our ability to classify genomic fragments. Further developments of this method might involve the combination of G+C content and profiles of tetranucleotides, etc.

ACKNOWLEDGEMENTS

The authors sincerely acknowledge Prof. Bailin Hao and Prof. Liping Zhao for valuable discussions on the manuscript. This work was supported by the Science Foundation of Zhejiang Sci-Tech University (ZSTU) under Grant No. 0613190-Y.

REFERENCES

[1] Boucher, Y.; Nesbo, C.L.; Doolittle, W.F. *Cur. Opin. Microbiol.*, **2001**, *4*, 285-289.
[2] Campbell, A.; Mrazek, J.; Karlin, S. *Proc. Natl. Acad. Sci. USA*, **1999**, *96*, 9184-9189.
[3] Dahllof, I. *Cur. Opin. Biotechnol.*, **2002**, *13*, 213-217.
[4] Frazer, K.A.; Elnitski, L.; Church, D.M.; Dubchak, I.; Hardison, R.C. *Genome Res.*, **2003**, *13*, 1-12.
[5] Garrity, G.M.; Bell, J.A.; Lilburn, T.G. Taxonomic Outline of The Prokaryotes, Bergey's Manual of Systematic Bacteriology. Springer-Verlag, New York, 2nd Ed. Release 5.0, **2004**.
[6] He, P.A. *Comb. Chem. High Throughput Screen.*, **2005**, *8*, 449-453.
[7] Jernigan, R.W.; Baran, R.H. *BMC Genomics*, **2002**, *3*, 23.
[8] Karlin, S. *Cur. Opin. Microbiol.*, **1998**, *1*, 598-610.
[9] Karlin, S.; Burge, C. *Trends Genet.*, **1995**, *11*, 283-290.
[10] Mira, A.; Ochman, H.; Moran, N.A. *Trends Genet.*, **2001**, *17*, 589-596.
[11] Mira, A.; Klasson, L.; Andersson, S.G.E. *Cur. Opin. Microbiol.*, **2002**, *5*, 506-512.
[12] Mrazek, J.; Karlin, S. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 3720-3725.
[13] Pride, D.T.; Meineramann, R.J.; Wassenaar, T.M.; Blaser, M.J. *Genome Res.*, **2003**, *13*, 145-158.
[14] Qi, J.; Wang, B.; Hao, B.L. *J. Mol. Evol.*, **2004**, *58*, 1-11.
[15] Reed, D.L.; Hafner, M.S. *Microbial Ecol.*, **2002**, *44*, 78-93.
[16] Rocap, G.; Larimer, F.W.; Larmerdin, J;. Malfatti, S.; Chain, P.; Ahlgren, N.A.; Arellano, A.; Coleman, M.; Hauser, L.; Hess, W.R.; Johnson, Z.I.; Land, M.; Lindell, D.; Post, A.F.; Regala, W.; Shah, M.; Shaw, S.L.; Steglich, C.; Sullivan, M.B.; Ting, C.S.; Tolonen, A.; Webb, E.A.; Zinser, E.R.; Chisholm, S.W. *Nature*, **2003**, *424*, 1042-104.
[17] Sandberg, R.; Winberg, G.; Branden, C.L.; Kaske, A.; Ernberg, I.; Coster, J. *Genome Res.*, **2001**, *11*, 1404-1409.
[18] Schloter, M.; Lebuhn, M.; Heulin, T.; Hartmann, A. *FEMS Microbiol. Rev.*, **2004**, *24*, 647-660.
[19] Stuart, G.W.; Moffett, K.; Baker, S. *Bioinformatics*, **2002**, *18*, 100-108.
[20] Teeling, H.; Meyerdierks, A.; Bauer, M.; Amann, R.; Glockner, F.O. *Environ. Microbiol.*, **2004**, *6*, 938-947.
[21] Tyson, G.W.; Chapman, J.; Hugenholtz, P.; Allen, E.E.; Ram, R.J.; Richardson, P.M.; Solovyev, V.V.; Rubin, E.M.; Rokhsar, D.S.; Banfield, J.F. *Nature*, **2004**, *428*, 37-43.
[22] Venter, J.C.; Remington, K.; Heidelberg, J. *Science*, **2004**, *304*, 66-74.
[23] Vinga, S.; Almeida, J. *Bioinformatics*, **2003**, *4*, 513-523.
[24] Wei, G.F.; Pan, L.; Du, H.M.; Chen, J.Y.; Zhao, L.P. *J. Microbiol. Methods*, **2004**, *59*, 91-108.
[25] Zhou, J.Z.; Miller, J.H. *J. Bacteriol.*, **2002**, *184*, 4327-4333.

Received: November 4, 2005 **Revised: March 14, 2007** **Accepted: April 18, 2007**