Histogram-based DNA analysis for the visualization of chromosome, genome and species information

António M. Costa, José T. Machado and Maria D. Quelhas

ABSTRACT

Motivation: We describe a novel approach to explore DNA nucleotide sequence data, aiming to produce high-level categorical and structural information about the underlying chromosomes, genomes and species.

Results: The article starts by analyzing chromosomal data through histograms using fixed length DNA sequences. After creating the DNA-related histograms, a correlation between pairs of histograms is computed, producing a global correlation matrix. These data are then used as input to several data processing methods for information extraction and tabular/graphical output generation. A set of 18 species is processed and the extensive results reveal that the proposed method is able to generate significant and diversified outputs, in good accordance with current scientific knowledge in domains such as genomics and phylogenetics.

Availability and implementation: Source code freely available for download at http://www4.dei.isep.ipp.pt/etc/dnapaper2010, implemented in Free Pascal and UNIX scripting tools. Study input data available online for download at University of California at Santa Cruz Genome Bioinformatics, http://hgdownload.cse.ucsc.edu/downloads.html.

Contact: acc@isep.ipp.pt

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 21, 2010; revised on March 1, 2011; accepted on March 7, 2011

1 INTRODUCTION

Phylogenetics concerns 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 such as the evolutionary tree of life (Maddison *et al*., 2007; Schuh and Brower, 2009), grouping of organisms, among many others. With the advent of genome sequencing and genome databases, a large volume of information is available for computational processing, allowing worldwide research on decoding the informational structure present in DNA sequences.

A massive amount of DNA information is being collected and decoded, as result of a large collaborative effort among many individuals and research institutions around the world, and is available for scientific research. In Machado (2010), this evolving area was addressed by applying mathematical tools to genome data, revealing new information patterns.

In this study, we analyze the DNA code in the perspective of identifying structural patterns in the nuclear and mitochondrial genomes. Understanding DNA may be one of the most challenging problems posed to the human knowledge (Nobel Prize Web site, http://nobelprize.org/nobel_prizes/medicine/laureates/1968/). The decoding of the DNA complex structure may not only have a primary level of biochemical detail, but also other levels of information (Seitz, 2007). This vision motivated the association of logical and mathematical concepts, namely, histogram, correlation and analysis/visualization tools such as multidimensional analysis, directed graphs and dendograms. Once established the methodology for several species, its DNA data is used to pursue the vision. In the chosen DNA repository a substantial part, corresponding to genes and short repetitive sequences (as defined in the University of California Santa Cruz Genome Bioinformatics web site), is organized into chromosomes, which is our input data. In this study, we consider the available nuclear and mitochondrial genomes of 18 species: 10 mammals, 2 birds (aves), 2 fishes, 1 insect, 2 nematodes and 1 fungus. We note that in most of the species the association of genes to chromosomes is not yet incomplete. In Table 1, we present the chromosomal characteristics of those species.

The DNA implements an alphabet composed by the symbols {T, C, A, G}. Any simple translation to a numerical counterpart may impose bias and destroy intrinsic information. Consequently, it was decided to directly process the non-numerical code. Due to the immense volume of information, a histogram-based measure was adopted. Nevertheless, in general, histograms do not capture dynamics. In order to overcome this limitation, a flexible pattern detection algorithm based on counting the sequence of symbols was considered (Vinga and Almeida, 2003). By 'flexible' we mean that the algorithm can count sequences of length n items, each one composed by one of the four base symbols.

With the exception of Yeast (Sc), the available chromosome data includes a fifth symbol ('N'), corresponding to masked DNA symbols not belonging to the genome, which typically appear in large contiguous sequences. For example, in the human Y chromosome file there are 59 373 566 bp, of which 33 710 000 bp are 'N' (56.78%) arranged in 17 sequences, the largest one with 30 000 000 symbols. Another example is the Chicken Ga25 chromosome, with 2 051 775 bp, of which 663 879 are 'N' (32.67%)

Table 1. Characteristics of 18 species and used chromosomes

Species	Group	Nuclear/mitochondrial chromosomes
Human (Ho)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y M
Chimpanzee (Ch)	Mammal	1 2a 2b 3 4 5 6 7 8 9 10 11 12 1 14 15 16 17 18 19 20 21 22 X Y M
Orangutan (Or)	Mammal	1 2a 2b 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X M
Pig (Po)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 X M
Opossum (Op)	Mammal	1 2 3 4 5 6 7 8 X M
Horse (Eq)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 X M
Dog (Dg)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 X M
Ox (Ox)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 X M
Mouse (Mm)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y M
Rat (Rn)	Mammal	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 X M
Chicken (Ga)	Ave	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 W Z M
Zebra Finch (Tg)	Ave	1a 1b 1 2 3 4 4a 5 6 7 8 9 10 11 12 13 14 15 17 18 19 20 21 22 23 24 25 26 27 28 Z M
Zebrafish (Zf)	Fish	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 M
Tetraodon (Tn)	Fish	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 M
Mosquito (Ag)	Insect	2l 2r 3l 3r X M
Caenorhabditis elegans (Ce)	Worm	1 2 3 4 5 X M
Caenorhabditis briggsae (Cb)	Worm	1 2 3 4 5 X M
Yeast (Sc)	Fungus	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 M

Chromosomes Ga32 and Tg16 were ignored due to their very small base pair count.

arranged in 274 sequences, the largest one with 500 000 symbols. HoY and Ga25 are just two examples of chromosomes with a percentage of 'N' symbols $>10\%$, but most of the chromosomes have smaller percentages.

We decided not to use 'N' in sequences as a fifth symbol or not to replace it by any of symbols {T, C, G, A}, because that would introduce an unknown bias in the sequence processing. We then considered two approaches: (a) removing all 'N' symbols in a preprocessing step or (b) process sequences but ignoring any sequence with an 'N'. Although (a) and (b) may seem different, we concluded that differences were minimal and that (a) could be advantageously used without compromising the quality of results and conclusions.

Table 2. Differences in two approaches for ignoring 'N' symbols

Chromosome	Sequences with 'N' removed [α]	Sequences with 'N' filtered [β]	$(\alpha-\beta)/\beta$ (%)
Ga25	1367889	1366030	0.136088
Ga3	110204947	110177075	0.025297
Tn1	20304845	20315377	0.051869
Tn15	6235253	6236842	0.025484
AgX	21470369	21477782	0.034527
Ag2l	48065434	48071405	0.012423
HoY	25653559	25653447	0.000437
Ho5	177695253	177695218	0.000020

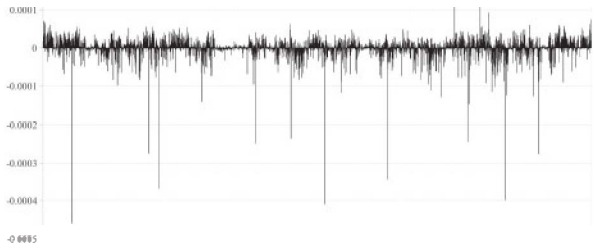

Fig. 1. Difference between relative frequencies of human X and Y chromosome's histograms for $n=6$ (4096 bins).

Using as examples {Ho, Ck, Tn, Ag} nuclear chromosomes and a sequence length of $n=8$, Table 2 rightmost column synthesizes the differences for the (a) and (b) approaches. For Ga25, the Pearson's correlation coefficient r between (a) and (b) sequences with length $n=8$ yields $r>0.9999717$, while for HoY the corresponding coefficient r is >0.9999999. We conclude that both approaches are statistically equivalent for the envisaged DNA decoding. Therefore, we opted to discard the 'N' symbol before histogram construction.

Different statistics may be produced 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 $m=4^8$ (65 536) states. For bin counting a one base sliding window (i.e. shift of one base and overlap of $n-1$ consecutive bases) method was adopted.

Figure 1 shows the differences between the relative frequencies of human X and human Y chromosome's histograms for $n=6$. The large number of bins (4096) visually helps understanding the differences in the DNA base sequences of both chromosomes. These differences are our main motivation to study the genomic DNA and find out if some high-level structural information will emerge.

In short, for our DNA sequence analysis process we adopted (i) the histogram for translating the T, C, A, G symbols into numerical values; (ii) the dynamical code characterization by means of n-tuple sequences; (iii) the sequence similarity comparison using a correlation method; and (iv) the identification of hidden patterns in the numerical sequence and subsequent high-level visualization of those patterns.

In this study, we demonstrate that a four-phase methodology for DNA sequence analysis is able to reveal unexpected structural patterns between nuclear chromosomes, either intra- or interspecies. The results also reveal important high-level relationships between chromosomes/species, showing the goodness of the proposed method and motivating further research involving more complete and extensive DNA data and other complementary scientific tools.

2 METHODS

After downloading the DNA data (in FASTA format) from the University of California Santa Cruz Genome Bioinformatics web site, each DNA chromosomal sequence was processed in order to remove the 'N' symbols and to convert all base symbols to the {A, C, G, T} alphabet.

For the bin counting two possible approaches were considered, namely windows without any overlapping, and windows with a partial overlapping of the n base sequence. Several tests revealed that both approaches tend to generate similar results, although some slight differences show up when processing smaller chromosomes. Therefore, to get a more robust counting, the one base sliding window (i.e. shift of one base and overlap of $n-1$ consecutive bases) was adopted. For that purpose we developed the '*genhists*' application, available in the Supplementary Material, which requires as parameters the chosen sequence length (n) and filtered DNA sequence files, generating the corresponding histogram files.

After obtaining the histograms for a given value of n and a set of chromosomes, the second step in our analysis is to evaluate their similarities. There are many methods for such task (Chaa and Srihari 2002; Ling and Okada 2006; Werman *et al*. 1985). In the end, we obtain a correlation matrix $\mathbf{S}=[s_{ij}]$, where s_{ij} is defined as $s_{ij}=f(H_i,\ H_j)\wedge i,\ j=1,\ldots,n$ ($s_{ij}=s_{ji}\wedge i$, $j=1,\ldots,n$; $s_{ii}=1\wedge i=1,\ldots,n$) in which H_i and H_j are two histograms of length m, the function $f(H_i,H_j)$ is real valued and $0\leq s_{ij}\leq 1$. Being interested in qualitative similarities, we opted for a method that measures the portion of ranks that match between any two histograms. As such, we adopted the statistical 'Kendall τ' rank correlation method (Kendall, 1938), which computes the correspondence between rankings of two histograms and assesses its significance based on the number of 'concordant pairs', 'discordant pairs' and 'ties' over all $\{H_i(a),H_j(a)\}$ and $\{H_i(b),H_j(b)\}$ pairs, where $a,\ b=1,\ldots,m$ and $a<b$. We note that the Kendall τ correlation method is computationally expensive. Therefore, we adopted the efficient algorithm described in Christensen (2005). To generate a correlation similarity matrix file, we developed the '*gentauk*' application, also available in the Supplementary Material, which requires as parameters the chosen sequence length (n) and the histogram files, generating the corresponding correlation matrix file.

The third step in the analysis consists in revealing embedded patterns in the correlation matrix data. For this purpose, we start by considering the multidimensional scaling (MDS) technique (Borg and Groenen, 2005; Cox and Cox, 2001; Kruskal and Wish, 1978; Shepard, 1962; Tzeng *et al*., 2008). The MDS is a mathematical tool that represents, in a lower dimensional map, a set of data points whose similarities (or alternatively distances) are defined in a higher dimensional space by means of a symmetric matrix $\mathbf{S}=[s_{ij}]$. In the case of similarities and classical MDS, the matrix main diagonal is composed of ones, while the rest of the matrix elements must obey the restriction $0\leq s_{ij}\leq 1 (s_{ij}\geq 0)$, $i,j=1,\ldots,m$. Usually, in order to facilitate the graphical representation, 2D and 3D MDS plots are used and its consistency verified by means of Shepard and/or stress charts. To create the MDS plots, we opted for the GGobi package, chosen due to its simplicity, speed and robustness (http://www.ggobi.org).

Other than for MDS plots, a correlation matrix can be used to produce graphs linking the most correlated items, in order to visualize the underlying patterns between them. As such, we chose the GraphViz package (http://www.graphviz.org), an open source software for representing structural information as diagrams of abstract graphs and networks, to create directed graphs that show how chromosomes or species are related. A correlation matrix can also be used to generate a dendogram, a tree-like diagram depicting clusters resulting from some hierarchical clustering method. To generate the dendograms in this study, we selected the MultiDendograms hierarchical clustering package, configured for the 'Joint Between Within' clustering method (Fernández and Gómez, 2008).

The Supplementary Material contains the source code and the executable files of our custom-developed applications, as well as input data files (sequence histograms), output data files (for use with the GGobi, GraphViz or MultiDendograms packages, images and videos) and some utility applications (mostly conversion scripts to be used in UNIX or GNU/Linux platforms).

3 RESULTS

3.1 Correlation analysis of the nuclear chromosomes

After generating the sets of 384 histograms for the nuclear chromosomes of 18 species using sequences of length $n=\{1,\ldots,8\}$, for each value of n, we applied the Kendall τ correlation method (Kendall, 1938) to generate the corresponding 384×384 $\mathbf{S}_{\text{nuclear}}$ similarity matrix. As previously mentioned, to create the MDS plots we use the GGobi package.

Figure 2 shows 3D MDS plots of the 384 nuclear chromosomes of 18 species for sequence lengths $n=\{3,6\}$. Even with a length as low as $n=3$, 3D patterns are easily noticeable. The same patterns can also be observed in the $n=6$ plot, but better defined and separated.

Figure 3 shows the result of performing a 3D MDS on the 384 nuclear chromosomes of 18 species for sequence length $n=8$, with two distinct bi-dimensional projections. The analysis of images in Figures 2 and 3 shows the emergence of spatial patterns strongly related with the chromosome grouping into species. Although MDS plots were created for $n=\{1,\ldots,8\}$, we note that larger values of n (7 or 8) generate MDS plots with better chromosome groupings and more clearly separated species. We also observe that the quality of chromosome grouping and species' separation improves as n increases, stabilizing around $n=8$.

The similarity matrix $\mathbf{S}_{\text{nuclear}}$ can also be used to produce graphs linking the most correlated chromosomes, in order to visualize the underlying structural patterns between them. The directed graph of Figure 4, generated by GraphViz, shows how the chromosomes of {Ho, Ch, Or} are correlated for $n=8$. A chromosome pointing to another chromosome by means of a continuous line (labeled '1') ending in an arrow represents the chromosome that is most correlated to the 'pointed' chromosome. If the line is a dashed one (labeled '2') then the pointing chromosome is the second most correlated. In Figure 4, only links with correlation $\geq 95\%$ are visible and it shows that {Ho, Ch, Or} chromosomes with the same 'number' are more

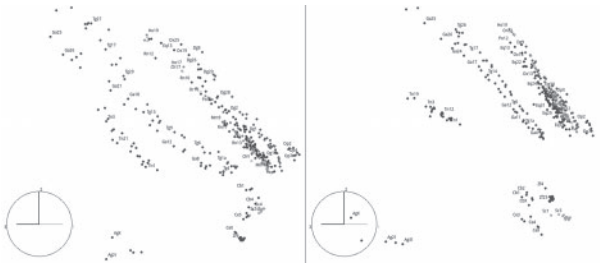

Fig. 2. 3D MDS plots of the 384 nuclear chromosomes for $n=3$ (left side) and $n=6$ (right side).

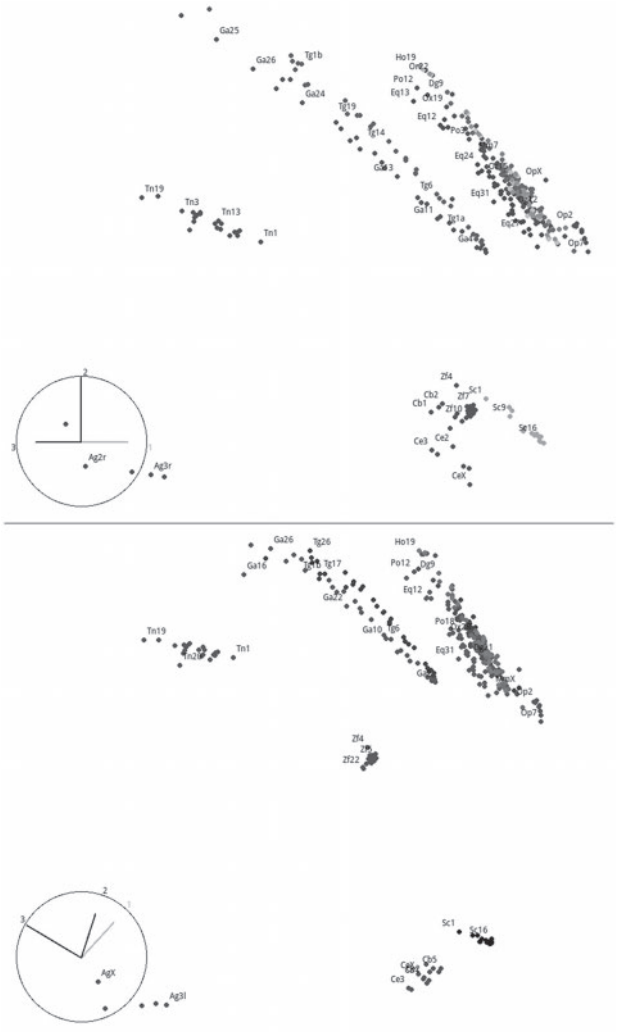

Fig. 3. Two distinct views of a 3D MDS plot of the 384 nuclear chromosomes for $n=8$. Shaded dots represent nuclear chromosomes, some labeled for readability.

correlated to each other, although there are some exceptions: for example, the Y chromosomes and the 2a and 2b chromosomes in {Ch, Or}.

The $\mathbf{S}_{\text{nuclear}}$ similarity matrix can also be used to generate a dendogram, a tree-like diagram depicting clusters resulting from some hierarchical clustering method. The dendogram of Figure 5 was created by the MultiDendograms hierarchical clustering package (Fernández and Gómez, 2008), using $n=8$ and the {Ho, Ch, Or} chromosomes. In Figure 5, we can observe several levels of {Ho, Ch, Or} clusterings (e.g. the clusters of chromosomes 10, 11, 1, 9, 15 in the left corner of Fig. 5).

3.2 Correlation analysis of the nuclear genome

Up to this point we concentrated on the nuclear chromosomes separately. Another approach is to consider the nuclear genome as a whole. This can be done by combining, for each species, all their nuclear chromosome histograms, thus originating the corresponding

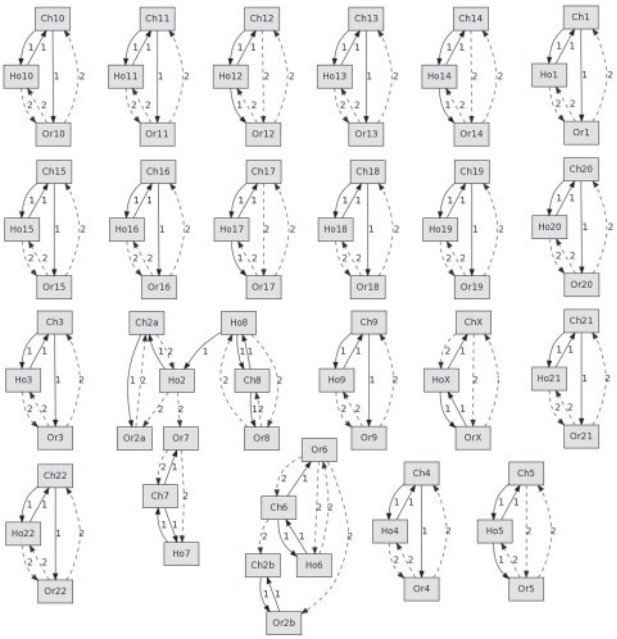

Fig. 4. Graph of the two most correlated chromosomes for the chromosomes of {Ho, Ch, Or} for $n=8$. Gray rectangle: chromosome, link r: connection of similarity r between two chromosomes.

'global nuclear histogram'. Subsequently, by applying the Kendall τ correlation method to 18 species' global nuclear histograms, a species similarity matrix $\mathbf{S}_{\text{global}}$ is produced, and then processed to generate some high-level visualizations.

Figure 6 presents the 3D MDS plot of 18 nuclear genomes for sequence length $n=8$ and shows a clear spatial organization of patterns involving species, particularly for the mammals, which are strongly clustered with the exception of the Opossum. We also note that the aves {Ga, Tg} are near each other, as well as the fishes {Zf, Tn}, and that the non-vertebrate species are far apart from the vertebrates.

Figure 7 depicts how 15 of the 18 species are most correlated to each other at primary level (continuous line, labeled '1') and at secondary level (dashed line, labeled '2'). Only links with correlation $\geq 75\%$ are visible. It also shows five clusters of species in the graph: {Ho, Ch, Or}, {Cb, Ce}, {Mn, Rn}, {Ga, Tg} and {Eq, Dg, Op, Po, Ox}.

3.3 Correlation analysis of the mitochondrial genome

Mitochondrial genomes have DNA sequence counts between 13 000 and 86 000 nt, while nuclear genomes have DNA sequences counts between 12 000 000 and 3 500 000 000 nt. This means that the mitochondrial data are very much smaller than the nuclear one.

After generating the sets of 18 histograms for the mitochondrial chromosomes of 18 species using sequences of variable length n, for each value of n we applied the Kendall τ correlation method to generate the corresponding 18×18 similarity matrix $\mathbf{S}_{\text{mito}}$.

Figure 8 reveals the result of performing a 3D MDS on the mitochondrial chromosomes of 18 species for $n=8$. It also shows a clear spatial organization of patterns involving mitochondrial chromosomes, particularly mammals, primates and aves.

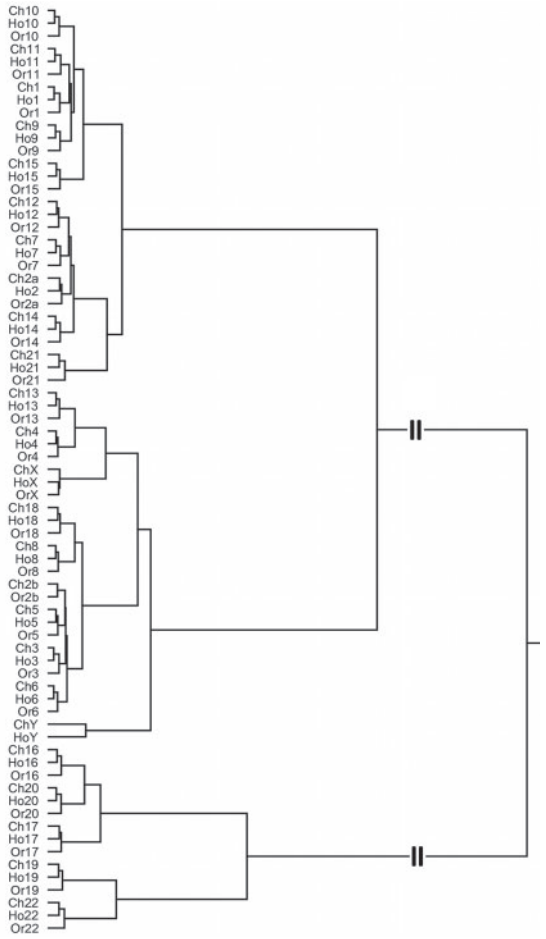

Fig. 5. Dendogram for the {Ho, Ch, Or} chromosomes with $n=8$ (rightmost clustering compressed).

Figure 9 depicts how some of 18 mitochondrial chromosomes are most related to each other at a primary (continuous line) and secondary (dashed line) levels. Only links with correlation $\geq 30\%$ are shown, for $n=8$. We can observe that there are five primary clusters of chromosomes in the graph: from left to right invertebrates, primates, {Rn, Mm, Op}, {Ox, Po, Eq, Dg} and {Ga, Tg}.

The dendogram of Figure 10 describes the hierarchical clustering of mitochondrial genomes using the 'Joint Between Within' clustering method, with the topmost clustering compressed for clarity. Only species that associate in groups of two or more are depicted in the dendogram of Figure 10 (i.e. 10 mammals, 2 aves, 2 fishes and 2 nematodes).

4 DISCUSSION

We described a novel approach to DNA analysis, based on a methodology that takes as input whole-genomic chromosome sequences and then, using alignment-free sequence techniques, extracts high-level information from histogram correlations. This information is used to generate several types of tabular and/or graphical outputs relating chromosomes or species.

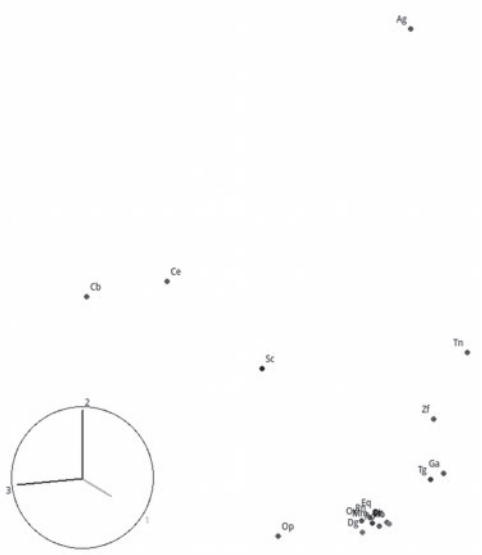

Fig. 6. 3D MDS plot of the nuclear genome of 18 species for $n=8$. Shaded dots represent species, all labeled for readability.

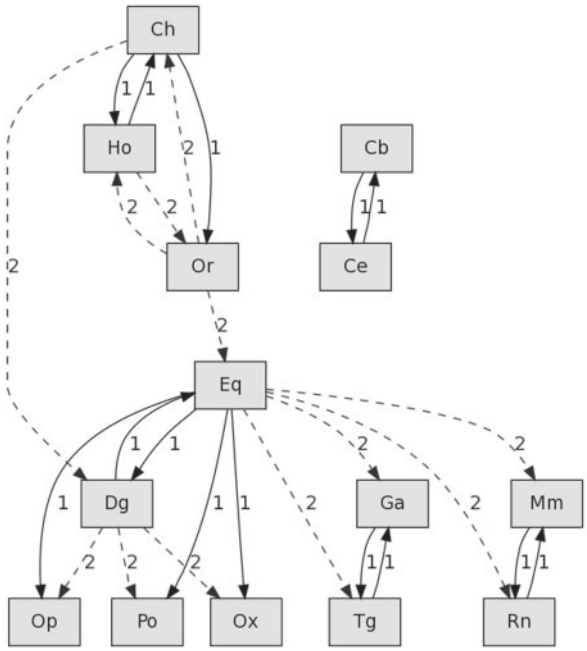

Fig. 7. Graph of the two most correlated species for $n=8$ (correlation $\geq 75\%$). Gray rectangle: chromosome, link r: connection of similarity r between two species.

In this study, an important parameter is the word length n, used in the sequence processing and histogram constructions steps. The $n=1$ case is just a mere counting of {T, C, A, G}, but when n goes from 2 to 8 the corresponding 3D MDS plots reveal increasingly evident spatial and structural patterns related to chromosomes, as shown in Figure 11 for $n=8$. In this 3D rendering (with 'shadows on the floor'), we can observe the individual chromosomes and many

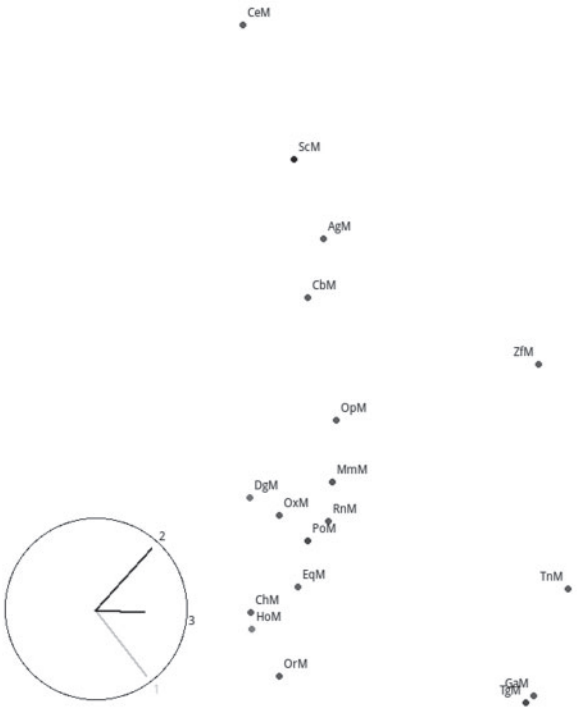

Fig. 8. 3D MDS plot of the mitochondrial genome of 18 species for $n=8$. Shaded dots represent species, all labeled for readability.

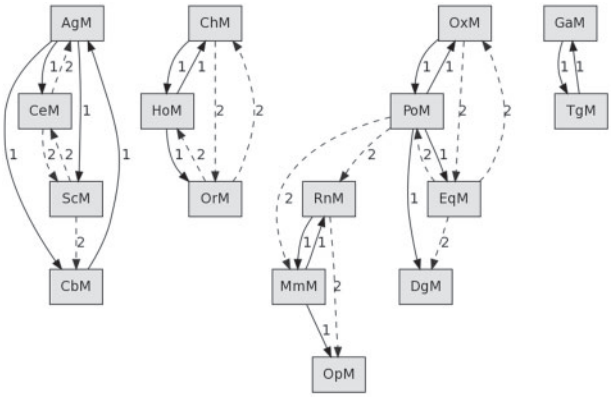

Fig. 9. Graph of the two most correlated mitochondrial chromosomes for $n=8$ (correlation $\geq 30\%$). Gray rectangle: chromosome, link r: connection of similarity r between two chromosomes.

spatial groupings: a big grouping including all the mammals (with Op slightly apart); a grouping with Ga and Tg; and a grouping with Ce and Cb. Fish species Zf and Tn are spatially far away from each other. It is also noteworthy the existence of spatial structure in most species: other than {Cb, Ce, Sc, Zf}, the remaining species reveal a chromosomal 'linear organization', with 'lines' mostly parallel between species. In most of the species, it is the sexual chromosome that lies more far apart from the corresponding species' 'line', but this also happens with other non-sexual chromosomes.

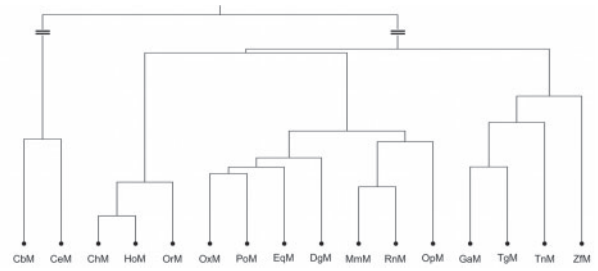

Fig. 10. Dendogram based on mitochondrial genomes for $n=8$ (topmost clustering compressed).

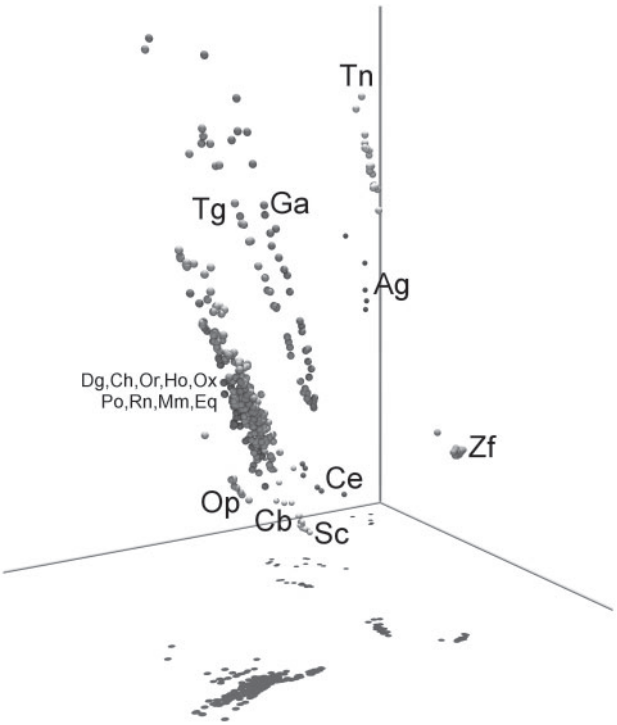

Fig. 11. 3D rendering of the MDS plot for the 384 nuclear chromosomes when $n=8$ (shadows visible in the bottom of the figure).

We do not have an immediate explanation for this remarkable structuring, but it may be related with higher levels of information in chromosomes and genomes. The emergence of this apparently new structural knowledge from chromosomal DNA seems to be unique to our approach and suggests novel ways to investigate the higher levels of information referred by (Seitz, 2007).

We also performed a study using nuclear and mitochondrial chromosomes from 18 species, with results at the chromosome, genome and species levels:

- 3D MDS plots of nuclear and mitochondrial chromosomes have shown the emergence of intra- and interspecies spatial patterns, based on an alignment-free sequencing and data mining of the whole-chromosomal DNA data, instead of genes or other small sequences.

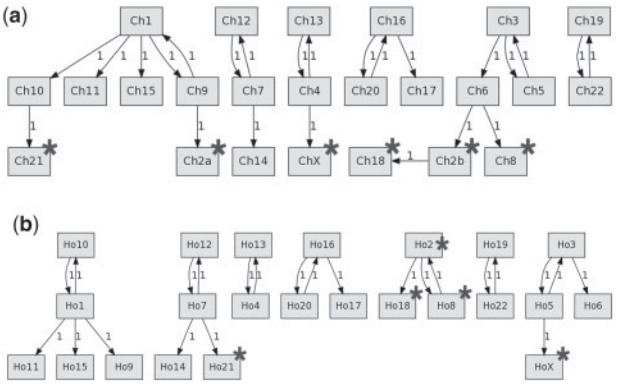

Fig. 12. Graphs of the most correlated nuclear chromosomes for $n=8$ in chimpanzee and human. Gray rectangle: chromosome, link: connection of highest similarity between two chromosomes. Asterisks mark the main differences in chromosomal relationships. (**a**) Chimpanzee ($n=8$, correlation $\geq 90\%$). (**b**) Human ($n=8$, correlation $\geq 90\%$).

- Directed graphs of nuclear and mitochondrial chromosomes demonstrated that chromosomal relationships can be derived from histogram correlations, as well as genomic/species relationships.
- Chromosomal-based dendograms have shown that histogram correlations can be used to compute diagrams depicting hierarchical clusterings of chromosomes and species.

In the dendogram of Figure 5, we observe that human, chimpanzee and orangutan chromosomes cluster into three main groups:

- 10 + 11 + 1 + 9 + 15 + 12 + 7 + 2/2a + 14 + 21
- 13 + 4 + X + 18 + 8 + 2b + 5 + 3 + 6 + Y
- 16 + 20 + 17 + 19 + 22

being the last one the most 'different'. There is no immediate explanation for this 16 + 20 + 17 + 19 + 22 cluster separation.

Using DNA base sequences partitioned into chromosomes, this study also showed the possibility of obtaining high-level chromosomal information like the most interspecies correlated chromosomes (Figs 4 and 7).

Looking at the dendogram of Figure 10 and considering the small amount of information stored in the mitochondrial genome, it seems that it can be regarded as a kind of species' signature. This dendogram depicts a hypothetical phylogenetic tree very similar, in qualitative terms, to those described in Wildman *et al*. (2007), Murphy *et al*. (2007), Zhao and Bourque (2009), Prasad and Allard (2008), Ebersberger *et al*. (2007), Dunn *et al*. (2008) and Hillier *et al*. (2004). It should be noted that the processes described by the aforementioned authors to generate the phylogenetic trees use portions of the DNA base sequence, and not a 'transformed' version like the one presented in this study. It should also be mentioned that whole DNA chromosomal sequences were used, not just portions like genes or other partial sequences.

Using the described histogram-correlation approach, it is also possible to generate outputs showing the most intraspecies correlated chromosomes and highlighting the differences. Figure 12 shows it for chimpanzee and human, with structural distinctions marked with an 'asterisk'.

In Figure 12, we can observe that chromosomes Ch21/Ho21 are linked to distinct chromosome groups (1 + 9 + 10 + 11 + 15 in chimpanzee, 7 + 12 + 14 in human), as well as ChX/HoX (4 + 13 in chimpanzee, 3 + 5 + 6 in human) and others.

All described results contribute to the notion that nuclear and mitochondrial genomes include structural information that allows chromosomal analysis and other high-level analysis, as well as species-related studies (e.g. evolutionary/comparative genomics and phylogenetic tree construction).

4.1 Open issues and future work

In this study, we have used chromosomal information that is incomplete, as explained in the UCSC Genome Bioinformatics web site. For many of the species referred in Table 1, there is a considerable amount of DNA sequence data that is not yet attached to chromosomes or, being associated to a certain chromosome, with its placement not yet defined. This informational uncertainty is undesirable and prone to contribute to misleading results, which are not caused by the mathematical and computational tools adopted.

For data processing we used, for DNA sequence lengths, values of $n=1,\ldots,8$. Although larger values of n are admissible, it should be noted that the total number of histogram bins is $m=4^n$ and, with a very large m, most of the histogram bins may become zero for the smaller chromosomes. For example, for the smallest nuclear chromosome considered (Sc1, with 231 k bases), for $n=8$ then $m=65\,536$ and there is an average of 3.5 samples per histogram bin, but for Sc1 with $n=10$ the average number of samples per histogram bin drops to 0.2 (i.e. 4/5 of bins equal to zero). From the empirical evidence gathered, the most promising values of n are 6, 7 and 8. Further research should address and clarify this issue.

The Kendall τ rank correlation method has proved to be adequate for generating the correlation matrix $\mathbf{S}$, but other correlation methods were also tested. This issue will be the subject of further research and evaluation.

Finally, the study should be repeated when complete genomic data becomes available, and extended to more species, eventually with greater 'biological diversity'. As soon as more DNA species' data are available or updated, this issue will be addressed.

ACKNOWLEDGEMENTS

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/genomes/view/pongo_abelii/
- Pig – The Swine Genome Sequencing Consortium, http://piggenome.org/
- Ox – The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/bovine/
- Dog Genome Sequencing Project – http://www.broad.mit.edu/mammals/dog/, Lindblad-Toh K, *et al*. Genome sequence,

comparative analysis and haplotype structure of the domestic dog. *Nature* 2005;**438**, 803–819.

- Horse – The Broad Institute, http://www.broad.mit.edu/mammals/horse/
- Mouse – Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002, **420**, 520–562, http://www.hgsc.bcm.tmc.edu/projects/mouse/
- Rat – The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/rat/, Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004, **428**(6982), 493–521.
- 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; **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/
- *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/

The authors would like to thank the reviewers for their comments and the opportunity to improve this article.

Conflict of Interest: none declared.

REFERENCES

Borg,I. and Groenen,P. (2005) *Modern Multidimensional Scaling-Theory and Applications*, 2nd edn. Springer, USA.

Chaa,S.-H. and Srihari,S.N. (2002) On measuring the distance between histograms. *Pattern Recogn*, **35**, 1355–1370.

Christensen,D. (2005) Fast algorithms for the calculation of Kendall's τ. *Comput. Stat.*, **20**, 51–62.

Cox,T. and Cox,M. (2001) *Multidimensional Scaling*, 2nd edn. Chapman & Hall/CRC, USA.

Dunn,C.W. *et al*. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452,** 745–750.

Ebersberger,I. *et al.* (2007) Mapping human genetic ancestry. *Mol. Biol. Evol.*, **24**, 2266–2276.

Fernández,A. and Gómez,S. (2008) Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *J. Classif.*, **25**, 43–65.

Hillier,L.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.

Kendall,M. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–89.

Kruskal,J. and Wish,M. (1978) *Multidimensional Scaling*. *Sage University Paper series on Quantitative Application in the Social Sciences, No. 07-011*. Sage Publications, London.

Ling,H. and Okada,K. (2006) Diffusion distance for histogram comparison. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, pp. 246–253.

Machado,J.T. *et al.* (2010) Fractional dynamics in DNA. *Commun Nonlinear Sci Numer Simulat.*, **16**, 2963–2969.

Maddison,D.R. *et al.* (2007) The tree of life web project. In linnaeus tercentenary: progress in invertebrate taxonomy (ed. Zhang ZQ, Shear WA). *Zootaxa*, **1668**, 1–766 (19–40).

Murphy,W.J. *et al.* (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**, 398–401.

Prasad,A.B. and Allard,M.W. (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol. Biol. Evol.*, **25**, 1795–1808.

Schuh,R.T. and Brower,A.V.Z. (2009) *Biological Systematics: Principles and Applications*, 2nd edn. Cornell University Press, Ithaca, USA.

Seitz,H. (2007) Analytics of protein-DNA interactions. In Seitz,H. (eds) *Advances in Biochemical Engineering Biotechnology*. Springer, Berlin/Heidelberg.

Shepard,R. (1962) The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, **27**, 219–246, 219–246.

Tzeng.J. *et al.* (2008) Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, **9**, 179.

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison - a review. *Bioinformatics*, **19,** 513–523.

Werman,M. *et al.* (1985) A distance metric for multidimensional histograms. *Comput. Vis. Graph. Image Process.*, **32**, 328–336.

Wildman,D.E. *et al.* (2007) Genomics, biogeography, and the diversification of placental mammals. *Proc. Natl Acad. Sci. USA*, **104**, 14395–14400.

Zhao,H. and Bourque,G. (2009) Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.*, **19**, 934–942.