

# Nonmetric Distances for Barcode of Life

Hisamitsu Akiba      Y-h. Taguchi  
tag@granular.com

Department of Physics, Faculty of Science and Technology, and Institute for Science and Technology, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan.

**Keywords:** nonmetric multidimensional scaling, clade, barcode of life

## 1 Introduction

Barcode of Life (BOL) project[4] is the project to enable us to recognize species easier. Although it is often troublesome to define what the species are, BOL can define species by simple DNA sequences. When it works, we do not have to consult with any other information than DNA sequences to decide if two individuals belong to the same species or not. If they share same BOL with each other, they belong to the same species undoubtedly.

In contrast to this, it is usually difficult to define what the higher clade are. We cannot expect that each individual which belong to the same upper Claude share the same BOL. Instead, we have to find how BOL of individuals which belong to distinct higher clade differ from each other. In this poster, we demonstrate how nonmetric measure of distances between BOL make easier to recognize if each belongs to common higher clade or not. We also show that usual hierarchical clustering like NJ method is not suitable to visualize relationships expressed by nonmetric measure and propose to usage of nonmetric multidimensional scaling (nMDS)[1, 2].

## 2 Method and Results

We have downloaded BOL sequences from The Barcode of Life Data Systems[4] and computed Kimura's two parameter distances between sequences after multiple alignment by clustal W[5]. In order to get nonmetric measure of distances, we have ranked distances and employed them as non-metric measures.

## 3 Discussion

In order to see how well higher clades can be distinguished from each other, it is suitable to require the following condition,

$$\frac{D_i + D_j}{2} < D_{i+j}, \quad (1)$$

where  $D_i$  is the mean distances between pairs within clade  $i$  and  $D_{i+j}$  is that within combined set of clades  $i$  and  $j$ . In Figs. 1, we have shown pairs of clades which violate this condition. Clearly, Kimura's two parameter distance violate eq. (1) more frequently than rank order. Similar tendency can be also found for both genus and class of Birds of North America (TZBNA)[4]. Thus, it is clear that rank (nonmetric measure) is better to distinguish between higher clades than Kimura's two parameter distances.

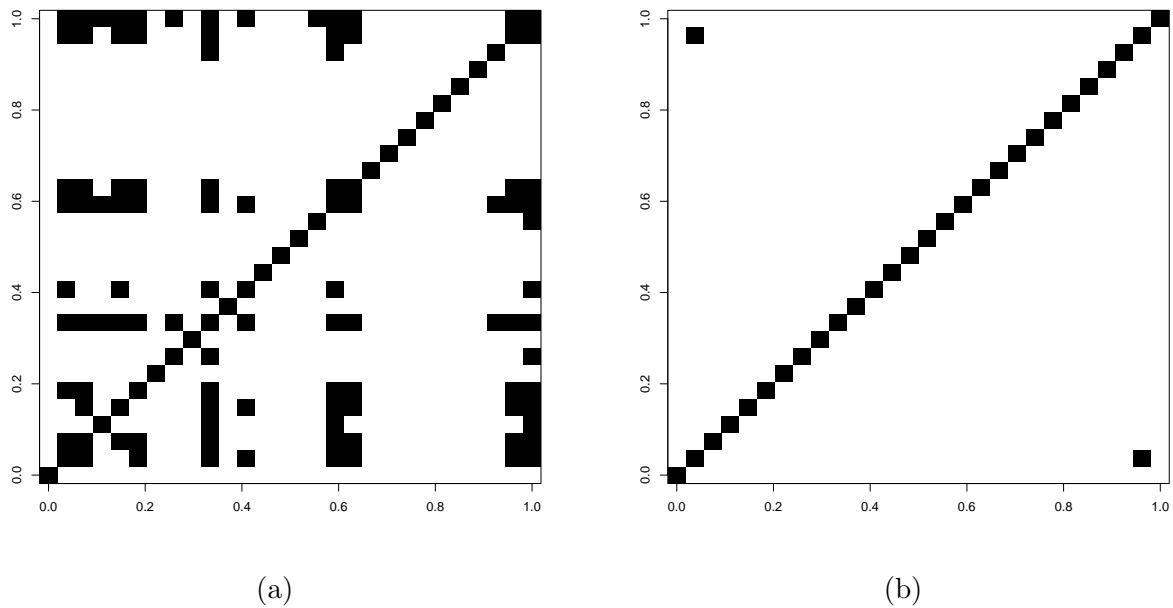


Figure 1: Pairs which violate eq. (1) for Ant Diversity in Northern Madagascar (JDWAM)[4] are indicated by filled squares. Both horizontal and vertical axes represent each genus. (a) Kimura's two parameter distance (b) rank order

Then, although we have tried to get phylogenetic tree by using NJ method applied to rank, it turns out that NJ cannot construct phylogenetic tree well by employing rank as distances. In order to overcome this difficulties, we have applied nMDS[1, 2] to visualize the relationship between higher clades. Since nMDS uses only rank order of dissimilarities, it is suitable to visualize the relationship obtained by rank order of distances. Although we cannot show here the nMDS results because of lack of space, nMDS turns out to be better to deal with rank order than NJ methods.

## References

- [1] Taguchi, Y-h., and Oono, Y., Nonmetric Multidimensional Scaling as a data-mining tool: new algorithm and new targets, *Advances in Chemical Physics*, 130B:315–351, 2005.
- [2] Taguchi, Y-h., and Oono, Y., Relational patterns of gene expression via non-metric multidimensional scaling analysis, *Bioinformatics*, 21(6):730–740, 2005.
- [3] <http://www.jsbi.org/>
- [4] <http://www.barcodinglife.org/>
- [5] <http://www.ebi.ac.uk/clustalw/>