ESPRIT-Tree: Hierarchical Clustering Analysis of Millions of 16S rRNA Pyrosequences in Quasilinear Computational Time (Supplementary Data)

Yunpeng Cai and Yijun Sun*

Interdisciplinary Center for Biotechnology Research
University of Florida, Gainesville, FL 32610

1 Probabilistic Matrix

Table 1: Illustration of the procedure of constructing a probabilistic sequence **x** by aligning two sequences **a** and **b**. '-' represents a gap and P(s=A) represents the probability of observing nucleotide A.

seq **a**	A	T	C	G	A	T	C	G	G	G	G
seq **b**	G	T	C	G	–	T	C	G	T	G	–
seq **x**	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
P(s=A)	0.5	0	0	0	0.5	0	0	0	0	0	0
P(s=T)	0	1	0	0	0	1	0	0	0.5	0	0
P(s=C)	0	0	1	0	0	0	1	0	0	0	0
P(s=G)	0.5	0	0	1	0	0	0	1	0.5	1	0.5
P(s=gap)	0	0	0	0	0.5	0	0	0	0	0	0.5

2 Normalized Mutual Information

Suppose that we have a sequence dataset consisting of N sequences. Let $\mathcal{C} = \{c_1, \cdots, c_J\}$ and $\Omega = \{\omega_1, \cdots, \omega_K\}$ be the clustering outcome and the ground-truth partition of the input sequences, respectively. NMI is computed as:

$$\text{NMI}(\Omega, \mathcal{C}) = \frac{2I(\Omega|\mathcal{C})}{H(\Omega) + H(\mathcal{C})} \ . \tag{1}$$

*Please address all correspondence to: Dr. Yijun Sun, Interdisciplinary Center for Biotechnology Research, University of Florida, P. O. Box 103622, Gainesville, FL 32610, USA. E-mail: sunyijun@biotech.ufl.edu. Tel: 352-273-8271, Fax: 352-273-8070. Y. Cai and Y. Sun contributed to the paper equally.

$I(\Omega|\mathcal{C})$ is the mutual information computed as

$$I(\Omega|\mathcal{C}) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{2}$$

where $|\omega_k \cap c_j|$ is the number of sequences included in the intersection of ω_k and c_j. $H(\Omega)$ is the entropy, given by

$$H(\Omega) = -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} \ . \tag{3}$$

$I(\Omega|\mathcal{C})$ measures the amount of information one has about the ground-truth partition Ω by knowing the clustering outcome $\mathcal{C}$. It is normalized by $(H(\Omega) + H(\mathcal{C}))/2$ so that clustering results with different numbers of clusters can be compared. NMI $= 1$ if $\Omega = \mathcal{C}$, and NMI $= 0$ if the sequences are randomly grouped since one gains no information about Ω from $\mathcal{C}$. For a more detailed description, interested reader may refer to [1, 2].

3 Additional Experiments

We performed additional experiments using other hypervariable regions and full-length 16S rRNA sequences. The experimental protocol is exactly the same as that used in the main text. Table 2 summaries the datasets we used. The benchmark results are given in Figure 1. Since CD-HIT performed worse than both UCLUST and ESPRIT-Tree, the results of CD-HIT are omitted. The results are consistent with those obtained on the human gut microbiota dataset. In all cases, ESPRIT-Tree performed similarly to ESPRIT-AL, and significantly better than UCLUST.

Table 2: Summary of Data Sets

Data	Region	# Reads	Num. Annotated	Ave. Len	# Species
ELDERMET (part) [3]	V4	333,383	143,687	242	332
Sea Water [4]	V6	22,2291	71,180	62	759
Crohn's Disease [5]	V6-V9	202,073	156,059	477	882
Bowel [6]	near full	45,351	23,871	1,081	868

References

[1] Manning CD, Raghavan P, Schutze H. (2008) *Introduction to Information Retrieval*. Cambridge University Press.

[2] Fred ALN, Jain AK. (2003) Robust data clustering. in *Proc. IEEE Conf Comp Vision Patt Recogn* **3**: 128-36.

[3] Claesson MJ, Cusack S, O'Sullivan O, Greene-Diniz R, de Weerd H, Flannery E, Marchesi JR, Falush D, Dinan T, Fitzgerald G, Stanton C, van Sinderen D, O'Connor M, Harnedy N, O'Connor K, Henry C, O'Mahony D, Fitzgerald A, Shanahan F, Twomey C, Hill C, Ross RP, O'Toole PW (2010) Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc Natl Acad Sci USA* doi: 10.1073/pnas.1000097107.

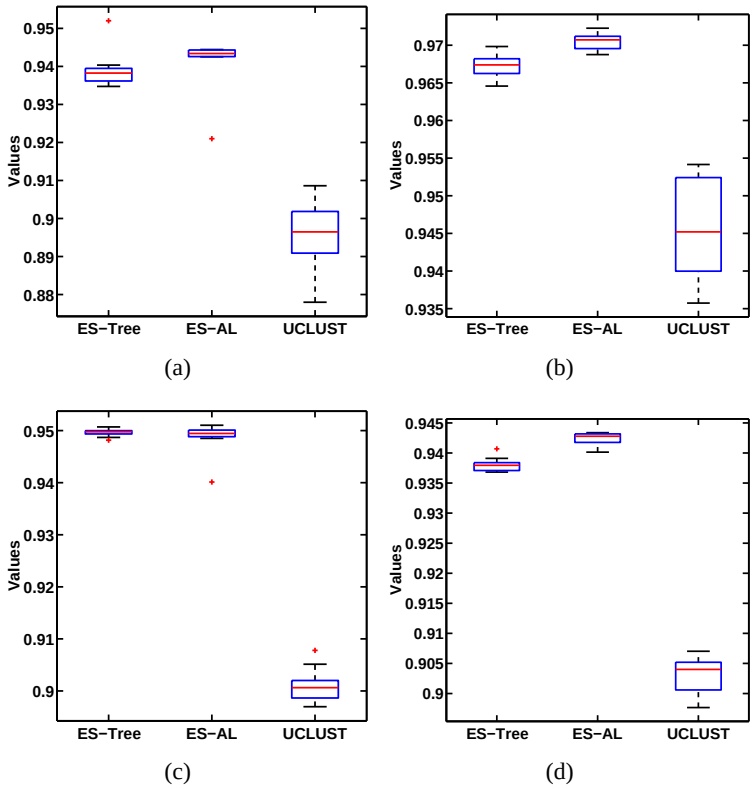

Figure 1: Box plots of the maximum NMI scores of ESPRIT-TREE, ESPRIT-AL and UCLUST obtained by using (a) v4, (b) v6, (c) v6-v9 hypervariable regions and (d) full-length reads of 16S rRNA. The species assignments of input sequences were used as ground truth.

[4] Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc Natl Acad Sci*. **103**:12115-20.

[5] Ellen Li, et. al. (2010) Effect of Crohn’s Disease Risk Alleles on Enteric Microbiota. *NIH Project No. 1UH2DK083994-01*. NCBI accesssion NO. SRX021354

[6] Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104**(34):13780-5.