Forensic Science International: Genetics 37 (2018) 13–20

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Optimizing body fluid recognition from microbial taxonomic profiles

Eirik Nataas Hanssen[a,b,*], Kristian Hovde Liland[c,d], Peter Gill[a,b], Lars Snipen[d,**]

[a] *Department of Forensic Biology, Oslo University Hospital, P.O. Box 4950 Nydalen, N-0424 Oslo, Norway*
[b] *Department of Forensic Medicine, University of Oslo, P.O. Box 4950 Nydalen, N-0424 Oslo, Norway*
[c] *Faculty of Science and Technology, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway*
[d] *Faculty of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway*

ARTICLE INFO

Keywords:
Forensics
Massive parallel sequencing
Microbiome
PLS
Discriminants

ABSTRACT

In forensics the DNA-profile is used to identify the person who left a biological trace, but information on body fluid can also be essential in the evidence evaluation process. Microbial composition data could potentially be used for body fluid recognition as an improved alternative to the currently used presumptive tests. We have developed a customized workflow for interpretation of bacterial 16S sequence data based on a model composed of Partial Least Squares (PLS) in combination with Linear Discriminant Analysis (LDA). Large data sets from the Human Microbiome Project (HMP) and the American Gut Project (AGP) were used to test different settings in order to optimize performance. From the initial cross-validation of body fluid recognition within the HMP data, the optimal overall accuracy was close to 98%. Sensitivity values for the fecal and oral samples were ≥ 0.99, followed by the vaginal samples with 0.98 and the skin and nasal samples with 0.96 and 0.81 respectively. Specificity values were high for all 5 categories, mostly > 0.99. This optimal performance was achieved by using the following settings: Taxonomic profiles based on operational taxonomic units (OTUs) with 0.98 identity (OTU98), Aitchisons simplex transform with $C = 1$ pseudo-count and no regularization ($r = 1$) in the PLS step. Variable selection did not improve the performance further. To test for robustness across sequencing platforms, we also trained the classifier on HMP data and tested on the AGP data set. In this case, the standard OTU based approach showed moderately decline in accuracy. However, by using taxonomic profiles made by direct assignment of reads to a genus, we were able to nearly maintain the high accuracy levels. The optimal combination of settings was still used, except the taxonomic level being genus instead of OTU98. The performance may be improved even further by using higher resolution taxonomic bins.

1. Background

DNA analysis is used to identify a person from a biological trace found at a crime scene. However, a trace might not be crime related and could be a result of ‘innocent’ activity or contamination [1–5]. In order to establish a picture of how the trace was deposited, it is helpful to determine from which part of the body the trace originates. Traditionally, presumptive tests have been used for body fluid recognition, but immunochromatographic lateral flow strip tests are also commonly used in forensic routine work [6,7]. In addition, some laboratories have also implemented gene expression based methods using mRNA and miRNA markers [8,9]. Although not yet ready for casework, a novel microbiota-based recognition method is a promising alternative [10].

Microbiota sequencing was made possible with the introduction of massively parallel sequencing (MPS). In the last few years this field has had exceptional interest, and both human [11,12] and environmental microbiota [13] have been thoroughly explored. For the majority of these studies, hypervariable regions of the 16S ribosomal gene have been used for taxonomic identification. The microbiota refers to the taxonomic composition of a microbial community and is typically mapped by 16S amplicon sequencing, as opposed to the microbiome, which is the community genome, typically mapped by shotgun sequencing. Large amounts of raw data (reads) from 16S amplicon samples are now publicly available, e.g. in the NCBI/SRA database (https://www.ncbi.nlm.nih.gov/sra).

Tools for handling, preparation and analysis of 16S amplicon data are numerous, e.g. [14–17]. It has been shown earlier that body fluids can be recognized using standard methods for raw data processing, using Principal Component Analysis (PCA) in combination with Linear Discriminant Analysis (LDA) for pattern recognition [10]. However, the

[*] Corresponding author at: Department of Forensic Biology, Oslo University Hospital, P.O. Box 4950 Nydalen, N-0424 Oslo, Norway.
[**] Principal corresponding author.
E-mail addresses: kjeirik@gmail.com (E.N. Hanssen), lars.snipen@nmbu.no (L. Snipen).

https://doi.org/10.1016/j.fsigen.2018.07.012
Received 26 March 2018; Received in revised form 12 July 2018; Accepted 13 July 2018

potential of this approach can only be revealed by a systematic evaluation over larger data sets exploring various settings in the data analysis.

Microbiota sequencing has primarily been conducted to explore new microbial communities. The data processing pipelines for discovering and exploring the ecology are not necessarily optimal when it comes to recognizing already characterized communities. To recognize a body fluid from microbiota data is a classical pattern recognition problem. Pattern recognition is a branch of machine learning where a model utilizes regularities in a training data set to classify samples in a new test data set. The training and test data sets need to have the same format, with the same predictor variables. In the case of using microbiota data for body fluid recognition, the predictor variables are the taxonomic bins that the reads are assigned to. The taxonomic resolution determines the number of predictor variables. The standard pipelines will in general cluster reads into Operational Taxonomic Units (OTUs) using an identity threshold of 97% [18]. Since we cannot hope to find taxa which are easily detected and unique to any body fluid [19,20], we must rely on fairly stable patterns of high or low abundances of several taxa. The optimal taxonomic resolution must be fine-grained enough to produce abundance patterns that permit discrimination between body fluids, but still coarse enough to yield reproducible results over many samples. In addition to the taxonomic resolution, there are several choices for data transformation that will affect the precision of a method for microbiota-based body fluid recognition [21,22].

We have performed a systematic study on data processing and pattern recognition approaches and quantified their effects on microbiota-based body fluid recognition. We have used downloaded pristine 16S sequence data to support customization and optimization of our proposed model. The use of such samples is ideal for measuring the best possible model performance and can act as a reference point when the model is to be validated on real casework samples. Our findings will be implemented in an R-package for microbial forensics that we are developing.

2. Methods

2.1. Data

Public data from the Human Microbiome Project (HMP) [23] were downloaded from http://hmpdacc.org/. The HMP sequenced 16S amplicons from various body sites of hundreds of people, and we assembled these into four body fluids (oral, nasal, vaginal, fecal). The data set also includes samples from human skin. Such samples may also be relevant in a forensic setting and are included as a fifth category in addition to the four body fluids. In Table 1 we show how our categories include the original body sites annotated by HMP. The 16S gene has 9 hypervariable regions designated V1–V9, and the HMP amplicons are from two distinct regions, V1–V3 and V3–V5. Reads for 7333 samples were downloaded and used in this analysis. These public data have been subject to a careful preprocessing (de-multiplexing, removing contaminants, etc), see protocols at http://hmpdacc.org/ for all details. Each sample was downloaded as a FASTA file of reads, using the SRA toolkit (https://www.ncbi.nlm.nih.gov/sra). Table 1 shows a summary of the data.

To test the body fluid recognition performance with these data, we used a 10-fold cross-validation. The data were split into 10 non-overlapping subsets or segments, where each segment in turn was used as a test-set and the remaining data as a training-set. The samples were first sorted by body fluid and then by person within each body fluid. Next, they were given a segment-number from 1 to 10 repeatedly throughout the data set to achieve maximum spread of body fluids across segments. However, within each body fluid all samples from the same person were given identical segment-number, to ensure data from the same person and body fluid was found within one segment.

As an external test-set we also downloaded public data from the American Gut Project (AGP) [24], again using the SRA toolkit. These data differ from HMP in several ways. Different sequencing technologies (Illumina, while HMP uses Roche 454) were used to obtain short reads (~120 bp) from 9500 samples from almost as many persons. The vast majority of these samples are from feces, as suggested by the project title, but the data set also includes some samples from all the other categories in our study. Another difference is the extraction protocols used by the two projects (AGP uses the Mobio MagAttract PowerSoil kit while HMP uses the Mobio DNeasy PowerSoil Kit). The AGP data are from region V4 of the 16S gene. The two consortia use comparable sampling techniques, mainly swabs (see https://www.hmpdacc.org/hmp/doc/ and appendix K in $HMP_{M}OP$. *pdf* and http://americangut.org/how-it-works/).

2.2. Taxon read-counts

The first question we addressed was how to group the reads in a sample into a set of taxa. We focused on two distinct approaches, one of clustering into OTUs and one of direct taxonomic classification of all reads. We also explored different taxonomic resolution for both approaches.

A standardized pipeline for OTU-finding was set up using the VSEARCH software [17]. The training data set was first de-replicated to find unique reads (VSEARCH settings: −derep_prefix and −minuniquesize 2). These reads were then filtered for chimera (VSEARCH setting: −db gold.fasta [25]), clustered to find centroid sequences and singleton clusters discarded. The clustering was done using a standard 0.97 identity-threshold. Next, all reads in each training set sample were assigned to OTUs by searching against the centroid sequences, using same similarity threshold for clustering (VSEARCH setting: −usearch_global). For each test set sample, the OTU clustering was repeated using the same centroid sequences and similarity threshold as for the training set. The whole process was repeated using clustering identity-thresholds of 0.98 and 0.99.

Table 1
A summary where the data are categorized after whether the V1–V3 or the V3–V5 amplicon have been sequenced. For a number of the samples both amplicons were sequenced, and these are counted twice, one for each amplicon. Each sample belongs to a body fluid indicated in the left column. Then we list, for each body fluid, the number of persons contributing, the total number of samples, median number of reads per sample, average read-length (bases) in a sample and finally the original body site annotations given by the Human Microbiome Project. The numbers for V1–V3 are given before the slash and V3–V5 after (V1–V3/V3–V5).

Body fluid	Persons	Samples	Reads	Length	Original body site
Fecal	139/212	177/313	9392/8640	405/409	Stool
Nasal	131/195	155/266	8218/6997	390/423	Anterior nares
Oral	143/214	1515/2687	8554/7679	410/429	Attached/Keratinized gingiva, Buccal mucosa, Hard palate, Palatine tonsils, Saliva, Subgivingal plaque, Supragingivl plaque, Throat, Tongue dorsum
Vaginal	72/99	243/385	8381/7562	422/427	Mid vagina, Posterior fornix, Vaginal introitus
Skin	163/222	651/941	8576/6500	401/405	Antecubital fossa (left or right), Retroauricular crease (left or right)