Mapping the Space of Genomic Signatures

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

Abstract

We propose a computational process to measure and simultaneously visualize the interrelationships among any number of DNA sequences which allows, for example, the examination of hundreds or thousands of complete mtDNA genomes. The process starts by computing an "image distance" between graphical representations of DNA sequences' composition and proceeds to visualize the inferred interrelationships as a *Molecular Distance Map*: Each point on the map represents a DNA sequence, and the spatial distance between any two points reflects the degree of structural similarity between the corresponding sequences. This is a general-purpose method that does not require DNA sequence homology and can thus be used to compare similar or vastly different DNA sequences, genomic or computer-generated, of the same length or of different lengths.

The graphical representation of DNA sequences utilized in this process is the Chaos Game Representation (CGR) of DNA sequences, which has been shown to be genome- and species-specific and can thus act as a genomic signature. Consequently, Molecular Distance Maps could inform taxonomic clarifications, species identification, placement of species in existing taxonomic categories, as well as studies of evolutionary history. The image distance employed in this process, the Structural Dissimilarity Index (DSSIM), implicitly compares the occurrences of oligomers of length up to a given k (herein $k = 9$) in the given DNA sequences. We computed such distances for more than 5 million pairs of complete mitochondrial sequences, and used Multi-Dimensional Scaling (MDS) to obtain Molecular Distance Maps that visually display the sequence relatedness in various taxonomic subsets of interest: phylum Vertebrata, (super)kingdom Protista, classes Amphibia-Insecta-Mammalia, class Amphibia only, and order Primates.

This process and analysis suggest that the presence (or absence) of oligomers in mitochondrial DNA sequences can be a source of phylogenetic information. In our dataset of 3,176 complete mitochondrial DNA sequences, this method also 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.

1 Introduction

In 2012 alone, biologists described between 16,000 and 20,000 new species [30]. Other findings, [31], 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. In the absence of a universal quantitative method to identify species' relationships, information for species classification has to be gleaned and combined from several sources, morphological, sequence-alignment-based phylogenetic anaylsis, and non-alignment based molecular information.

We propose a computational process that outputs,

[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

for any given dataset of DNA sequences, a concurrent display of the structural similarities among all sequences in the dataset. This is obtained by first computing an "image distance" for each pair of graphical representations of DNA sequences, and then visualizing the resulting interrelationships in a two-dimensional plane. The result of applying this method to a collection of DNA sequences is an easily interpretable *Molecular Distance Map* wherein sequences are represented by points in a common Euclidean plane, and the spatial distance between any two points reflects the differences in their subsequence composition.

The graphical representation we use is *Chaos Game Representation* (CGR) of DNA sequences, [21, 22], that simultaneously displays all subsequence frequencies of a given DNA sequence as a visual pattern. CGR has a remarkable ability to differentiate between genetic sequences belonging to different species, see Figure 1, and has thus been proposed as a *genomic signature*. Due to this characteristic, a *Molecular Distance Map* of a collection of genetic sequences may allow inferrences of relationships between the corresponding species.

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

(i) *Chaos Game Representation* (CGR), to graphically represent all subsequences of a DNA sequence s_i, $1 \leq i \leq n$, as pixels of one image, denoted by c_i;

(ii) *Structural Dissimilarity Index* (DSSIM), an "image-distance" measure, to compute the pairwise distances $\Delta(i,j)$, $1 \leq i,j \leq n$, for each pair of CGR images (c_i, c_j), and to produce a distance matrix;

(iii) *Multi-Dimensional Scaling* (MDS), an information visualization technique that takes as input the distance matrix and outputs a Molecular Distance Map in 2D, 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 other DNA sequences s_j in the dataset.

Note that the position of each point in a map is determined by *all* the distances between the sequence it represents and the other sequences in the dataset. For example, in Figure 2, the position of each sequence-point is determined by the 1,790 numerical distances between its sequence and all the other mtDNA vertebrate sequences in that dataset.

The main contributions of the paper are:

- The use of an "image distance" (designed to detect structural similarities between images) to compare the graphic signatures of two DNA sequences. For any given k, this distance simultaneously compares the occurrences of all subsequences of length up to k of the two sequences. In all computations of this paper we use $k = 9$.
- The use of a large dataset of 3,176 complete mitochondrial DNA sequences.
- The use of an information visualization technique to display the results as easily interpretable Molecular Distance Maps, wherein the spatial position of each sequence-point in relation to all other sequence-points is quantitatively significant.
- A method that is general-purpose, simple, computationally efficient and scalable. Since the compared sequences need not be homologous or of the same length, this method can be used to provide comparisons among any number of completely different DNA sequences: within the genome of an individual, across genomes within a single species, between genomes within a taxonomic category, and across taxa.
- An illustration of potential uses of this method by two taxonomical cases, the genus *Polypterus* and the family Tarsiidae.

This method may complement information obtained by using DNA barcodes [13] and Klee diagrams [41], since it is applicable to cases where barcodes may have limited effectiveness: plants and fungi for which different barcoding regions have to be used [23], [18], [39]; protists where multiple loci are generally needed to distinguish between species [17]; prokaryotes [43]; and artificial, computer-generated, DNA sequences. This method may also complement phylogenetic analyses by bringing in additional information gleaned from comparisons of non-homologous and non-coding sequences.

2 Methods

A CGR [21, 22] associates an image to each DNA sequence as follows. Start from a unit square with corners labelled *A, C, G,* and *T*. The starting point of the plot is the center of the square. To plot the CGR corresponding to a given DNA sequence, start reading the letters of the sequence from left to right, one by one. The point corresponding to the first letter is the point plotted in the middle of the segment determined by the center of the square and the corner labelled by the first letter. For example, if the center of the square is labelled "O" and the first letter of the sequence is "A", then the point of the plot coresponding to the first "A" is the point situated halfway between O and the corner A. Subsequent letters are plotted iteratively as the middle point between the previously-drawn-point and the corner labelled by the letter currently being read.

CGR images of genetic DNA sequences originating from various species show rich fractal patterns containing various motifs such as squares, parallel lines, rectangles, triangles and diagonal crosses, Figure 1. CGRs of genomic DNA sequences have been shown to be genome and species specific, [21, 22, 14, 15, 7, 6, 45]. Thus, sequences chosen from each genome as a basis for computing "distances" between genomes do not need to have any relation with one another from the point of view of their position or information content. In addition, this graphical representation facilitates easy visual recognition of global string-usage characteristics: Prominent diagonals indicate purine or pyrimidine runs, sparseness in the upper half indicates low G+C content, etc., see, e.g., [7].

If the generated CGR image has a resolution 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 this paper, for the CGR images of all 3,176 complete mtDNA sequences in our dataset, we used the value $k = 9$, that is, occurrences of subsequences of lengths up to 9 are being taken into consideration. In general, a length of the DNA sequence of 4,000 bp is necessary to obtain a sharply defined CGR, but in many cases 2,000 bp give a reasonably good approximation, [21]. In our case, we used the full length of all analyzed mtDNA sequences, which ranged from 288 bp to 1,555,935 bp, with an average of 28,000 bp.

Other visualizations of genetic data include the 2D rectangular walk [9] and methods similar to it in [32], [25], vector walk [28], cell [48], vertical vector [49], Huffman coding [37], and colorsquare [52] methods. Three-dimensional representations of DNA sequences include the tetrahedron [38], 3D-vector [51], and trinucleotide curve [50] methods. Among these visualization methods, CGR images arguably provide the most immediately comprehensible "signature" of a DNA sequence and a desirable genome-specificity, [21, 6]. In addition, the images produced using CGR are easy to compare, visually and computationally. Coloured versions of CGR, wherein the colour of a point corresponds to the frequency of the corresponding oligomer in the given DNA sequence (from red for high frequency, to blue for no occurrences) have also been proposed [29, 12].

Note that other alignment-free methods have been used for phylogenetic analysis of DNA strings, such as computing the Euclidean distance between frequencies of k-mers ($k \leq 5$) for the analysis of 125 GenBank DNA sequences from 20 bird species and the American alligator, [8]. Another study, [5], analyzed 459 dsDNA bacteriophage genomes and compared them with their host genomes to infer host-phage relationships, by computing Euclidean distances between frequencies of k-mers for $k = 4$. In [33], 75 complete HIV genome sequences were compared using the Euclidean distance between frequencies of 6-mers ($k = 6$), in order to group them in subtypes. In [35], 27 microbial genomes were analyzed to find implications of 4-mer frequencies ($k = 4$) on their evolutionary relationships. In [27], 20 mammalian complete mtDNA sequences were analyzed using the "similarity metric". Our method uses a larger dataset (3,176 complete mtDNA sequences), an "image distance" measure that was designed to capture structural similarities between images, as well as a value of $k = 9$.

Structural Similarity (SSIM) index is an image similarity index used in the context of image processing and computer vision to compare two images from the point of view of their structural similarities [47]. 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 (over 5 million) distances are between 0 and 1, with only half a dozen exceptions of distances between 1 and 1.0033.

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 spatial distances between points reflect the distances between the corresponding items in the distance matrix. Notable examples of molecular biology studies that used MDS are [26] (where it was used for the analysis of geographic genetic distributions of some natural populations), [13] (where it was used to provide a graphical summary of the distances among CO1 genes from various species), and [16] (where it was used to analyze and visualize relationships within collections of phylogenetic trees).

Classical MDS, which we use in this paper, receives as input an $n \times n$ distance matrix $(\Delta(i, j))_{1 \leq i,j \leq n}$ 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 spatial (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 spatial distance between the points p_i and p_j, and f is a function linear in $\Delta(i, j)$. Here, q can be at 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 example, in Figure 2 all mammalian mtDNA sequence-points are coloured red, while in Figure 6 the red points represent mtDNA sequences from the primate suborder Haplorhini and the green points represent mtDNA sequences from the primate suborder Strepshirrini. 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 spatial 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, [24]), 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 [24].

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 [42]. The prokaryotic origins and evolutionary history of mitochondrial genomes have long been extensively 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 2-6, 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, reedfish) 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 [11]. 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). Note that, in answer to the hypothesis in [8] regarding the diversity of signatures across vertebrates, in Figure 2 avian mtDNA signatures cluster neither with the mammals nor with the reptiles, and form a completely separate cluster of their own (albeit closer to reptiles than to mammals).

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 (and can be tested 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 [46], and MATLAB's built-in MDS function[6].

3 Results and Discussion

We applied our method to visualize the relationships among all represented species from the (super) kingdom Protista whose taxon, as defined in the legend of Figure 3, had more than one representative. As expected, the maximum distance between pairs of sequences in this map was higher than the maximum distances for the other maps in this paper, all at lower taxonomic levels.

The most obvious outlier in Figure 3 is *Haemoproteus* sp. jb1.JA27 (#1466), sequenced in [1] (see also [44]), and listed as an *unclassified* organism in the NCBI taxonomy. Note first that this species-point

[6] On-line Supplemental Material includes the annotated dataset and the DSSIM distance matrix, `http://www.csd.uwo.ca/~lila/MoDMap/`. An interactive web tool for easy exploration and navigation of the Molecular Distance Maps in this paper can be found at `http://www.csd.uwo.ca/~rkaramic/MoDMap/index.html`.

belongs to the same kingdom (Chromalveolata), superphylum (Alveolata), phylum (Apicomplexa), and class (Aconoidasida), as the other two species-points that appear grouped with it, *Babesia bovis* T2Bo (#1935), and *Theileria parva* (#3173). This indicates that its position is not fully anomalous. Moreover, as indicated by the high value of *Stress-1* for this figure, an inspection of DSSIM distances shows that this species-point may not be a true outlier, and its position may not be as striking in a higher dimensional version of the Molecular Distance Map. Overall, this map shows that our method allows an exploration of diversity at the level of super kingdom, obtains good clustering of known subtaxonomic groups, while at the same time indicating a lack of genome sequence information and paucity of representation that complicates analyses for this fascinating taxonomic group.

We then applied our method to visualize the relationships between all available complete mtDNA sequences from three classes, Amphibia, Insecta and Mammalia (Figure 4), as well as observe relationships within class Amphibia and three of its orders (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 spatial 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 spatial distances between points, and this can be achieved even when some of the points change their coordinates. In particular, the (x,y)-coordinates of a point representing an amphibian species in the amphibians-insects-mammals map (Figure 4) will not necessarily be the same as the (x,y)-coordinates of the same point when only amphibians are mapped (Figure 5).

In general, Molecular Distance Maps are in good agreement with classical phylogenetic trees at all scales of taxonomic comparisons, see Figure 5 with [36], and Figure 6 with [40]. 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, 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 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, [20]. According to [3], mitochondrial DNA evidence places tarsiiformes as a sister group to Strepsirrhini, while in contrast, [34] places tarsiers within Happlorhini. In Figure 6 the tarsiers are located within the Strepsrrhini cluster, thus agreeing with [3]. This may be partly because both this study and [3] used mitochondrial DNA, whose signature may be different from that of chromosomal DNA.

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, [19], but identical mitochondrial genomes.

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, [40]. 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 [10] 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" range $[0, 0.2]$. The exception is Figure 3 with *Stress-1* equal to 0.26. Note that *Stress-1* generally decreases with an increase in dimensionality, from $q = 2$ to $q = 3, 4, 5 \ldots$. Note also that, as suggested in [2], the *Stress-1* guidelines are not absolute: 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-points in a Molecular Distance Map. The choice of length of DNA sequences used may ultimately depend on the particular dataset and particular application.

We now discuss some limitations of the proposed methods. Firstly, DSSIM is very effective at picking up subtle differences between images. For example, all vertebrate CGRs present the triangular fractal structure seen in the human mtDNA, and are visually very similar. In spite of this, DSSIM is able to detect a range of differences that is sufficient for a good positioning of all 1,791 mtDNA sequences relative to each other. This being said, DSSIM may give too much weight to subtle differences, so that small and big differences in images produce distances that are numerically very close. This may be a useful feature for the analysis of datasets of closely related sequences. For large-scale taxonomic comparisons however, refinements of DSSIM or the use of other distances needs to be explored, that would space further apart the values of distances arising from small differences versus those arising from big-pattern differences between images.

Secondly, MDS always has some errors, in the sense that the spatial distance between two points does not always reflect the original distance in the distance matrix. For fine analyses, the placement of a sequence-point in a map has to be confirmed by checking the original distance matrix. Possible solutions include increasing the dimensionality of the maps to three-dimensional maps, which are still easily interpretable visually and have been shown in some cases to separate clusters which seemed incorrectly intermeshed in the two-dimensional version of the map. Other possibilities include a colour-scheme that would colour points with low stress-per-point differently from the ones with high stress-per-point, and thus alert the user to the regions where discrepancies between the spatial distance and the original distance exist.

Thirdly, we note that the use of the particular distance measure (DSSIM) or particular scaling technique (classical MDS) does not mean that these are the optimal choices in all cases.

Lastly, since the genomic signature of mtDNA can be very different from that of nuclear DNA of the same species, care must be employed in choosing the dataset and interpreting the results.

4 Conclusions

Our analysis suggests that the presence (or absence) of some oligomers in mitochondrial DNA sequences may contain some phylogenetic information. These results are of interest both because of the large dataset considered (see, e.g., the correct grouping in taxonomic categories of 1,791 mitochondrial genomes in Figure 2) and because this information has been extracted from DNA sequences that, by normal criteria, would be considered nonhomologous.

Potential applications of Molecular Distance Maps - when used on a dataset of genomic sequences – include clarification of taxonomic dilemmas, taxonomic classifications, species identification, studies of evolutionary history, as well as possible quantitative definitions of the notion of species and other taxa.

Possible extensions include generalizations of MDS, e.g., to 3-dimensional MDS, for improved accuracy. In addition, higher values of k may be used for comparisons of longer subsequences in case of whole genome analyses. 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 with vertices labelled 00, 01, 10, 11, and then DSSIM and MDS could be employed to compare and map them.

Acknowledgements. We thank Ronghai Tu for an early 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. 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*, page `http://git.io/Ypa_jA`, 2013.

[5] P. Deschavanne, M. DuBow, and C. Regeard. The use of genomic signature distance between bacteriophages and their hosts diplays evolutionary relationships and phage growth cycle determination. *Virology Journal*, 7(163), 2010.

[6] 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.

[7] 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.

[8] S. Edwards, B.Fertil, A.Girron, and P.Deschavanne. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic Biology*, 51(4):599–613, 2002.

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

[10] 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.

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

[12] B. Hao, H. Lee, and S. Zhang. Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*, 11(6):825–836, 2000.

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

[14] 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.

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

[16] D. Hillis, T. Heath, and K. St.John. Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482, 2005.

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

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

[19] 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.

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

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

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

[23] 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.

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

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

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

[27] M. Li, X. Chen, X. Li, B. Ma, and P. Vitany. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

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

[29] M. Makula and L. Benuskova. Interactive visualization of oligomer frequency in DNA. *Computing and Informatics*, 28(5):695–710, 2009.

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

[31] 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.

[32] 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.

[33] A. Pandit and S. Sinha. Using genomic signatures for HIV-1 subtyping. *BMC Bioinformatics*, 11(1), 2010.

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

[35] D. Pride, R. Meinersmann, T. Wassenaar, and M. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*, 13(2):145–158, 2003.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

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

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

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

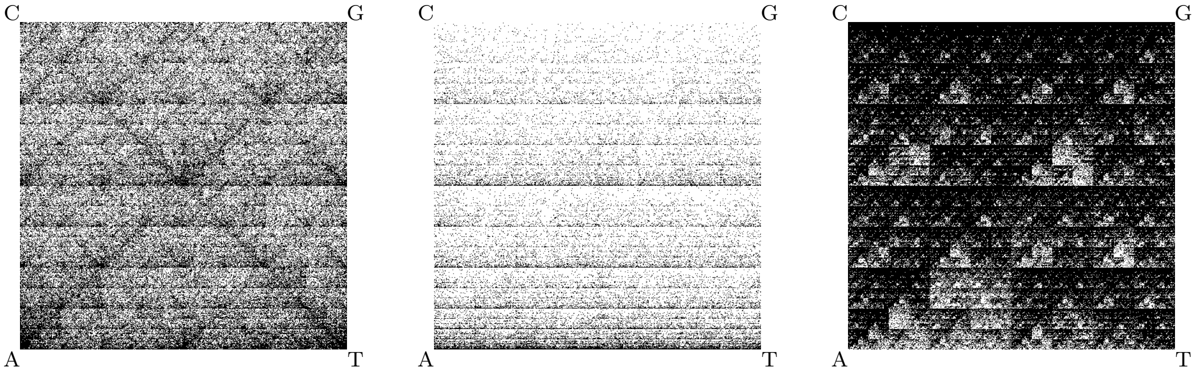

Figure 1: CGR images for three genomes. From left to right: (1) *Marchantia polymorpha* (liverwort) mtDNA, 186,609 bp; (2) *Malawimonas jakobiformis* (flagellate) mtDNA, 47,328 bp; (3) *Rhodobacter capsulatus*, full genome, 3,738,958 bp. Prominent diagonals are indicative of purine (A/G) and pyrimidine (C/T) runs, sparsness in the upper half indicates low G+C content.

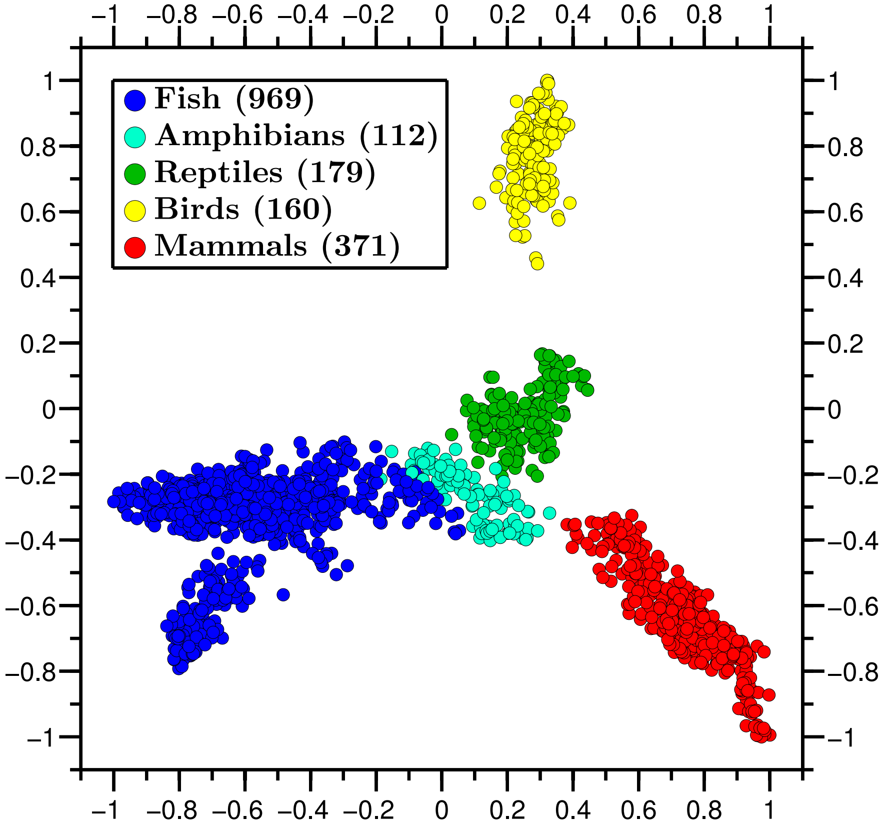

Figure 2: Molecular Distance Map of phylum Vertebrata (excluding the 5 represented jawless vertebrates), with its five subphyla. The total number of mtDNA sequences is 1,791, the average DSSIM distance is 0.8652, and the MDS *Stress-1* is 0.12. Fish species bordering amphibians include fish with primitive pairs of lungs (*Polypterus ornatipinnis* #3125, *Polypterus senegalus* #2868), a fish who can breathe atmospheric air using a pair of lungs (*Erpetoichthys calabaricus* #2745), a toadfish (*Porichtys myriaster* #2483), and all four represented lungfish (*Protopterus aethiopicus* #873, *Lepidosiren paradoxa* #2910, *Neoceratodus forsteri* #2957, *Protopterus doloi* #3119). Note that the question of whether species of the genus *Polypterus* are fish or amphibians has been discussed extensively for hundreds of years.

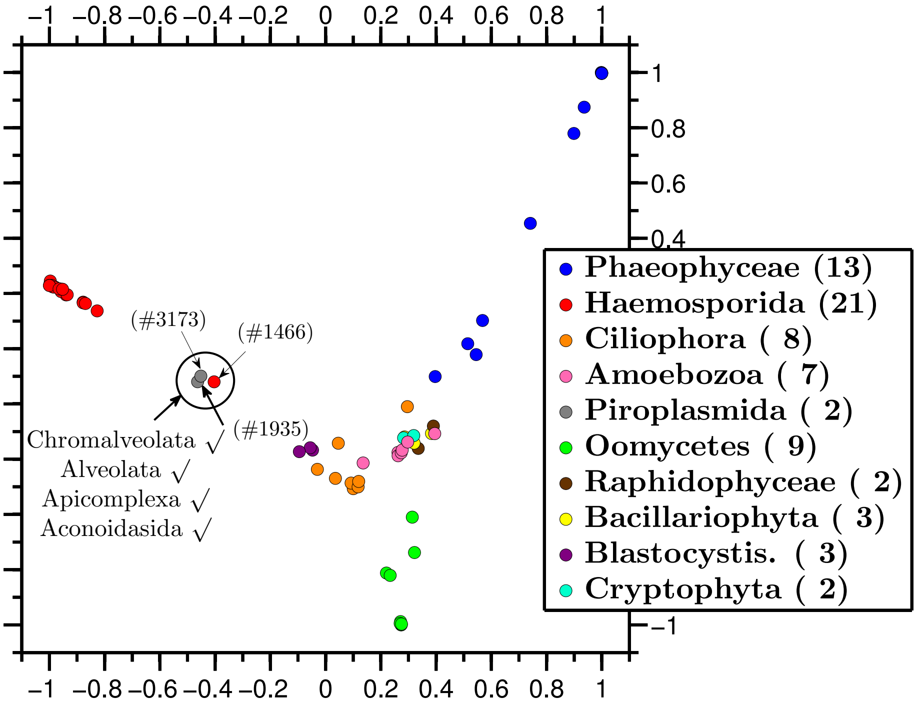

Figure 3: Molecular Distance Map of all represented species from (super)kingdom Protista and its orders. The total number of mtDNA sequences is 70, the average DSSIM distance is 0.8288, and the MDS *Stress-1* is 0.26. The sequence-point #1466 (red) is the unclassified *Haemoproteus* sp. jb1.JA27, #1935 (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, Chromalveolata, Alveolata, Apicomplexa, Aconoidasida, up to the order level.

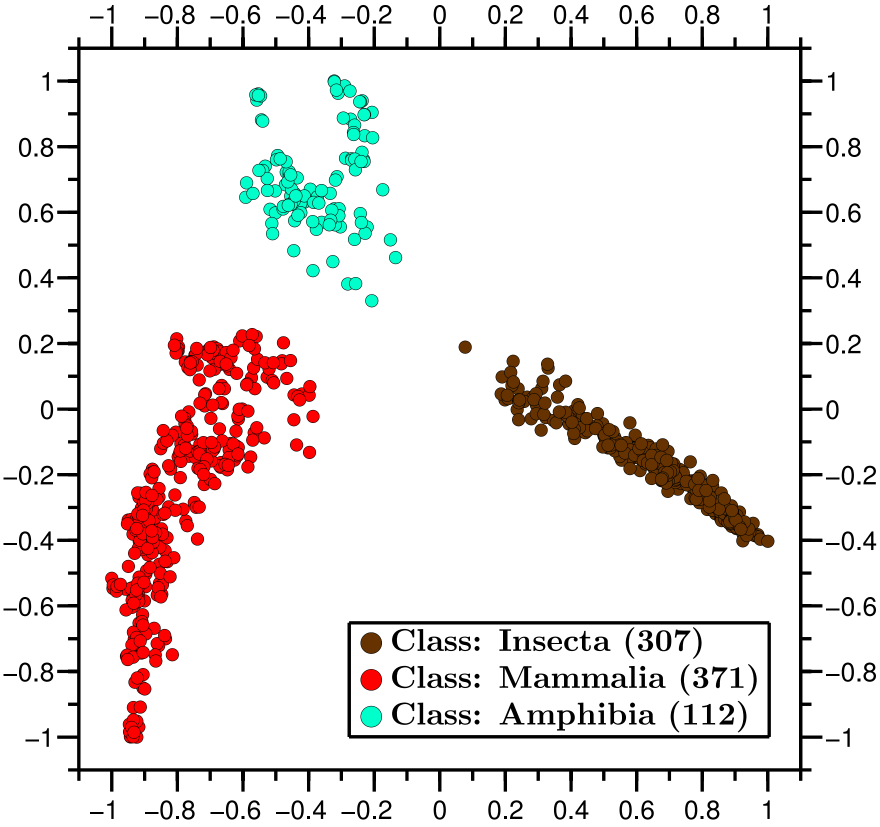

Figure 4: Molecular Distance Map of three classes: Amphibia, Insecta and Mammalia. Gaps and spaces in clusters, in this and other maps, may be due to sampling bias. The total number of mtDNA sequences is 790, the average DSSIM distance is 0.8139, and the MDS *Stress-1* is 0.16.

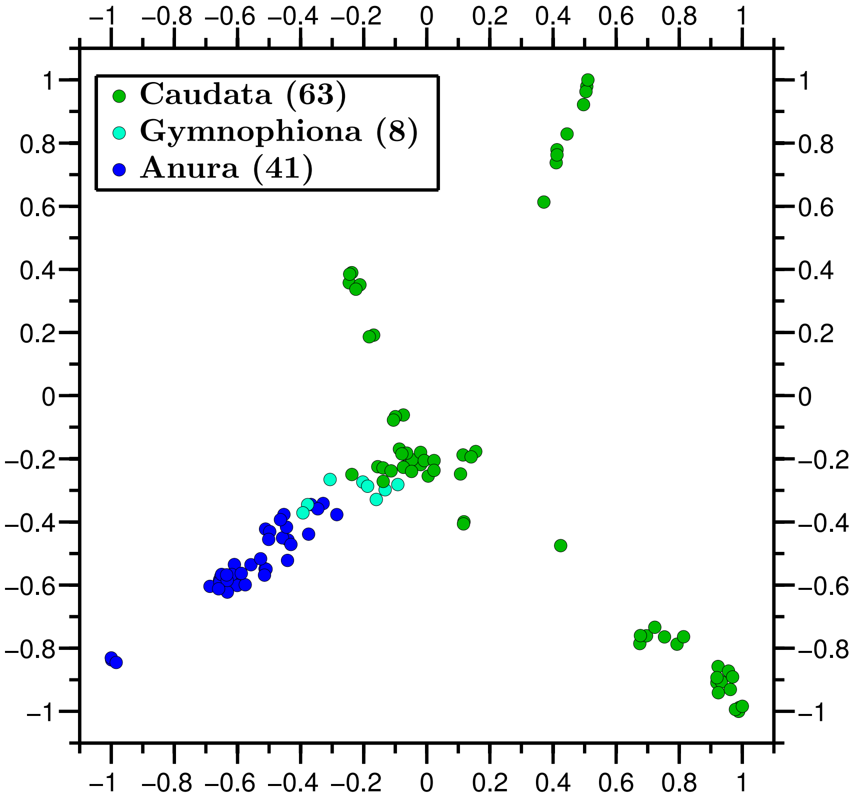

Figure 5: Molecular Distance Map of Class Amphibia and three of its orders. The total number of mtDNA sequences is 112, the average DSSIM distance is 0.8445, and the MDS *Stress-1* is 0.18. Note that the shape of the amphibian cluster and the (x, y)-coordinates of sequence-points are different here from those in Figure 4. This is because MDS outputs a map that aims to preserve pairwise distances between points, but not necessarily their absolute coordinates.

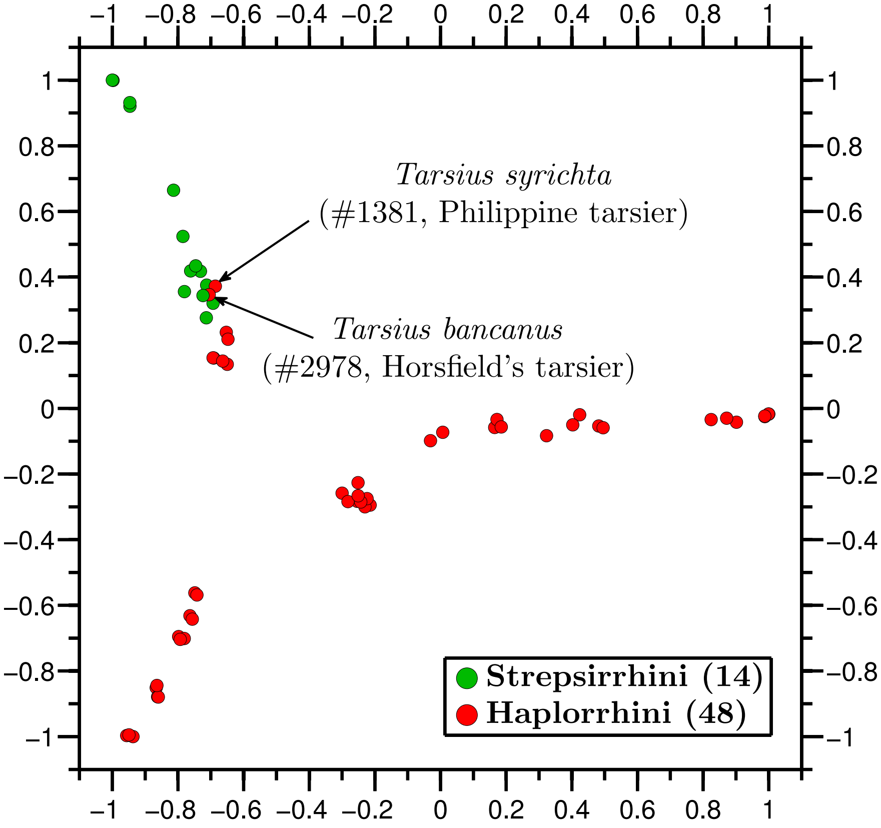

Figure 6: Order Primates and its suborders: Haplorhini (anthropoids and tarsiers), and Strepsirrhini (lemurs and lorises). The total number of mtDNA sequences is 62, the average DSSIM distance is 0.7733, and the MDS *Stress-1* is 0.19. The outliers are *Tarsius syrichta* #1381, and *Tarsius bancanus* #2978, whose phylogenetic placement within the order Primates has been subject of debate for over a century.

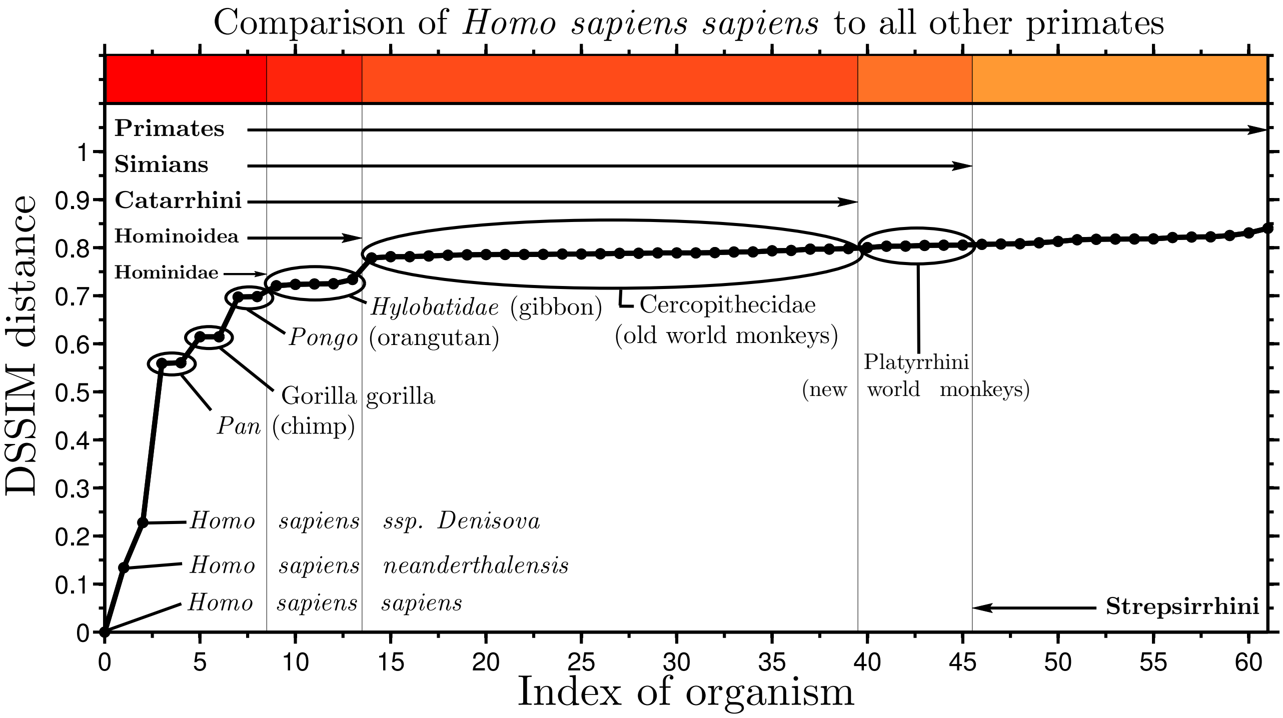

Figure 7: Graph of the DSSIM distances between the CGR images of *Homo sapiens sapiens* mtDNA and each of the 62 primate mitochondrial genomes (sorted by their distance from the human mtDNA). The distances are in accordance with established phylogenetic trees: The species with the smallest DSSIM distances from *Homo sapiens sapiens* are *Homo sapiens neanderthalensis*, *Home sapiens ssp. Denisova*, followed by the chimp.