Accurate Assignment of Short Pyrosequencing Reads in a 16S rRNA Taxonomy

José C. Clemente[1] jclement@lab.nig.ac.jp.jp
Jesper Jansson[2] jesper.jansson@ocha.ac.jp
Gabriel Valiente[3] valiente@lsi.upc.edu

[1] Center for Information Biology and DNA Databank of Japan National Institute of Genetics, Yata 1111, Mishima, Japan
[2] Graduate School of Humanities and Sciences, Ochanomizu University 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
[3] Algorithms, Bioinformatics, Complexity and Formal Methods Research Group Technical University of Catalonia, E-08034 Barcelona, Spain

Keywords: metagenomics, pyrosequencing, 16S rRNA, F-measure

1 Introduction

Metagenomic studies characterize bacterial communities using next generation sequencers to amplify the variety of 16S rRNAs present in a sample. The generated short fragments are assigned to the closest bacteria in a taxonomy obtained from full 16S rRNA sequences, but a significant proportion of the fragments can be assigned to more than one species. Previous studies have generally mapped those fragments to the lowest common ancestor (LCA) of the matched species in the taxonomy [3], which implicitly assumes that coverage should be maximized at the cost of minimizing accuracy. We present an assignment algorithm that maps each fragment to the node in the taxonomy that maximizes the F-measure in time linear in the size of the subtree rooted at the LCA of the matching sequences.

2 Method and Results

Given a reference taxonomy T, a set R of short reads, and a threshold value k of sequence similarity, let R_i be the ith read, let M_i be the leaves of T matching R_i with up to k mismatches, let T_i be the subtree of T rooted at the lowest common ancestor of M_i, and let N_i be the leaves of T_i not matching R_i with up to k mismatches. Let also $L_i = M_i \cup N_i$. Further, let $T_{i,j}$ be the subtree of T rooted at the jth node of T_i, let $M_{i,j}$ be the leaves of $T_{i,j}$ matching R_i with up to k mismatches, and let $N_{i,j}$ be the leaves of $T_{i,j}$ not matching R_i with up to k mismatches. For the ith read and the jth node of T_i, the leaves of T_i can be partitioned into the subsets of true positives ($TP_{i,j} = M_{i,j}$), false positives ($FP_{i,j} = N_{i,j}$), true negatives ($TN_{i,j} = N_i \setminus N_{i,j}$), and false negatives ($FN_{i,j} = M_i \setminus M_{i,j}$). The precision of classifying R_i as T_j is $P_{i,j} = |TP_{i,j}|/(|TP_{i,j}| + |FP_{i,j}|)$, and the recall is $R_{i,j} = |TP_{i,j}|/(|TP_{i,j}| + |FN_{i,j}|)$. The combined F-measure of precision and recall is $F_{i,j} = 2P_{i,j}R_{i,j}/(P_{i,j} + R_{i,j}) = 2|M_{i,j}|/(|L_{i,j}| + |M_i|)$.

Using a high quality bacterial taxonomy based on the 16S rRNA of 5,165 species [1] with a uniform scheme of seven taxonomic ranks (domain, phylum, class, order, family, genus, species), we mapped all reads from six different metagenomics studies to the taxonomy using the GEM tools [5]. Reads that could not be uniquely mapped to a single species were then assigned at the best taxonomic rank using both the LCA approach and our algorithm. As shown in Table 1, only 13.60% of the marine ambiguos reads (3,213 out of 23,612), 4.63% of the human gut ambiguous reads, 4.51% of the human twins gut (V2 region), 0.67% of the human twins gut (V6 region), 1.02% of the chicken gut ambiguous reads, and 0.15% of the rat gut ambiguous reads were actually assigned to the LCA of the matching

Table 1: Number of ambiguous pyrosequencing reads assigned at various taxonomic ranks using the LCA (left) and our algorithm (right) of the matching sequences in the reference bacterial taxonomy.

rank	number of reads											
	marine V6[6]	human twins V6,V3[2]	twins V2[7]	twins V6[7]	chicken V6[8]	rat V4[4]	marine V6[6]	human twins V6,V3[2]	twins V2[7]	twins V6[7]	chicken V6[8]	rat V4[4]
domain			40			1						
phylum	29	5,498	3	13,133	130	49						
class	12,099	2,354		1,854	154	3			2			
order	976	5	13	8	8	35			4			2
family	3,428	49,647	371	2,343	1,441	3,582	860	2,150	16	195	3	57
genus	7,089	33,831	349	77,661	2,662	27,839	17,705	8,441	411	2,353	210	3,622
species							5,056	80,744	343	92,451	4,182	27,828
	23,621	91,335	776	94,999	4,395	31,509	23,621	91,335	776	94,999	4,395	31,509

sequences using our method. The remaining ambiguous reads were assigned at a deeper taxonomic rank than the LCA of the matching sequences using our approach. While assigning a read to the LCA of the matching sequences tends to produce assignments at the ranks of class, order, family, and genus, the new method produces more accurate assignments at the ranks of genus and species.

3 Discussion

We have shown that our algorithm can accurately map each read to the node with best combined value of precision and recall, and that depending on the assignment schema for ambiguous fragments the distribution of taxonomic ranks varies greatly. This has important consequences for metagenomic studies drawing consequences from the distribution of bacterial species in an environment, such as the correlation between sick conditions and the diversity of bacteria in the human gut.

References

[1] Cole, J. R. et al. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, 37(D):141–145, 2009.

[2] Dethlefsen, L. et al. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology*, 6(11):e280, 2008.

[3] Huson, D. H. et al. MEGAN analysis of metagenomic data. *Genome Res.*, 17(3):377–386, 2007.

[4] Manichanh, C. Rat intestinal microbiota. Private communication, 2009.

[5] Ribeca, P. GEM—GEnomic Multi-tool. `http://gemlibrary.sourceforge.net/`, 2009.

[6] Sogin, M. L. et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA*, 103(32):12115–12120, 2006.

[7] Turnbaugh, P. J. et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

[8] VAMPS. Visualization and analysis of microbial population structure project. AGT_CKN_Bv6—Chicken intestinal microbiota, 2009. `http://vamps.mbl.edu/`