DNA BARCODING OF CLASS AVES USING CODON COUNTS WITH ARTIFICIAL NEURAL NETWORKS

Jaymar Soriano[1], Ted Guillano Chua[1], Nathan Lemuel Santi[1], Ric Janus Sapasap[1], Adrian Roy Valdez[1], and Ian Kendrich C. Fontanilla[2]

[1]Scientific Computing Laboratory, Department of Computer Science, University of the Philippines, Diliman, Quezon City

[2]Institue of Biology, University of the Philippines, Diliman, Quezon City
Corresponding authors: jbsoriano@gmail.com, adrez_13@yahoo.com

Abstract

DNA barcoding is currently used for taxonomical classification. For animals, the standard gene used for DNA barcoding is the mitochondrial cytochrome oxidase subunit I (COI) gene. Traditional methods require that the COI gene of an unknown DNA sequence be aligned first before classification. Commonly used algorithm for sequence alignment is the ClustalW. However, this method is computationally expensive since each sequence is aligned with all other sequences in the data. In this paper, we perform a family-level taxonomic classification of class Aves using codon counts from aligned and unaligned sequences as inputs for our neural network classifier. Our data consists of 17 families downloaded from the online database of National Center for Biotechnology Information. We find that the classification using unaligned sequences has a mean accuracy of 93.93%, a minimum of 90.30%, and a maximum of 96.46%. The results are comparable with that obtained from aligned sequences with mean accuracy of 98.38%. This result suggests that it may be possible to classify DNA sequences without the need for alignment that requires highly expensive computations.

Keywords: Bioinformatics, codon count, DNA barcoding, neural networks, soft computing

1. Introduction

Bioinformatics is a huge field of research that applies computers, together with algorithms, mathematics, and statistics, in analyzing floods of information that focuses primarily on subcellular and molecular structures composed of the DNA, RNA, and protein sequences. The DNA or the Deoxyribonucleic acid inside the cell of an organism contains the genetic material of the organism, which defines cell function and/or structure. Changes in the DNA can change a cell. These changes in turn, make up differences among different organisms. An extracted DNA sequence is usually more than a thousand base pairs. Nucleotide bases are one of the following: Adenine (A), Thymine (T), Guanine (G), and Cytosine (G). An arrangement of three nucleotide bases forms a codon. Each codon is translated to one of 20 amino acids. The translation differs between organisms. Part of a DNA sequence called Cytochrome Oxidase Subunit I (COI) is a gene from the mitochondrion, which is essential for cellular respiration. It is an ideal gene for DNA barcoding of animals. Moreover, it is robust in terms of readability and less prone to error, and has a greater number of signs for evolutionary differences. The difference between intra-species is minimal whereas for inter-species, the difference is large enough for classification.

DNA barcoding is a taxonomic method used to classify species using a database of COI sequences. Traditional method classifies unknown DNA sequences using Basic Local Alignment Tool (BLAST). After alignment, the percentage similarity of sequences per nucleotide is computed using the Kimura-2-parameter (K2P). The result of BLAST is a list of possible similar species. However, possible

pseudogenes in the sequences need to be determined. Pseudogenes are genes that are rendered dysfunctional by repeated and compounded mutations. These pseudogenes no longer code for proteins and therefore cannot be used for DNA barcoding. Common practice used is to manually identify and remove these pseudogenes from the sequences. The use of COI gene was shown significantly classify organism in different phyla in at least 98% of the species. It was also reported that intraspecific divergences are usually less than 1% and rarely greater than 2%.

In this study, DNA barcoding using the COI gene is done via codon counts. The codon counts of a subset of the database are then used to train an artificial neural network (ANN) and the result of which will be used to classify the remaining subset. The current technology in DNA classification requires a large database and pairwise alignment procedures. We also analyzed if DNA barcoding could be done if the sequences were not pairwise-aligned. If this alternative method on DNA classification works, the method itself may give a better understanding of organisms and their DNA. We show that for each family of the class Aves, the use of codon counts also significantly differentiates one family from another using aligned COI sequences. On the other hand, classification using unaligned sequences has practically good accuracy since the method is less involved.

2. DNA Barcoding using Codon Counts

The data used in the study is downloaded from the online database of the National Center for Biotechnology Information (NCBI). The data also contains the start codon based on which the sequences are processed and oriented correctly.

2.1. Alignment

For the first part of the study, we perform classification using pair-wise aligned sequences. This is done per family using the ClustalW algorithm. The sequences are trimmed by removing leading and trailing nucleotide bases. Potential pseudogenes are removed by determining which sequences contain extra nucleotides and therefore contribute to misalignment. Finally, sequences that contain less than 500 nucleotide bases are then removed.

2.2. Codon Counts

For each sequence, either aligned or unaligned, the number of occurrence of each of the 64 possible codons in the sequence is counted. The respective percentages are then computed.

2.3. Training in ANN

Thirty random samples from each family are used for training with an artificial neural network. The neural network has two hidden layers. The input layer has 64 nodes corresponding to the codon percentages while the output layer has number of nodes corresponding to number of represented families in class Aves. The training ends when any of the following conditions are met: the bitfail equals zero, the maximum epoch is reach, or the current error is lower than the desired error. The detailed algorithm for the ANN training is shown in Figure 1.

2.4. Classification of Unknown Sequences

After the training of the ANN, the weights are used to classify another set of sequences for testing. The classification is compared to the actual family and the accuracy of the classification is computed.

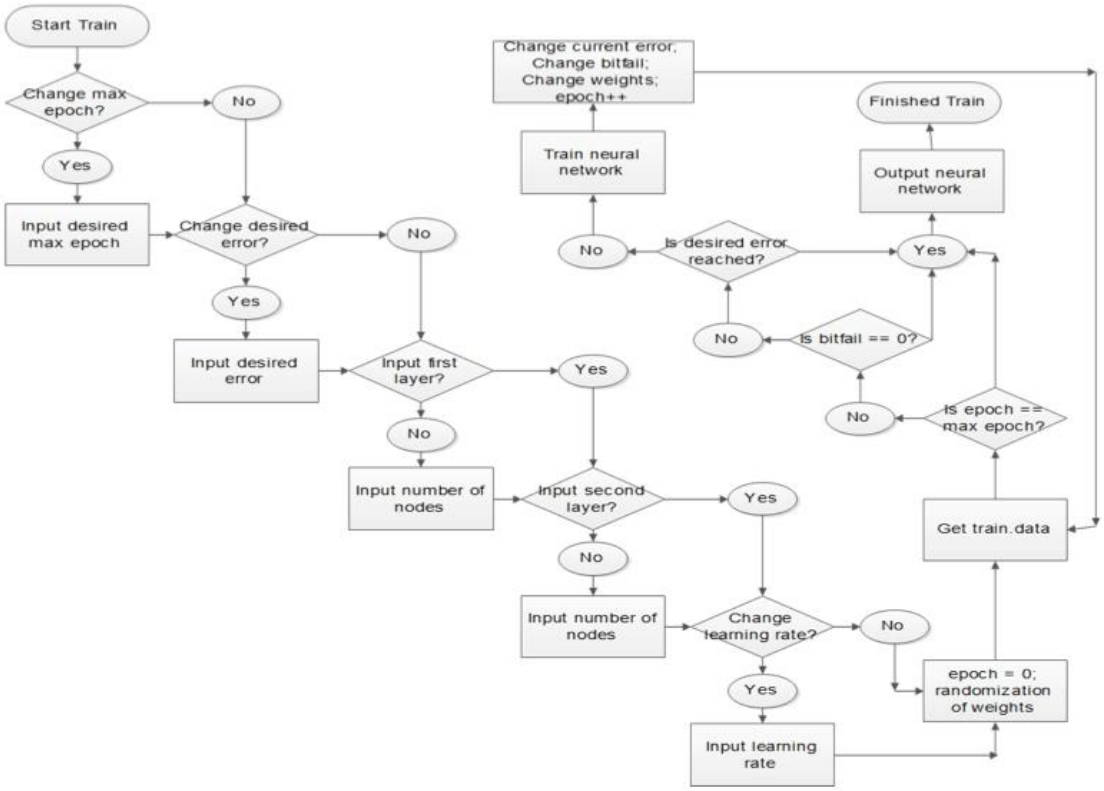

Figure 1: Flowchart for the ANN training.

3. Results and Discussion

We report in this study that codon count is able to discriminate among families of a class. Particularly, Table 1 compares the accuracy of classification for aligned and unaligned data sets of COI sequences of class Aves species. Average classification rate over thirty trainings of different ANN configurations for the aligned sequences using codon counts is 98.38% while 93.93% for unaligned sequences. Obviously, the classification is better for aligned sequences but it is remarkable that the latter does not require pair-wise alignment of sequences that incurs heavy computation. We can then say that the classification rate of for this data set is practically good. This finding also suggests that the distribution of codon counts among different families of class Aves vary in a nonlinear way that the ANN is able to discriminate, whether or not the sequences are pair-wise aligned.

Table 1: Classification rate using codon counts.

	Average	*Minimum*	*Maximum*
Aligned	98.38	97.64	98.79
Unaligned	93.93	90.30	96.46

We also report that an increase in the classification rate to an average of 98.37% is achieved with aligned sequences if protein counts are used instead of codon counts. Each codon corresponds to an amino acid in subjective manner. This implies that discrimination of COI sequences from different families via protein counts can be done if the sequences are pair-wise aligned. This is also remarkable since there is less number of possible proteins than number of possible codons implying ability to discriminate a little more accurate with less information. On the other hand, protein count seems not to be viable for classification with unaligned sequences.

Table 2: Classification rate using codon counts.

	Average	Minimum	Maximum
Aligned	98.64	98.37	98.97
Unaligned	70.25	67.35	73.63

4. Conclusion and Future Studies

DNA barcoding of class Aves species into different families was performed using codon counts as input to a neural network classifier. High classification rate with an average of 98.38% for aligned sequences and 93.93% for unaligned sequences over thirty different trainings is indicative of strong differences in the codon distributions of class aves species among different families. Moreover, the study concludes that it is possible to classify DNA sequences without the need for pair-wise alignment. It is also noteworthy that successful discrimination by the ANN does not account the order of the codons in the sequences. This may be considered for future studies if or not it will increase the classification rate, especially for the unaligned case.

The possibility of DNA barcoding of unaligned sequences will be very beneficial to taxonomists. However in this study, we assumed that the start codon is known from which we start counting the codons. It is also for future work to determine if or not the start codon may or may not be an issue for such a very promising alternative.

References

[1] Hebert, P., Cywinska, A., Ball, S., and deWaard, J. *Biological Identifications through DNA barcodes* in Proceedings of the Royal Society B: Biological Sciences, 270, pp. 313–321.

[2] Hebert PD, Ratnasingham S, and deWaard J. *Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species* in Proceedings of the Royal Society B: Biological Sciences 270 Suppl 1:S96-9.

[3] Jones, C., and Pevzner, P. An Introduction to Bioinformatics Algorithms. Cambridge, Massachusetts: MIT Press. 2004.

[4] Kress, W. J., & Erickson, D. *DNA barcodes: Genes, genomics, and bioinformatics* in Proceedings of the National Academy of Sciences, 105(8), 2761–2762, 2008.

[5] Sung, W. Algorithms in Bioinformatics A Practical Introduction. UK: CRC Press, Taylor & Francis Group, 2010.

[6] Wodehouse, P. G., *Bioinformatics and Pattern Recognition Come Together*, Journal of Pattern Recognition Research 1, pp. 37-41, 2006.