Spectral Representations of Alpha Satellite DNA

PETRE G. POP
Comm. Dept.
Technical University of Cluj-Napoca
G. Baritiu Str, 26-28, 400027
ROMANIA
petre.pop@com.utcluj.ro

Abstract: - Detection of tandem repeats can be used for phylogenic studies and disease diagnosis. The numerical representation of genomic signals is very important, as many of the methods for detecting repeated sequences are part of the DSP field. These methods involve the application of a kind of transformation. Applying a transform technique requires mapping the symbolic domain into the numeric domain in such a way that no additional structure is placed on the symbolic sequence beyond that inherent to it. Here we investigate the application of spectral analysis and spectrograms using a novel numerical representation to identify and study alpha satellite higher order repeats in human chromosomes 7 and 17.

Key-Words: - Sequence Repeats, DNA Representations, Alpha Satellite DNA, Spectral Analysis, Spectrograms.

1 Introduction

The presence of repeated sequences is a fundamental feature of genomes. From the genome explorer viewpoint, repeat is the simplest form of regularity and analyzing repeats gives first clues to discovering new biological phenomena. Repeats are two or more contiguous, approximate copies of a pattern of nucleotides. Duplication occurs because of mutational events in which an original segment of DNA, the pattern, is converted into a sequence of individual copies. Repeats, whose copies are distant in the genome, whether or not located on the same chromosome, are called distant repeats, while the repeats whose copies are adjacent on a chromosome are called tandem repeats (TR). Among those, biologists distinguish micro-satellites, mini-satellites, and satellites, according to the length of their repeated unit. Local repeats in the DNA arise, grow or disappear through molecular events that copy a contiguous segment on the DNA and insert one or many copies of it next to the original segment, or perform the dual operation. The repeated copies also change through point mutations: insertion, deletion or substitution of one base. Point mutations give rise to approximate tandem repeats (ATR) [1] and complicates the process of repeats detection and identification.

The interest in detecting tandem repeats can be summarized 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 (e.g. identification of dead corpse, in paternity testing).
- Medical interest: the appearance of specific kinds of tandem repeats has been linked to a number of 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 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 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 (megabase quantities at each centromere) and made up of ~171 bp 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].

Almost all DSP techniques require two parts: mapping the symbolic data to a numeric form in a

nonarbitrary manner and calculating a kind of transform of that numeric sequence. Therefore, the numerical representation of genomic signals is very important.

This paper presents results obtained by combining grey level spectrograms with a novel numerical representation to isolate position and length of DNA repeats for high-order and monomeric human alpha satellites in human chromosomes 7 and 17.

2 Assignment of Numerical Values

Biomolecular sequences are represented by character strings, in which each element is one out of a finite number of possible "letters" of an "alphabet." In the case of DNA, the alphabet has size 4 and consists of the letters A, T, C and G. corresponding to DNA nucleotides. Applying a transform technique requires mapping the symbolic domain into the numeric domain in such a way that no additional structure is placed on the symbolic sequence beyond that inherent to it.

One common representation is to map nucleotides to a set of indicator sequences. Consider a sequence (a_k), k=0,..,N-1 from the alphabet A_4={A, C, G, T}. For each different letter α in A we form an indicator sequence $x_{\alpha,k}$, k=0,..,N-1 such that:

$$x_{\alpha,k} = \begin{cases} 1, & if\ a_k = \alpha \\ 0, & otherwise \end{cases},\ \alpha \in \{A,T,G,C\} \quad (1)$$

And is obvious that:

$$\sum_j x_{j,k} = 1,\ for\ all\ k \quad (2)$$

This approach produces a four-dimensional representation yielding an efficient representation for spectral analysis.

One simple representation is to use numbers assigned to each nucleotide, such as A=0, G=1, C=2, T=3 and modulo operations, but this implies relations on nucleotides such that T > A and C > G.

Another representation use geometrical notations taken from telecommunication QPSK constellation: A=1+j, T=1-j, G=-1+j, C=-1-j [2]. This representation was useful for nucleotide quantization to amino acids and in autocorrelation analysis.

A representation which preserve DNA's reverse complementary properties [9] use discrete numerical sequence symmetric about y-axis, inspired from pulse amplitude modulation, in which A=-1.5, G=-0.5, C=0.5, T=1.5.

For statistical approaches using Markov models, a four Galois field assignment was used in which A=0, C=1, T=2, G=3 [8].

Another representation is based on the concept of categorical element applied to DNA sequences. This measures the existence of pairs of identical elements at a distance of k pairs in a DNA sequence [3].

All these representations have advantages for particular analyses but suggest some DNA properties beyond that inherent to them [11]. Starting from these representations, we introduced a novel representation to reduce the dimensionality of representation and generate only one numerical sequence for each DNA sequence.

3 DNA Spectral Analysis

Spectral analysis may be performed by taking the Discrete Fourier Transform (DFT) of each of the indicator sequences [4] [5]. Applying DFT definition to all indicator sequences, for alphabet A_4, we obtain another sequences $X_A[k]$, $X_C[k]$, $X_G[k]$, $X_T[k]$:

$$X_\alpha[k] = \sum_{n=0}^{N-1} (u_\alpha[n] - m_\alpha) e^{-j\frac{2\pi}{N}kn},\ k = 0,1,...,N-1 \quad (3)$$

Where:

$$m_\alpha = \frac{1}{N}\sum_{n=0}^{N-1} u_\alpha[n],\quad \alpha \in \{A,T,G,C\} \quad (4)$$

Subtracting of the mean of each indicator sequence allows avoiding interference from the dc component of the Fourier spectrum.

From (3) and (4) it follows that:

$$X_A[k] + X_C[k] + X_G[k] + X_T[k] = \begin{cases} 0, & k \neq 0 \\ N, & k = N \end{cases} \quad (5)$$

The sequences $X_A[k]$, $X_C[k]$, $X_G[k]$, $X_T[k]$ can be used to provide the total spectrum $T[k]$ of the DNA sequence.

One possibility is to compute a sum spectrum [4] [5]:

$$S[k] = |X_A[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 + |X_T[k]|^2 \quad (6)$$

Another possibility is to compute a product spectrum [6]:

$$P[k]=\prod_{\alpha\in\{A,T,G,C\}}(|X_\alpha[k]|+c),k=0,1,\ldots,N-1 \quad (7)$$

Where c is a small positive constant used to avoid the product spectrum cancellation when a nucleotide is, absent from the analysis window.
In most cases *T[k]* has a peak at the sample value $k=N/3$ (Fig. 1), as demonstrated in many papers [7].

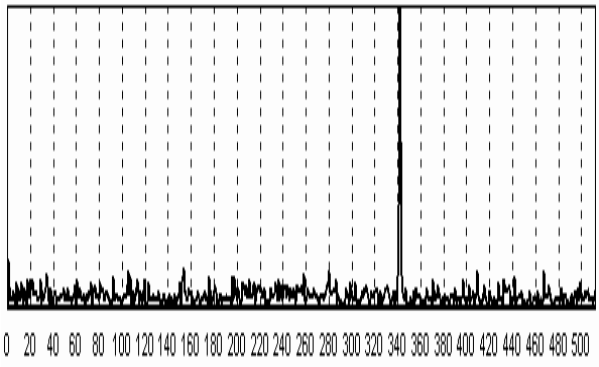

Fig. 1. T[k] showing a strong period-3 property.

This is often called a period-3 property of the DNA sequences and has often been attributed to the dominance of the base G at certain codon positions in the coding regions. This is the reason why the period-3 property was regarded to be a good (preliminary) indicator of gene location [7]. The periodic behavior indicates strong short-term correlation in the coding regions, in addition to the long-range correlation or 1/f-like behavior exhibited by DNA sequences in general.

If a period p repeat exists in the DNA sequence, T[k] should show a peak at frequencies *f =1/p, 2/p, 3/p*… (Fig. 2).

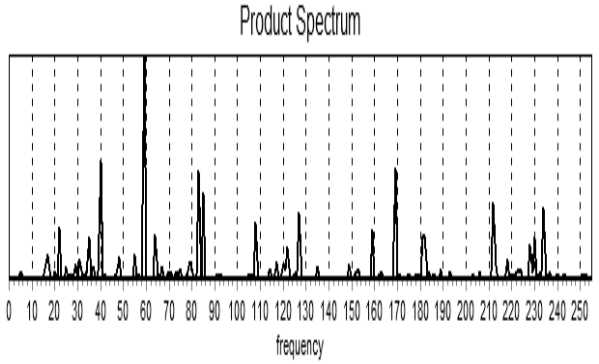

Fig. 2. P[k] showing peaks at different periods

The period *p* can thus be inferred from the peak location but the window length (N) limits the period. However, not all peaks are significant. A threshold T_h can be used to find peak candidates such that $T[k]/T_m>T_h$, where T_m is the frame spectral product average [6]. Now, the candidate peaks can be isolated and the length of TR, $N_i = 1/f_i$ can be estimated.

However, doing this on a frame-by-frame basis is difficult. A technique for detection of the beginning and end of the TRs regions is needed. Once we have detected a local TR and identified its fundamental period, we need to identify what subsequence in our window corresponds to the local TR. Instead, T[k] can be used to represent DNA sequence spectra in grey level spectrograms. In this case, TRs appear as horizontal lines (more or less continue) and frequency value indicates TRs length (Fig. 3). Horizontal positions of TRs indicate starting positions of windows for which DTF is calculated. This is approximate information about the location of repeats in original sequence.

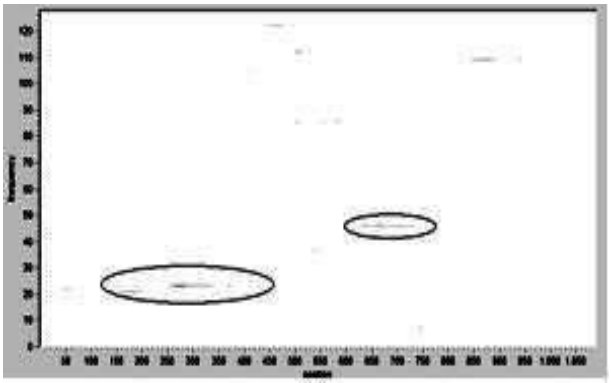
Fig. 3. Product spectrum grey level spectrogram

Spectrogram offers a global view of product spectrum but is difficult to estimate the exact location of TRs even if horizontal axis contains nucleotide position. This can be done calculating and representing the values of *T[fi]* in a sliding window along the sequence.

4 Reducing the Dimensionality

DNA Spectral analysis would be simpler if we could use only one numerical sequence instead of four indicator sequences. One way to do this was proposed in [4] [5] [14] as quartic mapping. In this case, the numerical sequence is given by:

$$x[n]=ax_A[n]+tx_T[n]+cx_C[n]+gx_G[n],\ n=1,2,\ldots,N-1 \quad (8)$$

Where a, t, c, and g are numerical values assigned to the nucleotides A, T, C, and G, respectively.

The numbers used in (8) can be:

- Consecutive integer values based on the nucleotides frequencies in the original sequence. For example, this values can be *a=4, t=3, c=2, g=1* if the nucleotides frequencies are in this order.
- Electron-ion interaction pseudopotential values for nucleotides (*a=0.1260, g=0.0806, c=0.1340, t=0.1335*) [8].

The new sequence, *x[n]* is then used to compute power spectrum.

In order to increase DNA spectral analysis accuracy for repeats detection, we propose a sequence representation [10], which takes into account the length of the expected repeats and the number of possible mismatches because of point mutations.
For a DNA sequence of length *L* a numerical value is associated in polynomial-like representation:

$$V = \sum_{k=0}^{L-1} V_{\alpha_k} 10^{k}, \quad \alpha \in \{A, G, C, T\} \quad (9)$$

Where V_α is the value of a single nucleotide as follows: *A=1, G=2, C=3, T=4*. Another possibility is to use consecutive natural numbers (1, 2, 3, and 4) based on nucleotides frequencies in the original DNA sequence.

The following input values are needed:

- A DNA sequence of length *N*;
- The length of expected repeated sequence, *L*;
- The maximum number of mismatches in the repeated sequences, M_m.

In passing from DNA sequence to numerical values, Hamming distance and consensus value are needed:

- Hamming distance measure the number of mismatches between sequences; if two sequences are identical the Hamming distance is zero;
- Given a number of sequences of same length, the consensus sequence is a sequence formed by the most popular nucleotide in the same positions.

The algorithm is summarized bellow:

- Consider all successive subsequences of length *L* in the initial DNA sequence;
- Determine all the positions (and the associated subsequences of length *L*) in original sequence for which the Hamming distance is less or equal the prefixed mismatches number;
- Determine the consensus sequence for all subsequences starting at these positions;
- Compute the value for consensus sequence and assign this value to all these positions.

As output, the algorithm generates a single vector *SeqVal* of (N-L) numerical values, each value being associated to a subsequence of length *L*. We also need a vector *Dist [N]* to store the distances for a sequence of length *L*, starting on a given position, to all other subsequences of same length *L*, starting on all possible positions.
The algorithm can be improved if the Hamming distance and the consensus sequences are evaluated only in forward direction (from the current position) and exclude first *L* subsequences starting from current position (for which is no sense to evaluate the distance).
Here is the pseudocode description of the algorithm:

```
foreach curr_pos in (0,..., N - L)
{
   foreach calc_pos in (curr_pos + L,...,N - L)
   {
    Dist[calc_pos]=GetDist(curr_pos, calc_pos, L);
       if (dist > Mm)
          Dist [calc_pos] = 0;
   }
   consensus = GetConsensus (Dist, L);
   val = GetVal (consesnsus, L);
   foreach calc_pos in (0,...,N-L)
   {
      if (Dist [calc_pos] != 0)
         SeqVal [calc_pos] = val;
   }
}
```

This algorithm has the advantage of simplicity. In addition, no additional structures or special memory requirements are needed. The main limitation is related to a priori information about repeat length and maximum number of mismatches but in many situations, biologists know this information in advance.

5 Alpha Satellite DNA Analysis

We used spectral analysis and grey level spectrograms to investigate periodicities for high-order and monomeric human alpha satellite DNA.

Our case study was the 16mer high order repeat in AC017075 from human chromosome 7 and the high-order repeat in AC136363 from human chromosome 17 (GenBank). In case of AC017075 high-order repeats were identified in the central domain (positions 31338 to 177434, total length 148147bp) while in the front domain of genomic sequence (31337 bp) and in the back domain (15843 bp), alpha satellite monomers were found [2] [14]. The AC136363 clone contain dispersed alphoid sequences, both higher-order and monomeric alpha-satellite [2].
DNA power spectrum was computed using algorithm based on (9) to obtain DNA numerical sequences, using different values for expected repeat length (L) and maximum number of mismatches (M_m).

Next figures (Fig.4 … Fig.14) show power spectrum grey-level spectrograms for AC017075 sequence from human chromosome 7 (GenBank).

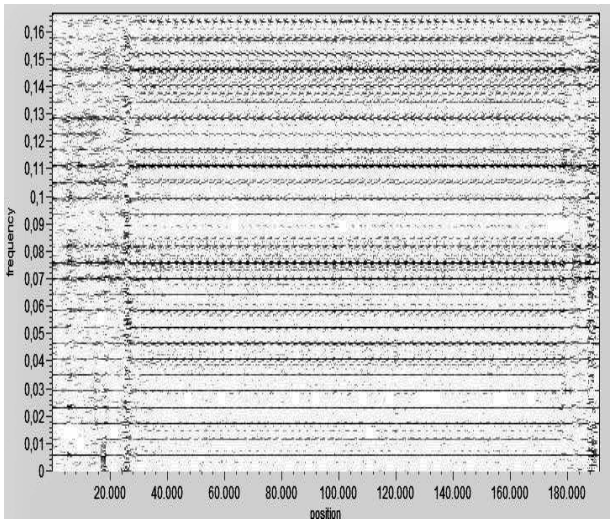

Fig. 4. Spectrogram for AC017075 using L=3, M_m=1.

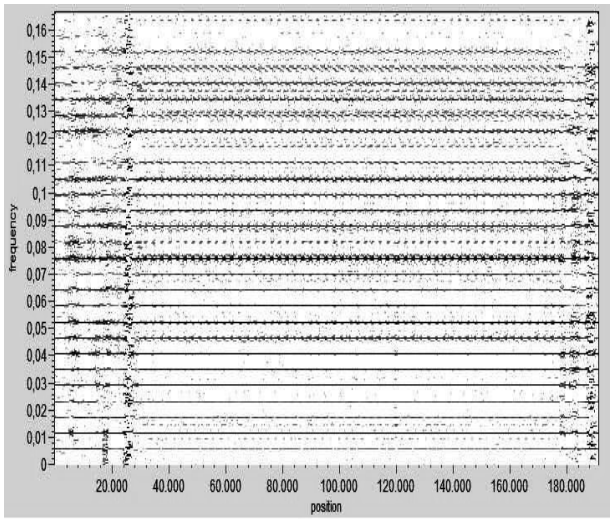

Fig. 5. Spectrogram for AC017075 using L=3, M_m=2

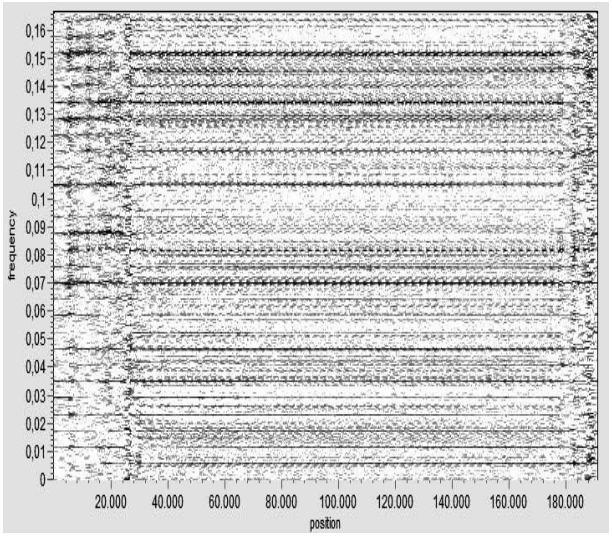

Fig. 6. Spectrogram for AC017075 using L=9, M_m=3.

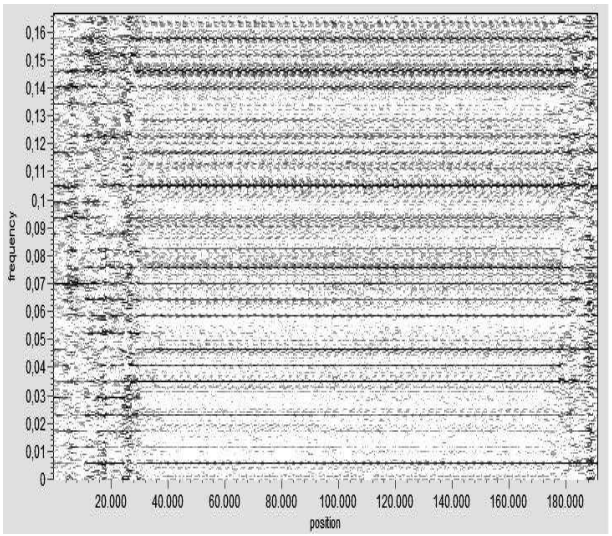

Fig. 7. Spectrogram for AC017075 using L=9, M_m=4.

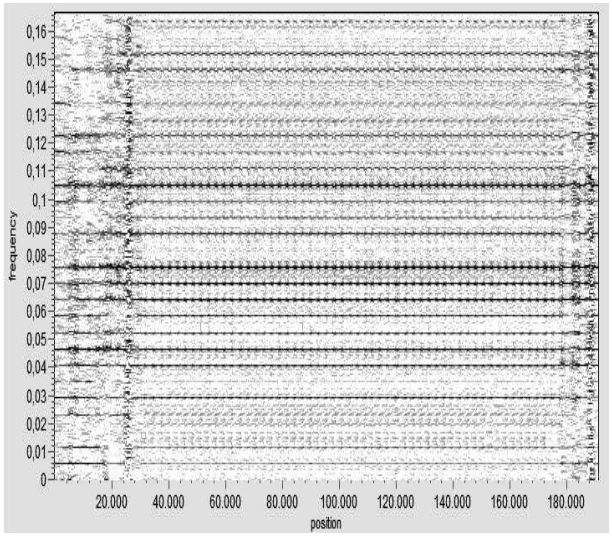

Fig. 8. Spectrogram for AC017075 using L=9, M_m=5

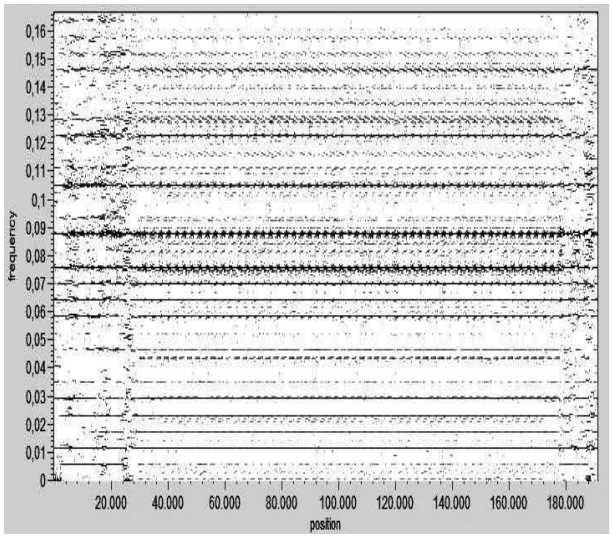

Fig. 9. Spectrogram for AC017075 using L=9, M_m=6

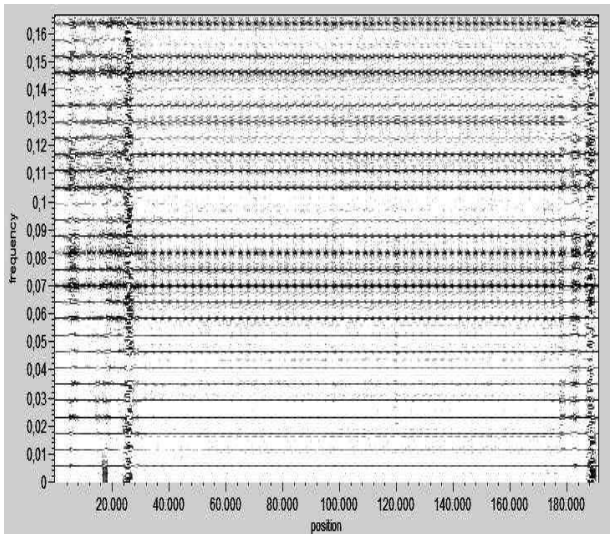

Fig. 10. Spectrogram for AC017075 using L=19, M_m=4.

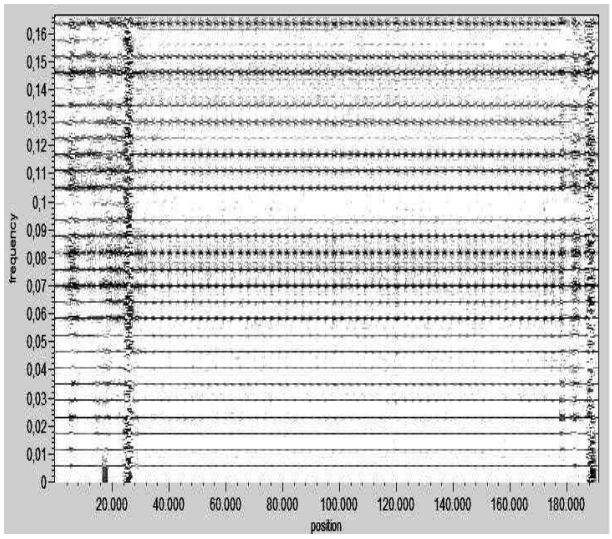

Fig. 11. Spectrogram for AC017075 using L=19, M_m=5.

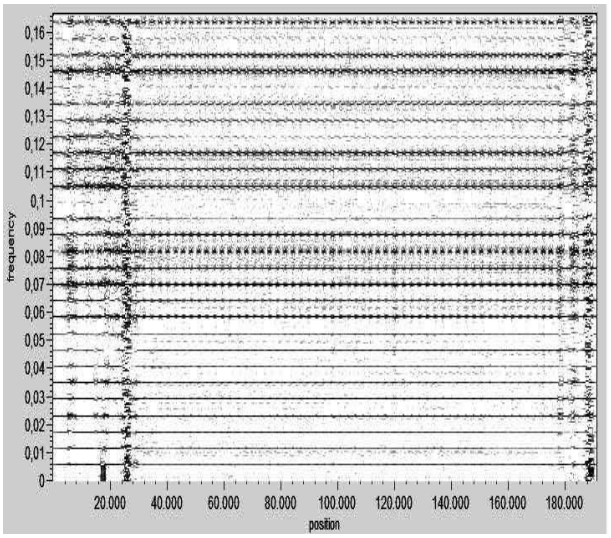

Fig. 12. Spectrogram for AC017075 using L=19, M_m=6.

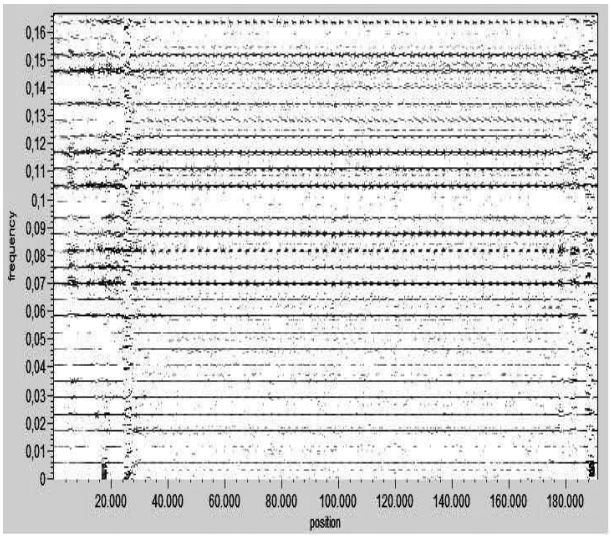

Fig. 13. Spectrogram for AC017075 using L=19, M_m=7.

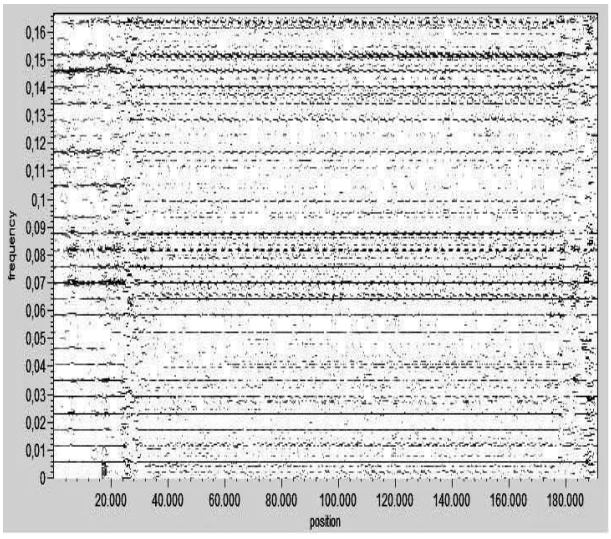

Fig. 14. Spectrogram for AC017075 using L=19, M_m=8.

Analyzing the above figures, we can formulate the following conclusions:

- All figures allows to isolate the area of a high-order repeat alpha satellite (27000bp-176000bp) and areas with monomeric alpha satellite in the front domain and back domain of genomic sequence;
- Repeats length (~171bp) is shown by the first horizontal line at a frequency *f=0.00585*, but this line is best shown in Fig. 4 (*L=3*), Fig. 7 (*L=9*) and Fig. 10,11,12 (L=19);
- Repetition number (16) is given by the number of equidistant lines starting from f=0.00585 but appears clearly only in Fig. 10 and 12 (*L=19*);
- It is sufficient to use divisors of repeat length (171: 3, 9, 19) for *L* values; this allows a significant reduction in the number of searches;
- Due to the large number of repeats contained in sequence good results are obtained even for small values of *L* parameter (*L=3* and Fig. 4);
- Mm values affect the quality of the results. Too small or too large values lead to the deterioration of results. The best results were obtained for values of 30-40% of *L* value. The values of this parameter are chosen on biological criteria.

Next figures (Fig.15…Fig.25) shows power spectrum grey-level spectrograms for AC136363 sequence from human chromosome 17 (GenBank), using different values for expected repeat length (*L*) and maximum number of mismatches (M_m).

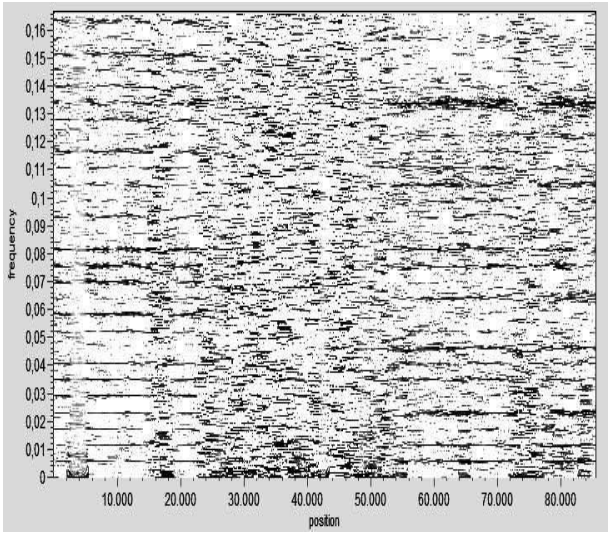

Fig. 15. Spectrogram for AC136363 using L=3, M_m=1.

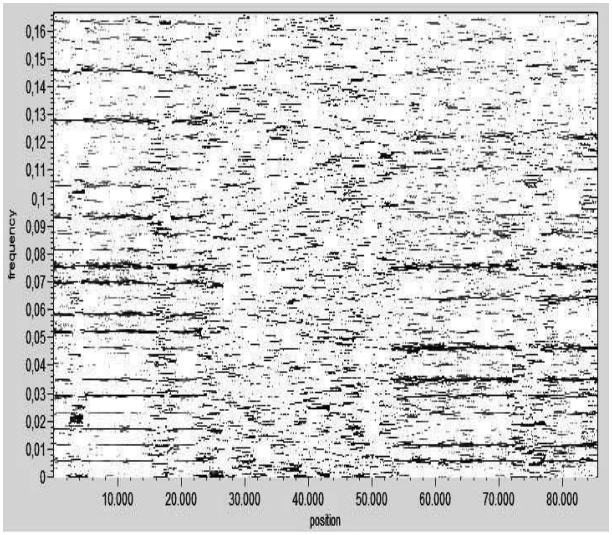

Fig. 16. Spectrogram for AC136363 using L=3, M_m=1.

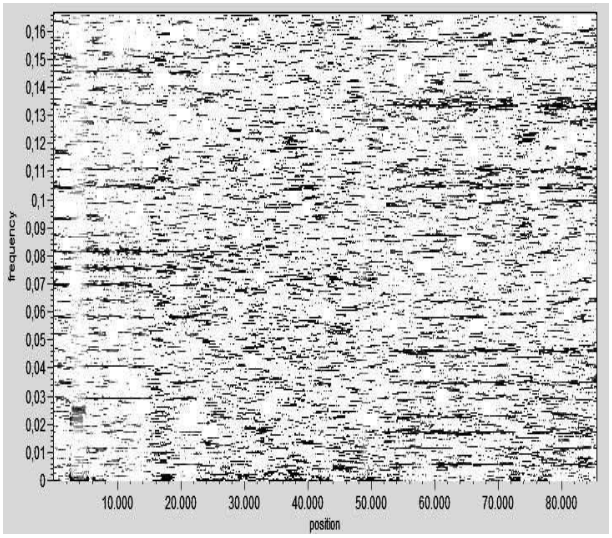

Fig. 17. Spectrogram for AC136363 using L=9, M_m=3.

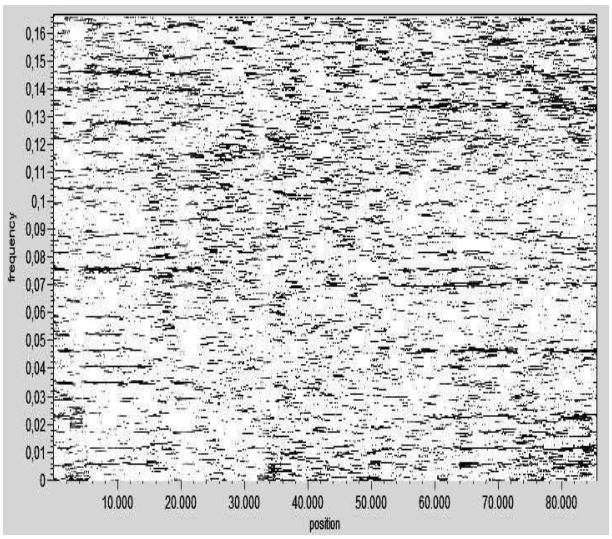

Fig. 18. Spectrogram for AC136363 using L=9, M_m=4.

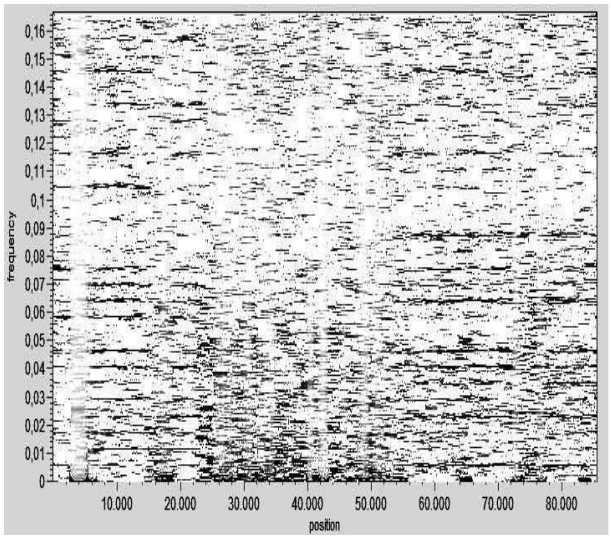

Fig. 19. Spectrogram for AC136363 using L=9, M_m=5.

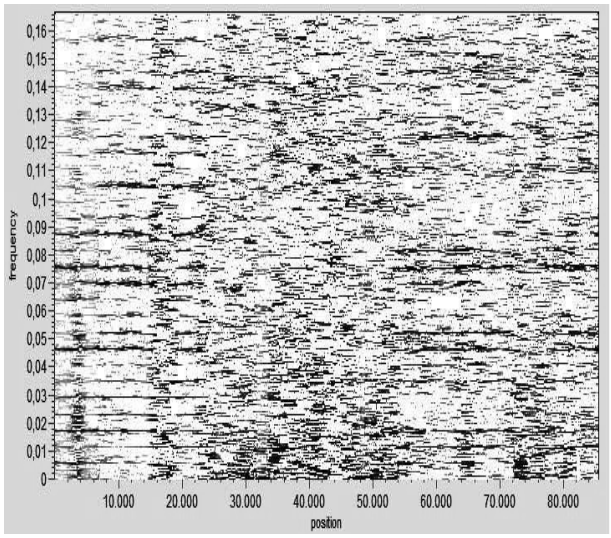

Fig. 20. Spectrogram for AC136363 using L=9, M_m=6

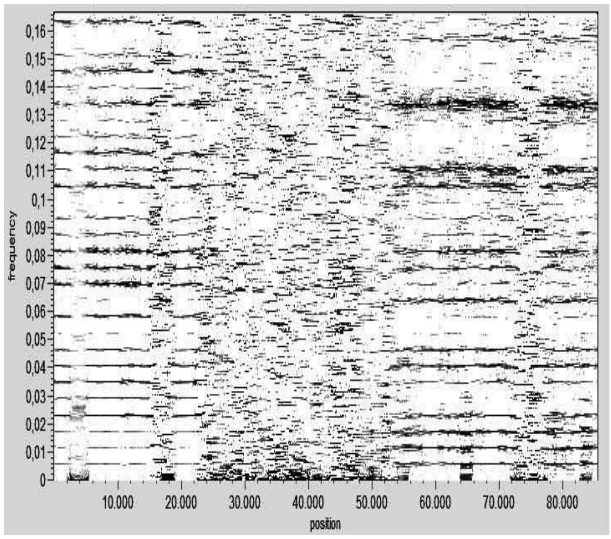

Fig. 21. Spectrogram for AC136363 using L=19, M_m=4.

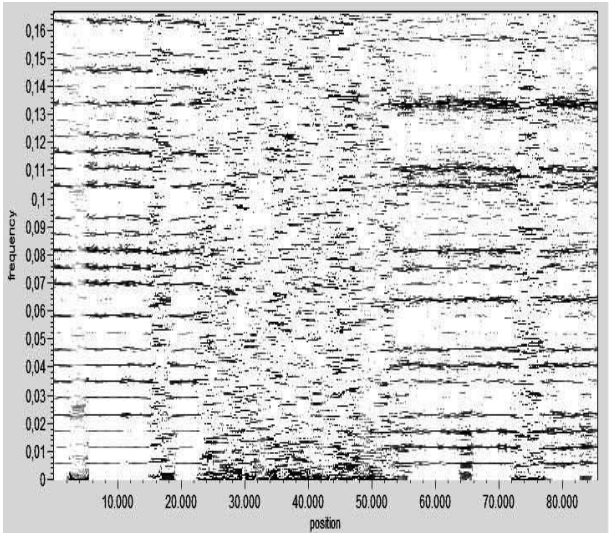

Fig. 22. Spectrogram for AC136363 using L=19, M_m=5.

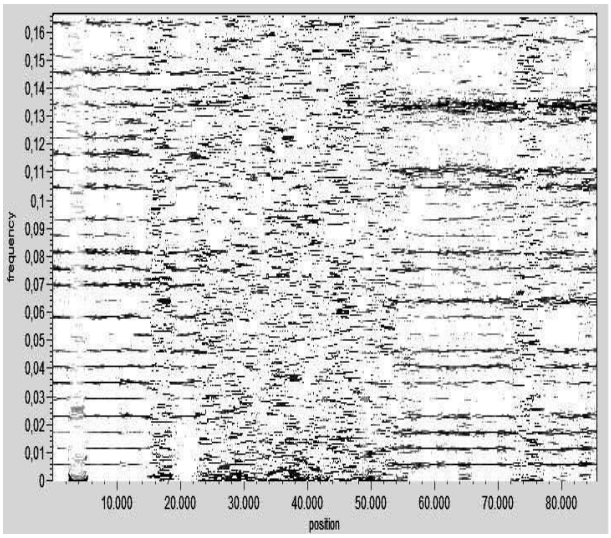

Fig. 23. Spectrogram for AC136363 using L=19, M_m=6.

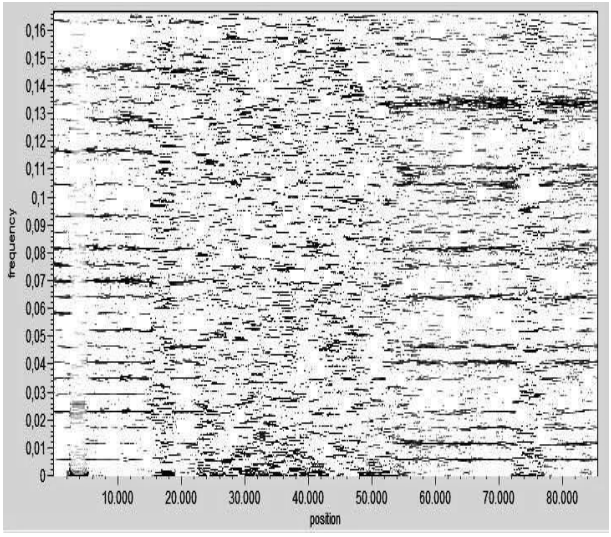

Fig. 24. Spectrogram for AC136363 using L=19, M_m=7.

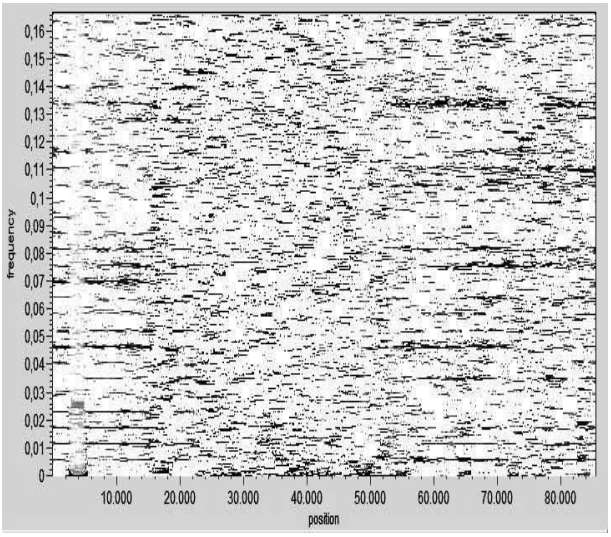

Fig. 25. Spectrogram for AC136363 using L=19, M_m=8

As one can see:

- Unlike the previous case, not all figures allows to easy isolate the areas of a high-order repeat and monomeric alpha satellite;
- Repeats length (~171bp) is shown by the first horizontal line at a frequency *f=0.00585*, but this line is best shown in Fig. 21…24 (*L=19*);
- Repetition number (16) is given by the number of equidistant lines starting from *f=0.00585* but appears clearly only in Fig. 16 (*L=3*), Fig. 21…25 (*L=19*);
- Due to the reduced number of repeats contained in sequence, small values for *L* allows only highlighting of zones with repeats. But large values for *L* allows determining repeat length and repeat number;
- Only value *L=19* (a divisor of 171) allows good results which imply increasing of calculation time.
- Again, the best results were obtained for Mm values of 25-40% of *L* value.

After the two experiments, we can formulate the following ideas:

- For small values of expected repeat length *L*, the algorithm allows to easily detect repeats and their associate information;
- For large values of expected repeat length *L*, the algorithm works well even with divisors of *L*; This allows reducing the number of attempts and computation time;
- Values for number of mismatches that allow obtaining good results are 25-40% from the value for calculation repeat length.

6 Conclusion

DFT and grey level spectrograms provides a robust detection method for higher order periodicity. Repeats are easily recognizable by regular horizontal lines, which give information about repeat length and number of repeats.

A polynomial-like representation of DNA sequences provides a single numerical sequence that can be used directly in spectral analysis and yields improved results. This method can be successfully applied to alpha satellite DNA detection and yields more accurate results.

Acknowledgement

This work has been partly supported by grant projects: PN2-IDEI-334/2007, PN2-PART-41082/2007.

References:

[1] A. Krishnan and F. Tang, Exhaustive Whole-Genome Tandem Repeats Search, *Bioinformatics Advance Access*, May 14, 2004.

[2] Rudd, MK, Wray GA, Willard HF, The evolutionary dynamics of a-satellite, *Genome Res* 16:88–96, 2006.

[3] Achuthsankar S.N., Mahalakshmi, T., Are categorical periodograms and indicator sequences of genomes spectrally equivalent ?, *In silico Biology*, 6, 0019, 2006.

[4] D. Anastassiou, Genomic signal processing, *IEEE Signal Process. Mag*., 18 (4) (2001) 8–20.

[5] Vera Afreixo, Paulo J.S.G. Fereira, Dorabella Santos, Fourier analysis of symbolic data: A brief review, *Digital Signal Processing*, 14(2004), pp. 523-530.

[6] V.A. Emanuele II, T.T. Tran, G.T. Zhou, A Fourier Product Method For Detecting Approximate Tandem Repeats In DNA, IEEE *Workshop on Statistical Signal Processing*, Bordeaux, July 17-20, 2005.

[7] H. Herzel, O. Weiss, and E.N. Trifonov, 10-11 bp periodicities in complete genomes reflect protein structure and protein folding, *Bioinformatics*, vol. 15, pp. 187-193, 1999.

[8] Achuthsankar S.N, Sivarama P.S, A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), *Bioinformation*, 1(6), 197-202 (2006).

[9] Chakravarthy, K. and et al., Autoregressive modeling and feature analysis of dna sequences, *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 13–28, 2004.

[10] Pop, G.P., Lupu, E., DNA Repeats Detection Using BW Spectrograms, *IEEE-TTTC Int. Conf. on Automation, Quality and Testing, Robotics, AQTR 2008*, May 22-25, 2008, Romania, Tome III, pp. 408-412

[11] Cristea P.D., Nucleic Acid Structural Properties Identified by Genomic Signal Analysis, *9th WSEAS Int. Conf. on Mathematics & Computers In Biology & Chemistry, (MCBC '08)*, Bucharest, Romania, June 24-26 (pp 182-187).

[12] Girish Rao, David K.Y. Chiu, Comparison of Genomes As 2-Level Pattern Analysis, *Proceedings of the 2006 WSEAS International Conference on Mathematical Biology and Ecology,* Miami, Florida, USA, January 18-20, 2006 (pp117-122).

[13] Kun-Lin Hsieh, Cheng-Chang Jeng, I-Ching Yang, Yan-Kwang Chen, Chun-Nan Lin, The Study of Bioinformatics Based on Codon Usage in DNA Sequence, *Proceedings of the 6th WSEAS Int. Conf. on Systems Theory & Scientific Computation*, Elounda, Greece, August 21-23, 2006 (pp33-38).

[14] Paar V, Pavin N, Basar I, Rosandic M, Gluncic M, Paar N, Hierarchical structure of cascade of primary and secondary periodicities in Fourier power spectrum of alphoid higher order repeats, *BMC Bioinformatics*, 2008 Nov 3; 9(1):466.