A Study of Compression–Based Methods for the Analysis of Barcode Sequences

Massimo La Rosa, Antonino Fiannaca Riccardo Rizzo, and Alfonso Urso

ICAR-CNR, National Research Council of Italy,
viale delle Scienze Ed.11, 90128 Palermo, Italy
{larosa,fiannaca,ricrizzo,urso}@pa.icar.cnr.it

Abstract. In this paper it is introduced a new methodology for the analysis of barcode sequences. Barcode DNA is a very short nucleotide sequence, corresponding for the animal kingdom to the mitochondrial gene cytochrome c oxidase subunit 1, that acts as a unique element for identification and taxonomic purposes. Traditional barcode analysis uses well consolidated bioinformatics techniques such as sequence alignment, computation of evolutionary distances and phylogenetic trees. The proposed alignment-free approach consists in the use of two different compression-based approximations of Universal Similarity Metric in order to compute dissimilarity matrices among barcode sequences of 20 datasets belonging to different species. From these matrices phylogenetic trees are computed and compared, in terms of topology and branch length, with trees built from evolutionary distance. The results show high similarity values between compression-based and evolutionary-based trees allowing us to consider the former methodology worth to be employed for the study of barcode sequences

Keywords: Barcode DNA, Compression–Based distances, Universal Similarity Metric, Phylogenetic trees.

1 Introduction

DNA barcoding aims at discovering and isolating a very short part of DNA of living organism for identification and taxonomic purposes [1, 2]. The very basic idea is to find and define, for each kingdom of life, such as animals, plants, fungi and so on, a single gene that works as a true "barcode" providing unique identification. In the animal kingdom, *mitochondrial gene cytochrome c oxidase subunit 1* (COI) has proven to be the best barcode sequence [3]. DNA barcoding has been used for the study of the biodiversity of several species, such as fishes, birds and some bugs [4–7].

The analysis of barcode sequences, both for identification and taxonomic purposes, is carried out by means of classic bioinformatics methodologies, based on sequence alignment and computation of dissimilarity matrices that can be used to build phylogenetic trees or to make identification of unknown species through well known threshold values [8].

L.E. Peterson, F. Masulli, and G. Russo (Eds.): CIBB 2012, LNBI 7845, pp. 105–116, 2013.

In this paper, an alignment-free methodological approach for the analysis of barcode sequences is proposed. It is based on compression-based distances derived from Universal Similarity Metric (USM) [9]. USM is a class of distance measures, founded on rigorous information theory concepts defined in the Kolmogorov complexity [10]. Unfortunately, Kolmogorov complexity is not computable, therefore there exists a set of USM approximations based on data compression. Compression-based methods have the advantage that they do not require a prior alignment of genomic sequences and above all they hold on strong theoretical assumptions. Evolutionary distances, in turn, are based on stochastic estimates and they do not define a distance metric.

In order to justify the use of compression–based distances for the study of barcode sequences, several datasets, belonging to different kinds of species, have been downloaded from Bold database [8]; for each dataset, a set of phylogenetic trees have been build according to the most common bioinformatics algorithms (see Section 3) a set of phylogenetic trees. Those trees, then, have been compared with phylogenetic trees obtained through state-of-the-art methods based on evolutionary distances [11].

2 Background

USM distance, as defined in [9], is "universal" in the sense it can be applied to different type of input data. In fact, it has been used for classification and clustering activities in several application domains, from text processing to language analysis, from music to image files [12]. A first attempt to use one of the USM approximations, called Normalized Compression Distance (NCD), for the study of genomic sequences has been done in [12]. The result of that work was a phylogenetic tree obtained considering complete mammalian mtDNA sequences of 24 species belonging to Eutherian order. In [13] another approximation of USM, based on GenCompress compressor [14], has been applied in order to compute a phylogenetic tree of a larger dataset containing mammalian mtDNA sequences of 34 taxa. The authors stated that USM is able to provide meaningful results when applied to very large genomic sequences and a small number of taxa.

A very important experimental assessment regarding the use of USM for different type of biological datasets has been carried out in [15]. By considering six different datasets, both of protein and genomic (complete mitochondrial genome) sequences, the authors tested two USM approximations, namely NCD and Universal Compression Dissimilarity (UCD), with several compressors in order to obtain phylogenetic trees. Those trees were then compared with gold standard taxonomies using classic tree comparison algorithms, F-measure [16] and Robinson metric [17], and they also concluded compression-based methods are allowed to be considered when dealing with biological datasets.

A different use of USM for clustering, through Self-Organizing Maps, and generation of topographic representations of bacteria datasets, considering 16S rRNA gene, was done in [18, 19], where topographic maps of three bacteria phyla were built from both evolutionary distance and NCD, showing similarities and differences between maps.

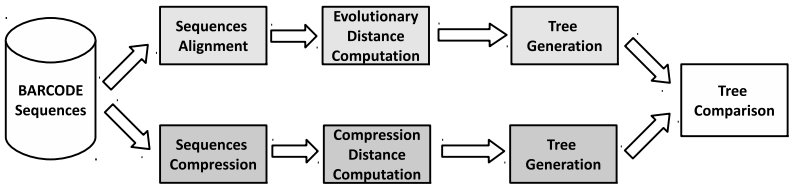

Fig. 1. Overall framework of the proposed methodology (lower workflow) compared with classic pipeline (upper workflow)

In our work, we want to demonstrate that USM, and in general compression–based distances, are also suited for the analysis of short barcode sequences, about 650 bp long, and for several datasets composed of very different species. Moreover, in order to compare phylogenetic trees obtained through evolutionary distances and compression-based methods, we adopt more recent and complete comparison tree algorithms that take into account relevant topological features of phylogenetic trees and not only basic different pairings.

3 Methods

In this Section it is presented the overall framework of our methodology; then in the following subsection, the tools and algorithms adopted in order to perform our experimental tests will be described in detail.

In Fig. 1 there are both the workflow of our proposed methodology, the lower one, and the classic workflow, the upper one, usually adopted for the analysis of gene sequences for phylogenetic purposes. After downloading barcode sequences from BOLD database [8], our approach consists in compressing the genomic sequences using GenCompress compressor [14], computing two different approximations of USM, as explained in Section 3.1, and finally building phylogenetic trees using state-of-the-art algorithms. On the other hand, classic methodology comprises sequence alignment, computation of evolutionary distance and finally generation of phylogenetic trees. Trees obtained with our and classic approach were then analyzed by means of two different tree comparison algorithms that consider different tree properties: topology and branch length.

3.1 Compression-Based Dissimilarity Measures

Universal Similarity Metric (USM) is a class of distance measures based on Kolmogorov complexity [10] and introduced by Li et al. [9]. USM allows to compare two generic data files and it has been demonstrated that it is a similarity metric, i.e the *identity axiom*, the *triangle inequality* and the *symmetry axiom* hold. The key idea of USM is to find a shared information content between two objects. Since it has been demonstrated Kolmogorov complexity is not computable, it needs to be approximated.

In our work, two different USM's approximations used for the comparison of genomic sequences have been considered: Normalized Compression Distance (NCD) [12] and the distance defined in [13] that for ease of explanation we call Information-Based Distance (IBD). In both kinds of distance, Kolmogorov complexity is approximated by means of the size of the compressed version of the sequence itself. NCD and IBD are defined respectively in Eq. 1 and Eq. 2:

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{1}$$

$$\mathrm{IBD}(x,y) = 1 - \frac{C(x) - C(x|y)}{C(xy)} \tag{2}$$

There $C(x)$ and $C(y)$ are the sizes, in bytes, of the compressed sequences x and y; whereas $C(xy)$ is the size of the compressed sequence obtained through the concatenation between x and y. $C(x|y)$ is the size of the compression of sequence x with respect to the reference sequence y, that is the information required to obtain x from y [20]. This kind of conditional compression is also known as vertical compression [20].

Both NCD and IBD's purpose is to find the shared information content between two sequences: NCD can be computed using a general purpose normal compressor; IBD has been introduced considering GenCompress compressor [14] to heuristically approximate Kolmogorov complexity.

GenCompress [14] is a compression algorithm optimized to work with DNA sequences. It follows the approach of Lempel and Ziv dictionary based compressors [21], taking advantage of the fact that a genomic sequence has just a four characters (a, c, g, t) dictionary. GenCompress, in fact, gives the best compression ratios only when dealing with DNA sequences: if it is applied to sequences containing more than the four nucleotide characters, it acts as a generic ascii-text compressor. GenCompress algorithm also implements a conditional version, i.e. it computes the compression of sequence x given another sequence as reference.

In this paper GenCompress is used when computing both NCD and IBD so that it is possible a direct comparison among results obtained by means of both kinds of distances.

3.2 Phylogenetic Inference

There are several methods to build a phylogenetic tree from molecular data [11, 22]. In our work two of the most used methods are considered: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [23] and Neighbor Joining (NJ) [24]. Both algorithms belong to the so called distance–based methods because they need a dissimilarity matrix among input sequences before building the tree. According to the adopted evolutionary distance model, like for instance Kimura 2–parameter [25], Tajima–Nei [26], Tamura–Nei [27], there can be different distance matrices and, consequently, different phylogenetic trees.

UPGMA is the simplest phylogenetic reconstruction algorithm, it creates an ultrametric tree (dendogram) and its basic assumption is that it builds a

correct tree if the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages.

NJ considers different rates of evolution among tree's branches and it is very reliable if the input dissimilarity matrix is very close to the true evolutionary distances among sequences. NJ uses a clustering algorithm that, starting from a star topology, at each iteration pairs the nearest elements, obtaining at the end a binary tree.

3.3 Comparison of Phylogenetic Trees

In phylogenetic studies, it is possible to obtain different phylogenetic trees according to the used algorithm or the considered gene or set of genes. For this reason, several algorithms for tree comparison have been developed. The most popular is the method proposed by Robinson and Foulds [17], also known as *symmetric distance*. It computes the distance between two phylogenetic trees by considering the number of transformations, or shifts, needed to reconstruct the first tree from the second one, or vice-versa. Symmetric distance can be seen as a generalization of edit metrics [28] to phylogenetic trees.

In order to compare phylogenetic trees obtained through evolutionary distances and compression based distances, two more recent comparison algorithms, whose approach is rather different from Robinson method, have been considered: the tool presented by Nye et al. [29] and the K tree score, introduced in [30].

Nye's algorithm aims at matching branches (edges) within two trees which share similar topological features. This topological feature is the partition of leaf elements created by every branch in a tree. The similarity score for each pair of edges between two trees is given comparing the shared leaf nodes belonging to the two corresponding partitions. This process builds a sort of alignment between the two trees to compare. While Robinson metric gives each topological difference the same penalty, in Nye's algorithm different pairings have a lesser penalty if their topological features are preserved, that is they belong to the same corresponding partitions. This way similarity between trees is not expressed by a mere number of edit operations, but by considering topological properties.

This fact is better explained looking at Fig. 2, where two phylogenetic trees obtained through evolutionary distance (on the left) and compression-based distance (on the right) are shown. Thicker branches in both trees highlight a lower similarity between the corresponding subtrees; whereas thin edges identify a perfect match among the two partitions. Using Robinson metric, on the other hand, all pairs of not corresponding leaf nodes are considered as a wrong pairing, ignoring any topological feature.

K score is an extension of branch length distance (BLD) defined in [31] and it allows to obtain a similarity score depending on the similarity between branch length of both trees. Once again it differs from symmetric distance because this one does not consider branch lengths when computing the similarity score. Those two algorithms offer two kinds of comparison: Nye tool gives a score based only on the similarity between trees topologies; K score takes into account the

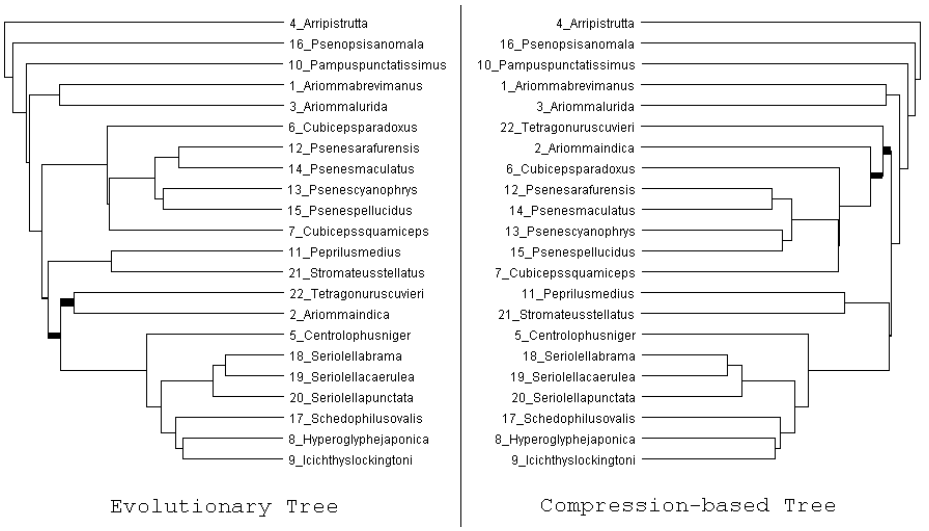

Fig. 2. Comparison between phylogenetic trees obtained from evolutionary distance (left) and compression-based distance (right). Thicker edges mean the corresponding partitions within the two tress have not exactly the same leaf nodes.

similarity between branch length's trees. This way we can test our results both in terms of topology and branch length similarity.

4 Results

In this section we report experimental tests used for evaluating two compression-based algorithms (NCD and IBD). The evaluation is based on the comparison of phylogenetic trees generated with both UPGMA and NJ algorithms.

4.1 Dataset Description

In order to test the performance of the discussed compression-based algorithms, we used 20 datasets from "Barcode of Life Data" Systems (BOLD) Project [8]. Among more than 1000 available datasets, we considered a subset composed by those datasets that respect two main criteria: first, for each dataset all the sequences (representing species or specimens) are the mitochondrial COI-5P gene and second, all the datasets belong to different *familia* of the animal kingdom. From this subset, we randomly selected 20 datasets.

All datasets used during experimental tests are reported in Table 1. The first column shows datasets acronyms, as reported in BOLD database. For each dataset, Table 1 reports four features: the number of specimens and species (respectively second and third column); the number of sequences having at least

Table 1. 20 Datasets selected from Barcode of Life Data System. Some datasets are clustered in 5 groups of distinctive features.

	Dataset	#Specimens (sequences)	#Species	Sequences with undefined bases	Length of sequences	Distinctive features				
						G1	G2	G3	G4	G5
1	**JTB**	225	53	1/225	658-899	✓				
2	**DLTC**	67	40	1/67	689-1821	✓				
3	**Onychophora**	210	52	2/210	451-884	✓				
4	**AGWEB**	33	33	29/33	460-890	✓	✓			✓
5	**GBFCJ**	202	61	14/202	537-1446	✓	✓			
6	**GZPSE**	78	23	6/78	601-658		✓			
7	**RDMYS**	37	6	12/37	636		✓	✓		
8	**ARCPU**	52	28	3/52	901			✓		
9	**BRBP**	106	17	0/106	658			✓	✓	
10	**AGFDO**	22	22	0/22	901			✓	✓	✓
11	**AECI**	30	30	0/30	605-679				✓	✓
12	**SIBHI**	85	38	0/85	673-694				✓	
13	**BLSPA**	86	86	4/86	604-658					✓
14	**ABSMC**	72	46	1/72	650-657					
15	**AGFSU**	48	42	1/48	605-680					
16	**AGLUO**	46	38	1/46	633-639					
17	**DSALA**	44	12	5/44	649-651					
18	**FBLOT**	64	34	2/64	419-658					
19	**MJMSL**	198	76	9/198	559-658					
20	**WXYZ**	34	9	1/34	650-680					

one undefined base (forth column) and the range of sequences' length (fifth column).

The last column is composed by 5 sub-columns that indicate a particular set of features. The meaning of each group is reported in the following, whereas the analogies among these datasets will be investigated in the next Section:

- G1: Datasets in this group contain some sequences much longer than the other ones of the same dataset;
- G2: In these datasets there is an high percentage of sequences with undefined bases;
- G3: All the sequences in these datasets have the same length;
- G4: Sequences in these datasets do not have undefined bases;
- G5: These datasets contain sequences with one specimen for each species.

The BOLD system provides, for each dataset, a distance matrix obtained by default using the Kimura 2-parameter distance model. With regards to the compression-based algorithms, since they require as input a list of sequences in order to generate the distance matrix, we also downloaded a list of (pre-aligned) COI-5P gene sequences for each dataset.

Table 2. Similarity and K-score among phylogenetic trees: evolutionary technique with Kimura 2-parameter distance versus compression based algorithms (both NCD and IBD)

	Dataset	Tree similarity (Nye et al.)				K-Score			
		NCD		IBD		NCD		IBD	
		UPGMA	NJ	UPGMA	NJ	UPGMA	NJ	UPGMA	NJ
1	**JTB**	0.75	0.61	**0.85**	0.59	0.1852	—	0.2090	—
2	**DLTC**	**0.86**	0.79	0.84	0.77	0.6842	—	0.6973	—
3	**Onychophora**	0.89	0.77	**0.92**	0.81	0.1165	—	0.1333	—
4	**AGWEB**	0.73	0.76	0.77	**0.89**	0.0667	0.0674	0.0775	0.0779
5	**GBFCJ**	0.80	0.72	**0.82**	0.72	0.1170	—	0.1155	—
6	**GZPSE**	0.84	**0.86**	0.85	0.85	0.0588	—	0.0714	—
7	**RDMYS**	0.78	0.60	0.81	**0.87**	0.0472	—	0.0587	—
8	**ARCPU**	0.87	**0.94**	0.87	0.87	0.0562	—	0.0720	—
9	**BRBP**	**0.99**	0.89	**0.99**	0.82	0.0772	—	0.1149	—
10	**AGFDO**	**0.92**	0.88	**0.92**	0.88	0.0315	0.0323	0.0607	0.0632
11	**AECI**	0.89	0.82	**0.90**	0.82	0.0517	—	0.0839	—
12	**SIBHI**	0.92	0.90	**0.94**	0.90	0.0581	—	0.0976	—
13	**BLSPA**	**0.88**	0.82	0.85	0.79	0.0424	0.0485	0.0621	0.0672
14	**ABSMC**	**0.97**	0.93	0.92	0.95	0.0720	—	0.1211	—
15	**AGFSU**	0.84	0.84	**0.89**	0.85	0.0609	0.0598	0.0910	0.0954
16	**AGLUO**	0.97	0.90	**0.98**	0.90	0.0394	0.0442	0.0615	0.0646
17	**DSALA**	**0.91**	0.88	**0.91**	0.88	0.0706	—	0.0940	—
18	**FBLOT**	0.89	0.81	**0.90**	0.82	0.0655	—	0.1011	—
19	**MJMSL**	**0.88**	0.82	**0.88**	0.79	0.0912	—	0.1211	—
20	**WXYZ**	0.92	0.79	**0.95**	0.76	0.0629	—	0.0926	—

4.2 Experimental Tests

Experimental tests aim to evaluate the quality of phylogenetic reconstructions obtained by means of compression-based algorithm. Using aforementioned evaluation techniques, we compare phylogenetic trees that have been generated with the two most used algorithms: UPGMA and NJ.

Results are reported in Table 2. This table is composed by three main columns: the first one contains dataset acronyms, the second one the "Tree Similarity" scores and the last one the "K Score". Second and third columns, in turn, contain sub-columns in order to show results obtained with UPGMA and NJ trees with both NCD and IBD distances. In the "Tree Similarity" columns, each number represents a percentage value, where "1" means trees have the same topology, that implies the compression-based algorithm preserves the evolutionary taxonomy. Numbers in bold type are the best scores for each dataset. The "K Score" column, instead, reports values that measure the difference between two trees in terms of branch length. In this case, lower values mean higher similarity among trees. Unfortunately, NJ algorithm sometimes generates trees with negative branches that can not be computed by K score algorithm: in fact, although

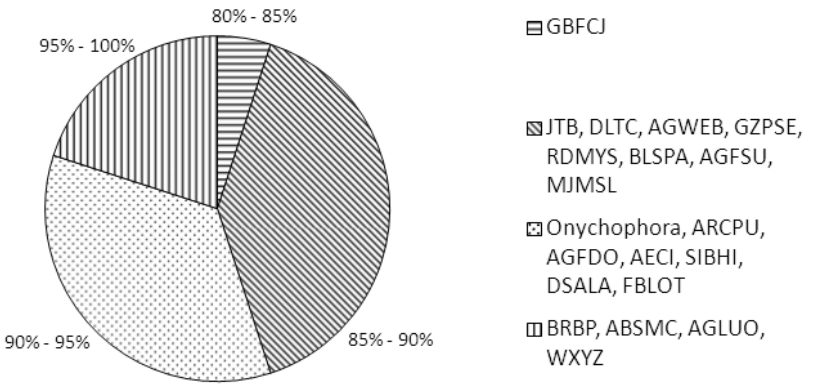

Fig. 3. Pie chart representing the percentage of similarity between evolutionary and compression-based tree for each datasets

mathematically the neighbor joining algorithm is admitted to produce negative values, biologically a tree with some negative branches is meaningless [31]. In this situation, a "−" symbol is reported.

First of all, considering the tree similarities results, we can state that the compared trees are quite similar. More in detail, considering the best values for each dataset, we obtained the pie chart in Fig. 3. The most of datasets are between 85% and 95% and four of them have a topological similarity greater than 95%. Only the fifth dataset (GBFCJ) gives the a result of 82%.

These results are not so surprising because the quality of retrieved datasets are different each other, looking at the five groups of datasets reported in Table 1. Group G1 contains some sequences much longer than the other ones: for these datasets, compression-based similarity algorithms give poor results, since they take into account mutual length of sequence.

Group G2 is composed by four datasets with several sequences that contain some special symbol (i.e. Y or N) to represent undefined nucleotides. In this case, as previously said in Section 3, the GenCompress algorithm works as a generic compressor of ASCII string, reducing its performance. For instance, the dataset GBFCJ, belonging both to first and to second group, shows the lowest value of tree similarity (82%). Another dataset belonging to G1 and G2 groups, is AGWEB. This dataset, with respect to GBFCJ, has an higher value of tree similarity (89%), because it has a lower spread in sequences length.

Groups G3 and G4 in Table 1 contain those datasets with respectively the same length for all sequences and with a complete COI-5P gene sequencing. BRBP dataset belongs to both these groups and represents the best one among datasets used in this paper, since it has no sequences with undefined bases and all the sequences have the same length of 658-bp, representing COI-5P gene length proposed as a potential 'barcode' in [2]. This dataset, composed of 106 elements, reaches a value of 99% tree similarity with its corresponding evolutionary tree. It is interesting to notice that the other datasets having sequence length close to

658-bp, and that do not belong to the first or the second group, score the best results, such as AGLUO (98%) and SIBHI (94%).

Group G5 reports datasets with a single specimen (sample) for each species. In terms of tree similarity, datasets with only a specimen for each species do not produce better results than datasets with more than a specimen for each specie, with both compression-based distances. In other words, compression algorithm works fine also at specimens level.

Considering the type of compression-based distances, obtained results demonstrate in the most of case (75%) IBD reaches highest values in terms of tree similarity, especially when UPGMA is used for generate trees, with the exception of datasets in G1 of Table 1. In fact, for datasets with an high percentage of sequences with undefined bases, NJ is able to better represent the evolutionary tree, for instance AGWEB has 87% of undefined bases and reaches the better value of similarity (89%) with IBD and NJ algorithm. This means that in all cases IBD algorithm is able to preserve the topology of an evolutionary tree of DNA barcode sequences.

As for the K-score column in Table 2, results confirm all the considerations previously said, except for compression distance algorithm analysis. In fact, in terms of differences in the relative tree branch length, it appears NCD algorithm works better than IBD. It is possible to notice that datasets in group G1 of Table 1 score lesser results, e.g. DLTC (0.684), whereas datasets in group G4 score the best results, e.g. AGFDO (0.031).

5 Conclusion

In this paper we presented a deep analysis about the use of compression-based methods, such as NCD and IBD, for the study of short DNA barcode sequences. NCD and IBD are both approximations of Universal Similarity Metric, that is a class of general-purpose distances based on non-computable Kolmogorov complexity. In previous works, USM and its approximations have been applied in the case of the analysis of complete mitochondrial genome of few species: there they showed how phylogenetic trees obtained through USM had a very similar topology to those ones obtained through classic bioinformatics methods based on sequence alignment and evolutionary distances computing. By employing compression-based methods there is no need to align input sequences and moreover USM represents a distance metric, whereas evolutionary distances are stochastic distance estimates lacking metric properties such as triangle inequality. In this work we extended the use of NCD and IBD to DNA barcode sequences, typically 650 bp long. We compared phylogenetic trees of 20 datasets obtained from NCD and IBD, using NJ and UPGMA algorithms, with trees of the same datasets obtained from Kimura 2-parameter evolutionary distance. The comparison was done by means of two different algorithms, considering both topological and branch length similarities. The results we presented show that trees obtained from compression-based methods are very similar (above 90%), and in some cases equal, to the ones built from classic distance. In few situations,

characterized by some flaws in input datasets, we obtained similarity scores of about 85%, demonstrating compression-based methods are robust enough to deal with noisy datasets. In the near future we are going to provide other comparisons between trees using other kinds of evolutionary distances and phylogenetic reconstruction algorithms so that we can definitively use compression-based methods for the study of phylogenetic relationships with DNA barcode sequences.

References

1. Savolainen, V., Cowan, R.S., Vogler, A.P., Roderick, G.K., Lane, R.: Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1805–1811 (2005)
2. Hebert, P.D.N., Cywinska, A., Ball, S.L., de Waard, J.R.: Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321 (2003)
3. Hebert, P.D.N., Ratnasingham, S., de Waard, J.R.: Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. Biol. Sci. 270(suppl. 1), 96–99 (2003)
4. Costa, F.O., Carvahlo, G.R.: The Barcode of Life Initiative: synopsis and prospective societal impacts of DNA barcoding of fish. Genomics, Society and Policy 3, 29–40 (2007)
5. Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., Francis, C.M.: Identification of Birds through DNA Barcodes. PLoS Biol. 2(10), e312 (2004)
6. Smith, M.A., Fisher, B.L., Hebert, P.D.N.: DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Phil. Trans. R. Soc. B 360, 1825–1834 (2005)
7. Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W., Hebert, P.D.N.: DNA barcodes distinguish species of tropical Lepidoptera. PNAS 103(4), 968–971 (2006)
8. Ratnasingham, S., Hebert, P.D.N.: BOLD: The Barcode of Life Data System. Molecular Ecology Notes 7, 355–364 (2007)
9. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.M.B.: The Similarity Metric. IEEE T. Inform. Theory 50(12), 3250–3264 (2004)
10. Li, M., Vitanyi, P.M.B.: An Introduction to Kolmogorov Complexity and its Applications, 2nd edn. Springer, New York (1997)
11. Makarenkov, V., Kevorkov, D., Legendre, P.: Phylogenetic network construction approaches. Applied Mycology and Biotechnology 6, 61–97 (2006)
12. Cilibrasi, R., Vitanyi, P.M.B.: Clustering by Compression. IEEE T. Inform. Theory 51(4), 1523–1545 (2005)
13. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17(2), 149–154 (2001)
14. Chen, X., Kwong, S., Li, M.: A compression algorithm for DNA sequences. IEEE Engineering in Medicine and Biology Magazine 20(4), 61–66 (2001)
15. Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., Valiente, G.: Compression-based classification of biological sequences and structures via the Universal Similarity Metric: Experimental assessment. BMC Bioinformatics 8(252) (2007)
16. van Rijsbergen, C.J.: Information Retireval. Butterworths, London (1979)
17. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical Biosciences 53(1), 131–147 (1981)

18. La Rosa, M., Rizzo, R., Urso, A., Gaglio, S.: Comparison of Genomic Sequences Clustering Using Normalized Compression Distance and Evolutionary Distance. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 740–746. Springer, Heidelberg (2008)
19. La Rosa, M., Gaglio, S., Rizzo, R., Urso, A.: Normalised compression distance and evolutionary distance of genomic sequences: comparison of clustering results. Int. J. Knowledge Engineering and Soft Data Paradigms 1(4), 345–362 (2009)
20. Grumbach, S., Tahi, F.: A new challenge for compression algorithms: genetic sequences. J. Information Processing and Management 30(6), 866–875 (1994)
21. Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Trans. Inform. Theory 23(3), 337–343 (1977)
22. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, New York (2000)
23. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy: The Principles and Practice of Numerical Classification. W.H. Freeman, San Francisco (1973)
24. Saitou, N., Nei, M.: The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. 4(4), 406–425 (1987)
25. Kimura, M.: Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. 78, 454–458 (1981)
26. Tajima, F., Nei, M.: Estimation of evolutionary distance between nucleotide sequences. Molecular Biology and Evolution 1, 269–285 (1984)
27. Tamura, K., Nei, M.: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10, 512–526 (1993)
28. Atallah, M.J., Blanton, M.: Algorithms and Theory of Computation Handbook. CRC Press LLC (1999)
29. Nye, T.M.W., Liò, P., Gilks, W.R.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22(1), 117–119 (2006)
30. Soria-Carrasco, V., Talavera, G., Igea, J., Castresana, J.: The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23(21), 2954–2956 (2007)
31. Kuhner, M.K., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468 (1994)