Comparison Of Three Numerical Representations Of Mitochondrial COI Genes For Species Similarity/Dissimilarity Analysis

Maděránková D [1], Provazník I [1]
[1] Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering, Brno, Czech Republic
maderankova@phd.feec.vutbr.cz

We report comparison of three numerical representations of DNA sequences applied on mitochondrial COI genes of nine bird species for similarity/dissimilarity analysis based on distance matrix method. Mitochondrial COI genes are used for DNA barcoding that means quick species identification of unknown sample. Another intent of DNA barcoding is identification of new and close related species. Numerical representation of DNA sequences is appropriate step before computational processing of the sequences. Results of similarity/dissimilarity analysis of sequences in numerical representations were also compared with phylogenetic three based on Jukes-Cantor method.

1 Introduction

Intent of DNA barcoding is large-scale screening of mitochondrial gene for cytochrome c oxidase I (COI) of all species and creating databases of COI sequences for assigning of unknown individuals to species and discovery of new species. Concept of DNA barcoding is not fundamentally new but usage of mitochondrial COI gene has many advantages. COI is very short genetic sequence of 648 base-pairs which can be easily extracted from cells. This gene is suitable for identification of animals, especially birds, fish and insect, but it is not suitable for plants. As COI is mitochondrial gene, it evolves quickly and therefore it is possible to identify close related species and new species [1].

Numerical representation of DNA sequences is appropriate step before computational processing of the sequences which can be very effective and quick, particularly for great data volume (number of sequences or length of sequence). There is a number of numerical maps that can be used but the best ones should keep the informative content of sequences, especially biochemical characteristic of nucleotides. The nucleotides can be sorted into three classes according to: 1) molecular structure – adenine and guanine are purines (R), cytosine and thymine are pyrimidines (Y); 2) strength of links – adenine and thymine are linked by two hydrogen bonds (W), cytosine and guanine are linked by three hydrogen bonds (S); 3) radical content – adenine and cytosine contain the amino group (M), thymine and guanine contain keto group (K) [2].

In this paper, we report comparison of three numerical representations of COI gene sequences of nine bird species on the basis of similarity/dissimilarity analysis based on distance matrix and band average widths. The results of the three numerical representations were also compared with phylogenetic three.

2 Numerical representations of DNA sequences

Three numerical representations of DNA sequences were chosen. The first numerical representation creates 4-D vector and assign coordinates to each nucleotide as follows [3]:

$$A \to (1,0,0,0) \quad C \to (0,1,0,0) \quad G \to (0,0,1,0) \quad T \to (0,0,0,1)$$

The assignment of nucleotide *i* is added to previous assignment *i*-1. Table 1 shows example of this first numerical representation for short sequence 'ACGT'.

Sequence	A	C	G	T
A	1	0	0	0
C	1	1	0	0
G	1	1	1	0
T	1	1	1	1

Tab 1. Example of the first numerical representation.

The order of nucleotides in columns is arbitrary and it does not carry any information.

The second numerical representation creates 3-D vector and assigns coordinates to each nucleotide as follows [4]:

$$A \to (0,0,0) \quad C \to (1,0,1) \quad G \to (0,1,1) \quad T \to (1,1,0)$$

This numerical representation carry information about chemical characteristics of the nucleotides. The first coordinate indicates R/Y characteristic, the second coordinate indicates M/K and the third indicates W/S. The order of the coordinates is arbitrary and it doesn´t affect subsequent similarity/dissimilarity analysis. As in the first representation, the assignment of nucleotide *i* is added to previous assignment *i*-1. Table 2 shows example of the second numerical representation for short sequence 'ACGT'.

Sequence	R/Y	M/K	W/S
A	0	0	0
C	1	0	1
G	1	1	2
T	2	2	2

Tab 2. Example of the second numerical representation.

The third numerical representation is almost the same as the second one but it adds forth coordinate which extends along sequence [4]. Table 3 shows example of the third numerical representation for short sequence 'ACGT'.

Sequence	R/Y	M/K	W/S	k
A	0	0	0	1
C	1	0	1	2
G	1	1	2	3
T	2	2	2	4

Tab 3. Example of the third numerical representation.

3 Similarity/dissimilarity analysis

When the DNA sequence is numerically represented, distance matrix based on Euclidean distances is created:

$$D_{i,j} = \sqrt{\sum_n \left(x_{ni} - x_{nj}\right)^2} ,$$

where n is number of coordinates, x_{ni} and x_{nj} are the nth coordinate of the ith and jth nucleotide in the sequence.

Band average widths are calculated from the distance matrix as sums of elements in diagonals parallel to the main diagonal divided by a number of elements [4].

Similarity/dissimilarity can be quantitatively measured by Euclidean distances between sequences band average widths:

$$s_{a,i} - s_{b,i} = \Delta s_i = \sqrt{\Delta s_1^2 + \Delta s_2^2 + ...\Delta s_n^2} ,$$

where $i = 1, 2, \ldots, n$ and $n = 10$ (average window).

4 Results

The three numerical representations were applied on COI gene sequences of nine bird species. Then the similarity/dissimilarity analysis was done. The sliding average window n=10 was used. The table 4, 5 and 6 show relative similarities of the nine bird sequences as sums of Euclidean distances between band average widths. The reference sequences are in the first row of the tables. The most similar sequences to the reference sequence are in the upper part of the tables, the less similar sequences are in the bottom part.

(Legend of tables: Ac_gen = Accipiter gentilis; Ac_nis = Accipiter nisus; Ae_mon = Aegypius monachus; Bu_bub = Bubo bubo; Bu_but = Buteo buteo; Fa_tin = Falco tinnunculus; Ni_scu = Ninox scutulata; Ot_lem = Otus lempiji; Ot_sco = Otus scops.)

Ac_gen	**Ac_nis**	**Ae_mon**	**Bu_bub**	**Bu_but**	**Fa_tin**	**Ni_scu**	**Ot_lem**	**Ot_sco**
Bu_but	Ot_lem	Ac_gen	Fa_tin	Fa_tin	Bu_but	Bu_but	Ot_sco	Ot_lem
Fa_tin	Ot_sco	Ni_scu	Bu_but	Ac_gen	Ac_gen	Ae_mon	Ac_nis	Ae_mon
Ae_mon	Ae_mon	Ot_sco	Ac_gen	Ni_scu	Bu_bub	Ac_gen	Ae_mon	Ni_scu
Ni_scu	Ni_scu	Bu_but	Ni_scu	Ae_mon	Ni_scu	Fa_tin	Ni_scu	Ac_nis
Bu_bub	Ac_gen	Ot_lem	Ae_mon	Bu_bub	Ae_mon	Ot_sco	Ac_gen	Ac_gen
Ot_sco	Bu_but	Fa_tin	Ot_sco	Ot_sco	Ot_sco	Bu_bub	Bu_but	Bu_but
Ot_lem	Fa_tin	Ac_nis	Ot_lem	Ot_lem	Ot_lem	Ot_lem	Fa_tin	Fa_tin
Ac_nis	Bu_bub	Bu_bub	Ac_nis	Ac_nis	Ac_nis	Ac_nis	Bu_bub	Bu_bub

Tab 4. Relative similarities of the sequences for the first numerical representation.

Ac_gen	**Ac_nis**	**Ae_mon**	**Bu_bub**	**Bu_but**	**Fa_tin**	**Ni_scu**	**Ot_lem**	**Ot_sco**
Fa_tin	Fa_tin	Bu_but	Ot_sco	Ae_mon	Ac_gen	Ac_nis	Bu_bub	Bu_but
Ac_nis	Ni_scu	Ot_sco	Ot_lem	Ot_sco	Ac_nis	Fa_tin	Ot_sco	Bu_bub
Ae_mon	Ac_gen	Ac_gen	Bu_but	Bu_bub	Ni_scu	Ac_gen	Bu_but	Ae_mon
Bu_but	Ae_mon	Bu_bub	Ae_mon	Ac_gen	Ae_mon	Ae_mon	Ae_mon	Ot_lem
Ni_scu	Bu_but	Fa_tin	Ac_gen	Ot_lem	Bu_but	Bu_but	Ac_gen	Ac_gen
Ot_sco	Ot_sco	Ot_lem	Fa_tin	Fa_tin	Ot_sco	Ot_sco	Fa_tin	Fa_tin
Bu_bub	Bu_bub	Ac_nis	Ac_nis	Ac_nis	Bu_bub	Bu_bub	Ac_nis	Ac_nis
Ot_lem	Ot_lem	Ni_scu	Ni_scu	Ni_scu	Ot_lem	Ot_lem	Ni_scu	Ni_scu

Tab 5. Relative similarities of the sequences for the second numerical representation.

Ac_gen	Ac_nis	Ae_mon	Bu_bub	Bu_but	Fa_tin	Ni_scu	Ot_lem	Ot_sco
Fa_tin	Fa_tin	Bu_but	Ot_sco	Ae_mon	Ac_gen	Ac_nis	Bu_bub	Bu_but
Ac_nis	Ni_scu	Ot_sco	Ot_lem	Ot_sco	Ac_nis	Fa_tin	Ot_sco	Bu_bub
Ae_mon	Ac_gen	Ac_gen	Bu_but	Bu_bub	Ni_scu	Ac_gen	Bu_but	Ae_mon
Bu_but	Ae_mon	Bu_bub	Ae_mon	Ac_gen	Ae_mon	Ae_mon	Ae_mon	Ot_lem
Ni_scu	Bu_but	Fa_tin	Ac_gen	Fa_tin	Bu_but	Bu_but	Ac_gen	Ac_gen
Ot_sco	Ot_sco	Ot_lem	Fa_tin	Ot_lem	Ot_sco	Ot_sco	Fa_tin	Fa_tin
Bu_bub	Bu_bub	Ac_nis	Ac_nis	Ac_nis	Bu_bub	Bu_bub	Ac_nis	Ac_nis
Ot_lem	Ot_lem	Ni_scu	Ni_scu	Ni_scu	Ot_lem	Ot_lem	Ni_scu	Ni_scu

Tab 6. Relative similarities of the sequences for the third numerical representation.

5 Discussion

The sequences of COI gene of nine bird species can be devided into two main groups: species order *Falconiformes/Accipitiformes* and order *Strigiformes*. We expected that the similarity/dissimilarity analysis of COI sequences represented by three numerical representations will at least differentiate between the species orders.

The first numerical representation method provides modest results. Species cannot be correctly classified into orders. For example, Accipiter nisus (Sparrowhawk) is incorrectly closely related with genus Otus (Scops-owels).

The second numerical representation method correctly classified 6 species of 9. The third numerical representation method gives the same results. It suggests that the forth coordinate indicating order of nucleotides in sequence does not add any relevant information into numerical representation and it is redundant.

Figure 1 shows a corresponding phylogenetic tree of the sequences based on Jukes-Cantor method (Matlab Bioinformatics toolbox was used).

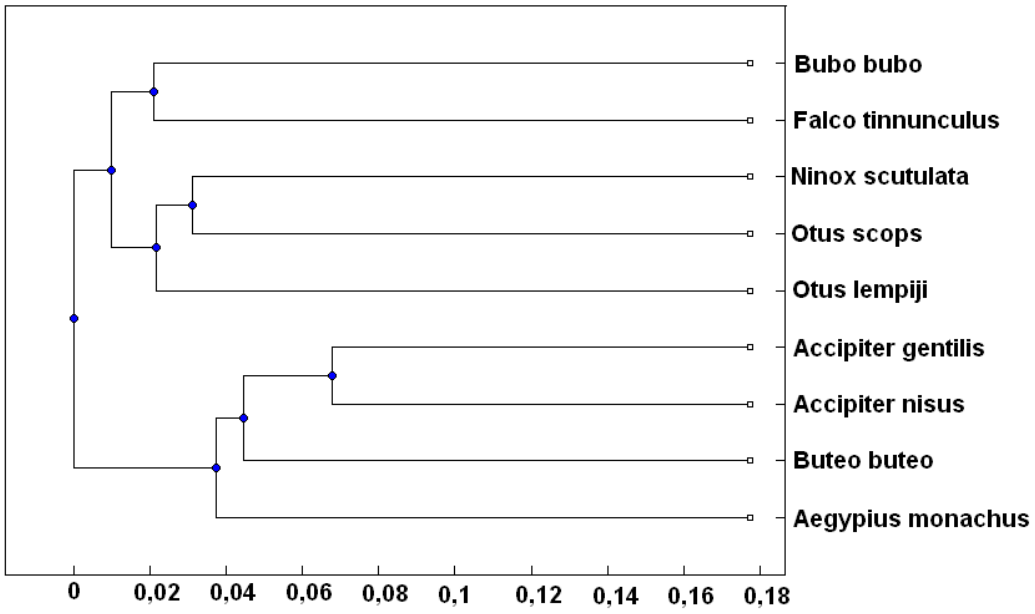

Fig 1. Phylogenetic tree of nine bird species.

6 Conclusions

We applied three numerical representations on short DNA sequences of mitochondrial COI gene of nine bird species. The numerical representations were compared through similarity/dissimilarity analysis based on simple distance matrix method. The results of analysis are strongly dependent on the applied numerical representation. Expected differentiation between species orders was not achieved. The best result gave the second and the third numerical representation. The differentiation was 66 % that connot be considered as suffitient result. The fist numerical representation which was not able to differentiate species orders at all.

Acknowledgement

This work was supported from: GAČR 102/07/1473, GAČR 102/09/H083 and MSM0021630513.

References

[1] Moritz C, Cicero C. DNA Barcoding: Promise and Pitfalls. PLoS Biology 2004;2:1529 – 1531.
[2] Dougherty, E, Shmulevich, I, Chen, J, Wang, Z. Genomic Signal Processing and Statistics. Hindawi Publishing Corporation, 2005.
[3] Randić M, Balaban A. On A Four-Dimensional Representation of DNA Primary Sequences. J. Chem. Inf. Comput. Sci. 2003;43:532-539.
[4] Chi R, Ding Kequan. Novel 4D numerical representation of DNA sequences. Chemical Physics Letters 2005;470:63-67.