Data Format for Tree Structured Data.

Bayesian Approaches for Flexible and Informative Clustering of Microbiome Data by Yushu Shi, Liangliang Zhang, Kim-Anh Do, Robert Jenq and Christine Peterson.

1. Data Structure. We denote the original $p \times n$ OTU matrix as $\mathbf{Y}$, where p is the number of OTUs, and n is the number of observations. There is also a tree T describing the relationship of OTUs. The total number of tree nodes is denoted as d. Given the tree structure, we can expand the OTU table $\mathbf{Y}$ to a 3-dimensional $(d, d+p-1, n)$ matrix $\mathbf{X}$. $\mathbf{X}$ can be considered as a collection of n sets of $p \times (d+p-1)$ 2-dimensional matrices, where n is the number of subjects. The first dimension, denotes d parent nodes. Second dimension, denotes $d+p-1$ children node or leaves. (Minus 1 because the root could not be children node of other nodes). For each given subject i, the j,kth entry of his/her matrix is the number of sequences passing from node j to node/leave k.

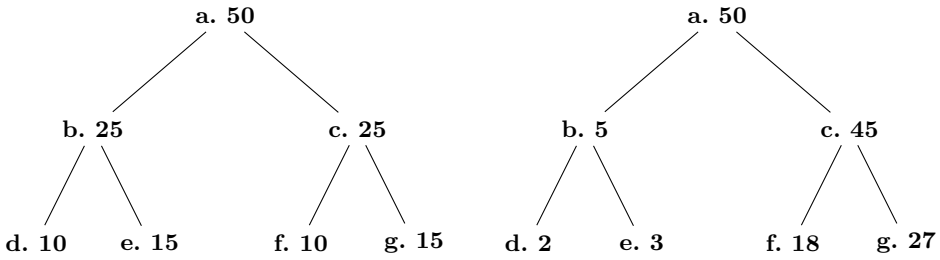

Fig 1: The above figures show the sequence allocation of two example observations with the same tree

We use a toy example with two observations to illustrate how we convert the OTU phylogeny information into matrices. Figure 1 shows the sequence allocation from the root to each OTU leaf for two observations. Two tables below give the created corresponding sparse matrices. For convenience, one can save the data of each observation as a single matrix.

	b	**c**	**d**	**e**	**f**	**g**
a	25	25	0	0	0	0
b	0	0	10	15	0	0
c	0	0	0	0	10	15

	b	**c**	**d**	**e**	**f**	**g**
a	5	45	0	0	0	0
b	0	0	2	3	0	0
c	0	0	0	0	18	27