

September 2016

Molecular Distance Maps: An alignment-free computational tool for analyzing and visualizing DNA sequences' interrelationships

Rallis Karamichalis

The University of Western Ontario

Supervisor

Prof. Lila Kari

The University of Western Ontario

Graduate Program in Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

© Rallis Karamichalis 2016

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>

 Part of the [Other Computer Sciences Commons](#)

Recommended Citation

Karamichalis, Rallis, "Molecular Distance Maps: An alignment-free computational tool for analyzing and visualizing DNA sequences' interrelationships" (2016). *Electronic Thesis and Dissertation Repository*. 4071.
<https://ir.lib.uwo.ca/etd/4071>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact tadam@uwo.ca.

Abstract

In an attempt to identify and classify species based on genetic evidence, we propose a novel combination of methods to quantify and visualize the interrelationships between thousand of species. This is possible by using Chaos Game Representation (CGR) of DNA sequences to compute genomic signatures which we then compare by computing pairwise distances. In the last step, the original DNA sequences are embedded in a high dimensional space using Multi-Dimensional Scaling (MDS) before everything is projected on a Euclidean 3D space.

To start with, we apply this method to a mitochondrial DNA dataset from NCBI containing over 3,000 species. The analysis shows that the oligomer composition of full mtDNA sequences can be a source of taxonomic information, suggesting that this method could be used for unclassified species and taxonomic controversies.

Next, we test the hypothesis that CGR-based genomic signature is preserved along a species' genome by comparing inter- and intra-genomic signatures of nuclear DNA sequences from six different organisms, one from each kingdom of life. We also compare six different distances and we assess their performance using statistical measures. Our results support the existence of a genomic signature for a species' genome at the kingdom level.

In addition, we test whether CGR-based genomic signatures originating only from nuclear DNA can be used to distinguish between closely-related species and we answer in the negative. To overcome this limitation, we propose the concept of “composite signatures” which combine information from different types of DNA and we show that they can effectively distinguish all closely-related species under consideration. We also propose the concept of “assembled signatures” which, among other advantages, do not require a long contiguous DNA sequence but can be built from smaller ones consisting of 100-300 base pairs.

Finally, we design an interactive webtool MoDMaps3D for building three-dimensional Molecular Distance Maps. The user can explore an already existing map or build his/her own using NCBI's accession numbers as input. MoDMaps3D is platform independent, written in Javascript and can run in all major modern browsers.

Keywords: comparative genomics, genomic signature, species classification, alignment-free, chaos game representation, additive DNA signatures, Molecular Distance Maps

Co-authorship statement

This thesis consists of three published articles. The article in chapter three was published in the journal PloS ONE, while the articles in chapter four and five were published in the journal BMC Bioinformatics. The author order follows the conventions of the field: For papers in chapter 4 and 5 the order is alphabetical; The paper in chapter 3 has two senior authors (LK, KAH) and the author order reflects the contributions of the authors. The major individual contributions are listed below.

Chapter 3 contains the article “Mapping the space of genomic signatures” by Lila Kari, Kathleen A. Hill, Abu S. Sayem, Rallis Karamichalis, Nathaniel Bryans, Katelyn Davis, Nikesh S. Dattani. The individual contributions are as follows. Analyzed the data: NB, KAH, NSD. Wrote the paper: LK. Designed the software: NB, NSD, RK, ASS. In-depth analysis: LK, KAH, ASS, RK, NB, KD, NSD.

Chapter 4 contains the article “An investigation into inter- and intragenomic variations of graphic genomic signatures” by Rallis Karamichalis, Lila Kari, Stavros Konstantinidis and Steffen Kopecki. The individual contributions are as follows. RK: data collection; data analysis, methodology and result interpretation; manuscript draft; manuscript editing; software design. LK: data analysis, methodology and result interpretation; manuscript draft; manuscript editing. S.Kon: data analysis, methodology and result interpretation; manuscript editing. S.Kop: data analysis, methodology and result interpretation; manuscript editing.

Chapter 5 contains the article “Additive methods for genomic signatures” by Rallis Karamichalis, Lila Kari, Stavros Konstantinidis, Steffen Kopecki, Stephen Solis-Reyes. The individual contributions are as follows. RK: data collection, data analysis, methodology and result interpretation, manuscript tables and figures, manuscript editing, software design and implementation. LK: data analysis, methodology and result interpretation, manuscript draft, manuscript editing. S.Kon: data analysis, methodology and result interpretation, manuscript editing. S. Kop: data analysis, methodology, result interpretation. S. Solis-Reyes: manuscript

draft (part of Section 1), data collection and analysis (plant experiments), software performance enhancements, language editing.

Acknowledgements

First and foremost I would like to express my sincere gratitude to my supervisor Prof. Lila Kari for her guidance and continuous support throughout my Ph.D. studies. Her research ideas and vision for research topics, together with her enthusiasm, were contagious and motivational for me even during tough times. Her advice on both research and personal career has been invaluable, encouraging me to successfully complete my Ph.D. program.

Besides my supervisor, I would like to thank Prof. Stavros Konstantinidis and Dr. Stefan Kopecki for their insightful comments and the hard questions they have posed during the writing of our research articles. I also thank my fellow labmates Manasi Kulkarni, Amirhossein Simjour, Srujan Enaganti and Stephen Solis-Reyes for the stimulating discussions and the friendly atmosphere maintained in the lab for the last four years.

Special thanks to Marina Chadou for her help and support during my studies in spite of being thousands of kilometers away.

Last but not least, I thank my parents for their continuous encouragement and support in all possible ways.

*To Marina Chadou, who gave me 4×365 proofs of the inequality **LOVE** > **DISTANCE***

Contents

Abstract	ii
Co-authorship Statement	iv
Acknowledgements	vi
List of Figures	xi
List of Tables	xviii
1 Introduction	1
2 Genomic signatures	5
2.1 Literature review	5
2.1.1 Graphical representations of DNA	6
2.1.2 <i>k</i> -mer frequency-based methods	8
2.1.3 Other representations of DNA sequences	11
2.2 Methods	13
2.2.1 Chaos Game Representation (CGR)	13
2.2.2 Distance Measures	17
2.2.3 Multi-Dimensional Scaling (MDS)	23
3 Mapping the space of genomic signatures	48
3.1 Introduction	48
3.2 Methods	50

3.3	Results and Discussion	56
3.4	Conclusions	68
4	An investigation into inter- and intragenomic variations of graphic genomic signatures	75
4.1	Introduction	75
4.2	Methods	80
4.2.1	Dataset	80
4.2.2	Overview	82
4.2.3	Distances	85
4.3	Analysis and Results	94
4.3.1	Quality measures for distances	100
4.3.2	Distance comparison results	107
4.4	Discussion and Conclusions	109
5	Additive methods for genomic signatures	121
5.1	Background	121
5.2	Results	124
5.2.1	Composite DNA signatures	132
5.2.2	Assembled DNA signatures	135
5.2.3	Composite-assembled DNA signatures	142
5.3	Conclusions	142
5.4	Methods	145
6	Molecular Distance Maps 3D (MoDMaps3D)	167
6.1	Introduction	167
6.2	Methods	167
6.3	Software Description	168

7 Conclusion	172
A Copyright Releases	174
B Supplemental Material - Appendices per chapter	176
C Errata	177
Curriculum Vitae	178

List of Figures

- 2.1 The Chaos Game Representation (CGR) of the DNA sequence ACGCTG. . . . 15
- 2.2 The Frequency Chaos Game Representation (FCGR) of the sequence ACGCTGC, for $k = 2$. The first $k - 1$ (here $2 - 1 = 1$) points do not alter the FCGR matrix. . 16
- 2.3 Chaos Game Representation (CGR) of nDNA fragments from *E. coli*, *S. cerevisiae*, *A. thaliana*, *P. falciparum*, *P. furiosus* and *H. sapiens*. 18
- 2.4 An example of computing SSIM similarity values for a set of 6 images. Upper left corner is the original image, hence similarity value is equal to 1. The rest of the images have various pixel values changed, blurred and distorted. The similarity between these images and the original image decreases. This example is part of the examples demonstrated in <https://ece.uwaterloo.ca/~z70wang/research/ssim/> 22
- 2.5 Multi-Dimensional Scaling (MDS) example. The red points are the real positions of 10 big cities in North America. The blue points are the positions of these cities as output of MDS. 24

- 3.1 **CGR images for three DNA sequences.** (a) *Homo sapiens sapiens* mtDNA, 16,569 bp; (b) *Homo sapiens sapiens* chromosome 11, beta-globin region, 73,308 bp; (c) *Polypterus endlicherii* (fish) mtDNA, 16,632 bp. Observe that chromosomal and mitochondrial DNA from the same species can display different patterns, and also that mtDNA of different species may display visually similar patterns that are however sufficiently different as to be computationally distinguishable. 51

3.2 Molecular Distance Map of phylum Vertebrata (excluding the 5 represented jawless vertebrates), with its five subphyla. (a) This Molecular Distance Map comprises 1,791 mtDNA sequences, 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 (*Porichthys 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. Note also that gaps and spaces in clusters, in this and other maps, may be due to sampling bias. (b) Screenshot of the zoomed-in rectangular region outlined in Figure 3.2(a), obtained using the interactive web tool *MoD Map* [35]. 58

3.3 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. . . . 60

3.4	Molecular Distance Map of three classes: Amphibia, Insecta and Mammalia. The method successfully clusters taxonomic groups also at the Class level. Gaps and spaces in clusters, in this and other maps, may be due to sampling bias. A topic of further exploration would be to understand the cluster shapes and nature of the distribution of sequences in this figure. The total number of mtDNA sequences is 790, the average DSSIM distance is 0.8139, and the MDS <i>Stress-1</i> is 0.16.	61
3.5	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 <i>Stress-1</i> 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 3.4. This is because MDS outputs a map that aims to preserve pairwise distances between points, but not necessarily their absolute coordinates.	62
3.6	Molecular Distance Map of order Primates and its suborders: Haplorrhini (anthropoids and tarsiers), and Strepsirrhini (lemurs, lorises, etc.). The total number of mtDNA sequences is 62, the average DSSIM distance is 0.7733, and the MDS <i>Stress-1</i> is 0.19. The outliers are <i>Tarsius syrichta</i> #1381, and <i>Tarsius bancanus</i> #2978, whose placement within the order Primates has been subject of debate for over a century.	64
3.7	Graph of the DSSIM distances between the CGR images of <i>Homo sapiens sapiens</i> mtDNA and the CGR images of 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 <i>Homo sapiens sapiens</i> are <i>Homo sapiens neanderthalensis</i> , <i>Homo sapiens ssp. Denisova</i> , followed by the chimp.	66
4.1	$2^9 \times 2^9$ CGR images of 150 kbp genomic DNA sequences from <i>H. sapiens</i> , <i>E. coli</i> , <i>S. cerevisiae</i> , <i>A. thaliana</i> , <i>P. falciparum</i> , and <i>P. furiosus</i>	83

- 4.2 The first experiment: Two-dimensional Molecular Distance Maps of 150 kbp genomic sequences spanning one complete chromosome from each of six organisms, representing all kingdoms of life. The MoD Maps were obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. Each point corresponds to one 150 kbp genomic sequence from: *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange). . . . 95
- 4.3 The first experiment: Three-dimensional Molecular Distance Maps of 150 kbp genomic sequences spanning one complete chromosome from each of six organisms, representing all kingdoms of life. The MoD Maps were obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. Each point corresponds to one 150 kbp genomic sequences from: *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange). . . . 96
- 4.4 The first experiment (150 kbp fragments spanning one complete chromosome per each of the six organisms): Histograms of pairwise intragenomic and intergenomic distances among the DNA sequences from *H. sapiens* and *A. thaliana*. 98
- 4.5 The second experiment: Two-dimensional Molecular Distance Maps of DNA genomic sequences sampled from the entire genomes of all six organisms, obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. The dataset consists of 10 randomly sampled fragments from each chromosome of multi-chromosome genomes, and all complete fragments from the genomes of *E. coli* and *P. furiosus*, for a total of 526 fragments. Each point corresponds to one such 150 kbp fragment from *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange). 101

4.6	The second experiment: Three-dimensional Molecular Distance Maps of genomic DNA sequences sampled from the genomes of all six chosen organisms, obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. The dataset consists of 10 randomly sampled fragments from each chromosome of multi-chromosome genomes, and all complete fragments from the genomes of <i>E. coli</i> and <i>P. furiosus</i> , for a total of 526 fragments. Each point corresponds to one such 150 kbp fragment from <i>H. sapiens</i> (blue), <i>E. coli</i> (green), <i>S. cerevisiae</i> (red), <i>A. thaliana</i> (turquoise), <i>P. falciparum</i> (magenta), and <i>P. furiosus</i> (orange).	102
4.7	The preview experiment: Two-dimensional Molecular Distance Maps of 150 kbp genomic DNA sequences, randomly sampled from each chromosome (10 fragments per chromosome) of <i>H. sapiens</i> (blue), <i>M. musculus</i> (fuchsia) using the six distances.	111
4.8	The preview experiment: Three-dimensional Molecular Distance Maps of 150 kbp genomic DNA sequences, randomly sampled from each chromosome (10 fragments per chromosome) of <i>H. sapiens</i> (blue), <i>M. musculus</i> (fuchsia) using the six distances.	112
5.1	3D Molecular Distance Map illustrating interrelationships among conventional nDNA signatures of 480 randomly sampled 150 kbp nuclear genomic fragments from <i>H. sapiens</i> (blue) and 128 randomly sampled 150 kbp nuclear genomic fragments from <i>D. melanogaster</i> (orange). The accuracy of separation is 97.2%.	126
5.2	3D Molecular Distance Map illustrating interrelationships among conventional nDNA signatures of 480 randomly sampled nuclear genomic fragments from <i>H. sapiens</i> (blue) and 500 randomly sampled nuclear genomic fragments from <i>P. troglodytes</i> (red). All fragments are 150 kbp long, the accuracy of separation is 52.34%, and no separation plane could be found.	130

5.3	3D Molecular Distance Map illustrating interrelationships among composite DNA signatures using nDNA and mtDNA, of 480 DNA fragments from <i>H. sapiens</i> (blue) and 500 DNA fragments from <i>P. troglodytes</i> (red). The accuracy of separation is 100%.	133
5.4	3D Molecular Distance Map illustrating interrelationships among signatures of 380 DNA fragments from <i>B. napus</i> (magenta) and 180 DNA fragments from <i>B. oleracea</i> (brown) using (a) conventional nDNA signatures, (b) composite DNA signatures using nDNA and mtDNA, (c) composite DNA signatures using nDNA and cpDNA, and (d) composite DNA signatures using nDNA, mtDNA, and cpDNA. The accuracy of separation is 63.03% for (a), and 100% for each of (b), (c), and (d).	134
5.5	3D Molecular Distance Map illustrating interrelationships among 480 composite (respectively 480 composite-assembled) DNA signatures, each using one nDNA fragment and the mtDNA genome from <i>H. sapiens</i> , blue (resp. green); and 500 composite (resp. 500 composite-assembled) DNA signatures, each using one nDNA fragment and the mtDNA genome from <i>P. troglodytes</i> , red (resp. turquoise); For the composite-assembled DNA signatures, the length of contigs was $n = 100$, while the number of contigs was 4,500 for each 150 kbp nDNA fragment, and 497 (resp. 496) for the human (resp. chimp) mtDNA genome. The accuracy of separation between the <i>H. sapiens</i> and the <i>P. troglodytes</i> sequences was 58%, but the existence of a separation plane was verified.	143
5.6	Conventional nDNA signatures of 150 kbp sequences of the pivot organisms from Kingdom (a) Animalia, (b) Fungi, (c) Plantae, (d) Protista, (e) Bacteria, and (f) Archaea.	146

5.7	(a) Conventional nDNA signatures, and (b) composite (nDNA + cpDNA) signatures of <i>Capsicum annuum</i> L, cultivar <i>Zunla-1</i> (domesticated) shown in light green, and <i>Capsicum annuum</i> var. <i>glabriusculum</i> , cultivar <i>Chiltepin</i> (wild) shown in grey.	158
6.1	Molecular Distance Map of Phylum Vertebrata, consisting of 3,158 mtDNA sequences. Blue, cyan, green, red and yellow, represent fish, amphibia, reptiles, mammals and birds mtDNA genomes respectively. Enlarged version of left and right panel can be found in Figure 6.2	170
6.2	Enlarged version of left and right panel of MoDMaps3D.	171

List of Tables

4.1	Dataset for the first experiment: NCBI accession numbers of the complete chromosomes considered, in increasing order of their NCBI accession number.	81
4.2	The first experiment: Organisms considered, total length of the chromosome (respectively genome), number of ignored letters “N”, and number of DNA fragments (sequences) obtained by splitting a single complete chromosome per organism into consecutive, non-overlapping, equal length (150 kbp) contiguous fragments.	81
4.3	The first experiment: Mean and standard deviation of distances between clusters $C_i - C_j$ for $i, j = 1, \dots, 6$	99
4.4	The first experiment: Summary of quality measures for the performances of six distances (DSSIM, descriptor, Euclidean, Manhattan, Pearson, approximated information distance) on a dataset of 508 genomic DNA sequences spanning one complete chromosome for multi-chromosomes organisms and the complete genome otherwise, of one organism from each kingdom of life. \mathcal{D}_α is the correlation to an idealized cluster, \mathcal{A}_α the silhouette cluster accuracy, and O_α the relative overlap. Higher is better.	107

- 4.5 The second experiment: Summary of quality measures for the performances of six distances (DSSIM, descriptor, Euclidean, Manhattan, Pearson, approximated information distance) on a dataset of 526 genomic DNA sequences sampled randomly (10 fragments per chromosome for multi-chromosome organisms, and all fragments of the genome otherwise) from the genomes of organisms from each kingdom of life. \mathcal{D}_α is the correlation to an idealized cluster, \mathcal{A}_α the silhouette cluster accuracy, and O_α the relative overlap. Higher is better. 108
- 4.6 The preview experiment: Summary of quality measures for the performances of six distances (DSSIM, descriptor, Euclidean, Manhattan, Pearson, approximated information distance) on a dataset of 450 DNA sequences, sampled from the entire genome (10 fragments per chromosome) of *H. sapiens* and *M. musculus*. \mathcal{D}_α is the correlation to an idealized cluster, \mathcal{A}_α is the silhouette cluster accuracy, and O_α is the relative overlap. Higher is better. 110
- 5.1 Each subtable summarizes, for a given kingdom, the results of pairwise comparisons between DNA signatures of fragments from a pivot organism (blue) and those from one other organism, at increasing levels of relatedness. The first two result columns indicate the outcome of the comparisons of conventional nDNA signatures, and the last two columns the comparisons of composite DNA signatures. Green indicates that separation was achieved with AID, red indicates that separation was not achieved with any of the six distances listed in Section 5.2, and yellow (Y/N) or Y* indicate results discussed in the text. The columns labelled Acc % indicate the accuracy of the separations listed immediately at their left: Acc > 85% was considered separation. A dash indicates that no sequenced data was available on NCBI/GenBank at the time of this submission. The corresponding 3D Molecular Distance Maps for each of the comparisons can be found in [58]. 127

5.2 (A) through (C) – Distances between the conventional nDNA signature of a fragment and its assembled DNA signatures, for various numbers r of contigs of the same length n : (A) distances to fully-assembled DNA signatures; (A') theoretical upper bounds for (A); (B) distances to assembled DNA signatures; (C) same as (B), when tripling the number of contigs. (B') through (C') – Distances between the conventional nDNA signature of a fragment and its assembled DNA signatures, using variable-length contigs taken from a normal distribution $N(n, \sigma)$, with mean n and variance $\sigma = 40$. The nDNA fragment used was from *H. sapiens*, chromosome 21, fragment 20 (from position 2,850,001 to 3,000,000 after removing all Ns in the original sequence). 136

5.3 Each subtable summarizes, for a given kingdom, the results of pairwise comparisons between DNA signatures of fragments from a pivot organism (blue) and those from one other organism, at increasing levels of relatedness. The first two result columns indicate the outcome of the comparisons of conventional nDNA signatures, and the last two columns the comparisons of composite DNA signatures. Green indicates that separation was achieved with AID, red indicates that separation was not achieved with any of the six distances listed in Section 5.2, and yellow (Y/N) or Y* indicate results discussed in the text. The columns labelled Acc % indicate the accuracy of the separations listed immediately at their left: Acc > 85% was considered separation. A dash indicates that no sequenced data was available on NCBI/GenBank at the time of this submission. The corresponding 3D Molecular Distance Maps for each of the comparisons can be found in [58]. 139

Chapter 1

Introduction

Classic alignment-based methods (DNA barcoding [3], Klee diagrams [12], multiple sequence alignments [11]) have been used extensively for classification and identification of genomic sequences. Alignment-free methods provide an alternative for this task while having a few other advantages in terms of speed and applicability. After Karlin *et al.* suggested in [10] that k -mer frequencies can play the role of a genomic signature, there was an increasing interest in the bioinformatics community to further explore and analyze genomic signatures. Jeffrey in [4, 5] introduced the use of Chaos Game Representation (CGR) of a DNA sequence giving a visual aspect to its structural properties, while later the study of CGRs was standardized by Deschavanne *et al.* in [2, 1].

Our goal is to find a general, universal method of classification based on the structural composition of genomic DNA. In this thesis, we continue the exploration of genomic signatures using CGRs and we extend results from other studies which were either qualitative or very limited in scope. In particular, we investigate whether or not CGR-based signatures can indeed act as genomic signatures. We also investigate the hypothesis that genomic signatures are preserved along a species' genome. Finally, we test the discriminating power of CGRs for closely related species and we generalize genomic signatures by introducing two new types: “composite signature” and “assembled signature”.

Our proposed algorithm consists of three main components. Firstly, Chaos Game Representation (CGR) is being used to visualize and quantitatively express the syntactic structure of a DNA sequence. Secondly, a distance measure is employed to compute distances between CGRs of different DNA sequences. Notable distances being used in this thesis are Approximated Information Distance (AID), Structural Similarity Index (SSIM) and Descriptors-based distance along with classical numerical distances such as Euclidean, Manhattan and Pearson distance. Finally, Multi-Dimensional Scaling (MDS) is used to reduce the dimensionality and efficiently embed the datapoints representing DNA sequences into a 2D or 3D Euclidean space, producing a Molecular Distance Map (MoDMap).

Our research findings are organized in the following way. Chapter 3 contains the article “Mapping the space of genomic signatures” [9] in which we perform an analysis of a mitochondrial (mtDNA) dataset from the National Center for Biotechnology Information (NCBI) exploring phylogenetic relationships and getting deeper insights on various unclassified species. Results of this extensive analysis confirm that the oligomer composition of full mtDNA sequences can be a source of taxonomic information. Chapter 4 contains the article “An investigation into inter- and intragenomic variations of graphic genomic signatures” [7], in which we test the hypothesis of DNA genomic signatures being preserved along the genome of the same organism, while being dissimilar for DNA sequences originating from different organisms, at the kingdom level. We also assess six different distance measures and rank their performance based on statistical measures. Results suggest that several distances outperform the Euclidean distance, which has so far been almost exclusively used for such studies. Chapter 5 contains the article “Additive methods for genomic signatures” [8], in which we test the discriminating power of conventional CGR signatures of nuclear DNA sequences and we find, unexpectedly, that they do not suffice for distinguishing closely related species, for example *H.sapiens* and *P.troglodytes*. To overcome these limitations we extend the notion of genomic signature by proposing the use of composite signature which combines in general information from various types of DNA. We also propose the notion of assembled signature and we show that it can effi-

ciently distinguish genomes even using less information than conventional genomic signatures. Finally, in Chapter 6 we present a web tool to explore, analyze and visualize genomic diversity on various DNA datasets [6]. The tool is written in JavaScript, is platform independent and can run in any modern web browser.

We conclude this thesis with Chapter 7, which contains a discussion about possible extensions of current work, including the search for a “representative” genomic signature of a species and haplogroup identification using human mitochondrial DNA data to track maternal lineage. Far more challenging tasks include backtracking paternal lineage in the Y chromosome and testing the ability of this method to distinguish between healthy and unhealthy populations with large scale mutations.

Bibliography

- [1] P. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, and B. Fertil. Genomic signature is preserved in short DNA fragments. *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, 2000.
- [2] 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.
- [3] Paul DN Hebert, Alina Cywinska, Shelley L Ball, et al. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):313–321, 2003.
- [4] H. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.

- [5] H. Jeffrey. Chaos game visualization of sequences. *Computers & Graphics*, 16(1):25–33, 1992.
- [6] R. Karamichalis. Molecular Distance Maps 3D. <https://github.com/rallis/MoDMaps3D>, 2016.
- [7] R. Karamichalis, L. Kari, S. Konstantinidis, and S. Kopecki. An investigation into inter- and intragenomic variations of graphic genomic signatures. *BMC Bioinformatics*, 16:246, 2015.
- [8] R. Karamichalis, L. Kari, S. Konstantinidis, S. Kopecki, and S. Solis-Reyes. Additive methods for genomic signatures. *BMC Bioinformatics*, 17:313, 2017.
- [9] L. Kari, K. Hill, A. Sayem, R. Karamichalis, N. Bryans, K. Davis, and N. Dattani. Mapping the space of genomic signatures. *PloS ONE*, 10(5):e0119815, 2015.
- [10] S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, 11(7):283–290, 1995.
- [11] Dan E Krane. *Fundamental concepts of bioinformatics*. Pearson Education India, 2003.
- [12] Lawrence Sirovich, Mark Y Stoeckle, and Yu Zhang. Structural analysis of biodiversity. *PLoS One*, 5(2):e9266, 2010.

Chapter 2

Genomic signatures

This chapter is organized in the following way. Firstly, an extensive literature review of research on genomic signatures is given, presenting a timeline of research in this area. Various methods and tools that have been used for analyzing DNA sequences are presented. Secondly, the main methods employed throughout this thesis are presented. We give the definition of Chaos Game Representation (CGR) of a DNA sequence and its improved version, Frequency Chaos Game Representation (FCGR), together with an example. We also present various distance measures to compute distances between CGRs. Finally, we present Multi-Dimensional Scaling (MDS) which is used for dimensionality reduction in the final stage of our proposed algorithm.

2.1 Literature review

Deoxyribonucleic Acid (DNA) is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms. As such, DNA has become a subject of both theoretical and applied studies for the last decades. DNA is a polymer of nucleotides. Nucleotides are the building blocks, or monomers of DNA. Each nucleotide is made of a phosphate, a deoxyribose sugar and a nitrogen base. The four different nucleotides of DNA are Adenine (A), Cytosine (C), Guanine (G), Thymine (T). DNA can be single stranded

or double stranded. In a double stranded DNA, the nucleotides are pairwise complementary, A is complementary to T, C is complementary to G. With this in mind, we can represent any DNA sequence as a string over a 4-letter alphabet consisting of letters A,C,G and T. In this thesis, we use various types of DNA, namely nuclear DNA (nDNA) which is the DNA located in the nucleus of eukaryotic cells, mitochondrial DNA (mtDNA) which is the DNA located in an organelle called mitochondrion and is responsible for energy production of the cell, chloroplast DNA (cpDNA) which is the DNA located in chloroplasts found mainly in land plants and algae, and plasmid DNA (pDNA) which is the DNA of plasmids mainly found in prokaryotes.

2.1.1 Graphical representations of DNA

After the first successful sequencing method reported by Fred Sanger in 1977 [149], various methods have been proposed to represent, explore and analyze DNA data. Initial studies proposed simple pictorial representations where each nucleotide is replaced with a specific symbol to represent it [121], or gap plots which visualize positional correlations and periodic patterns in a DNA sequence [102]. They were followed by representations of DNA sequences as random walks in 2D [121, 54, 124, 102, 118] where a DNA sequence is represented as a curve in a plane where the four possible moves, left/right and up/down, encode the four nucleotides. Enhanced versions of the 2D random walk were later proposed, namely the DB-curve which reduced degeneracy [185], and a modified version which eliminated degeneracy completely [192, 106]. Another approach, this time in 3D, introduced by Zhang *et al.* [204], was the Z-curve which was used to recognize coding protein genes in yeast [203], to build a database of Z-curves for more than 1000 genomes [202], and to build phylogenetic trees for 24 coronaviruses [205]. A variation of Z-curve, the L-curve, was introduced in [110]. 4D curves based on Z-curves were also introduced in [162].

Applications of random walks to DNA datasets include the construction of phylogenetic trees for 12 primates using mtDNA and Euclidean distance [109], the construction of a similarity matrix between 11 mammals based on the coding sequences of the first exon of the

β -globin gene using both Euclidean and a custom Cosine distance in [108], and the use of the Euclidean distance in 3-component vectors in [191]. Other studies perform analysis computing similarity matrices using random walks, as for example for 8 coding sequences of eukaryotes to reduce degeneracy of 2D curves [114], 50 β -globin genes and 127 protein kinase C enzymes to build “moment vectors” [193], 7 mammalian sequences using a custom distance called “Similar Factor” [75], ND5 proteins from 9 mammals [180], 35 mammalian mtDNA, 33 primate lentiviruses and 30 coronaviruses using Euclidean distance [194], and 35 mammals using whole mtDNA and Euclidean distance [76]. A Huffman-encoded version of 2D walk for 1-mers for the first exon of β -globin gene from 11 mammals was reported in [141].

Another approach to depict a DNA sequence is using a cell representation. Randic *et al.* was one of the first to use this method in order to overcome the problem of arbitrary assignment of the four nucleotides to symbols. His approach was based on the construction of a 12-component vector whose components are the leading eigenvalues of the L/L matrices (length/length) associated with the DNA sequence. This method was used for the first exon of β -globin region of 11 mammals [145]. Various modifications along with tweaks and optimizations followed this study [112, 111, 190, 28, 107, 140, 142, 195] working mainly on small sets (less than 15) of exons of β -globin regions of mammals.

Finally, another visual attempt to represent and analyze DNA sequences was accomplished with the introduction of two 2D diagrams. These methods, although visually informative do not always provide quantitative structural insights of the DNA sequences under consideration. Examples of such studies is in [143] where a 2D diagram is formed in which all “spots” have integer coordinates and by using only distances between spots having the same x or the same y coordinate one can construct a “map profile”. This method is applied on DNA sequences of the first exon of human β -globin. Similarly, in [144] a four-color map representation of DNA and RNA sequences along with a similarity measure were introduced and illustrated with the coding sequence of the first exon of the human β -globin gene, and in [207] where a “ColorSquare” representation is introduced and its efficiency is explored for a set consisting of

the first exon of β -globin gene for 10 mammals.

Detailed reviews about most of the graphical representations described above can be found in [147, 125].

2.1.2 *k*-mer frequency-based methods

Many other tools have been used to perform statistical analysis and explore properties of DNA sequences in terms of *k*-mer composition. Philips *et al.* used Markov Chain of order 3 in nDNA of *E.coli* [135], and later the same method was used for nDNA of yeast [10]. Many studies that followed, suggested possible explanations for asymmetric *k*-mer frequencies among which are that scarcity of CG in nDNA may reflect a requirement for mRNA stability [15], that scarcity of GATC in enterobacteriophages may be a result of the methyl-directed mismatch repair system [39], that scarcity of 4-mer and 6-mer palindromes in bacteria and bacteriophages may be because of restriction/methylation regimes, recombination and transcription processes [91], that changes in 4-mer frequencies in nDNA of *E.coli* may have been altered by “Very Short Patch” repair process [16], that excess/scarcity of some 2, 3, 4-mers in gDNA, mtDNA and virusDNA may be due to DNA/RNA structure and regulatory mechanisms [25], that excess/scarcity of some 2, 3, 4-mers in phages, animal mtDNA, bacteria nDNA, vertebrate nDNA and chloroplasts cpDNA may be because of DNA structures (dinucleotide stacking energies, DNA curvature and superhelicity, nucleosome organization), context-dependent mutational events, methylation effects and processes of replication and repair [94, 90, 95] and that scarcity of 4, 5, 6-mer palindromes in bacterial and archaeal nDNA may be due to restriction enzymes [55].

Based on these observations, Karlin and Burge suggested, in [90], that *k*-mer frequency can play the role of a *genomic signature* and since then, *k*-mer frequencies have been widely used as a means to compare genomic sequences. Dinucleotide Relative Abundance Profiles (DRAP) were proposed by Burge *et al.* to explore evolutionary relationships using 2-mers [25]. Relative Abundance Profiles for various *k*-mer lengths ($k = 2, 3, 4$) were used extensively

along with Manhattan and/or Mahalanobis distance for analyzing DNA datasets such as 19 eukaryotes [93], *E. coli* and phage nDNA [94], nDNA from 51 prokaryotes and cpDNA from 11 plants [89], prokaryote nDNA, pDNA and mtDNA sequences [26], 504 bacterial pDNA and 230 nDNA [160], nDNA from 50 microbes [19], eukaryote nDNA [56], nDNA from 22 different species [83], and HIV-1 genomes from different years [131]. Dinucleotide Relative Abundance Profiles were later generalized to tetranucleotides (4-mers) and used to classify 27 microbial nDNA sequences [136], and to study inter-genomic distances among 636 prokaryotes [20].

As researchers started using frequency vectors of longer k -mers of lengths ranging from $k = 4$ to $k = 8$ the applications of this method became apparent. In the majority of these studies weighted or standardized Euclidean distance have been used, or slightly modified versions of them [68, 182, 183]. Typical examples include building phylogenetic trees from nDNA sequences of 8 amniotes [152], 20 mammalian mtDNA sequences and 48 Hepatitis E genomes [32], and RAG1 genes from 46 vertebrates, 18S rRNA sequences from 93 plants and nDNA from 16 proteobacteria [27].

Based on k -mer vectors, more sophisticated approaches were used over time. Blaisdell *et al.* in [18] used Euclidean distance between k -mer frequencies as “multiplet distribution distance” and counting of bases not occurring in significantly long common words as “complements of long words” to generate trees for mammalian α and β globin genes. A method used in stochastic processes, “Return Time Distribution”, was used to build a phylogenetic tree from 100 sets of *Flaviridae* nDNA in [96]. In the context of DNA sequences, the return time is the number of nucleotides in between the reappearance of a particular nucleotide. Moreover, comparing sorted k -mer frequency vectors was later proposed in [189] and was tested in constructing phylogenies for 48 *Hepatitis E* genomes and 42 HIV-1 genomes, while in [132] machine learning methods were employed to “learn” a distance for classifying more than 1,000 microbial sequences. Jensen-Shannon distance between k -mer vectors for some specifically selected k -mers, named “feature frequency profiles”, was proposed in [153] and used for producing

phylogenetic trees from intron sequences of 10 mammals. The same technique was used also in [181] for 142 dsDNA eukaryote viruses, in [154] for 36 nDNA sequences taken from *E. coli* and some *Shigella* species, in [100] for 27 primate mtDNA and 13 Malvidae/plant nDNA, in [167] for 377 *H.pylori* genomes, and a webtool based on this was build in [74]. Using vectors of k -mer occurrences one can also use the number of k -word matches as a distance between two sequences. This distance which is usually denoted by D_2 is equivalent to the dot product of the k -mer occurrence vectors [113]. A variation of this, that considers up to t mismatches, is denoted by D_2^t [24], while also some standardized and weighted versions exist [86, 146, 115, 14]. Many studies exist that study the statistical distribution properties [113, 24, 49, 48, 146, 84] and statistical power [173, 115, 157] of these distances. Such distances have been used for comparison of regulatory sequences [86] and for construction of phylogenetic trees for nDNA of 5 mammals and 13 tree species from NGS reads [157]. A few other studies used methods derived from k -mer vectors and usually mixed with custom distance measures. An example of such a method is the “discrimination measure” introduced in [46], which uses primitive discrimination substrings and was illustrated for 10 mammalian whole β -globin and 24 coronaviruses. Another example is the “natural vectors” introduced in [35], which are based on normalized central moments and were tested with 51 influenza viruses, 99 human rhinovirus and 31 mammalian mtDNA. One more example is the “underlying approach” in [31], which is based on subword composition and tested with 54 H1N1 viruses, 18 prokaryote nDNA and 5 *Plasmodium* nDNA. Finally, a last example is the entropy of Gamma distribution in [188], which is based on a k -mer voting model and was presented and tested with 30 mammalian mtDNA.

The range of applications of k -mer vectors is far from being limited to DNA. A significant number of studies also focuses on using k -mer vectors for peptide sequences. A widely used method named CVTree [61] constructs phylogenetic trees using cosine distance between “composition vectors”. Applications of this method with minor variations, using values of $k = 5$ or 6 include building phylogenies for protein transcripts of over 80 archaea and bacteria nDNA

sequences [61], 442 proteins of 34 mammals [158], 21 plant chloroplasts together with several eubacteria, archaea, and eukaryote sequences [29], 139 prokaryotes [137], 16 archaea, 87 bacteria, and 6 eukaryotes [138], 124 dsDNA viruses [51], 82 fungi [174] and 109 eukaryotes, 34 plant chloroplasts and 62 alpha-proteobacteria [201]. Successful applications of k -mer vectors have also been reported in metagenomics for the classification of bacterial nDNA fragments from different species [148, 164]. In addition to this, k -mer vectors have been used along with Self-Organizing Maps (SOM), in order to classify hundreds of thousands of short prokaryote sequences into different phylogenetic groups in [3, 2, 120] and for various *Drosophila* genomes in [1]. Other applications of k -mer frequencies include detection of horizontal transfers [79]. Efficient algorithms for parallel counting of k -mers have been developed as well [119]. Logic Alignment Free (LAF) was also introduced and applied to bacterial genomes in [179]. LAF combines alignment-free techniques and rule-based classification to assign biological samples to their taxa, by searching for a minimal subset of k -mers whose relative frequencies are used to build classification models as disjunctive-normal-form logic formulas. Finally, other studies transform the problem of classifying DNA sequences based on k -mer frequencies into classifying signals coming out of those. Initial studies analyzed small sets of genes [9], while later whole genome comparison using genomic signals was tested for eukaryotes in [155] and for prokaryotes in [151].

Many of the distances described for k -mer vectors have been compared and benchmarked in [184, 73, 72, 33, 34, 58, 80, 63], and detailed reviews of the literature can be found in [92, 169, 123, 21, 150, 156].

2.1.3 Other representations of DNA sequences

Apart from the above representations which have been studied in depth, there is a number of studies which used different approaches to analyze and compare DNA sequences. Markov models have been used to cluster about 30 coding nDNA sequences in [17] and to explore within-sequence variations for animal mtDNA and some viruses in [30]. A distance between

Markov models was used to build phylogenetic trees for genes of *E.coli* and *S.flexneri* in [134], while a “weighted relative entropy” between Markov models was used to build phylogenies of 48 Hepatitis E viruses.

A few other studies focus on the multifractal analysis of the “measure representation” of DNA sequences, and were applied to bacterial nDNA in [197] and bacteria whole peptide transcript in [198]. Multifractal analysis has also been applied to exon and intron sequences [186] and to the human genome [53, 122]. Many statistical properties have been investigated in [159], and a fractal model to simulate phylogenetic relationships has been proposed in [199].

Some other studies perform Lempel-Ziv complexity-based analysis of DNA, where distance measures are defined based on complexity measures of the sequences analyzed. Such methods were used to build phylogenies for 30 mammalian mtDNA in [127] using 4 distances based on Lempel-Ziv complexity, for *Candida cytochrome b* and 18S rRNA for some medically relevant Fungi in [12], for 26 placental mammal mtDNA in [104], for the first exon of β -globin gene for 10 mammals together with 12 H5N1 genomes in [114], for various protein datasets in [4], for 48 Hepatitis E viruses and 18 mammal mtDNA in [78], for 38 mammal mtDNA in [175] and for 16 rRNA ITS region of *Galanthus* plants in [11]. Various modifications of complexity-based distances have been defined also in [97, 47].

Another type of studies focus on the average length of the longest common substrings (and various modifications of it) in order to extract information for a set of sequences. In [166] authors use a custom distance measure between sequences which is related to information-theoretic tools (KullbackLeibler relative entropy) and use it to construct phylogenies for 34 mammal mtDNA, 191 proteomes and many ssRNA viruses. A generalization of this method also exists, that considers the average length of the longest common substrings with k mismatches, and was illustrated for 27 primate mtDNA and 32 *Roseobacter* nDNA in [101]. Guyon *et al.* proposed an evolutionary distance on maximal unique matches between sequences and applied that to *Gammaproteobacteria* [59]. Similar to this idea is the use of the number of substitutions and/or mismatches per site. In [65] an estimator of the number of substitu-

tions was derived and used for 27 primate mtDNA, 8 *Streptococcus agalactiae* strains and 12 *Drosophila* nDNA to compare the results against three other measures, while in [42] a faster version of [65] was reported and illustrated for 825 HIV-1 strains and 13 *E.coli* strains. A similar estimator, this time for pairwise mismatches, was presented in [66] for 37 *D.melanogaster* strains, and an improvement of it in [64] for 21 *Drosophila* species.

Various other representations combined with custom distances have been used in literature to analyze DNA datasets. A “Standardized Hasse” distance between Hasse matrices, based on partial ordering rules, was used for the first exon of β -globin genes from 8 mammals in [165]. Pattern-comparison based on linear predictive coding and its spectral distortion measure was used to classify genes from *E.coli* and *S.flexneri* in [133]. The Euclidean distance between 16-dimensional vectors, called “base-base correlations”, were used for 48 Hepatitis E viruses and many prokaryote nDNA in [116] and for many coronaviruses in [117]. The Pearson distance between “average mutual information” profiles was used for classifying HIV-1 genomes by subtype in [13], and the Euclidean distance between information correlation matrices like [13] was tested for 218 dsDNA viruses in [52]. Finally, adjacency matrices of weighted digraphs were used with Euclidean, Cosine and Pearson distance for mtDNA genes of 12 primates in [139], the difference in free energy of nearest-neighbour interaction was presented in [206], Fourier transforms were proposed as a genomic signal processing distance and tested for 26 eukaryote 18S rRNA in [23], and “variable length local decoding” based on prefix codes was illustrated for 117 Hepatitis C viruses in [41].

2.2 Methods

2.2.1 Chaos Game Representation (CGR)

Chaos Game Representation (CGR) of a DNA sequence is a graphical representation of a sequence. Visual patterns produced in a CGR image depict the distribution of frequencies of all k -mers in the sequence. Originally proposed by Jeffrey, [81, 82], as a means of visual inspection

of DNA sequences, CGRs were initially used to analyze DNA sequences qualitatively [99, 69, 70, 60]. This led to the hypothesis expressed by Dutta *et al.* and Goldman, that CGR images represent no more information than second-order Markov chains [44, 57]. This hypothesis was later disproven by Almeida *et al.* [6, 5] and others [176, 88]. Deschavanne *et al.* [38, 37] were the first to suggest that CGR is a good candidate for the role of genomic signature. The way a CGR of a DNA sequence is produced is as follows. Starting with a unit square with its four vertices labelled A, C, G, and T, clockwise starting from the bottom-left corner, we plot the very first point in the center of the square. We then read the sequence from left to right letter by letter, until the end of the sequence. For each letter being read, we plot a point in the middle of the segment connecting our currently drawn point and the vertex labelled with the letter we just read. An example demonstrating the procedure of plotting a CGR for a DNA sequence can be found in Figure 2.1. A set of various CGR plots can be found in Figure 2.3.

It turned out however, that representing a DNA sequence as a set of points being plotted in a unit square has its own limitations in terms of resolution and computer precision. This problem was later solved by Deschavanne *et al.* suggesting Frequency CGR (FCGR) as an extension of conventional CGR, where unit squares are in fact matrices of dimensions $2^k \times 2^k$, where k is the resolution of FCGR. Each matrix entry represents the number of occurrences of a specific substring of length k in the original sequence. This way, FCGR can quantitatively express the structure and complexity of a DNA sequence as it contains the frequencies all k -mers of length up to a certain length k . An example of FCGR plot can be found in Figure 2.2.

CGR has been used extensively in literature as the main or as a complementary tool for analyzing DNA sequences. Initially, CGR was used to build phylogenetic trees for various datasets, using commonly-used distance measures. Notable examples of these are the use of nDNA fragments from various domains [38], 125 nDNA fragments from several bird genomes [45], 27 genomes from various genera [37], 26 mtDNA sequences [176], 4 bacteria and about 200 phages [36], 75 HIV-1 genomes [130], 10 mtDNA sequences and 14 nDNA sequences from plants in the *Brassicales* order [62]. In later years, other distances have been

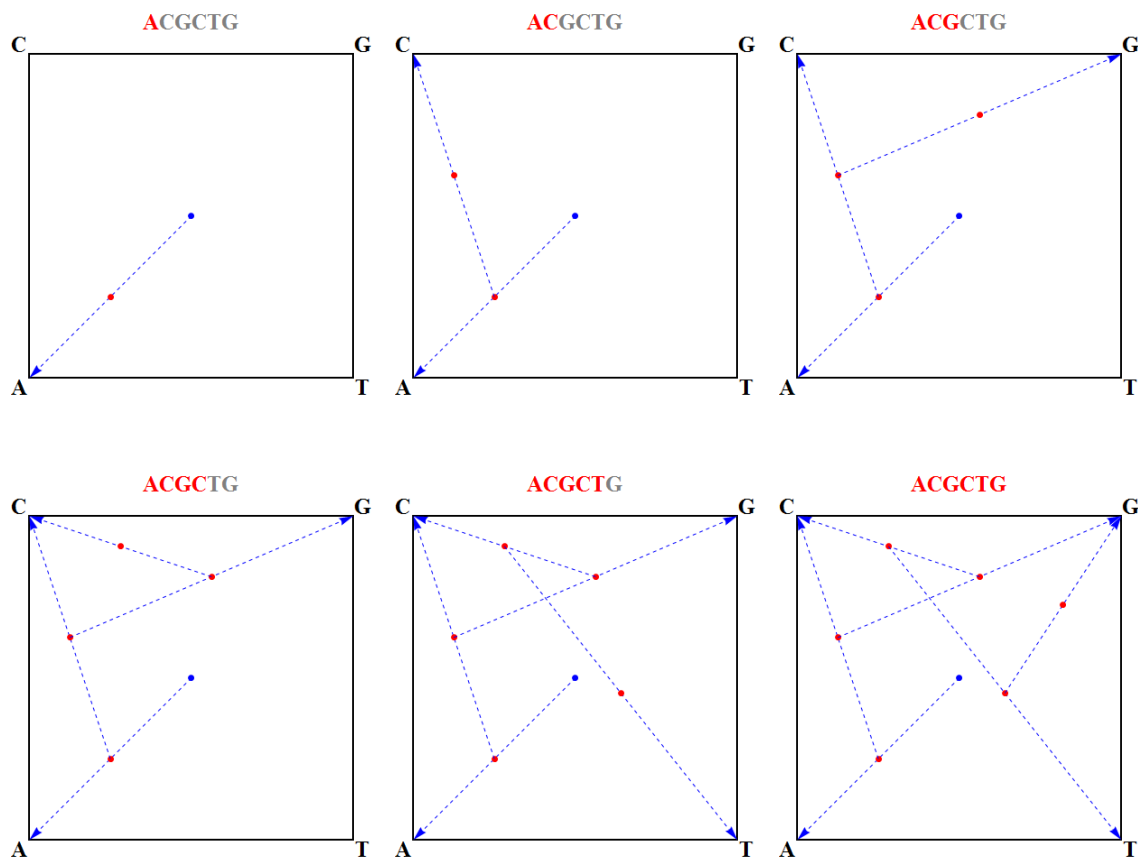
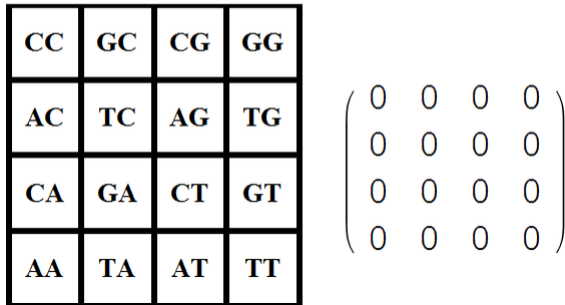
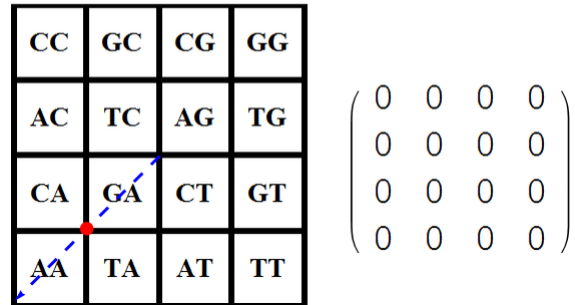


Figure 2.1: The Chaos Game Representation (CGR) of the DNA sequence ACGCTG.

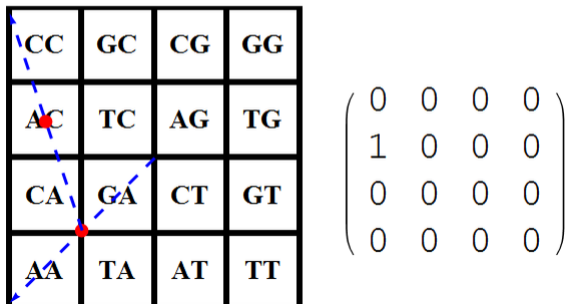
ACGCTGC...



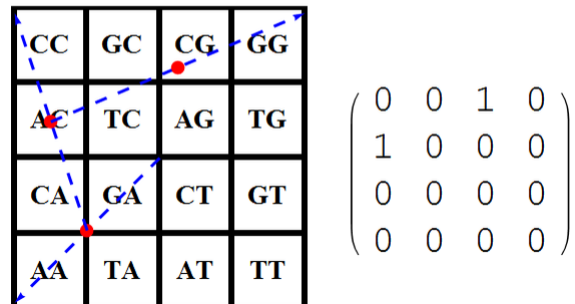
ACGCTGC...



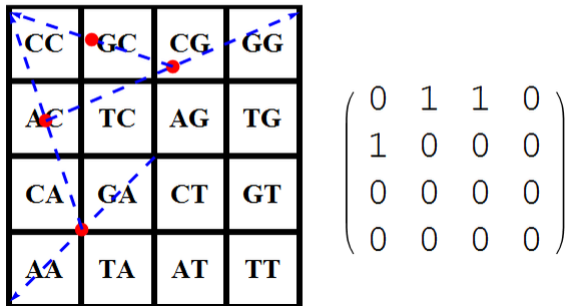
ACGCTGC...



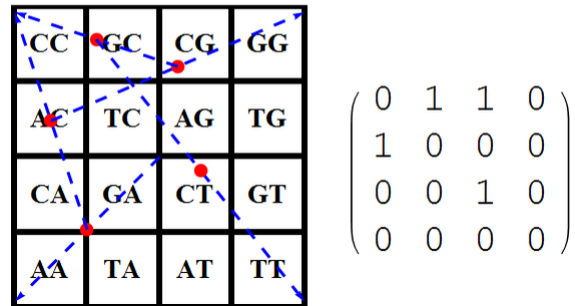
ACGCTGC...



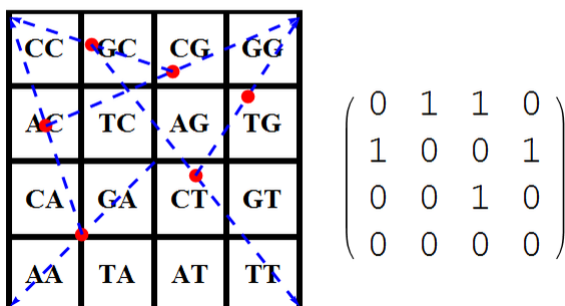
ACGCTGC...



ACGCTGC...



ACGCTGC...



ACGCTGC...

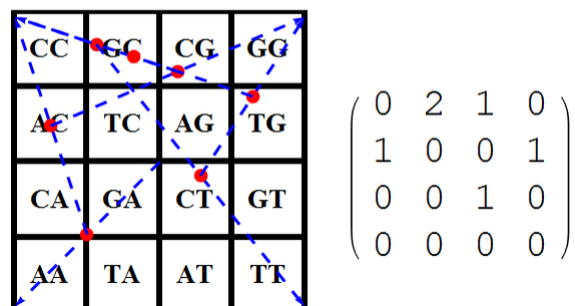


Figure 2.2: The Frequency Chaos Game Representation (FCGR) of the sequence ACGCTGC, for $k = 2$. The first $k - 1$ (here $2 - 1 = 1$) points do not alter the FCGR matrix.

used, some of which proving to perform better than those typically used. Pearson distance, along with a custom image distance have been used in [176] for 26 mtDNA sequences, DSSIM image distance has been used for over 3,000 mtDNA sequences in [88], and six different distances have been used in [87] for nDNA fragments from organisms of all major kingdoms of life.

Other uses of CGRs have been reported in the literature as well. Deschavanne *et al.* in [40] used CGRs to classify functional families of proteins using reverse encoding of amino acids into nucleic sequences. An extension of CGR called Universal Sequence Map (USM) has been reported in [7, 8], which can be used for any size of alphabet. A three-dimensional CGR has been proposed in [163]. CGRs have also been used in studies to measure the degree of self-similarity within images (multifractal analysis) e.g., [196, 178, 50, 168, 122, 129, 128], to estimate sequence entropy [126, 170, 171], to detect horizontal transfers in prokaryotes in [43], to speed up local-alignment algorithms [85], to classify HPV genomes by genotype (together with Neural Networks) in [161], to propose a Recurrent Iterated Function Systems (RIFS) model tested in 50 eukaryotes in [200], to construct CGRs of multiple resolutions (in combination with Neural Networks) in [77] and to refactor foundational string problems using CGR-based algorithms in [172]. Protein sequence analysis using modified CGR and physico-chemical properties has been studied in [187].

2.2.2 Distance Measures

In this thesis, there are six distance measures being used and here we give the definition and a short description for each one of them. All of the distances are being applied to CGRs/FCGRs, that is, to $2^k \times 2^k$ matrices with non-negative integer entries. Let $X = [x_{ij}], Y = [y_{ij}]$ with $i, j = 1, 2, \dots, 2^k$ be matrices with non-negative integer values, that is $X, Y \in \mathbb{Z}_{\geq 0}^{2^k \times 2^k}$. In order to compute the Euclidean, Manhattan and Pearson distances, we first convert the matrices $X, Y \in \mathbb{Z}_{\geq 0}^{2^k \times 2^k}$ into 1×4^k vectors. Now, for two vectors $x, y \in \mathbb{R}^n$, their Euclidean distance $d_E(x, y)$ and their Manhattan distance $d_M(x, y)$ are computed as

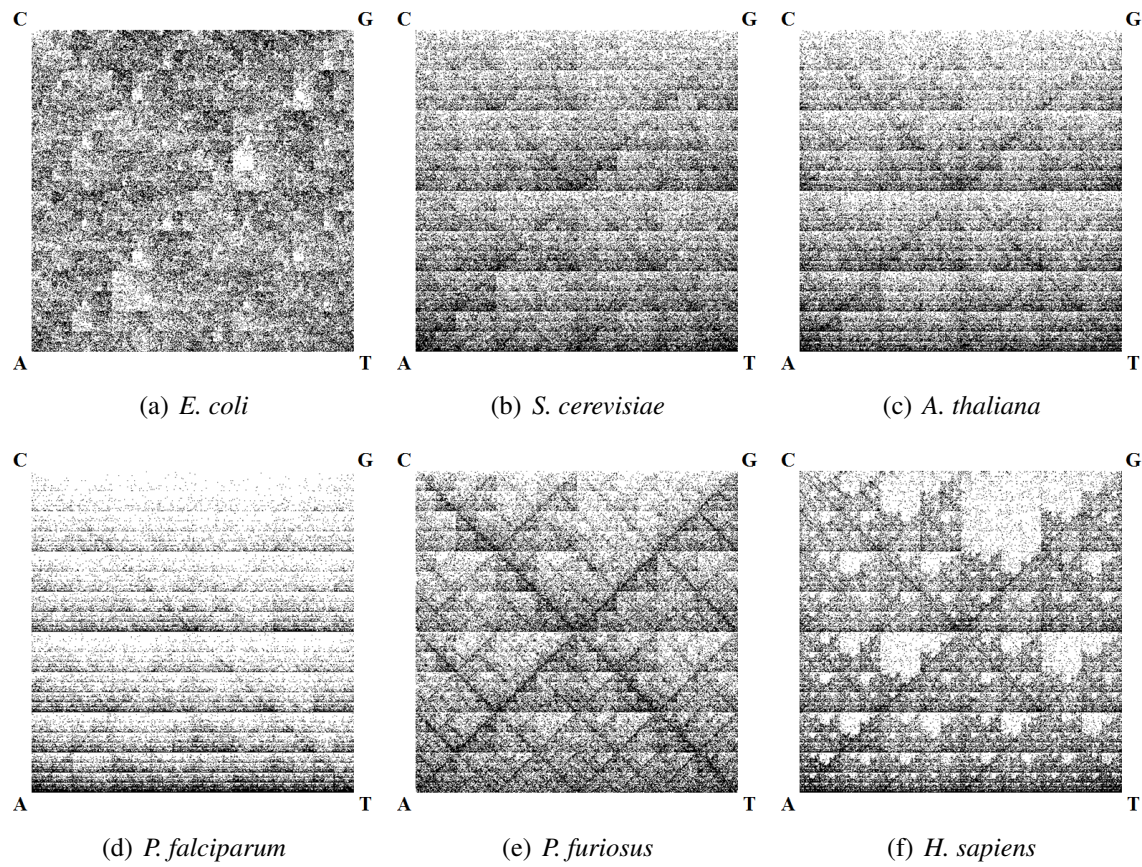


Figure 2.3: Chaos Game Representation (CGR) of nDNA fragments from *E. coli*, *S. cerevisiae*, *A. thaliana*, *P. falciparum*, *P. furiosus* and *H. sapiens*.

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

$$d_M(x, y) = \sum_{i=1}^n |x_i - y_i|,$$

while their Pearson distance $d_P(x, y)$ is defined as

$$d_P(x, y) = 1 - \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where

$$\mu_x = \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)^2},$$

$$\sigma_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y).$$

In general, $\frac{\sigma_{xy}}{\sigma_x \sigma_y}$ ranges in the interval $[-1, 1]$, and as a result the Pearson distance ranges in the interval $[0, 2]$. Euclidean and Manhattan distances are metrics (they are non-negative, symmetric and satisfy the triangle inequality). Pearson distance is not a metric.

Also, we define Approximated Information Distance d_{AID} based on the Information Distance defined in [105]. The Normalized Information Distance in [105] was based on the uncomputable notion of Kolmogorov complexity. Using k -mers, the information distance between two strings x, y was defined as

$$d(x, y) = \frac{N_k(x|y) + N_k(y|x)}{N_k(xy)}$$

with

$$N_k(x|y) = N_k(xy) - N_k(x)$$

where $N_k(x)$ is the number of different, possibly overlapping, k -mers that occur in x .

The Approximated Information Distance (AID) for $X, Y \in \mathbb{Z}_{\geq 0}^{2^k \times 2^k}$ is a modification of the

previous distance, and is defined as

$$d_{AID}(X, Y) = 2 - \frac{f(X) + f(Y)}{f(X + Y)}$$

where $f(X) = \text{SumOfElements}(\text{Unitize}(X))$. By unitization of a matrix X we mean that every non-zero entry becomes 1, while zeros remain 0. The reason behind this modification was that we wanted to avoid counting possible “extra” k -mers that are produced by the concatenation of two strings x and y . This way, we are also guaranteed that

$$d_{AID}(X, Y) = d_{AID}(Y, X)$$

and

$$d_{AID}(X, X) = 0$$

which are properties that were not present before. Approximated Information Distance is a metric.

The *descriptor distance* between two FCGRs $X, Y \in \mathbb{Z}_{\geq 0}^{2^k \times 2^k}$ aims to compare properties of the two given FCGRs of different scales. A *descriptor* is a vector characterized by the parameters m which is the size of the non-overlapping windows in which the FCGR is divided, r which is the number of intervals in the analysis, and r intervals which define the numbers of k -mer occurrences that are considered significant.

For given $m \leq k$ and r , and intervals $[a_0, a_1), [a_1, a_2), \dots, [a_{r-1}, a_r)$ such that $\bigcup_{i=0}^{r-1} [a_i, a_{i+1}) = [0, \infty)$ and $[a_i, a_{i+1}) \cap [a_j, a_{j+1}) = \emptyset \forall i, j$ with $i \neq j$, we construct a descriptor in the following way. Firstly, we divide each of the two FCGR matrices X and Y into non-overlapping submatrices of size $2^m \times 2^m$, resulting in 4^{k-m} submatrices X_{ij} and Y_{ij} with $i, j = 1, \dots, 2^{k-m}$, which will be pairwise compared. Secondly, we compute for every X_{ij} a vector $\text{vec}X_{ij} = \frac{1}{(2^m \times 2^m)}(b_1, b_2, \dots, b_r)$ where $b_i = |\{x \in X_{ij} : a_{i-1} \leq x < a_i\}|$. The same procedure is performed for Y_{ij} , resulting in the vector $\text{vec}Y_{ij}$. Thirdly, we append all vectors $\text{vec}X_{ij}$ to form a new

vector $\text{vec}X^{m,r}$ and, using the same order of appending, we append all vectors $\text{vec}Y_{ij}$ to form a new vector $\text{vec}Y^{m,r}$. For these parameter values of m , r and the r chosen intervals, $\text{vec}X^{m,r}$ and $\text{vec}Y^{m,r}$ are the “descriptors” of the FCGR matrices X and Y . Finally, we can combine descriptors $\text{vec}X^{m,r}$ (respectively $\text{vec}Y^{m,r}$) for several values of m and r by appending them one after another, in the same order, to obtain the vector $\text{vec}X$ (respectively $\text{vec}Y$). The *descriptor distance* between the two FCGRs X and Y is now defined as the Euclidean distance (and as a result is a metric) between the vectors $\text{vec}X$ and $\text{vec}Y$

$$d_D(X, Y) = d_E(\text{vec}X, \text{vec}Y).$$

Finally, another distance measure being used in this thesis is derived from the Structural Similarity Index, SSIM, which was introduced in [177] for the purpose of assessing the degree of similarity between two images. Given two images X, Y as $n \times n$ matrices, SSIM computes the luminance, contrast and structure of these images and combines them to obtain a similarity value. However, instead of computing a global similarity between the two images, each image is divided into 11×11 sliding square windows which move pixel by pixel to eventually cover the entire image. The SSIM similarity of any given pair of images is then computed by comparing their corresponding square windows. In theory, SSIM values range in the interval $[-1, 1]$ with the similarity being 1 between two identical images, 0, for example, between a black image and a white image, and -1 if the two images are negatively correlated. To compute the distance rather than the similarity between two images, we calculate $\text{DSSIM}(X, Y) = 1 - \text{SSIM}(X, Y)$ and therefore the range of DSSIM is the interval $[0, 2]$. DSSIM is not a metric since it does not satisfy the triangle inequality. An example of the SSIM similarity measure being applied to a set of 6 images is depicted in Figure 2.4.

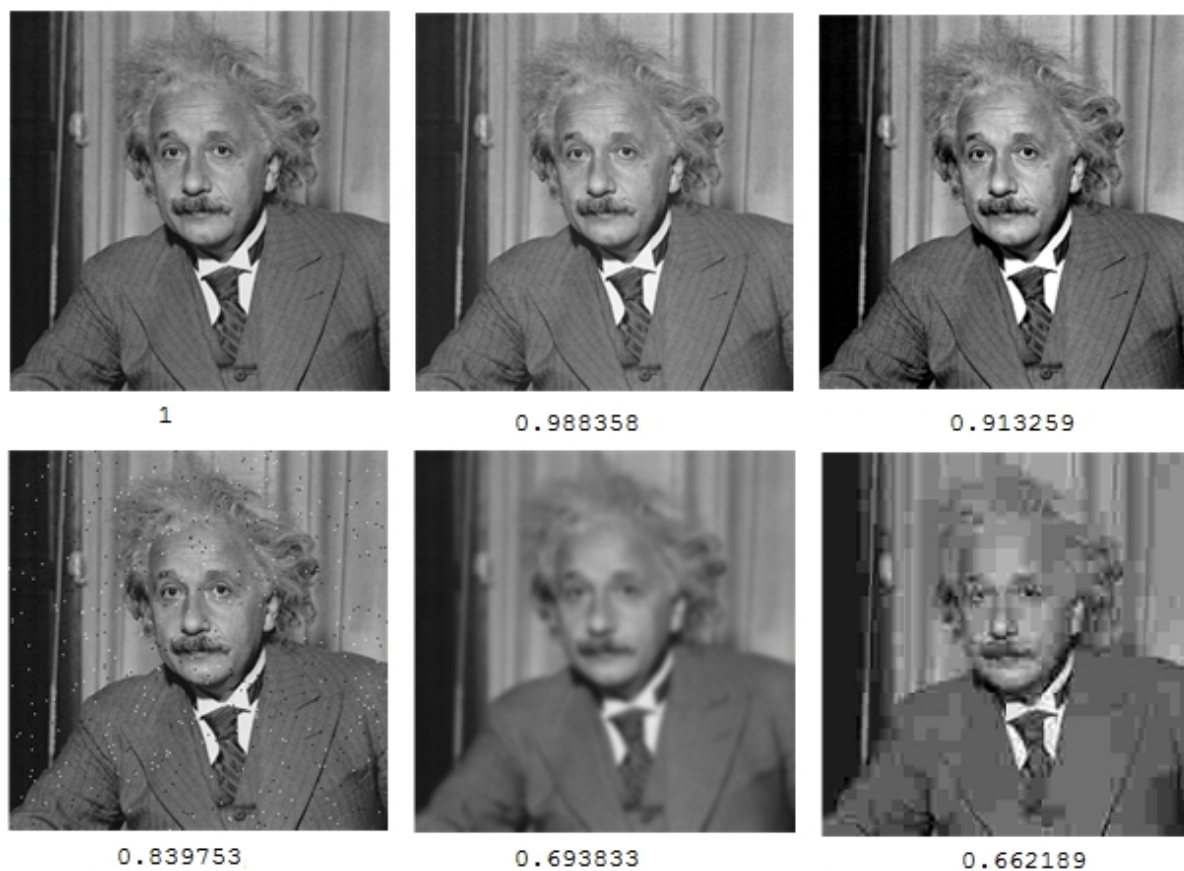


Figure 2.4: An example of computing SSIM similarity values for a set of 6 images. Upper left corner is the original image, hence similarity value is equal to 1. The rest of the images have various pixel values changed, blurred and distorted. The similarity between these images and the original image decreases. This example is part of the examples demonstrated in <https://ece.uwaterloo.ca/~z70wang/research/ssim/>

2.2.3 Multi-Dimensional Scaling (MDS)

Multi-Dimensional Scaling (MDS) is a statistical method [98] that has been used to visualize the degree of similarity between individual objects in a given dataset. MDS has been used extensively in various fields such as cognitive science, information science, psychometrics, marketing, ecology, social science, and other areas of study [22]. Applications of MDS to molecular biology studies can be found in [103] where it was used for the analysis of geographic genetic distributions of some natural populations, in [67] where it was used to visualize distances among COI genes from various species, and in [71] where it was used to analyze and visualize relationships within collections of phylogenetic trees.

Given two points a, b in the r -dimensional Euclidean space we can directly compute their Euclidean distance as $d(a, b) = \sqrt{\sum_{i=1}^r (a_i - b_i)^2}$. MDS tries to solve the inverse problem. Given all the pairwise distances d_{ij} ($i, j = 1, \dots, n$) between n objects, it tries to determine a set of vectors (that is, a set of points in the r -dimensional space) that have these distances as their distances. MDS finds a set of points in the r -dimensional Euclidean space such that the Euclidean distance between two objects is similar to the distances between the corresponding objects in the input distance matrix d_{ij} .

More precisely, classical MDS, 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. MDS will return n points $p_1, p_2, \dots, p_n \in \mathbb{R}^r$ such that $d(i, j) = \|p_i - p_j\| \approx f(\Delta(i, j))$ for all $i, j \in \{1, \dots, n\}$ where $d(i, j)$ is the spatial distance between the points p_i and p_j , and f is a function linear in $\Delta(i, j)$. The value of r can be at most $n - 1$, while in most of the cases the value of either $r = 2$ or $r = 3$ is used to produce a visualization in 2D or 3D space, respectively.

A relatively simple example of MDS can be seen in Figure 2.5. The red points represent 10 big cities in North America. A 10×10 distance matrix $D = [d_{ij}]$ ($i, j = 1, \dots, 10$) was formed where the d_{ij} element of this matrix was the distance in kilometers between the i -th and j -th city. $D = [d_{ij}]$ was the input to MDS, and the output (taken for $r = 2$ dimensions) is the set of blue points. We can see that blue points, although not identically placed as the red

points, approximate fairly well the topology of the red points. Some misplacement of points is expected because of measurement errors, mainly because roads connecting cities are not straight lines and because the Earth is not flat.

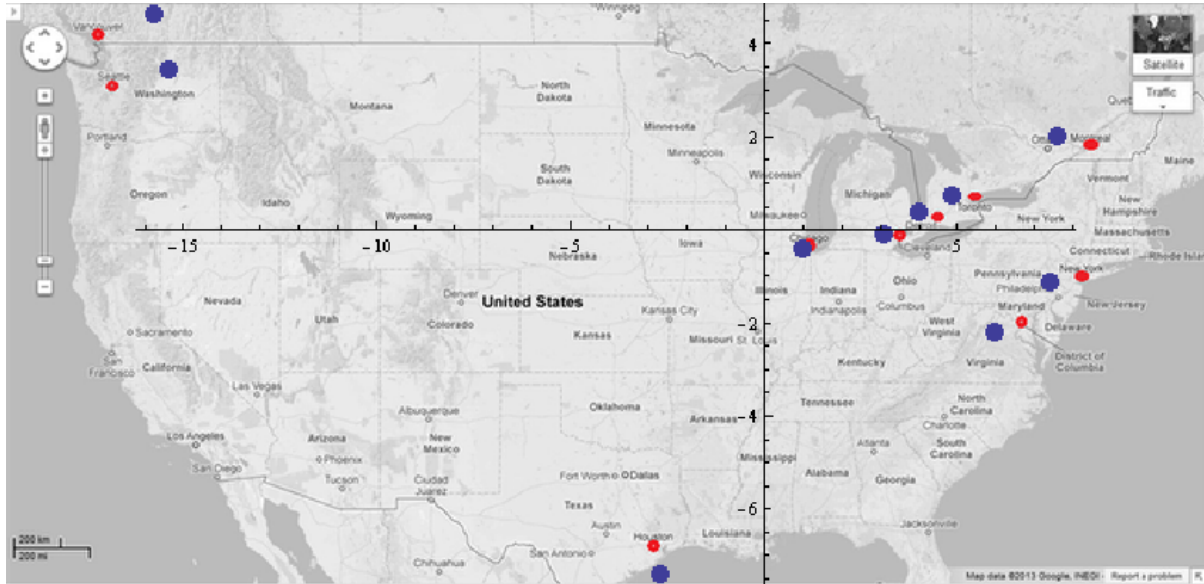


Figure 2.5: Multi-Dimensional Scaling (MDS) example. The red points are the real positions of 10 big cities in North America. The blue points are the positions of these cities as output of MDS.

Bibliography

- [1] T. Abe, Y. Hamano, and T. Ikemura. Visualization of genome signatures of eukaryote genomes by batch-learning self-organizing map with a special emphasis on drosophila genomes. *BioMed Research International*, 2014, 2014.
- [2] T. Abe, H. Sugawara, S. Kanaya, and T. Ikemura. A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of uncultured environmental microbes. *Polar Research*, 20:103–112, 2006.
- [3] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Research*, 12(5):281–290, 2005.
- [4] A. Albayrak, H. Otu, and U. Sezerman. Clustering of protein families into functional subtypes using Relative Complexity Measure with reduced amino acid alphabets. *BMC Bioinformatics*, 11:428, 2010.
- [5] J. Almeida. Sequence analysis by iterated maps, a review. *Briefings in Bioinformatics*, 15(3):369–375, 2014.
- [6] J. Almeida, J. Carrio, A. Marezek, P. Noble, and M. Fletcher. Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437, 2001.
- [7] J. Almeida and S. Vinga. Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, 3:6, 2002.
- [8] J. Almeida and S. Vinga. Computing distribution of scale independent motifs in biological sequences. *Algorithms for Molecular Biology*, 1:18, 2006.
- [9] A. Arneodo, Y. d'Aubenton Carafa, E. Bacry, P. Graves, J. Muzy, and C. Thermes. Wavelet based fractal analysis of DNA sequences. *Physica D: Nonlinear Phenomena*, 96(1):291–320, 1996.

- [10] J. Arnold, A. Cuticchia, D. Newsome, W. Jennings, and R. Ivarie. Mono- through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. *Nucleic Acids Research*, 16(14):7145–7158, 1988.
- [11] Y. Bakış, H. Otu, N. Taşçı, C. Meydan, N. Bilgin, S. Yüzbaşıoğlu, and O. Sezerman. Testing robustness of relative complexity measure method constructing robust phylogenetic trees for *Galanthus L.* using the relative complexity measure. *BMC Bioinformatics*, 14(1):1–12, 2013.
- [12] D. Bastola, H. Otu, S. Doukas, K. Sayood, S. Hinrichs, and P. Iwen. Utilization of the relative complexity measure to construct a phylogenetic tree for fungi. *Mycological Research*, 108(Pt 2):117–125, 2004.
- [13] M. Bauer, S. Schuster, and K. Sayood. The average mutual information profile as a genomic signature. *BMC Bioinformatics*, 9:48, 2008.
- [14] E. Behnam, M. Waterman, and A. Smith. A geometric interpretation for local alignment-free sequence comparison. *Journal of Computational Biology*, 20(7):471–485, 2013.
- [15] E. Beutler, T. Gelbart, J. Han, J. Koziol, and B. Beutler. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proceedings of the National Academy of Sciences of the United States of America*, 86(1):192–196, 1989.
- [16] A. Bhagwat and M. McClelland. DNA mismatch correction by very short patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. *Nucleic Acids Research*, 20(7):1663–1668, 1992.
- [17] B. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 83(14):5155–5159, 1986.

- [18] B. Blaisdell. Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *Journal of Molecular Evolution*, 29(6):526–537, 1989.
- [19] J. Bohlin. Genomic signatures in microbes – properties and applications. *The Scientific World Journal*, 11:715–725, 2011.
- [20] J. Bohlin and E. Skjerve. Examination of genome homogeneity in prokaryotes using genomic signatures. *PLoS ONE*, 4(12), 2009.
- [21] O. Bonham-Carter, J. Steele, and D. Bastola. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in Bioinformatics*, page bbt052, 2013.
- [22] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2nd edition, 2010.
- [23] E. Borrayo, E. Mendizabal-Ruiz, H. Vélez-Pérez, R. Romo-Vázquez, A. Mendizabal, and J. Morales. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. *PLoS ONE*, 9(11), 2014.
- [24] C. Burden, S. Forêt, and S. Wilson. k-Word matches: an alignment-free sequence comparison method. *Supplementary Conference Proceedings PRIB 2008*, pages 235–238, 2008.
- [25] C. Burge, A. Campbell, and S. Karlin. Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 89(4):1358–1362, 1992.
- [26] A. Campbell, J. Mrázek, and S. Karlin. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16):9184–9189, 1999.

- [27] C. Chapus, C. Dufraigne, S. Edwards, A. Giron, B. Fertil, and P. Deschavanne. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evolutionary Biology*, 5(1):1–18, 2005.
- [28] R. Chi and K. Ding. Novel 4D numerical representation of DNA sequences. *Chemical Physics Letters*, 407(1-3):63–67, 2005.
- [29] K. Chu, J. Qi, Z.-G. Yu, and V. Anh. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Molecular Biology and Evolution*, 21(1):200–206, 2004.
- [30] G. Churchill. Hidden Markov chains and the analysis of genome structure. *Computers & Chemistry*, 16(2):107–115, 1992.
- [31] M. Comin and D. Verzotto. Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms for Molecular Biology : AMB*, 7(1):34, 2012.
- [32] Q. Dai, X. Liu, Y. Yao, and F. Zhao. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. *Journal of Theoretical Biology*, 276(1):174–180, 2011.
- [33] Q. Dai and T. Wang. Comparison study on k-word statistical measures for protein: from sequence to 'sequence space'. *BMC Bioinformatics*, 9:394, 2008.
- [34] Q. Dai, Y. Yang, and T. Wang. Markov model plus k-word distributions: A synergy that produces novel statistical measures for sequence comparison. *Bioinformatics*, 24(20):2296–2302, 2008.
- [35] M. Deng, C. Yu, Q. Liang, R. He, and S. Yau. A novel method of characterizing genetic sequences: Genome space with biological distance and applications. *PLoS ONE*, 6(3), 2011.

- [36] P. Deschavanne, M. DuBow, and C. Regeard. The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virology Journal*, 7:163, 2010.
- [37] P. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, and B. Fertil. Genomic signature is preserved in short DNA fragments. *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 161–167, 2000.
- [38] 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.
- [39] P. Deschavanne and M. Radman. Counterselection of GATC sequences in enterobacteriophages by the components of the methyl-directed mismatch repair system. *Journal of Molecular Evolution*, 33(2):125–132, 1991.
- [40] P. Deschavanne and P. Tuffery. Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie*, 90(4):615–625, 2008.
- [41] G. Didier, E. Corel, I. Laprevotte, A. Grossmann, and C. Landès-Devauchelle. Variable length local decoding and alignment-free sequence comparison. *Theoretical Computer Science*, 462:1–11, 2012.
- [42] M. Domazet-Lošo Mirjana and B. Haubold. Efficient estimation of pairwise distances between genomes. *Bioinformatics*, 25(24):3221–3227, 2009.
- [43] C. Dufraigne, B. Fertil, S. Lespinats, A. Giron, and P. Deschavanne. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research*, 33(1):e6, 2005.
- [44] C. Dutta and J. Das. Mathematical characterization of Chaos Game Representation. New

- algorithms for nucleotide sequence analysis. *Journal of Molecular Biology*, 228(3):715–719, 1992.
- [45] S. Edwards, B. Fertil, A. Giron, and P. Deschavanne. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic Biology*, 51(4):599–613, 2002.
- [46] J. Feng, Y. Hu, P. Wan, A. Zhang, and W. Zhao. New method for comparing DNA primary sequences based on a discrimination measure. *Journal of Theoretical Biology*, 266(4):703–707, 2010.
- [47] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente. Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, 8:252, 2007.
- [48] S. Forêt, S. Wilson, and C. Burden. Characterizing the D2 statistic: word matches in biological sequences. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 43, 2009.
- [49] S. Forêt, S. Wilson, and C. Burden. Empirical distribution of k-word matches in biological sequences. *Pattern Recognition*, 42(4):539–548, 2009.
- [50] W. Fu, Y. Wang, and D. Lu. Multifractal analysis of genomes sequences' CGR graph. *Journal of Biomedical Engineering*, 24(3):522–525, 2007.
- [51] L. Gao and J. Qi. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evolutionary Biology*, 7(1):41, 2007.
- [52] Y. Gao and L. Luo. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene*, 492(1):309–314, 2012.
- [53] S. Garte. Fractal properties of the human genome. *Journal of Theoretical Biology*, 230(2):251–260, 2004.

- [54] M. Gates. A simple way to look at DNA. *Journal of Theoretical Biology*, 119(3):319–328, 1986.
- [55] M. Gelfand and E. Koonin. Avoidance of palindromic words in bacterial and archaeal genomes: A close connection with restriction enzymes. *Nucleic Acids Research*, 25(12):2430–2439, 1997.
- [56] A. Gentles and S. Karlin. Genome-scale compositional comparisons in eukaryotes. *Genome Research*, 11(4):540–546, 2001.
- [57] 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.
- [58] F. Guyon, C. Brochier-Armanet, and A. Guénoche. Comparison of alignment free string distances for complete genome phylogeny. *Advances in Data Analysis and Classification*, 3(2):95–108, 2009.
- [59] F. Guyon and A. Guénoche. An evolutionary distance based on maximal unique matches. *Communications in Statistics Theory and Methods*, 39(3):385–397, 2010.
- [60] 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.
- [61] B. Hao and J. Qi. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *Journal of Bioinformatics and Computational Biology*, 02(01):1–19, 2004.
- [62] K. Hatje and M. Kollmar. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Frontiers in Plant Science*, 3, 2012.
- [63] B. Haubold. Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics*, 15(3):407–18, 2014.

- [64] B. Haubold and P. Pfaffelhuber. Alignment-free population genomics: An efficient estimator of sequence diversity. *G3: Genes—Genomes—Genetics*, 2(8):883–889, 2012.
- [65] B. Haubold, P. Pfaffelhuber, M. Domazet-Loso, and T. Wiehe. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, 16(10):1487–1500, 2009.
- [66] B. Haubold, F. Reed, and P. Pfaffelhuber. Alignment-free estimation of nucleotide diversity. *Bioinformatics*, 27(4):449–455, 2011.
- [67] P. Hebert, A. Cywinska, S. Ball, and J. deWaard. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512):313–321, 2003.
- [68] W. Hide, J. Burke, and D. Davison. Biological evaluation of D2, an algorithm for high-performance sequence comparison. *Journal of Computational Biology*, 1(3):199–215, 1994.
- [69] 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. *Journal of Molecular Evolution*, 35(3):261–269, 1992.
- [70] K. Hill and S. Singh. The evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes. *Genome*, 40(3):342–356, 1997.
- [71] D. Hillis, T. Heath, and K. John. Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482, 2005.
- [72] M. Höhl and M. Ragan. Is multiple-sequence alignment required for accurate inference of phylogeny? *Systematic Biology*, 56(2):206–221, 2007.
- [73] M. Höhl, I. Rigoutsos, and M. Ragan. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evolutionary Bioinformatics Online*, 2(2003):359–375, 2006.

- [74] S. Horwege, S. Lindner, M. Boden, K. Hatje, M. Kollmar, C. Leimeister, and B. Morgenstern. Spaced words and kmacs: Fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Research*, 42(W1), 2014.
- [75] G. Huang, B. Liao, Y. Li, and Y. Yu. Similarity studies of DNA sequences based on a new 2D graphical representation. *Biophysical Chemistry*, 143(1-2):55–59, 2009.
- [76] G. Huang, H. Zhou, Y. Li, and L. Xu. Alignment-free comparison of genome sequences by a new numerical characterization. *Journal of Theoretical Biology*, 281(1):107–112, 2011.
- [77] X. Huang, D.-S. Huang, H.-Q. Wang, and X.-M. Zhao. Representation of DNA sequences with multiple resolutions and BP neural network based classification. In *Neural Networks, 2004*, volume 2, pages 1185–1189. IEEE, 2004.
- [78] Y. Huang, L. Yang, and T. Wang. Phylogenetic analysis of DNA sequences based on the generalized pseudo-amino acid composition. *Journal of Theoretical Biology*, 269(1):217–223, 2011.
- [79] K. Jaron, J. Moravec, and N. Martínková. Sighunt: horizontal gene transfer finder optimized for eukaryotic genomes. *Bioinformatics*, 30(8):1081–1086, 2014.
- [80] R. Jayalakshmi, R. Natarajan, M. Vivekanandan, and G. Natarajan. Alignment-free sequence comparison using N-dimensional similarity space. *Current Computer-Aided Drug Design*, 6(4):290–296, 2010.
- [81] H. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.
- [82] H. Jeffrey. Chaos game visualization of sequences. *Computers & Graphics*, 16(1):25–33, 1992.

- [83] R. Jernigan and R. Baran. Pervasive properties of the genomic signature. *BMC Genomics*, 3(1):23, 2002.
- [84] J. Jing, C. Burden, S. Forêt, and S. Wilson. Statistical considerations underpinning an alignment-free sequence comparison method. *Journal of the Korean Statistical Society*, 39(3):325–335, 2010.
- [85] J. Joseph and R. Sasikumar. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, 7:243, 2006.
- [86] M. Kantorovitz, G. Robinson, and S. Sinha. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249–i255, 2007.
- [87] R. Karamichalis, L. Kari, S. Konstantinidis, and S. Kopecki. An investigation into inter- and intragenomic variations of graphic genomic signatures. *BMC Bioinformatics*, 16:246, 2015.
- [88] L. Kari, K. Hill, A. Sayem, R. Karamichalis, N. Bryans, K. Davis, and N. Dattani. Mapping the space of genomic signatures. *PLoS ONE*, 10(5):e0119815, 2015.
- [89] S. Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Current Opinion in Microbiology*, 1(5):598–610, 1998.
- [90] S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, 11(7):283–290, 1995.
- [91] S. Karlin, C. Burge, and A. Campbell. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Research*, 20(6):1363–1370, 1992.
- [92] S. Karlin, A. Campbell, and J. Mrázek. Comparative DNA analysis across diverse genomes. *Annual Review of Genetics*, 32:185–225, 1998.

- [93] S. Karlin and I. Ladunga. Comparisons of eukaryotic genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12832–12836, 1994.
- [94] S. Karlin, I. Ladunga, and B. Blaisdell. Heterogeneity of genomes: measures and values. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12837–12841, 1994.
- [95] S. Karlin, J. Mrzek, and A. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, 179(12):3899–3913, 1997.
- [96] P. Kolekar, M. Kale, and U. Kulkarni-Kale. Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping. *Molecular Phylogenetics and Evolution*, 65(2):510–522, 2012.
- [97] N. Krasnogor and D. Pelta. Measuring the similarity of protein structures by means of the Universal Similarity Metric. *Bioinformatics*, 20(7):1015–1021, 2004.
- [98] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.
- [99] B. Kumar, R. Alok, B. Sk, and D. Jayant. Genome analysis: A new approach for visualization of sequence organization in genomes. *Journal of Biosciences*, 17(4):395–411, 1992.
- [100] C. Leimeister, M. Boden, S. Horwege, S. Lindner, and B. Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30(14):1991–1999, 2014.
- [101] C. Leimeister and B. Morgenstern. Kmacs: The k-mismatch average common substring

- approach to alignment-free sequence comparison. *Bioinformatics*, 30(14):2000–2008, 2014.
- [102] P. Leong and S. Morgenthaler. Random walk and gap plots of DNA sequences. *Computer Applications in the Biosciences*, 11(5):503–507, 1995.
- [103] E. Lessa. Multidimensional analysis of geographic genetic structure. *Systematic Zoology*, 39(3):242–252, 1990.
- [104] B. Li, Y. Li, and H. He. LZ complexity distance of DNA sequences and its application in phylogenetic tree reconstruction. *Genomics, Proteomics & Bioinformatics*, 3(4):206–212, 2005.
- [105] M. Li, X. Chen, X. Li, B. Ma, and P. Vitany. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.
- [106] B. Liao. A 2D graphical representation of DNA sequence. *Chemical Physics Letters*, 401(1-3):196–199, 2005.
- [107] B. Liao, R. Li, W. Zhu, and X. Xiang. On the similarity of DNA primary sequences based on 5-D representation. *Journal of Mathematical Chemistry*, 42(1):47–57, 2007.
- [108] B. Liao, M. Tan, and K. Ding. A 4D representation of DNA sequences and its application. *Chemical Physics Letters*, 402(4-6):380–383, 2005.
- [109] B. Liao, M. Tan, and K. Ding. Application of 2-D graphical representation of DNA sequence. *Chemical Physics Letters*, 414(4-6):296–300, 2005.
- [110] B. Liao and T. Wang. 3-D graphical representation of DNA sequences and their numerical characterization. *Journal of Molecular Structure*, 681(1-3):209–212, 2004.
- [111] B. Liao and T. Wang. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. *Journal of Chemical Information and Computer Sciences*, 44(5):1666–1670, 2004.

- [112] B. Liao and T.-M. Wang. New 2D graphical representation of DNA sequences. *Journal of Computational Chemistry*, 25(11):1364–1368, 2004.
- [113] R. Lippert, H. Huang, and M. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):13980–13989, 2002.
- [114] X. Liu and Y. Li. Some notes on 2-D graphical representation of DNA sequence. In *Proceedings of the 27th Chinese Control Conference, CCC*, pages 303–305, 2008.
- [115] X. Liu, L. Wan, J. Li, G. Reinert, M. Waterman, and F. Sun. New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *Journal of Theoretical Biology*, 284(1):106–116, 2011.
- [116] Z.-H. Liu, H.-D. Liu, J.-R. Li, X. Sun, and D. Jiao. Base-base correlation a novel sequence feature and its applications. In *1st International Conference on Bioinformatics and Biomedical Engineering*, pages 370–373, July 2007.
- [117] Z.-H. Liu and X. Sun. Coronavirus phylogeny based on base-base correlation. *International Journal of Bioinformatics Research and Applications*, 4(2):211–220, 2008.
- [118] J. Lobry. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, 78(5):323–326, 1996.
- [119] G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.
- [120] C. Martin, N. Diaz, J. Ontrup, and T. Nattkemper. Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. *Bioinformatics*, 24(14):1568–1574, 2008.
- [121] E. Mizraji and J. Ninio. Graphical coding of nucleic acid sequences. *Biochimie*, 67(5):445–448, 1985.

- [122] P. Moreno, P. Vélez, E. Martínez, L. Garreta, N. Díaz, S. Amador, I. Tischer, J. Gutiérrez, A. Naik, F. Tobar, and F. García. The human genome: a multifractal analysis. *BMC Genomics*, 12(1):506, 2011.
- [123] O. Nalbantoglu and K. Sayood. Computational genomic signatures. *Synthesis Lectures on Biomedical Engineering*, 6(2):1–129, 2011.
- [124] A. Nandy. A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes. *Current Science*, 66(4):309–314, 1994.
- [125] A. Nandy, M. Harle, and S. Basak. Mathematical descriptors of DNA sequences: development and applications. *Arkivoc*, 2006(9), 2006.
- [126] J. Oliver, P. Bernaola-Galván, J. Guerrero-García, and R. Román-Roldán. Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretical Biology*, 160(4):457–470, 1993.
- [127] H. Otu and K. Sayood. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(16):2122–2130, 2003.
- [128] M. Pal, B. Satish, K. Srinivas, P. Rao, and P. Manimaran. Multifractal detrended cross-correlation analysis of coding and non-coding DNA sequences through chaos-game representation. *Physica A: Statistical Mechanics and its Applications*, 436:596–603, 2015.
- [129] A. Pandit, A. Dasanna, and S. Sinha. Multifractal analysis of HIV-1 genomes. *Molecular Phylogenetics and Evolution*, 62(2):756–763, 2012.
- [130] A. Pandit and S. Sinha. Using genomic signatures for HIV-1 sub-typing. *BMC Bioinformatics*, 11 Suppl 1:S26, 2010.
- [131] A. Pandit, J. Vadlamudi, and S. Sinha. Analysis of dinucleotide signatures in HIV-1 subtype B genomes. *Journal of Genetics*, 92(3):403–412, 2013.

- [132] K. Patil and A. McHardy. Alignment-free genome tree inference by learning group-specific distance metrics. *Genome Biology and Evolution*, 5(8):1470–1484, 2013.
- [133] T. Pham. Spectral distortion measures for biological sequence comparisons and database searching. *Pattern Recognition*, 40(2):516–529, 2007.
- [134] T. Pham and J. Zuegg. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics*, 20(18):3455–3461, 2004.
- [135] G. Phillips, J. Arnold, and R. Ivarie. Mono-through hexanucleotide composition of the *Escherichia coli* genome: A Markov chain analysis. *Nucleic Acids Research*, 15(6):2611–2626, 1987.
- [136] D. Pride, R. Meinersmann, T. Wassenaar, and M. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*, 13(2):145–156, 2003.
- [137] J. Qi, H Luo, and B. Hao. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, 32(suppl 2):W45–W47, 2004.
- [138] J. Qi, B. Wang, and B. Hao. Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach. *Journal of Molecular Evolution*, 58(1):1–11, 2004.
- [139] X. Qi, Q. Wu, Y. Zhang, E. Fuller, and C.-Q. Zhang. A novel model for DNA sequence similarity analysis based on graph theory. *Evolutionary Bioinformatics online*, 7:149–58, 2011.
- [140] X.-Q. Qi, J. Wen, and Z.-H. Qi. New 3D graphical representation of DNA sequence based on dual nucleotides. *Journal of Theoretical Biology*, 249(4):681–690, 2007.
- [141] 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.

- [142] Z. Qi and X. Qi. Novel 2D graphical representation of DNA sequence based on dual nucleotides. *Chemical Physics Letters*, 440(1-3):139–144, 2007.
- [143] M. Randić. Graphical representations of DNA as 2-D map. *Chemical Physics Letters*, 386(4-6):468–471, 2004.
- [144] M. Randić, N. Lerš, D. Plavšić, S. Basak, and A. Balaban. Four-color map representation of DNA or RNA sequences and their numerical characterization. *Chemical Physics Letters*, 407(1-3):205–208, 2005.
- [145] M. Randić, M. Vračko, N. Lerš, and D. Plavšić. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chemical Physics Letters*, 371(1-2):202–207, 2003.
- [146] G. Reinert, D. Chew, F. Sun, and M. Waterman. Alignment-free sequence comparison (I): statistics and power. *Journal of Computational Biology*, 16(12):1615–1634, 2009.
- [147] A. Roy, C. Raychaudhury, and A. Nandy. Novel techniques of graphical representation and analysis of DNA sequences: A review. *Journal of Biosciences*, 23(1):55–71, 1998.
- [148] R. Sandberg, G. Winberg, C. Bränden, A. Kaske, I. Ernberg, and J. Cöster. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Research*, 11(8):1404–1409, 2001.
- [149] F. Sanger, S. Nicklen, and A. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, 1977.
- [150] I. Schwende and T. Pham. Pattern recognition and probabilistic measures in alignment-free sequence analysis. *Briefings in Bioinformatics*, 15(3):354–68, 2014.

- [151] K. Sedlar, H. Skutkova, M. Vitek, and I. Provaznik. Prokaryotic DNA signal downsampling for fast whole genome comparison. In *Information Technologies in Biomedicine, Volume 3*, pages 373–383. Springer, 2014.
- [152] A. Shedlock, C. Botka, S. Zhao, J. Shetty, T. Zhang, J. Liu, P. Deschavanne, and S. Edwards. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8):2767–2772, 2007.
- [153] G. Sims, S.-R. Jun, G. Wu, and S.-H. Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2677–2682, 2009.
- [154] G. Sims and S.-H. Kim. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences of the United States of America*, 108(20):8329–8334, 2011.
- [155] H. Skutkova, M. Vitek, P. Babula, R. Kizek, and I. Provaznik. Classification of genomic signals using dynamic time warping. *BMC Bioinformatics*, 14(Suppl 10):S1, 2013.
- [156] K. Song, J. Ren, G. Reinert, M. Deng, M. Waterman, and F. Sun. New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, 15(3):343–353, 2014.
- [157] K. Song, J. Ren, Z. Zhai, X. Liu, M. Deng, and F. Sun. Alignment-free sequence comparison based on next-generation sequencing reads. *Journal of Computational Biology*, 20(2):64–79, 2013.
- [158] G. Stuart, K. Moffett, and S. Baker. Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, 18(1):100–108, 2002.

- [159] T. Sun, L. Zhang, J. Chen, and Z. Jiang. Statistical properties and fractals of nucleotide clusters in DNA sequences. *Chaos, Solitons & Fractals*, 20(5):1075–1084, 2004.
- [160] H. Suzuki, M. Sota, C. Brown, and E. Top. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Research*, 36(22), 2008.
- [161] W. Tanchotsrinon, C. Lursinsap, and Y. Poovorawan. A high performance prediction of HPV genotypes by chaos game representation and singular value decomposition. *BMC Bioinformatics*, 16(1), 2015.
- [162] X. Tang, P. Zhou, and W. Qiu. On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. *Chinese Science Bulletin*, 55(8):701–704, 2010.
- [163] I. Tavassoly, O. Tavassoly, M. Rad, and N. Dastjerdi. Three dimensional chaos game representation of genomic sequences. In *Proceedings of the Frontiers in the Convergence of Bioscience and Information Technologies, FBIT 2007*, pages 219–223, 2007.
- [164] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6(9):938–947, 2004.
- [165] R. Todeschini, V. Consonni, A. Mauri, and D. Ballabio. Characterization of DNA primary sequences by a new similarity/diversity measure based on the partial ordering. *Journal of Chemical Information and Modeling*, 46(5):1905–1911, 2006.
- [166] I. Ulitsky, D. Burstein, T. Tuller, and B. Chor. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13(2):336–350, 2006.
- [167] A. Van Vliet and J. Kusters. Use of alignment-free phylogenetics for rapid genome

- sequence-based typing of *Helicobacter pylori* virulence markers and antibiotic susceptibility. *Journal of Clinical Microbiology*, 53(9):2877–2888, 2015.
- [168] P. Vélez, L. Garreta, E. Martínez, N. Díaz, S. Amador, I. Tischer, J. Gutiérrez, and P. Moreno. The *Caenorhabditis elegans* genome: A multifractal analysis. *Genetics and Molecular Research*, 9(2):949–965, 2010.
- [169] S. Vinga and J. Almeida. Alignment-free sequence comparison - A review. *Bioinformatics*, 19(4):513–523, 2003.
- [170] S. Vinga and J. Almeida. Rényi continuous entropy of DNA sequences. *Journal of Theoretical Biology*, 231(3):377–388, 2004.
- [171] S. Vinga and J. Almeida. Local Rényi entropic profiles of DNA sequences. *BMC Bioinformatics*, 8:393, 2007.
- [172] S. Vinga, A. Carvalho, A. Francisco, Luís M. Russo, and J. Almeida. Pattern matching through chaos game representation: bridging numerical and discrete data structures for biological sequence analysis. *Algorithms for Molecular Biology*, 7(1):1, 2012.
- [173] L. Wan, G. Reinert, F. Sun, and M. Waterman. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *Journal of Computational Biology*, 17(11):1467–1490, 2010.
- [174] H. Wang, Z. Xu, L. Gao, and B. Hao. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evolutionary Biology*, 9(1):195, 2009.
- [175] W. Wang and T. Wang. Conditional LZ complexity and its application in mtDNA sequence analysis. *MATCH - Communications in Mathematical and in Computer Chemistry*, 66(1):425–443, 2011.

- [176] 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.
- [177] 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.
- [178] F. Weijuan, W. Yuanyuan, and L. Daru. Multifractal analysis of genomic sequences CGR images. In *IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4783–4786, 2005.
- [179] E. Weitschek, F. Cunial, and G. Felici. Laf: Logic alignment free and its application to bacterial genomes classification. *BioData mining*, 8(1):1, 2015.
- [180] J. Wen and Y. Zhang. A 2D graphical representation of protein sequence and its numerical characterization. *Chemical Physics Letters*, 476(4-6):281–286, 2009.
- [181] G. Wu, S.-R. Jun, G. Sims, and S.-H. Kim. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31):12826–12831, 2009.
- [182] T. Wu, J. Burke, and D. Davison. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, 53(4):1431–1439, 1997.
- [183] T. Wu, Y. Hsieh, and L. Li. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, 57(2):441–448, 2001.
- [184] T. Wu, Y. Huang, and L. Li. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics*, 21(22):4125–4132, 2005.

- [185] Y. Wu, A. Liew, H. Yan, and M. Yang. DB-Curve: A novel 2D method of DNA sequence visualization and representation. *Chemical Physics Letters*, 367(1-2):170–176, 2003.
- [186] Y. Xiao, R. Chen, R. Shen, J. Sun, and J. Xu. Fractal dimension of exon and intron sequences. *Journal of Theoretical Biology*, 175(1):23–26, 1995.
- [187] C. Xu, D. Sun, S. Liu, and Y. Zhang. Protein sequence analysis by incorporating modified chaos game and physicochemical properties into Chou's general pseudo amino acid composition. *Journal of Theoretical Biology*, 2016.
- [188] L. Yang, X. Zhang, and H. Zhu. Alignment free comparison: K word voting model and its applications. *Journal of Theoretical Biology*, 335:276–282, 2013.
- [189] X. Yang and T. Wang. A novel statistical measure for sequence comparison on the basis of k-word counts. *Journal of Theoretical Biology*, 318:91–100, 2013.
- [190] 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.
- [191] Y.-H. Yao, X.-Y. Nan, and T. Wang. A new 2D graphical representation-classification curve and the analysis of similarity/dissimilarity of DNA sequences. *Journal of Molecular Structure*, 764(1-3):101–108, 2006.
- [192] S. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, and Y.-K. Ho. DNA sequence representation without degeneracy. *Nucleic Acids Research*, 31(12):3078–3080, 2003.
- [193] S. Yau, C. Yu, and R. He. A protein map and its application. *DNA and Cell Biology*, 27(5):241–250, 2008.
- [194] 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.

- [195] 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.
- [196] Z. Yu, V. Anh, and K. Lau. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *Journal of Theoretical Biology*, 226(3):341–348, 2004.
- [197] Z.-G. Yu, V. Anh, and K.-S. Lau. Measure representation and multifractal analysis of complete genomes. *Physical Review E*, 64(3 Pt 1):031903, 2001.
- [198] Z.-G. Yu, V. Anh, and K.-S. Lau. Multifractal and correlation analyses of protein sequences from complete genomes. *Physical Review E*, 68(2 Pt 1):021913, 2003.
- [199] Z.-G. Yu, V. Anh, K.-S. Lau, and K.-H. Chu. The genomic tree of living organisms based on a fractal model. *Physics Letters A*, 317(3):293–302, 2003.
- [200] Z.-G. Yu, L. Shi, Q.-J. Xiao, V. Anh, et al. Simulation for chaos game representation of genomes by recurrent iterated function systems. *Journal of Biomedical Science and Engineering*, 1(01):44, 2008.
- [201] Z.-G. Yu, X.-W. Zhan, G.-S. Han, R. Wang, V. Anh, and K. Chu. Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. *International Journal of Molecular Sciences*, 11(3):1141–1154, 2010.
- [202] C. Zhang, R. Zhang, and H. Ou. The Z curve database: A graphic representation of genome sequences. *Bioinformatics*, 19(5):593–599, 2003.
- [203] C.-T. Zhang and J Wang. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Research*, 28(14):2804–2814, 2000.

- [204] R. Zhang and C. Zhang. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. *Journal of Biomolecular Structure & Dynamics*, 11(4):767–782, 1994.
- [205] R. Zhang and C.-T. Zhang. Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea*, 1(5):335–346, 2005.
- [206] Y. Zhang and W. Chen. A measure of DNA sequence dissimilarity based on free energy of nearest-neighbor interaction. *Journal of Biomolecular Structure & Dynamics*, 28(4):557–565, 2011.
- [207] Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, and L. Pan. ColorSquare : A colorful square visualization of DNA sequences. *MATCH Communications in Mathematical and in Computer Chemistry*, 68:621–637, 2012.

Chapter 3

Mapping the space of genomic signatures¹

3.1 Introduction

Even though every year biologists discover and classify thousands of new species, it is estimated that as many as 86% of existing species on Earth and 91% of species in the oceans have not yet been classified and catalogued, [1]. 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, e.g., morphological, sequence-alignment-based phylogenetic analysis, and non-alignment-based molecular information.

We propose a computational process that outputs, 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.

¹A version of this chapter was published (L. Kari, K. Hill, A. Sayem, R. Karamichalis, N. Bryans, K. Davis, N. Dattani, “Mapping the space of genomic signatures”, *PLoS One* 10(5): e0119815 (2015))

The graphical representation we use is *Chaos Game Representation* (CGR) of DNA sequences, [2, 3], 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, 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 inferences of relationships between the corresponding species.

Concretely, to compute and visually display relationships within a given set $S = \{s_1, s_2, \dots, s_n\}$ of n DNA sequences, we propose a computational process that uses:

(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.

We apply this method to analyze and visualize several different taxonomic subsets of a dataset of 3,176 complete mtDNA sequences: phylum Vertebrata, (super)kingdom Protista, classes Amphibia-Insecta-Mammalia, class Amphibia only, and order Primates. We illustrate the usability of this approach by discussing, e.g., the placement of the genus *Polypterus* within phylum Vertebrata, of the unclassified organism *Haemoproteus* sp. jbl.JA27 within the (super)kingdom Protista, and the placement of the family Tarsiidae within the order Primates. We also provide an interactive web tool, *MoD Map* (*Molecular Distance Map*), that allows an in-depth exploration of all Molecular Distance Maps in the paper, complete with zoom-in features, search options, and easily accessible additional information for each sequence-representing

point (called hereafter sequence-point).

Overall, this method groups mtDNA sequences in correct taxonomic groups, from the kingdom level down to the order and family level. These results are of interest both because of the size of the dataset and because this information was extracted from DNA sequences that normally would not be considered in alignment-based comparison methods. Our analysis confirms that sequence composition (presence or absence of oligomers) contains taxonomic information that could be relevant to species identification, taxonomic classification, and identification of large evolutionary lineages. Last but not least, the appeal of this method lies in its simplicity, robustness, and generality, whereby exactly the same measuring tape can automatically yield meaningful measurements between non-specific DNA sequences of species as distant as those of the anatomically modern human and a cucumber, and as close as those of the anatomically modern human and the Neanderthal.

3.2 Methods

A CGR [2, 3] associates an image to each DNA sequence as follows. Begin with a unit square with corners labelled *A*, *C*, *G*, and *T*, clockwise starting from the bottom-left corner. The first point of any CGR 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 corresponding 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 di-

agonal crosses, see, e.g., Figure 3.1. CGRs of genomic DNA sequences have been shown to be genome- and species-specific, [2, 3, 4, 5, 6, 7, 8]. 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 for example [6].

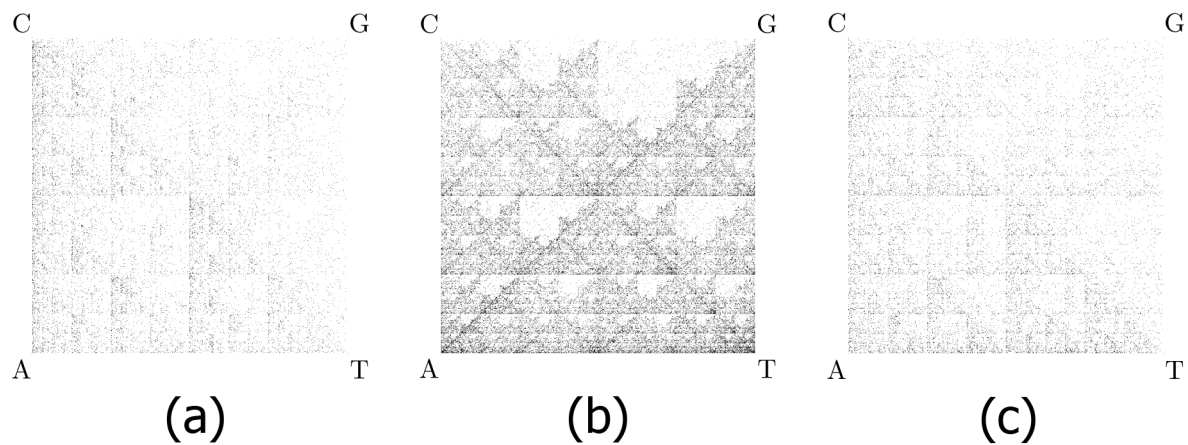


Figure 3.1: **CGR images for three DNA sequences.** (a) *Homo sapiens sapiens* mtDNA, 16,569 bp; (b) *Homo sapiens sapiens* chromosome 11, beta-globin region, 73,308 bp; (c) *Polypterus endlicherii* (fish) mtDNA, 16,632 bp. Observe that chromosomal and mitochondrial DNA from the same species can display different patterns, and also that mtDNA of different species may display visually similar patterns that are however sufficiently different as to be computationally distinguishable.

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 were being taken into consideration. In general, a length of DNA sequence of about 4,000 bp is necessary to obtain a well-defined CGR, but a length of 2,000 bp can sometimes give a good approximation, [2]. 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 [10, 11], vector walk [12], cell [13], vertical vector [14], Huffman coding [15], and colorsquare [16] methods. Three-dimensional representations of DNA sequences include the tetrahedron [17], 3D-vector [18], and trinucleotide curve [19] methods. Among these visualization methods, CGR images arguably provide the most immediately comprehensible “signature” of a DNA sequence and a desirable genome-specificity, [2, 7]. 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 [20, 21].

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, [22]. Another study, [23], 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 [24], 75 complete HIV genome sequences were compared using the Euclidean distance between frequencies of 6-mers ($k = 6$), in order to group them into subtypes. In [25], 27 microbial genomes were analyzed to find implications of 4-mer frequencies ($k = 4$) on their evolutionary relationships. In [26], 20 mammalian complete mtDNA sequences were analyzed using a so-called “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 [27]. 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 for example between 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 [28]. 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 [29] (where it was used for the analysis of geographic genetic distributions of some natural populations), [30] (where it was used to provide a graphical summary of the distances among CO1 genes from various species), and [31] (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 $p_1, p_2, \dots, p_n \in \mathbb{R}^q$ such that $d(i, j) = \|p_i - p_j\| \approx f(\Delta(i, j))$ for all $i, j \in \{1, \dots, n\}$ 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 classical 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 and the colour associated to that taxon in that 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 $x_{\text{sca}} = 2 \cdot \left(\frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}\right) - 1$, $y_{\text{sca}} = 2 \cdot \left(\frac{y - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}}\right) - 1$, 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, [32]), is defined in our case as

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

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 subset 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 [32].

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 [33]. The prokaryotic origins and evolutionary history of mitochondrial genomes have long been extensively studied, which will allow comparison of our results with known relatedness of species. Lastly, this genome dataset permits testing of both recent and deep rooted species relationships, providing fine resolution of species differences.

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 OpenMDM [34]. This program makes use of an open source MATLAB program for SSIM, [27], and MATLAB's built-in MDS function. The interactive web tool *MoD Map*, [35], allows an in-depth exploration and navigation of the Molecular Distance Maps in this paper. When using the web tool *MoD Map*, clicking on the "Draw MoD Map" button allows the selection of any of the five maps presented in the paper, each with features such as zoom-in and search by scientific name of the species or the NCBI accession number of its mtDNA. On any given Molecular Distance Map, clicking on a sequence-point displays its full mtDNA sequence information such as its unique identifier in this analysis, NCBI accession number, scientific name, common name, length of mtDNA sequence, taxonomy, CGR image, as well as a link to the corresponding NCBI entry. Clicking on the "From here" and "To here" buttons displays the image distance between the CGR images of two selected sequence-points, as a number between 0 and 1.

3.3 Results and Discussion

The Molecular Distance Maps we analyzed, of several different taxonomic subsets (phylum Vertebrata, (super)kingdom Protista, classes Amphibia-Insecta-Mammalia, class Amphibia only, and order Primates), confirm that the presence or absence of oligomers in mtDNA sequences may contain information that is relevant to taxonomic classifications. These results are relevant because they are the output of a method that bypasses the need of sequence alignment and uses as input DNA sequences that would not generally be considered by other, alignment-based, methods. The main contributions of the paper are the following:

- 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 $k = 9$. This image distance (with parameter set to $k = 9$) is highly sensitive and succeeds to successfully group hundreds of CGRs that are visually similar, such as the ones in Figure 3.1(a) and Figure 3.1(c), into correct taxonomic categories.
- 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. This is augmented by an interactive web tool which allows an in-depth exploration of the Molecular Distance Maps in this paper, with features such as zoom-in, search by scientific name or NCBI accession number, and quick access to complete information for each of the full mtDNA sequences in the map.
- 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.

- The use of a large dataset of 3,176 complete mitochondrial DNA sequences.
- An illustration of potential uses of this approach by the discussion of several case studies such as the placement of the genus *Polypterus* within phylum Vertebrata, of the unclassified organism *Haemoproteus* sp. jb1.JA27 (#1466) within the (super)kingdom Protista, and the placement of the family Tarsiidae within the order Primates.

This method could complement information obtained by using DNA barcodes [30] and Klee diagrams [36], since it is applicable to cases where barcodes may have limited effectiveness: plants and fungi for which different barcoding regions have to be used [37, 38, 39]; protists where multiple loci are generally needed to distinguish between species [40]; prokaryotes [41]; and artificial, computer-generated, DNA sequences. This method may also complement other taxonomic analyses by bringing in additional information gleaned from comparisons of non-homologous and non-coding sequences.

An example of the CGR/DSSIM/MDS approach is the Molecular Distance Map in Figure 3.2 which depicts the complete mitochondrial DNA sequences of all 1,791 jawed vertebrates in our dataset. (In the legends of all Molecular Distance Maps in this paper, the number of represented mtDNA sequences in each category is listed in parenthesis after the category name.) 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. In the case of Figure 3.2, the position of each sequence-point is determined by the 1,790 numerical distances between its sequence and all the other vertebrate mtDNA sequences in that dataset.

Observe that 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*

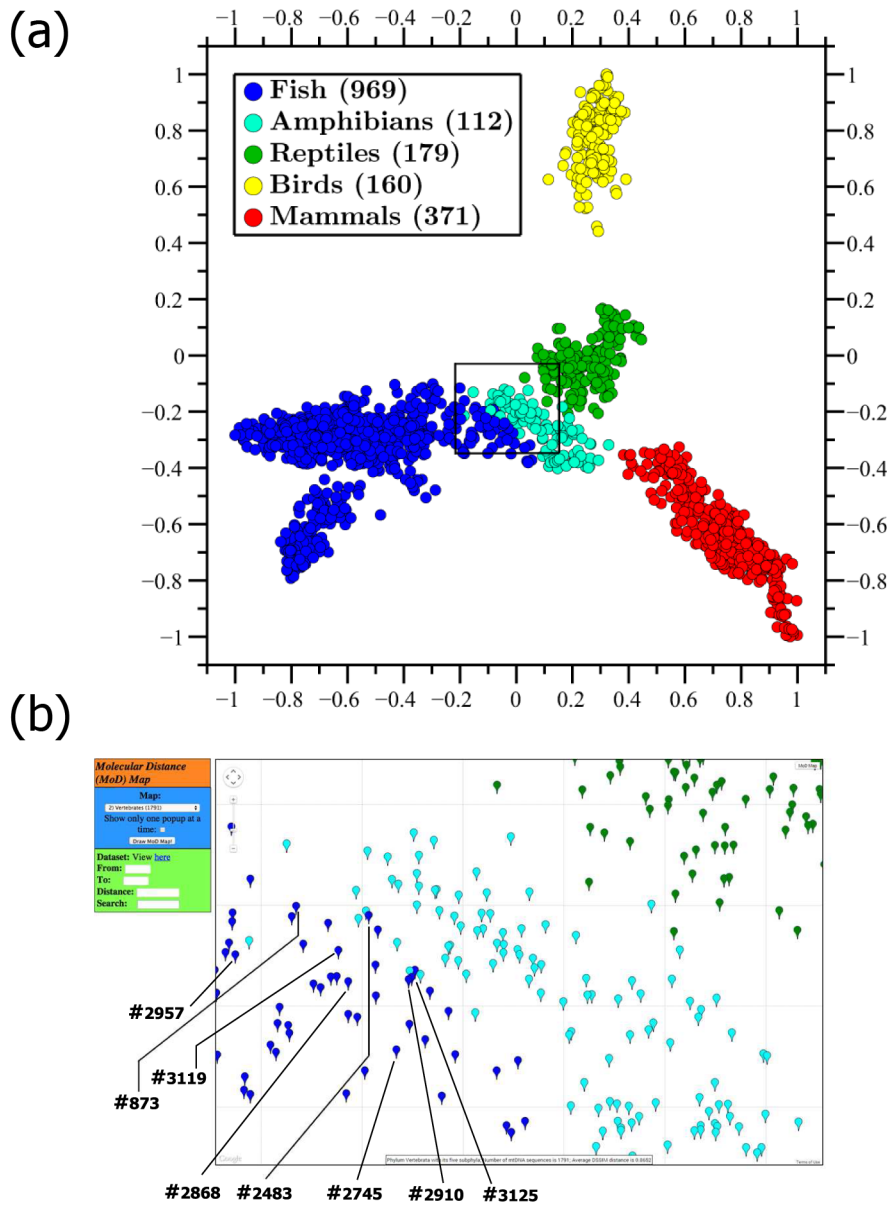


Figure 3.2: **Molecular Distance Map of phylum Vertebrata (excluding the 5 represented jawless vertebrates), with its five subphyla.** (a) This Molecular Distance Map comprises 1,791 mtDNA sequences, the average DSSIM distance is 0.8652, and the MDS *Stress-1* is 0.12. Fish species bordering amphibians 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 (*Porichthys myriaster* #2483), and all four represented lungfish (*Protopterus aethiopicus* #873, *Lepidosiren paradoxa* #2910, *Neoceratodus forsteri* #2957, *Protopterus doloii* #3119). Note that the question of whether species of the genus *Polypterus* are fish or amphibians has been discussed extensively for hundreds of years. Note also that gaps and spaces in clusters, in this and other maps, may be due to sampling bias. (b) Screenshot of the zoomed-in rectangular region outlined in Figure 3.2(a), obtained using the interactive web tool *MoD Map* [35].

(#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 [42]. 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). In answer to the hypothesis in [22] regarding the diversity of signatures across vertebrates, we note that in Figure 3.2 the 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).

We further 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 3.3, 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 3.3 is *Haemoproteus* sp. jb1.JA27 (#1466), sequenced in [43] (see also [44]), and listed as an *unclassified* organism in the NCBI taxonomy. Note first that this sequence-point 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 sequence-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

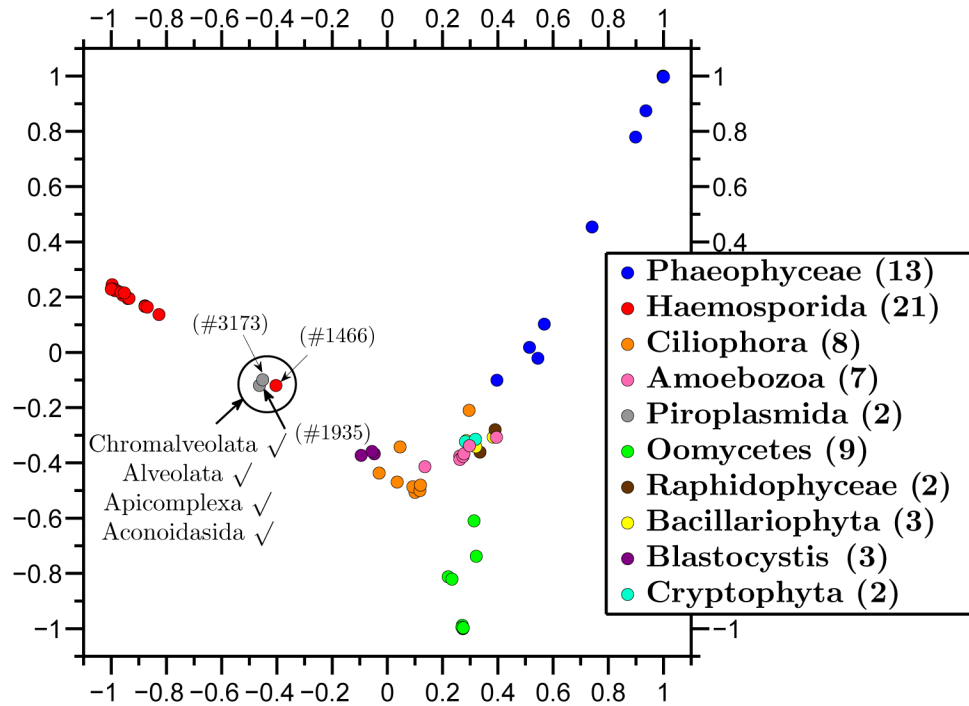


Figure 3.3: **Molecular Distance Map of all represented species from (super)kingdom Protista and its orders.** The total number of mtDNA sequences is 70, the average DSSIM distance is 0.8288, and the MDS *Stress-1* is 0.26. The sequence-point #1466 (red) is the unclassified *Haemoproteus* sp. jbl.JA27, #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.

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 3.4), as well as to observe relationships within class Amphibia and three of its orders (Figure 3.5).

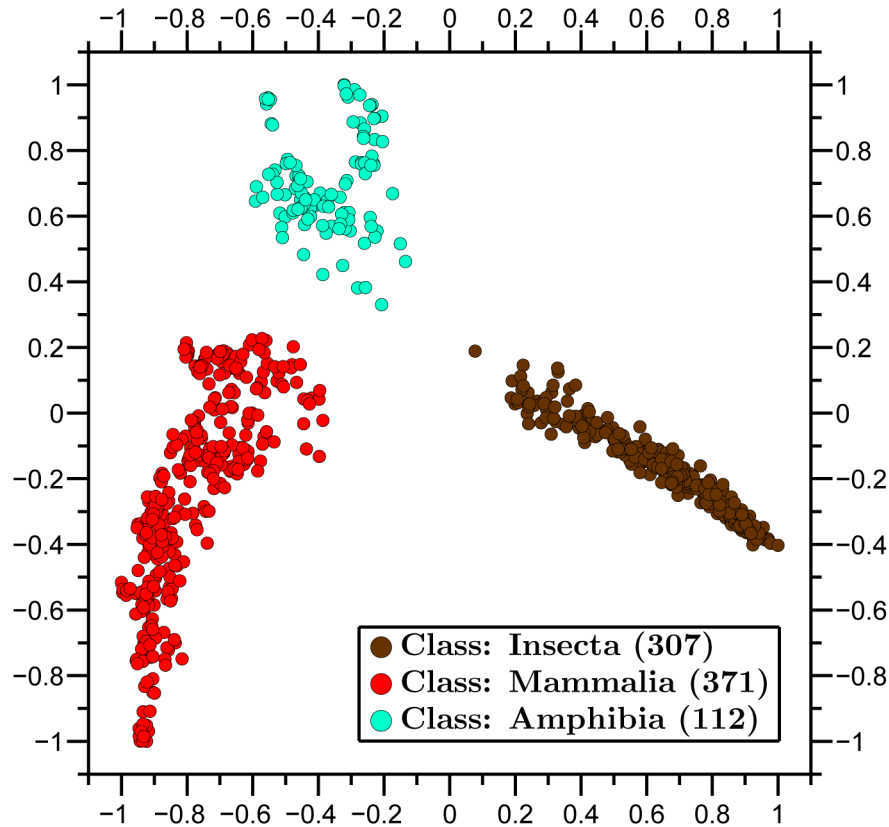


Figure 3.4: **Molecular Distance Map of three classes: Amphibia, Insecta and Mammalia.** The method successfully clusters taxonomic groups also at the Class level. Gaps and spaces in clusters, in this and other maps, may be due to sampling bias. A topic of further exploration would be to understand the cluster shapes and nature of the distribution of sequences in this figure. The total number of mtDNA sequences is 790, the average DSSIM distance is 0.8139, and the MDS *Stress-1* is 0.16.

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

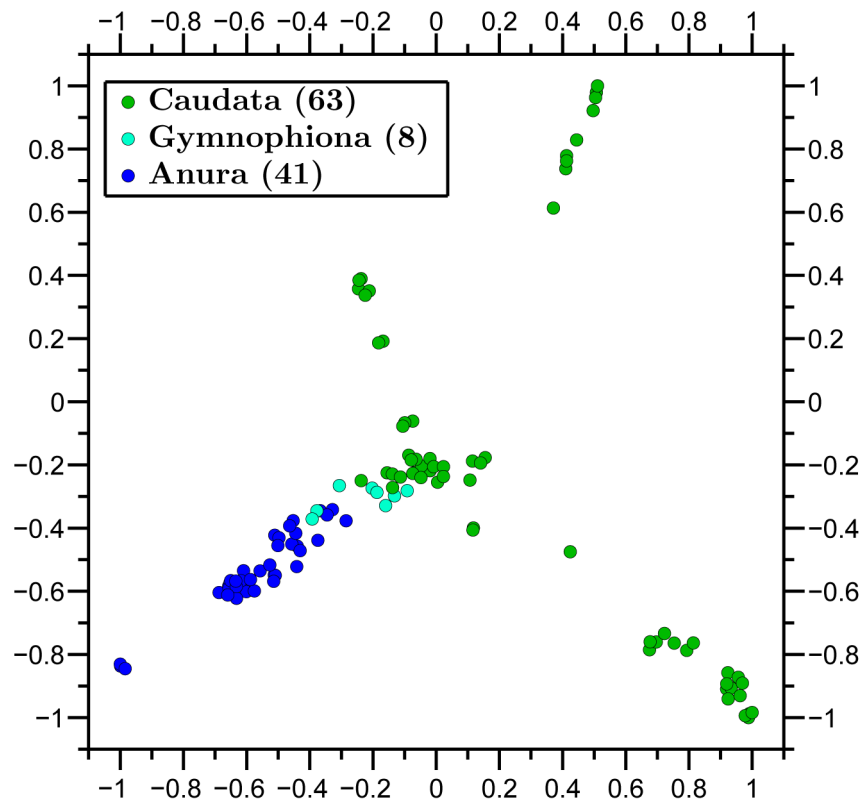


Figure 3.5: **Molecular Distance Map of class Amphibia and three of its orders.** The total number of mtDNA sequences is 112, the average DSSIM distance is 0.8445, and the MDS *Stress-1* is 0.18. Note that the shape of the amphibian cluster and the (x, y) -coordinates of sequence-points are different here from those in Figure 3.4. This is because MDS outputs a map that aims to preserve pairwise distances between points, but not necessarily their absolute coordinates.

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 the mtDNA sequence of an amphibian species in the Amphibia-Insecta-Mammalia map (Figure 3.4) will not necessarily be the same as the (x, y) -coordinates of the same point when only amphibians are mapped (Figure 3.5).

In general, Molecular Distance Maps are in good agreement with classical phylogenetic trees at all scales of taxonomic comparisons, see Figure 3.5 with [45], and Figure 3.6 with [46]. 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, discussed below.

Zooming in, we observed the relationships within an order, Primates, with its suborders (Figure 3.6). 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, Haplorrhini (dry-nosed primates comprising anthropoids and tarsiers) and Strepsirrhini (wet-nosed primates including lemurs and lorises). Figure 3.6 shows a clear separation of these suborders, with the top-left arm of the map comprising the Strepsirrhini. However, there are two Haplorrhini 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, [47]. According to [48], mitochondrial DNA evidence places tarsiiiformes as a sister group to Strepsirrhini, while in contrast, [49] places tarsiers within Haplorrhini. In Figure 3.6 the tarsiers are located within the Strepsirrhini cluster, thus agreeing with [48]. This may be partly because both this study and [48] used mitochondrial DNA, whose signature may be different from that of chromosomal DNA as seen in Figure 3.1(a) and Figure 3.1(b).

The DSSIM distances computed for 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)

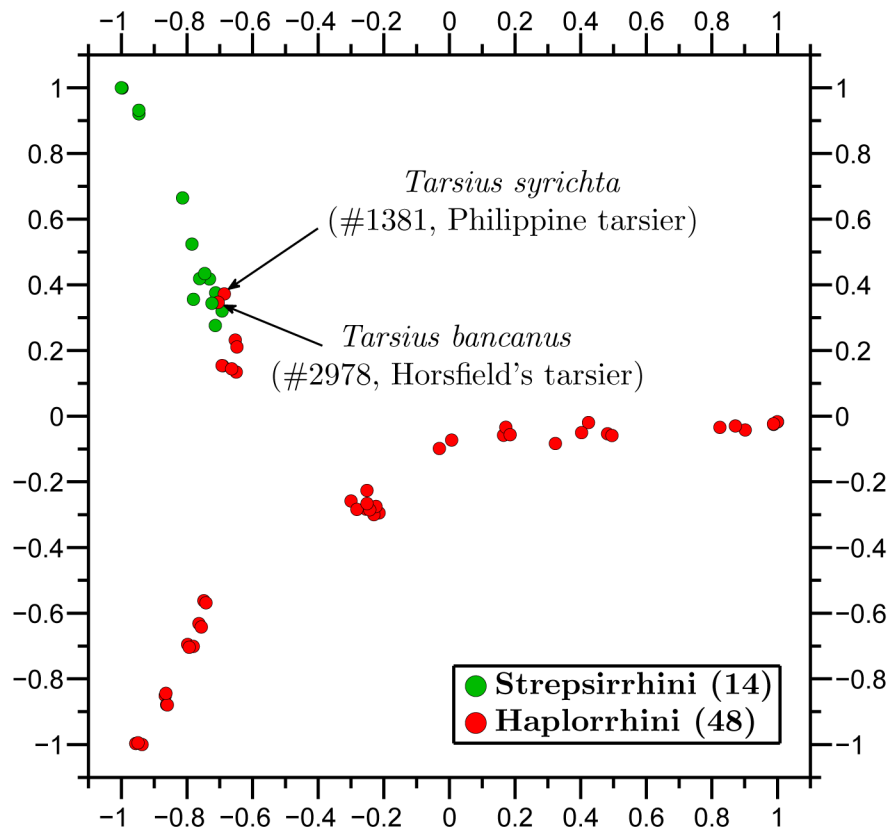


Figure 3.6: **Molecular Distance Map of order Primates and its suborders: Haplorrhini (anthropoids and tarsiers), and Strepsirrhini (lemurs, lorises, etc.).** The total number of mtDNA sequences is 62, the average DSSIM distance is 0.7733, and the MDS *Stress-1* is 0.19. The outliers are *Tarsius syrichta* #1381, and *Tarsius bancanus* #2978, whose placement within the order Primates has been subject of debate for over a century.

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. Subsequently, 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, [50], 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). Interestingly, the pair with the maximum distance $\Delta(\#2656, \#954) = 1.0033$ featured neither the longest mitochondrial DNA sequence, with the darkest CGR (*Cucumis sativus*, #533, cucumber, length 1,555,935 bp), nor the shortest mitochondrial DNA sequence, with the lightest CGR (*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 3.7 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, [46]. 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 *Cucumis sativus* (#533, cucumber, length 1,555,935 bp).

In addition to comparing real DNA sequences, this method can compare real DNA sequences to computer-generated sequences. As an example, we compared the mtDNA genome

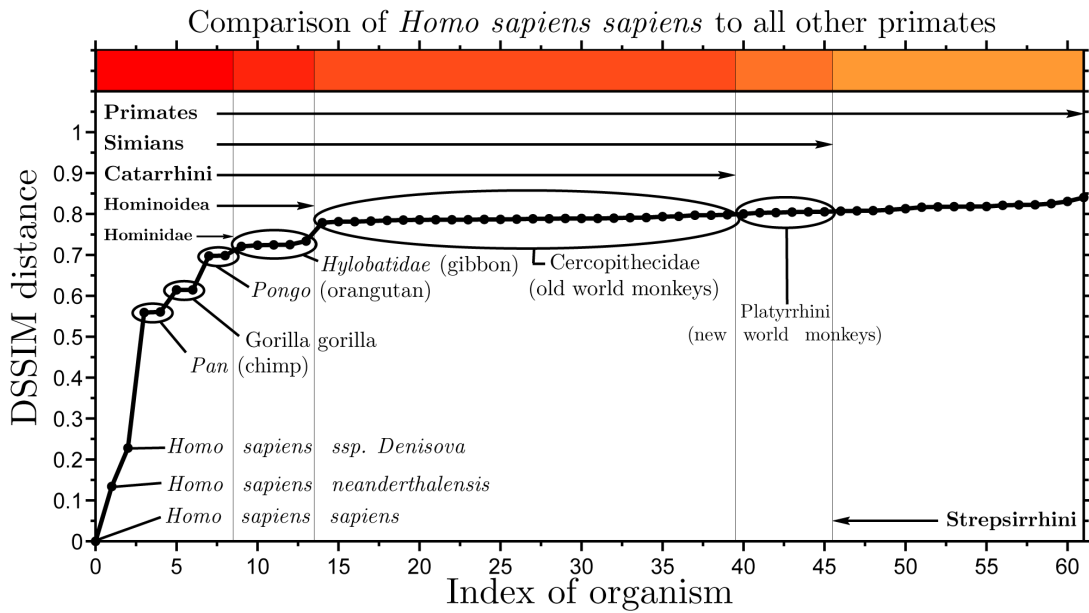


Figure 3.7: Graph of the DSSIM distances between the CGR images of *Homo sapiens sapiens* mtDNA and the CGR images of each of the 62 primate mitochondrial genomes (sorted by their distance from the human mtDNA). The distances are in accordance with established phylogenetic trees: The species with the smallest DSSIM distances from *Homo sapiens sapiens* are *Homo sapiens neanderthalensis*, *Homo sapiens ssp. Denisova*, followed by the chimp.

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 [51] that “CGR gives no further 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 being Figure 3.3 with *Stress-1* equal to 0.26. However, note that *Stress-1* generally decreases with an increase in the map’s dimensionality, from two to three or to a higher number of dimensions. In addition, as suggested in [28], 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. 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 method. 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, as seen in Figure 3.1(a) and Figure 3.1(c). 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 reader 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, as seen in Figure 3.1(a) and Figure 3.1(b), the genomic signature of mtDNA can be very different from that of nuclear DNA of the same species and care must be employed in choosing the dataset and interpreting the results.

3.4 Conclusions

Our analysis suggests that the oligomer composition of mitochondrial DNA sequences can be a source of taxonomic 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 3.2), and because this method circumvents the need for sequence similarity and extracts information from DNA sequences that normally would not be considered when

using local, homology-based comparisons.

Potential applications of Molecular Distance Maps - when used on a dataset of genomic sequences, whether coding or non-coding, homologous or not homologous, of the same length or vastly different lengths – include identification of large evolutionary lineages, taxonomic classifications, species identification, as well as quantitative definitions of the notion of species and other taxa.

Possible extensions include generalizations of MDS, such as 3-dimensional MDS, for improved visualization, and the use of increased oligomer length (higher values of k) for comparisons of longer subsequences in case of whole chromosome or whole genome analyses. Lastly, it is worth mentioning 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.

Bibliography

- [1] Mora C, Tittensor D, Adl S, Simpson A, Worm B (2011) How many species are there on earth and in the ocean? *PLoS Biology* 9: 1-8.
- [2] Jeffrey H (1990) Chaos Game Representation of gene structure. *Nucleic Acids Research* 18: 2163–2170.
- [3] Jeffrey H (1992) Chaos game visualization of sequences. *Comput Graphics* 16: 25-33.
- [4] Hill K, Schisler N, Singh S (1992) Chaos Game Representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *J Mol Evol* 35: 261-9.
- [5] Hill K, Singh S (1997) Evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes. *Genome* 40: 342-356.
- [6] Deschavanne P, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by Chaos Game Representation of sequences. *Molecular Biology and Evolution* 16: 1391–1399.
- [7] Deschavanne P, Giron A, Vilain J, Dufraigne C, Fertil B (2000) Genomic signature is preserved in short DNA fragments. In: *IEEE Intl. Symposium on Bio-Informatics and Biomedical Engineering*. pp. 161–167.
- [8] Wang Y, Hill K, Singh S, Kari L (2005) The spectrum of genomic signatures: From dinucleotides to Chaos Game Representation. *Gene* 346: 173–185.
- [9] Gates M (1986) A simple way to look at DNA. *J Theor Biology* 119: 319–328.
- [10] Nandy A (1994) A new graphical representation and analysis of DNA sequence structure: Methodology and application to globin genes. *Current Science* 66: 309 - 314.

- [11] Leong P, Morgenthaler S (1995) Random walk and gap plots of DNA sequences. Computer applications in the biosciences : CABIOS 11: 503-507.
- [12] Liao B (2005) A 2D graphical representation of DNA sequence. Chemical Physics Letters 401: 196–199.
- [13] Yao Y, Wang T (2004) A class of new 2D graphical representation of DNA sequences and their application. Chemical Physics Letters 398: 318–323.
- [14] Yu C, Liang Q, Yin C, He R, Yau S (2010) A novel construction of genome space with biological geometry. DNA Research 17: 155-168.
- [15] Qi Z, Li L, Qi X (2011) Using Huffman coding method to visualize and analyze DNA sequences. Journal of Computational Chemistry 32: 3233-3240.
- [16] Zhang Z, et al. (2012) Colorsquare: A colorful square visualization of DNA sequences. Comm in Math and in Comp Chemistry 68: 621-637.
- [17] Randic M, Vracko M, Nandy A, Basak S (2000) On 3D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inf and Comp Sci 40: 1235-1244.
- [18] Yuan C, Liao B, Wang T (2003) New 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters 379: 412 - 417.
- [19] Yu J, Sun X, Wang J (2009) TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. Journal of Theoretical Biology 261: 459 - 468.
- [20] Makula M, Benuskova L (2009) Interactive visualization of oligomer frequency in DNA. Computing and Informatics 28: 695-710.
- [21] Hao B, Lee H, Zhang S (2000) Fractals related to long DNA sequences and complete genomes. Chaos, Solitons and Fractals 11: 825-836.

- [22] Edwards S, Fertil B, Girron A, Deschavanne P (2002) A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic Biology* 51: 599-613.
- [23] Deschavanne P, DuBow M, Regard C (2010) The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virology Journal* 7: 163.
- [24] Pandit A, Sinha S (2010) Using genomic signatures for HIV-1 subtyping. *BMC Bioinformatics* 11: S26.
- [25] Pride D, Meinersmann R, Wassenaar T, Blaser M (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research* 13: 145-158.
- [26] Li M, Chen X, Li X, Ma B, Vitany P (2004) The similarity metric. *IEEE Transactions on Information Theory* 50: 3250-3264.
- [27] Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13: 600-612.
- [28] Borg I, Groenen P (2010) *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2nd edition.
- [29] Lessa E (1990) Multidimensional analysis of geographic genetic structure. *Systematic Zoology* 39(3): 242-252.
- [30] Hebert P, Cywinska A, Ball S, Dewaard J (2003) Biological identifications through DNA barcodes. *Proc Biol Sci* 270: 313-321.
- [31] Hillis D, Heath T, StJohn K (2005) Analysis and visualization of tree space. *Systematic Biology* 54: 471-482.
- [32] Kruskal J (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29: 1-27.

- [33] Sloan D, et al. (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology* 10: e1001241.
- [34] Dattani N, Sayem A, Tu R, Bryans N (2014) OpenMDM. Computer Program : <http://dx.doi.org/10.6084/m9.figshare.1243376>.
- [35] Karamichalis R (2014) *MoD-Map*. Web Tool : <https://github.com/rallis/MoDMap>.
- [36] Sirovich L, Stoeckle M, Zhang Y (2010) Structural analysis of biodiversity. *PLoS ONE* 5: e9266.
- [37] Kress W, Wurdack K, Zimmer E, Weigt L, Janzen D (2005) Use of DNA barcodes to identify flowering plants. *PNAS* 102: 8369–8374.
- [38] Hollingsworth P, et al. (2009) A DNA barcode for land plants. *PNAS* 106: 12794-2797.
- [39] Schoch C, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *PNAS* 109: 6241-6246.
- [40] Hoef-Emden K (2012) Pitfalls of establishing DNA barcoding systems in protists: the Cryptophyceae as a test case. *PLoS One* 7: e43652.
- [41] Unwin R, Maiden M (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol* 11: 479–487.
- [42] Hall B (2001) John Samuel Budgett (1872-1904): In pursuit of *Polypterus*. *BioScience* 51: 399-407.
- [43] Beadell J, Fleischer R (2005) A restriction enzyme-based assay to distinguish between avian hemosporidians. *Journal of Parasitology* 91: 683-685.

- [44] Valkiunas G, et al. (2010) 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.
- [45] Pyron R, Wiens J (2011) 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.
- [46] Shoshani J, et al. (1996) Primate phylogeny: morphological vs molecular results. *Molecular Phylogenetics and Evolution* 5: 102-154.
- [47] Jameson N, et al. (2011) Genomic data reject the hypothesis of a prosimian primate clade. *Journal of Human Evolution* 61: 295-305.
- [48] Chatterjee H, Ho S, Barnes I, Groves C (2009) Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evolutionary Biology* 9: 259.
- [49] Perelman P, Johnson W, Roos C, Seuánez H, Horvath J, et al. (2011) A molecular phylogeny of living primates. *PLoS Genetics* 7: e1001342.
- [50] Hu J, et al. (2012) Characteristics of diploid and triploid hybrids derived from female *Megalobrama amblycephala* Yih × male *Xenocypris davidi* Bleeker. *Aquaculture* 364-365: 157-164.
- [51] Goldman N (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in Chaos Game Representations of DNA sequences. *Nucleic Acids Research* 21: 2487-2491.

Chapter 4

An investigation into inter- and intragenomic variations of graphic genomic signatures¹

4.1 Introduction

Alongside DNA barcoding, [1] and Klee diagrams [2], Chaos Game Representation (CGR) patterns of genomic segments have been proposed as another method for the classification and identification of genomic sequences [3, 4]. The concept of *genomic signature* was first introduced in [5], as being any specific quantitative characteristic of a DNA genomic sequence that is pervasive along the genome of the same organism, while being dissimilar for DNA sequences originating from different organisms. Initial studies [3, 6] suggesting that short fragments of genomic sequences retain most of the characteristics of the genome of origin indicated that such genomic signatures exist. In particular, the Chaos Game Representation (CGR) of a DNA sequence, a graphic representation of its sequence composition, was proposed

¹A version of this chapter was published (R. Karamichalis, L. Kari, S. Konstantinidis and S. Kopecki, “An investigation of inter and intragenomic variations of graphic genomic signatures”, *BMC Bioinformatics*, 16:246 (2015))

in [3] as having both the pervasiveness and differentiability properties necessary for it to qualify as a genomic signature. Indeed, CGRs of genomic DNA sequences have been shown to be genome- and species-specific, see, e.g., [3, 6, 7, 8, 4, 9, 10, 11, 12]. Note that CGR patterns of mtDNA sequences can be different from those of DNA sequences from the major genome of the same organism, and that large scale quantitative analyses, at all taxonomic levels, of the hypothesis that CGR can play the role of a genomic signature for genomic sequences have not, to our knowledge, been performed. The long term objective of this research is to find out whether CGR can play the role of genomic signature for genomic DNA sequences, and can be used to identify and classify genomic sequences at all taxonomic levels. To this end, the objective of this study is to quantitatively assess the usability of CGR for classification of genomic sequences at the kingdom level, as well as to assess various distances that can be used to compare CGRs of genomic sequences for this purpose.

We first analyze 508 fragments, 150 kbp (kilo base pairs) long, spanning single complete chromosomes of six organisms, each representing a different kingdom: chromosome 21 of *Homo sapiens*, chromosome 4 of *Saccharomyces cerevisiae*, chromosome 1 of *Arabidopsis thaliana*, chromosome 14 of *Plasmodium falciparum*, the genome of *Escherichia coli*, and the genome of *Pyrococcus furiosus*, for a total length of 76,200 kbp analyzed. We analyze the intergenomic and intragenomic variation of CGR genomic signatures of these sequences by using six different distances: Structural Dissimilarity Index (DSSIM) [13], Euclidean distance, Pearson correlation distance [14], Manhattan distance [15], approximated information distance [16], and a distance defined here, based on an idea from computer vision, called *descriptor distance*. For each of the six distances, we visualize the results by computing Molecular Distance Maps, [12], which represent sequences as points in a two-dimensional or three-dimensional space, and thus display all their interrelationships simultaneously. The resulting Molecular Distance Maps show a good clustering, with genomic sequences originating from the same genome being largely grouped together, and separated from sequences belonging to genomes of different organisms. We observe that, in some of the cases where the clustering is subopti-

mal, the computation of three-dimensional Molecular Distance Maps resolves what appeared to be cluster overlaps in the two-dimensional Molecular Distance Maps. Using the “ground-truth” that sequences from the same genomes should have similar structural characteristics and thus be grouped together, while those from genomes of different organisms should be separated, we assess the six distances by combining three different quality measures: correlation to an idealized cluster distance, silhouette accuracy, and histogram overlap. We conclude that, for this dataset, DSSIM and the descriptor distance perform best according to these measures.

To maximize the diversity within each species, we also analyze a set of 526 fragments, 150 kbp long, sampled from the entire genomes of the aforementioned six organisms, for a total length of 78,900 kbp analyzed. The resulting Molecular Distance Maps are very similar to the ones in the first experiment, and the distance ranking is also the same, confirming the preceding results.

Lastly, we provide some preliminary evidence of this method’s applicability to classifying genomic DNA sequences at lower taxonomic levels by comparing 240 genomic sequences, 150 kbp long, sampled from the entire genome of *Homo sapiens* (class Mammalia, order Primates) with 210 genomic sequences, 150 kbp long, sampled from the entire genome of *Mus musculus* (mouse, class Mammalia, order Rodentia) for an additional length of 67,500 kbp analyzed. While a clear separation of sequences by genome is indeed achieved, we observe that the distance ranking is quite different compared to the previous two experiments, indicating that different distances may have to be used for comparing genomic sequences at different taxonomic levels.

Note that early analyses of genomic sequences with regard to similarities in the relative abundances of oligonucleotides of lengths $k = 1, \dots, 6$ exists and include [17, 18, 19, 20, 21, 22, 23, 24, 25]. Also, several alignment-free methods that use fixed-length word frequencies have been used for phylogenomic analysis of DNA sequences, [26, 27, 28]. These methods include statistical studies of word frequency within a DNA sequence [29, 5, 30, 31, 32, 33, 34], or employ k -words and the Markov model to obtain information about DNA sequences [35, 36,

37, 38, 39]. Iterated map methods for DNA sequence comparison include CGR-based analyses, see [3, 40, 41, 42, 43, 44, 45, 46], and such alignment-free methods have been successfully applied for sequence comparison [4, 47, 48, 49, 50, 11, 51, 52, 12, 53].

The initial reports on CGRs of genomic sequences [3, 6] contained mostly qualitative assessments of CGR patterns of whole genes. In [54], several comparisons of eukaryotic genomic sequences, including within-species comparisons, were reported, using di-, tri-, and tetranucleotide relative abundance distance ($k = 2, 3, 4$). In [25] di- and tetranucleotide abundance profiles ($k = 2, k = 4$) were compared for genomic collections from genomes of 5 gram-negative proteobacteria (including 2 complete genomes), 3 gram-positive bacteria, 2 mycoplasmas (complete genomes), 2 cyanobacteria (1 complete genome), and 3 thermophilic archaea (1 complete genome), using the δ^* distance which computes the average absolute difference of the dinucleotide relative abundance values. In [4], several datasets of up to 36 genomic DNA sequences were analyzed, and in [9] some various-length sequences were analyzed based on computing Euclidean distances between frequencies of their k -mers, for $k = 1, \dots, 8$. Subsequently, [10] computed 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. In [47], 27 microbial genomes were analyzed to find implications of 4-mer frequencies ($k = 4$) on their evolutionary relationships. In [16], 20 mammalian complete mtDNA sequences were analyzed using the “similarity metric”, for $k = 7$. In [50] a multigene dataset of 33 genes for 9 bacteria and one archaea species, as well as the whole genomes of a set of 16 γ -proteobacteria were analyzed, using values of k between 1 and 10, and Euclidean and χ^2 distances. In [11] a collection of 26 complete mitochondrial genomes was analyzed, using the Euclidean distance and an “image distance”, with a value of $k = 10$. In [55] a megabase-scale phylogenomic analysis of the Reptilia was reported, that compared frequency distributions of 8-mer oligonucleotides ($k = 8$) using Euclidean distance. Another study, [56], analyzed 459 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 [57], 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 [58] several datasets were analyzed (109 complete genomes of prokaryotes and eukaryotes, 34 prokaryote and chloroplast genomes, mitochondrial genomes of 64 vertebrates, and 62 complete genomes of alpha proteobacteria) using values of $k = 5, 6$ for protein-coding genes and $k = 11, 12$ for whole genomes, with two distances: chord distance and piecewise distance. In [12] a dataset of 3,176 complete mtDNA sequences was analyzed using an image distance, DSSIM, and a value of $k = 9$, and several Molecular Distance Maps were obtained which displayed sequences' interrelationships at several taxonomic levels (phylum Vertebrata, kingdom Protista, classes Amphibia-Insecta-Mammalia, class Amphibia, and order Primates).

The main contributions of this paper are:

- We tested and confirmed for an extensive dataset, of a total length of approximately 174Mbp, the hypothesis that CGR images of *genomic* DNA sequences can play the role of a (*graphic*) *genomic signature*, meaning that they have a desirable genome- and species-specificity. The dataset comprised 150 kbp fragments taken from genomes of six organisms, one from each of the six kingdoms of life. This was augmented by a set 150 kbp fragments randomly sampled from all chromosomes of *M. musculus*, as a test-case of this method's applicability at lower taxonomic levels.
- We assessed the performance of six different distances in this context, and this analysis included both same-genome and different-genome DNA fragment pairs. For several of these distances, the intragenomic values were overall smaller than intergenomic values, suggesting that this method could separate DNA genomic fragments belonging to different genomes, based on their CGRs.
- We showed that several distances outperform the Euclidean distance, which has so far been almost exclusively used for such studies. In particular, we determined that the DSSIM distance and the descriptor distance, adapted from computer vision for this ap-

plication, were best able to differentiate sequences originating from different genomes at the kingdom level. Both these distances essentially compare the k -mer composition of DNA sequences (herein $k = 9$).

- Based on preliminary data, we suggested the use of three-dimensional Molecular Distance Maps for improved visualization of the simultaneous interrelationships within a given set of genomic sequences.

Further analysis is needed to explore this method's potential to differentiate genomic sequences originating from closely related species (e.g. within the same order). Additional refinements of the distances considered may have to be defined for optimal genomic DNA sequence identification and classification at very low taxonomic levels.

4.2 Methods

In this section we first describe the dataset used for our analysis, then present an overview of the three main steps of the method, and conclude with a description of the six distances that we considered.

4.2.1 Dataset

We used the complete genomes from six organisms, each representing one of the six kingdoms of life. For the first experiment, we used one complete chromosome from each genome, see Table 4.1. For additional information about the dataset see [59], Appendix B.

In order to have relatively comparable numbers of DNA sequences for each organism, we chose the longest chromosomes for all organisms except *H. sapiens*, for which the shortest chromosome was chosen.

The DNA sequences in the NCBI database are represented as strings of letters “A”, “C”, “G”, “T”, and “N” which represent the four nucleotides Adenine, Cytosine, Guanine, Thymine,

	Organism	NCBI Acc. Nr.
1	<i>H. sapiens</i> , chrom. 21 (Animalia)	NC_000021.8
2	<i>E. coli</i> (Bacteria)	NC_000913.3
3	<i>S. cerevisiae</i> , chrom. 4 (Fungi)	NC_001136.10
4	<i>A. thaliana</i> , chrom. 1 (Plantae)	NC_003070.9
5	<i>P. falciparum</i> , chrom. 14 (Protista)	NC_004317.2
6	<i>P. furiosus</i> (Archaea)	NC_018092.1

Table 4.1: Dataset for the first experiment: NCBI accession numbers of the complete chromosomes considered, in increasing order of their NCBI accession number.

and “unidentified Nucleotide”, respectively. For our analysis we ignored all letters “N”. In *S. cerevisiae* and *E. coli* there were no ignored letters, and in *P. falciparum* and *P. furiosus* the number of ignored letters is of the order of 0.001% of the length of the sequence. In *H. sapiens* this number is 27%, and in *A. thaliana* is 0.54%. In *H. sapiens*, in particular, 96.4% of these ignored letters exist in centromeric and telomeric regions of the chromosome.

The resulting genomic DNA sequences were divided into successive, non-overlapping, contiguous fragments, each 150 kbp long. When the last sequence was shorter than 150 kbp, it was not included in the analysis. This resulted in 234 fragments for *H. sapiens*, 30 fragments for *E. coli*, 10 fragments for *S. cerevisiae*, 201 fragments for *A. thaliana*, 21 fragments for *P. falciparum*, and 12 fragments for *P. furiosus*, for a total of 508 DNA fragments, see Table 4.2.

Organism	Length(bp)	# Letters “N”	# Fragments
<i>H. sapiens</i>	48,129,895	13,023,253	234
<i>E. coli</i>	4,641,652	0	30
<i>S. cerevisiae</i>	1,531,933	0	10
<i>A. thaliana</i>	30,427,671	164,359	201
<i>P. falciparum</i>	3,291,871	37	21
<i>P. furiosus</i>	1,909,827	10	12

Table 4.2: The first experiment: Organisms considered, total length of the chromosome (respectively genome), number of ignored letters “N”, and number of DNA fragments (sequences) obtained by splitting a single complete chromosome per organism into consecutive, non-overlapping, equal length (150 kbp) contiguous fragments.

To maximize the diversity within each species, the dataset of the second experiment comprised fragments randomly sampled from each chromosome of the six chosen organisms, as

follows. After deleting all “N” nucleotides, each chromosome was divided into successive, non-overlapping, contiguous fragments, each 150 kbp long. When the last fragment was shorter than 150 kbp, it was not included in the analysis. Next, for each chromosome we selected randomly 10 such fragments to represent the chromosome, see [59], Appendix B. In the cases where there were fewer than 10 fragments in a chromosome, all of them were considered. In the cases of *E. coli* and *P. furiosus*, we retained all complete fragments of the genome. This resulted in 240 fragments for *H. sapiens*, 30 fragments for *E. coli*, 73 fragments for *S. cerevisiae*, 50 fragments for *A. thaliana*, 121 fragments for *P. falciparum*, and 12 fragments for *P. furiosus*, for a total of 526 fragments.

4.2.2 Overview

The method we used to analyze and classify genomic sequences has three steps: (i) generate graphical representations (images) of each DNA sequence using Chaos Game Representation (CGR), (ii) compute all pairwise distances between these images, and (iii) visualize the interrelationships implied by these distances as two- or three-dimensional maps, using Multi-Dimensional Scaling (MDS).

CGR is a method introduced by Jeffrey [3] in 1990 and studied in, e.g., [3, 6, 60, 61, 7, 62, 63, 11] as a way to visualize the structure of a DNA sequence. This method associates an image to each DNA sequence as follows. Starting from a unit square with corners 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 corner 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 *k*-mer: A pixel is black if the *k*-mer it represents occurs in the DNA sequence, otherwise it is white. CGR images of genetic DNA sequences originating from various species show patterns such as squares, parallel lines, rectangles, triangles, and also complex fractal patterns, Figure 5.6.

For step (i), a slight modification of the original CGR was used, introduced by Deschavanne

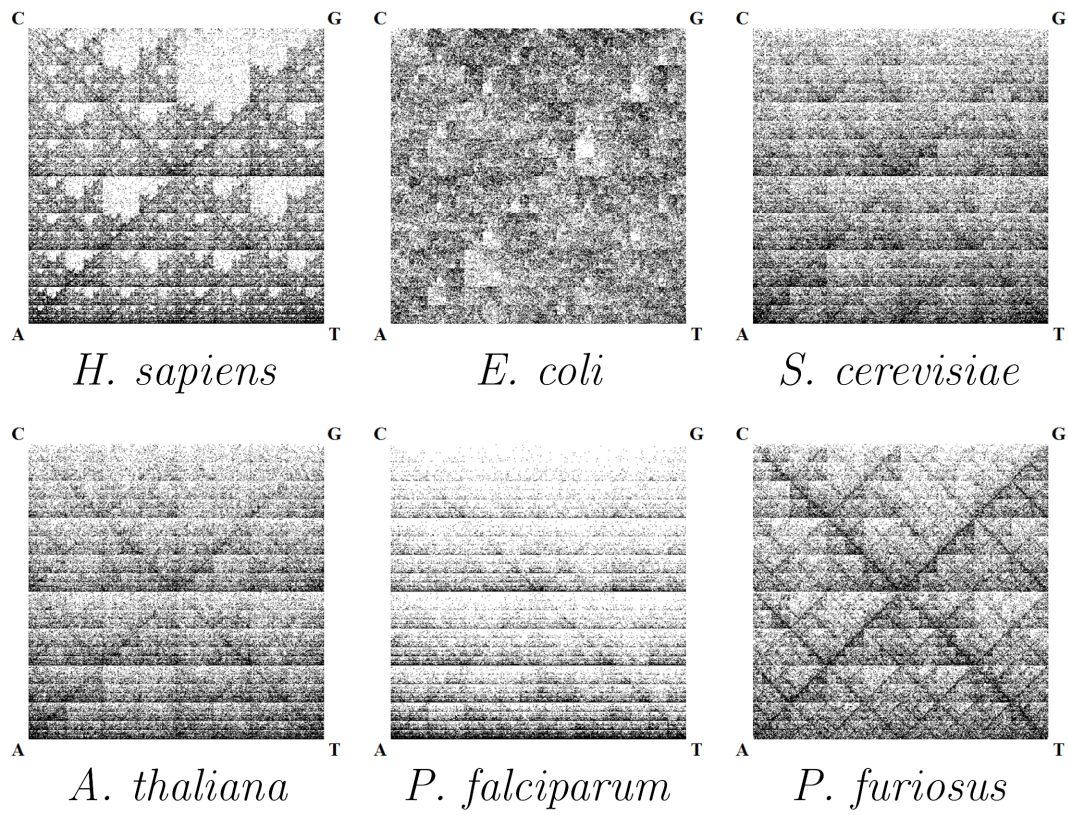


Figure 4.1: $2^9 \times 2^9$ CGR images of 150 kbp genomic DNA sequences from *H. sapiens*, *E. coli*, *S. cerevisiae*, *A. thaliana*, *P. falciparum*, and *P. furiosus*.

[4]: a k -th order FCGR (frequency CGR) is a $2^k \times 2^k$ matrix that can be constructed by dividing the CGR plot into a $2^k \times 2^k$ grid, and defining the element a_{ij} as the number of points that are situated in the corresponding grid square. A second order FCGR is shown below, where N_w is the number of occurrences of the oligonucleotide w in the sequence s .

$$FCGR_2(s) = \begin{pmatrix} N_{CC} & N_{GC} & N_{CG} & N_{GG} \\ N_{AC} & N_{TC} & N_{AG} & N_{TG} \\ N_{CA} & N_{GA} & N_{CT} & N_{GT} \\ N_{AA} & N_{TA} & N_{AT} & N_{TT} \end{pmatrix}$$

The $(k+1)$ -th order $FCGR_{k+1}(s)$ can be obtained by replacing each element N_X in $FCGR_k(s)$ with four elements

$$\begin{pmatrix} N_{CX} & N_{GX} \\ N_{AX} & N_{TX} \end{pmatrix}$$

where X is a sequence of length k over the alphabet $\{A, C, G, T\}$.

For step (ii), after computing the FCGR matrices for each of the 150 kbp sequences in a given dataset, the goal was to measure “distances” between every two CGR images. There are many distances that can be defined and used for this purpose, [64]. One of the goals of this study was to identify what distance is better able to differentiate the structural differences of various genomic DNA sequences and classify them based on the species they belong to. In this paper we use six different distances: Structural Dissimilarity Index (DSSIM), descriptor distance (adapted from computer vision for this application), Euclidean distance, Manhattan distance, Pearson correlation distance, and approximated information distance.

For step (iii), after computing all possible pairwise distances we obtained six different distance matrices. To visualize the inter-relationships between sequences implied by each of the distance matrices, and to thus visually assess each of the distances, we used Multi-Dimensional Scaling (MDS). MDS is an information visualization technique introduced by Kruskal in [65]. MDS takes as input a distance matrix that contains the pairwise distances among a set of items (here the items are the 150 kbp DNA sequences analyzed). The output of MDS is a spatial representation of the items in a common Euclidean space, wherein each item is represented as

a point and the spatial distance between any two points corresponds to the distance between the items in the distance matrix. Objects with a small pairwise distance will result in points that are close to each other, while objects with a large pairwise distance will become points that are far apart.

The combination of CGR/DSSIM/MDS was first proposed in [66], [12] as a tool to quantitatively measure and display the interrelationships among a set of complete mitochondrial sequences. The outputs of this method, called Molecular Distance Maps, are two-dimensional maps wherein each point represents a mitochondrial genome, and the spatial distances between any two points correspond to the differences between the structural composition of the corresponding DNA sequences. The ideal Molecular Distance Map is a placement of n items as points in an $(n - 1)$ -dimensional space. The two-dimensional Molecular Distance Map is simply an approximation, a flattening of this highly-dimensional space onto the plane, which may sometimes result in erroneous positioning of some points. Increasing the dimensionality of the Molecular Distance Map often results in a more accurate representation of the real interrelationships between sequences, as embodied in the original distance matrix.

4.2.3 Distances

In this section we describe and formally define each of the six distances used in our analysis: DSSIM, descriptor distance (adapted from computer vision for this application), Euclidean, Manhattan, Pearson, and approximated information distance.

Structural Similarity Index, SSIM, was introduced in [13] for the purpose of assessing the degree of similarity between two images. Given two images X, Y as $n \times n$ matrices having as elements integers ranging in the interval $[0, L]$, SSIM computes three factors (luminance, contrast and structure) and combines them to obtain a similarity value. However, instead of computing a global similarity between the two images, each image is divided into 11×11 sliding square windows $X^{ij}(Y^{ij}$ respectively) with $i, j = 1, \dots, n - 10$ which move pixel by pixel to eventually cover the entire image. The SSIM similarity of any given pair of images

is then computed by comparing their corresponding square windows. In addition, an 11×11 circular symmetric Gaussian weighting function $W \in \mathbb{R}^{11 \times 11}$ with a fixed standard deviation of 1.5, normalized to unit sum ($\sum_{p=1}^{11} \sum_{q=1}^{11} W_{pq} = 1$), is used. Then, the mean $\mu_{x,i,j}$ ($\mu_{y,i,j}$ for Y), variance $\sigma_{x,i,j}$ ($\sigma_{y,i,j}$ for Y) and correlation $\sigma_{xy,i,j}$ are computed, as follows:

$$\mu_{x,i,j} = \sum_{p=1}^{11} \sum_{q=1}^{11} W_{pq} X_{pq}^{ij}$$

$$\sigma_{x,i,j} = \sqrt{\sum_{p=1}^{11} \sum_{q=1}^{11} W_{pq} (X_{pq}^{ij} - \mu_{x,i,j})^2}$$

$$\sigma_{xy,i,j} = \sum_{p=1}^{11} \sum_{q=1}^{11} W_{pq} (X_{pq}^{ij} - \mu_{x,i,j})(Y_{pq}^{ij} - \mu_{y,i,j})$$

where A_{pq} denotes the (p, q) element of the matrix A . Based on these values, the luminance $l(X^{ij}, Y^{ij})$, contrast $c(X^{ij}, Y^{ij})$ and structure $s(X^{ij}, Y^{ij})$ are computed as

$$l(X^{ij}, Y^{ij}) = \frac{2\mu_{x,i,j}\mu_{y,i,j} + C_1}{\mu_{x,i,j}^2 + \mu_{y,i,j}^2 + C_1}$$

$$c(X^{ij}, Y^{ij}) = \frac{2\sigma_{x,i,j}\sigma_{y,i,j} + C_2}{\sigma_{x,i,j}^2 + \sigma_{y,i,j}^2 + C_2}$$

$$s(X^{ij}, Y^{ij}) = \frac{\sigma_{xy,i,j} + C_3}{\sigma_{x,i,j}\sigma_{y,i,j} + C_3}$$

where $C_1 = (0.01)^2$, $C_2 = (0.03)^2$, $C_3 = \frac{C_2}{2}$. Then, these three factors are combined to get

$$SSIM(X^{ij}, Y^{ij}) = l(X^{ij}, Y^{ij})c(X^{ij}, Y^{ij})s(X^{ij}, Y^{ij})$$

and finally, the SSIM index used to evaluate the overall image similarity is computed as

$$SSIM(X, Y) = \frac{1}{(n-10)^2} \sum_{i=1}^{n-10} \sum_{j=1}^{n-10} SSIM(X^{ij}, Y^{ij}).$$

In theory, the values for SSIM range in the interval $[-1, 1]$ with the similarity being 1

between two identical images, 0, for example, between a black image and a white image, and -1 if the two images are negatively correlated; that is, $\text{SSIM}(X, Y) = -1$ if and only if X and Y have the same luminance μ and every pixel x_i of image X has the inverted value of the corresponding pixel $y_i = 2\mu - x_i$ in Y .

To compute the distance rather than the similarity between two images, we calculate $\text{DSSIM}(X, Y) = 1 - \text{SSIM}(X, Y)$. Consequently, the range of DSSIM is the interval $[0, 2]$: two identical images will result in a DSSIM distance of 0, while two images that are the negatives of each other would result in a DSSIM distance of 2.

For defining the *descriptor distance* we adapted for this application the spatial pyramid matching approach of [67], which is used to calculate hierarchical image descriptors. The *descriptor distance* between two FCGRs $X, Y \in \mathbb{N}^{2^k \times 2^k}$ aims to compare a combination of several different “descriptors”, that is, a combination of several different aspects, of the two given FCGRs.

A *descriptor* is a vector characterized by parameters m and r , as well as r intervals, where m is the size of the non-overlapping windows in which the FCGR is divided (scale of the comparison), and the r intervals represent the “granularity” of the analysis, in that they define the intervals of numbers of k -mer occurrences that are considered significant.

For a given $m \leq k$ and r , and intervals $[a_0, a_1), [a_1, a_2), \dots, [a_{r-1}, a_r)$ such that $\bigcup_{i=0}^{r-1} [a_i, a_{i+1}) = [0, \infty)$ and $[a_i, a_{i+1}) \cap [a_j, a_{j+1}) = \emptyset \forall i, j$ with $i \neq j$, a descriptor is constructed as follows.

Starting from the top-left corner, we divide each of the two FCGR matrices X and Y into non-overlapping submatrices² of size $2^m \times 2^m$. This procedure results in 4^{k-m} submatrices X_{ij} and Y_{ij} with $i, j = 1, \dots, 2^{k-m}$, which will be pairwise compared.

The choice of the r intervals, called “bins”, points to the fact that, rather than considering the finest granularity, we are interested in a coarser comparison. This means that, instead of a computationally expensive pairwise comparison of all possible numbers of occurrences of k -mers, we are interested only in certain “bins” of such numbers. For example, in our

²In general, these windows (submatrices) can be overlapping, but in this paper we made the choice of using non-overlapping windows.

case, we use $r = 5$ and consider only 5 different bins, that is only k -mers with number of occurrences: 0 (not occurring), 1 (one occurrence), 2 (two occurrences), between 2 and 5, between 5 and 20, and greater than 20 (most frequent). Formally, we use $r = 5$ and $[0, \infty) = [0, 1) \cup [1, 2) \cup [2, 5) \cup [5, 20) \cup [20, \infty)$ as the 5 bins.

Afterwards, we compute for every X_{ij} a vector $\text{vec}X_{ij} = \frac{1}{(2^m \times 2^m)}(b_1, b_2, \dots, b_r)$ where $b_i = |\{x \in X_{ij} : a_{i-1} \leq x < a_i\}|$. In our case, for each X_{ij} , we compute a five-tuple wherein, for example, the 4th element represents the number of 9-mers whose number of occurrences is in the 4th bin, that is, at least 5 but less than 20. The division to $2^m \times 2^m$ is to obtain a probability distribution for each submatrix. The same procedure is performed for Y_{ij} , resulting in the vector $\text{vec}Y_{ij}$.

We further append all vectors $\text{vec}X_{ij}$ and form a new vector $\text{vec}X^{m,r}$ and, using the same order of appending, we append all vectors $\text{vec}Y_{ij}$ forming a new vector $\text{vec}Y^{m,r}$. These two vectors are the “descriptors” of the FCGR matrices X and Y for the parameters m , r and the r chosen bins.

As a last step, we combine descriptors $\text{vec}X^{m,r}$ (respectively $\text{vec}Y^{m,r}$) for several values of m and r by appending them one after another, in the same order, to obtain the vector $\text{vec}X$ (respectively $\text{vec}Y$).

The *descriptor distance* between the two FCGRs X and Y is now defined as the Euclidean distance between the vectors $\text{vec}X$ and $\text{vec}Y$

$$d_D(X, Y) = d_E(\text{vec}X, \text{vec}Y).$$

In our case we computed descriptors for $m = 4, 5, 6$ therefore forming vectors $\text{vec}X$ and $\text{vec}Y$ of length $5((\frac{512}{64})^2 + (\frac{512}{32})^2 + (\frac{512}{16})^2) = 6720$. In general, for a given r , the length of the vectors compared is $r((2^{k-m_1})^2 + (2^{k-m_2})^2 + \dots + (2^{k-m_p})^2)$, where m_1, m_2, \dots, m_p are the values used for m . The choice of m for this study was made to balance the computational cost of calculating the vector of descriptors with the ability to compare the two matrices at various scales: large ($m = 6$, that is, compare windows of size 64×64), medium ($m = 5$, windows of

size 32×32) and small ($m = 4$, windows of size 16×16). The parameter $r = 5$ and the 5 bins were kept constant throughout our calculations but, in general, these parameters can also be varied, and the resulting vectors for each value added to the vector of descriptors, resulting in a larger vector.

In principle, the descriptor distance between two given FCGRs effectively compares the distribution of frequencies of k -mers between the corresponding submatrices X_{ij} and Y_{ij} , and does that for several values of m , that is, at several different scales. (Note that, in each window X_{ij} , all k -mers have the same suffix of length $k - m$.)

We now illustrate the *descriptor distance* by an example wherein $k = 3$, $m = 2$, $r = 3$, and the 3 bins are $[0, 15) \cup [15, 30) \cup [30, \infty)$. Since $k = 3$, the FCGR table will contain the number of occurrences of all 3-mers in a DNA sequence, as follows:

CCC	GCC	CGC	GGC	CCG	GCG	CGG	GGG
ACC	TCC	AGC	TGC	ACG	TCG	AGG	TGG
CAC	GAC	CTC	GTC	CAG	GAG	CTG	GTG
AAC	TAC	ATC	TTC	AAG	TAG	ATG	TTG
CCA	GCA	CGA	GGA	CCT	GCT	CGT	GGT
ACA	TCA	AGA	TGA	ACT	TCT	AGT	TGT
CAA	GAA	CTA	GTA	CAT	GAT	CTT	GTT
AAA	TAA	ATA	TTA	AAT	TAT	ATT	TTT

Take the two FCGRs $X, Y \in \mathbb{N}^{8 \times 8}$, ($k = 3$, thus $2^3 \times 2^3$) corresponding to two genomic 150 kbp sequences of our dataset (one human and one bacterial), respectively. In order to use small numbers throughout the example, we divide all elements of the obtained matrices by 100 and take the integer part of each element, obtaining:

$$X = \begin{pmatrix} 42 & 33 & 9 & 33 & 14 & 10 & 15 & 45 \\ 22 & 30 & 26 & 25 & 9 & 5 & 37 & 37 \\ 32 & 21 & 33 & 19 & 44 & 35 & 41 & 35 \\ 17 & 9 & 13 & 21 & 23 & 10 & 22 & 18 \\ 37 & 26 & 6 & 32 & 34 & 24 & 9 & 23 \\ 29 & 24 & 31 & 27 & 19 & 27 & 18 & 28 \\ 21 & 23 & 10 & 9 & 19 & 17 & 21 & 15 \\ 35 & 15 & 14 & 14 & 19 & 12 & 17 & 30 \end{pmatrix},$$

$$Y = \begin{pmatrix} 18 & 34 & 40 & 27 & 30 & 36 & 27 & 12 \\ 27 & 18 & 27 & 32 & 24 & 23 & 15 & 23 \\ 24 & 17 & 13 & 17 & 36 & 12 & 32 & 18 \\ 27 & 17 & 28 & 26 & 18 & 8 & 22 & 25 \\ 32 & 32 & 23 & 16 & 16 & 25 & 23 & 22 \\ 20 & 29 & 18 & 25 & 16 & 16 & 15 & 17 \\ 25 & 25 & 7 & 16 & 26 & 27 & 20 & 25 \\ 32 & 21 & 20 & 21 & 25 & 18 & 27 & 34 \end{pmatrix}.$$

Thus, in the human DNA sequence, the triplet CCC appears about 42 x 100 times, the triplet GCC appears about 33 x 100 times, the triplet CGC appears about 9 x 100 times, etc.

Since $m = 2$, we divide each of the matrices X and Y into non-overlapping submatrices of size 4×4 ($2^2 \times 2^2$). For X we thus obtain $X_{11}, X_{12}, X_{21}, X_{22}$

$$\begin{pmatrix} 42 & 33 & 9 & 33 \\ 22 & 30 & 26 & 25 \\ 32 & 21 & 33 & 19 \\ 17 & 9 & 13 & 21 \end{pmatrix}, \begin{pmatrix} 14 & 10 & 15 & 45 \\ 9 & 5 & 37 & 37 \\ 44 & 35 & 41 & 35 \\ 23 & 10 & 22 & 18 \end{pmatrix},$$

$$\begin{pmatrix} 37 & 26 & 6 & 32 \\ 29 & 24 & 31 & 27 \\ 21 & 23 & 10 & 9 \\ 35 & 15 & 14 & 14 \end{pmatrix}, \begin{pmatrix} 34 & 24 & 9 & 23 \\ 19 & 27 & 18 & 28 \\ 19 & 17 & 21 & 15 \\ 19 & 12 & 17 & 30 \end{pmatrix}.$$

and similarly for Y .

Since the $r = 3$ bins are $[0, 15) \cup [15, 30) \cup [30, \infty)$, we will count, for each submatrix, the number of 3-mers for which the number of occurrences is less than 15, between 15 and 30, and greater than or equal to 30. Thus we obtain $\text{vec}X_{11} = \frac{1}{16}(3, 7, 6)$ which has as elements the number of elements of X_{11} which belong in each of the intervals selected, divided by the total number of elements of X_{11} . We proceed similarly for $\text{vec}X_{12} = \frac{1}{16}(5, 4, 7)$, $\text{vec}X_{21} = \frac{1}{16}(5, 7, 4)$, $\text{vec}X_{22} = \frac{1}{16}(2, 12, 2)$ and we form $\text{vec}X$ by appending these vectors one after the other, that is

$$\text{vec}X = \frac{1}{16} (3, 7, 6, 5, 4, 7, 5, 7, 4, 2, 12, 2).$$

We apply exactly the same procedure for the matrix Y and we get

$$\text{vec}Y = \frac{1}{16} (1, 12, 3, 3, 9, 4, 1, 12, 3, 0, 15, 1).$$

The descriptor distance between these two FCGRs is computed as the Euclidean distance between $\text{vec}X$ and $\text{vec}Y$, in this case $d_D(X, Y) \approx 0.718$. Note that, since we started by dividing the number of 3-mer occurrences by 100, as well as because of the bin selection, this is a fictitious example. The real value of the descriptor distance between the mentioned human and bacterial sequences is 8.66, and the range of the descriptor distance for this dataset of DNA sequences is $[0, 13.17]$. In general, the descriptor distance has a variable range, that depends on the choices of parameters used.

To compute the Euclidean, Manhattan and Pearson distances, we first convert the matrices $X, Y \in \mathbb{N}^{n \times n}$ into $1 \times n^2$ vectors. For two vectors $x, y \in \mathbb{R}^n$, their Euclidean distance $d_E(x, y)$ and their Manhattan distance $d_M(x, y)$ are computed as

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

$$d_M(x, y) = \sum_{i=1}^n |x_i - y_i|,$$

while their Pearson distance $d_P(x, y)$ is defined as

$$d_P(x, y) = 1 - \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where

$$\mu_x = \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)^2},$$

$$\sigma_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y).$$

In theory, the correlation coefficient $\frac{\sigma_{xy}}{\sigma_x \sigma_y}$ ranges in the interval $[-1, 1]$, and therefore the Pearson distance ranges in the interval $[0, 2]$.

The last distance we considered is based on the information distance defined in [16]. The use of this distance is motivated computationally since it is easily computed from FCGRs as it tracks only the number of different k -mers for a sequence instead of the actual set. In [16], for a given k , the information distance for two strings x, y is defined as

$$d_{AID}(x, y) = \frac{N_k(x|y) + N_k(y|x)}{N_k(xy)}$$

with

$$N_k(x|y) = N_k(xy) - N_k(x)$$

where $N_k(x)$ is the number of different k -mers (possibly overlapping) which occur in x . We go one step further and modify this in order to avoid the creation of “unwanted” k -mers from the concatenation xy of x and y . We now show how to compute $N_k(x)$ for a sequence x . For a sequence x , first we build its FCGR(x) = $X \in \mathbb{N}^{2^k \times 2^k}$, which is a matrix of $2^k \times 2^k$ with

element values in \mathbb{N} . Then we unitize X , that is every non-zero entry becomes 1, while zeros remain 0. $N_k(x)$ is now computed as the sum of the elements of this unitized FCGR, that is, $N_k(x) = f(X) = \text{SumOfElements}(\text{Unitize}(X))$. For two strings x and y , with FCGRs X and Y respectively, we define $N_k(x|y)$ as:

$$N_k(x|y) = f(X + Y) - N_k(x) \quad (4.1)$$

This slight modification of the information distance gives us also the desired properties of $d(x, x) = 0$ and $d(x, y) = d(y, x)$ which were not satisfied before. Using (4.1), we now define the *approximated information distance* (AID) as:

$$d_{AID}(x, y) = 2 - \frac{f(X) + f(Y)}{f(X + Y)} \quad (4.2)$$

where x, y are the strings and $X, Y \in \mathbb{N}^{2^k \times 2^k}$ their FCGRs, respectively. It also turns out that this distance is in fact the normalized Hamming Distance of the unitized FCGRs X and Y . Note that, for two sets \mathcal{X} and \mathcal{Y} , the normalized Hamming distance is $\frac{|\mathcal{X} \Delta \mathcal{Y}|}{|\mathcal{X} \cup \mathcal{Y}|} = 2 - \frac{|\mathcal{X}| + |\mathcal{Y}|}{|\mathcal{X} \cup \mathcal{Y}|}$ where Δ denotes the symmetric difference.

Online Material, [59], includes the code used, the distance matrices, and an Appendix (Appendix A with details about accessing the online resources, Appendix B with information about the dataset, and Appendix C with additional histograms for the first experiment). The code, written in Wolfram Mathematica version 9, was used (and can be tested) for the generation of CGR images, the calculation of distance matrices, and the creation of 2D and 3D Molecular Distance Maps. The interactive webtool ModMap, [68], allows in-depth exploration of the 2D Mod Maps (Molecular Distance Maps) in this paper. When using the interactive webtool MoDMap, clicking on a distance underneath a dataset will result in plotting the MoD Map of the dataset computed with that distance. On any particular MoD Map, clicking on a point will display a window with information about the subsequence represented by that point: its NCBI accession number, scientific name of the organism it originates from, and its CGR pattern.

Clicking on the “From here” and “To here” buttons on two such selected windows will display the distance between the corresponding genomic subsequences in the distance matrix.

4.3 Analysis and Results

For our dataset, we use $k = 9$, that is, each DNA sequence was represented as a $2^9 \times 2^9$ FCGR matrix. In practice, this means that the FCGR of a DNA sequence contains the full information regarding its k -mer sequence composition, for $k = 1, 2, \dots, 9$. The length choice of 150 kbp and value of $k = 9$ is partly justified by the fact that, for a random sequence of length 150 kbp, its CGR at resolution $2^9 \times 2^9$ has around half of the pixels black, and half white, and partly justified by the fact that it empirically produced good results while at the same time being computationally inexpensive.

Figure 4.2 depicts two-dimensional Molecular Distance Maps obtained from the first experiment, using one complete chromosome for each organism, computed using the DSSIM distance, descriptor distance, Euclidean distance, Manhattan distance, Pearson distance and approximated information distance, respectively. Figure 4.3 depicts the corresponding three-dimensional Molecular Distance Maps for the same dataset. The projection of each three-dimensional map is chosen by hand in order to visually separate clusters of points which appear to be overlapping in the two-dimensional maps, as discussed below.

We note that MDS is not a clustering method, as the clusters are defined beforehand by the coloring scheme used (blue for *H. sapiens*, green for *E. coli*, and so on). MDS simply tries to display visually the interrelationships between the given items, based on the pairwise distances in the distance matrix which is its input. Note also that an increase in dimensionality from 2 to 3 can lead to a better cluster visualization. For example, if we compare the two-dimensional and the three-dimensional Molecular Distance Maps obtained using DSSIM, we see that points that appeared to be erroneously mixed with each other in the two-dimensional map, Figure 4.2(a), (*S. cerevisiae* and *P. falciparum* sequences mixed in with *A. thaliana* sequences) are in

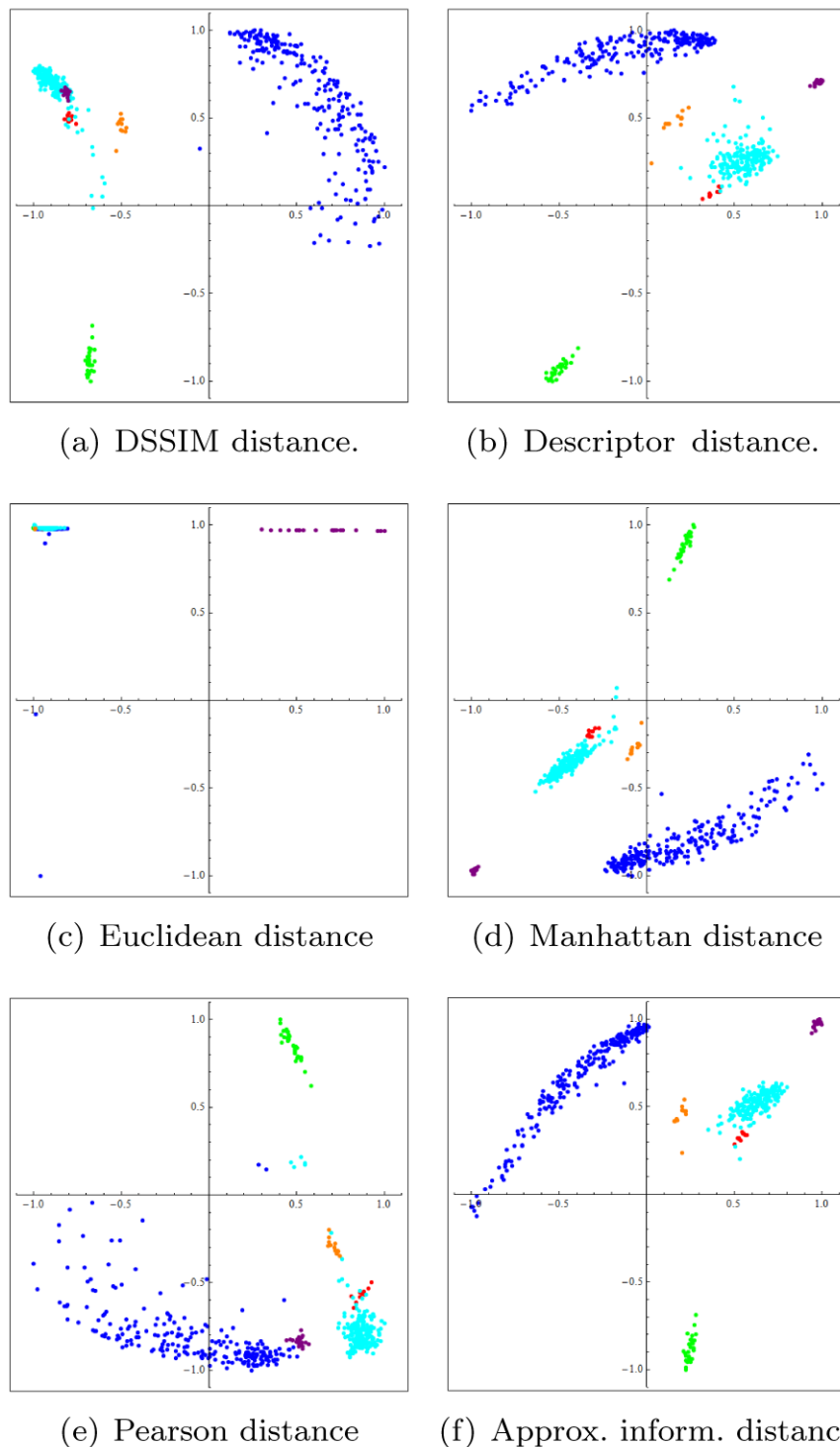


Figure 4.2: The first experiment: Two-dimensional Molecular Distance Maps of 150 kbp genomic sequences spanning one complete chromosome from each of six organisms, representing all kingdoms of life. The MoD Maps were obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. Each point corresponds to one 150 kbp genomic sequence from: *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange).

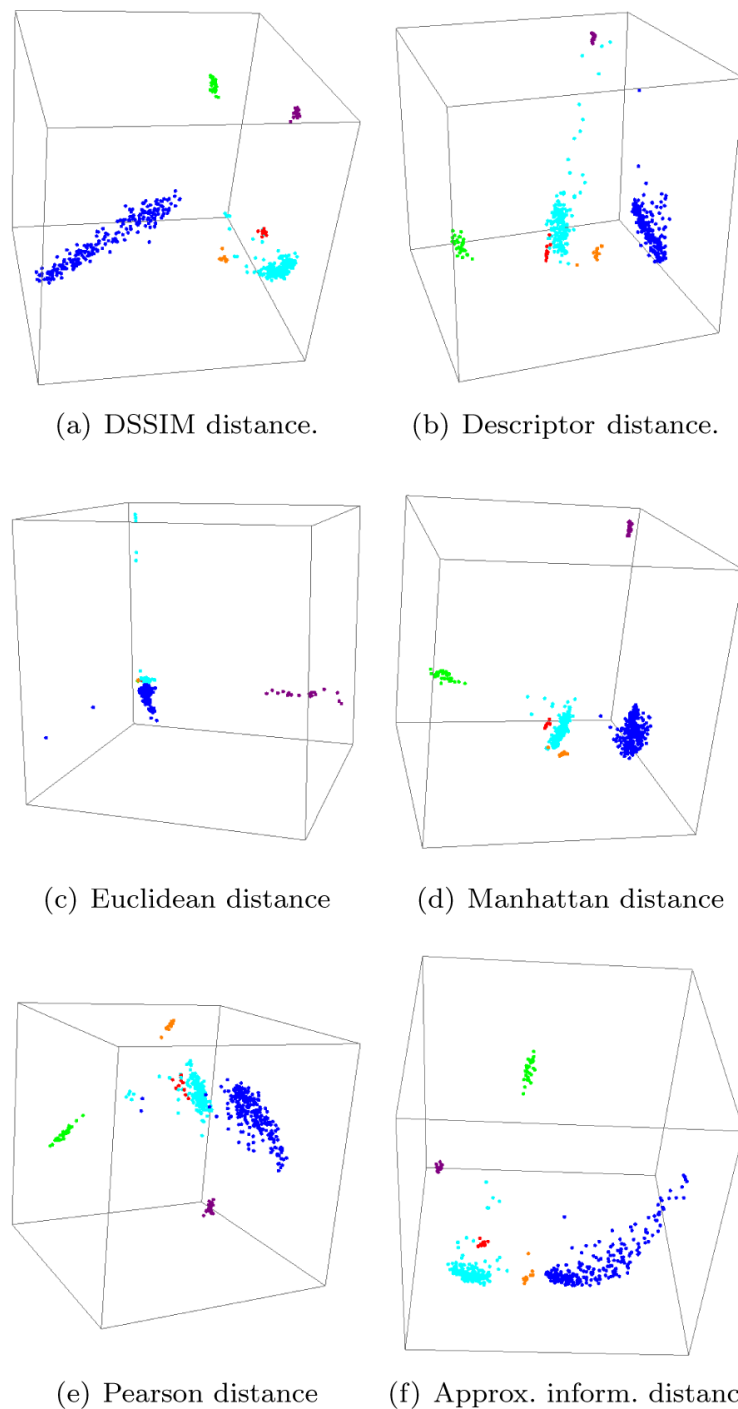


Figure 4.3: The first experiment: Three-dimensional Molecular Distance Maps of 150 kbp genomic sequences spanning one complete chromosome from each of six organisms, representing all kingdoms of life. The MoD Maps were obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. Each point corresponds to one 150 kbp genomic sequences from: *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange).

fact clearly separated from each other in Figure 4.3(a), the three-dimensional version of the Molecular Distance Map.

Figure 4.4 displays the histograms of the pairwise intragenomic distances (dark blue and turquoise) and intergenomic distances (grey) of DNA sequences from *H. sapiens* and *A. thaliana*, obtained using each of the six distances. As noted, some distances seem to perform better than others. Visually, the poorest performer for these two sets of sequences (from *H. sapiens* and *A. thaliana*) seems to be the Euclidean distance wherein the intragenomic distances are as high as intergenomic distances, and no separation is visible. In contrast, DSSIM gives – for the same data – intergenomic distances that are overall much higher than intragenomic distances, resulting in a clear classification of DNA sequences into the species they belong to.

Table 4.3 displays the mean and standard deviation of distances between clusters C_i and C_j , $1 \leq i, j \leq 6$, where a cluster C_ℓ is defined as the set of all genomic sequences from the genome of organism ℓ , as labelled in Table 4.1. In each subtable, the diagonals represent the means and standard deviation for intragenomic distances, while the other entries are all intergenomic distances.

From this table we see that for DSSIM, Manhattan and approximated information distance, the maximum of all the averages of intragenomic distances in this dataset is strictly smaller than the minimum of all the averages of intergenomic distances. For the descriptor distance and Pearson distance the previous statement does not hold but, for each pair of organisms, the two averages of intragenomic distances (e.g., *H. sapiens* - *H. sapiens* and *A. thaliana* - *A. thaliana*) are both lower than the average of the intergenomic distances (*H. sapiens* - *A. thaliana*). For the Euclidean distance, none of the previous statements holds: For example, the average of the *A. thaliana* - *A. thaliana* intragenomic distances (element 4-4 in the Euclidean distance subtable of Table 4.3) is 723, a value which is larger than 672, the average of the *S. cerevisiae* - *A. thaliana* intergenomic distances (element 3-4 in the Euclidean distance subtable of Table 4.3). The complete histograms of all pairwise comparisons $C_i - C_j$ can be found in [59], Appendix C.

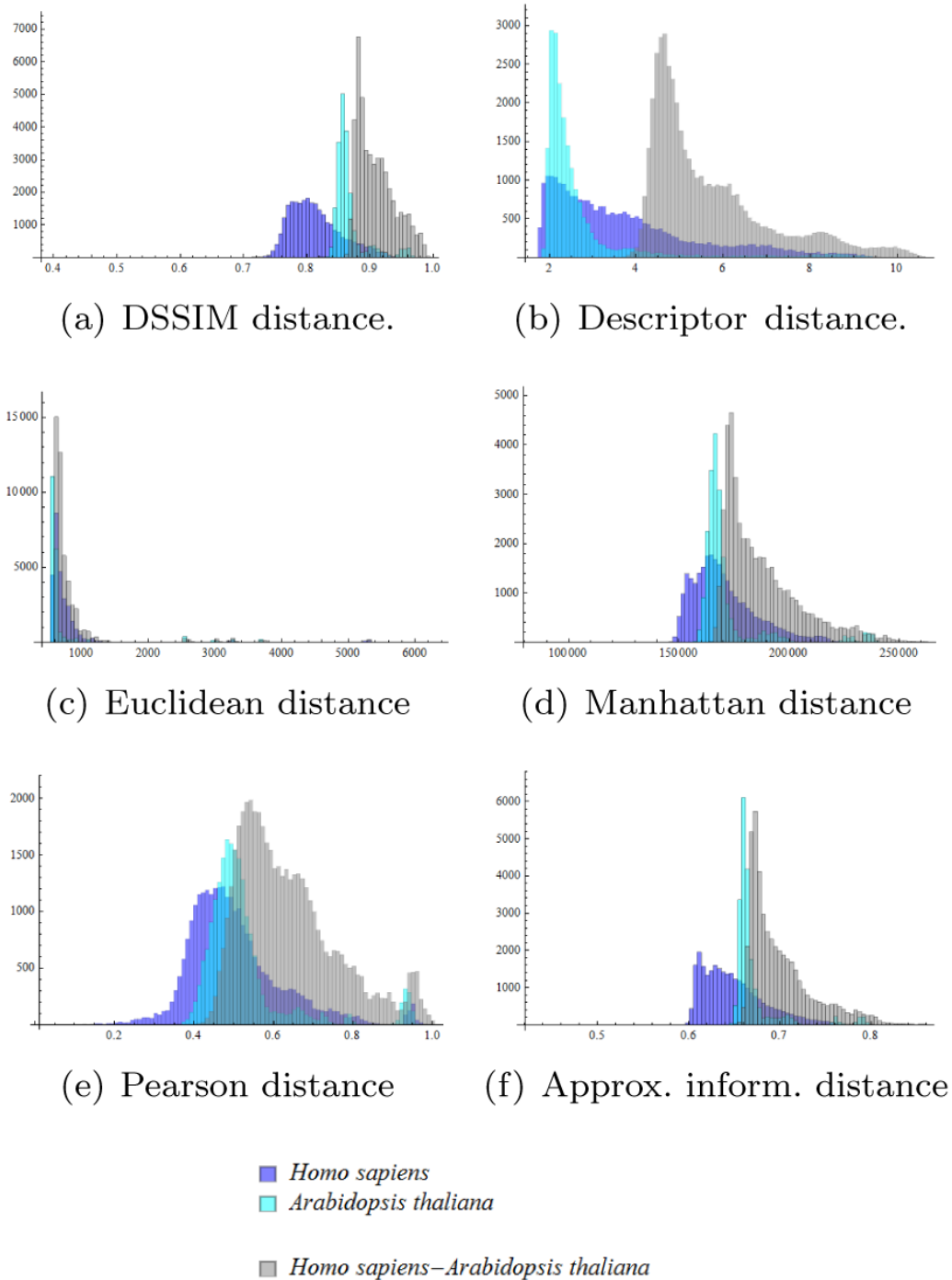


Figure 4.4: The first experiment (150 kbp fragments spanning one complete chromosome per each of the six organisms): Histograms of pairwise intragenomic and intergenomic distances among the DNA sequences from *H. sapiens* and *A. thaliana*.

-	1	2	3	4	5	6	-	1	2	3	4	5	6
1	0.81 ± 0.04	0.99 ± 0.01	0.92 ± 0.02	0.91 ± 0.03	0.92 ± 0.03	0.91 ± 0.02	1	171 ± 15	222 ± 5	189 ± 13	188 ± 17	213 ± 20	191 ± 9
2	-	0.85 ± 0.01	0.97 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.	2	-	175 ± 2	209 ± 4	219 ± 8	252 ± 4	218 ± 3
3	-	-	0.87 ± 0.01	0.89 ± 0.02	0.91 ± 0.	0.91 ± 0.01	3	-	-	171 ± 2	177 ± 10	206 ± 2	184 ± 2
4	-	-	-	0.87 ± 0.03	0.9 ± 0.02	0.91 ± 0.01	4	-	-	-	172 ± 16	200 ± 11	188 ± 9
5	-	-	-	-	0.74 ± 0.01	0.94 ± 0.	5	-	-	-	-	105 ± 3	224 ± 2
6	DSSIM					0.83 ± 0.01	6	Manhattan (in thousands)					167 ± 3
1	3.76 ± 1.69	9.74 ± 0.66	5.92 ± 1.14	5.71 ± 1.41	9.33 ± 1.23	5.44 ± 0.92	1	0.5 ± 0.12	0.97 ± 0.02	0.69 ± 0.1	0.64 ± 0.12	0.65 ± 0.09	0.81 ± 0.06
2	-	2.5 ± 0.28	8.05 ± 0.39	9.1 ± 0.55	12.67 ± 0.19	9.38 ± 0.41	2	-	0.71 ± 0.02	0.93 ± 0.02	0.96 ± 0.02	0.98 ± 0.01	0.99 ± 0.02
3	-	-	2.12 ± 0.08	3.42 ± 1.05	9.48 ± 0.31	4.6 ± 0.09	3	-	-	0.6 ± 0.02	0.6 ± 0.07	0.71 ± 0.03	0.75 ± 0.02
4	-	-	-	2.75 ± 1.33	8.23 ± 0.94	4.94 ± 0.76	4	-	-	-	0.53 ± 0.11	0.63 ± 0.09	0.76 ± 0.04
5	-	-	-	-	1.53 ± 0.14	9.99 ± 0.28	5	-	-	-	-	0.02 ± 0.01	0.94 ± 0.01
6	Descriptor					2.4 ± 0.32	6	Pearson					0.64 ± 0.03
1	756 ± 498	856 ± 349	756 ± 361	818 ± 514	3914 ± 510	812 ± 356	1	0.65 ± 0.03	0.78 ± 0.01	0.7 ± 0.03	0.7 ± 0.03	0.76 ± 0.04	0.69 ± 0.02
2	-	558 ± 5	674 ± 17	802 ± 366	4102 ± 466	696 ± 18	2	-	0.67 ± 0.	0.75 ± 0.01	0.77 ± 0.02	0.85 ± 0.01	0.77 ± 0.01
3	-	-	564 ± 11	672 ± 383	3964 ± 472	633 ± 20	3	-	-	0.67 ± 0.01	0.68 ± 0.02	0.74 ± 0.	0.69 ± 0.
4	-	-	-	723 ± 535	3923 ± 506	748 ± 372	4	-	-	-	0.67 ± 0.03	0.73 ± 0.02	0.69 ± 0.02
5	-	-	-	-	999 ± 276	4085 ± 468	5	-	-	-	-	0.64 ± 0.01	0.76 ± 0.01
6	Euclidean					585 ± 24	6	Approx. Information					0.65 ± 0.01

Table 4.3: The first experiment: Mean and standard deviation of distances between clusters $C_i - C_j$ for $i, j = 1, \dots, 6$.

To maximize the diversity within each species, we performed a second experiment, with similar parameters as the first, but in which the fragments analyzed were randomly sampled from the entire genomes. The Molecular Distance Maps for this experiment are presented in Figure 4.5 and Figure 4.6. Note that the separation of sequences by the organism they belong to is even more clear than in the previous experiment, that used one complete chromosome from each organism. This suggests that (for this dataset), the CGR pattern is a genome-wide characteristic.

4.3.1 Quality measures for distances

In this section we present three quality measures that each evaluates the quality of the six distances considered. In the data mining literature a wide range of quality measures for a given clustering has been defined; see for example [69, 70]. Most of these measures are designed to assess the quality of different automated clustering methods while using the same distance. Our set-up is different, as we use different distances while the clustering is fixed and given by the initial colour-coding of the sequence-representing points. Thus, we have to use other approaches to compare the distances we analyze. In particular, as the six distances have different ranges, we have to use assessment methods which are invariant to the scale of the distance.

The “ground-truth” that we use as a basis for our distance assessment is the fact that the “ideal” clustering of DNA sequences and the points that represent them is known: sequences from the same organism should be close to one another and far from sequences originating from other organisms. (This assumption is justified – for this dataset – as the six organisms considered are very different from one another, belonging to different kingdoms of life.) Thus, an optimal distance should yield a relatively small value for two FCGRs which were generated from the DNA sequences originating from the same organism, and relatively high values for two FCGRs originating from DNA sequences coming from different organisms.

In order to assess each of the six distances quantitatively, we computed three quality measures which rate different features of a distance:

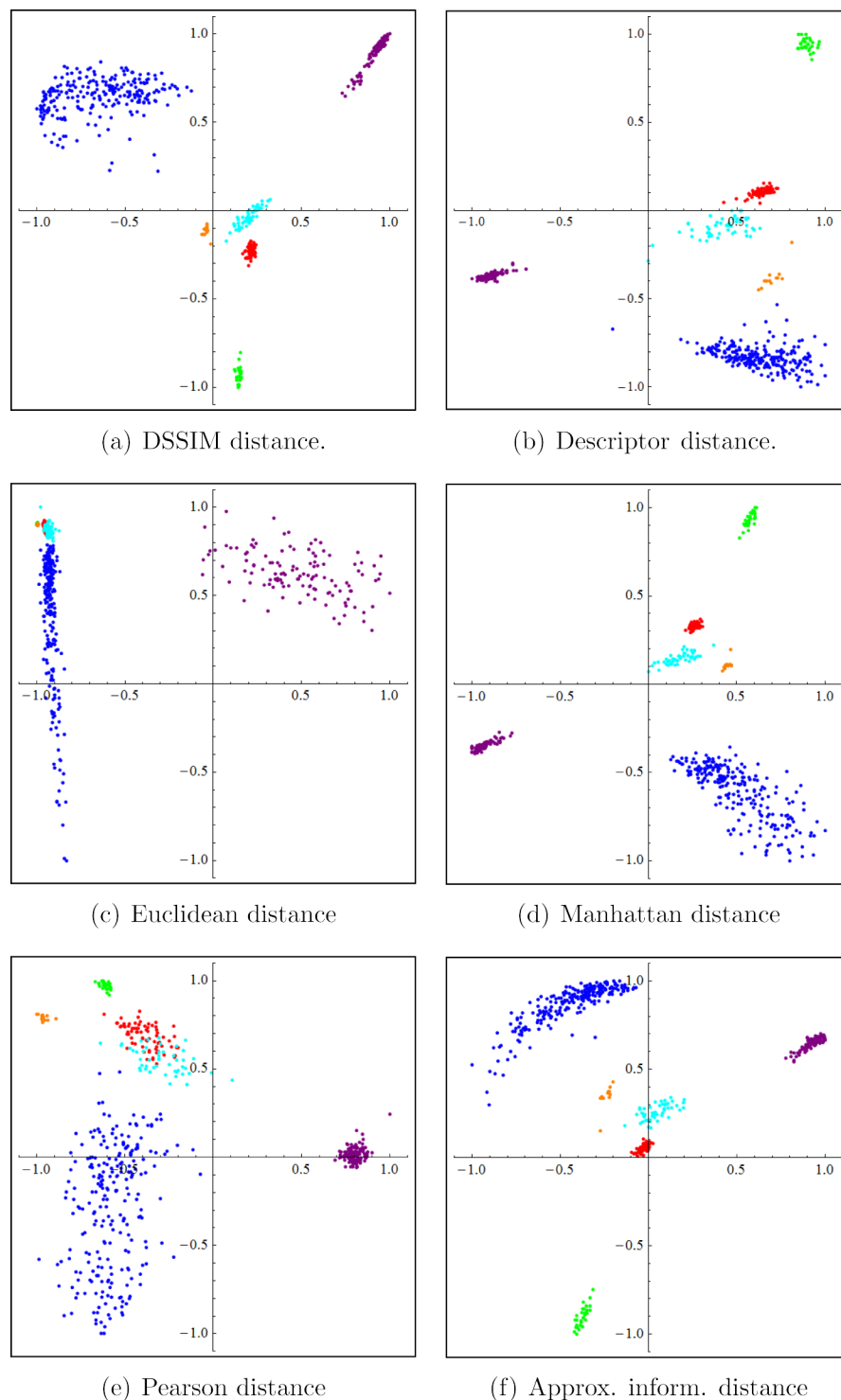


Figure 4.5: The second experiment: Two-dimensional Molecular Distance Maps of DNA genomic sequences sampled from the entire genomes of all six organisms, obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. The dataset consists of 10 randomly sampled fragments from each chromosome of multi-chromosome genomes, and all complete fragments from the genomes of *E. coli* and *P. furiosus*, for a total of 526 fragments. Each point corresponds to one such 150 kbp fragment from *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange).

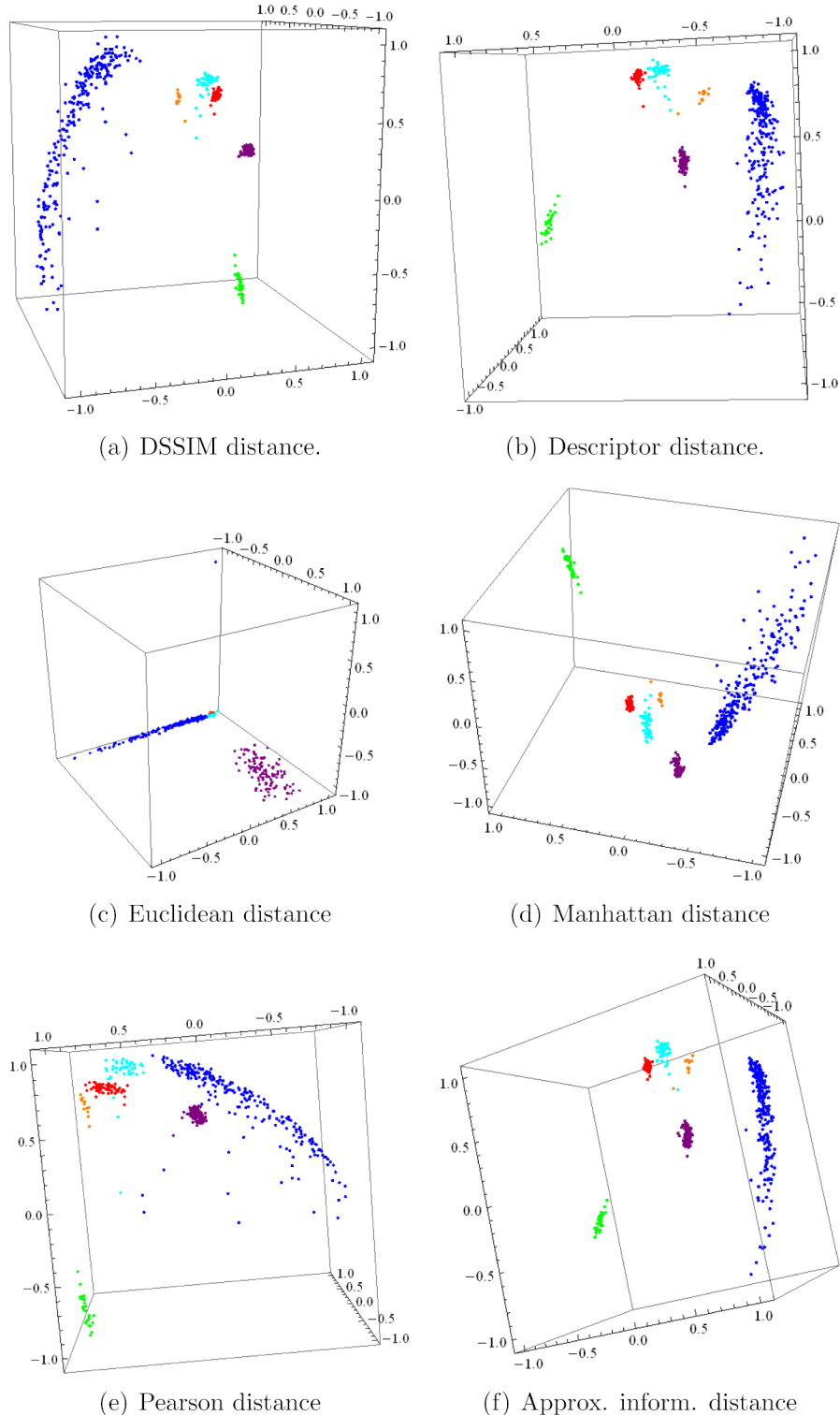


Figure 4.6: The second experiment: Three-dimensional Molecular Distance Maps of genomic DNA sequences sampled from the genomes of all six chosen organisms, obtained using DSSIM, descriptor, Euclidean, Manhattan, Pearson and approximated information distance, respectively. The dataset consists of 10 randomly sampled fragments from each chromosome of multi-chromosome genomes, and all complete fragments from the genomes of *E. coli* and *P. furiosus*, for a total of 526 fragments. Each point corresponds to one such 150 kbp fragment from *H. sapiens* (blue), *E. coli* (green), *S. cerevisiae* (red), *A. thaliana* (turquoise), *P. falciparum* (magenta), and *P. furiosus* (orange).

- the correlation to an idealized cluster distance
- the silhouette cluster accuracy
- the relative overlap between the intragenomic and intergenomic distance histograms.

Let us stress that all three quality measures of the six distances are based on the distance matrices which we computed and not on their MDS plots. We will define the three quality measures such that their expected values range in the interval $[0, 1]$ where higher values correspond to better performance.

Let us first describe the three quality measures informally. An idealized distance is a distance that would be able to differentiate DNA sequences by species, that is, a distance δ for which $\delta(x, y) = 0$ if x and y are sequences from the same species and $\delta(x, y) = 1$ otherwise. The first quality measure, the *correlation to an idealized cluster distance*, measures how well a distance is linearly correlated to the idealized distance δ . The second quality measure, *silhouette cluster accuracy*, is the percentage of points that are best embedded in the cluster they belong to. The third quality measure quantifies the “visual overlap” between the intragenomic and intergenomic distance histograms. Given our dataset, it is reasonable to expect that a good distance gives a low value if applied to FCGRs of genomic sequences of the same organism, and a high value when applied to FCGRs of genomic sequences from two different organisms, thus separating the histograms of intragenomic distances from that of intergenomic distances. This is illustrated by the histograms in Figure 4.4, where a high overlap between the graph of intragenomic distances (dark blue and turquoise) and the graphs of intergenomic distances (grey) is an indication of a poorly performing distance. In a theoretically optimal situation, there would exist a value c such that all distances that are smaller than c are intragenomic distances and all distances that are larger than c are intergenomic distances. This can usually not be expected from real data, but a low overlap between histograms is nevertheless indicative of a “good” distance.

In order to formally define the three quality measures, we consider a dataset V which is

partitioned into p non-overlapping clusters C_1, \dots, C_p for which a distance $d_\alpha: V \times V \rightarrow \mathbb{R}_{\geq 0}$ exists. The cardinalities of the sets are $|V| = m$ and $|C_i| = m_i$ for $i = 1, \dots, p$. In our analysis, $p = 6$ and C_1 contains all FCGRs generated from genomic DNA sequences from *H. sapiens*, C_2 contains all FCGRs generated from genomic sequences of *E. coli*, and so on, according to the order in Table 4.1. The distance d_α is one of the six distances $\alpha \in \{\text{DSSIM, D, E, M, P, AID}\}$.

The *correlation to an idealized cluster distance* is computed as follows. We define the *idealized cluster distance* as a function (or matrix) $\delta: V \times V \rightarrow \{0, 1\}$ such that $\delta(x, y) = 0$ if and only if x and y belong to the same cluster, and $\delta(x, y) = 1$ otherwise. Because we can view d_α and δ as discrete, symmetric functions which have the same domain, we can compute their correlation coefficient. We define the correlation of δ to d_α to be the Pearson correlation of δ and d_α . More precisely, the upper triangular part of the matrix corresponding to a distance d_α is interpreted as a vector (x_1, \dots, x_n) and compared with the corresponding values (y_1, \dots, y_n) given by δ . We obtain the δ -correlation as

$$\mathcal{D}_\alpha = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

The correlation ranges in the interval $[-1, 1]$: a value of 1 means that d_α and δ are linearly correlated, and a value of 0 means that they are unrelated. In other words, if the value obtained by measuring the *correlation* of a given distance *to the idealized cluster distance* is close to 1, this means that the given distance is closer to the idealized cluster distance, and hence, performs well. Note that negative values for this measure are not expected as this would imply that d_α and δ were negatively related (d_α would perform worse than a matrix containing random entries).

The *silhouette cluster accuracy* is based on the *silhouette coefficient* defined in [71] as a measure that determines how well a single point is embedded in the cluster to which it belongs. For a point x from cluster C_i we define a_x as the average distance of this point to all other points in C_i , that is,

$$a_x = \frac{1}{m_i - 1} \sum_{y \in C_i, y \neq x} d_\alpha(x, y),$$

and we define b_x as the minimum over the average distances of x to all points of a different cluster

$$b_x = \min_{j=1, j \neq i}^K \left\{ \frac{1}{m_j} \sum_{y \in C_j} d_\alpha(x, y) \right\}.$$

The silhouette coefficient of x is defined as

$$S_\alpha(x) = \frac{b_x - a_x}{\max\{a_x, b_x\}}.$$

If a point x has a silhouette coefficient $S_\alpha(x) \leq 0$, then x is at least as close to a cluster to which it does not belong than to its own cluster. The *silhouette cluster accuracy* \mathcal{A}_α denotes the percentage of points with a silhouette coefficient greater than 0, that is the percentage of points which are well-embedded in their own cluster,

$$\mathcal{A}_\alpha = \frac{|\{x \in V \mid S_\alpha(x) > 0\}|}{m}.$$

Obviously, the silhouette cluster accuracy ranges in $[0, 1]$ with a high accuracy being desirable.

For assessing the *relative overlap* of the histograms, consider any two clusters C_i and C_j with $i \neq j$ (for example, C_1 is the *H. sapiens* cluster and C_4 the *A. thaliana* cluster). We compare the two sets of intragenomic distances C_i-C_i and C_j-C_j with the set of intergenomic distances C_i-C_j . For a distance d_α , we divide the range from $\min(d_\alpha)$ to $\max(d_\alpha)$ in this dataset into 100 bins of size $r = \frac{\max(d_\alpha) - \min(d_\alpha)}{100}$ and count the distances which fall into this bin: $c_{i,i}[\ell]$ denotes bin ℓ containing distances from C_i-C_i and $c_{i,j}[\ell]$ denotes bin i containing distances from C_i-C_j . For $\ell = 1, \dots, 100$ we let

$$c_{i',j'}[\ell] = |\{\{x, y\} \mid x \in C_{i'}, y \in C_{j'} \text{ and } x \neq y \\ \text{and } (\ell - 1) \cdot r < d_\alpha(x, y) \leq \ell \cdot r\}|.$$

By $s_{i',j'}$ we denote the sum over all $c_{i',j'}$ -bins, that is, $s_{i',j'} = \sum_{\ell=1}^{100} c_{i',j'}[\ell]$. We define the relative overlap $O_\alpha(i, j)$ of $C_i - C_i$ (intragenomic distances) with $C_i - C_j$ (intergenomic distances) as

$$O_\alpha(i, j) = \frac{\max\{s_{i,i}, s_{i,j}\}}{\min\{s_{i,i}, s_{i,j}\}} \cdot \frac{\sum_{i=1}^{100} \min\{c_{i,i}, c_{i,j}\}}{\sum_{i=1}^{100} \max\{c_{i,i}, c_{i,j}\}}.$$

The relative overlap $O_\alpha(j, i)$ of $C_j - C_j$ with $C_i - C_j$ is defined analogously; note that $O_\alpha(i, j) \neq O_\alpha(j, i)$ in general. The overlap is normalized to the range $[0, 1]$ where 0 means no overlap of elements of bins between intra- and intergenomic distances, and 1 means that one of the histograms completely “covers” the other. Also note that we are not interested in the overlap of $C_i - C_i$ with $C_j - C_j$ as both sets of distances are intragenomic distances.

Since we intend to define a quality measure where a value close to 1 should represent a small overlap, we will use $1 - O_\alpha(i, j)$. Furthermore, we combine these quantities for all possible pairs of clusters C_i and C_j , obtaining *the relative overlap* as:

$$O_\alpha = 1 - \frac{1}{p(p-1)} \sum_{i=1}^p \sum_{j=1, i \neq j}^p O_\alpha(i, j).$$

For example, in Figure 4.4, for each of the considered distance, the dark blue histograms depict the $C_1 - C_1$ (*H. sapiens* – *H. sapiens*) intragenomic distances, the turquoise histograms the $C_4 - C_4$ (*A. thaliana* – *A. thaliana*) intragenomic distances, and grey histograms the $C_1 - C_4$ (*H. sapiens* – *A. thaliana*) intergenomic distances. As seen from this figure, the descriptor distance appears to visually perform best at separating the two intragenomic distance histograms from the intergenomic histogram, while the Euclidean distance has the weakest performance. The relative overlap attempts to quantify this by computing the overlaps of each of the two

pairs of histograms (dark blue with grey, and turquoise with grey). Note that small visual histogram overlaps will result in a high numerical *relative overlap*, and is indicative of a better performing distance.

4.3.2 Distance comparison results

For the first experiment (one complete chromosome from each organism) the results of ranking the six distances, using the three quality measures, are listed in Table 4.4. Recall that all quality measures have an expected range of $[0, 1]$ where larger values imply better performance.

	\mathcal{D}_α	\mathcal{A}_α	\mathcal{O}_α	z -score sum	Rank
DSSIM	0.627	1.000	0.965	1.895	2nd
Descriptor	0.639	0.976	0.988	2.509	1st
Euclidean	0.231	0.325	0.907	-4.831	6th
Manhattan	0.527	1.000	0.951	0.84	3rd
Pearson	0.536	0.980	0.888	-0.875	5th
Approx. Inf.	0.527	1.000	0.937	0.462	4th

Table 4.4: The first experiment: Summary of quality measures for the performances of six distances (DSSIM, descriptor, Euclidean, Manhattan, Pearson, approximated information distance) on a dataset of 508 genomic DNA sequences spanning one complete chromosome for multi-chromosomes organisms and the complete genome otherwise, of one organism from each kingdom of life. \mathcal{D}_α is the correlation to an idealized cluster, \mathcal{A}_α the silhouette cluster accuracy, and \mathcal{O}_α the relative overlap. Higher is better.

To compare each distance relative to all the other distances, we compute for each quality measure (each column) the *standard scores* (z -scores) of each distance d_α , where $\alpha \in \{\text{DSSIM, D, E, M, P, AID}\}$, as $z(d_\alpha) = \frac{d_\alpha - \mu}{\sigma}$ where μ is the mean and σ is the deviation for that particular quality measure (column). A positive value of the standard score will mean that a distance performs above average (in this category) and a negative value that it performs below average. Finally, we compute the sum of the z -scores for each quality measure as seen in Table 4.4, second last column. Note that the total of z -scores for a distance represents the performance of that distance relative to the other distances, and indicates its relative ranking.

Table 4.5 contains the results of the distance comparison for the second experiment, that

sampled 10 fragments from each chromosome. Interestingly, the ranking of distances is the same for both experiments.

	\mathcal{D}_α	\mathcal{A}_α	\mathcal{O}_α	z-score sum	Rank
DSSIM	0.729	1.000	0.964	1.980	2nd
Descriptor	0.726	0.998	0.984	2.336	1st
Euclidean	0.438	0.608	0.861	-5.292	6th
Manhattan	0.662	1.000	0.955	1.172	3rd
Pearson	0.639	0.949	0.875	-0.954	5th
Approx. Inf.	0.637	1.000	0.946	0.759	4th

Table 4.5: The second experiment: Summary of quality measures for the performances of six distances (DSSIM, descriptor, Euclidean, Manhattan, Pearson, approximated information distance) on a dataset of 526 genomic DNA sequences sampled randomly (10 fragments per chromosome for multi-chromosome organisms, and all fragments of the genome otherwise) from the genomes of organisms from each kingdom of life. \mathcal{D}_α is the correlation to an idealized cluster, \mathcal{A}_α the silhouette cluster accuracy, and \mathcal{O}_α the relative overlap. Higher is better.

The conclusion of these analyses is that the best performing distances for this dataset are the descriptor distance and DSSIM. The Manhattan, Pearson, and approximate information distances perform well in some categories but not so well in other categories. For this dataset and value of k , the Euclidean distance had the weakest performance in all measured categories, which confirms the visual assessment of the MDS plots obtained by using the Euclidean distance, as seen in Figure 4.2 and Figure 4.3.

It is worth noting that the two distances which perform best (DSSIM and descriptor) treat FCGR matrices as two-dimensional maps in which the local arrangement of the cells (matrix entries) influences the computed distance, whereas the other distances treat the FCGR matrices as linear vectors. This suggests that the organization of the k -mer tallies (in this paper $k = 9$) of a DNA sequence as an FCGR matrix, rather than a simple vector, reveals structural properties of the DNA sequence that could be utilized in order to identify and classify genomic DNA sequences.

4.4 Discussion and Conclusions

In this study we test, at the kingdom level, the hypothesis that CGR-based genomic signatures of genomic DNA sequences are indeed species and genome-specific. With this goal in mind we first analyzed over five hundred 150 kbp DNA genomic sequences spanning one complete chromosome from each of six organisms, representing all kingdoms of life. We then separately analyzed over five hundred 150 kbp genomic sequences randomly sampled from the complete genomes of all organisms considered.

Our quantitative comparison of six different distances suggests that several other distances outperform the Euclidean distance, which has been until now almost exclusively used in such studies. Our preliminary results show that two of these distances, DSSIM and descriptor distance (introduced here) when applied to CGR-based genomic signatures, have indeed the ability to differentiate between DNA sequences coming from different species at this taxonomic level. This indicates that the k -mer sequence composition (where $k = 1, 2, \dots, 9$) of genomic sequences contains taxonomic information which could potentially aid in the identification, comparison and classification of species based on molecular evidence. The two-dimensional and three-dimensional Molecular Distance Maps we obtain, which visualize the simultaneous intragenomic and intergenomic interrelationships among the sequences in our dataset, show this method's potential.

Further analysis is needed to explore this method's applicability to the genomic species identification and classification at lower taxonomic levels. As a preview experiment, we applied it to 240 fragments, randomly sampled from the entire genome of *H. sapiens* (10 fragments per chromosome), and 210 fragments randomly sampled from the entire genome of *M. musculus* (10 fragments per chromosome). See [59], Appendix B, for dataset details.

The Molecular Distance Maps of these 450 DNA sequences, 150 kbp each (see Figure 4.7 and Figure 4.8) suggest that several of the distances are able to differentiate between DNA sequences at lower taxonomic levels. As seen in Table 4.6, the Euclidean distance was again outperformed by other distances, when assessed with the quality measures we described. How-

ever, we note a change in the distance rankings, with Pearson and DSSIM ranking first and respectively second, and the descriptor distance ranking last. This may be because the descriptor distance is able to identify large pattern-differences between CGR images, which may be more suitable when comparing genomic sequences at high taxonomic levels, while DSSIM is good at picking up subtle differences between similar CGR images and thus it may be better suited to comparing genomic sequences from more closely related species. Overall, this suggests that different distances may have to be chosen, depending on the taxonomic level of the analysis.

	\mathcal{D}_α	\mathcal{A}_α	\mathcal{O}_α	z-score sum	Rank
DSSIM	0.422	1.000	0.618	3.014	2nd
Descriptor	0.032	0.560	0.063	-3.347	6th
Euclidean	0.079	0.658	0.318	-1.558	4th
Manhattan	0.209	0.969	0.336	0.601	3rd
Pearson	0.531	0.993	0.647	3.643	1st
Approx. Inf.	0.101	0.578	0.195	-2.353	5th

Table 4.6: The preview experiment: Summary of quality measures for the performances of six distances (DSSIM, descriptor, Euclidean, Manhattan, Pearson, approximated information distance) on a dataset of 450 DNA sequences, sampled from the entire genome (10 fragments per chromosome) of *H. sapiens* and *M. musculus*. \mathcal{D}_α is the correlation to an idealized cluster, \mathcal{A}_α is the silhouette cluster accuracy, and \mathcal{O}_α is the relative overlap. Higher is better.

Further large-scale computational experiments have to be carried out to confirm these preliminary results and establish their validity, as well as to establish the applicability of this method to genomic sequences identification and classification at lower taxonomic levels. Such experiments could provide additional insights regarding the choice of optimal distance for structural genomic sequence comparisons in different settings.

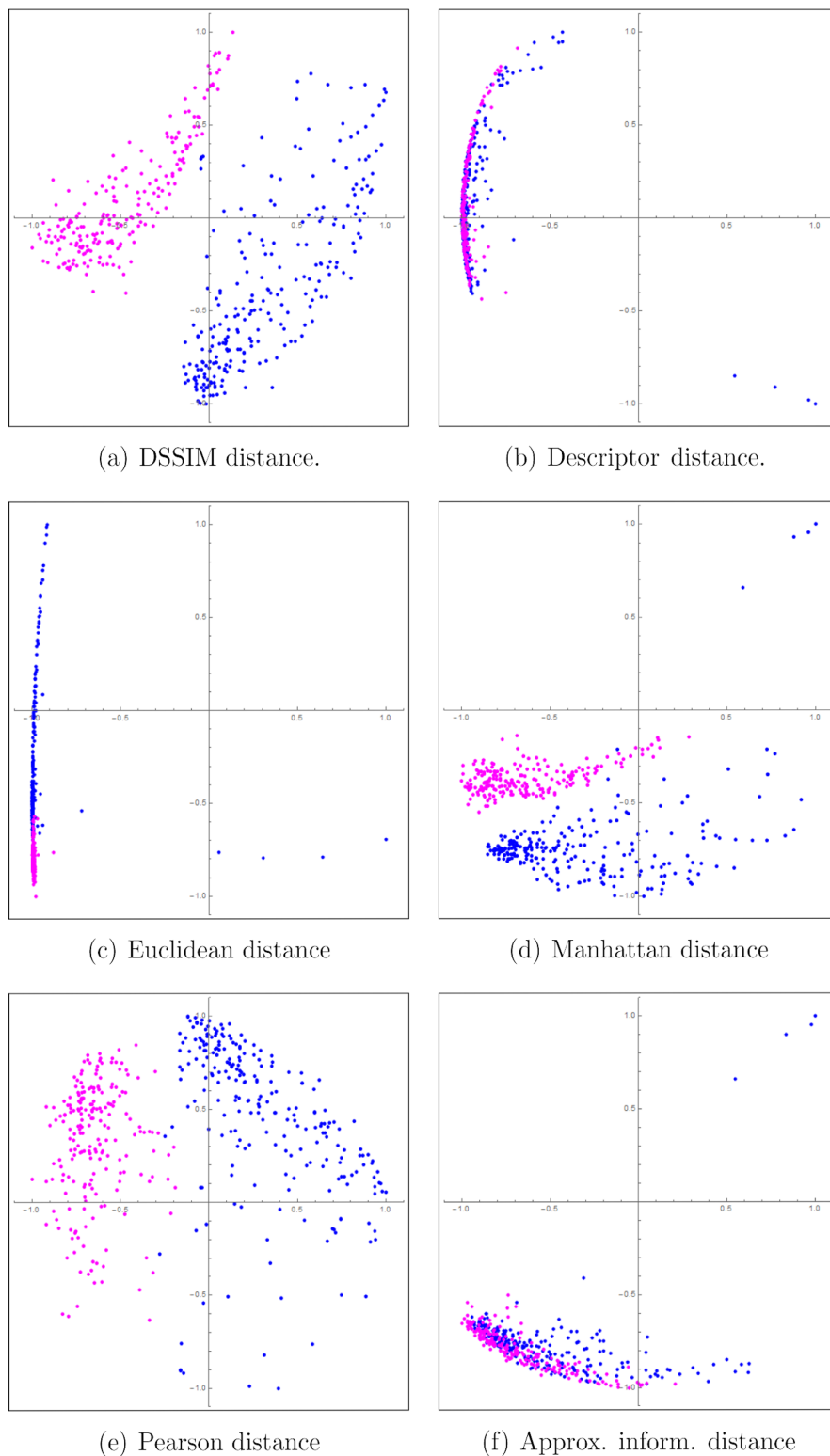
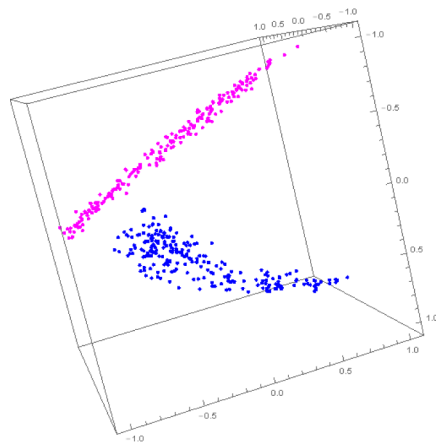
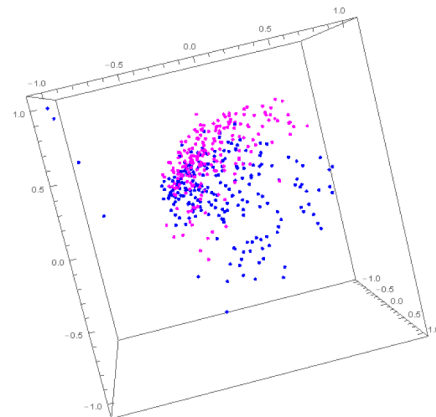


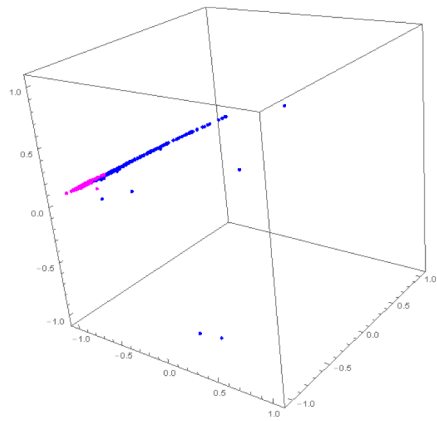
Figure 4.7: The preview experiment: Two-dimensional Molecular Distance Maps of 150 kbp genomic DNA sequences, randomly sampled from each chromosome (10 fragments per chromosome) of *H. sapiens* (blue), *M. musculus* (fuchsia) using the six distances.



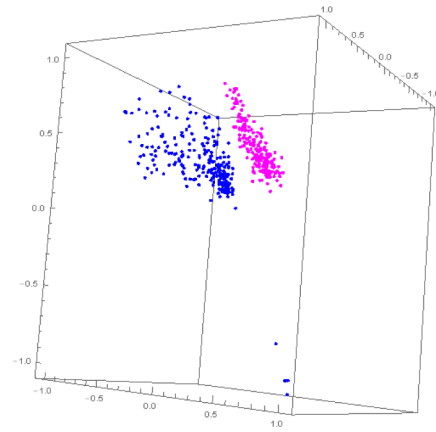
(a) DSSIM distance.



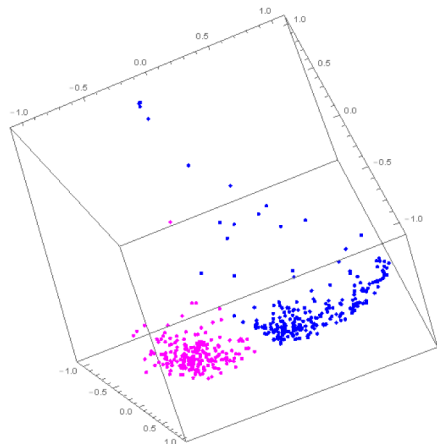
(b) Descriptor distance.



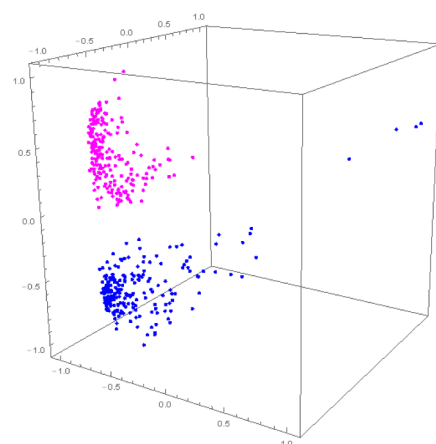
(c) Euclidean distance



(d) Manhattan distance



(e) Pearson distance



(f) Approx. inform. distance

Figure 4.8: The preview experiment: Three-dimensional Molecular Distance Maps of 150 kbp genomic DNA sequences, randomly sampled from each chromosome (10 fragments per chromosome) of *H. sapiens* (blue), *M. musculus* (fuchsia) using the six distances.

Bibliography

- [1] Hebert, P.D., Cywinska, A., Ball, S.L., *et al.*: Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**(1512), 313–321 (2003)
- [2] Sirovich, L., Stoeckle, M.Y., Zhang, Y.: Structural analysis of biodiversity. *PLoS One* **5**(2), 9266 (2010)
- [3] Jeffrey, H.: Chaos game representation of gene structure. *Nucleic Acids Research* **18**(8), 2163–2170 (1990)
- [4] Deschavanne, P., Giron, A., Vilain, J., Fagot, G., Fertil, B.: Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution* **16**(10), 1391–1399 (1999)
- [5] Karlin, S., Burge, C.: Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* **11**(7), 283–290 (1995)
- [6] Jeffrey, H.: Chaos game visualization of sequences. *Computers & Graphics* **16**(1), 25–33 (1992)
- [7] Hill, K., Schisler, N., Singh, S.: Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *Journal of Molecular Evolution* **35**(3), 261–9 (1992)
- [8] Hill, K., Singh, S.: Evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes. *Genome* **40**, 342–356 (1997)
- [9] Deschavanne, P., Giron, A., Vilain, J., Dufraigne, C., Fertil, B.: Genomic signature is preserved in short DNA fragments. In: *Proceedings of IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pp. 161–167 (2000)

- [10] Edwards, S., Fertil, B., Girron, A., Deschavanne, P.: A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic Biology* **51**(4), 599–613 (2002)
- [11] Wang, Y., Hill, K., Singh, S., Kari, L.: The spectrum of genomic signatures: From dinucleotides to chaos game representation. *Gene* **346**, 173–185 (2005)
- [12] Kari, L., Hill, K.A., Sayem, A.S., Karamichalis, R., Bryans, N., Davis, K., Dattani, N.S.: Mapping the Space of Genomic Signatures. *PLoS One* **10**(5) (2015)
- [13] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
- [14] Iversen, G.R., Gergen, M., Gergen, M.M.: *Statistics: The Conceptual Approach*. Springer, Berlin Heidelberg (1997)
- [15] Krause, E.F.: *Taxicab Geometry: An Adventure in Non-Euclidean geometry*. Courier Dover Publications, Mineola, New York (2012)
- [16] Li, M., Chen, X., Li, X., Ma, B., Vitany, P.: The similarity metric. *IEEE Transactions on Information Theory* **50**(12), 3250–3264 (2004)
- [17] Phillips, G.J., Arnold, J., Ivarie, R.: Mono-through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. *Nucleic Acids Research* **15**(6), 2611–2626 (1987)
- [18] Beutler, E., Gelbart, T., Han, J., Koziol, J.A., Beutler, B.: Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proceedings of the National Academy of Sciences* **86**(1), 192–196 (1989)
- [19] Deschavanne, P., Radman, M.: Counterselection of GATC sequences in enterobacteriophages by the components of the methyl-directed mismatch repair system. *Journal of Molecular Evolution* **33**(2), 125–132 (1991)

- [20] Bhagwat, A.S., McClelland, M.: DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. *Nucleic Acids Research* **20**(7), 1663–1668 (1992)
- [21] Burge, C., Campbell, A.M., Karlin, S.: Over-and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences* **89**(4), 1358–1362 (1992)
- [22] Karlin, S., Burge, C., Campbell, A.M.: Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Research* **20**(6), 1363–1370 (1992)
- [23] Blaisdell, B.E., Rudd, K.E., Matin, A., Karlin, S.: Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome: several new groups. *Journal of Molecular Biology* **229**(4), 833–848 (1993)
- [24] Gelfand, M.S., Koonin, E.V.: Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Research* **25**(12), 2430–2439 (1997)
- [25] Karlin, S., Mrazek, J., Campbell, A.M.: Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology* **179**(12), 3899–3913 (1997)
- [26] Vinga, S., Almeida, J.: Alignment-free sequence comparison a review. *Bioinformatics* **19**(4), 513–523 (2003)
- [27] Bonham-Carter, O., Steele, J., Bastola, D.: Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in Bioinformatics* **15**(6), 890–905 (2014)
- [28] Almeida, J.S.: Sequence analysis by iterated maps, a review. *Briefings in Bioinformatics* **15**(3), 369–375 (2014)

- [29] Blaisdell, B.E.: A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences* **83**(14), 5155–5159 (1986)
- [30] Sitnikova, T., Zharkikh, A.: Statistical analysis of L-tuple frequencies in eubacteria and organelles. *Biosystems* **30**(1), 113–135 (1993)
- [31] Wu, T.-J., Burke, J.P., Davison, D.B.: A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, 1431–1439 (1997)
- [32] Wu, T.-J., Hsieh, Y.-C., Li, L.-A.: Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics* **57**(2), 441–448 (2001)
- [33] Stuart, G.W., Moffett, K., Baker, S.: Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics* **18**(1), 100–108 (2002)
- [34] Qi, J., Wang, B., Hao, B.-I.: Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *Journal of Molecular Evolution* **58**(1), 1–11 (2004)
- [35] Pham, T.D., Zuegg, J.: A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* **20**(18), 3455–3461 (2004)
- [36] Pham, T.D.: Spectral distortion measures for biological sequence comparisons and database searching. *Pattern Recognition* **40**(2), 516–529 (2007)
- [37] Kantorovitz, M.R., Robinson, G.E., Sinha, S.: A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* **23**(13), 249–255 (2007)
- [38] Van Helden, J.: Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics* **20**(3), 399–406 (2004)
- [39] Dai, Q., Yang, Y., Wang, T.: Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **24**(20), 2296–2302 (2008)

- [40] Almeida, J.S., Carrico, J.A., Marezek, A., Noble, P.A., Fletcher, M.: Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* **17**(5), 429–437 (2001)
- [41] Almeida, J.S., Vinga, S.: Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics* **3**(1), 6 (2002)
- [42] Almeida, J.S., Vinga, S.: Computing distribution of scale independent motifs in biological sequences. *Algorithms for Molecular Biology* **1**, 18 (2006)
- [43] Almeida, J.S., Vinga, S.: Biological sequences as pictures—a generic two dimensional solution for iterated maps. *BMC Bioinformatics* **10**(1), 100 (2009)
- [44] Feng, J., Hu, Y., Wan, P., Zhang, A., Zhao, W.: New method for comparing DNA primary sequences based on a discrimination measure. *Journal of Theoretical Biology* **266**(4), 703–707 (2010)
- [45] Pandit, A., Dasanna, A.K., Sinha, S.: Multifractal analysis of HIV-1 genomes. *Molecular Phylogenetics and Evolution* **62**(2), 756–763 (2012)
- [46] Pandit, A., Vadlamudi, J., Sinha, S.: Analysis of dinucleotide signatures in HIV-1 subtype B genomes. *Journal of Genetics* **92**(3), 403–412 (2013)
- [47] Pride, D., Meinersmann, R., Wassenaar, T., Blaser, M.: Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research* **13**(2), 145–158 (2003)
- [48] Sandberg, R., Bränden, C.-I., Ernberg, I., Cöster, J.: Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. *Gene* **311**, 35–42 (2003)
- [49] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glöckner, F.O.: TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**(1), 163 (2004)

- [50] Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., Deschavanne, P.: Exploration of phylogenetic data using a global sequence analysis method. *BMC Evolutionary Biology* **5**(1), 63 (2005)
- [51] Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., Deschavanne, P.: Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research* **33**(1), 6 (2005)
- [52] Joseph, J., Sasikumar, R.: Chaos game representation for comparison of whole genomes. *BMC Bioinformatics* **7**(1), 243 (2006)
- [53] Tanchotsrinon, W., Lursinsap, C., Poovorawan, Y.: A high performance prediction of HPV genotypes by chaos game representation and singular value decomposition. *BMC Bioinformatics* **16**(1), 71 (2015)
- [54] Karlin, S., Ladunga, I.: Comparisons of eukaryotic genomic sequences. *Proceedings of the National Academy of Sciences* **91**(26), 12832–12836 (1994)
- [55] Shedlock, A.M., Botka, C.W., Zhao, S., Shetty, J., Zhang, T., Liu, J.S., Deschavanne, P.J., Edwards, S.V.: Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences* **104**(8), 2767–2772 (2007)
- [56] Deschavanne, P., DuBow, M., Regard, C.: The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virology Journal* **7**(1), 163 (2010)
- [57] Pandit, A., Sinha, S.: Using genomic signatures for HIV-1 subtyping. *BMC Bioinformatics* **11**(Suppl 1), 26 (2010)

- [58] Yu, Z.-G., Zhan, X.-W., Han, G.-S., Wang, R.W., Anh, V., Chu, K.H.: Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. *International Journal of Molecular Sciences* **11**(3), 1141–1154 (2010)
- [59] Online Material. https://github.com/rallis/intraSupplemental_Material
- [60] Burma, P.K., Raj, A., Deb, J.K., Brahmachari, S.K.: Genome analysis: a new approach for visualization of sequence organization in genomes. *Journal of Biosciences* **17**(4), 395–411 (1992)
- [61] Dutta, C., Das, J.: Mathematical characterization of chaos game representation: New algorithms for nucleotide sequence analysis. *Journal of Molecular Biology* **228**(3), 715–719 (1992)
- [62] Goldman, N.: Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Research* **21**(10), 2487–2491 (1993)
- [63] Oliver, J., Bernaola-Galvan, P., Guerrero-García, J., Roman-Roldan, R.: Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretical Biology* **160**(4), 457–470 (1993)
- [64] Deza, M.M., Deza, E.: *Encyclopedia of Distances*. Springer, Berlin Heidelberg (2009)
- [65] Kruskal, J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**(1), 1–27 (1964)
- [66] Kari, L., Sayem, A.S., Dattani, N., Hill, K.: Map of life: Measuring and visualizing species relatedness with genome distance maps. University of Western Ontario Technical Report 756, 978-0771430220 (April 2013)

- [67] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference On, vol. 2, pp. 2169–2178 (2006)
- [68] Karamichalis, R.: Molecular Distance Map Interactive Webtool (2014). <https://github.com/rallis/intraMoDMap>
- [69] Pang-Ning, T., Steinbach, M., Kumar, V., *et al.*: Introduction to data mining. In: Library of Congress (2006)
- [70] Zhao, Y., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning* **55**(3), 311–331 (2004)
- [71] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987)

Chapter 5

Additive methods for genomic signatures¹

5.1 Background

Motivated by the general need to identify and classify species based on molecular evidence, alignment-free genome comparisons have been proposed, based on comparing Chaos Game Representations (CGR) of genomic DNA sequences. The CGR of a DNA sequence, proposed by Jeffrey [1, 2], is a graphical representation of a DNA sequence, where the patterns in the image correspond to the frequencies of k -mers in the sequence. Deschavanne *et al.* [3, 4] were the first to suggest that CGR is a good candidate for the role of “genomic signature” defined by Karlin and Burge [5] as any specific quantitative characteristic of a sequence that is pervasive along the genome, while being dissimilar for sequences originating from organisms of different species.

CGR is one of a variety of alignment-free methods (see [6, 7, 8, 9, 10, 11] for detailed literature reviews) that have been proposed for sequence and genome comparisons, as a computationally efficient approach that performs well even with DNA sequences that have nothing or little in common. (We use the following notational conventions for genomic DNA: nDNA (nuclear/nucleoid DNA), mtDNA (mitochondrial DNA), cpDNA (chloroplast DNA), and pDNA

¹A version of this chapter was published (R. Karamichalis, L. Kari, S. Konstantinidis, S. Kopecki and S. Solis-Reyes, “Additive methods for genomic signatures”, *BMC Bioinformatics*, 17:313 (2016))

(plasmid DNA.)

Initially, CGR images were only qualitatively analyzed [12, 13, 14], and Dutta *et al.* and Goldman both advanced the suggestion that CGR images represent no more information than second-order Markov chains [15, 16], which was later disproven by Almeida *et al.* [17, 18] and others [19, 20]. CGR has been applied extensively to phylogenetics together with the Euclidean distance, for instance on nDNA fragments from various domains [3], 27 genomes from various genera [4], 125 nDNA fragments from several bird genomes [21], 26 mtDNA sequences (also with the Pearson distance and a custom image distance) [19], 4 bacteria and about 200 phages [22], 75 HIV-1 genomes [23], 10 mtDNA sequences and 14 nDNA sequences from plants in the *Brassicales* order [24]. Other distances have also been used, for instance the DSSIM image distance on a set of 3,176 mtDNA sequences [20], and six different distances on 174 million base pairs of sampled nDNA fragments from organisms of all major kingdoms of life [25]. The performance of several distance functions has also been compared and benchmarked on their accuracy in constructing phylogenetic trees [26, 27, 28, 29, 30, 31, 32]. Initially, CGR was used only for strings over a 4-letter alphabet (like DNA), but generalizations have been proposed to peptide sequences [33, 34, 35, 36, 37, 38], and Almeida and Vinga proposed a derivative of CGR called the Universal Sequence Map (USM), which is suitable for alphabets of any size [39, 40]. CGRs have also been subjected to multifractal analysis (which measures the degree of self-similarity within the image), see, e.g., [35, 41, 42, 43, 44, 45, 46]. Lastly, CGR has been used to estimate sequence entropy [47, 48, 49], to speed up local-alignment algorithms [50], and has been used together with neural networks to classify HPV genomes by genotype [51].

Several CGR studies [13, 52, 20] observed that CGR patterns of nuclear and organellar DNA sequences of the same organism can be completely different. While the hypothesis that CGRs of mitochondrial DNA sequences can play the role of genomic signatures was tested and validated on the set of all 3,176 sequenced mitochondrial genomes (totalling 91.3 megabase pairs) available in the NCBI GenBank sequence database in July 2012 [20], to our knowledge

no such extensive analysis of CGRs of nuclear/nucleoid genomic sequences exists to date.

The main contributions of this paper are:

- We present an extensive analysis of the hypothesis that conventionally computed (called herein “conventional”) nDNA signatures can play the role of genomic signatures at multiple taxonomic levels, from kingdom to species. Our dataset totals 1.45 gigabase pairs of nDNA sequences from 42 different genomes, from all major kingdoms of life.
- Our analysis indicates that conventional nDNA signatures of two different origins cannot always be differentiated, especially if they originate from closely related organisms. To address this issue, we propose taking into account information obtained from organellar DNA, in addition to nDNA. More generally, we propose the concept of an additive DNA signature of a set (collection) of DNA sequences, and define two particular instances: composite DNA signatures and assembled DNA signatures.
- We explore composite DNA signatures, which combine conventional nDNA signatures with organellar DNA signatures (mtDNA, cpDNA, or pDNA) of the same organism. We demonstrate that, in this dataset, the composite DNA signatures originating from two different organisms can be differentiated in all cases, including those where the use of conventional nDNA signatures failed. In particular, composite DNA signatures from genomes of species as closely related as *H. sapiens* and *P. troglodytes*, or *E. coli* and *E. fergusonii*, can be successfully separated.
- We explore assembled DNA signatures, which combine information from many short contigs (e.g., 100 bp) of a DNA fragment to produce a recognizable signature. This is in contrast to conventional DNA signatures wherein one single long (thousand to hundreds of thousands of basepairs) DNA sequence is needed to generate a recognizable signature.

The enhanced discriminating power of composite DNA signatures, and the ability of assembled DNA signatures to operate with scattered and reduced sequence data, open the possibility of

practical applications including aiding species identification or classification, and comparisons of DNA fragments of various origins such as genomes of extinct organisms, synthetic genomes, raw unassembled next-generation sequencing (NGS) read data, or even computer-generated DNA sequences.

5.2 Results

The first objective of this study was to test, on a comprehensive dataset, the hypothesis that conventional nDNA signatures can be used to differentiate between nuclear DNA sequences originating from different organisms, spanning all major kingdoms of life, at multiple taxonomic levels.

To this end, the following computational experiment was performed, for each of the major kingdoms of life, at various taxonomic levels. We chose a pivot organism (e.g., *H. sapiens* for Kingdom Animalia) and proceeded to use conventional nDNA signatures to compare fragments of its nuclear/nucleoid genome with fragments of the nuclear/nucleoid genome of one other organism from the same kingdom. The process was then repeated with the second organism being at increasing degrees of relatedness to the pivot organism.

More precisely, for each such pairwise comparison, the following three-step process was implemented.

Step 1. Randomly sample 150 kbp nDNA fragments from every chromosome (20 per chromosome, or all fragments if fewer) of the two genomes involved in the comparison. For each such nDNA fragment, construct its corresponding conventional nDNA signature using the process described in Section 5.4.

Step 2. Compute pairwise distances for all pairs of conventional nDNA signatures generated in Step 1. The distance used to start with was an approximated information distance (AID), formally defined in Section 5.4 (see also [53, 25]), since it is computationally simple and uses the least amount of sequence information. If separation was not achieved using

AID, five other distance measures were used: Structural Dissimilarity Index (DSSIM) [54], Euclidean distance, Pearson correlation distance [55], Manhattan distance [56], and descriptor distance [25].

Step 3. Use the distance matrix obtained in **Step 2** as input to a Multi-Dimensional Scaling (MDS) algorithm to produce a 3D Molecular Distance Map [25]: Each point in the map corresponds to (the conventional nDNA signature of) an nDNA fragment from **Step 1**, and the geometric distance between every two points corresponds to the distance between the respective conventional nDNA signatures in the distance matrix. Assess, for each Molecular Distance Map, whether or not separation between conventional nDNA signatures of DNA fragments from the pivot organism and those from the other organism was achieved, by using either *k*-means clustering [57] or by verifying the existence of a separating plane.

Figure 5.1 illustrates an example of the end result of this three-step process: A three-dimensional Molecular Distance Map that displays the conventional nDNA signatures of the pivot organism of Kingdom Animalia, *H. sapiens*, plotted together with the conventional nDNA signatures of *D. melanogaster*.

The results for all kingdoms are presented in Table 5.2 (the first two result columns) and the corresponding 3D Molecular Distance Maps can be found in [58]. For Kingdom Animalia, the approximated information distance succeeded to separate *H. sapiens* (24 chromosomes, 480 fragments) conventional nDNA signatures from those of other organisms, down to and including from *M. murinus* (grey mouse lemur, same order but different suborder) and *T. syrichta* (Phillipine tarsier, same suborder but different infraorder). In the cases marked Y* in Table 5.2, while the accuracy was less than the threshold for separation (85%), the existence of a separating plane was verified. See discussion in Section 5.4 for details.

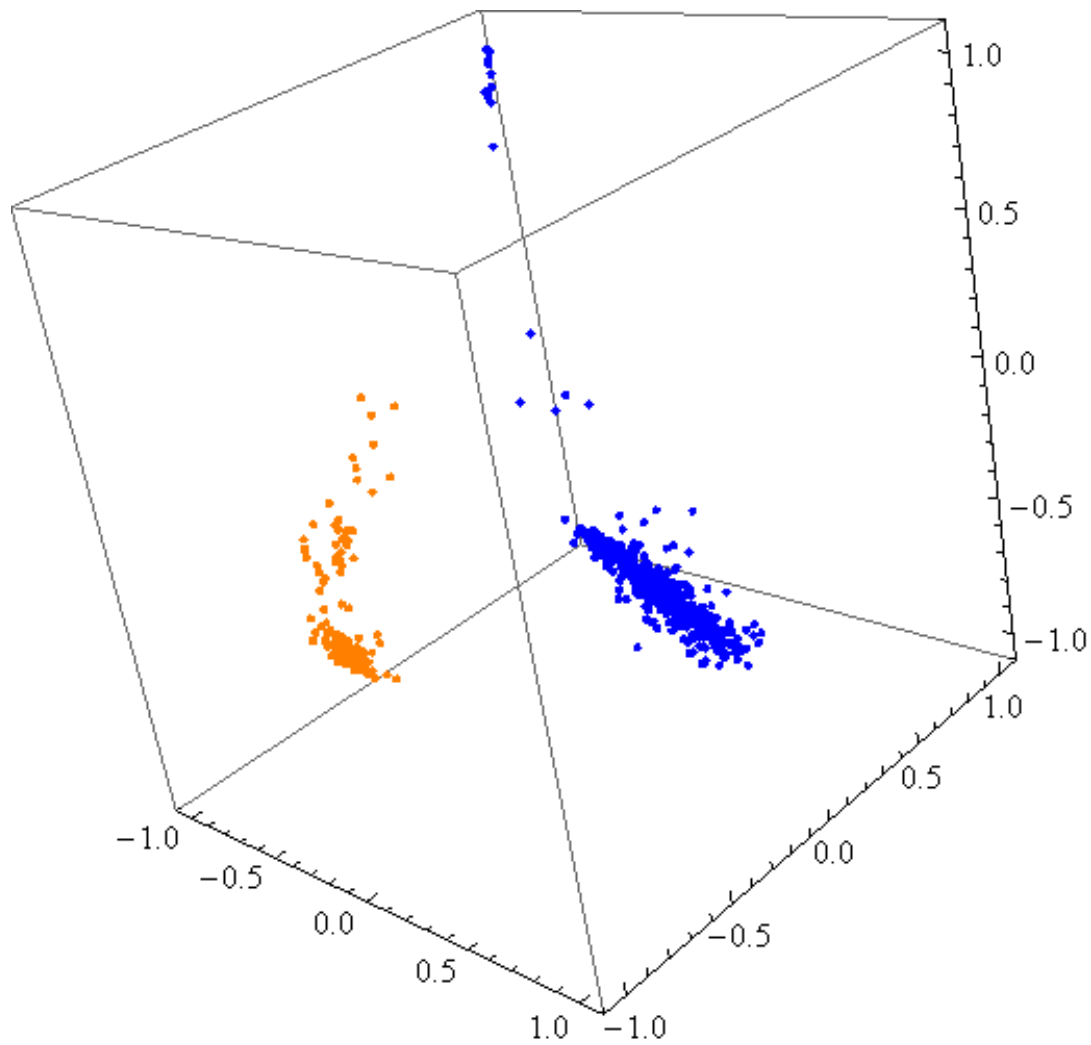


Figure 5.1: 3D Molecular Distance Map illustrating interrelationships among conventional nDNA signatures of 480 randomly sampled 150 kbp nuclear genomic fragments from *H. sapiens* (blue) and 128 randomly sampled 150 kbp nuclear genomic fragments from *D. melanogaster* (orange). The accuracy of separation is 97.2%.

Table 5.1: Each subtable summarizes, for a given kingdom, the results of pairwise comparisons between DNA signatures of fragments from a pivot organism (blue) and those from one other organism, at increasing levels of relatedness. The first two result columns indicate the outcome of the comparisons of conventional nDNA signatures, and the last two columns the comparisons of composite DNA signatures. Green indicates that separation was achieved with AID, red indicates that separation was not achieved with any of the six distances listed in Section 5.2, and yellow (Y/N) or Y* indicate results discussed in the text. The columns labelled Acc % indicate the accuracy of the separations listed immediately at their left: $Acc > 85\%$ was considered separation. A dash indicates that no sequenced data was available on NCBI/GenBank at the time of this submission. The corresponding 3D Molecular Distance Maps for each of the comparisons can be found in [58].

Animalia

<i>H. sapiens</i> vs.	Common taxon	Different taxon	nDNA	Acc %	nDNA+mtDNA	Acc%
<i>D. melanogaster</i>	Kingdom: Animalia	Phylum: Arthropoda	Y	97.2	Y	100
<i>G. gallus</i>	Phylum: Chordata	Class: Aves	Y*	65.25	Y	100
<i>M. musculus</i>	Class: Mammalia	Order: Rodentia	Y*	50.6	Y	100
<i>M. murinus</i>	Order: Primates	Suborder: Strepsirrhini	Y*	57.04	Y	100
<i>T. syrichta</i>	Suborder: Haplorhini	Infraorder: Tarsiiformes	Y*	62.65	Y	100
<i>C. jacchus</i>	Infraorder: Simiiformes	Parvorder: Callitrichidae	N	50.36	Y	100
<i>P. anubis</i>	Parvorder: Catarrhini	Family: <i>Cercopithecidae</i>	N	51	Y	100
<i>N. leucogenys</i>	Superfamily: Hominoidea	Family: <i>Hylobatidae</i>	N	52.9	Y	100
<i>P. abelii</i>	Family: <i>Hominidae</i>	Subfamily: <i>Ponginae</i>	N	50.41	Y	100
<i>G. gorilla gorilla</i>	Subfamily: <i>Homininae</i>	Genus: <i>Gorilla</i>	N	50.72	Y	100
<i>P. troglodytes</i>	Tribe: <i>Hominini</i>	Genus: <i>Pan</i>	N	52.34	Y	100

Fungi

<i>S. cerevisiae</i> vs.	Common taxon	Different taxon	nDNA	Acc %	nDNA+mtDNA	Acc%
<i>C. gattii</i>	Kingdom: Fungi	Phylum: Basidiomycota	Y	100	—	—
<i>F. oxysporum</i>	Phylum: Ascomycota	Class: Sordariomycetes	Y	100	Y	100
—	Class: Saccharomycetes	Order: —	—	—	—	—
<i>K. pastoris</i>	Order: Saccharomycetales	Family: <i>Phaffomycetaceae</i>	Y*	65.6	—	—
<i>C. dubliniensis</i>	Family: <i>Saccharomycetaceae</i>	Genus: <i>Candida</i>	Y	100	—	—
<i>S. arboricola</i>	Genus: <i>Saccharomyces</i>	Species: <i>S. arboricola</i>	Y/N	59	Y	100

Plantae

<i>B. napus</i> vs.	Common taxon	Different taxon	nDNA	Acc %	nDNA+mtDNA/ nDNA+cpDNA	Acc%
<i>M. pusilla</i>	Kingdom: Plantae	Phylum: Chlorophyta	Y	98.04	Y/Y	100
<i>P. patens</i>	Unranked: Embryophyta	Unranked: Bryophyta	Y	98.26	Y/Y	100
<i>M. domestica</i>	Unranked: Rosids	Unranked: Fabids	Y	100	Y/Y	100
<i>C. papaya</i>	Order: Brassicales	Family: <i>Caricaceae</i>	Y	99.67	Y/Y	100
<i>A. thaliana</i>	Family: <i>Brassicaceae</i>	Tribe: <i>Camelineae</i>	N	70	Y/Y	100
<i>R. sativus</i>	Tribe: <i>Brassicaceae</i>	Genus: <i>Raphanus</i>	N	65.4	Y/Y	100
<i>B. oleracea</i>	Genus: <i>Brassica</i>	Species: <i>B. oleracea</i>	N	62.85	Y/Y	100

Protista

<i>P. falciparum</i> vs.	Common taxon	Different taxon	nDNA	Acc %	nDNA+mtDNA	Acc%
<i>O. trifallax</i>	Kingdom: Protista	Phylum: Ciliophora	Y	100	—	—
<i>T. gondii</i>	Phylum: Apicomplexa	Class: Conoidasida	Y	100	—	—
<i>T. orientalis</i>	Class: Aconoidasida	Order: Piroplasmida	Y	100	—	—
—	Order: Haemosporida	Family: —	—	—	—	—
—	Family: <i>Plasmodiidae</i>	Genus: —	—	—	—	—
<i>P. vivax</i>	Genus: <i>Plasmodium</i>	Species: <i>P. vivax</i>	Y	99.65	Y	99.65

Bacteria

<i>E. coli</i> vs.	Common taxon	Different taxon	nDNA	Acc %	nDNA+pDNA	Acc%
<i>S. aureus</i>	Kingdom: Bacteria	Phylum: Firmicutes	Y	100	—	—
<i>H. pylori</i>	Phylum: Proteobacteria	Class: Epsilonproteobacteria	Y	100	—	—
<i>A. baumannii</i>	Class: Gammaproteobacteria	Order: Pseudomonadales	Y	100	Y	100
—	Order: Enterobacteriales	Family: —	—	—	—	—
<i>S. enterica</i>	Family: <i>Enterobacteriaceae</i>	Genus: <i>Salmonella</i>	Y	87.5	Y	100
<i>E. fergusonii</i>	Genus: <i>Escherichia</i>	Species: <i>E. fergusonii</i>	N	50	Y	100

Archaea

<i>P. furiosus</i> vs.	Common taxon	Different taxon	nDNA	Acc %	—	—
<i>S. islandicus</i>	Kingdom: Archaea	Phylum: Crenarchaeota	Y	100	—	—
<i>M. smithii</i>	Phylum: Euryarchaeota	Class: Methanobacteria	Y	100	—	—
—	Class: Thermococci	Order: —	—	—	—	—
—	Order: Thermococcales	Family: —	—	—	—	—
<i>Thermococcus</i> sp. AM4	Family: <i>Thermococcaceae</i>	Genus: <i>Thermococcus</i>	Y	100	—	—
<i>P. yayanosii</i>	Genus: <i>Pyrococcus</i>	Species: <i>P. yayanosii</i>	Y	100	—	—

The use of conventional nDNA signatures failed to achieve separation for genomes of more closely related species. In particular, it failed to separate conventional nDNA signatures of *H. sapiens* from those of *C. jacchus* (common marmoset, same infraorder), *P. anubis* (Anubis baboon, same parvorder), *N. leucogenys* (northern white-cheeked gibbon, same superfamily), *P. abelii* (Sumatran orangutan, same family), *G. gorilla* (gorilla, same subfamily, and *P. troglodytes* (chimpanzee, same tribe, see Figure 5.2). For those organisms where separation was not achieved with approximated information distance, we performed the comparisons with the other five distances. The results of these multiple computations were that, in all cases where approximated information distance failed to achieve separation, the other distances also failed.

For Kingdom Fungi, the pivot organism is the model organism *Saccharomyces cerevisiae* (16 chromosomes, 73 fragments), a species of yeast instrumental to winemaking, baking, and brewing. Separation of its conventional nDNA signatures was achieved down to and including separation from *C. dubliniensis* (same family, different genus). In the case of the comparison with *K. pastoris*, marked with Y* in Table 5.2, the accuracy score was lower than 85%: This

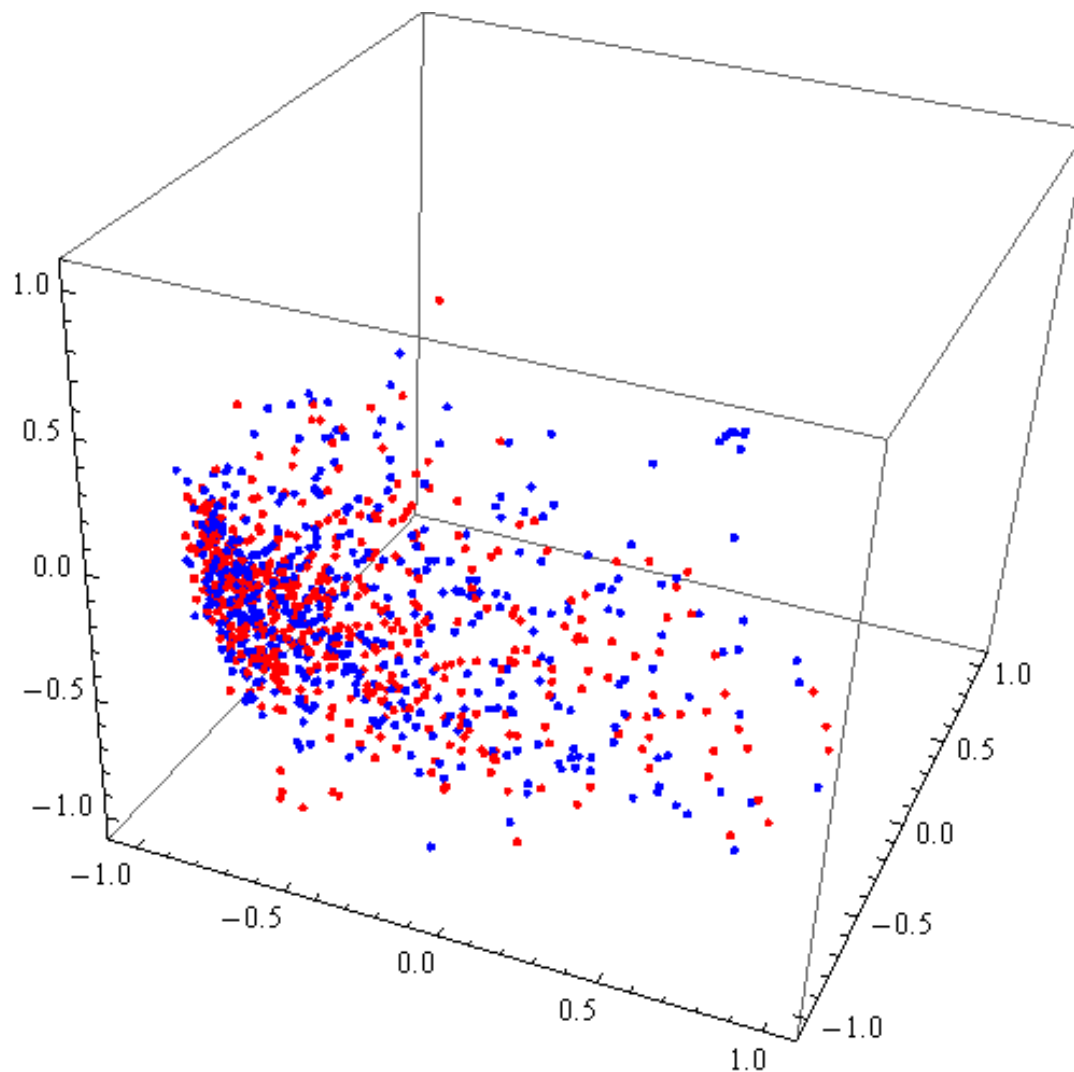


Figure 5.2: 3D Molecular Distance Map illustrating interrelationships among conventional nDNA signatures of 480 randomly sampled nuclear genomic fragments from *H. sapiens* (blue) and 500 randomly sampled nuclear genomic fragments from *P. troglodytes* (red). All fragments are 150 kbp long, the accuracy of separation is 52.34%, and no separation plane could be found.

is an artifact of the shape of the 3D Molecular Distance Map wherein one of the clusters has a trailing set of points that become erroneously separated by k -means from all the rest of the points. Because of this, and since the use of k -means on the 2D Molecular Distance Map of the same dataset resulted in an accuracy score of 100%, we interpreted this comparison as resulting in separation. The results of the comparison between the conventional nDNA signatures of the pivot organism and those of *S. arboricola* (same genus, different species), were inconclusive: The use of Euclidean and Pearson distances resulted in separation (both with accuracy of 88.48%), while the use of the other four distances (DSSIM, Manhattan, descriptor, approximated information distance) did not result in separation.

For Kingdom Plantae, the pivot organism is the model organism *Brassica napus* (19 chromosomes, 380 DNA fragments), rapeseed, a flowering member of the family *Brassicaceae* (mustard or cabbage family). Separation of its conventional nDNA signatures was achieved down to and including separation from *C. papaya* (papaya, same order, different family). For the comparisons with *A. thaliana* (thale cress, same family, different tribe) and *R. sativus* (radish, same tribe, different genus), cluster separation was visually observed but not quantitatively confirmed by either k -means or plane separation. The comparison with *B. oleracea* (wild cabbage, same genus, different species) did not result in separation, with any of the six distances.

For Kingdom Protista, the pivot organism is the model organism *Plasmodium falciparum*, a protozoan parasite (14 chromosomes, 149 DNA fragments), one of the species of *Plasmodium* that cause malaria in humans. Separation of its conventional nDNA signatures from those of other organisms from the same kingdom was achieved at all taxonomic levels, down to and including separation from *P. vivax* (same genus, different species).

For Kingdom Bacteria, the pivot organism is the model organism *Escherichia coli* (20 genomic DNA fragments), a bacterium commonly found in the lower intestine of warm-blooded organisms. Separation of its conventional nDNA signatures from those of other bacteria was successful down to and including separation from *S. enterica* (same family, different genus),

but failed with all six distances in the comparison with *E. fergusonii* (same genus, different species).

For Kingdom Archaea, the pivot organism is the model organism *Pyrococcus furiosus* (12 genomic DNA fragments), an extremophilic species of Archaea. Separation of its conventional nDNA signatures from those of other archaea was successful at all levels, down to and including separation from *P. yayanosii* (same genus, different species).

The above results indicate that, especially in kingdom Animalia, conventional nDNA signatures cannot always be used to differentiate nuclear/nucleoid genomic sequences originating from two different genomes. This suggests that conventional nDNA signatures cannot always play the role of a “genomic signature”, particularly when the genomes being compared belong to closely related species.

5.2.1 Composite DNA signatures

To enhance the discriminating power of conventional nDNA signatures, our second objective was to introduce and explore the concept of composite DNA signatures, which combine conventional nuclear/nucleoid DNA signatures with signatures of organellar genomes (mtDNA, cpDNA, or pDNA).

To test the discriminating power of composite DNA signatures, we repeated all previous pairwise comparisons (where sequenced organellar DNA was available), using this time composite DNA signatures. The results are presented in the last two columns of Table 5.2.

For Kingdoms Animalia, Fungi and Protista we used composite DNA signatures combining the conventional nDNA signature of each nuclear/nucleoid genomic fragment with that of the mtDNA of the same organism (when available). Using such composite DNA signatures, differentiation of DNA signatures by organism was successful in all cases, including all cases where the use of conventional nDNA signature previously failed or was inconclusive. See Figure 5.2 (*H. sapiens* vs. *P. troglodytes* conventional nDNA signatures, no separation) versus Figure 5.3 (*H. sapiens* vs. *P. troglodytes* composite DNA signatures using nDNA and mtDNA, complete

separation).

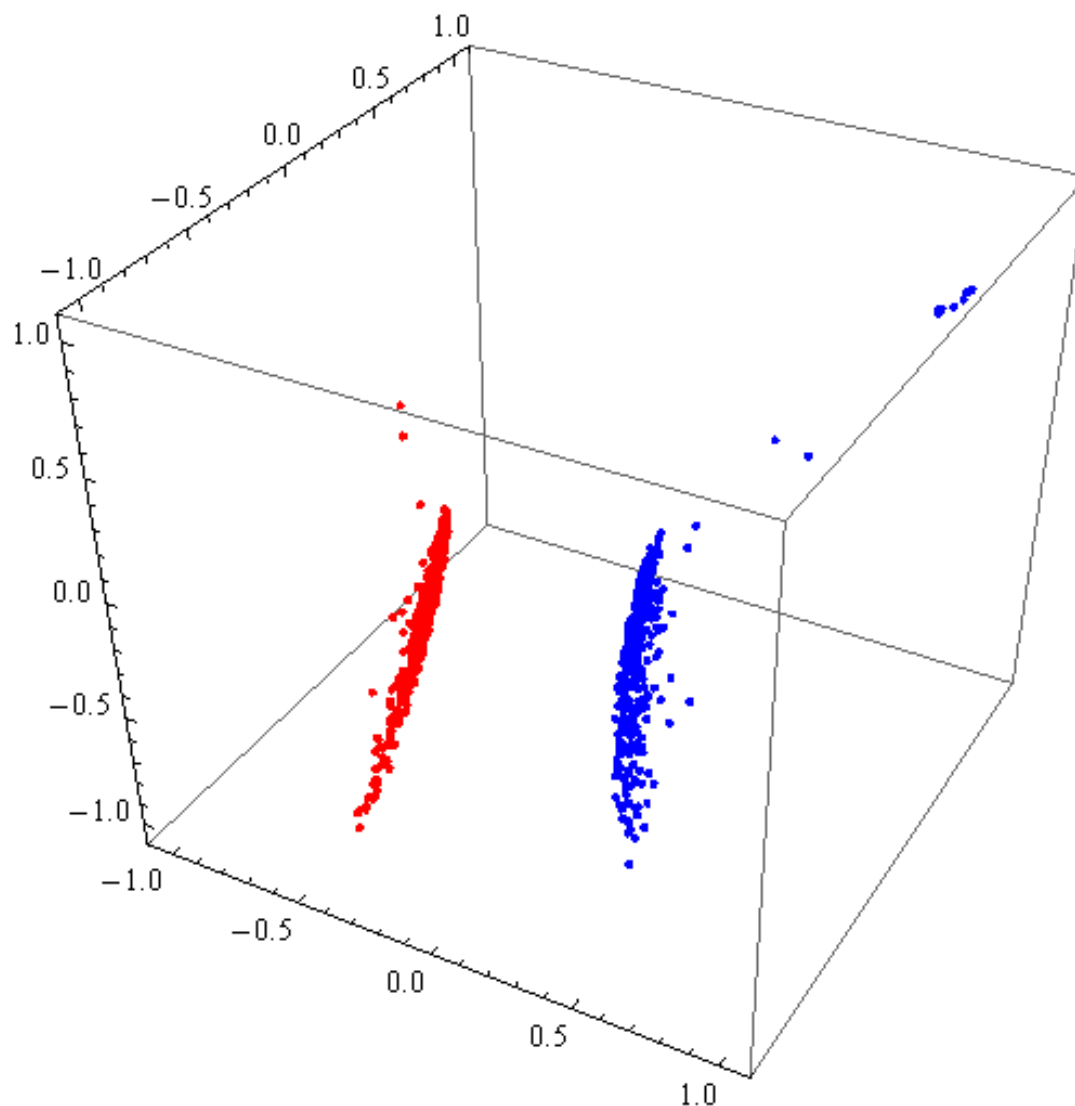


Figure 5.3: 3D Molecular Distance Map illustrating interrelationships among composite DNA signatures using nDNA and mtDNA, of 480 DNA fragments from *H. sapiens* (blue) and 500 DNA fragments from *P. troglodytes* (red). The accuracy of separation is 100%.

To test the discriminating power of composite DNA signatures using nDNA, mtDNA and cpDNA, we employed them to perform comparisons for all genome pairs from Kingdom Plantae. Separation was achieved using all of: composite DNA signatures using nDNA and mtDNA, composite DNA signatures using nDNA and cpDNA, and composite DNA signatures using nDNA, mtDNA, and cpDNA. See Figure 5.4 for the Molecular Distance Maps illustrat-

ing the relationships between these signatures for *B. napus* and *B. oleracea*.

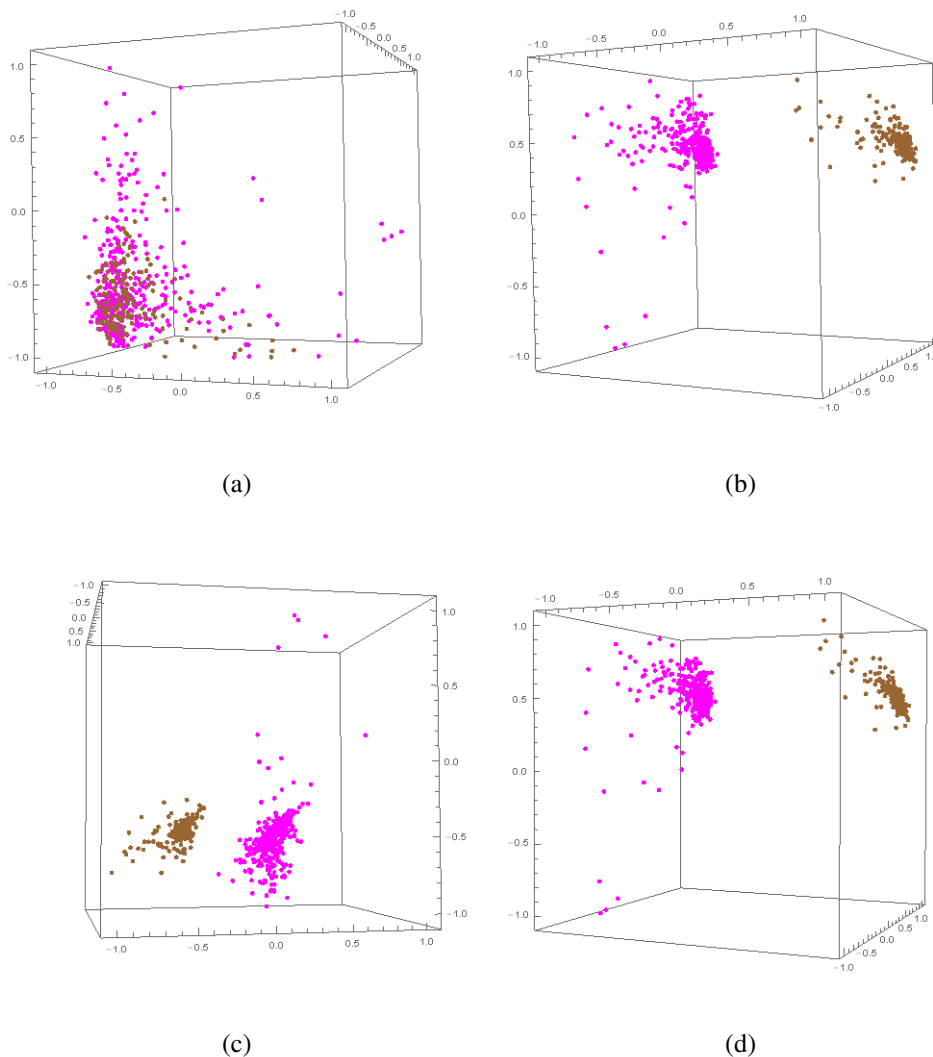


Figure 5.4: 3D Molecular Distance Map illustrating interrelationships among signatures of 380 DNA fragments from *B. napus* (magenta) and 180 DNA fragments from *B. oleracea* (brown) using (a) conventional nDNA signatures, (b) composite DNA signatures using nDNA and mtDNA, (c) composite DNA signatures using nDNA and cpDNA, and (d) composite DNA signatures using nDNA, mtDNA, and cpDNA. The accuracy of separation is 63.03% for (a), and 100% for each of (b), (c), and (d).

For Kingdom Bacteria, the use of composite DNA signatures combining nDNA and pDNA (when available) resulted in separation in all cases.

Overall, the use of composite DNA signatures resulted in separation in all pairwise comparisons in Table 5.2 (where organellar DNA sequencing data was available), including in those

where the use of conventional nDNA signature failed or resulted in inconclusive separations.

5.2.2 Assembled DNA signatures

As the third objective of this study, we explored a way to enhance the practical applicability of conventional DNA signatures. Recall that, to produce a recognizable visual pattern that can be reliably used to represent a genome, a conventional DNA signature needs as input a long contiguous (two to several hundred kilobase pairs) DNA fragment. This assumes a high quality and reliability of sequencing and assembly, which are not always available. We propose instead to approximate a conventional signature by an assembled DNA signature, which combines the conventional DNA signatures of many short contigs (e.g., 100 bp) of the given fragment. Note that these contigs need not cover the entire DNA fragment.

In what follows, we denote by $|s|$ the length of the sequence s . Given a DNA fragment s , an assembled DNA signature of s , using r equi-length contigs of length n (subfragments of the sequence s), is defined as the sum of the conventional DNA signatures of all of the r contigs. A particular case of assembled DNA signature is where the fragment s is partitioned into equi-length, consecutive, non-overlapping contigs, that is, $s = s_1 s_2 \dots s_r s_{r+1}$, and $|s_i| = n$ for $1 \leq i \leq r$, with $|s_{r+1}| < n$. In this case, we call the assembled signature a fully-assembled DNA signature of the sequence s , using equi-length contigs of length n .

Table 5.2 ((A) through (C)) presents a comparison between the conventional nDNA signature of a given DNA fragment and its assembled DNA signatures, as well as fully-assembled DNA signatures, for various values of contig length n , and number of contigs r . The DNA fragment used is from *H. sapiens*, chromosome 21, fragment 20 (from position 2,850,001 to 3,000,000 after removing all *N*s in the original sequence), and the distance used is approximated information distance between CGRs with $k = 9$. For example, the distance between the conventional nDNA signature and the fully-assembled DNA signature of the same fragment, that uses 1,000 contigs of length 150 bp each, is 0.03 (row 2, column (A)). This value is very small, given that approximated information distance theoretically ranges between 0 and

n	r	(A)	(A')	(B)	r	(C)	r	(B')	r	(C')
100	1500	0.05	0.13	0.29	4500	0.042	1475	0.32	4434	0.041
150	1000	0.03	0.09	0.29	3000	0.034	1000	0.29	2999	0.040
200	750	0.02	0.07	0.28	2250	0.033	750	0.29	2250	0.038
300	500	0.02	0.04	0.28	1500	0.030	500	0.28	1500	0.038
500	300	0.01	0.03	0.26	900	0.037	300	0.28	900	0.033
1000	150	0.005	0.01	0.30	450	0.030	150	0.25	450	0.039
2000	75	0.003	0.007	0.30	225	0.041	75	0.26	225	0.023
3000	50	0.002	0.004	0.25	150	0.044	50	0.29	150	0.021
10000	15	0.0004	0.001	0.30	45	0.053	15	0.25	45	0.045
15000	10	0.0003	0.0008	0.24	30	0.12	10	0.23	30	0.079
30000	5	0.0001	0.0004	0.36	15	0.13	5	0.41	15	0.058

Table 5.2: (A) through (C) – Distances between the conventional nDNA signature of a fragment and its assembled DNA signatures, for various numbers r of contigs of the same length n : (A) distances to fully-assembled DNA signatures; (A') theoretical upper bounds for (A); (B) distances to assembled DNA signatures; (C) same as (B), when tripling the number of contigs. (B') through (C') – Distances between the conventional nDNA signature of a fragment and its assembled DNA signatures, using variable-length contigs taken from a normal distribution $N(n, \sigma)$, with mean n and variance $\sigma = 40$. The nDNA fragment used was from *H. sapiens*, chromosome 21, fragment 20 (from position 2,850,001 to 3,000,000 after removing all *Ns* in the original sequence).

1. This suggests that, for these parameter values ($n = 150$ and $r = 1,000$), a fully-assembled DNA signature can be an excellent approximation of the conventional DNA signature of the same fragment. This was expected, given that the only information lost in the computation of a fully-assembled DNA signature, when using the approximated information distance, is the information about the k -mers situated at the borders between contigs.

Also as expected, for the same values of n and r , the distance between an assembled DNA signature and the conventional nDNA signature of the same fragment (Table 5.2, Column (B)) is higher than the one between a fully-assembled DNA signature and the conventional nDNA signature of the same fragment (Table 5.2, (A)). This indicates that the assembled DNA signature is less performant than the fully-assembled DNA signature as an approximation of a conventional nDNA signature. The reason is that, given a fixed number r of contigs, in the case of an assembled DNA signature the contigs are allowed to overlap and need not cover the entire fragment. This can be compensated by increasing the coverage, that is, the number

r of contigs. Table 5.2, (C) shows that tripling the number of contigs results in significantly smaller differences between assembled DNA signatures and the conventional DNA signature of the same fragment which they were meant to approximate.

The results in Table 5.2 suggest that assembled DNA signatures have the potential to play the role of “genomic signatures”, and be used directly on raw unassembled next-generation sequencing read data, or in cases where other methods are not directly applicable because high-quality sequencing data is not available. To test this hypothesis, we considered the organism pairs in Table 5.2 for which separation was obtained using conventional nDNA signatures, and attempted to reproduce these successful separations using assembled DNA signatures instead. In addition, we empirically sought to find, in each case, the coverage (amount of sequence data) needed to achieve separation, as a percentage of total fragment length.

To determine the threshold interval where separation between assembled DNA signatures of a given pair of organisms was achieved, when contigs of length $n = 300$ were used, the following process was employed. For various values of t , $0 \leq t \leq 1$ (representing the fragment coverage, e.g., $t = 0.5$ means that 50% of the fragment data was used), we attempted to see if separation of assembled DNA signatures from the two organisms was achieved, in the following way.

For each of the 150 kbp fragments s from the two genomes, q random positive integers were picked from the interval 1 to $|s| - n + 1 = (150,000 - 300 + 1)$, where $q = \lfloor t * |s|/n \rfloor$, that is, the integer part of $t * |s|/n$. These q numbers represent the start positions of the q chosen contigs. For each contig start position, a contig of length $n = 300$ was read and used for the assembled DNA signature of the fragment s .

For each value of t , the corresponding 3D Molecular Distance Map of the assembled DNA signatures of the two organisms was then analyzed, by verifying the existence (or absence) of a separating plane.

The results are summarized in Table 5.2.2 and can be interpreted as follows. In the comparison between *H. sapiens* and *D. melanogaster* the threshold interval is 1% - 5%. The lower limit

of this interval is 1%, and this means that in the computation using the coverage value $t = 0.01$ (implying $q = \lfloor 0.01 * 150,000/300 \rfloor = 5$), separation was not achieved. That is, for each of the 150 kbp nDNA fragments available (480 from *H. sapiens* and 128 from *D. melanogaster*), when employing assembled nDNA signatures using only 5 contigs per fragment (for a maximum of 1% of each fragment length, that is, 1,500 bp per fragment), separation was not achieved. The upper limit of the interval is 5%, and this means that in the computation using the coverage value $t = 0.05$ (implying $q = 25$), separation was achieved. That is, when employing assembled nDNA signatures using 25 contigs per fragment (for a maximum of 5% of each fragment length, that is, 7,500 bp per fragment), separation was achieved.

Table 5.3: Each subtable summarizes, for a given kingdom, the results of pairwise comparisons between DNA signatures of fragments from a pivot organism (blue) and those from one other organism, at increasing levels of relatedness. The first two result columns indicate the outcome of the comparisons of conventional nDNA signatures, and the last two columns the comparisons of composite DNA signatures. Green indicates that separation was achieved with AID, red indicates that separation was not achieved with any of the six distances listed in Section 5.2, and yellow (Y/N) or Y* indicate results discussed in the text. The columns labelled Acc % indicate the accuracy of the separations listed immediately at their left: Acc > 85% was considered separation. A dash indicates that no sequenced data was available on NCBI/GenBank at the time of this submission. The corresponding 3D Molecular Distance Maps for each of the comparisons can be found in [58].

Animalia

<i>H. sapiens</i> vs.	Different taxon	Thresh.
<i>D. melanogaster</i>	Phylum: Arthropoda	1% - 5%
<i>G. gallus</i>	Class: Aves	3% - 10%
<i>M. musculus</i>	Order: Rodentia	10% - 20%
<i>M. murinus</i>	Suborder: Strepsirrhini	60% - 80%
<i>T. syrichta</i>	Infraorder: Tarsiiformes	20% - 40%

Fungi

<i>S. cerevisiae</i> vs.	Different taxon	Thresh.
<i>C. gattii</i>	Phylum: Basidiomycota	0.5% - 2%
<i>F. oxysporum</i>	Class: Sordariomycetes	0.5% - 2%
<i>K. pastoris</i>	Family: <i>Phaffomycetaceae</i>	2% - 10%
<i>C. dubliniensis</i>	Genus: <i>Candida</i>	2% - 10%

Plantae

<i>B. napus</i> vs.	Different taxon	Thresh.
<i>M. pusilla</i>	Phylum: Chlorophyta	2% - 3%
<i>P. patens</i>	Unranked: Bryophyta	3% - 4%
<i>M. domestica</i>	Unranked: Fabids	4% - 5%
<i>C. papaya</i>	Family: <i>Caricaceae</i>	4% - 5%

Protista

<i>P. falciparum</i> vs.	Different taxon	Thresh.
<i>O. trifallax</i>	Phylum: Ciliophora	0.5% - 2%
<i>T. gondii</i>	Class: Conoidasida	0.5% - 2%
<i>T. orientalis</i>	Order: Piroplasmida	0.5% - 2%
<i>P. vivax</i>	Species: <i>P. vivax</i>	0.5% - 2%

Bacteria

<i>E. coli</i> vs.	Different taxon	Thresh.
<i>S. aureus</i>	Phylum: Firmicutes	0.5% - 2%
<i>H. pylori</i>	Class: Epsilonproteobact.	0.5% - 2%
<i>A. baumannii</i>	Order: Pseudomonadales	0.5% - 2%
<i>S. enterica</i>	Genus: <i>Salmonella</i>	10% - 20%

Archaea

<i>P. furiosus</i> vs.	Different taxon	Thresh.
<i>S. islandicus</i>	Phylum: Crenarchaeota	0.5% - 2%
<i>M. smithii</i>	Class: Methanobacteria	0.5% - 2%
<i>Thermococcus</i>	Genus: <i>Thermococcus</i>	0.5% - 2%
<i>P. yayanosii</i>	Species: <i>P. yayanosii</i>	0.5% - 2%

The actual threshold values lie in the intervals listed, and may be subject to the quality of the sequencing. As expected, in general, the thresholds needed for separation increase with the increase in the degree of relatedness of the organisms being compared. This suggests that nDNA

sequences from closely related organisms require a higher coverage (that is, a higher amount of information from each sequence) to be separated. The only exception to this trend, in this dataset, were the pairs *H. sapiens* with *M. murinus* (gray mouse lemur) requiring 60% - 80% sequence coverage, and *H. sapiens* and *T. syrichta* (Philippine tarsier) requiring 20% - 40% sequence coverage. Thus, the (human, lemur) pair required higher sequence coverage to achieve separation than the (human, tarsier) pair, even though the gray mouse lemur belongs to a different primate suborder (Haplorrhini) than the modern human, while the tarsier belongs to the same primate suborder as the modern human (Strepsirrhini), and thus one would expect that more information would be needed to achieve the latter separation. This apparent anomaly may be partly related to the fact that the phylogenetic placement of tarsiers within the order Primates has been controversial for over a century [59]: In [60] tarsiers are placed within Haplorrhini, while according to [61, 20], mitochondrial DNA evidence places tarsiiiformes as a sister group to Strepsirrhini.

Table 5.2.2 indicates that the amount of DNA fragment information needed to achieve separation, at the same taxonomic level, can differ from one kingdom to another. For example, in Kingdom Animalia, conventional nDNA signatures of organisms from two species of a different species (*H. sapiens* and *P. troglodytes*) could not be separated even though we use 100% of the DNA fragment information. In contrast, in Kingdom Fungi, assembled nDNA signatures from two organisms of a different genus (*S. cerevisiae* and *C. dubliniensis*) could be separated even when using only 10% of DNA fragment data. Similarly, in Kingdom Bacteria, assembled nDNA signatures from two organisms of different genus (*E. coli* and *S. enterica*) could be separated even when using only 20% of DNA fragment data. The situation is even more extreme in Kingdom Protista and Kingdom Archaea, where even organisms belonging to the same genus could be separated with very little sequence coverage. Indeed, in Kingdom Protista, assembled nDNA signatures of two organisms of the same genus (*P. falciparum* and *P. vivax*) could be separated using only 2% of DNA fragment data. Similarly, in Kingdom Archaea, assembled nDNA signatures from two organisms of the same genus (*P. furiosus* and

P. yananosii) could also be separated using only 2% of DNA fragment data. This suggests that some taxonomic categories, such as “genus”, do not necessarily reflect the same degree of structural similarity of genomic sequences uniformly across kingdoms.

5.2.3 Composite-assembled DNA signatures

We now briefly explore the potential of combining the approach of composite DNA signatures with that of assembled DNA signatures. A composite-assembled DNA signature is produced by combining information from the assembled DNA signatures of two (or more) different types of DNA fragments. For example, a composite-assembled signature using nDNA and mtDNA is obtained by combining the assembled nDNA signature of one 150 kbp nDNA fragment, with the assembled mtDNA signature of the mtDNA genome of the same organism.

Figure 5.5 plots together composite DNA signatures and composite-assembled DNA signatures using nDNA and mtDNA from *H. sapiens* and *P. troglodytes*. Note that composite-assembled DNA signatures and composite DNA signatures of fragments (using nDNA and mtDNA), from the same species are closely clustered together. On the other hand, all DNA signatures of *H. sapiens* are separated from all DNA signatures of *P. troglodytes*, and the existence of a separating plane was verified. These results suggest that composite-assembled DNA signatures could also be potential candidates for the role of “genomic signature”, as they have in general better discriminating power than conventional nDNA signatures while using scattered and potentially less sequence information.

5.3 Conclusions

The first objective of this paper was to conduct a comprehensive analysis, on a dataset totalling 1.45 Gb, of the hypothesis that Chaos Game Representations of nuclear/nucleoid genomic sequences can play the role of “genomic signatures”, that is, that they are genome- and species-specific. Our results suggest that this hypothesis is not always valid, in that nuclear/nucleoid

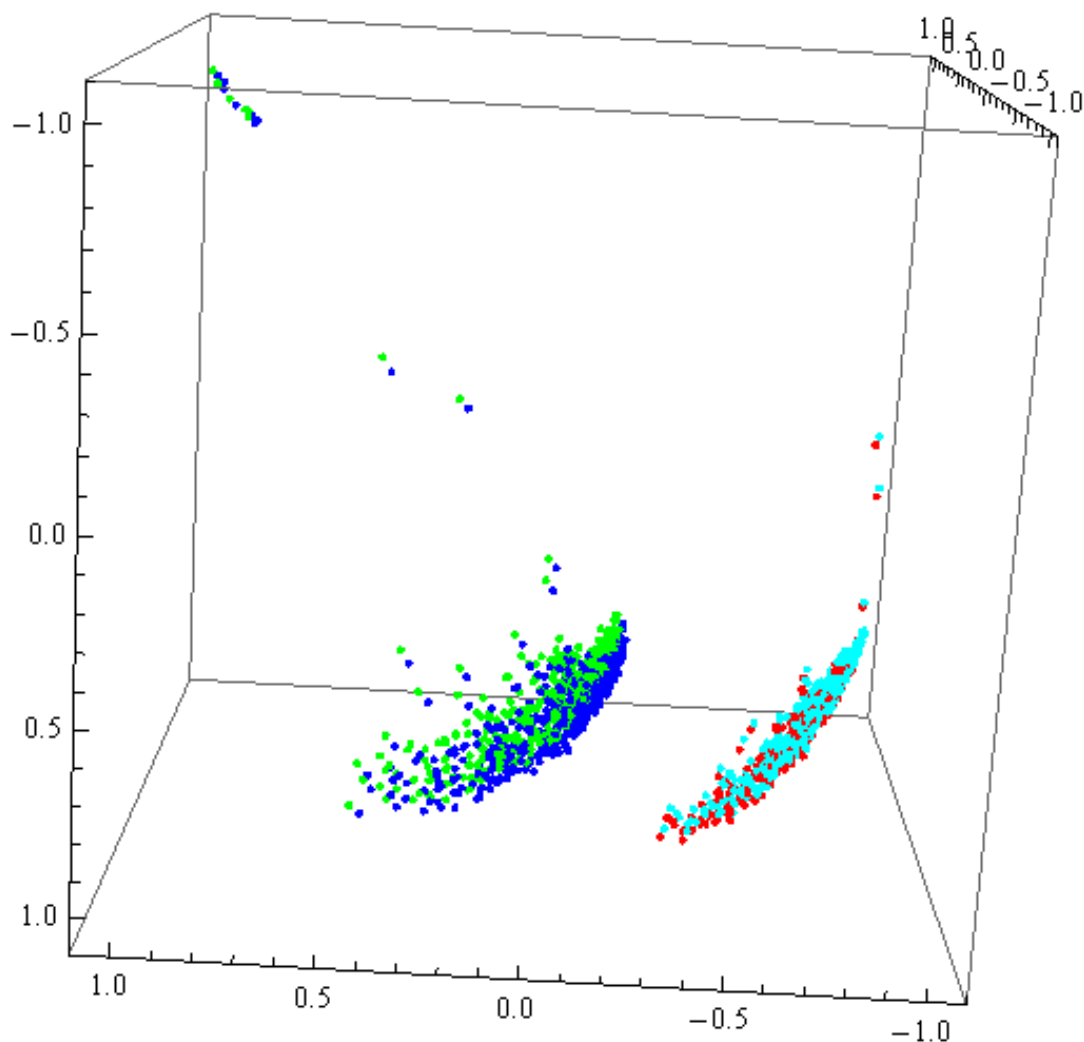


Figure 5.5: 3D Molecular Distance Map illustrating interrelationships among 480 composite (respectively 480 composite-assembled) DNA signatures, each using one nDNA fragment and the mtDNA genome from *H. sapiens*, blue (resp. green); and 500 composite (resp. 500 composite-assembled) DNA signatures, each using one nDNA fragment and the mtDNA genome from *P. troglodytes*, red (resp. turquoise); For the composite-assembled DNA signatures, the length of contigs was $n = 100$, while the number of contigs was 4,500 for each 150 kbp nDNA fragment, and 497 (resp. 496) for the human (resp. chimp) mtDNA genome. The accuracy of separation between the *H. sapiens* and the *P. troglodytes* sequences was 58%, but the existence of a separation plane was verified.

DNA sequences belonging to closely related species such as *H. sapiens* and *P. troglodytes* or *E. coli* and *E. fergusonii* cannot always be separated using conventionally computed CGR signatures.

To address this issue, as a second objective, we propose the use of composite DNA signatures, which combine information from the nuclear/nucleoid genome with that from one or more organellar genomes (mtDNA, cpDNA and/or pDNA). Composite DNA signatures were found, in this study, to result in successful separation of DNA sequences by organism in all cases, including those where conventional nDNA signatures failed.

As a third objective, we propose the use of assembled DNA signatures, which combine information from short contigs (subfragments) of a DNA fragment, rather than using the entire contiguous fragment, to produce its signature. We show that assembled DNA signatures can be successful replacements of conventional DNA signatures, and also that the composite and assembled DNA signature approaches can be used simultaneously.

Mathematically, composite and assembled DNA signatures are both particular cases of a general concept, namely that of an additive DNA signature of a set of DNA sequences (see Section 5.4). Our results indicate that such additive DNA signatures could be considered as potential candidates for the role of “genomic signatures” at various taxonomic levels, from distant to closely related species, and thus complement other methods for species identification and classification.

Several directions of future research stem from the fact that existing literature indicates that the oligomer composition of nuclear/nucleoid DNA sequences and mitochondrial DNA sequences can be a source of taxonomic information. Such directions include testing the discriminating power of additive DNA signatures in large-scale multi-genome comparisons, and exploring their utility in practical applications such as DNA sequence identification and classification (including directly on raw unassembled NGS read data or when high-quality sequencing data is not available), metagenomics, and synthetic genomes.

5.4 Methods

Dataset

The dataset, totalling 1.45 Gb, comprised whole nuclear/nucleoid genomes and organellar genomes of 42 organisms, spanning all major kingdoms of life (see Appendix for the scientific name, NCBI accession number, chromosome number, and number of fragments sampled). In our analysis, for each complete genomic sequence, all letters other than *A, C, G, T* were ignored, and the resulting DNA sequence was divided into successive, non-overlapping, contiguous fragments, each 150 kbp long (when the last portion was shorter than 150 kbp, it was not included in the analysis). The choice of fragment length, 150 kbp, was due to our choice of CGR image resolution (namely $2^9 \times 2^9$, that is, $k = 9$), empirical testing, and computational efficiency reasons, see [25].

Subsequently, 20 such 150 kbp fragments were randomly sampled from each chromosome and, for each such fragment, a corresponding conventional nDNA signature was constructed, as described below. (If there were fewer than 20 fragments, all fragments in the chromosome were chosen.) In the cases where the genome assembly of the organism was at the contig/scaffold level, the contigs/supercontigs of the assembly were sorted by length and the first 500 contigs/supercontigs were selected. (If there were fewer than 500 contigs/supercontigs, all were selected.) From each contig/supercontig, only the first 150 kbp fragment was considered.

We note that this method is alignment-free, and that its approach contrasts typical biodiversity and species identification research [62, 63, 64, 65] in that it uses randomly selected DNA sequences rather than specific marker genes for identification and classification of species. This approach is somewhat similar to novel approaches in metagenomics, metatranscriptomics, and viromics [66], but there are also substantial differences such as that metatranscriptomics is based on RNA rather than DNA and that it groups sequences based on functionality rather than oligomer composition.

Chaos Game Representation (CGR)

CGR is a method introduced by Jeffrey [1] as a way to visualize the structural composition of a DNA sequence. This method associates an image to each DNA sequence as follows: Starting from a square with corners 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 corner 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 *k*-mer: A pixel is black if the *k*-mer it represents occurs in the DNA sequence, otherwise it is white. CGR images of genetic DNA sequences originating from various species show patterns such as squares, parallel lines, rectangles, triangles, and also complex fractal patterns, as shown in Figure 5.6.

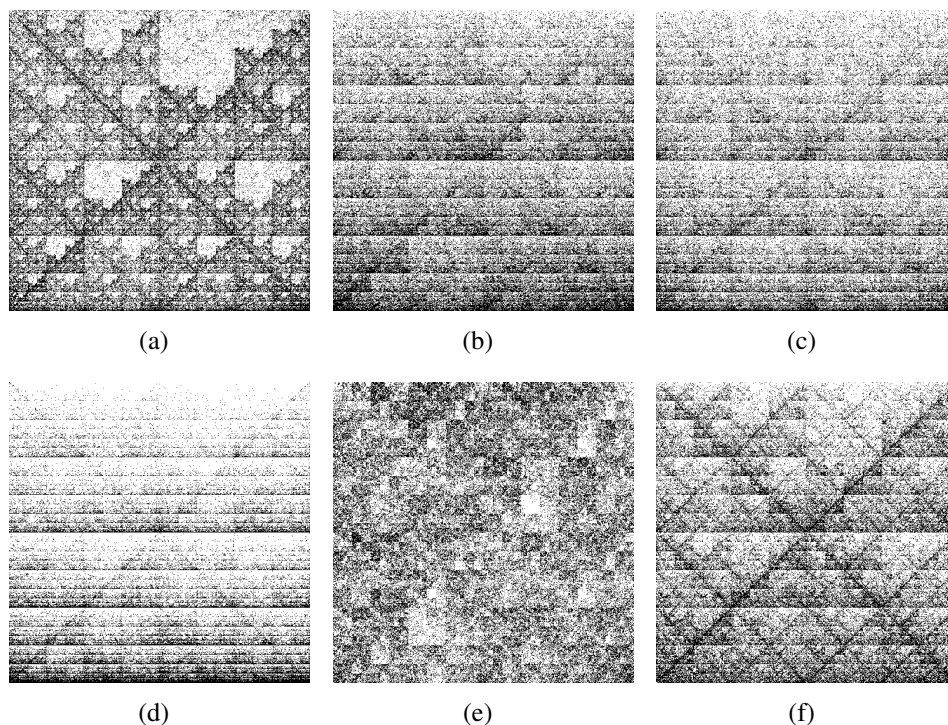


Figure 5.6: Conventional nDNA signatures of 150 kbp sequences of the pivot organisms from Kingdom (a) Animalia, (b) Fungi, (c) Plantae, (d) Protista, (e) Bacteria, and (f) Archaea.

We used a modification of the original CGR, introduced by Deschavanne [3]: a *k*-th order

FCGR (frequency CGR) of a sequence s , denoted by $FCGR_k(s)$, is a $2^k \times 2^k$ matrix that can be constructed by dividing the CGR image of the sequence s into a $2^k \times 2^k$ grid, and defining the element a_{ij} of the matrix $FCGR_k(s)$ as the number of points that are situated in the corresponding grid square.

We now formally define the conventional DNA signature of a sequence s to be the matrix $FCGR_k(s)$, which records the numbers of occurrences of all possible k -mers in the sequence s . Throughout this paper, the parameter k is assumed to be a fixed constant. In particular, similar to [25], in all computational experiments in this paper the value used was $k = 9$.

For computing composite and assembled DNA signatures, we introduce the general concept of additive DNA signature of a set of sequences, formally defined as follows.

Definition The additive DNA signature of a set of sequences $S = \{s_1, s_2, \dots, s_r\}$, $r \geq 1$, is defined as

$$FCGR_k(S) = FCGR_k(s_1) + \dots + FCGR_k(s_r).$$

Note that the notions of conventional DNA signature, composite DNA signature, assembled DNA signature, and fully-assembled DNA signature, are all particular cases of additive DNA signatures, as follows:

- The conventional DNA signature of a sequence s is the additive DNA signature of the set $\{s\}$ consisting of a single sequence s , that is, $FCGR_k(s) = FCGR_k(\{s\})$.
- The composite DNA signature using two DNA sequences s_1, s_2 , of two different types, is

$$FCGR_k(\{s_1, s_2\}) = FCGR_k(s_1) + FCGR_k(s_2),$$

- An assembled signature of a sequence s , using r equi-length contigs of length n , is

$$FCGR_k(\{s_1, s_2, \dots, s_r\}) = \sum_{i=1}^r FCGR_k(s_i),$$

where $s = \alpha_i s_i \beta_i$, $|s_i| = n$, for $1 \leq i \leq r$.

- The fully-assembled DNA signature of a sequence s , using equi-length contigs of length n , is

$$FCGR_k(\{s_1, s_2, \dots, s_r\}) = \sum_{i=1}^r FCGR_k(s_i), \text{ where } r = \lfloor |s|/n \rfloor, s = s_1 s_2 \dots s_r s_{r+1}, \text{ and } |s_i| = n \text{ for } 1 \leq i \leq r, \text{ while } |s_{r+1}| < n.$$

To compute the fully-assembled DNA signature of a sequence s , using equi-length contigs of length n , one adds the $FCGR_k$ of all the adjacent consecutive contigs of length n that cover s (except possibly a short tail of length less than n), where the first contig starts at the beginning of the sequence. In contrast, to compute an assembled signature of s using equi-length contigs of length n , one has the freedom to set the number of such contigs as an additional parameter r , and then add the $FCGR_k$ of r contigs sampled randomly from the sequence s . Thus, for a given n , a sequence s has only one fully-assembled DNA signature, but many different assembled signatures, each depending on both the choice of parameter r , and the particular sampling of the r sequences (which may overlap or be identical).

For example, if s is the DNA sequence

$$s = AAAAAACCCCGGGGGTTT,$$

of length 18, and if we consider contigs of length $n = 5$, then the fully-assembled DNA signature of s is unique and is obtained by adding the $FCGR_k$ of the following $r = \lfloor 18/5 \rfloor = 3$ contigs

$$\{AAAAA, CCCCC, GGGGG\}$$

that cover s (except the discarded remainder TTT).

For the same sequence s and contig length $n = 5$, many different assembled DNA signatures can be computed. For example, an assembled DNA signature of s using $r = 3$ equi-length contigs of length $n = 5$ could use contigs $\{AAACC, CCCGG, CCCGG\}$, while another could use contigs $\{AACCC, CCCCG, GGTTT\}$. In addition, other assembled DNA signatures of s with equi-length contigs of length $n = 5$ exist, depending on the parameter r . For example, an

assembled DNA signature of s using $r = 5$ equi-length contigs of length $n = 5$ could use the contigs

$$\{AAAAA, AAACC, CGGGG, GGGGT, GGTTT\}.$$

Approximated Information Distance (AID)

For a finite set X , we denote by $|X|$ the cardinality of X , that is the number of elements in X . Given a set of sequences $S = \{s_1, s_2, \dots, s_n\}$ we denote by $M_k(S)$ the set of all distinct k -mers that occur in all the sequences of S . In the case of a set consisting of a single sequence s , we write $M_k(s)$ to denote $M_k(\{s\})$.

The approximated information distance between two sequences s and t (introduced in [25] as a slight modification of a distance used in [53]) is defined as:

$$d_{\text{AID}}^k(s, t) = \frac{|M_k(s) \setminus M_k(t)| + |M_k(t) \setminus M_k(s)|}{|M_k(\{s, t\})|},$$

where for two sets X and Y , $X \setminus Y$ denotes the set difference between X and Y , that is, the set of elements that belong to X but not to Y .

The distance $d_{\text{AID}}^k(s, t)$ was used for most of the computations of pairwise distances between conventional DNA signatures in this paper.

The notion of approximated information distance between two sequences can now be extended to that of generalized approximated information distance between two sets of sequences S and T , as:

$$d_{\text{AID}}^k(S, T) = \frac{|M_k(S) \setminus M_k(T)| + |M_k(T) \setminus M_k(S)|}{|M_k(S \cup T)|}.$$

This generalization of the approximated information distance preserves the original meaning of the concept as the ratio between the number of noncommon k -mers of the two sets S and T and the total number of k -mers that occur in S or in T (or both). This distance was used to compute distances between conventional, composite and assembled DNA signatures in this

paper.

The next Proposition leads to a formula for the computation of the generalized approximated information distance, as well as gives a theoretical upper bound for the generalized approximated information distance in the case of fully-assembled DNA signatures. The following auxiliary lemma follows from standard set theory arguments.

Lemma 5.4.1 *Let s be a sequence and S, T be two finite sets of sequences over the DNA alphabet $\{A, C, G, T\}$, and let $k \geq 2$ be an integer. The following statements hold true.*

1. *If $S \subseteq T$ then $|M_k(S)| \leq |M_k(T)|$ and*

$$|M_k(S \cup T)| = |M_k(T)|,$$

2. *If every sequence in S is a subsequence of a given sequence s , then $|M_k(S) \cup M_k(s)| = |M_k(s)|$,*

3. *The number of distinct k -mers that occur in S but not in T is $|M_k(S) \setminus M_k(T)| = |M_k(S \cup T)| - |M_k(T)|$,*

4. $|M_k(S)| = \#FCGR_k(S)$,

where for a numerical matrix A we denote by $\#(A)$ or $\#A$ the number of non-zero entries of A .

Proposition 5.4.2 *Let s be a sequence and let S, T be two sets of sequences. The following statements hold true.*

1. $d_{\text{AID}}^k(S, T) = 2 - \frac{|M_k(S)| + |M_k(T)|}{|M_k(S \cup T)|}$

2. *If $s = s_1 s_2 \dots s_r$ and each s_i is of length n , $n > k$, then*

$$d_{\text{AID}}^k(\{s_1 s_2 \dots s_r\}, s) \leq \frac{\min\{(r-1)(k-1), |M_k(s)|\}}{|M_k(s)|}.$$

3. *There is a sequence s for which the above relation holds with “=”.*

Proof The first statement follows from Lemma 5.4.1.3, by noting that $d_{\text{AID}}^k(S, T)$ equals

$$\frac{(|M_k(S \cup T)| - |M_k(T)|) + (|M_k(S \cup T)| - |M_k(S)|)}{|M_k(S \cup T)|}$$

which is indeed equal to the required formula.

For the second statement, let $S = \{s_1, s_2, \dots, s_r\}$ and $T = \{s\}$. By the definition of the generalized information distance, $d_{\text{AID}}^k(\{s_1, \dots, s_r\}, s)$ equals a fraction, where the numerator is the sum between the number of distinct k -mers that appear in $\{s_1, \dots, s_r\}$ but not in s , and the number of distinct k -mers that appear in s but not in $\{s_1, \dots, s_r\}$. The first term of this sum is obviously zero, since s_i are contigs that span the sequence s . Thus, the numerator of this fraction is the second term of the sum, namely the number of distinct k -mers that appear in $s = s_1 s_2 \dots s_r$ but not in $\{s_1, \dots, s_r\}$. We can count these k -mers by noticing that the only k -mers that appear in s but not in $\{s_1, \dots, s_r\}$, are the ones that span consecutive contigs.

We now note that each joint of two contigs $s_i s_{i+1}$ contains at most $(k - 1)$ distinct k -mers that span both contigs, and that s contains $(r - 1)$ such joints $s_i s_{i+1}$. Thus, the total number of k -mers of s , that are in s but not in $\{s_1, \dots, s_r\}$, is at most $(r - 1) \cdot (k - 1)$.

Since the denominator of the fraction is, by Lemma 5.4.1.2, $|M_k(s) \cup M_k(\{s_1, s_2, \dots, s_r\})| = |M_k(s)|$, we have that

$$d_{\text{AID}}^k(\{s_1, \dots, s_r\}, s) \leq \frac{0 + (r - 1)(k - 1)}{|M_k(s)|}.$$

Since the approximated information distance ranges between 0 and 1, the required inequality follows.

For the third statement, an example of a sequence where the upper bound of the distance between the conventional DNA signature of the sequence and the fully-assembled DNA signature of its contigs is reached is the sequence

$$s = \text{AAAACCCCGGGGTTTT},$$

with $k = 3$ and $n = r = 4$. Then s contains exactly 10 different 3-mers, that is, $|M_3(s)| = 10$, and $(r-1) \cdot (k-1) / |M_3(s)| = 0.6$. On the other hand, let $s_1 = AAAA$, $s_2 = CCCC$, $s_3 = GGGG$, $s_4 = TTTT$. Then we have $|M_3(\{s_1, s_2, s_3, s_4\})| = 4$, since only 4 distinct 3-mers, namely AAA, CCC, GGG and TTT can be found in this set, and thus

$$d_{AID}^3(\{s_1, s_2, s_3, s_4\}, s) = 2 - \frac{4 + 10}{10} = 0.6,$$

which equals the given upper bound.

Remark that, by Proposition 5.4.2.1, the generalized approximated distance between two sets of sequences S and T can be now computed as

$$d_{AID}^k(S, T) = 2 - \frac{\#FCGR_k(S) + \#FCGR_k(T)}{\#(FCGR_k(S) + FCGR_k(T))},$$

which is the formula that was used for all generalized approximated information distance calculations in this paper.

Remark also that the upper bound determined in Proposition 5.4.2.2 for the generalized approximated information distance, in the case of the comparison between the conventional DNA signature of a sequence and the fully-assembled DNA signature of its r contigs of length n , is the one illustrated in Column (A') of Table 5.2.

Multi-Dimensional Scaling and separation assessment

To visualize the interrelationships among DNA signatures originating from a pair of genomes, and thus to visually assess whether separation was achieved, we used Multi-Dimensional Scaling (MDS). MDS is an information visualization technique introduced by Kruskal in [67]. MDS takes as input a distance matrix that contains the pairwise distances among a set of items (here the items are DNA signatures), and outputs a spatial representation of the items in a common Euclidean space. Each item is represented as a point, and the spatial distance between any

two points corresponds to the distance between the items in the distance matrix. Objects with a smaller pairwise distance will result in points that are close to each other, while objects with a larger pairwise distance will become points that are far apart.

Concretely, classical MDS, which we use in this paper, receives as input an $m \times m$ distance matrix $(\Delta(i, j))_{1 \leq i, j \leq m}$ of the pairwise distances between any two items in the set. The output of classical MDS consists of m 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 m points $p_1, p_2, \dots, p_m \in \mathbb{R}^q$ such that $d(i, j) = \|p_i - p_j\| \approx f(\Delta(i, j))$ for all $i, j \in \{1, \dots, m\}$ 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 $(m - 1)$ and the points are recovered from the eigenvalues and eigenvectors of the input $m \times m$ distance matrix. If we choose $q = 3$, the result of classical MDS is an approximation of the original $(m - 1)$ -dimensional space as a three-dimensional map, such as the Molecular Distance Maps in this paper. Throughout the paper, for consistency, all Molecular Distance Maps have been scaled so that the x -, y -, and z - coordinates always span the interval $[-1, 1]$. The formula used for scaling is $x_{\text{sca}} = 2 \cdot \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) - 1$, 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_{sca} and z_{sca} . In all Molecular Distance Maps displayed in this paper, the origin of coordinates $(0, 0, 0)$ is the center of the depicted cube, and the parallel edges of the cube are parallel to one of the x -, y -, and z - axis respectively. The maps have been rotated for optimal visualization and, for each of the axes, the length units are displayed only on one of the four edges of the cube that are parallel to it.

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 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, z) -coordinates of a point representing the DNA signature of a particular DNA fragment of *H. Sapiens* in Figure 5.1 will not be the same as the (x, y, z) -coordinates of the point representing the same DNA fragment in Figure 5.2.

For a given Molecular Distance Map, *k-means clustering* [57] was used to assess whether separation of the DNA sequences by organism was achieved. The reason for this choice were that in all computed Molecular Distance Maps the number of clusters was known *a priori*, $k = 2$ (not to be confused with *k*-mers, where *k* has a different meaning), that the clusters had approximately the same number of points and thus the prior probability of the two clusters was the same, and that in most cases the clusters were somewhat spherical in shape. Moreover, the use of *k*-means yielded satisfactory results in the majority of cases.

The *k*-means clustering algorithm proceeds as follows. Suppose S_1 is the set of points originating from the genome of one of the organisms, and S_2 is the set of points originating from the second one. *k*-means assigns labels *A* and *B* to all given points, in its attempt to cluster them into two clusters, *A* and *B*. The *k*-means accuracy score is computed by counting how many points were assigned correctly to their cluster, that is,

$$Acc = \frac{\max\{|A_{S_1}| + |B_{S_2}|, |B_{S_1}| + |A_{S_2}|\}}{|S_1| + |S_2|}$$

where A_{S_1} is the set of points in the cluster *A* that belong to the set S_1 , and B_{S_2} is the set of points in the cluster *B* that belong to the set S_2 (B_{S_1} and A_{S_2} are defined similarly). If label *A* would correspond to species S_1 , and *B* to species S_2 , the quantity $|A_{S_1}| + |B_{S_2}|$ would represent the number of points that have been correctly classified in this Molecular Distance Map, while $|B_{S_1}| + |A_{S_2}|$ would represent the number of points that have been incorrectly classified. As a number, *Acc* is a quantity between 0.5 and 1, with 50% indicating the worst clustering, and 100% indicating perfect clustering. For this paper, any Molecular Distance Map with an accuracy greater than 85% was interpreted as achieving separation of points by species.

In some cases the accuracy was less than 85% in spite of the fact that separation of clusters could clearly be observed visually. A closer look at those cases revealed that they were gen-

erally plots similar to Figure 5.3, that is, consisting of two long and thin clusters. In addition, in those plots the clusters were closer to each other than in Figure 5.3. In such cases, k -means erroneously labelled the top halves of the two clusters by A , and the two bottom halves by B . For such situations, where the k -means clustering algorithm had a relatively low accuracy score but visual separation was nevertheless observed, we verified the existence of a plane that completely separated the two clusters. That is, if cluster S_1 had n_1 points of coordinates $(x_{i_1}, x_{i_2}, x_{i_3})$, where $1 \leq i \leq n_1$, and cluster S_2 had n_2 points $(y_{j_1}, y_{j_2}, y_{j_3})$, where $1 \leq j \leq n_2$, then our Mathematica-based code [68] was used to find one (out of possibly infinitely many) solutions to the system of equations with unknowns a, b, c, d :

$$\begin{cases} a \cdot x_{i_1} + b \cdot x_{i_2} + c \cdot x_{i_3} + d > 0, & i = 1, \dots, n_1 \\ a \cdot y_{j_1} + b \cdot y_{j_2} + c \cdot y_{j_3} + d < 0, & j = 1, \dots, n_2 \end{cases}$$

that is, it found the equation $ax + by + cz + d = 0$ of a plane with the property that the points of the cluster S_1 are situated on one of its sides, while those of cluster S_2 are situated on the other. For example, in Figure 5.5, the equation of a plane computed by this method, that completely separates the points originating from *H. sapiens* from those originating from *P. troglodytes*, is $x + 0.918y + 0.37z + 0.0002 = 0$.

For Molecular Distance Maps with more complex cluster shapes, where k -means accuracy is low and separating planes do not exist, the use of other clustering methods such as density-based spatial clustering of applications with noise (DBSCAN) [69] would have to be explored to see if separation is achieved.

The webtool MoDMap3D, [58], illustrates the 3D Molecular Distance Maps that correspond to each of the comparisons listed in Table 5.2, in the same way the Molecular Distance Map in Figure 5.1 illustrates the positive separation result listed in Table 5.2, subtable Animalia, line 1. The webtool MoDMap3D is, moreover, interactive, and allows for an in-depth exploration of each particular 3D Molecular Distance Map. After first selecting the pair of genomes to be compared, the user can navigate in the three-dimensional space of their DNA

signatures: clicking on any point in the map will display information about the DNA fragment represented by that point, such as its NCBI accession number or assembly number, scientific name of the organism it originates from, chromosome or contig/scaffold number, length of the subsequence in bp, and fragment number from the original sequence.

Software

The code for running the experiments [68] was written in Wolfram Mathematica, and was used for the generation of FCGRs, the computation of composite and assembled DNA signatures, the calculation of distance matrices, the creation of the 3D Molecular Distance Maps, and the computation of the separating planes.

Remarks

One observation should be made about the genome assemblies at contig/scaffold level in the dataset. The general intent was for the 150 kbp DNA fragments from a given genome not to be overlapping. This is because sequence overlaps could result in artificially smaller intragenomic distances due to the increase in sequences' similarities, and this could potentially lead to false positive cluster separations. However, some overlap may have been unavoidable in the cases where only contig/scaffold level data was available. The availability of contig/scaffold data only may thus explain why in Table 5.2 the accuracy scores do not always decrease uniformly, as expected, when one compares the pivot organism with organisms more and more closely related to it.

Another observation should be made about the length of sequences analyzed. When computing composite DNA signatures, the signature of the mitochondrial genome (or entire chloroplast or plasmid) was appended to that of each 150 kbp nDNA fragment. This, in some sense, magnifies the role of the organellar genome in the composite signature. Depending on the application, one can generalize Definition 5.4 to a weighted additive DNA signature which gives different weights to the different types of DNA that compose it.

We now discuss some limitations of the proposed methods. First, note that assembled DNA signatures as defined here use equi-length contigs. Preliminary computational experiments, illustrated in Table 5.2, columns (B') and (C'), show the results of comparisons between a conventional nDNA signature and variable-length assembled DNA signatures of the same fragment. In those experiments, contig lengths are drawn from a normal distribution $N(\mu, \sigma)$ with mean $\mu = n$ (the length of the contig in the corresponding equi-length contig experiment) and variance $\sigma = 40$. The table shows that the performance of assembled DNA signatures using variable-length contigs is comparable with the performance of those using equi-length contigs. This indicates that both equi-length and variable-length contigs assembled DNA signatures could be reliable approximations of conventional genomic signatures, depending on the application. Additional exploration is needed to confirm this hypothesis.

Second, every computational experiment in this study is a comparison between DNA signatures of genomic sequences belonging to two different organisms. Further analysis is needed to determine if the positive preliminary results on the discriminating power of composite and composite-assembled DNA signatures extend successfully to multi-genome comparisons. A necessary step for such an experiment would be a thorough investigation of intragenomic variations of FCGRs and finding a method to determine, for each genome, a single “representative” FCGR matrix to successfully represent that genome.

Third, we mention a case where separation by organism could not be achieved, even when using composite DNA signatures (nDNA and cpDNA). This is the pairwise comparison between a cultivated pepper *Capsicum annuum* L, cultivar *Zunla-1* (domesticated) and its wild progenitor *Capsicum annuum* var. *glabriusculum*, cultivar *Chiltepin* (wild), see Figure 5.7.

Several directions of future research stem from the observation that the function $FCGR_k$ is a quasi-homomorphism from the set of all DNA sequences with the operation of catenation, to the set of $2^k \times 2^k$ matrices with the operation of addition, in the sense that for sequences s, t , we have

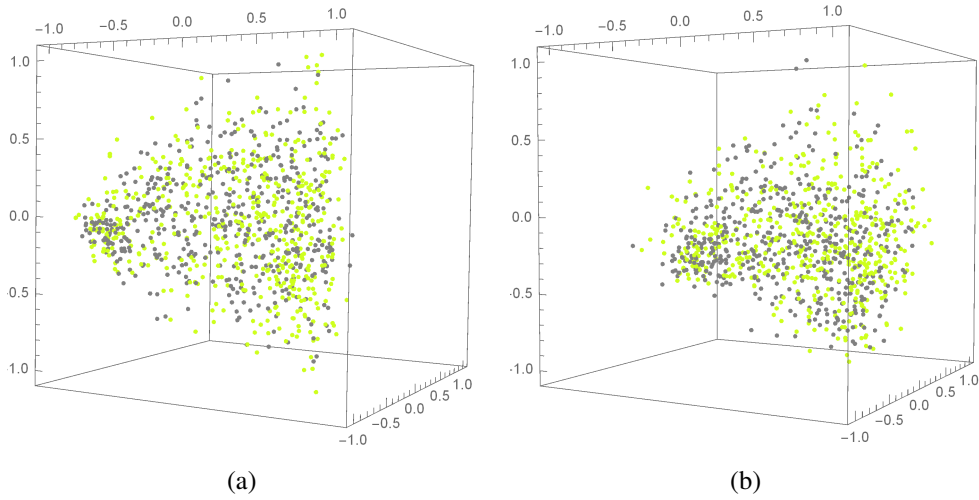


Figure 5.7: (a) Conventional nDNA signatures, and (b) composite (nDNA + cpDNA) signatures of *Capsicum annuum* L, cultivar *Zunla-1* (domesticated) shown in light green, and *Capsicum annuum* var. *glabriusculum*, cultivar *Chiltepin* (wild) shown in grey.

$$FCGR_k(st) \approx FCGR_k(s) + FCGR_k(t).$$

The definition of $FCGR_k$ can be easily modified to make it an exact homomorphism by, e.g., defining a marked catenation of sequences s and t as $s \cdot t = s\$t$, with $\$$ a new symbol, and constructing $FCGR_k$ so as to not count any k -mer that includes the symbol $\$$. Next steps in the exploration of the mathematical properties of additive DNA signatures include studying the implications of the homomorphic, structure-preserving, nature of $FCGR_k$, as well as extensions of the concept of additive DNA signature, to, e.g., weighted additive DNA signatures which would give different weights to the different types of DNA that compose it.

Bibliography

- [1] Jeffrey, H.J.: Chaos game representation of gene structure. *Nucleic Acids Research* **18**(8), 2163–2170 (1990)
- [2] Jeffrey, H.J.: Chaos game visualization of sequences. *Computers & Graphics* **16**(1), 25–33 (1992)
- [3] Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B.: Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution* **16**(10), 1391–1399 (1999)
- [4] Deschavanne, P.J., Giron, A., Vilain, J., Dufraigne, C., Fertil, B.: Genomic signature is preserved in short DNA fragments. In: *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pp. 161–167 (2000)
- [5] Karlin, S., Burge, C.: Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* **11**(7), 283–290 (1995)
- [6] Karlin, S., Campbell, A.M., Mrázek, J.: Comparative DNA analysis across diverse genomes. *Annual Review of Genetics* **32**, 185–225 (1998)
- [7] Vinga, S., Almeida, J.S.: Alignment-free sequence comparison - A review. *Bioinformatics* **19**(4), 513–523 (2003)
- [8] Nalbantoglu, O.U., Sayood, K.: Computational Genomic Signatures. *Synthesis Lectures on Biomedical Engineering* **6**(2), 1–129 (2011)
- [9] Bonham-Carter, O., Steele, J., Bastola, D.: Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in Bioinformatics* **15**(6), 890–905 (2013)
- [10] Schwende, I., Pham, T.D.: Pattern recognition and probabilistic measures in alignment-free sequence analysis. *Briefings in Bioinformatics* **15**(3), 354–68 (2014)

- [11] Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M.S., Sun, F.: New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. *Briefings in Bioinformatics* **15**(3), 343–353 (2014)
- [12] Burma, P.K., Raj, A., Deb, J.K., Brahmachari, S.K.: Genome analysis: A new approach for visualization of sequence organization in genomes. *Journal of Biosciences* **17**(4), 395–411 (1992)
- [13] Hill, K.A., Singh, S.M.: The evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes. *Genome* **40**(3), 342–356 (1997)
- [14] Hao, B., Lee, H.C., Zhang, S.-y.: Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals* **11**(6), 825–836 (2000)
- [15] Dutta, C., Das, J.: Mathematical characterization of Chaos Game Representation. New algorithms for nucleotide sequence analysis. *Journal of Molecular Biology* **228**(3), 715–719 (1992)
- [16] Goldman, N.: Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Research* **21**(10), 2487–2491 (1993)
- [17] Almeida, J.S., Carriço, J.a.A., Marezek, A., Noble, P.A., Fletcher, M.: Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* **17**(5), 429–437 (2001)
- [18] Almeida, J.S.: Sequence analysis by iterated maps, a review. *Briefings in Bioinformatics* **15**(3), 369–75 (2014)
- [19] Wang, Y., Hill, K., Singh, S., Kari, L.: The spectrum of genomic signatures: From dinucleotides to chaos game representation. *Gene* **346**, 173–185 (2005)
- [20] Kari, L., Hill, K.A., Sayem, A.S., Karamichalis, R., Bryans, N., Davis, K., Dattani, N.S.: Mapping the Space of Genomic Signatures. *PLoS ONE* **10**(5) (2015)

- [21] Edwards, S.V., Fertil, B., Giron, A., Deschavanne, P.J.: A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic Biology* **51**(4), 599–613 (2002)
- [22] Deschavanne, P., DuBow, M.S., Regeard, C.: The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virology Journal* **7**, 163 (2010)
- [23] Pandit, A., Sinha, S.: Using genomic signatures for HIV-1 sub-typing. *BMC Bioinformatics* **11**(Suppl 1), 26 (2010)
- [24] Hatje, K., Kollmar, M.: A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method. *Frontiers in Plant Science* **3**(192) (2012)
- [25] Karamichalis, R., Kari, L., Konstantinidis, S., Kopecki, S.: An investigation into inter- and intragenomic variations of graphic genomic signatures. *BMC Bioinformatics* **16**(1), 246 (2015)
- [26] Wu, T.J., Huang, Y.H., Li, L.A.: Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics* **21**(22), 4125–4132 (2005)
- [27] Höhl, M., Rigoutsos, I., Ragan, M.A.: Pattern-based phylogenetic distance estimation and tree reconstruction. *Evolutionary Bioinformatics* **2**, 359–375 (2006)
- [28] Höhl, M., Ragan, M.A.: Is multiple-sequence alignment required for accurate inference of phylogeny? *Systematic Biology* **56**(2), 206–221 (2007)
- [29] Dai, Q., Yang, Y., Wang, T.: Markov model plus k-word distributions: A synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **24**(20), 2296–2302 (2008)

- [30] Guyon, F., Brochier-Armanet, C., Guénoche, A.: Comparison of alignment free string distances for complete genome phylogeny. *Advances in Data Analysis and Classification* **3**(2), 95–108 (2009)
- [31] Jayalakshmi, R., Natarajan, R., Vivekanandan, M., Natarajan, G.S.: Alignment-free sequence comparison using N-dimensional similarity space. *Current Computer-Aided Drug Design* **6**(4), 290–296 (2010)
- [32] Haubold, B.: Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics* **15**(3), 407–18 (2014)
- [33] Fiser, A., Tusnády, G.E., Simon, I.: Chaos game representation of protein structures. *Journal of Molecular Graphics* **12**(4), 302–304 (1994)
- [34] Basu, S., Pan, A., Dutta, C., Das, J.: Chaos game representation of proteins. *Journal of Molecular Graphics and Modelling* **15**(5), 279–289 (1997)
- [35] Yu, Z.-G., Anh, V., Lau, K.-S.: Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *Journal of Theoretical Biology* **226**(3), 341–348 (2004)
- [36] Yang, J.-Y., Peng, Z.-L., Yu, Z.-G., Zhang, R.-J., Anh, V., Wang, D.: Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *Journal of Theoretical Biology* **257**(4), 618–626 (2009)
- [37] Randić, M., Novič, M., Vikić-Topić, D., Plašić, D.: Novel numerical and graphical representation of DNA sequences and proteins. *SAR and QSAR in Environmental Research* **17**(6), 583–595 (2006)
- [38] Almeida, J.S., Vinga, S.: Biological sequences as pictures: a generic two dimensional solution for iterated maps. *BMC Bioinformatics* **10**, 100 (2009)

- [39] Almeida, J.S., Vinga, S.: Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics* **3**, 6 (2002)
- [40] Almeida, J.S., Vinga, S.: Computing distribution of scale independent motifs in biological sequences. *Algorithms for Molecular Biology* **1**, 18 (2006)
- [41] Fu, W., Wang, Y., Lu, D.: Multifractal analysis of genomic sequences CGR images. In: *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 5, pp. 4783–4786 (2005)
- [42] Fu, W., Wang, Y., Lu, D.: Multifractal analysis of genomes sequences' CGR graph. *Journal of Biomedical Engineering* **24**(3), 522–525 (2007)
- [43] Vélez, P.E., Garreta, L.E., Martínez, E., Díaz, N., Amador, S., Tischer, I., Gutiérrez, J.M., Moreno, P.A.: The *Caenorhabditis elegans* genome: A multifractal analysis. *Genetics and Molecular Research* **9**(2), 949–965 (2010)
- [44] Moreno, P.A., Vélez, P.E., Martínez, E., Garreta, L.E., Díaz, N., Amador, S., Tischer, I., Gutiérrez, J.M., Naik, A.K., Tobar, F., García, F.: The human genome: a multifractal analysis. *BMC Genomics* **12**(1), 506 (2011)
- [45] Pandit, A., Dasanna, A.K., Sinha, S.: Multifractal analysis of HIV-1 genomes. *Molecular Phylogenetics and Evolution* **62**(2), 756–763 (2012)
- [46] Pal, M., Satisha, B., Srinivas, K., Madhusudana Rao, P., Manimaran, P.: Multifractal detrended cross-correlation analysis of coding and non-coding DNA sequences through chaos-game representation. *Physica A: Statistical Mechanics and its Applications* **436**, 596–603 (2015)
- [47] Oliver, J.L., Bernaola-Galván, P., Guerrero-García, J., Román-Roldán, R.: Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretical Biology* **160**(4), 457–470 (1993)

- [48] Vinga, S., Almeida, J.S.: Rényi continuous entropy of DNA sequences. *Journal of Theoretical Biology* **231**(3), 377–388 (2004)
- [49] Vinga, S., Almeida, J.S.: Local Rényi entropic profiles of DNA sequences. *BMC Bioinformatics* **8**, 393 (2007)
- [50] Joseph, J., Sasikumar, R.: Chaos game representation for comparison of whole genomes. *BMC Bioinformatics* **7**, 243 (2006)
- [51] Tanchotsrinon, W., Lursinsap, C., Poovorawan, Y.: A high performance prediction of HPV genotypes by Chaos game representation and singular value decomposition. *BMC Bioinformatics* **16**(1) (2015)
- [52] Campbell, A.M., Mrázek, J., Karlin, S.: Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America* **96**(16), 9184–9189 (1999)
- [53] Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.M.B.: The similarity metric. *Information Theory, IEEE Transactions on* **50**(12), 3250–3264 (2004)
- [54] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
- [55] Iversen, G.R., Gergen, M., Gergen, M.M.: *Statistics: The Conceptual Approach*. Springer, Berlin Heidelberg (1997)
- [56] Krause, E.F.: *Taxicab Geometry: An Adventure in Non-Euclidean geometry*. Courier Dover Publications, Mineola, New York (2012)
- [57] Lloyd, S.: Least squares quantization in pcm. *IEEE Transactions on Information Theory* **28**(2), 129–137 (1982)

- [58] Karamichalis, R.: Molecular Distance Map Interactive Webtool. <https://github.com/rallis/ModMap3D> (2015)
- [59] Jameson, N.M., Hou, Z.-C., Sterner, K.N., Weckle, A., Goodman, M., Steiper, M.E., Wildman, D.E.: Genomic data reject the hypothesis of a prosimian primate clade. *Journal of Human Evolution* **61**(3), 295–305 (2011)
- [60] Perelman, P., Johnson, W.E., Roos, C., Seunez, H.N., Horvath, J.E., Moreira, M.A.M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M.P.C., Silva, A., O'Brien, S.J., Pecon-Slattery, J.: A molecular phylogeny of living primates. *PLoS Genet* **7**(3), 1001342 (2011)
- [61] Chatterjee, H., Ho, S., Barnes, I., Groves, C.: Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evolutionary Biology* **9**(1), 259 (2009)
- [62] Li, H., Homer, N.: A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* **11**(5), 473–483 (2010)
- [63] Thompson, J.D., Linard, B., Lecompte, O., Poch, O.: A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS ONE* **6**(3), 18093 (2011)
- [64] Grossmann, L., Jensen, M., Heider, D., Jost, S., Glücksman, E., Hartikainen, H., Mahamdallie, S.S., Gardner, M., Hoffmann, D., Bass, D., et al.: Protistan community analysis: key findings of a large-scale molecular sampling. *The ISME journal* (Feb 2016)
- [65] Lange, A., Jost, S., Heider, D., Bock, C., Budeus, B., Schilling, E., Strittmatter, A., Boenigk, J., Hoffmann, D.: Ampliconduo: A split-sample filtering protocol for high-throughput amplicon sequencing of microbial communities. *PLoS ONE* **10**(11), 0141590 (2015)

- [66] Bikel, S., Valdez-Lara, A., Cornejo-Granados, F., Rico, K., Canizales-Quinteros, S., Soberón, X., Del Pozo-Yauner, L., Ochoa-Leyva, A.: Combining metagenomics, meta-transcriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Computational and Structural Biotechnology Journal* **13**, 390–401 (2015)
- [67] Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**(1), 1–27 (1964)
- [68] Karamichalis, R.: Source code for computing FCGR matrices, distance matrices, MultiDimensional Scaling and separation planes. <https://github.com/rallis/GenomicSignatures> (2015)
- [69] Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. *Conference on Knowledge Discovery and Data Mining* **96**(34), 226–231 (1996)

Chapter 6

Molecular Distance Maps 3D

(MoDMaps3D)

6.1 Introduction

In order to identify and classify species based on genetic evidence, various alignment-free methods for genome comparison have been proposed. In an effort to visualize these interrelationships between DNA sequences, we propose the interactive webtool Molecular Distance Maps (MoDMaps3D).

6.2 Methods

The flow of our proposed algorithm is as follows. We start with n DNA sequences. First, we compute the CGR of each one of them CGR_1, \dots, CGR_n . Secondly we compute all pairwise distances between these CGRs, and we store the values in a (symmetric) distance matrix $D = [d_{ij}]$ with $i, j \in \{1, \dots, n\}$. Finally, we run MDS with input this distance matrix D and we get as output a $n \times 5$ vector representing a 5D representation of the original DNA sequences. Then, we can visualise any triplet of these 5 dimensions in a regular 3D space, producing eventually a three dimensional molecular distance map.

MoDMaps3D can be used in two main ways: (i) Exploring pre-build maps of various DNA sequences or, (ii) Building a new Molecular Distance Map which involves three computational steps.

6.3 Software Description

MoDMaps3D has been written in Javascript. MoDMaps3D uses jQuery (a free open-source cross-platform Javascript library), Bootstrap (a free open-source collection of tools for web applications), and Three.js (a cross-browser JavaScript library using WebGL for displaying animated 3D computer graphics in a web browser). In addition to these, the Parallel.js library is used for parallel computation when applicable.

For exploring a Molecular Distance Map, the user can select a category from the drop-down menu, view the header of the file for relevant information and then visualize it. There, MoDMaps3D uses internally the Three.js library for building 3D computer graphics. On Figure 6.1, you can see a screenshot of MoDMap for Phylum Vertebrata from the webtool MoDMaps3D. Each map is accompanied by a set of panels, left and right. Both of the panels can be minimized/maximized at any time either by clicking the top button of each panel or by using the shortcuts available (Ctrl+Alt+N for left, Ctrl+Alt+M for right).

On the left panel, from top to bottom, one can find the basic keyboard commands to navigate through the MoDMap. You can move in the four directions using keyboard keys A, D, W, S for left, right, up and down respectively. You can zoom in by E and Q. You can rotate the map by left click and drag accordingly. You can also navigate using Ctrl + ArrowKey. A detailed list of shortcuts can be found <https://github.com/rallis/MoDMaps3D>. Next, one can see the triplet of selected dimensions (out of 5) that are being plotted at any given time. User can also increase/decrease the radius of each point depicted from the radius field. By changing any of these values and hitting the button Re-Draw a new map with the settings specified will be drawn. Underneath, user can select to see CGR image of a selected point (it will be visible on

the right panel upon selection of a point), change the highlight color for selecting/highlighting points on the map, see the distance menu to compute the distance between any pair of points, and enable/disable selection of points while hovering the mouse instead of double clicking on them. Under this, user can find a legend of the map containing a short description and a list of colour, name and number of points for each category of sequences being plotted.

Another feature of MoDMap is that one can search and highlight any subset of points (DNA sequences) which have a common label in their description. On the right panel, user can search for any keyword (complete or partial, as infix) in the dataset of the map under consideration. A shortcut for moving to search field is also available (Ctrl+F). Under this, the user can see the CGR image of a point, if this option has been checked from the left panel, and if a point has been selected on the map. Under CGR, the user can change the highlight color, if this option has been checked from the left panel. Under this menu, is the Point Info Panel. This panel contains all available information for the currently selected point: its index in the current map, its accession number in NCBI (a direct link to that page is also provided), its full scientific name (with the option of one-click search in Google images), its length in base pairs and its taxonomy (if applicable). Depending on the type of map, fragment or haplogroup fields may be available as well.

For building a Molecular Distance Map, the user can input the sequences (as NCBI accession numbers) to be included in the map, adjust the parameters (number of different colors, length of k -mers, taxonomy information to be included or not) and build the map. All three steps described in Chapter 2.2 are computed on the user's browser and when the computation has finished, the user is prompted to visualize the MoDMap produced. In case of large datasets (greater than 500 sequences), the user is prompted to continue the computation off the browser, in either Wolfram Mathematica or Python. All files necessary to continue the computation are provided to the user during runtime.

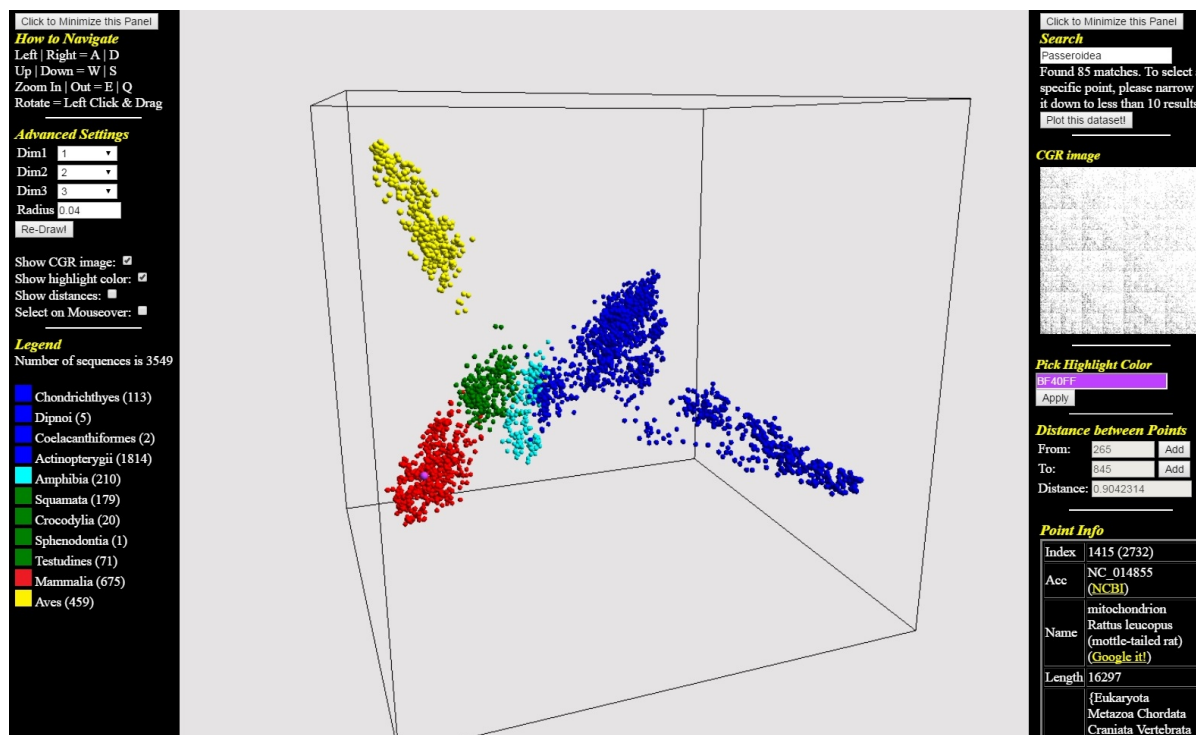


Figure 6.1: Molecular Distance Map of Phylum Vertebrata, consisting of 3,158 mtDNA sequences. Blue, cyan, green, red and yellow, represent fish, amphibia, reptiles, mammals and birds mtDNA genomes respectively. Enlarged version of left and right panel can be found in Figure 6.2

Click to Minimize this Panel

How to Navigate
 Left | Right = A | D
 Up | Down = W | S
 Zoom In | Out = E | Q
 Rotate = Left Click & Drag

Advanced Settings

Dim1

Dim2

Dim3

Radius

Show CGR image:

Show highlight color:

Show distances:

Select on Mouseover:

Legend
 Number of sequences is 3549

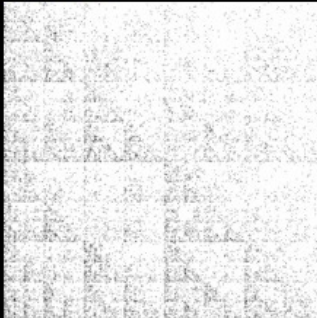
- Chondrichthyes (113)
- Dipnoi (5)
- Coelacanthiformes (2)
- Actinopterygii (1814)
- Amphibia (210)
- Squamata (179)
- Crocodylia (20)
- Sphenodontia (1)
- Testudines (71)
- Mammalia (675)
- Aves (459)

Click to Minimize this Panel

Search

 Found 85 matches. To select a specific point, please narrow it down to less than 10 results.

CGR image



Pick Highlight Color

Distance between Points

From:

To:

Distance:

Point Info

Index	1415 (2732)
Acc	NC_014855 (NCBI)
Name	mitochondrion Rattus leucopus (mottle-tailed rat) (Google it!)
Length	16297
	{Eukaryota Metazoa Chordata Craniata Vertebrata

Figure 6.2: Enlarged version of left and right panel of MoDMaps3D.

Chapter 7

Conclusion

In this thesis, we continue the exploration of the concept of genomic signature and we show that Chaos Game Representation (CGR) can quantitatively express the syntactical properties of genomic DNA sequences. We also show in chapter 4 that intragenomic variation of genomic signatures, although distance dependent, is in general smaller than intergenomic one, giving support to the idea that genomic signature as a quantitative characteristic is maintained throughout the genome of a species. These findings suggest that CGR can play the role of genomic signature. In chapter 5, we extend the notion of genomic signature to overcome its limitations and to make it applicable to more real-case scenarios using smaller sequence length (compared to that produced from next generation sequencing) while still maintaining its discriminative power. An interactive webtool is also developed to support a user-friendly exploration of the datasets and the methods presented in this thesis.

As a future work, we focus on building a representative signature of a species. Such a signature will incorporate information from various loci (and possibly different types of DNA as well) to uniquely identify any species. An extension of it would be to build, if possible, representative signatures of taxonomic categories (similar to signatures with wildcards where mismatches are allowed).

Another topic we are interested in is haplogroup identification using genomic signatures on

H.sapiens mtDNA. Preliminary results suggest that it is possible to identify haplogroups and track maternal lineage using genomic signature of mitochondrial DNA. A further extension would be to look at paternal lineage as well by using the Y chromosome. This is more challenging computationally than mtDNA analysis, due to the sequence size differences. (human mtDNA is about 16,000 bp long while the Y chromosome is about 60,000,000 bp long).

Another venue of future research is to explore the possibility of using this method to distinguish between healthy and unhealthy individuals for a disease with a genetic cause (preferably with large scale mutations), based solely on genomic signatures. Finally, increasing the resolution of a CGR (value of k) significantly ($k = 20$) could potentially help identifying Single Nucleotide Polymorphism (SNP) in the human genome.

Appendix A

Copyright Releases

Chapter 3 contains the article “Mapping the space of genomic signatures” from *PLoS ONE*. According to their website <http://journals.plos.org/plosone/s/content-license>

“PLOS applies the Creative Commons Attribution (CC BY) license to works we publish. This license was developed to facilitate open access namely, free immediate access to, and unrestricted reuse of, original works of all types. Under this license, *authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose. Anyone may copy, distribute or reuse these articles, as long as the author and original source are properly cited.*”

“Using PLOS Content: *No permission is required from the authors or the publishers to reuse or repurpose PLOS content provided the original article is cited. In most cases, appropriate attribution can be provided by simply citing the original article. If the item you plan to reuse is not part of a published article (e.g., a featured issue image), then indicate the originator of the work, and the volume, issue, and date of the journal in which the item appeared. For any reuse or redistribution of a work, you must also make clear the license terms under which the work was published.*”

Chapters 4 and 5 contain the articles “An investigation of inter- and intragenomic variations of graphic genomic signatures” and “Additive methods for genomic signatures” from *BMC*

Bioinformatics. According to their website <https://bmcbioinformatics.biomedcentral.com/submission-guidelines/copyright>

“Copyright on any open access article in a journal published by BioMed Central is retained by the author(s). Authors grant BioMed Central a license to publish the article and identify itself as the original publisher. Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified. The Creative Commons Attribution License 4.0 formalizes these and other terms and conditions of publishing articles.”

and <http://old.biomedcentral.com/bmcbioinformatics/about/faq/journal-copyright-policy>

“All articles published in BMC Bioinformatics are open access, which means the articles are universally and freely available online. In addition, *the authors retain copyright of their article, and grant any third party the right to use reproduce and disseminate the article*, subject to the terms of our copyright and license agreement. Allowing the authors to retain copyright of their work permits wider distribution of their work on the condition it is correctly attributed to the authors.”

Appendix B

Supplemental Material - Appendices per chapter

Here is a list of supplemental material and appendices per chapter.

Supplemental material for chapter 3 can be found

https://github.com/rallis/Supplemental_Material_Mapping_the_Space_of_Genomic_Signatures

Supplemental material and appendix for chapter 4 can be found

https://github.com/rallis/intraSupplemental_Material

Appendix for chapter 5 can be found

https://github.com/rallis/Thesis_Appendices

Appendix C

Errata

Since this thesis is formatted as integrated-article, the content of specific chapters should be exactly the same as those of published articles and no change is allowed.

Therefore, here we list all the modifications according to the comments provided by the thesis examiners.

Page 84, “The ideal Molecular Distance Map is a placement of n items as points in an $(n-1)$ -dimensional space.” should be removed.

Page 105, correct formula for $O_\alpha(i, j)$ is

$$O_\alpha(i, j) = \frac{\max\{s_{i,i}, s_{i,j}\}}{\min\{s_{i,i}, s_{i,j}\}} \cdot \frac{\sum_{l=1}^{100} \min\{c_{i,i}[l], c_{i,j}[l]\}}{\sum_{l=1}^{100} \max\{c_{i,i}[l], c_{i,j}[l]\}}.$$

Curriculum Vitae

Name: Rallis Karamichalis

Post-Secondary Education and Degrees: PhD in Computer Science
University of Western Ontario
London, Canada, 2012-2016

MSc in Theoretical Computer Science and Control Theory
Aristotle University of Thessaloniki
Thessaloniki, Greece, 2010-2012

BSc in Mathematics
Aristotle University of Thessaloniki
Thessaloniki, Greece, 2006-2010

Honours and Awards: First prize in Bioinformatics (×3)
UWO Research in Computer Science (UWORCS) 2014, 2015, 2016

Western Graduate Research Scholarship (WGRS)
University of Western Ontario, 2012-16

Erasmus Lifelong Learning Scholarship
Lund University, Sweden, 2009-10

Greek State Scholarships Foundation
Aristotle University, Greece, 2007-08

Bronze medal in the 47th International Mathematical Olympiad (IMO)
Bronze medal in the 9th Mediterranean Mathematical Olympiad (MMO)
Bronze medals (3) in the 20th, 22nd, 23rd Greek Mathematical Olympiad

Related Work Experience: Graduate Teaching Assistant
The University of Western Ontario
2012-2016

Publications:

1. Rallis Karamichalis, Lila Kari, Stavros Konstantinidis, Steffen Kopecki and Stephen Solis-Reyes, “Additive methods for genomic signatures”, BMC Bioinformatics, 17:313 (2016)
2. Rallis Karamichalis, Lila Kari, Stavros Konstantinidis and Steffen Kopecki, “An investigation of inter- and intragenomic variations of graphic genomic signatures”, BMC Bioinformatics, 16:246 (2015)
3. Lila Kari, Kathleen A. Hill, Abu S. Sayem, Rallis Karamichalis, Nathaniel Bryans, Katelyn Davis, Nikesh S. Dattani, “Mapping the space of genomic signatures”, PLoS One 10(5): e0119815 (2015)
4. Nicholas P.Karampetakis and Rallis Karamichalis, “Discretization of singular systems and error estimation”, Journal of Applied Mathematics and Computer Science (AMCS) Vol.24, No.1, pp.65-73, 2014
5. Nicholas P.Karampetakis and Rallis Karamichalis, “Discretization of Singular Systems and Error Estimation”, International Conference on Control, Decision and Information Technologies (CoDIT) 2013, IEEE Control Systems Society