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 59,737,566 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

<table>
<thead>
<tr>
<th>Species</th>
<th>Group</th>
<th>Nuclear/mitochondrial chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (Ho)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X Y M</td>
</tr>
<tr>
<td>Chimpanzee (Ch)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 14 15 16 17 18 19 20 21 22</td>
</tr>
<tr>
<td>Orangutan (Or)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13 14 15 16 17 18 19 20 21</td>
</tr>
<tr>
<td>Pig (Po)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td>Opossum (Op)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 X M</td>
</tr>
<tr>
<td>Horse (Eq)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 15 16 17 18 19 20 21 22</td>
</tr>
<tr>
<td>Dog (Dg)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 15 16 17 18 19 20 21 22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23 24 25 26 27 28 29 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31 X M</td>
</tr>
<tr>
<td>Ox (Ox)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 15 16 17 18 19 20 21 22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23 24 25 26 27 28 29 X M</td>
</tr>
<tr>
<td>Mouse (Mm)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 15 16 17 18 19 20 21 22</td>
</tr>
<tr>
<td>Rat (Rn)</td>
<td>Mammal</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 15 16 17 18 19 20 21 22</td>
</tr>
<tr>
<td>Chicken (Ga)</td>
<td>Ave</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td>Zebra Finch (Tg)</td>
<td>Ave</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td>Zebrafish (Zf)</td>
<td>Fish</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td>Tetraodon (Tn)</td>
<td>Fish</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td>Mosquito (Ag)</td>
<td>Insect</td>
<td>21 2r 3l 3r X M</td>
</tr>
<tr>
<td>Caenorhabditis elegans (Ce)</td>
<td>Worm</td>
<td>1 2 3 4 5 X M</td>
</tr>
<tr>
<td>Caenorhabditis brigaeae (Ch)</td>
<td>Worm</td>
<td>1 2 3 4 5 X M</td>
</tr>
<tr>
<td>Yeast (Sc)</td>
<td>Fungus</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 15 16 M</td>
</tr>
</tbody>
</table>

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

Table 2. Differences in two approaches for ignoring ‘N’ symbols

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Sequences with ‘N’ removed</th>
<th>Sequences with ‘N’ filtered</th>
<th>(a−β)/β (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ga25</td>
<td>1367889</td>
<td>1366030</td>
<td>0.136088</td>
</tr>
<tr>
<td>Ga3</td>
<td>110204947</td>
<td>110177075</td>
<td>0.025297</td>
</tr>
<tr>
<td>Tn1</td>
<td>20304845</td>
<td>20315377</td>
<td>0.051869</td>
</tr>
<tr>
<td>Tn15</td>
<td>6235253</td>
<td>6236842</td>
<td>0.025484</td>
</tr>
<tr>
<td>AgX</td>
<td>21470369</td>
<td>21477782</td>
<td>0.034527</td>
</tr>
<tr>
<td>Ag2l</td>
<td>48065344</td>
<td>48071405</td>
<td>0.012423</td>
</tr>
<tr>
<td>HoY</td>
<td>25653559</td>
<td>25653447</td>
<td>0.000437</td>
</tr>
<tr>
<td>Ho5</td>
<td>177695253</td>
<td>177695218</td>
<td>0.000020</td>
</tr>
</tbody>
</table>

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

Advantageously used without compromising the quality of results and conclusions.

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 \( r > 0.9999999 \). 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 \( m = 4 \) 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, we developed the 'point' sliding window approach, which requires as parameters the chosen sequence length (n) and filtered DNA sequence files, generating the corresponding histogram files.

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 'Point Between Within' clustering method (Fernandez and Gomez, 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, 2, 3, 4, 5, 6, 7, 8}, for each value of n, we applied the Kendall τ correlation method (Kendall, 1938) to generate the corresponding 384 × 384 \( S_{\text{nucl}} \) 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 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, 2, 3, 4, 5, 6, 7, 8}, we noted 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 \( S_{\text{nucl}} \) 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 pointed chromosome is the second most correlated. In Figure 4, only links with correlation ≥ 95% are visible and it shows that (Ho, Ch, Or) chromosomes with the same ‘number’ are more similar.
correlated to each other, although there are some exceptions: for example, the Y chromosomes and the 2a and 2b chromosomes in {Ch, Or}.

The $S_{nuclear}$ 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

... continue with the rest of the text...
Dendrogram 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, Mn, Op\}, \{Ox, Po, Eq, Dg\} and \{Ga, Tg\}.

The dendrogram 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 dendrogram 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.

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 \{Ch, 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.
Figure 12 shows it possible to generate outputs showing the most intraspecies correlated like genes or other partial sequences. Whole DNA chromosomal sequences were used, not just portions like the one presented in this study. It should also be mentioned that portions of the DNA base sequence, and not a ‘transformed’ version the aforementioned authors to generate the phylogenetic trees use et al. (2004). It should be noted that the processes described by (2007), Dunn et al. (2008) and Hillier et al. (2007), Zhao and Bourque (2009), Prasad and Allard (2007), Murphy et al. (2007), 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 chromosomal differences and highlighting the differences. Figure 12 shows it possible to generate outputs showing the most intraspecies correlated like genes or other partial sequences.

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 \( \tau \) rank correlation method has proved to be adequate for generating the correlation matrix \( S \), 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:

- Common Chimpanzee – Chimpanzee Genome Sequencing Consortium
- Orangutan – Genome Sequencing Center at WUSTL, http://genome.wustl.edu/genomes/view/pongo_abeli/
- 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,

- Horse – The Broad Institute, http://www.broad.mit.edu/mammals/horse/
- Opossum – The Broad Institute, http://www.broad.mit.edu/mammals/opossum/
- Zebra Finch – Genome Sequencing Center at Washington University in St Louis School of Medicine
- Zebratsih – The Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/Projects/D_riego/
- Gambiace Mosquito – The International Anopheles Genome Project
- Briggsae nematode – Genome Sequencing Center at Washington University in St Louis School of Medicine

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

**Conflict of Interest:** none declared.

**REFERENCES**


