Map of Life: Measuring and Visualizing Species' Relatedness with *Molecular Distance Maps*

Lila Kari[1,2] Kathleen A. Hill[3] Abu S. Sayem[2] Nathaniel Bryans[4]
Katelyn Davis[3] Nikesh S. Dattani[5]

October 10, 2013

Abstract

We propose a novel combination of methods that *(i)* portrays quantitative characteristics of a DNA sequence as a black-and-white image, *(ii)* computes pair-wise distances between these images, and *(iii)* uses these distances to output a map wherein each sequence is a point in a common Euclidean space. In the resulting *Molecular Distance Map* each point signifies a DNA sequence, and the geometric distance between any two points reflects the degree of relatedness between the corresponding sequences and species.

Molecular Distance Maps present compelling visual representations of relationships between species and could be used for taxonomic clarifications, for species identification, placement of species in existing taxonomic categories, as well as for studies of evolutionary history. One of the advantages of this method is its general applicability since, as sequence alignment is not required, the DNA sequences chosen for comparison can be completely different regions in different genomes. In fact, this method can be used to compare *any* two DNA sequences. For example, in our dataset of 3,176 mitochondrial DNA sequences, it correctly finds the mtDNA sequences most closely related to that of the anatomically modern human (the Neanderthal, the Denisovan, and the chimp), and it finds that the sequence most different from it belongs to a cucumber. Furthermore, our method can be used to compare real sequences to artificial, computer-generated, DNA sequences. For example, it is used to determine that the distances between a *Homo sapiens sapiens* mtDNA and artificial sequences of the same length and same trinucleotide frequencies can be larger than the distance between the same human mtDNA and the mtDNA of a fruitfly.

We demonstrate this method's promising potential for taxonomical clarifications by applying it to a diverse variety of cases that have been historically controversial, such as the genus *Polypterus*, the family Tarsiidae, and the vast (super)kingdom Protista.

1 Introduction

In 2012 alone, biologists described between 16,000 and 20,000 new species [27]. Recent findings, [28], suggest that as many as 86% of existing species on Earth and 91% of species in the oceans have not yet been classified and catalogued. It is thus imperative to find a comprehensive, quantitative, general-purpose method to reliably identify the relationships among the 1.2 million species that have already been catalogued, [28], as well as the vastly larger numbers of those that have not.

We propose a combination of three techniques to efficiently measure distances between DNA se-

[1]Corresponding author, email lila.kari@uwo.ca
[2]Department of Computer Science, University of Western Ontario, London, ON, N6A 5B7 Canada
[3]Department of Biology, University of Western Ontario, London, ON, N6A 5B7 Canada
[4]Microsoft Corporation, Redmond, WA, 98052, USA
[5]Physical and Theoretical Chemistry Laboratory, Dept. of Chemistry, Oxford University, Oxford, OX1 3QZ, UK

quences, and to simultaneously visualize the relationships among all the DNA sequences within any given dataset. The result of applying this method to a collection of DNA sequences is a *Molecular Distance Map* that allows the visualization of the sequences as points in a common two-dimensional Euclidean space, wherein the geometric distance between any two points reflects the differences in composition between all the subsequences of the two sequences. The proposed method is based on the *Chaos Game Representation* (CGR) of DNA sequences, [18, 19], a genomic signature that has a remarkable ability to differentiate between genetic sequences belonging to different species, see Figure 1. Due to this characteristic, a *Molecular Distance Map* of a collection of genetic sequences allows the inferrence of relationships between the corresponding species.

Concretely, to compute and visually display relationships between DNA sequences in a given set $S = \{s_1, s_2, ..., s_n\}$ of n DNA sequences, we propose the following combination of three techniques:

- *Chaos Game Representation* (CGR), to deterministically represent each DNA sequence s_i, $1 \leq i \leq n$, as a two-dimensional black-and-white image denoted by c_i;
- *Structural Dissimilarity Index* (DSSIM), an image-distance measure, to compute the distances $\Delta(i, j)$, $1 \leq i, j \leq n$, between pairs of CGR images (c_i, c_j), and to produce a distance matrix;
- *Multi-Dimensional Scaling* (MDS), applied to the distance matrix to produce a map in the Euclidean space wherein each plotted point p_i with coordinates (x_i, y_i) represents the DNA sequence s_i whose CGR image is c_i. The position of the point p_i in the map, relative to all the other points p_j, reflects the distances between the DNA sequence s_i and the DNA sequences s_j in the dataset.

The combination of CGR, DSSIM, and MDS applied to any given set of genetic sequences yields a *Molecular Distance Map* which visually illustrates the quantitative relationships and patterns of proximities among the given genetic sequences and, accordingly, among the species they represent.

Besides presenting compelling visual pictures of the relatedness between species as seen in, e.g., Figure 2, there are several advantages of our proposed method over other methods in computational phylogenetics.

Advantages over alignment-based methods include the fact that our method allows comparison between *any* two DNA sequences. In particular, it allows comparisons within the genome of an individual, across genomes within a single species, between genomes within a taxonomic category, and across taxa, while also allowing the use of completely different portions of the genomes for comparison. In addition, sequence alignment analyses use only a limited aspect of the compared sequences, namely the identity of characters at each position. In contrast, our method is based on significantly more information by simultaneously comparing all subsequences, up to a certain length, of the given sequences.

An advantage over phylogenetic trees [22] pertains to the fact that, in a phylogenetic tree, adjacency of two species-representing leaves is not always meaningful since one can rotate branches about the nodes of the tree. In contrast, in a Molecular Distance Map the distance on the plane between any two species-representing points has a concrete and fixed meaning. Furthermore, our DSSIM distance matrices can be used, [7], to calculate phylogenetic trees in addition to Molecular Distance Maps.

Advantages over DNA barcodes [11] and Klee diagrams [36] include the fact that our method is not alignment-based, and that it is applicable to cases where barcodes may have limited effectiveness: Plants and fungi for which different barcoding regions have to be used [20], [15], [34]; protists where multiple loci are generally needed to distinguish between species [14]; prokaryotes [38]; and artificial, computer-generated, DNA sequences.

Finally, our proposed approach addresses the need for a visual representation of species relatedness that is easily interpretable, as well as capable of including a vast amount of data rather than selective or regional datasets.

2 Methods

A CGR of a genomic DNA sequence is a genomic signature that utilizes all subsequence composition structures of DNA sequences, and is genome and species specific, [18, 19, 12, 13, 6, 5, 40]. The sequences chosen from each genome as a basis for computing "distances" between species do not need to have any relation with one another from the point of view of their position or information content. A CGR [18, 19] associates an image to each DNA sequence as follows: Starting from a unit square with vertices labelled *A, C, G,* and *T*, and the center of the square as the starting point, the image is obtained by successively plotting each nucleotide as the middle point between the current point and the vertex labelled by the nucleotide to be plotted. If the generated square image has a size of $2^k \times 2^k$ pixels, then every pixel represents a distinct DNA subsequence of length k: a pixel is black if the subsequence it represents occurs in the DNA sequence, otherwise it is white. In our case, k is approximately 9. In general 4,000 bp are necessary for a sharply defined image, but in many cases 2,000 bp give a reasonably good approximation, [18]. CGR images of genetic DNA sequences originating from various species show interesting patterns such as squares, parallel lines, rectangles, triangles, and also complex fractal patterns, Figure 1.

Other visualizations of genetic data have been proposed, such as the 2D rectangular walk [8] and methods similar to it in [29], [23], vector walk [25], cell [43], vertical vector [44], Huffman coding [32], and colorsquare [47] methods. Three-dimensional representations of DNA sequences include the tetrahedron [33], 3D-vector [46], and trinucleotide curve [45] methods. Among these visualization methods, CGR images arguably provide the most immediately comprehensible "signature" of a DNA sequence and a desirable genome-specificity, [18, 5]. In addition, the black-and-white images produced using CGR are easy to compare, visually and computationally.

Structural Similarity (SSIM) index is an image similarity index used in the context of image processing and computer vision to compare two black-and-white images from the point of view of their structural similarities [42]. SSIM combines three parameters - luminance distortion, contrast distortion, and linear correlation - and was designed to perform similarly to the human visual system, which is highly adapted to extract structural information. Originally, SSIM was defined as a similarity measure $s(A, B)$ whose theoretical range between two images A and B is $[-1, 1]$ where a high value amounts to close relatedness. We use a related *DSSIM distance* $\Delta(A, B) = 1 - s(A, B) \in [0, 2]$, with the distance being 0 between two identical images, 1 between e.g. a black image and a white image, and 2 if the two images are negatively correlated; that is, $\Delta(A, B) = 2$ if and only if every pixel of image A has the inverted value of the corresponding pixel in image B while both images have the same luminance (brightness). For our particular dataset of genetic CGR images, almost all distances were between 0 and 1, with the exceptions being very close to 1.

MDS has been used for the visualization of data relatedness based on distance matrices in various fields such as cognitive science, information science, psychometrics, marketing, ecology, social science, and other areas of study [2]. MDS takes as input a distance matrix containing the pairwise distances between n given items and outputs a two-dimensional map wherein each item is represented by a point, and the geometric distances between points reflect the distances between the corresponding items in the distance matrix. Two notable examples of molecular biology studies that used MDS are [24] (where it was used for the analysis of geographic genetic distributions of some natural populations) and [11] (where it was used to provide a graphical summary of the distances among CO1 genes from various species).

Classical MDS, which we use in this paper, receives as input an $n \times n$ distance matrix $(\Delta(i, j))_{1 \le i,j \le n}$ of the pairwise distances between any two items in the set. The output of classical MDS consists of n points in a q-dimensional space whose pairwise Euclidean distances are a linear function of the distances between the corresponding items in the input distance matrix. More precisely, MDS will return n points $p_1, p_2, \ldots, p_n \in \mathbb{R}^q$ such that $d(i, j) = ||p_i - p_j|| \approx f(\Delta(i, j))$ for all $i, j \in \{1, \ldots, n\}$ where $d(i, j)$ is the Euclidean distance between the points p_i and p_j, and f is a function linear in $\Delta(i, j)$. Here, q can be at

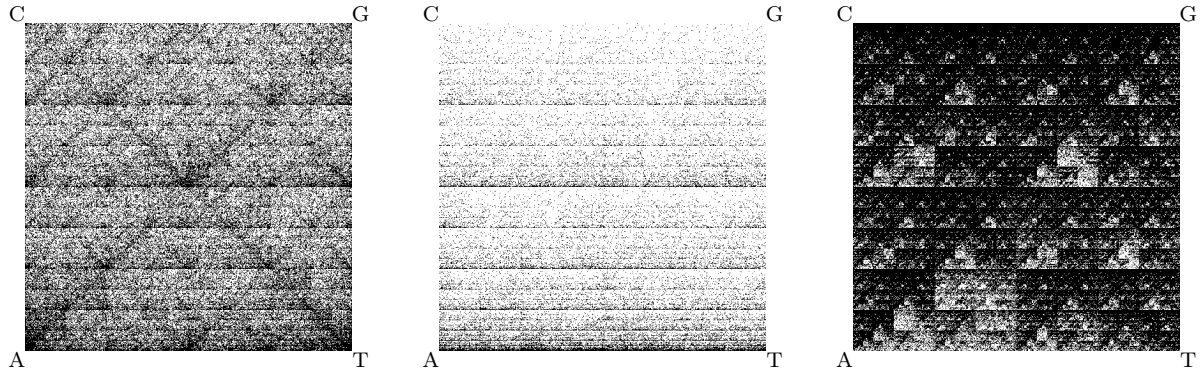

Figure 1: CGR images for various genomes. From left to right: (1) *Marchantia polymorpha* (liverwort) mtDNA, 186,609 bp; (2) *Malawimonas jakobiformis* (flagellate) mtDNA, 47,328 bp; (3) *Rhodobacter capsulatus*, full genome, 3,738,958 bp.

most $n-1$ and the points are recovered from the eigenvalues and eigenvectors of the input $n \times n$ distance matrix. If we choose $q = 2$ (respectively $q = 3$), the result of classic MDS is an approximation of the original $(n-1)$-dimensional space as a two- (respectively three-) dimensional map.

In this paper all Molecular Distance Maps consist of coloured points, wherein each point represents an mtDNA sequence from the dataset. Each mtDNA sequence is assigned a unique numerical identifier retained in all analyses, e.g., #1321 is the identifier for the *Homo sapiens sapiens* mitochondrial genome. The colour assigned to a sequence-point may however vary from map to map, and it depends on the taxon assigned to the point in a particular Molecular Distance Map. For consistency, all maps are scaled so that the x- and the y-coordinates always span the interval $[-1, 1]$. The formula used for scaling is $x_{\text{sca}} = 2 \cdot (\frac{x - x_{\min}}{x_{\max} - x_{\min}}) - 1$, $y_{\text{sca}} = 2 \cdot (\frac{y - y_{\min}}{y_{\max} - y_{\min}}) - 1$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum of the x-coordinates of all the points in the original map, and similarly for $y_{\min}$ and $y_{\max}$.

Each Molecular Distance Map has some error, that is, the Euclidean distances $d_{i,j}$ are not exactly the same as $f(\Delta(i,j))$. When using the same dataset, the error is in general lower for an MDS map in a higher-dimensional space. The *Stress-1* (Kruskal stress, [21]), is defined in our case as

$$\textit{Stress-1} = \sigma_1 = \sqrt{\frac{\Sigma_{i<j}[f(\Delta(i,j)) - d_{i,j}]^2}{\Sigma_{i<j} d_{i,j}^2}}$$

where the summations extend over all the sequences considered for a given map, and $f(\Delta(i,j)) = a \times \Delta(i,j) + b$ is a linear function whose parameters $a, b \in \mathbb{R}$ are determined by linear regression for each dataset and corresponding Molecular Distance Map. A benchmark that is often used to assess MDS results is that *Stress-1* should be in the range $[0, 0.20]$, see [21].

The dataset consists of the entire collection of complete mitochondrial DNA sequences from NCBI as of 12 July, 2012. This dataset consists of 3,176 complete mtDNA sequences, namely 79 protists, 111 fungi, 283 plants, and 2,703 animals. This collection of mitochondrial genomes has a great breadth of species across taxonomic categories and great depth of species coverage in certain taxonomic categories. For example, we compare sequences at every rank of taxonomy, with some pairs being different at as high as the (super)kingdom level, and some pairs of sequences being from the exact same species, as in the case of *Silene conica* for which our dataset contains the sequences of 140 different mitochondrial chromosomes [37]. The prokaryotic origins and evolutionary history of mitochondrial genomes have long been ex-

tensively studied, which will allow comparison of our results with both phylogenetic trees and barcodes. Lastly, this genome dataset permits testing of both recent and deep rooted species relationships, providing fine resolution of species differences.

An example of the CGR/DSSIM/MDS approach is the Molecular Distance Map in Figure 2 which depicts the complete mitochondrial DNA sequences of all 1,791 jawed vertebrates in our dataset. (In the legends of Figures 2-7, the number of represented mtDNA sequences in each category is listed in paranthesis after the category name.) All five different subphyla of jawed vertebrates are separated in non-overlapping clusters, with very few exceptions. Examples of fish species bordering or slightly mixed with the amphibian cluster include *Polypterus ornatipinnis* (#3125, ornate bichir), *Polypterus senegalus* (#2868, Senegal bichir), both with primitive pairs of lungs; *Erpetoichthys calabaricus* (#2745, snakefish) who can breathe atmospheric air using a pair of lungs; and *Porichtys myriaster* (#2483, specklefish midshipman) a toadfish of the order Batrachoidiformes. It is noteworthy that the question of whether species of the *Polypterus* genus are fish or amphibians has been discussed extensively for hundreds of years [10]. Interestingly, all four represented lungfish (a.k.a. salamanderfish), are also bordering the amphibian cluster: *Protopterus aethiopicus* (#873, marbled lungfish), *Lepidosiren paradoxa* (#2910, South American lungfish), *Neoceratodus forsteri* (#2957, Australian lungfish), *Protopterus doloi* (#3119, spotted African lungfish).

The creation of the datasets, acquisition of data from NCBI's GenBank, generation of the CGR images, calculation of the distance matrix, and calculation of the Molecular Distance Maps using MDS, were all done with the free open-source MATLAB program OpenMPM [4]. This program makes use of an open source MATLAB program for SSIM written by Z.Wang [41], and MATLAB's built-in MDS function[6].

[6]Supplemental Material including the annotated dataset, its DSSIM distance matrix, and full-size versions of Figures 2-7, with numerical mtDNA sequence identifiers, is available at http://www.csd.uwo.ca/~lila/MapOfLife/

3 Results and Discussion

We applied our method to visualize all available complete mtDNA sequences from *three classes*, Amphibia, Insecta and Mammalia, in Figure 3. On a finer scale, we applied this method to observe relationships *within a class*: class Amphibia and three of its orders in Figure 4, and class Insecta grouped in nine categories in Figure 5. Note that a feature of MDS is that the points p_i are not unique. Indeed, one can translate or rotate a map without affecting the pairwise Euclidean distances $d(i,j) = ||p_i - p_j||$. In addition, the obtained points in an MDS map may change coordinates when more data items are added to or removed from the dataset. This is because the output of the MDS aims to preserve only the pairwise Euclidean distances between points, and this can be achieved even when some of the points change their coordinates. In particular, while the position within a taxonomic category may be correctly preserved, the (x, y) coordinates of a point representing an amphibian species in the amphibians-insects-mammals map (Figure 3) will not necessarily be the same as the (x, y) coordinates of the same point when only amphibians are mapped (Figure 4).

In general, Molecular Distance Maps are in good agreement with classical phylogenetic trees at all scales of taxonomic comparisons, see Figure 5 with [26], Figure 4 with [31], and Figure 6 with [35]. Our approach may also provide supporting evidence for a DNA sequence-based classification for certain species where morphology-based taxonomy is uncertain. In addition, our approach may be able to weigh in on conflicts between taxonomic classifications based on morphological traits and those based on more recent molecular data, as in the case of tarsiers (Figure 6), and protists (Figure 7), as seen below.

Zooming in, we observed the relationships within an order, Primates, with its suborders (Figure 6). Notably, two extinct species of the genus *Homo* are represented: *Homo sapiens neanderthalensis* and *Homo sapiens ssp. Denisova*. Primates can be classified into two groups, Haplorhini (dry-nosed primates comprising anthropoids and tarsiers) and Strepsirrhini (wet-nosed primates including lemurs and lorises). The map shows a clear separation of

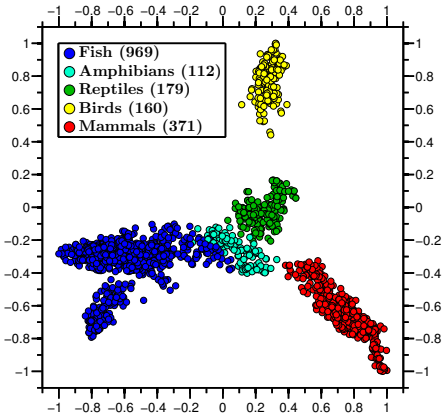

Figure 2: Molecular Distance Map of phylum Vertebrata (excluding the 5 represented jawless vertebrates), with its five subphyla. The total number of mtDNA sequences is 1,791, the average DSSIM distance is 0.8652, and the MDS *Stress-1* is 0.12.

these suborders, with the top-left arm of the map in Figure 6, comprising the Strepsirrhini. However, there are two Haplorhini placed in the Strepsirrhini cluster, namely *Tarsius bancanus* (#2978, Horsfield's tarsier) and *Tarsius syrichta* (#1381, Philippine tarsier). The phylogenetic placement of tarsiers within the order Primates has been controversial for over a century, [17]. According to [3], mitochondrial DNA evidence places tarsiiformes as a sister group to Strepsirrhini, while in contrast, [30] places tarsiers within Happlorhini. In Figure 6 the tarsiers are located within the Strepsrrhini cluster, which may be due to the fact that they evolved independently for millions of years, [35].

Finally, we applied our method to a group whose taxonomic classification has historically been challenging: Figure 7 depicts all protist species in the dataset whose taxon (as defined in the legend of the figure) contained more than one representative. The obvious outlier is *Haemoproteus* sp. jb1.JA27 (#1466), sequenced in [1] (see also [39]), and listed as an *unclassified* organism in the NCBI taxonomy. Interestingly this outlier is in the same kingdom (Chromalveolata), superphylum (Alveolata), phylum (Apicomplexa), and class (Aconoidasida) as the two mtDNA sequences *Babesia bovis* T2Bo (#1935), and *Theileria parva* (#3173), that appear grouped with it. It therefore seems plausible that Molecular Distance Maps may shed light on the taxonomical ambiguity of *Haemoproteus* sp. jb1.JA27.

The DSSIM distances computed between all pairs of complete mtDNA sequences varied in range. The minimum distance was 0, between two pairs of identical mtDNA sequences. The first pair comprised the mtDNA of *Rhinomugil nasutus* (#98, shark mullet, length 16,974 bp) and *Moolgarda cunnesius* (#103, longarm mullet, length 16,974 bp). A base-to-base sequence comparison between these sequences (#98, NC_017897.1; #103, NC_017902.1) showed that the sequences were indeed identical. However, after completion of this work, the sequence for species #103 was updated to a new version (NC_017902.2), on 7 March, 2013, and is now different from the sequence for species #98 (NC_017897.1). The second pair comprises the mtDNA sequences #1033 and #1034 (length 16,623 bp), generated by crossing female *Megalobrama amblycephala* with male *Xenocypris davidi* leading to the creation of both diploid (#1033) and triploid (#1034) nuclear genomes, [16], but identical mitochondrial genomes.

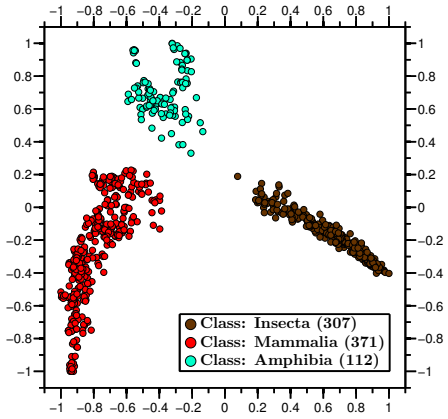

Figure 3: Molecular Distance Map of three classes: Amphibia, Insecta and Mammalia. The total number of mtDNA sequences is 790, the average DSSIM distance is 0.8139, and the MDS *Stress-1* is 0.16.

The maximum distance was found to be between *Pseudendoclonium akinetum* (# 2656, a green alga, length 95,880) and *Candida subhashii* (#954, a yeast, length 29,795). Thus, the pair with the maximum distance $\Delta(\#2656, \#954) = 1.0033$ featured neither the longest mitochondrial sequence, belonging to *Cucumas sativus* (#533, cucumber, length 1,555,935 bp), nor the shortest mitochondrial sequence, belonging to *Silene conica* (#440, sand catchfly, a plant, length 288 bp).

An inspection of the distances between *Homo sapiens sapiens* and all the other primate mitochondrial genomes in the dataset showed that the minimum distance to *Homo sapiens sapiens* was $\Delta(\#1321, \#1720) = 0.1340$, the distance to *Homo sapiens neanderthalensis* (#1720, Neanderthal), with the second smallest distance to it being $\Delta(\#1321, \#1052) = 0.2280$, the distance to *Homo sapiens ssp. Denisova* (#1052, Denisovan). The third smallest distance was $\Delta(\#1321, \#3084) = 0.5591$ to *Pan troglodytes* (#3084, chimp). Figure 8 shows the graph of the distances between the *Homo sapiens sapiens* mtDNA and each of the primate mitochondrial genomes. With no exceptions, this graph is in full agreement with established phylogenetic trees, [35]. The largest distance between the *Homo sapiens sapiens* mtDNA and another mtDNA sequence in the dataset was 0.9957, the distance between *Homo sapiens sapiens* and *Cucumas sativus* (#533, cucumber, length 1,555,935 bp),

In addition to comparing real DNA sequences, our method can compare real DNA sequences to computer-generated sequences. As an example, we compared the mtDNA genome of *Homo sapiens sapiens* with one hundred artificial, computer-generated, DNA sequences of the same length and the same trinucleotide frequencies as the original. The average distance between these artificial sequences and the original human mitochondrial DNA is 0.8991. This indicates that all "human" artificial DNA sequences are more distant from the *Homo sapiens sapiens* mitochondrial genome than *Drosophila melanogaster* (#3120, fruit fly) mtDNA, with $\Delta(\#3120, \#1321) = 0.8572$. This further implies that trinucleotide frequencies may not contain sufficient information to classify a genetic sequence, suggesting that Goldman's claim [9] that "CGR gives no futher insight into the structure of the DNA sequence than is given by the dinucleotide and trinucleotide frequencies" may not hold in general.

The *Stress-1* values for all but one of the Molecular Distance Maps in this paper were in the "acceptable"

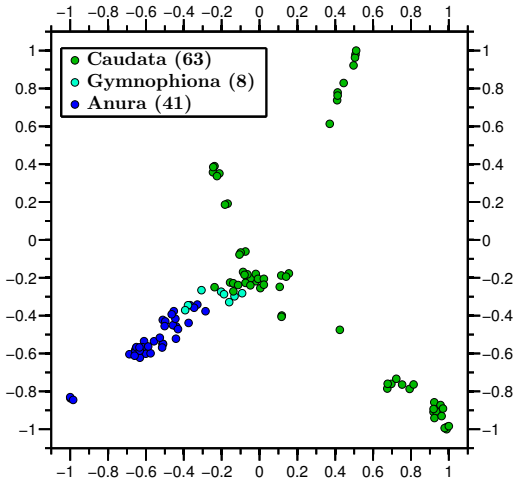

Figure 4: Molecular Distance Map of Class Amphibia and three of its orders. The total number of mtDNA sequences is 112, the average DSSIM distance is 0.8445, and the MDS *Stress-1* is 0.18.

range $[0, 0.2]$. The exception is Figure 7 with *Stress-1* equal to 0.26. Note that *Stress-1* generally decreases with an increase in dimensionality, from $q = 2$ to $q = 3, 4, 5...$, and that in [2] it is suggested that it is not always the case that only MDS representations with *Stress-1* under 0.2 are acceptable, nor that all MDS representations with *Stress-1* under 0.05 are good.

In all the calculations in this paper, we used the full mitochondrial sequences. However, since the length of a sequence can influence the brightness of its CGR and thus its Molecular Distance Map coordinates, further analysis is needed to elucidate the effect of sequence length on the positions of sequence-representing points in a Molecular Distance Map. The choice of length of DNA sequences used may ultimately depend on the particular dataset and particular application.

4 Conclusions

Molecular Distance Maps combine three methods to measure distances between *any* DNA sequences (real or computer-generated) of a given dataset, and visually display their interrelationships. Applications of Molecular Distance Maps include clarification of taxonomic dilemmas, taxonomic classifications, species identification, studies of evolutionary history, as well as possible quantitative definitions of the notions of species and other taxa.

Possible extensions include generalizations of CGR images to three-dimensional CGR images mapped using a regular tetrahedron, as well as MDS generalizations, e.g., to three-dimensional Molecular Distance Maps, for improved accuracy. We note also that this method can be applied to analyzing sequences over other alphabets. For example binary sequences could be imaged using a square or a tetrahedron with vertices labelled 00, 01, 10, 11, and then DSSIM and MDS could be employed to compare and map them. Lastly, we note that the use of the particular image distance measure (DSSIM) or particular scaling technique (classical MDS) does not mean that these are the best choices in all cases, and other image distance measures as well as refinements of classical MDS may be explored for optimal results.

Acknowledgements. We thank Ronghai Tu for an early

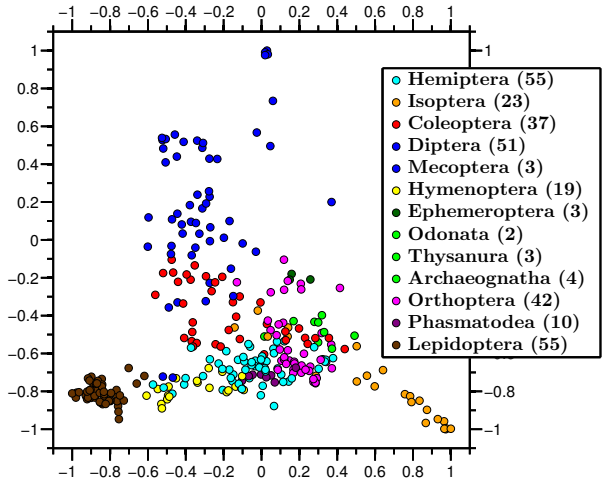

Figure 5: Molecular Distance Map of Class Insecta. The total number of mtDNA sequences is 307, the average DSSIM distance is 0.6855, and the MDS *Stress-1* is 0.14.

version of our MATLAB code to generate CGR images, Tao Tao for assistance with NCBI's GenBank, Steffen Kopecki for generating artificial sequences and discussions, and Rallis Karamichalis for assistance with calculations. We also thank Andre Lachance, Jeremy McNeill, and Greg Thorn for resources and discussions on taxonomy. We thank the Oxford University Mathematical Institute for the use of their Windows compute server Pootle/WTS. This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants to L.K. and K.H.; Oxford University Press Clarendon Fund and NSERC USRA, PGSM, and PGSD3 awards to N.D.; NSERC USRA award to N.B.

References

[1] J. Beadell and R. Fleischer. A restriction enzyme-based assay to distinguish between avian hemosporidians. *Journal of Parasitology*, 91:683–685, 2005.

[2] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2nd edition, 2010.

[3] H. Chatterjee, S. Ho, I. Barnes, and C. Groves. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evolutionary Biology*, 9(259), 2009.

[4] N. Dattani, A. Sayem, R. Tu, and N. Bryans. OpenMPM. *Computer Program*, pages https://github.com/ndattani/Dattani–Sayem–Bryans–genomicMDS, 2013.

[5] P. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, and B. Fertil. Genomic signature is preserved in short DNA fragments. In *IEEE Intl. Symposium on Bio-Informatics and Biomedical Engineering*, pages 161–167, 2000.

[6] P. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by Chaos Game Representation of sequences. *Molecular Biology and Evolution*, 16(10):1391–1399, 1999.

[7] W. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760):279–284, 1967.

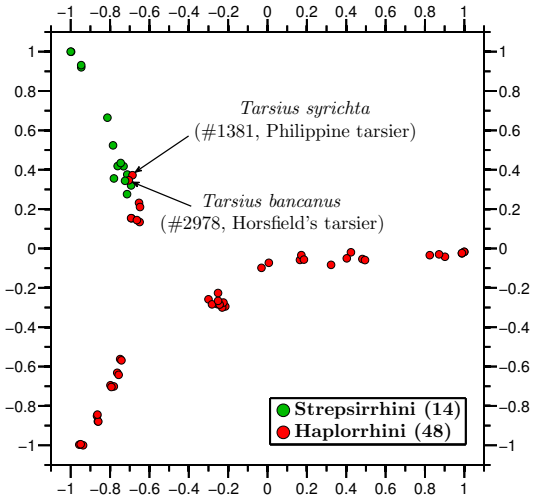

Figure 6: Order Primates and its suborders: Haplorhini (anthropoids and tarsiers), and Strepsirrhini (lemurs and lorises). The total number of mtDNA sequences is 62, the average DSSIM distance is 0.7733, and the MDS *Stress-1* is 0.19.

[8] M. Gates. A simple way to look at DNA. *J. Theor. Biology*, 119(3):319–328, 1986.

[9] N. Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in Chaos Game Representations of DNA sequences. *Nucleic Acids Research*, 21(10):2487–2491, 1993.

[10] B. Hall. John Samuel Budgett (1872-1904): In pursuit of P*olypterus*. *BioScience*, 51(5):399–407, 2001.

[11] P. Hebert, A. Cywinska, S. Ball, and J. Dewaard. Biological identifications through DNA barcodes. *Proc. Biol. Sci*, 270:313–321, 2003.

[12] K. Hill, N. Schisler, and S. Singh. Chaos Game Representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *J. Mol. Evol.*, 35(3):261–9, 1992.

[13] K. Hill and S. Singh. Evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes. *Genome*, 40:342–356, 1997.

[14] K. Hoef-Emden. Pitfalls of establishing DNA barcoding systems in protists: the Cryptophyceae as a test case. *PLoS One*, 7:e43652, 2012.

[15] P. Hollingsworth et al. A DNA barcode for land plants. *PNAS*, 106(31):12794–2797, 2009.

[16] J. Hu et al. Characteristics of diploid and triploid hybris derived from female Megalobrama amblycephala Yih × male Xenocypris davidi Bleeker. *Aquaculture*, 364-365:157–164, 2012.

[17] N. Jameson et al. Genomic data reject the hypothesis of a prosimian primate clade. *Journal of Human Evolution*, 61:295–305, 2011.

[18] H. Jeffrey. Chaos Game Representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.

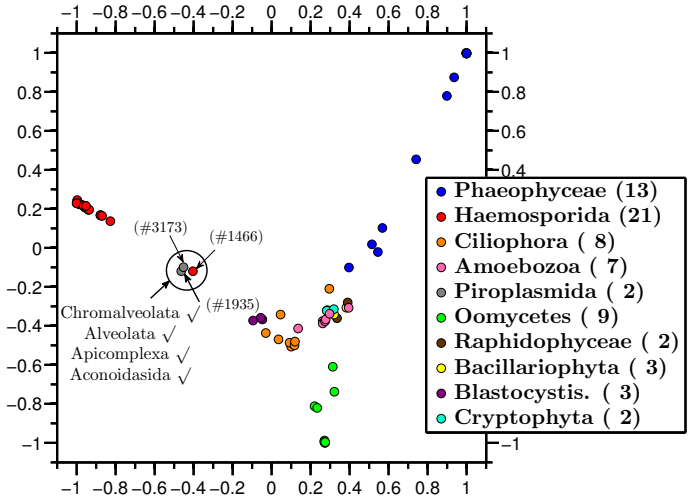

Figure 7: Molecular Distance Map of all represented species from (super)kingdom Protista and its orders. The sequence-point #1466 (red) is *Haemoproteus* sp. jb1.JA27, #1935 (grey) is *Babesia bovis T2Bo*, and #3173 (grey) is *Theileria parva*. The annotation shows that all these three species belong to the same taxonomic groups, up to the order level. The total number of mtDNA sequences is 70, the average DSSIM distance is 0.8288, and the MDS *Stress-1* is 0.26.

[19] H. Jeffrey. Chaos game visualization of sequences. *Comput. Graphics*, 16(1):25–33, 1992.

[20] W. Kress, K. Wurdack, E. Zimmer, L. Weigt, and D. Janzen. Use of DNA barcodes to identify flowering plants. *PNAS*, 102(23):8369–8374, 2005.

[21] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[22] P. Lemey, M. Salemi, and A. Vandamme, editors. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge Univ. Press, 2nd edition, 2009.

[23] P. Leong and S. Morgenthaler. Random walk and gap plots of DNA sequences. *Computer applications in the biosciences : CABIOS*, 11(5):503–507, 1995.

[24] E. Lessa. Multidimensional analysis of geographic genetic structure. *Systematic Zoology*, 39(3):242–252, 1990.

[25] B. Liao. A 2D graphical representation of DNA sequence. *Chemical Physics Letters*, 401(1–3):196–199, 2005.

[26] S. Marshall. *Insects: Their natural history and diversity*. Firefly Books, 2006.

[27] S. Milius. New species of the year. *Science News*, 182(13):30, 2012.

[28] C. Mora, D. Tittensor, S. Adl, A. Simpson, and B. Worm. How many species are there on earth and in the ocean? *PLoS Biology*, 9(8):1–8, 2011.

[29] A. Nandy. A new graphical representation and analysis of DNA sequence structure: Methodology and application to globin genes. *Current Science*, 66(4):309 – 314, 1994.

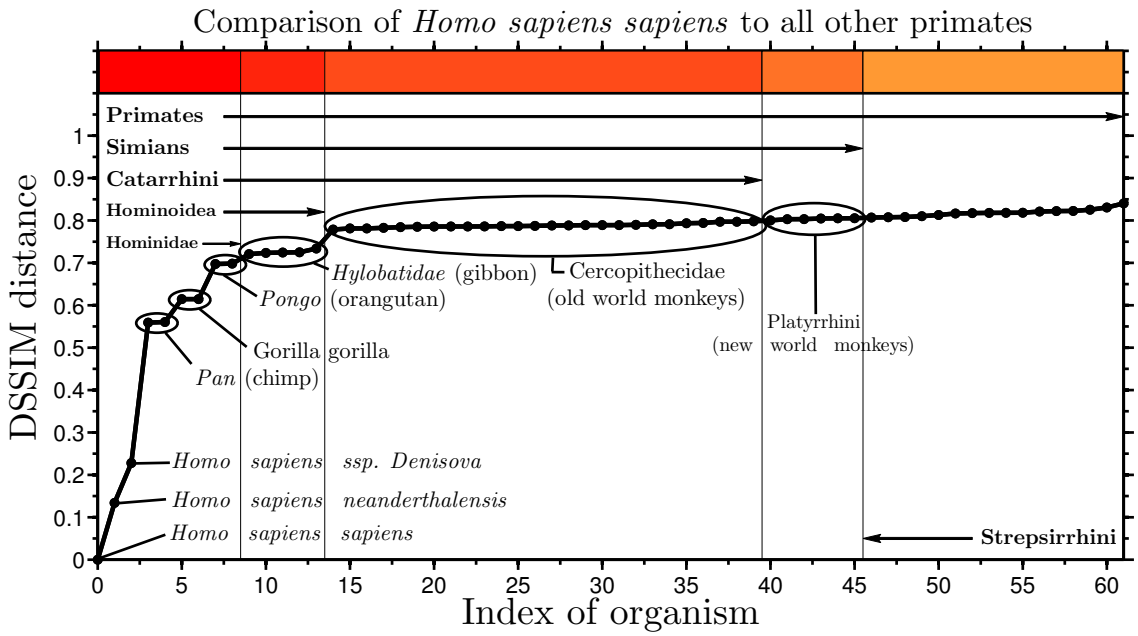

Figure 8: Graph of the DSSIM distances between the CGR images of *Homo sapiens sapiens* mtDNA and each of the 62 primate mitochondrial genomes (sorted).

[30] P. Perelman et al. A molecular phylogeny of primates. *PLoS Genetics*, 7(3), 2011. e1001342.

[31] R. Pyron and J. Wiens. A large-scale phylogeny of amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, 61:543–583, 2011.

[32] Z. Qi, L. Li, and X. Qi. Using Huffman coding method to visualize and analyze DNA sequences. *Journal of Computational Chemistry*, 32(15):3233–3240, 2011.

[33] M. Randic, M. Vracko, A. Nandy, and S. Basak. On 3D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. and Comp. Sci.*, 40(5):1235–1244, 2000.

[34] C. Schoch et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *PNAS*, 109(16):6241–6246, 2012.

[35] J. Shoshani et al. Primate phylogeny: morphological vs molecular results. *Molecular Phylogenetics and Evolution*, 5(1):102–154, 1996.

[36] L. Sirovich, M. Stoeckle, and Y. Zhang. Structural analysis of biodiversity. *PLoS ONE*, 5(2):e9266, 2010.

[37] D. Sloan et al. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology*, 10:e1001241, 2012.

[38] R. Unwin and M. Maiden. Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.*, (11):479–487, 2003.

[39] G. Valkiunas et al. A new Haemoproteus species (Haemosporida: Haemoproteidae) from the endemic Galapagos dove Zenaida galapagoensis,

with remarks on the parasite distribution, vectors, and molecular diagnostics. *Journal of Parasitology*, 96:783–792, 2010.

[40] Y. Wang, K. Hill, S. Singh, and L. Kari. The spectrum of genomic signatures: From dinucleotides to Chaos Game Representation. *Gene*, 346:173–185, 2005.

[41] Z. Wang. SSIM index. *Computer Program*, page https://ece.uwaterloo.ca/ z70wang/research/ssim/, 2003.

[42] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[43] Y. Yao and T. Wang. A class of new 2D graphical representation of DNA sequences and their application. *Chemical Physics Letters*, 398(4–6):318–323, 2004.

[44] C. Yu, Q. Liang, C. Yin, R. He, and S. Yau. A novel construction of genome space with biological geometry. *DNA Research*, 17(3):155–168, 2010.

[45] J. Yu, X. Sun, and J. Wang. TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. *Journal of Theoretical Biology*, 261(3):459 – 468, 2009.

[46] C. Yuan, B. Liao, and T. Wang. New 3D graphical representation of DNA sequences and their numerical characterization. *Chemical Physics Letters*, 379:412 – 417, 2003.

[47] Z. Zhang et al. Colorsquare: A colorful square visualization of DNA sequences. *Comm. in Math. and in Comp. Chemistry*, 68(2):621–637, 2012.