Analysis and visualization of chromosome information

J.A. Tenreiro Machado, António C. Costa, Maria Dulce Quelhas

a b s t r a c t

This paper analyzes the DNA code of several species in the perspective of information content. For that pur- pose several concepts and mathematical tools are selected towards establishing a quantitative method with- out a priori distorting the alphabet represented by the sequence of DNA bases. The synergies of associating Gray code, histogram characterization and multidimensional scaling visualization lead to a collection of plots with a categorical representation of species and chromosomes.

Keywords:
DNA, Chromosome, Phylogenetics, Correlation, Visualization, Multidimensional scaling

1. Introduction

System dynamics studies the behavior of complex systems over time. During the last decades system dynamics evolved and presently addresses many different research topics. In fact, besides the classical areas of physics and engineering, we observe novel research directions such as economy and finance (Tenreiro Machado et al., 2011) or complex systems (Ignazio and Ammar, 2008), just to mention a few.

Phylogenetics is the study of the evolutionary relations between groups of organisms. Nowadays phylogenetics benefits from molecular sequencing data techniques to gather extensive data for analyses, aiming to improve research in areas like the evolutionary tree of life (Schuh and Brower, 2009; The Tree of Life Project) and organisms grouping, among many others. With the advent of genome sequencing and genome databases, considerable new information is available for computational processing, allowing worldwide research on decoding and understanding the informational structure present on DNA sequences. As such, this was the authors' main motivation to address this exciting and evolving area by applying sophisticated well known mathematical tools to genome data, hoping to identify new information structure and patterns.

The present paper studies the deoxyribonucleic acid (DNA) code in the perspective of system dynamics (Machado et al., 2011a,b,c). In fact, it is presently realized that the understanding of the DNA may be one of the most challenging problems posed to the human knowledge (The Official Web Site of the Nobel Prize). Decoding of this complex structure has not only a first level of biochemical detail, but also a second level of information (Harald, 2007). It is believed that, besides the information about the "structural construction" of a given species, DNA also includes the history of evolution towards the particular species and the "recipe" for the growth of each individual during its lifetime. These two different time scales, one of the backbones of Darwin's theory of evolution, reveal that we are in the presence of a complex system (may be the utmost one) with a complicated dynamics, and that the analysis tools developed in the scope of classical nonlinear systems may prove to be helpful in this context. This global observation motivated the association of several logical and mathematical concepts, namely, Gray code, histogram comparison and multidimensional visualization. Once established the analysis methodology, it is considered a collection of fifteen species and its corresponding DNA data. The results reveal important relationships between chromosomes and species, demonstrating the goodness of the proposed method, and motivating further research with the usual formalisms of system dynamics.

Having these ideas in mind, this paper is organized as follows. Section 2 briefly presents the main biological concepts and mathematical tools, and formulates their application in the framework of the DNA sequence decoding. Section 3 evaluates the correlation between chromosomes, investigates the data representation using multidimensional scaling, and compares several groups of species and chromosomes. Finally, Section 4 outlines the main conclusions.

2. Mathematical tools and DNA decoding

A gene is "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions" (Pearson, 2006). DNA is made up of two polymers connected by the bonding of hydrogen atoms, leading to a double helix structure. Each polymer contains nucleotides that can be classified into three types: deoxyribose (a five carbon sugar), a phosphate group, and a nitrogenous base. There are four different nitrogenous bases: thymine, cytosine, adenine, and guanine, often represented as T, C, A, and G. Each type of base on one strand forms a bond with just one type of base on the other strand. This arrangement is called "base pairing", with A bonding only to T, and C bonding only to G. The four bases are the foundation of the genetic code and act as the cell's memory, instructing it on how to synthesize enzymes and proteins. For example, in a human being, each cell holds twenty-three pairs of separate DNA-protein complexes (chromosomes), each containing, on average, 160 million nucleotide pairs. This massive amount of information is being collected and decoded during the last years, as the result of a large collaborative effort among many individuals and at research institutions around the world, and is available (UCSC Genome Bioinformatics) for scientific research. At the present moment of history of human development, it is known that we are merely starting to scratch the surface of a complex code and, in spite of the considerable efforts, a long way must be carried out to catch a glimpse of the logic beneath the mere chemical implementation of DNA.

Bearing these ideas in mind, this study analyses the DNA code of several species by applying dynamic and statistical mathematical tools. From the available DNA sequences a substantial part, corresponding to genes and short repetitive sequences (UCSC Genome Bioinformatics), organized into chromosomes, has been used in this study. For capturing the dynamics of the DNA code it is first observed that we are handing an alphabet, composed by the symbols {T, C, A, G}, and that any translation to a numerical counterpart may impose, since the initial conception, a bias towards some limited, or even wrong, direction. Consequently, it was decided to tackle directly the non-numerical code. Due to the vast volume of information, it was adopted a histogram-based measure. Nevertheless, in general, histograms do not capture dynamics and, in order to overcome this limitation, a scalable pattern detection algorithm based on counting the sequence of symbols was adopted. By "scalable" it is meant that the algorithm can count sequences of length n composed of the four base symbols. The available chromosome data includes a fifth symbol, represented by "N", which has no practical meaning for the DNA coding. Therefore, this symbol was discarded during the histogram bin construction.

We have different statistics when considering the length ranging from $n=1$, representing merely a static counting of $m=4^1$ states, up to $n=8$, representing the dynamics of a system with $m=4^8$ (65536) states. It must be noted that we are handling non-numerical quantities. Therefore, in order to prevent implicitly inserting a numerical order, it was decided to adopt bins in the histograms, according to the binary Gray encoding (Black, 2009) applied to the DNA four base alphabet. Since the standard Gray binary code changes only one digit between adjacent states, in our case we implement a one base change per state. For example, we get the consecutive bins {A} {C} {G} {T}, and {AA} {AC} {AG} {AT} {CT} {CG} {CC} {CA} {GA} {GC} {GG} {GT} {TT} {TG} {TC} {TA} for $n=1$ and $n=2$, respectively. Bins for larger values of n are not presented due to space limitations. During the bin counting two possible schemas may be considered, namely windows without any overlapping, and windows with a partial overlapping of the n base sequence. Therefore, we tested two extreme opposite cases, namely successive counting windows with zero and with $n-1$ adjacent bases in the DNA. In the first case, for a DNA strand of length L and Gray sequences of *length n,* results a total of L/n counting windows, while for the second it yields $L-n+1$ counting windows. Several tests revealed that both schemes lead to similar qualitative results, although some slight changes occurred with the smaller chromosomes. Therefore, to get a more robust counting, in the sequel is adopted the one-base sliding window, that is, the overlapping of $n-1$ consecutive bases.

Once the histograms of the sequence are obtained, the second step in our analysis consists in evaluating their similarities. There are several methods for such task (Chaa and Srihari, 2002; Haibin and Kazunori, 2006; Werman et al., 1985), but in the present case we notice that the histogram is the first numerical representation of the DNA code, or, by other words, that there is a relationship between two variables and, as such, it can be handled by a signal analysis tool. In this perspective it was considered the expression of the cosine correlation r_{ij} given by:

$$r_{ij}=\frac{\sum_{t=1}^{m}x_i(t)x_j(t)}{\sqrt{\sum_{t=1}^{m}x_1^2(t)\sum_{t=1}^{m}x_1^2(t)}},\ i,j=1,\ldots,p \qquad (1)$$

where x_i and x_j are two "signals" (histograms), m represents the number of bins and p denotes the total number of signals under comparison (in our case the total number of chromosomes). Eq. (1) is the normalized inner product and called the cosine coefficient because it measures the angle between two vectors and, thus, often denoted the angular metric (Deza and Deza, 2006; Sung-Hyuk, 2008).

The third stage of the analysis consists in revealing patterns embedded in the data. For this purpose we adopted the multidimensional scaling (MDS) technique (Borg and Groenen, 2005; Cox and Cox, 2001; Kruskal and Wish, 1978; Shepard, 1962; Tzeng et al., 2008). Briefly, MDS is a mathematical tool that represents, in a low dimensional map, a set of data points whose similarities are defined in a higher dimensional space, by means of a symmetric matrix, either of similarities $\mathsf{S}=[s_{ij}]$, or of distances $\mathsf{D}=[d_{ij}]$. In the case of similarities and classical MDS, the main diagonal of the matrix, S, is composed of ones, while the rest of the matrix elements must obey the restriction $0 \le s_{ij} \le 1$, $i,j=1,\ldots,p$. Alternatively, when comparing points in the viewpoint of distances, the main diagonal of matrix D is composed of zeros while the matrix elements obey the restriction $d_{ij} \ge 0$.

It should be noted that MDS works with relative measurements. Therefore, MDS maps are not sensitive to translations or rotations. The axes have only the meaning and units (if any) of the measuring index and packages usually apply a heuristic procedure to center the chart. In practical terms, this means that MDS maps are analyzed on the basis of proximity of (or, alternatively, of distance between) points, and comparison of the resulting cloud of points.

MDS is not an "exact" procedure, but instead tries to rearrange objects so as to arrive at a configuration that best approximates the observed similarities (or, alternatively, distances). A common measure for evaluating how accurately a particular configuration reproduces the D matrix is the raw stress measure defined by $\sigma=[d_{ij}-f(d_{ij})]^2$ where d_{ij} stands for the reproduced distances, given the respective number of dimensions, and d_{ij} represents the input data (i.e., the observed distances). The expression $f(d_{ij})$ indicates a non-metric, monotone transformation of the input data. Thus, the smaller the stress value σ, the better is the fit between the reproduced and the observed distance matrices. Plotting σ versus the number of dimensions leads usually to a monotonic decreasing plot and we can choose the "best dimension" as a compromise between stress reduction and number of dimension for the map representation. We can also analyze the goodness-of-fit by means of the Shepard diagram that, for a given number of dimensions, depicts the reproduced distances against the observed input data. Therefore, a narrow scatter around the 45 degree line indicates a good fit of the distances to the dissimilarities, while a large scatter indicates a lack of fit.

In the present case of DNA analysis, each element of matrix S is obtained with the cosine correlation r_{ij} (1) yielding a matrix of $p \times p$

similarities. The representation consists of three-dimensional plots and the consistency of the map is verified by means of the stress and Shepard charts.

In synthesis, for the DNA sequence analysis is adopted (*i*) the histogram for translating the thymine, cytosine, adenine, and guanine symbols without introducing a numerical bias, due to a Gray encoding of successive bins; (*ii*) the dynamical characterization of the code by means of *n*-tuple sequences; (*iii*) the similarity comparison of histograms using the cosine correlation; and (*iv*) the classical MDS for the emergence of patterns hidden in the high dimensional space.

Once defined the steps for attacking the problem and the corresponding mathematical tools, we had to consider the subjects to focus upon. In the natural world there is a huge number of species, but presently the number of decoded genomes is still limited to some hundred. Even so, the amount of combinations is considerable and, therefore, given the data sets available for processing, we decided:

- A first direction of study, limiting the set of species to six mammals, relatively close in phylogenetic terms, and aiming to explore the variation of dynamic analysis by changing the sequence length in the range $n=\{1, ..., 8\}$;
- A second direction of study, with fifteen species and more variation between them, using $n=6$ in the dynamic analysis.

In the first direction of study we consider six mammals, namely Human, Common Chimpanzee, Orangutan, Rhesus monkey, Pig and Opossum, denoted by the tags {Ho, Ch, Or, Rm, Po, Op}. Therefore, we get a total of $p = 122$ chromosomes while the number of histogram bins varies from $m=4^1$ up to $m=4^8$. The calculation cost of expression (1) increases significantly with n, both in the viewpoint of required memory for histogram construction and computational load. On the other hand, the visual structure of the MDS map evolves rapidly from the case of $n=1$, with unclear patterns, up to $n=8$, where pattern formation has clearly stabilized. To avoid defining further performance measures, the grouping of points in the MDS representation was not measured quantitatively. The qualitative visualization of the MDS charts demonstrated that $n = 6$ leads to a good visualization and, therefore, it is adopted as the default value in the second direction of study.

In the second case we have fifteen species, including the six mammals {Ho, Ch, Or, Rm, Po, Op}, two birds, Chicken and Zebra Finch {Ck, Tg}, two fishes, Zebrafish and Tetraodon {Zf, Tn}, two insects, Gambiae mosquito and Honeybee {Ag, Am}, two nematodes, *Caenorhabditis elegans* and *Caenorhabditis briggsae* {Ce, Cb}, and one fungus, Yeast {Sc}. This set of beings makes a grand total of $p=281$ chromosomes that, as mentioned earlier, is compared for a DNA sequence length of $n = 6$. The chromosomes characteristics of each DNA species are presented in Table 1.

Table 1
Species and their chromosomes.

Specie	Tag	Group	Number of chromosomes
Human	Ho	Mammal	24
Chimpanzee	Ch	Mammal	25
Orangutan	Or	Mammal	24
Rhesus	Rm	Mammal	21
Pig	Po	Mammal	19
Opossum	Op	Mammal	9
Chicken	Ga	Bird	30
Zebra Finch	Tg	Bird	31
Zebrafish	Zf	Fish	25
Tetraodon	Tn	Fish	21
Mosquito (*Anopheles gambiae*)	Ag	Insect	6
Honeybee (*Apis mellifera*)	Am	Insect	16
Caenorhabditis elegans	Ce	Nematode	6
Caenorhabditis briggsae	Cb	Nematode	6
Yeast (*Saccharomyces cerevisiae*)	Sc	Fungus	16

(Note: chromosomes Ck32 and Tg16 were ignored due to their very small base pair count)

3. Multidimensional analysis of DNA

This section explores the visualization of that information by means of MDS of relationships between the chromosomes included in the two research directions (*i*) the set of six species {Ho, Ch, Or, Rm, Po, Op} with sequence DNA lengths $n=\{1, ..., 8\}$ and (*ii*) the set of fifteen species {Ho, Ch, Or, Rm, Po, Op, Ck, Tg, Zf, Tn, Ag, Am, Ce, Cb, Sc} with $n=6$.

There are several packages available, either proprietary, or open source (MathWorks; The R Project for Statistical Computing) for the calculation of MDS. After several tests, we adopted the GGobi package due to its simplicity, speed and robustness (GGobi - Interactive and dynamic graphics). In all experiments the MDS maps were tested using the stress chart and the Shepard diagram for evaluating the dimension. Small variations occurred between experiments, leading to a minimum number of dimension varying between two and three. Nevertheless, since the default GGobi visualization adopts three dimensions, for the sake of simplifying comparisons are adopted those maps in all cases.

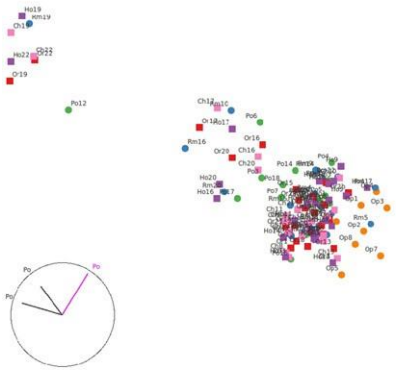
The comparison of group of six mammals {Ho, Ch, Or, Rm, Po, Op} leads to a matrix S with 122×122 elements. Fig. 1 depicts the MDS charts for a three-dimensional representation, and statistics with Gray code and sequence lengths from $n=1$ up to $n=8$ (i.e., number of histogram bins from $m=4^1$ up to $m=4^8$), respectively. There are infinitely many projection views that can be done to depict a three-dimensional representation. The authors have chosen *one* projection that, somehow, preserves the main visualization properties.

It is easily noticeable that for $n=1$ we get a kind of "cloud" where it is difficult to devise any structure, but when n increases a clear pattern emerges, the effect being stabilized around $n=6$. For the highest vales of n we observe two types of objects in the MDS representations, namely radial vectors (RV) and sets of close points (SP). The RV object seems to be associated with "complex beings". For each of those the chromosomes are somehow different between themselves, therefore originating distinct points in a "direction" that represents that particular species. The SP object seems to be associated with "less complex beings" and is formed by the corresponding group of chromosomes, but are not enough different to define a "direction". For both objects the classification is relative and the results depend on the types of species represented in the MDS plot. In terms of dynamical analysis, the RV shows not only that the Gray encoding fond significant variations of the base alignment in each chromosome (the importance of such phenomenon requires space in the MDS), but also that there is a common logic. Otherwise we would get more "fuzzy" associations.

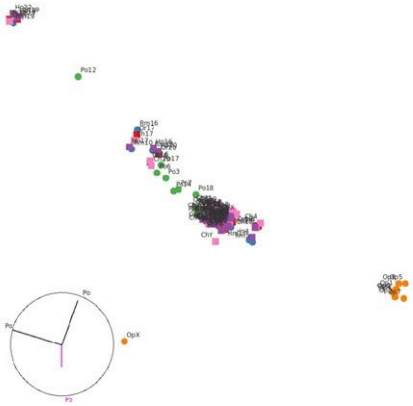
In global terms we verify the close relationship in the RVs object representing {Ho, Ch, Or}, Ch being slightly closer to Ho than Or. The smaller RV for Po and the SP for Op are further away from Ho, being Op the most far away. These results are close to what is known from phylogenetics (Dunn et al., 2008; Ebersberger et al., 2007; Hillier et al., 2004Murphy et al., 2007; Prasad and Allard, 2008; Sims et al., 2009; Zhao and Bourque, 2009). Although nothing absolutely new, these results show that the produced MDS charts agree with current scientific knowledge and support our analysis.

In Fig. 2a) a more detailed observation reveals the emergence of clusters with chromosome points. Several clusters of chromosomes are visible with similar identifiers for the species {Ho, Ch, Or}, as well as the RVs for the {Ho, Ch, Or} group and {Rm}. It is worth mentioning that some Rm chromosomes shown up inside the {Ho, Ch, Or} RV. It is also interesting to note that {Ho, Ch, Or} X chromosomes are slightly apart from the rest. The HoY and ChY chromosomes (there is no OrY chromosome) are the farthest away, both of them being very much different in size and content from the other chromosomes.

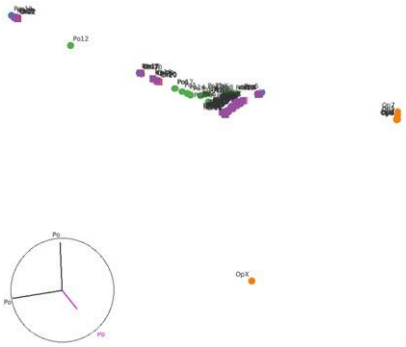
n = 1



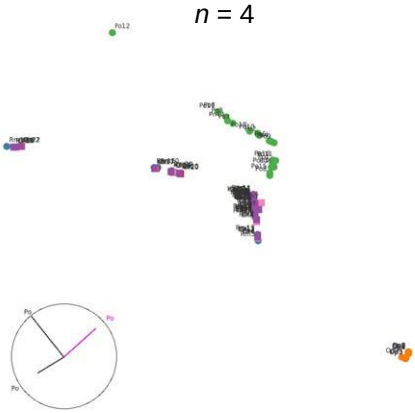
n = 2



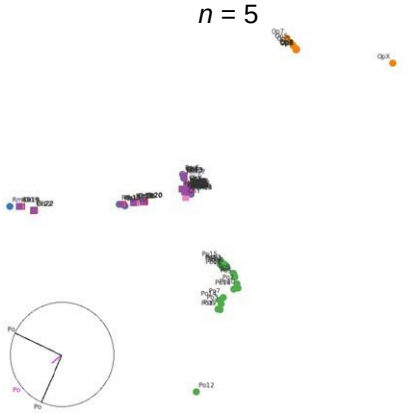
n = 3



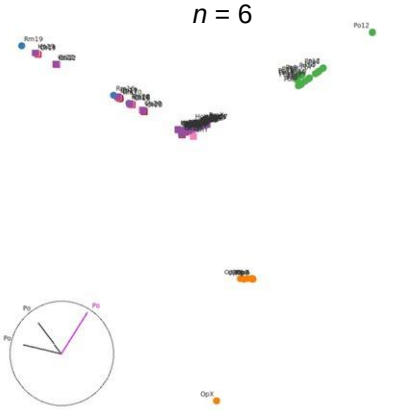
n = 4



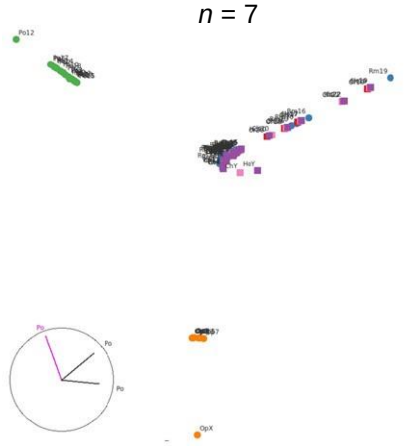
n = 5



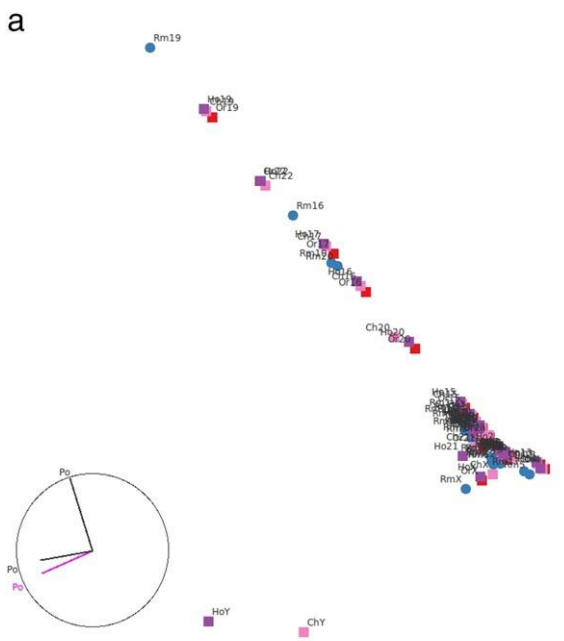
n = 6



n = 7



n = 8



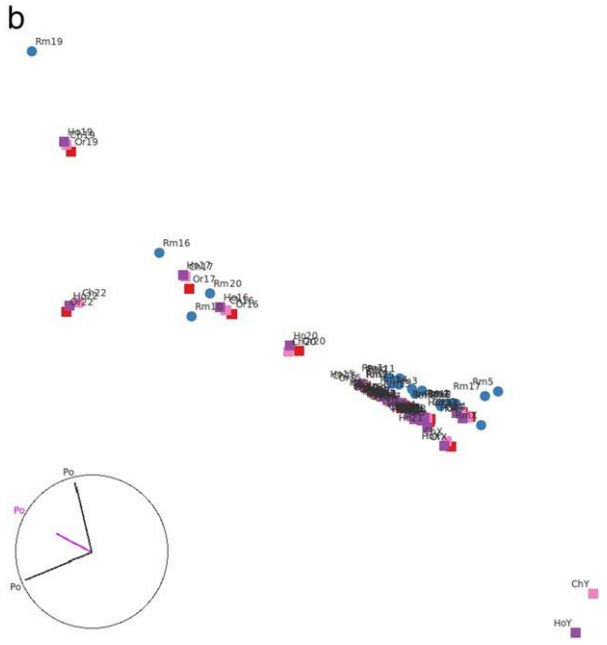


Fig. 2. Three dimensional MDS plot for DNA sequence length $n=8$ for the species {Ho, Ch, Or, Rm}: a) zoom for of the MDS plot with {Ho, Ch, Or, Rm }={violet □, pink □, red □, blue ○}, $p=122$; b) MDS plot with {Ho, Ch, Or, Rm, Po, Op}={violet □, pink □, red □, blue ○}, $p=94$.

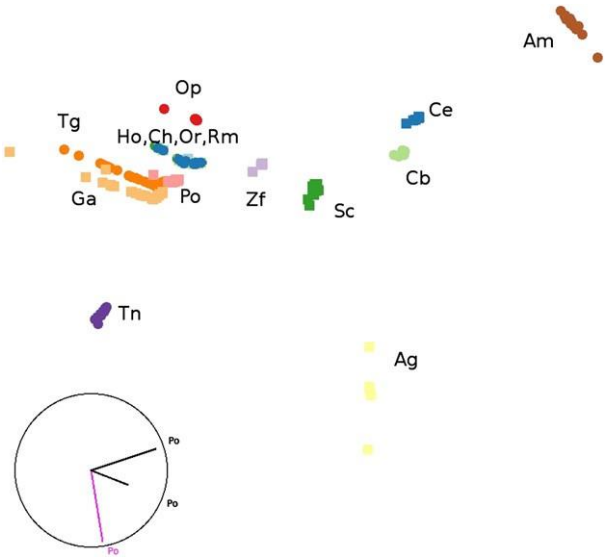


Fig. 3. Three dimensional MDS plot for the fifteen species when and $n=6$. {Ho, Ch, Or, Rm, Po, Op, Ga, Tg, Zf, Tn, Ag, Am, Ce, Cb, Sc}={light blue □, blue ○, light green □, green ○, pink □, red ○, light brown □, orange ○, gray □, violet ○, light yellow □, brown ○, blue □, light green ○, green □}, $p=281$.

It should be noted that the zoom of a MDS plot with $p=122$ (Fig. 2a) is not identical to the MDS plot of only the four species {Ho, Ch, Or, Rm} and $p=94$ (Fig. 2b), but globally the conclusions are identical.

We now consider the group of fifteen species {Ho, Ch, Or, Rm, Po, Op, Ck, Tg, Zf, Tn, Ag, Am, Ce, Cb, Sc} and DNA sequence length $n=6$.

Fig. 3 depicts the three dimensional MDS chart for histograms with $n=6$ and Fig. 4 shows a zoom in one of the most dense areas. We verify that the six mammals are very close in the map because their relative differences are very small, when compared with the non-mammal species represented in the MDS plot.

In a global perspective Fig. 3 reveals that, as expected, the mammals' species are close together. The points corresponding to {Ho, Ch, Or, Rm} superimpose over the same RV, while {Po, Op} originate two slightly distinct sets. The {Ck, Tg} originate two distinct RVs, somewhat close but not superimposing. The {Zf, Tn}, {Ag, Am}, {Ce, Cb} define small distinct SPs. It is interesting to note that {Sc} is somehow in between RV and SP, because it is a long chain of points but not so well defined as a RV. Moreover, in relative terms, {Sc} seems to be in-between mammals/aves and nematodes.

Fig. 4 reveals that, as expected, the primate species are close together. The points corresponding to {Ho, Ch, Or, Rm} superimpose over the same RV, while {Ck, Tg} originate two slightly distinct, but close and parallel sets. It is also worth mention that in {Ck, Tg} the chromosome numbering is very similar from top to bottom in the image.

In conclusion, MDS plots resulting from correlation histogram comparison show to be helpful in identifying relevant patterns, potentially leading to new and important observations.

4. Conclusions

In this paper it was verified that chromosomes have a code based on a for symbol alphabet. This information can be analyzed with tools usually adopted in the study of complex systems. Nevertheless, a quantitative analysis must avoid introducing assumptions that may a priori distort all subsequent numerical processing. The proposed methodology, by embedding a Gray-like encoding into a histogram, avoids quantification assumptions and provides data in a numerical format suitable for further processing with mathematical tools. A cosine correlation for comparing histograms was adopted, as well as a multidimensional scaling procedure for visualizing and understanding

Fig. 1. Three dimensional MDS plots for the set of six species and DNA sequence lengths $n=\{1, 2, ..., 8\}$. {Ho, Ch, Or, Rm, Po, Op}={violet □, pink □, red □, blue ○, green ○, orange ○}, $p=122$.

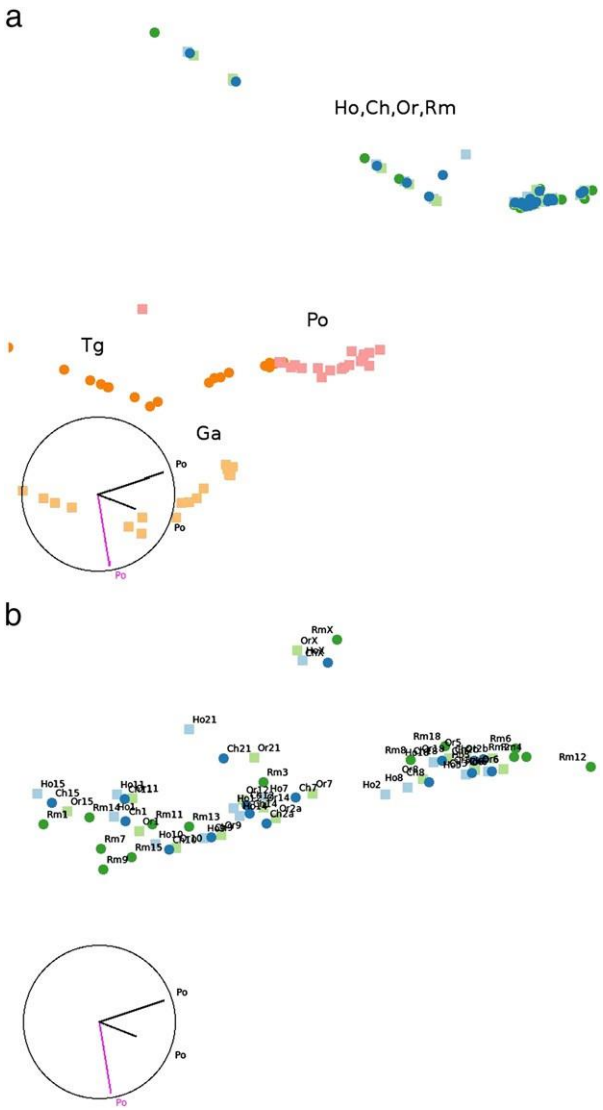


Fig. 4. Two zooms of the MDS plot for $n=6$. {Ho, Ch, Or, Rm, Po, Op, Ga, Tg, Zf, Tn, Ag, Am, Ce, Cb, Sc} = {light blue □, blue ○, light green □, green ○, pink □, red ○, light brown □, orange ○, gray □, violet ○, light yellow □, brown ○, blue □, light green ○, green □}, $p = 281$.

results. These tools revealed important relationships, but we believe that the weak restrictions assumed in the quantifying study may have reduced the code dynamics. Even so, the main merit of the overall processing described is the identification of hidden patterns that may open new research directions to pursuit.

Acknowledgments

We thank the following organizations for allowing access to genome data:

- Human – Genome Reference Consortium, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
- Common Chimpanzee – Chimpanzee Genome Sequencing Consortium
- Orangutan – Genome Sequencing Center at WUSTL, http://genome.wustl.edu/genome.cgi GENOME=Pongo%20abelii
- Rhesus – Macaque Genome Sequencing Consortium, http://www.hgsc.bcm.tmc.edu/projects/rmacaque/
- Pig – The Swine Genome Sequencing Consortium, http://piggenome.org/
- Opossum – The Broad Institute, http://www.broad.mit.edu/mammals/opossum/
- Chicken – International Chicken Genome Sequencing Consortium Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004 Dec 9; 432(7018): 695-716. PMID: 15592404
- Zebra Finch – Genome Sequencing Center at Washington University St. Louis School of Medicine
- Zebrafish – The Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/Projects/D_rerio/
- Tetraodon – Genoscope, http://www.genoscope.cns.fr/
- Honeybee – The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/honeybee/
- Gambiae Mosquito – The International Anopheles Genome Project
- Elegans nematode – Wormbase, http://www.wormbase.org/
- Briggsae nematode – Genome Sequencing Center at Washington University in St. Louis School of Medicine
- Yeast – Sacchromyces Genome Database, http://www.yeastgenome.org/

References

Black, Paul E., 2009. Gray code. In: Black, Paul E. (Ed.), Dictionary of Algorithms and Data Structures [online], 31. U.S. National Institute of Standards and Technology. August.

Borg, I., Groenen, P., 2005. Modern Multidimensional Scaling-Theory and Applications, 2nd edition. Springer-Verlag, New York.

Chaa, Sung-Hyuk, Srihari, Sargur N., 2002. On measuring the distance between histograms. Pattern Recognit. 35, 1355-1370.

Cox, T., Cox, M., 2001. Multidimensional scaling, 2nd edition. Chapman & Hall/CRC.

Deza, E., Deza, M.M., 2006. Dictionary of Distances. Elsevier.

Dunn, Casey W., et al., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745-750.

Ebersberger, Ingo, Galgoczy, Petra, Taudien, Stefan, Taenzer, Simone, Platzer, Matthias, von Haeseler, Arndt, 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24 (10), 2266-2276.

GGobi - Interactive and dynamic graphics http://www.ggobi.org/.

Haibin, Ling, Kazunori, Okada, 2006. Diffusion Distance for Histogram Comparison. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Analytics of Protein-DNA Interactions. In: Harald, Seitz (Ed.), Advances in Biochemical Engineering Biotechnology. Springer.

Hillier, LaDeana W., et al., 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, International Chicken Genome Sequencing Consortium. Nature 432, 695-716.

Ignazio, Licata, Ammar, Sakaji (Eds.), 2008. Physics of Emergence and Organization. World Scientific.

Kruskal, J., Wish, M., 1978. Multidimensional Scaling. Sage Publications, Inc, Newbury Park, CA.

Machado, J.A. Tenreiro, Costa, António C., Quelhas, Maria Dulce, 2011a. Fractional dynamics in DNA. Commun. Nonlinear Sci. Numer. Simul. 16 (8), 2963-2969 August.

Machado, J.A. Tenreiro, Costa, António C., Quelhas, Maria Dulce, 2011b. Entropy analysis of DNA code dynamics in human chromosomes. Comput. Math. Appl. 62 (3), 1612-1617 Aug.

Machado, J.A. Tenreiro, Costa, António C., Quelhas, Maria Dulce, 2011c. Shannon, Rényie and Tsallis entropy analysis of DNA using phase plane. Nonlinear Anal. Ser. B Real World Appl. 12 (6), 3135-3144 December.

MathWorks http://www.mathworks.com/.

Murphy, William J., Pringle, Thomas H., Crider, Tess A., Springer, Mark S., Miller, Webb, 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res. 17, 413-421.

Pearson, H., 2006. Genetics: what is a gene? Nature 441 (7092), 398-401.

Prasad, Arjun B., Allard, Marc W., 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol. Biol. Evol. 25 (9), 1795-1808.

Schuh, R.T., Brower, A.V.Z., 2009. Biological Systematics: principles and applications, 2nd edition. Cornell University Press.

Shepard, R., 1962. The analysis of proximities: multidimensional scaling with an unknown distance function, I and II. Psychometrika 27, 219-246 and 219-246.

Sims, Gregory E., Jun, Se-Ran, Wu, Guohong A., Kim, Sung-Hou, 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Nat. Acad. Sci. U. S. A. 106 (8), 2677-2682.

Sung-Hyuk, Cha, 2008. Taxonomy of Nominal Type Histogram Distance Measures. American Conference on Applied Mathematics (MATH '08), Harvard, Massachusetts, USA, March 24-26.

Tenreiro Machado, J.A., Duarte, Gonçalo M., Duarte, Fernando B., 2011. Identifying economic periods and crisis with the multidimensional scaling. Nonlinear Dyn. 63 (4), 611–622 Springer.
The Official Web Site of the Nobel Prizehttp://nobelprize.org/nobel_prizes/medicine/laureates/1968/.
The R Project for Statistical Computinghttp://www.r-project.org/.
The Tree of Life Projecthttp://tolweb.org/tree/home.pages/abouttol.html.
Tzeng, J., Horng-Shing Lu, H., Li, Wen-Hsiung, 2008. Multidimensional scaling for large genomic data sets. BMC Bioinforma. 9, 179.
UCSC Genome Bioinformaticshttp://hgdownload.cse.ucsc.edu/downloads.html.
Werman, M., Peleg, S., Rosenfeld, A., 1985. A distance metric for multidimensional histograms. Comput. Vision Graphics Image Proc. 32, 328–336.
Zhao, Hao, Bourque, Guillaume, 2009. Recovering genome rearrangements in the mammalian phylogeny. Genome Res. 19, 934–942.