2007－BIO－9 (5)
2007／6／14

Barcode of Lifeによる非計量法の利用

秋葉寿光[1], 田口善弘[1,2]

[1] 中央大学理工学部物理学科,[2] 中央大学理工学研究所

Barcode of Life(BOL)projectとは生物の持つDNA配列を用いることで、他の情報を必要とせず種の識別を容易に行う計画である。我々はこのBOLにさらに高次の分類群が含まれていることを仮定した。今研究ではBOLに非計量の距離法を用いることで、高次分類群の区別が容易になり、また個体間の関係性を可視化するために、非計量多次元尺度構成法(nMDS)への応用が有用であることを示している。

Nonmetric distances for Barcode of Life

Hisamitsu Akiba[1], Y-h. Taguchi[2]

[1]Dept. Phys., Chuo. Univ., akiba@phys.chuo-u.ac.jp
[2]Dept. Phys., Chuo. Univ., tag@granular.com

Barcode of Life (BOL) project aims to identify species with no other information than DNA sequence. We assume that BOL includes information about higher taxa. In this study, we compute non-metric distance from BOL bar codes. This enables us to recognize higher taxa more easily. Non-metric multidimensional scaling method turns out to be useful to visualize the relationship among individual organism hidden in non-metric distances.

1 Introduction

Barcode of Life (BOL) project[3] 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 taxa are. We cannot expect that each individual which belongs to the same higher taxa share the same BOL. Instead, we have to find how BOL of individuals which belong to distinct higher taxa differ from each other. In this paper, we demonstrate how nonmetric measure of distances between BOL make easier to recognize if each belongs to common higher taxa 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 Materials and Methods

We have downloaded BOL sequences from The Barcode of Life Data Systems[3]. Used data sets are Ant Diversity in Northern Madagascar (JDWAM), Survey of Chelicerates (RBCH), and Birds of North America (TZBNA) (Table 1). Then Kimura's two parameter distances d_{ij}, between ith and jth sequences are computed after multiple alignment by

General Projects	Code	Sequences (N)	Species
Ant Diversity in Northern Madagascar	JDWAM	268	86
Survey of Chelicerates	RBCH	214	110
Birds of North America	TZBNA	437	263

Table 1: BOL[3] sequences used in this study

clustal W[4]. In order to get nonmetric measure of distances, we have ranked distances as $n_{ij}, (1 \leq n_{ij} \leq N(N-1)/2)$, where N are total number of sequences. These rank n_{ij}s are employed as nonmetric measures.

3 Results

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

$$D_I + D_J < D_{I+J}, \tag{1}$$

where D_I is the mean distances between pairs within taxon i,

$$D_I(\delta_{ij}) \equiv \frac{1}{N_I(N_I-1)/2} \sum_{(i,j)\in I} \delta_{ij},$$

where δ_{ij} is distance (i.e., d_{ij} or n_{ij}) between ith sequence and jth sequence, N_I is the number of sequences in Ith higher taxon, and D_{I+J} is that between sets of taxa I and J,

$$D_{I+J}(\delta_{ij}) \equiv \frac{1}{N_I N_J} \sum_{i\in I} \sum_{j\in J} \delta_{ij}.$$

In Figs. 1, we have shown pairs of taxa which violate the condition eq. (1). Clearly, Kimura's two parameter distance violate eq. (1) more frequently than rank order. This tendency can be found independent of the taxa considered (genus, order or family) and kinds of organism (insect, bird, and so on). In Figs. 2 we have shown direct comparison between $D_I(\delta_{ij})$ computed from Kimura's two parameter distance d_{ij} and that from rank order n_{ij}. In these figures, we have shown normalized distances,

$$\hat{D}_I(\delta_{ij}) \equiv \frac{D_I(\delta_{ij})}{\frac{1}{N}\sum_{ij}\delta_{ij}}$$

instead of D_I. Thus, $0 \leq \tilde{D}_I(\delta_{ij}) \leq 1$ and it is possible to compare mean distance given by d_{ij} with that by n_{ij}. In these figures, almost always,

$$\hat{D}_I(n_{ij}) < \hat{D}_I(d_{ij})$$

stands. Thus, it is clear that rank (nonmetric measure) is better to distinguish between higher taxa than Kimura's two parameter distances.

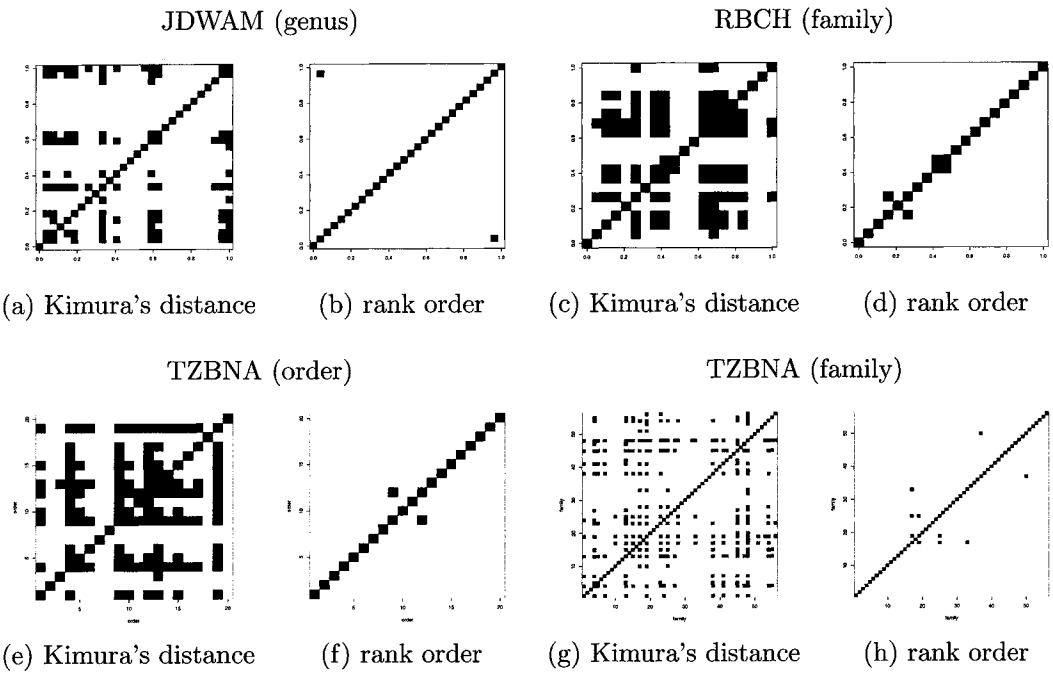

Figure 1: Pairs which violate eq. (1) for Ant Diversity in Northern Madagascar (JDWAM), Survey of Chelicerates (RBCH), Birds of North America (TZBNA) are indicated by filled squares, Both horizontal and vertical axes represent each higer taxon (genus, family, or order). (a), (c), (e), (g) : Kimura's two parameter distance, (b), (d), (f), (h) : rank order. (a) and (c) : JDWAM(genus), (d) and (e) : RBCH(family), (e) and (f) : TZBNA (order), (g) and (h) : TZBNA (family)

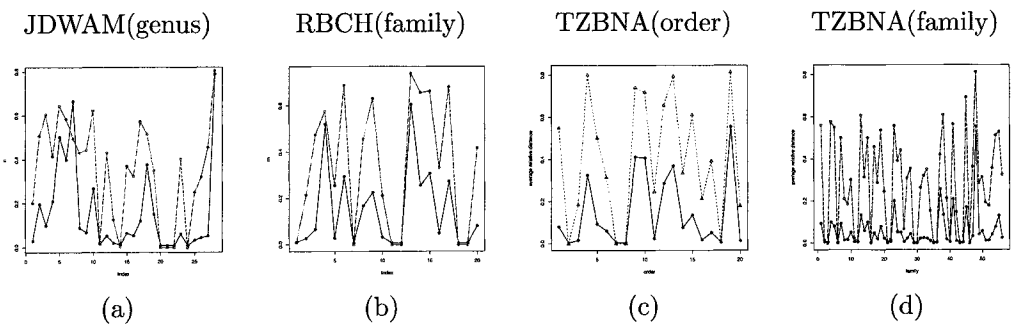

Figure 2: Comparison between mean Kimura distance $\hat{D}_I(d_{ij})$ within each taxon (thin lines) and mean rank $\hat{D}_I(n_{ij})$ within each taxon (solid lines) for Ant Diversity in Northern Madagascar (JDWAM), Survey of Chelicerates (RBCH), Birds of North America (TZBNA). The vartical axes are normalized by the mean distance over all pairs.

d	number of correctly discriminated sequences	leave one out cross validarion
10	188 [88 %]	177 [85 %]
15	194 [91 %]	166 [81 %]
20	194 [91 %]	170 [83 %]
30	208 [97 %]	183 [89 %]
40	208 [97 %]	181 [88 %]
MP	199 [93 %]	
NJ	200 [93 %]	

Table 2: Number of sequences correctly discriminated by LDA as a function of embedding dimension d for RBCH data set (Survey of Chelicerates). Leave one out cross validation results are shown, too. For the leave one out cross validation, 5 families with only one sequence are excluded. For comparison, results for maximum parsimony (MP) and neighboring joining (NJ) with visual inspection is shown.

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 taxa. Since nMDS uses only rank order of dissimilarities, it is suitable to visualize the relationship obtained by rank order of distances.

By applying nMDS to Kimura's two parameter distances for RBCH data set, we get d dimensional vector for each sequence. We have checked if each taxon in d dimensional space is separated well, linear discriminant analysis (LDA) is applied to obtained d dimensional embedding. For LDA, we have employed lda module in R[5]. For $d \geq 30$, number of sequences which are discriminated correctly is 208 among 214. Thus, 97 % of sequences is separated well so as to be included into correct taxon. In Table 2, we have shown that dependency of number of correctly discriminated sequences upon embedding dimension d. Leave one out cross validation results show the sensitivity that unclassified sequence is discriminated correctly[1]. When embedding dimension is larger than 30, almost 90 % of sequences can be correctly classified. Thus, nMDS with Kimura's two parameter distance turns out to be useful tool to clarify sequence into one of higher taxa.

In Table 3, we have shown that true vs predicted family by LDA for RBCH data set (Survey of Chelicerates), for $d = 40$. In the last row and column, we have shown sensitivities TP/(TP+FN) and precision TP/(TP+FP) respectively. TP is the number of sequences which belong to a family and are classified into the family, FN is the number of sequences which belong to a family but are not classified into the family, and FP is the number of sequences which do not belong to a family but are classified into the family. Clearly, almost all family are 100 % correctly discriminated.

4 Discussion

It is natural to wonder why rank order works better than raw distance. In Fig. 3, we have shown typical example of histograms of intra-order, inter-order, and all over distances for

[1]5 families with only one sequence are excluded for the leave one out cross validation

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	TP/(TP+FN)
1	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	**14**	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	93.3
4	0	0	0	**9**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	90
5	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	**34**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	100
9	0	0	1	0	0	0	0	0	**24**	0	0	0	0	0	0	0	0	0	0	2	96
10	0	0	0	0	0	0	0	0	0	**5**	0	0	0	0	0	0	0	0	0	0	100
11	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	100
12	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	100
13	0	0	0	0	0	0	0	0	0	0	0	0	**25**	0	0	0	0	0	0	0	100
14	0	0	0	0	0	0	0	0	0	1	0	0	0	**28**	0	0	0	0	0	0	96.5
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**11**	0	0	0	0	0	100
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	100
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**20**	0	0	0	100
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	100
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	100
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**13**	100
TP/(TP+FP)	100	100	93.3	100	100	100	100	85.7	96	83.3	100	100	100	100	100	100	100	100	100	86.7	

Table 3: True (row) vs predicted (column) family for RBCH data set (Survey of Chelicerates) in 40 dimensional embedding space by nMDS. Each family is numbered so as to be compared with phylogenetic tree (Fig.4). Bold number is TP for each family.

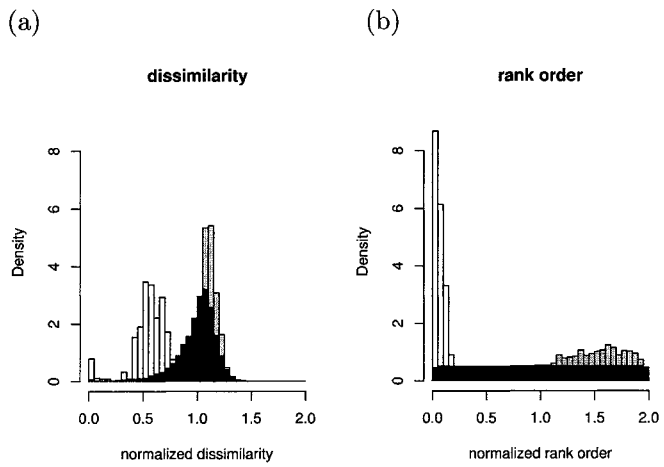

Figure 3: A typical example of histogram of intra-order (white), inter-order (grey), and over all (black) distances for Birds of North America (TZBNA). (a) Kimura's two parameter (b) Rank.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	?	TP/(TP+FN)
1	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
2	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
3	0	0	**15**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
4	0	0	0	**2**	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	20
5	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
6	0	0	0	0	0	**34**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
7	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
8	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0	0	2	66.7
9	0	0	0	0	0	0	0	0	**25**	0	0	0	0	0	0	0	0	0	0	2	0	92.6
10	0	0	0	0	0	0	0	0	0	**5**	0	0	0	0	0	0	0	0	0	0	0	100
11	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	100
12	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	100
13	0	0	0	0	0	0	0	0	0	0	0	0	**25**	0	0	0	0	0	0	0	0	100
14	0	0	0	0	0	0	0	0	0	1	0	0	0	**28**	0	0	0	0	0	0	0	96.6
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	2	81.2
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	100
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**20**	0	0	0	0	100
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	100
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	100
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**13**	0	100
TP/(TP+FP)	100	100	100	100	100	85	100	100	100	83.3	100	100	100	100	100	100	100	100	100	100		

Table 4: True (row) vs predicted (column) family for RBCH data set (Survey of Chelicerates) by the visual inspection for MP result. Each family is numbered so as to be compared with phylogenetic tree (Fig. 4). Bold number is TP for each family. The column labeled by ? indicates the sequences which do not seem to belong to any groups.

Birds of North America (TZBNA). In order to draw this histogram, we have pick up two orders, and distribution of distance within one of orders is drawn as an example of intra-order distance, and that between two orders is drawn as an example of inter-order distance. For raw distance (Kimura's two parameter), all of them obeys Gaussian like distributions and differences between intra-order and inter-order distances are small. However, once raw distances are converted to rank order, histogram of over all distances is flatten, thus the difference between intra-order and inter-order distances are enhanced. As can be seen in Figs. 2 (c), mean intra-order distance is substantially smaller than over all distances. Under such a circumstance, converting raw distance into rank enable us to discriminate higher taxa (e. g., family).

In order to see if standard phylogenetic tree can discriminate taxon better than proposed method, we have used PAUP[6] to apply maximum parsimony (MP) method to RBCH data set (Survey of Chelicerates). In Fig. 4, we have shown resultant phylogenetic tree. Although generally each family is separated with each other very well, there are some exceptions. For example, family 4 is completely in the family 6, thus it cannot be regarded as independent family. However, in Table 2, TP/(TP+FN) and TP/(TP+FP) are 90 % and 100 % respectively. This means, our method can discriminate family batter than MP.

In order to see this point more clearly, we have divided tree into a set of monophylic groups such that each family is equal to each of these sets (by visual inspection). These procedures enable us to have Table like Table 3 for MP results (See Table 4). If we compare Table 4 with Fig. 4, one can understand the problem of phylogenetic tree. For example, most sequences in 4th family(6 out of 10) are under 6th family group in Fig. 4, upper limit of TP/(TP+FN) is 40 % for 4th family. Furthermore, the remaining 4 sequences are not grouped together, TP/(TP+FN) is limited to 20 % for MP result. The same occurs

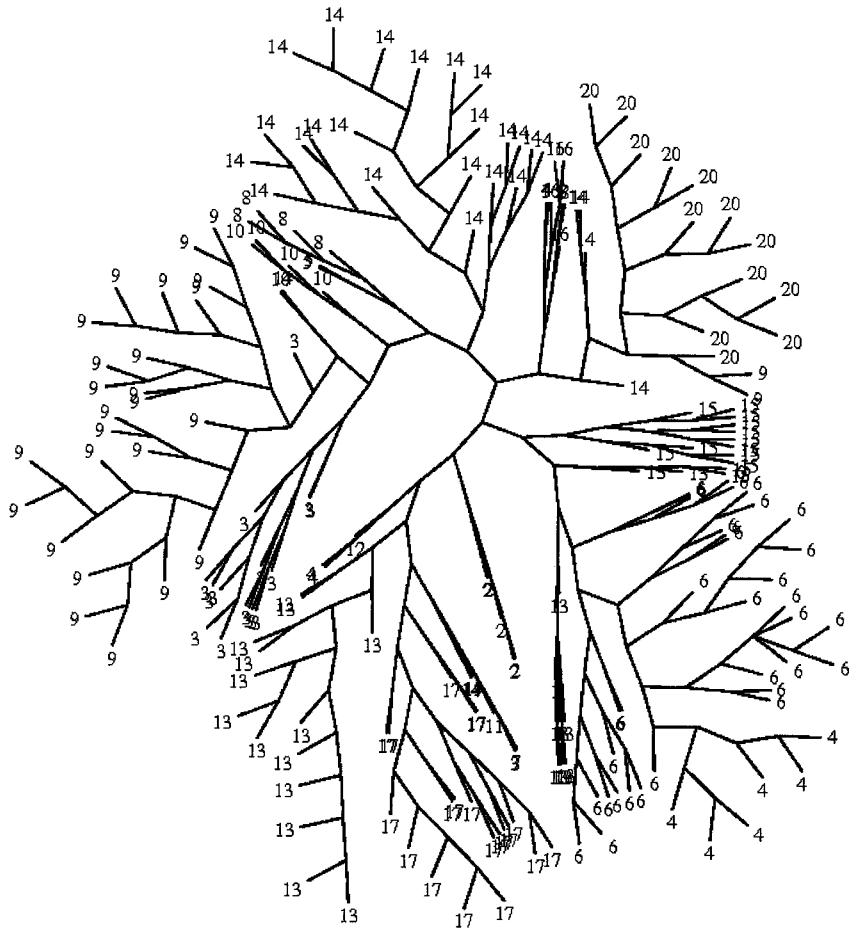

Figure 4: Phylogentic tree for RBCH data set (Survey of Chelicerates) obtained by MP method. The number indicates family.

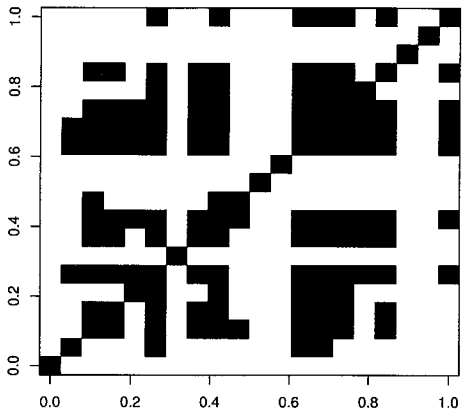

Figure 5: Pairs which violate eq. (1) for RBCH data set (Survey of Chelicerates) when employing distances along phylogenetic tree, Fig. 4.

for 2 out of 27 sequences in 9 he family is located next to 20th family. On the other hand, some sequences cannot be repeated to belong to any of groups (the column label by ? in Table 4). These do not happen for Table 3 (our result). As a result, the number of pairs of orders which violate eq. (1) are more than that in Fig. 1 (d) (see Fig. 5). Thus, we can conclude that our method discriminate order better than MP method.

5 Acknowledgement

This work has been partially supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.

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.barcodinglife.org/

[4] http://www.ebi.ac.uk/clustalw/

[5] http://www.R-project.org/

[6] http://paup.csit.fsu.edu/