DNA Patterns Localization Using a Dedicated Dot-Plot and Image Analysis

Petre G. Pop

***Abstract*— DNA repeats are believed to play significant roles in genome evolution and diseases. Many of the methods for finding repeated sequences are part of the digital signal processing (DSP) field and most of these methods use distances, similarities and consensus sequences to generate candidate sequences. This paper presents results obtained using a dedicated numerical representation with a mapping algorithm (using DNA distances and consensus types) and a custom dot-plot analysis (using similarities to represent DNA patterns) combined with image processing techniques, to visual isolate the position of DNA patterns with different lengths. The final images that best put in evidence the presence of repeated sequences were obtained using weighted cosine cross-correlation, Jukes-Cantor distance and Motyka similarity.**

***Keywords*— DNA repeats, DNA numerical representations, DNA distances, dot-plot analysis, numerical similarities.**

I. Introduction

The existence of DNA repeats is a fundamental characteristic of all biological genomes. A repeat is the simplest form of regularity and the detection of repeats (pattern, length, position, repeat number) is important in biology and medicine as it can be used for phylogenic studies and disease diagnosis. A major difficulty in identification of DNA repeats is caused by the fact that the repeat units can be of different lengths and either in tandem or dispersed or exact or imperfect. Although research in this domain is made for many years, the problem is not yet considered solved. The main approaches for repeats detection are reviewed in [1].

DNA or nucleotides sequences are represented by sequences of the characters A, T, C and G, corresponding to nucleotides A (adenine), T (thymine), C (cytosine) and G (guanine). DNA repeats are sequences that are present in more than one copy. An exact repeat is a short string of nucleotides repeated contiguously at least twice. An approximate repeat is a string of nucleotides repeated consecutively with differences between the instances (mainly due to different types of mutations). Repeats, whose copies are distant in the genome, are named distant or dispersed repeats. Among those, we can distinguish micro-satellites, mini-satellites, and satellites, depending on the length of their repeated unit [1].

The interest in finding DNA repeats may be theoretical, technical or medical, as follows:

- Theoretical interest: related to their role in the structure and evolution of the genome.
- Technical interest: repeats can be used as polymorphic markers, either to trace the propagation of genetic traits in populations or as genetic identifiers in forensic studies.
- Medical interest: the presence of specific types of tandem repeats has been associated to different severe diseases (e.g. Huntington's disease, myotonic dystrophy). In healthy individuals, the tandem repeat size varies around a few tens of copies, while in affected individuals the number of copies at the same locus reaches hundreds or even a thousand in some cases.

The centromere of most complex eukaryotic chromosomes is a specialized locus comprised of repetitive DNA that is responsible for chromosome segregation at mitosis and meiosis. Alpha satellite DNA is composed of a tandem array of repeat units and has been identified at every human centromere. There are two major types of alpha satellite, higher-order and monomeric [2]. Higher-order alpha satellite is the predominant type in the genome and made up of ~171 bp (base pairs) monomers organized in arrays of multimeric repeat units that are highly homogeneous. Monomeric alpha satellite lies at the edges of higher-order arrays and lacks any higher-order periodicity; its monomers are only on average ~70% identical to each other [2]. Our present research was focused on determining these alpha satellites DNA.

The numerical representation of DNA sequences becomes very important as almost all DSP techniques require two parts: mapping the symbolic sequence (letters corresponding to nucleotides) into a numeric form and calculating a kind of transform of the resulting numeric sequence [3]. Most of the numerical representations associate a single numerical value to one position in the sequence using numerical values associated to each nucleotide and, finally, reflect the presence or the absence of a certain nucleotide in a specific position (e.g. indicator sequences) [3]. Another approach could be to include information about the number and type of consecutive nucleotides and to generate only one numerical value for each DNA subsequence which may be associated with a

Manuscript received November 15, 2015, revised December 21, 2015

Petre G. Pop is with the Technical University of Cluj-Napoca, Romania, Communication department (e-mail: petre.pop@com.utcluj.ro).

doi: 10.11601/ijates.v4i3.140

repeat. In this regard, we have introduced a representation which considers the length of the expected repeats and the number of possible mismatches, based on polynomial-like representation [4]. This representation needs a mapping algorithm which uses distances and then evaluates a consensus sequence to generate candidates.

This paper presents results obtained using this dedicated numerical representation with associated mapping algorithm (that uses DNA distances and consensus types) combined with a custom dot-plot analysis (involving similarities to represent DNA patterns) to generate in-memory dot-plot images which are then processed to highlight the position of repeats with different lengths in DNA sequences.

II. Numerical Representation And Mapping Algorithm

In a previous work [4] we have proposed a DNA numerical representation and a mapping algorithm, which includes both the length of the DNA repeats and the number of mismatches due to point mutations. For a DNA sequence of length *L*, a numerical value is associated using a polynomial-like representation:

$$V = \sum_{k=0}^{L-1} V_{\alpha_k} 10^k, \quad \alpha \in \{A,G,C,T\}\,, \tag{1}$$

where V_α is the value of a single nucleotide. These coefficients should be different natural numbers such that the resulting numerical value is unique for a given DNA subsequence.

However, in the case of two DNA sequences with a high degree of similarity (which differ, for instance, by a single nucleotide) representation (1) will produce two very different values. Therefore we need an algorithm that allows finding similar sequences (considering possible mismatches), determine the associate consensus sequence and then generates a single numerical value for these similar sequences, using (1) for calculated consensus sequence.

To pass from initial DNA sequence to a single, final sequence of numerical values, we need both a type of distance and a type of consensus:

- The distance should measures the number of mismatches between DNA subsequences of the same length; if two subsequences are identical, this distance should be zero.
- Given a number DNA of subsequences of the same length, the consensus sequence is a sequence pattern derived from multiple, similar DNA sequences that represents the nucleotide most likely to occur at each position in analyzed sequences.

The proposed mapping algorithm has the following steps:

- Step-1: Consider all successive DNA subsequences of same length *L*;
- Step-2: Determine all the positions (and the associated subsequences) in the original DNA sequence for which the distance (against a subsequence from Step-1) is less or equal to the prefixed maximum mismatches allowed number M_m;
- Step-3: Determine the consensus sequence for all (similar) subsequences from Step-2; Calculate the distance between the consensus sequence and each associated subsequence; those subsequences whose distance is greater than M_m must be reassign; Determine again the consensus sequence for remaining sequences.
- Step-4: Using (1), compute the numerical value for consensus sequence and assign this value to all starting positions of subsequences determined in Step-2.

As output, the algorithm generates a single sequence of numbers; each number is associated to a unique subsequence of length *L* (possible a repeat unit).

An important property of this mapping algorithm is that if the *L* value is a prime factor of repeated sequence length then the entire repeated sequence will be emphasized. This allows a significant reduction of the computational effort in case of long repeats. On the other hand, the final numerical values contain information about the structure of associated consensus sequence, which can be used to specify the structure of detected repeated sequences.

Finding similar sequences in Step-2 and determination of consensus sequence requires evaluating the distance between two DNA subsequences. In our experiments we used Hamming distance and Jukes-Cantor distance (an evolutionary distance).

The Hamming distance determines the number of different nucleotides between two equal length DNA sequences.

Let *x* and *y*, two DNA sequences of length *n*. The Jukes–Cantor distance between DNA sequences is defined by [5]:

$$d_{JC}(x,y) = -\frac{3}{4}\ln\left(1-\frac{4}{3}\frac{\sum 1_{x_i \neq y_i}}{n}\right). \tag{2}$$

We also used these distances in Step-3 to determine the distance between the consensus sequence and candidate DNA subsequences.

Also, in Step-3 we determine the consensus sequence for all similar DNA subsequences determined in Step-2 and, on this basis, we calculate the associated numerical value (using (1)).

We used the following types of consensus:

- Most frequently occurring nucleotide (in each position or column), even if it is not the majority.
- Majority with fixed cutoff: use the fraction of nucleotides in a position to establish majority for that position, provided that the fraction is greater than the cutoff parameter.

- Majority with global appearing frequency cutoff: same as previous case but the cutoff for each nucleotide is computed as the appearing frequency in the original sequence.
- Majority with local appearing frequency cutoff: same as previous case but the cutoff for each nucleotide is computed as the appearing frequency in the analyzed similar subsequences.

For last three consensus types, if there is no nucleotide that exceeds the threshold we consider that we have no valid consensus sequence and those subsequences must be reassign. In case that more than one nucleotide is calculated to have the same confidence, and this exceeds the consensus threshold, the nucleotides are assigned in descending order of global appearing frequency precedence [5].

III. Numerical Similarities And Dot-Plot Analysis

In case of DNA sequences, dot plots are two-dimensional representations where each axis represents a sequence (possible the same) and the plot itself shows a comparison of analyzed sequences by a calculated score for each position of the sequences. Most of time dot plots are used to determine regions of similarity within a single DNA sequence (i.e. repeats) or between two different sequences.

Some important characteristics of patterns appearing in DNA dot plots are [6]:

- Parallels to the main diagonal indicate repeated regions on different parts of the analyzed sequences (Fig. 1-a).
- Blocks of parallel lines indicate tandem repeats in both sequences and the distance between the lines equals the distance of the repeats (Fig. 1-b).

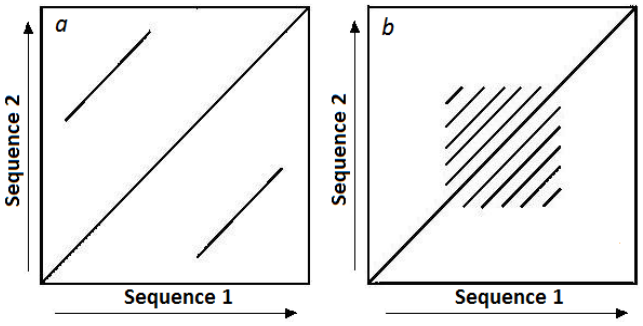

Fig. 1. Characteristic patterns appearing in dot plots.

If the length of sequences is large, windows are used to perform the analysis. In this case, if a window of fixed size on one sequence (one axis) match another window (possible of different size) on the other sequence, a dot is drawn at the plot.

To evaluate the results obtained in the experiments we need a customized dot-plot analysis as:

- The analyzed sequence is not symbolic but a numerical one (the output of mapping algorithm).
- In most cases the length of analyzed sequence far exceeds the number of points represented on each axis. For this reason, the dot plot analysis is done using windows.
- Due to the large number of values resulted from mapping and the different resolution on each axis, we need to determine the degree of similarity between windows of different lengths to decide if a dot will be plot or not.

To determine the degree of similarity between two numerical sequences of length *m* and *n*, a similarity coefficient is calculated for two equal sequences of length *n*, *(m-n)* times, with a sliding window, then determine the average value.

In our experiments we used several similarities: cross-correlation, Motyka, Bray-Curtis, Kulczynski-1, Kulczynski-2, Ruzicka, Roberts, and Baroni [7]. The best results were obtained using: Motyka, Bray-Curtis and Kulczynski-2 similarities.

The Motyka similarity is defined by [7]:

$$s_{Mot}(x,y) = \frac{\sum \min(x_i, y_i)}{\sum (x_i + y_i)} . \tag{3}$$

Bray–Curtis similarity is defined by [7]:

$$\frac{2}{n(\bar{x}+\bar{y})} \sum \min(x_i, y_i) . \tag{4}$$

The definition of Kulczynski similarity 2 is [7]:

$$\frac{n}{2}\left(\frac{1}{\bar{x}}+\frac{1}{\bar{y}}\right) \sum \min(x_i, y_i) . \tag{5}$$

The average similarity coefficient is scaled in the interval (0, 1) and based on its values, a dot with a grey level value between 0 and 255 will be plot.

Finally, we obtain a gray level image to which we applied a threshold based on the mean (μ) and image variance (σ), using the formula [8]:

$$T_h = k_1 \cdot \mu + k_2 \cdot \sigma . \tag{6}$$

The constants k_1 and k_2 are image type dependent and, after some experiments, we used $k_1 = 1$ and $k_2 = 1.75$.

Following this procedure we obtain a graphical representation for each combination of parameters *L* (length of the searched repeated sequence), M_m (maximum number of allowed mismatches). The quality of representation varies depending on the number of similar subsequences of the same length that are found in the original sequence.

The graphical representation allows locating repeated sequences of a certain length using position of segments parallel to the main diagonal, position represented on each axis. Using a prime factor of repeated sequence length for parameter *L* allows highlighting of patterns in the dot-plot image. In this case it is necessary to run the application using multiple combinations for parameters values, followed by analysis of the images obtained to determine the length of repeated sequences. Therefore, a priori information is needed about the domain values for the length of repeated sequences.

To determine the length of repeated sequences, we used another approach that allows calculation of dot-plot images, in memory, for several values of the parameter L (length of repeated sequences). Then extract some features from each image that allows the localization of repeated sequences, features that are represented in the form of intensities on a single line, for each value of L.

For this purpose we extracted next features from each in-memory dot-plot image:

- The average density for each column of the image:

$$D_i = \frac{\sum_{j=1}^{N} I(i,j)}{N - n^*}, i = 1, N \ , \qquad (7)$$

where, n^* is the number of column not null values.

- Lagged autocorrelation of image values from each column of the image:

$$r_{k,i} = \frac{\sum_{j=1}^{N-k} (I(i,j) - \overline{I_i})(I(i+k,j) - \overline{I_i})}{\sum_{j=1}^{N} (I(i,j) - \overline{I_i})^2}, i = 1, N \ . \qquad (8)$$

- Cross-correlation between two adjacent columns of the image, expressed by the Pearson similarity coefficient:

$$r_i = \frac{\sum_{j=1}^{N} (I(i,j) - \overline{I_i})(I(i+1,j) - \overline{I_{i+1}})}{\sqrt{\sum_{j=1}^{N} (I(i,j) - \overline{I_i})^2 \sum_{j=1}^{N} (I(i+1,j) - \overline{I_{i+1}})^2}}, i = 1, N-1 \cdot \qquad (9)$$

- Cross-correlation between two adjacent columns of the image, expressed by the cosine similarity coefficient:

$$s_i = \frac{\sum_{j=1}^{N} I(i,j) I(i+1,j)}{\sum_{j=1}^{N} I(i,j)^2 \sum_{j=1}^{N} I(i+1,j)^2}, i = 1, N-1 \ . \qquad (10)$$

Following this procedure we can represent results for a set of values of the L parameter (length of searched repeated sequences) as lines (or segmented lines) and the intensities and positions of these lines give information about the presence of repeated sequences and their positions. Our goal was to obtain a synthetic image for a range of values (as widely as possible) of the L parameter so we can easily appreciate the presence of repeated sequences of a certain length and their position.

IV. Experiments And Results

Our case study was the high order repeats in AC010523 from Homo sapiens chromosome 19 (GenBank) and in AC136363 from human chromosome 17 (GenBank) which contain both higher-order and monomeric DNA alpha-satellite [9], of approximately 171 bp (base pairs), arranged in tandem, in a head-to-tail fashion. High-order repeats were identified in the front domain while in the back and central domain, alpha satellite monomers were found [10] (Fig. 4, Fig. 5). Numerical representation based on (1) and associated mapping algorithms were used to obtain the associated numerical sequence.

Several experiments were performed using several combinations of parameters L and M_m, combinations of Hamming and Jukes-Cantor distances (in Step-2 and Step-3 of mapping algorithm) and different similarities (Motyka,

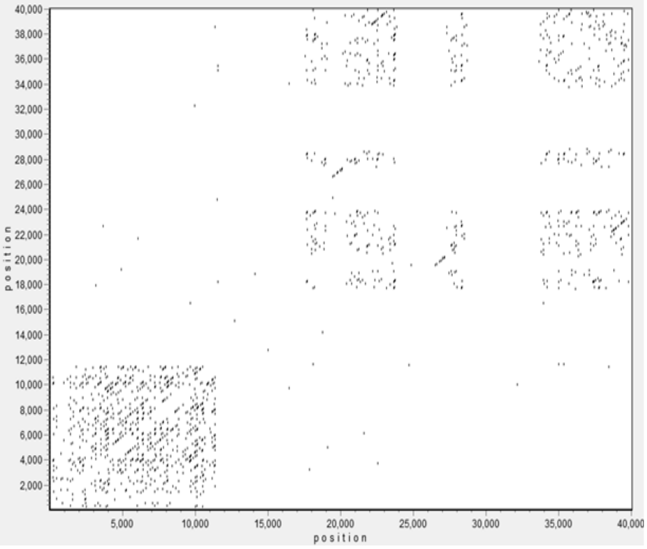

Fig. 2. Lower quality dot-plot image, with $L = 7$, $M_m = 2$, for AC010523

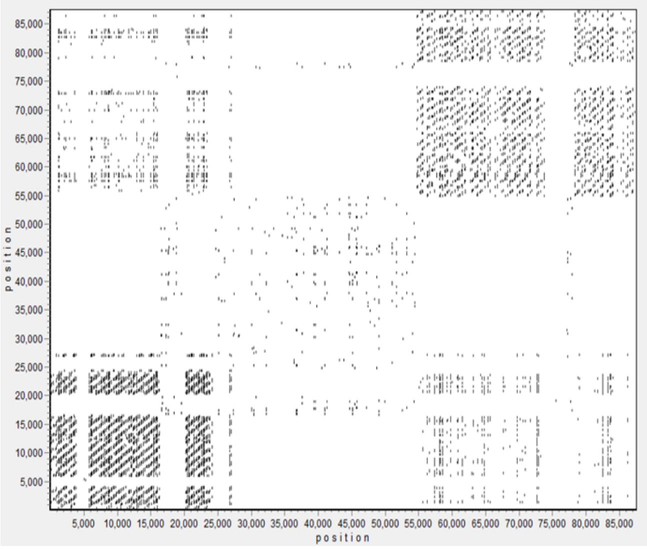

Fig. 3. Lower quality dot-plot image, with $L = 7$, $M_m = 2$, for AC136363

Bray-Curtis, Kulczynsky-2) for in-memory dot plot image construction, as they were introduced in section III.

For consensus sequence evaluation we used majority with local appearing frequency cutoff (use the fraction of nucleotides in a position to establish majority for that position, provided that the fraction is greater than a cutoff, computed as the appearing frequency in the analyzed similar DNA subsequences) [6].

Figures 2, 3, 4, 5 show dot plots results of different qualities using different values for L parameter. As one can see, the value of L = 19 gives better results compared to L = 7 as 19 is a prime factor of the length of the satellites (~171 bp). As mentioned previously, we need more pictures like this, to be able to tell what kind of patterns are present in the analyzed DNA sequence.

The following figures (Fig. 6-13) shows the best results obtained with the second approach using parameter values for L between 5 and 25 (represented on vertical axis) for

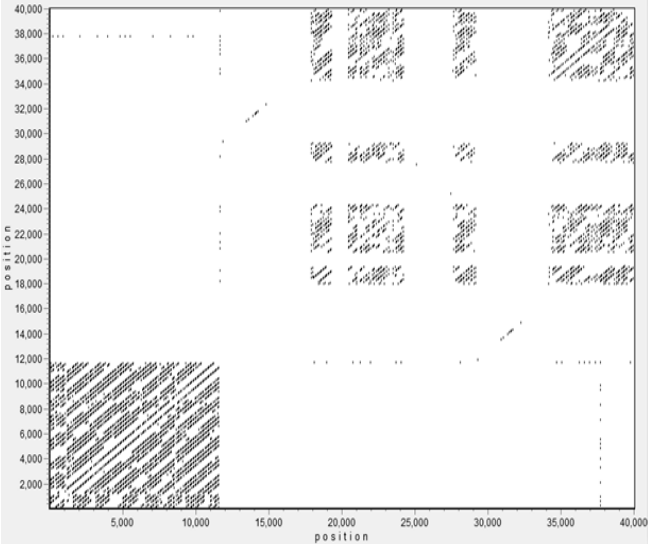

Fig. 4. Higher quality dot-plot image, with L = 19, M_m = 4, for AC010523

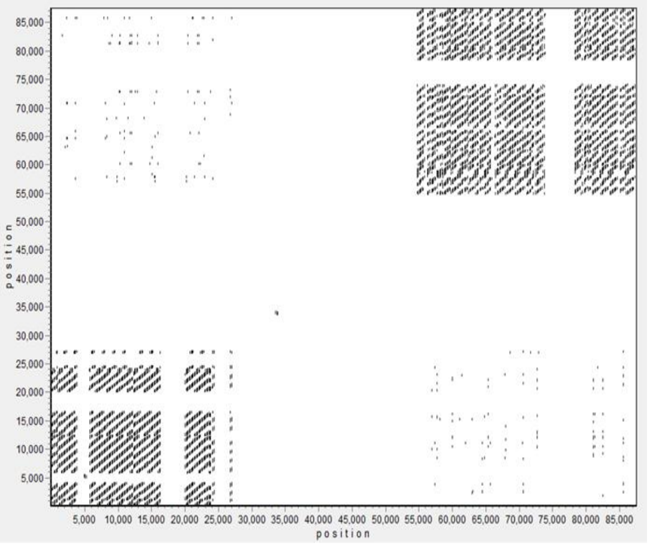

Fig. 5. Higher quality dot-plot image, with L = 19, M_m = 4, for AC136363

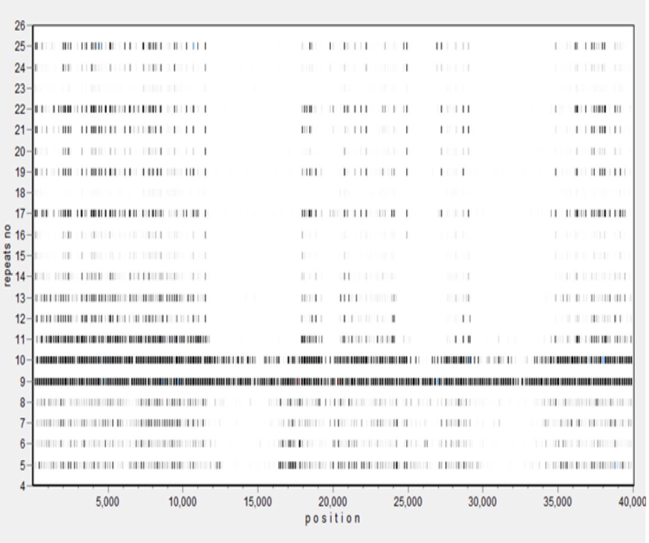

Fig. 6. Results obtained using average density, Hamming-Hamming distances, Motyka similarity (AC010523).

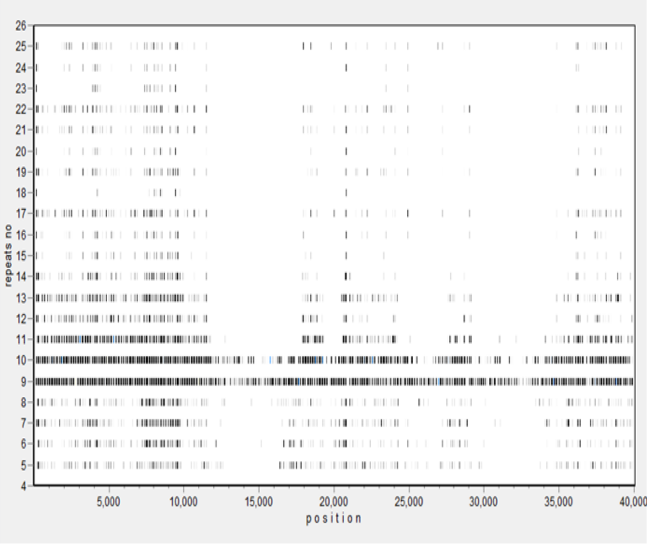

Fig. 7. Results obtained using auto-correlation, Hamming-Jukes Cantor distances, Motyka similarity (AC010523).

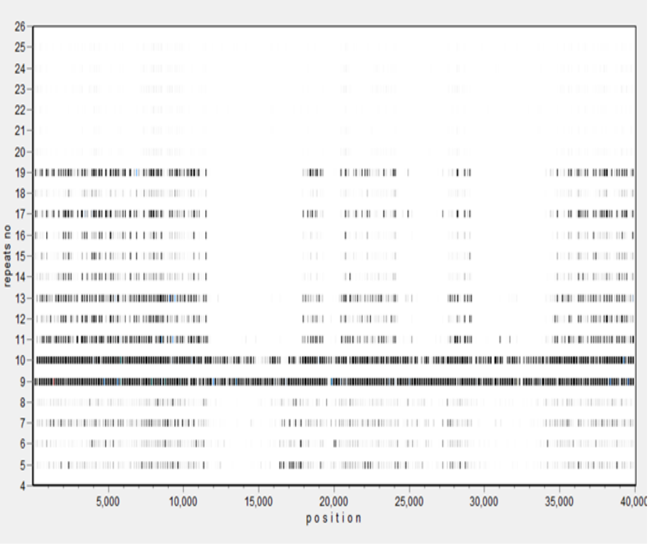

Fig. 8. Results obtained using auto-correlation, Jukes Cantor-Hamming distances, Kulczynski-2 similarity (AC010523).

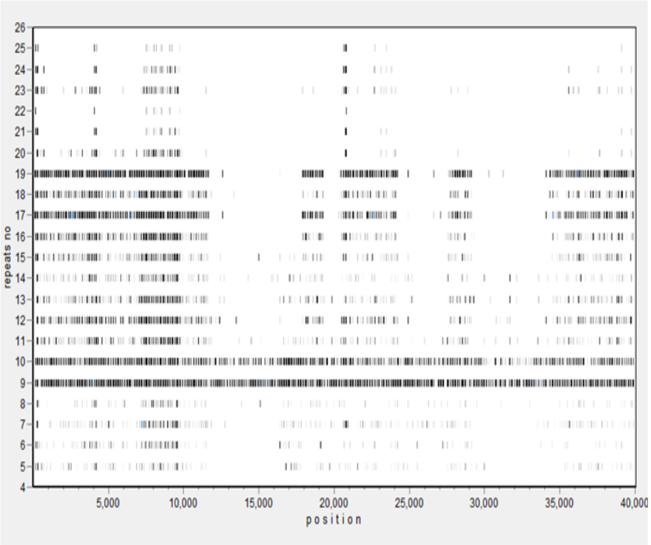

Fig. 9. Results obtained using auto-correlation, Jukes Cantor-Jukes Cantor distances, Kuczynski-2 similarity (AC010523).

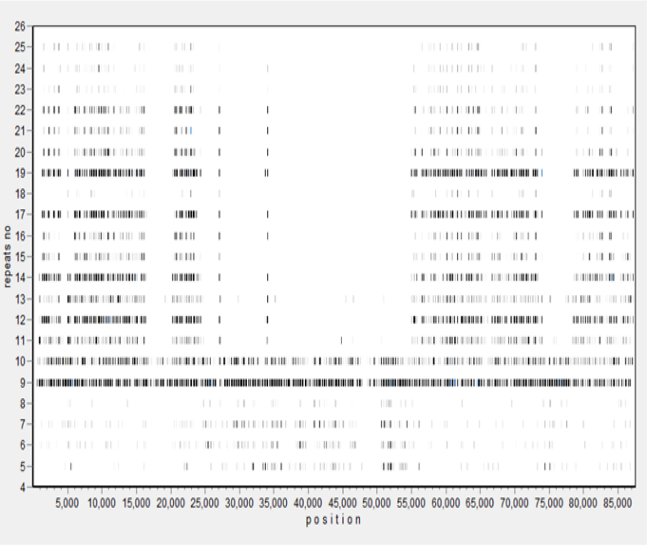

Fig. 12. Results obtained using auto-correlation, Jukes Cantor-Hamming distances, Bray Curtis similarity (AC136363).

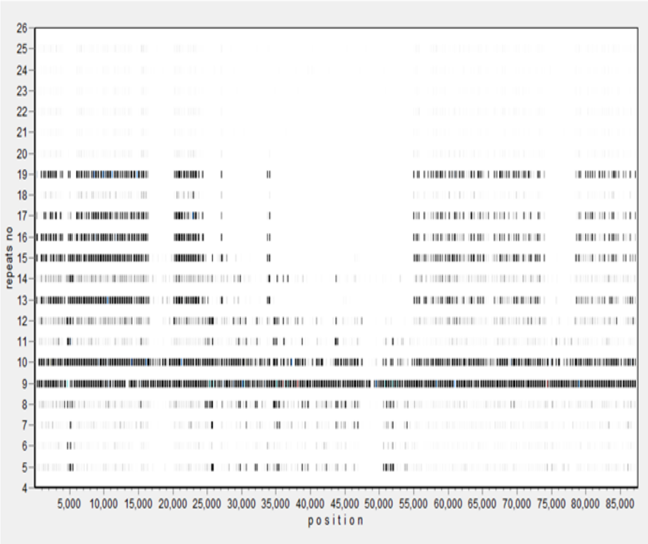

Fig. 10. Results obtained using average density, Hamming-Hamming distances, Kuczynski-2 similarity (AC136363).

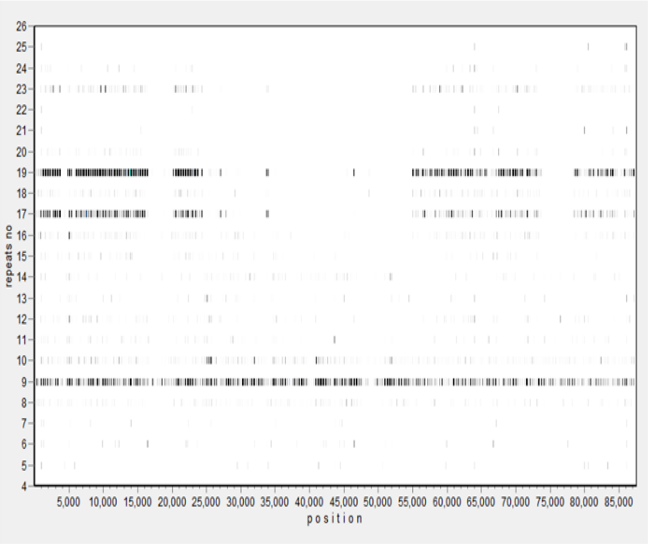

Fig. 13. Results obtained using cosine cross-correlation, Jukes Cantor-Jukes Cantor distances, Motyka similarity (AC136363).

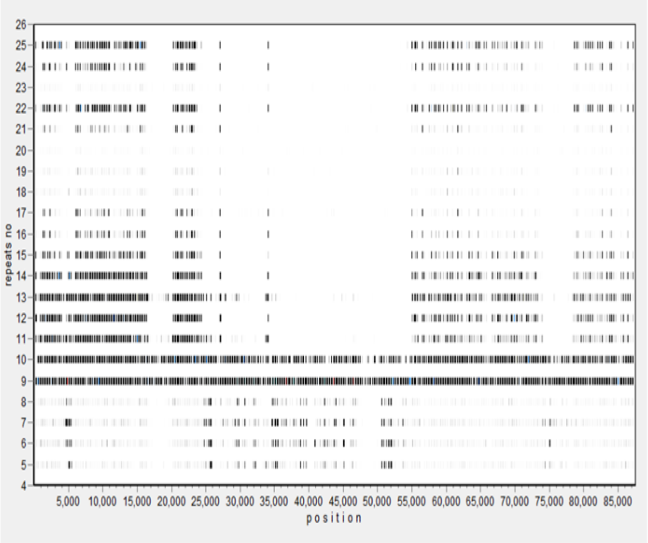

Fig. 11. Results obtained using average density, Hamming-Jukes Cantor distances, Kuczynski-2 similarity (AC136363).

our case studies (AC010523 and AC136363). As a final feature extracted from the in-memory dot-plot images we used: the average density, autocorrelation (with the target repeat length as lag) and cross-correlation (weighted Pearson and cosine similarity with the rate of common not null values), as they were introduced in section III. The images represent the best results for each type of distances combination in Step-2 and Step-3 from the mapping algorithm.

Image analysis presented in previous figures allows us to make the following statements.

When using Hamming-Hamming distances combination: (Fig. 6, Fig. 10):

- Average density allows best results combined with Motyka or Kulczynski-2 similarity.
- Values for L = 9, L = 19 are more visible, both in the front domain and in central and back domain; this can be explained by the fact that both values are the prime factors of 171 (DNA satellite

length).

- Besides the values for L = 9, L = 19, are also highlighted values for L = 10, L = 17 but at a lower intensity.

When using Hamming-Jukes Cantor distances combination: (Fig. 7, Fig. 11):

- Average density and autocorrelation allow best results combined with Motyka or Kulczynski-2 similarity.
- Values for L = 9, L = 19 are more visible, both in the front domain and in central and back domain.
- Other values, such as for L = 10, L = 17, even if they are prime factors of 170 (close to 171), are not highlighted.

When using Jukes Cantor-Hamming distances combination: (Fig. 8, Fig. 12):

- Auto-correlation allows best results combined with Bray-Curtis or Kulczynski-2 similarity.
- Values for L = 9, L = 19 are quite visible, both in the front domain and in central and back domain.
- Values for L = 10, L = 17, are also highlighted but at a lower intensity, as they are prime factors of 170 (close to 171, the DNA satellite length).

When using Jukes Cantor-Jukes Cantor distances combination: (Fig. 9, Fig. 13):

- Auto-correlation and cosine cross-correlation allow best results combined with Motyka or Kulczynski-2 similarity.
- Values for L = 9, L = 19 are more visible, both in the front domain and in central and back domain.
- Values for L = 10, L = 17, are also highlighted but at a lower intensity.
- Other values are less visible (in particular in Fig. 13).

Most images also highlights DNA satellite values in the front domain and the back domain for other values of L parameter. This suggests that there are repeated sequences of higher lengths in those areas.

If using weighted cosine similarity (Fig. 13), the values of the parameter L that are closer to the divisors of length of DNA satellite, are favored particularly those areas with a higher similarity patterns (front domain) while the other values parameter L and remaining areas are less highlighted.

V. Conclusion

An original DNA sequence representation (which includes information about repeats length and the number of mismatches) and a mapping algorithm are used to provide a single associated numerical sequence which is then used for a dedicated in-memory dot-plot representations of DNA patterns. These images can be used to extract some features allowing a suggestive graphical representation of the position and length of repeated sequences. These final images provide visual and numerical information about the length of repeats and their approximate position.

We investigated the effect of several features, distances and similarities on final images. The images that best put in evidence the presence of repeated sequences were obtained using weighted cosine cross-correlation, Jukes-Cantor distance and Motyka similarity.

While final images are based on numerical features, assessment of these images is visual and, consequently, the assessment results (the presence and position of a certain length patterns based on identification of line segments more intense) are approximate. But this information can be very useful in case of long initial sequences for the use of more accurate methods is cumbersome and time consuming (because they generate very long list of candidate sequences, with associated information: score, position, difficult to interpret). This visual information can be used to dramatically narrow the search domain for more accurate methods. Looking ahead, we intend to try to automatic determine the position of those line segments whose intensity exceeds a certain threshold and extracting pattern structure information, stored in the numerical value (associated with a consensus sequence), which is represented in the in-memory images.

References

[1] K. G. Lim, C. K. Kwoh, L. Y. Hsu, A. Wirawan, "Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance," *Brief Bioinform.*, vol. 14, no. 1, pp. 67-81, Jan. 2013.

[2] M. K. Rudd, G. A. Wray, H. F. Willard, The evolutionary dynamics of a-satellite, *Genome Res*, 16:88-96, 2006.

[3] J. V. Lorenzo-Ginori, A. Rodriguez-Fuentes, R. G. Abalo, R. S. Rodrigues, "Digital signal processing in the analysis of genomic sequences," *Curr. Bioinformatics*, vol. 4, pp. 28-40, 2009.

[4] P. G. Pop, A. Voina, "Numerical Representations Involved in DNA Repeats Detection Using Spectral Analysis," *Studies in Informatics and Control*, vol. 20, no. 2, pp. 163-180, 2011.

[5] D. Yankov, E. Keogh, S. Lonardi, "Dot plots for time series analysis," *Proc. 17th IEEE Int. Conf. on Tools with Artificial Intelligence*, pp. 159-168, 2005.

[6] C. Sung-Hyuk, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *Int. J. of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 2, pp. 300-307, 2007.

[7] T. D. Schneider, "Consensus Sequence Zen," *Applied Bioinformatics*, vol. 1, no. 3, pp: 111-119, 2002.

[8] N. Hamadani, "Automatic target cueing in IR imagery", *Master Thesis, Air Force Institute of Technology*, -WPAFB. Ohio.

[9] C. Alkan, M. Ventura, N. Archidiacono, M. Rocchi, S. C. Sahinalp, E. E. Eichler, "Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data," *PLoS Comput Biol.*, vol. 3, no. 9, pp. 1807-1818, Sep. 2007.

[10] V. Paar, N. Pavin, I. Basar, M. Rosandic, M. Gluncic, N. Paar, Hierarchical structure of cascade of primary and secondary periodicities in Fourier power spectrum of alphoid higher order repeats, *BMC Bioinformatics*, vol. 9, no. 1, 466, 2008.