Numerical Representations Involved in DNA Repeats Detection Using Spectral Analysis

Petre G. Pop[1], Alin Voina[2]

Technical University of Cluj-Napoca, Comm. Dept.,
26-28, G. Baritiu, Cluj-Napoca, 400027, Romania,
petre.pop@com.utcluj.ro, alin.voina@com.utcluj.ro

Abstract: Sequence repeats are the simplest form of regularity and the detection of repeats is important in biology and medicine as it can be used for phylogenic studies and disease diagnosis. A major difficulty in identification of repeats is caused by the fact that the repeat units can be of unknown length and either exact or imperfect, in tandem or dispersed. Many of the methods for detecting repeated sequences are part of the digital signal processing (DSP) field. These methods involve a transformation which has as main goal the mapping of the symbolic domain into the numeric domain without adding structure information to the symbolic sequence beyond that inherent to it. Therefore, the numerical representation of genomic signals is very important. This paper presents the results obtained by using different numerical representations (including two novel) and spectral analysis to isolate the position and length of DNA repeats in short sequences containing microsatellites and on long sequences with alpha DNA repeats.

Keywords: genomic signal processing, sequence repeats, DNA representations, Fourier analysis, spectrograms.

1. Introduction

Over the past few decades, major progress in the field of molecular biology, combined with the advances in genomic technologies, have led to a fulminating growth in the biological information generated by scientists. There are databases which contains hundreds of billions of bases and sequence records. Therefore, computers have become an indispensable tool for biological research as they provide the means for storing large quantities of data and revealing the relationships between them.

A surprising genetic difference among species is the size of their genomes. Relatively simple organisms may have much larger genomes than complex organisms. These major differences might be due to the presence of repeats. In general, for eukaryotes duplicated genetic material is abundant and can represent up to 60% of the genome. Although some of the mechanisms that generate these repeats are known, from an evolutionary point of view, the reasons for such redundancy remains unknown [1]. The presence of repeated sequences is a fundamental feature of all genomes.

A repeat is the simplest form of regularity and analyzing repeats can lead to first clues to discovering new biological phenomena. Tandem repeats are two or more contiguous, approximate copies of a pattern of nucleotides. Tandem duplication occurs as a result of mutational events in which an original segment of DNA (the pattern) is converted into a sequence of individual copies.

The centromere of most complex eukaryotic chromosomes is a specialized locus made up of repetitive DNA which is responsible for chromosome segregation at mitosis and meiosis.

A major challenge in genomic signal processing is to understand the information contained in the biological genomes. Almost all DSP techniques require two parts: mapping the symbolic data (symbols for nucleotides) into a numeric form in a non-arbitrary manner and calculating a kind of transform of that numeric sequence. Consequently, the numerical representation of genomic signals becomes very important.

Fourier spectral analysis is used to reveal periodicity in symbolic sequences because they are rather robust in the presence of substitutions, insertions and deletions and may identify approximate periodicities in DNA sequences.

This paper presents results obtained using different numerical representations (including two new) and spectral analysis to isolate the position and length of DNA repeats in short sequences containing microsatellites and on long sequences with alpha DNA repeats.

Most of the numerical representations used for repeats detection associate a 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. In order to include information about the number of consecutive

nucleotides and to generate only one numerical sequence for each DNA subsequence which may be associated with a repeat [9, 10], we've introduced two novel representations. Therefore, to emphasize subsequences with consecutive repeats of the same nucleotide, we used a modified form of indicator sequences which includes the repeating factor. Then, we proposed a novel sequence representation and a mapping algorithm which takes into account the length of the expected repeats and the number of possible mismatches due to point mutations, based on polynomial-like representation.

Grey-levels spectrograms were used to validate numerical representations because they provide an overview of the informational content of the analyzed sequence and allow a fast and easy determination of the presence of repeated sequences. In addition, spectrograms do not need to specify the length, the pattern or the number of mismatches for target repeats. Thus, the spectrogram can be used for a qualitative assessment of numerical representation. The main focus was on numerical representations and on qualitative differences that occur in spectrograms and not on spectral analysis itself. Our goal was not the comparison of the different ways for identifying the repeated sequences but the comparison of the different numeric representations using one of the frequent used methods.

Interests in DNA Repeats

Nucleotide sequences contain patterns or motifs that have been preserved throughout evolution because of their importance to the structure or function of the DNA molecule. Nucleotide sequences outside the coding regions generally tend to be less conserved among organisms, except where they have a functional importance, like the involvement in gene expression regulation. Motifs discovery in protein and nucleotide sequences can lead to determination of function and to the elucidation of the evolutionary relationships among sequences.

The interest in detecting tandem repeats can be summarized as follows [2]:

- Theoretical interest: regarding their role in the structure and evolution of the genome.
- Technical interest: they 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 appearance of specific tandem repeats has been linked to a number of different severe diseases (e.g. Huntington's disease). In healthy individuals, the repeat size varies around a few tens of copies, while in affected individuals the number of copies at the same locus reaches at least hundreds.

Definitions

Nucleotide and protein sequences are represented by character strings, in which each element is one out of a finite number of possible symbols of an "alphabet." In the case of DNA sequences, the alphabet has four symbols and consists of the letters A, T, C and G, corresponding to Adenine, Thymine, Cytosine and Guanine nucleotides.

A perfect (exact) repeat is a string that can be represented as a smaller string repeated contiguously twice or more. For example, ACACAC is a repeat, as it can be represented as string AC repeated three times. The length of the repeated pattern is called the period (2 for the case of ACACAC), and the number of pattern copies is called the exponent (3 for ACACAC). If the exponent is 2 or more, the repeat is usually called a tandem repeat (TR). Repeats, whose copies are distant in the genome, whether or not located on the same chromosome, are called distant/dispersed repeats. Among those, biologists distinguish micro-satellites, mini-satellites, and satellites, according to the length of their repeated unit.

However, perfect tandem repeats are of limited biological interest, since different biological events will often render the copies imperfect [3]. The result is an approximate tandem repeat (ATR), defined as a string of nucleotides repeated consecutively at least twice with small differences between the instances. The role of ATRs discovered by using some of the algorithmic approaches is limited by constraints on the input data, search parameters, the type of allowed mutations and the number of such mutations. In other ATRs, time requirements render the algorithm

infeasible for the analysis of whole genomes containing millions of base pairs (bp).

The centromere of most complex eukaryotic chromosomes is a specialized locus comprised of repetitive DNA that is responsible for chromosome segregation during 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 [4]. 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 [4].

2. Methods

Applying a transform technique requires mapping the symbolic domain into the numeric domain such 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 in order to indicate the presence or absence of a nucleotide in a certain position [5]. 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 as:

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

And it 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 which preserve the DNA's reverse complementary properties [6], such as:

$$A=0,\quad G=1,\quad C=2,\quad T=3 \quad (3)$$

Or

$$A=1,\quad G=2,\quad C=3,\quad T=4 \quad (4)$$

Another representation uses geometrical notations taken from the telecommunication QPSK constellation [7]:

$$A=1+j,\ T=1\text{-}j,\ G=-1+j,\ C=-1\text{-}j \quad (5)$$

This representation was useful for nucleotide quantization to amino acids and in autocorrelation analysis.

A different representation - inspired from pulse amplitude modulation (PAM) - which preserves the DNA's reverse complementary properties [8] uses discrete numerical values which are symmetric about y-axis:

$$A=-1.5,\ G=-0.5,\ C=0.5,\ T=1.5 \quad (6)$$

All previous representations are punctual ones: they associate a 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.

Starting from these representations we introduced two novel representations to include information about the number of consecutive nucleotides and to reduce the dimensionality of representation by obtaining only one numerical sequence for each DNA subsequence which may be associated with a repeat [9, 10].

Often, the pattern of the repeats contains repeated subsequences of the same nucleotide. For example, 11mer repeats from Table 1, shows subsequences of repeating nucleotides like CC, TTT, GGG. In order to emphasize these subsequences we used a modified form of indicator sequences.

First, the indicator sequences are modified to include the repeating factor *m* as the number of consecutive positions with the same values in the sequence [9]:

$$u_\alpha[k]=\begin{cases}m, & if\ \ a_k=\alpha\\ 0, & otherwise\end{cases} \quad (7)$$

Consider the nucleotide sequence: TGACTTTGGGG. The modified indicator sequences which include the repeating factors are:

$$
\begin{aligned}
u_A[n] &= 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\\
u_C[n] &= 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\\
u_G[n] &= 0\ 1\ 0\ 0\ 0\ 0\ 0\ 4\ 4\ 4\ 4\\
u_T[n] &= 1\ 0\ 0\ 0\ 3\ 3\ 3\ 0\ 0\ 0\ 0
\end{aligned} \tag{8}
$$

Second, the expected repeated factors in the repeat sequence, for each nucleotide, are included in the indicator sequences by limiting the initial repeat factor to the expected repeat factor in the repeat sequence. Assuming the next expected repeating factors for each nucleotide: r_A=1, r_C=2, r_G=3, r_T=2 then the final indicator sequences become:

$$
\begin{aligned}
u_A[n] &= 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\\
u_C[n] &= 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\\
u_G[n] &= 0\ 1\ 0\ 0\ 0\ 0\ 0\ 3\ 3\ 3\ 3\\
u_T[n] &= 1\ 0\ 0\ 0\ 2\ 2\ 2\ 0\ 0\ 0\ 0
\end{aligned} \tag{9}
$$

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, 7, 11] as quartic mapping. In this case, the numerical sequence is given by:

$$
\begin{aligned}
&x[k] = a\,x_A[k] + t\,x_T[k] + c\,x_C[k] + g\,x_G[k],\\
&k = 1,2,\ldots,N-1
\end{aligned} \tag{10}
$$

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

The coefficients used in (10) can be:

- Consecutive integer values based on the nucleotides appearing frequencies in the original sequence. For example, these values can be a=4, t=3, c=2, g=1 if the nucleotides frequencies are in this order.
- Consecutive integer values like in (3) and (4).
- Numerical values from (6).
- Electron-ion interaction pseudo-potential values for nucleotides [12]:

$$
\begin{aligned}
&a = 0.1260,\quad g = 0.0806,\\
&c = 0.1340,\quad t = 0.1335
\end{aligned} \tag{11}
$$

In order to increase DNA spectral analysis accuracy for repeats detection, we proposed a sequence representation and a mapping algorithm [9, 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 a polynomial-like representation:

$$
V = \sum_{k=0}^{L-1} V_{\alpha_k} 10^{k}, \quad \alpha \in \{A, G, C, T\} \tag{12}
$$

where V_α is the value of a single nucleotide.

One possibility is to use consecutive natural numbers, preserving DNA's reverse complementary properties, such as A=1, G=2, C=3, T=4. Another possibility is to use consecutive natural numbers (e.g. 1, 2, 3, and 4) based on nucleotides frequencies in the original DNA sequence.

But for two very similar sequences (which differ, for instance, by a single nucleotide) will get two very different numbers. So it takes an algorithm that allows finding similar sequences and then generates single numerical values for these sequences.

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.

To pass from DNA sequence to numerical values, Hamming distance and consensus value are needed:

- Hamming distance measures the number of mismatches between sequences of the same length [13]; if two sequences are identical then the Hamming distance is zero;
- Given a number of sequences of the same length, the consensus sequence is a sequence formed by the most frequent nucleotide in the same positions.

The mapping algorithm is summarized bellow:

- Step 1: Consider all successive subsequences of length L in the initial DNA sequence;
- Step 2: Determine all the positions (and the associated subsequences of length L) in the original sequence for which the Hamming distance (against a sequence from Step1) is less or equal to the prefixed mismatches number M_m;

- Step 3: Determine the consensus sequence for all subsequences from Step2, starting at these positions;
- Step 4: Compute the numerical value for consensus sequence (using 12) and assign this value to all these positions.

As output, the algorithm generates a single vector, *SeqVal*, of (N-L) numerical values; each value it is associated to a unique subsequence of length L (possible a repeat unit). 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.

This mapping algorithm has the following properties:

- Even for an L value smaller than the actual length of repeated sequence, the final numerical sequence will highlight a repeat.
- 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.

The algorithm can be improved if the Hamming distance and the consensus sequences are evaluated only in the forward direction (from the current position) and exclude first L subsequences starting from current position (these ones makes 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] =
GetDistance(curr_pos,calc_pos, L);
     if (Dist[calc_pos] > Mm)
       Dist [calc_pos] = -1;
  }
  consensus = GetConsensus (Dist, L);
  val = GetValue (consesnsus, L);
  foreach calc_pos in (curr_pos + L,...,N-L)
  {
    if (Dist [calc_pos] >= 0)
     SeqVal [calc_pos] = val;
  }
}
```
(13)

Function *GetDistance()* computes the Hamming distance between the subsequence starting at current position and other subsequence (of the same length) which has different position (in forward direction). Function *GetConsensus()* determines the consensus sequence for all subsequences of length L, whose positions are stored in the *Dist[]* vector; function *GetValue()* computes the associated numerical value for the consensus sequence, using (12).

The algorithm can be further improved if we consider:

- Using negative values to reflect available positions in *Dist[]* vector;
- Storing positions of similar subsequences in another vector (*PozSimilar[]*);
- Replacing already calculated distances only if the newer distance is lower than the existing distances.

Below is the pseudocode description of the mapping algorithm which includes the improvements described above:

```
foreach curr_pos in (0,..., N - L)
{
  Dist[curr_pos] = -1;
}
foreach curr_pos in (0,..., N - L)
{
  if (seqVal[curr_pos] >= 0)
   continue;
  cnt = 0;
  foreach calc_pos in (curr_pos + L,...,N - L)
  {
    dd = GetDistance(curr_pos, calc_pos, L);
    if (dd < Mm)
    {
      if (Dist[calc_pos] < 0) OR (dd <
        Dist[calc_pos]))
      {
        Dist[calc_pos] = dd;
        PozSimilar[cnt] = calc_pos;
        cnt++;
      }
    }
  }
  if (cnt > 0)
  {
    consensus = GetConsensus (Dist, L);
    val = GetVal (consesnsus, L);
    foreach i in (0,...,cnt-1)
    {
      SeqVal [PozSimilar[i]] = val;
    }
  }
  SeqVal[curr_pos] = val;
}
```
(14)

To analyze the algorithm's complexity we have considered major operations categories of the algorithm:

- Calculating distance between two subsequences:
  - In the worst case, no similar subsequences at all, and the complexity is $O(n^2)$; each distance implies L comparisons between two characters;
  - Once you have determined several similar subsequences and a numerical value was associated, the corresponding positions are no longer used in the outer iteration. Increasing the number of similar subsequences decreases the number of distance evaluations. In this way the number of such operations will decrease.
- Determining the consensus sequence for similar subsequences founded at a time:
  - In the worst case, one evaluation for each current position, the complexity is O(n);
  - If there are repeated sequences, the number of calls decreases but is applicable to a growing number of subsequences; if there are no repeated sequences, the number of calls is higher but applies to a small number of subsequences;
  - Again, once you have determined several similar subsequences and a numerical value was associated, the corresponding positions are no longer used in the outer iteration, hence the number of such operations will decrease.
  - Determination of consensus sequence among n subsequences of length L involves counting types of nucleotides in each position and then, at least three comparisons between these numbers.
- Calculating the numerical value associated to the consensus sequence:
  - The first considerations in the previous case apply also here;
  - Determining the numerical value associated to a subsequence of length L involves calculating the value of a polynomial function of degree L.

M_m parameter influences the algorithm's performance. Thus, a high value increases the probability of finding similar subsequences when low values significantly reduce the number of similar subsequences.

Our implementation uses a threshold to prevent the determination of consensus sequence and the associated numerical value in case of a small number of similar subsequences.

The proposed algorithm has the advantage of simplicity. Generates only one numerical sequence containing embedded information about the repeated sequences searched for (length and number of mismatches) which can be exploited later. Also, no additional structures or special memory requirements are needed and if the length of the repeated sequence admits divisors, computing effort can be reduced substantially. The main limitation is related to increased complexity ($O(n^2)$ related to distance evaluation) and a priori information about the length of the repeat and the maximum number of mismatches (in most of the situations, biologists know this information in advance.

3. Results

Case study

Our intention was to study numerical representations on short sequences with a small number of short repeats and on large sequences with alpha satellites. Well known and analyzed sequences were used to validate the new numeric representations, without introducing new information regarding repeated sequences from these nucleotide sequences. Our case study was the human microsatellite sequence M65145 (GenBank) and the 16mer high order repeat in AC017075 sequence from human chromosome 7 (GenBank). Table 1 lists the repeats values and positions from M65145 microsatellite [14].

Table 1. - 11mer Repeats in the Microsatellite M65145 [14]

Position	Sequence
131–141	T G A C C T T T G G G
157–167	T G A C C T T G G G G
256–266	T G A C T T T A G G G
300–310	T T T C T T T G G G G
322–332	T G A C T T T G G G G
346–356	T G A T T T T G A G G
411–421	T G A C T T T G A A G
458–468	T G A C T C T G G G G
634–644	T G G C T T G G G G G
738–748	T G T C T C T G G G G
Consensus sequence	T G A C T T T G G G G

In case of AC017075 high-order repeats (highly homogeneous, organized in arrays of multimeric repeat units) were identified in the central domain (positions 31338 to 177434, total length 148147 bp) while in the front domain of genomic sequence (31337 bp) and in the back domain (15843 bp), alpha satellite monomers (which exhibit substantial mutual sequence divergence) were found [4, 11].

In the next sections, DNA power spectrum was computed and then represented as spectrograms using different numerical representations in order to detect repeats in micro-satellite M65145 and in alpha satellite DNA AC017075. The spectrograms were obtained using a custom application developed in Delphi.

DNA Spectral Analysis Using Indicator Sequences

Spectral analysis may be performed by taking the Discrete Fourier Transform (DFT) of each of the indicator sequences [6, 7]. Applying DFT definition to all indicator sequences, for alphabet A_4, we obtain other sequences $X_A[k]$, $X_C[k]$, $X_G[k]$, and XT [k]:

$$X_\alpha[k]=\sum_{n=0}^{N-1}(x_\alpha[n]-m_\alpha)e^{-j\frac{2\pi}{N}kn},k=0,1,...,N-1 \quad (15)$$

Where:

$$m_\alpha=\frac{1}{N}\sum_{n=0}^{N-1}x_\alpha[n],\quad \alpha\in\{A,T,G,C\} \quad (16)$$

Subtracting of the mean of each indicator sequence is used to avoid interference from the dc component of the Fourier spectrum.

The sequences $X_A[k]$, $X_C[k]$, $X_G[k]$, and XT [k] provide the total spectrum of the DNA sequence [8, 14, 15, 16]:

$$S[k]=|X_A[k]|^2+|X_C[k]|^2+|X_G[k]|^2+|X_T[k]|^2 \quad (17)$$

In most cases S[k] has a peak at the sample value k=N/3, as shown in other papers [18, 19]. This is the so called 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 period-3 component seems to appear because of the codon structure involved in the translation of base sequences into amino acids. For eukaryotes (cells with nucleus) this periodicity has mostly been observed within the exons (coding subregions inside the genes) and not within the introns (noncoding subregions in the genes). This is the reason why the period-3 property was regarded to be a good (preliminary) indicator of gene location [17, 18, 19]. The periodic behaviour indicates strong short-term correlation in the coding regions, in addition to the long-range correlation or 1/f-like behaviour exhibited by DNA sequences in general.

The spectrum peak at 1/3 indicates the presence of exons. The window length N should be large enough so that the periodicity effect dominates the background 1/f spectrum. However, a long window compromises the base-domain resolution in predicting the exon location.

These spectra can also be used to compute a Fourier product spectrum [20, 21] such as:

$$P[k]=\prod_{\alpha\in\{A,T,G,C\}}|X_\alpha[k]|,k=0,1,...,N-1 \quad (18)$$

where $X_\alpha[k]$ is the DFT of the mean removed indicator sequence.

Multiplication as a nonlinear operation is used to enhance peaks in a product spectrum. If a period p repeat exists in the DNA sequence, P[k] should show a peak at frequencies f =1/p, 2/p, 3/p,… The period p can thus be inferred from the peak location but the period is limited by the window length (N).

When a nucleotide is absent from a given (windowed) DNA sequence, one of the indicator sequences will be zero for all n. Thus, the product defined by (18) will be equal to zero. To avoid this, a modified product spectrum is defined, as:

$$P[k]=\prod_{\alpha\in\{A,T,G,C\}}(|X_\alpha[k]|+c),k=0,1,...,N-1 \qquad (19)$$

where c is a small positive constant.

The product spectrum of a genome sequence enables to detect the presence of repeats based on the spectral peaks. But not all peaks are significant. A threshold T can be used to find peak candidates such that $P[k]/P_m>T$, where P_m is the frame spectral product average [16]. Now, the candidate peaks can be isolated and the length of repeat, $N_i=1/f_i$ can be estimated.

represent DNA sequence spectra in another way, namely in grey level spectrograms. Colour and grey-levels spectrograms were used to analyze nucleotide sequences [7, 16] due to their capacity to provide a global view of categories of spectra for sequences of high length. At the same time they allow easy observation of patterns that appear in the spectra and do not require a priori information related to the repeated sequences (length, structure, mutations number or type). So, grey-levels spectrograms can be used to validate previous numerical representation because it provides an overview and allow rapid and easy determination of the presence of repeated sequence.

Figure 1 shows sum and product spectrum spectrograms using 256 DFT for microsatellite M65145 sequence (GenBank).

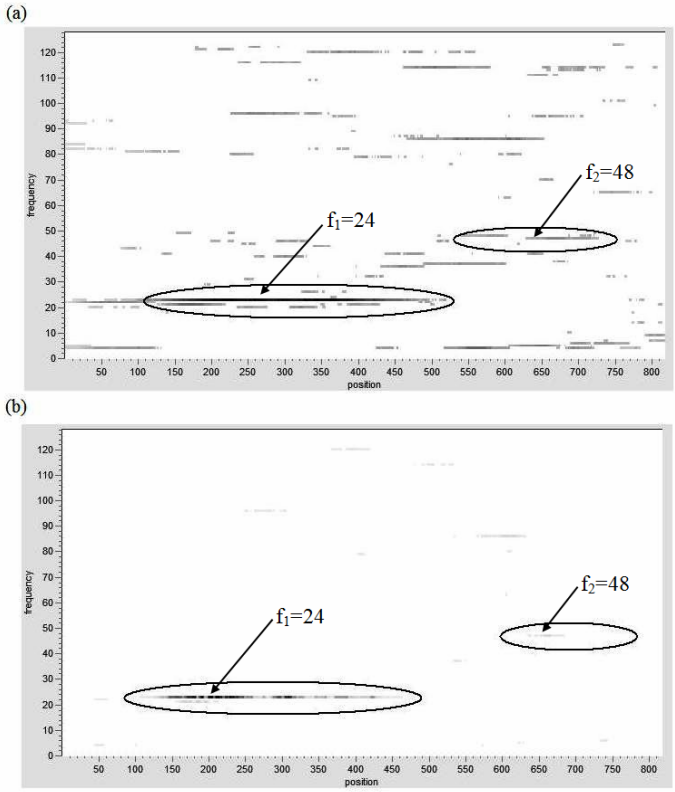

Figure 1. Spectrograms for M65145 microsatellite (256 DFT) a) Sum spectrum spectrogram. b) Product spectrum spectrogram.

However, doing this on a frame-by-frame analysis is difficult. A detection technique for the start and end position of the repeats regions is needed. Once we have detected a local repeat and we've identified its fundamental period, we need to identify what subsequence in our window corresponds to the local repeat. Instead, P[k] can be used to

Spectrograms were generated using a threshold value set to T=3.5 and a global normalization for image. In this way, only significant peaks from P[k] will be represented and it is easier to identify the presence of repeats and the associated length. In this case repeats appear as horizontal lines (more or less continue) at frequencies values

f_1=24, f_2=48 (frequency value indicates repeats length). Value f_1=24 correspond to a 11mer repeats (256 div 24) while the line at f_2=48 suggest that some 5mer repeats are part of 11mer repeats. Horizontal positions of repeats indicate starting positions of windows for which DTF is calculated. This is approximate information about the location of repeats in original sequence.

Comparing the two types of spectrum it follows that:

- Sum spectrum appears "dirty" with much additional information such as repeats are harder to detect.
- Product spectrum is clean, with line segments directly related to repeats, but repeats are quite attenuated compared to the sum spectrum.
- In case of product spectrum, repeats in the upper area (634-644, 738-748) are not well highlighted.

done by calculating and representing the values of $P[f_i]$ in a sliding window along the sequence [15, 16, 22].

Figure 2.a presents the product spectrum values $P[f_1]$ of the same sequence using threshold T=3.5 to eliminate weak peaks. In this case, is easy to identify the regions containing the repeats (11mer repeat) as those where peaks are significant.

Figure 2.b presents the product spectrum values $P[f_2]$ of the same sequence. In this case, the peak positions indicate that 11mer and 5mer repeats are located in the same region and some 5mer repeats may be part of 11mer repeats. All repeats from Table 1 can be found among maxima of $P[f_1=24]$ and some of them are also present among maxima of $P[f_2=48]$.

Since the length of the repeat ($1/f_i$) and the region containing the repeats are both completely specified, the actual repeats can be identified by a heuristic local alignment method.

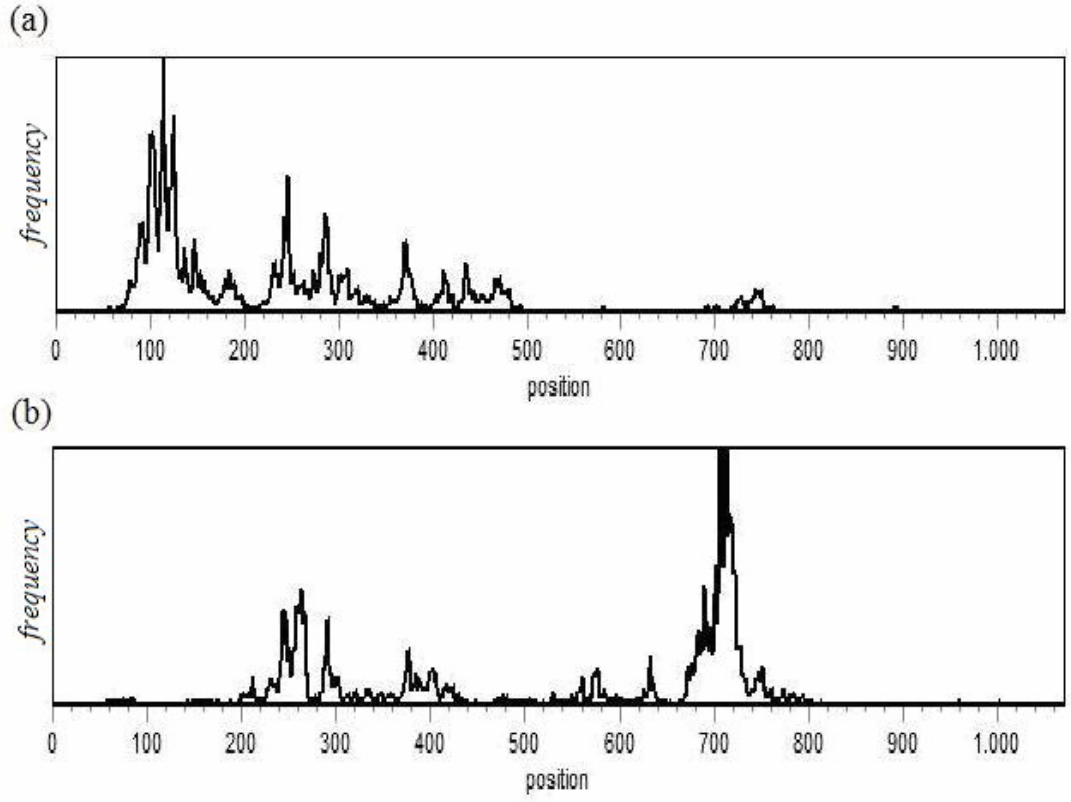

Figure 2. Product spectrum values in a sliding window along the DNA sequence (for M65145). a) $P[f_1]$ along DNA sequence; b) $P[f_2]$ along DNA sequence.

Spectrogram offers a global view of product spectrum but it is difficult to estimate the location of repeats even if horizontal axis contains nucleotide position. This can be

Next Figure (3) shows sum spectrum and product spectrum grey-level spectrograms for alpha satellite DNA AC017065 sequence (GenBank).

– Product spectrum does not reveal the number of repetitions but allows a better localization of areas with repetitions.

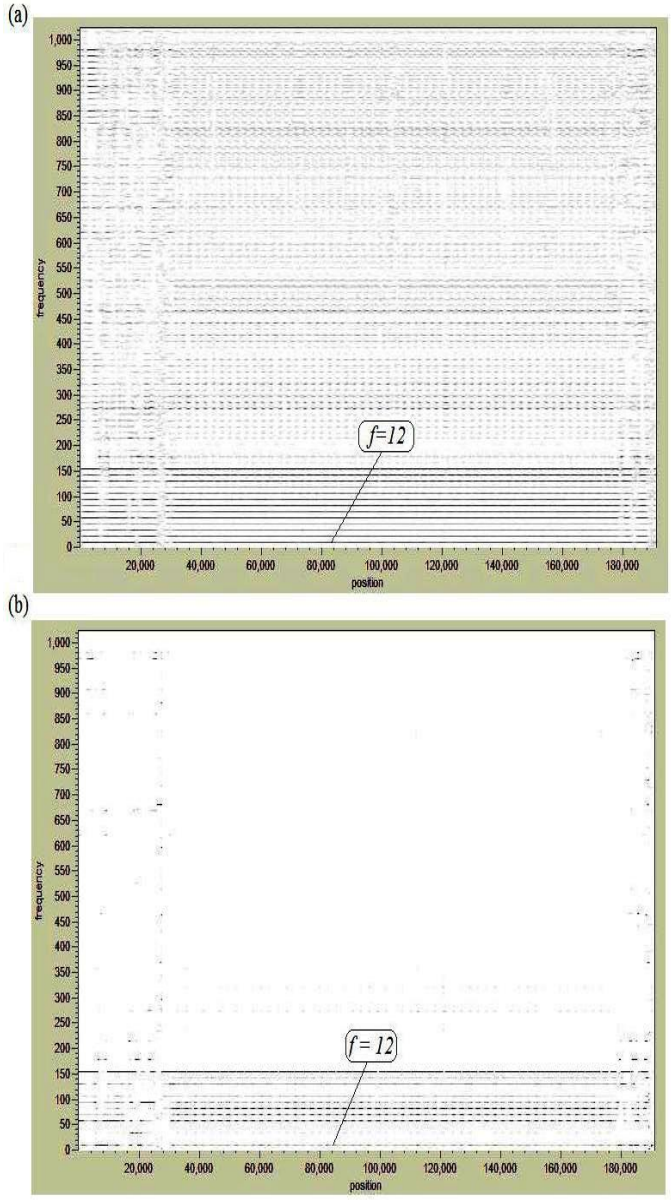

Figure 3. Grey-level spectrograms for AC017065 (2048 DFT) a) Sum spectrum spectrogram b) Product spectrum spectrogram

In this case:

– Sum spectrum allows a good estimate of the repeat length (≈171bp) by the position of first horizontal line, f≈12 (171 ≈ 2048 div 12) and number of repetitions (number of equidistant horizontal lines);

DNA Spectral Analysis Using Modified Indicator Sequences

In this case the new indicator sequences are calculated using (7) and with different repeating factors for each nucleotide. Then, the total spectrum is computed using (17) or (19) and represented in grey-level spectrograms.

Figure 4 shows sum spectrum grey-level spectrograms for microsatellite M65145 sequence (GenBank) using modified indicator sequences with different values for the expected nucleotide repeating factors r_G and r_T. As it can be seen, an increase of repeats factors r_G and r_T allows a decrease of residual information but repeats are still hard to locate (especially in the upper area).

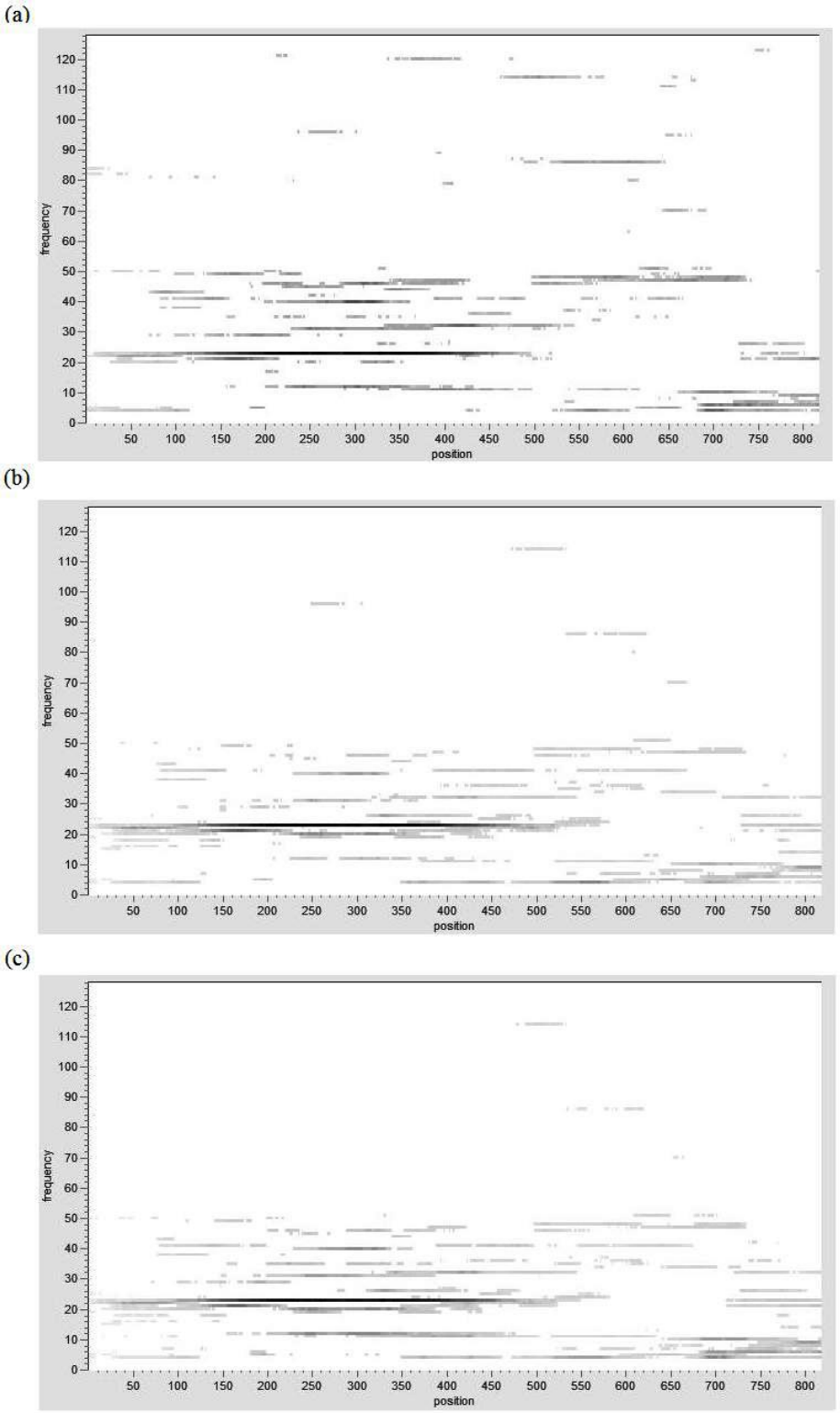

Figure 4. Sum spectrum spectrograms using modified indicator sequences (for M65145, 256 DFT). a) r_A=1, r_G=1, r_C=1, r_T=2; b) r_A=1, r_G=2, r_C=1, r_T=2; c) r_A=1, r_G=2, r_C=1, r_T=3.

Figure 5 shows product spectrum grey-level spectrograms for the same microsatellite M65145 sequence (GenBank) using modified indicator sequences with different values for expected nucleotide repeating factors r_G and r_T. As one can see, f_1 and f_2 frequencies are more highlighted as r_G and r_T repeating factors are increased. These values correspond to the repeating factors of nucleotides G and T in the consensus sequence from Table 1. On the other hand, the product spectrum values for other frequencies are diminished such that spectrogram zones associated with repeats can be more easily localized.

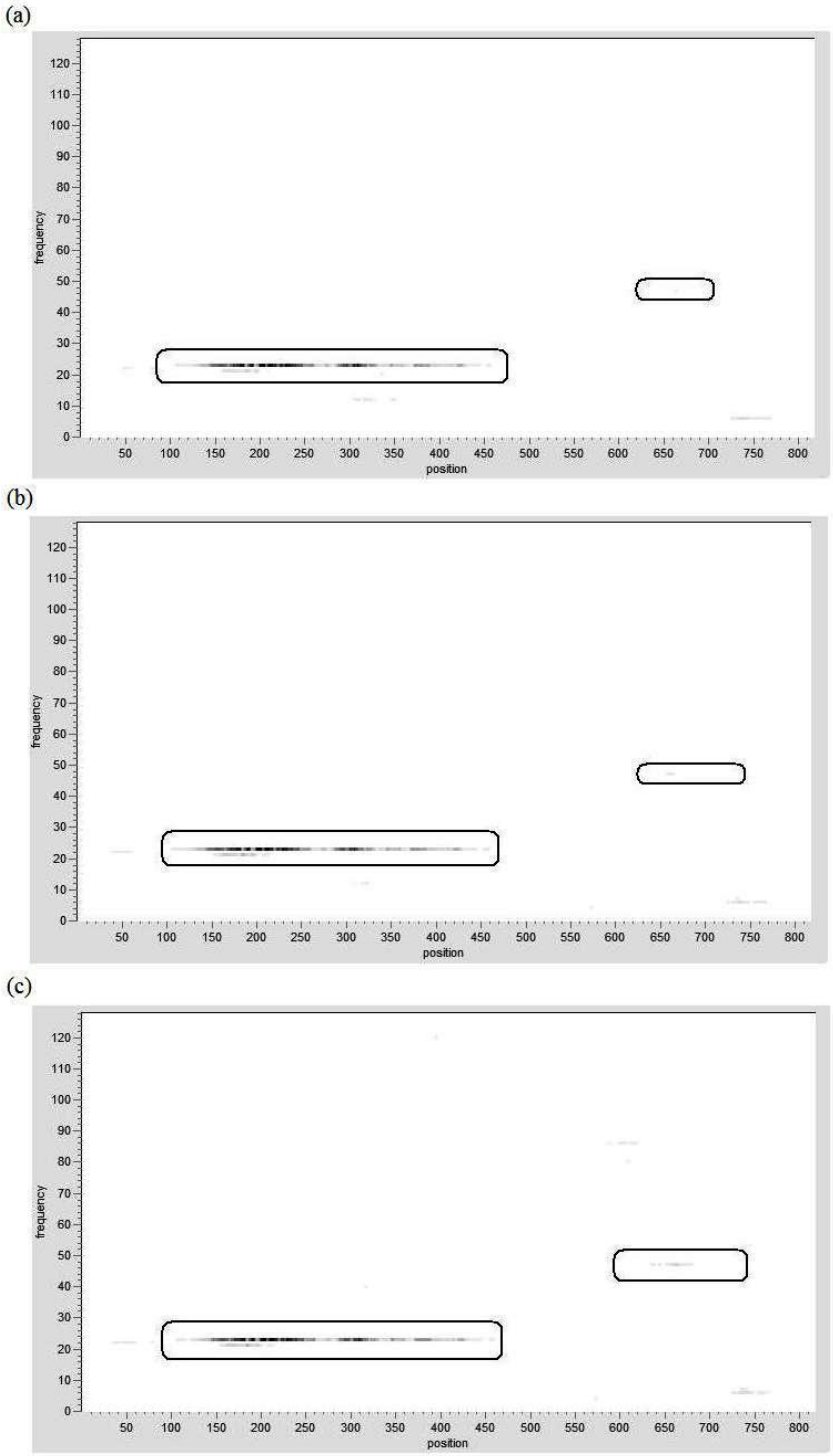

Figure 5. Product spectrum spectrograms using modified indicator sequences (for M65145, 2048 DFT). a) r_A=1, r_G=1, r_C=1, r_T=2; b) r_A=1, r_G=2, r_C=1, r_T=2; c) r_A=1, r_G=2, r_C=1, r_T=3.

DNA Spectral Analysis Using Quartic Mapping

The new sequence, x[n] is calculated using (10) and different values for a, c, g, t coefficients from (3), (4), (6) and (11). Finally x[n] is used to compute power spectrum which is represented using grey-level spectrograms.

Next figure (6) shows grey-level spectrograms for microsatellite M65145 sequence (GenBank) using different values for coefficients implied in (10).

In this case:

– Repeats of the lower zone are well highlighted by all variants;

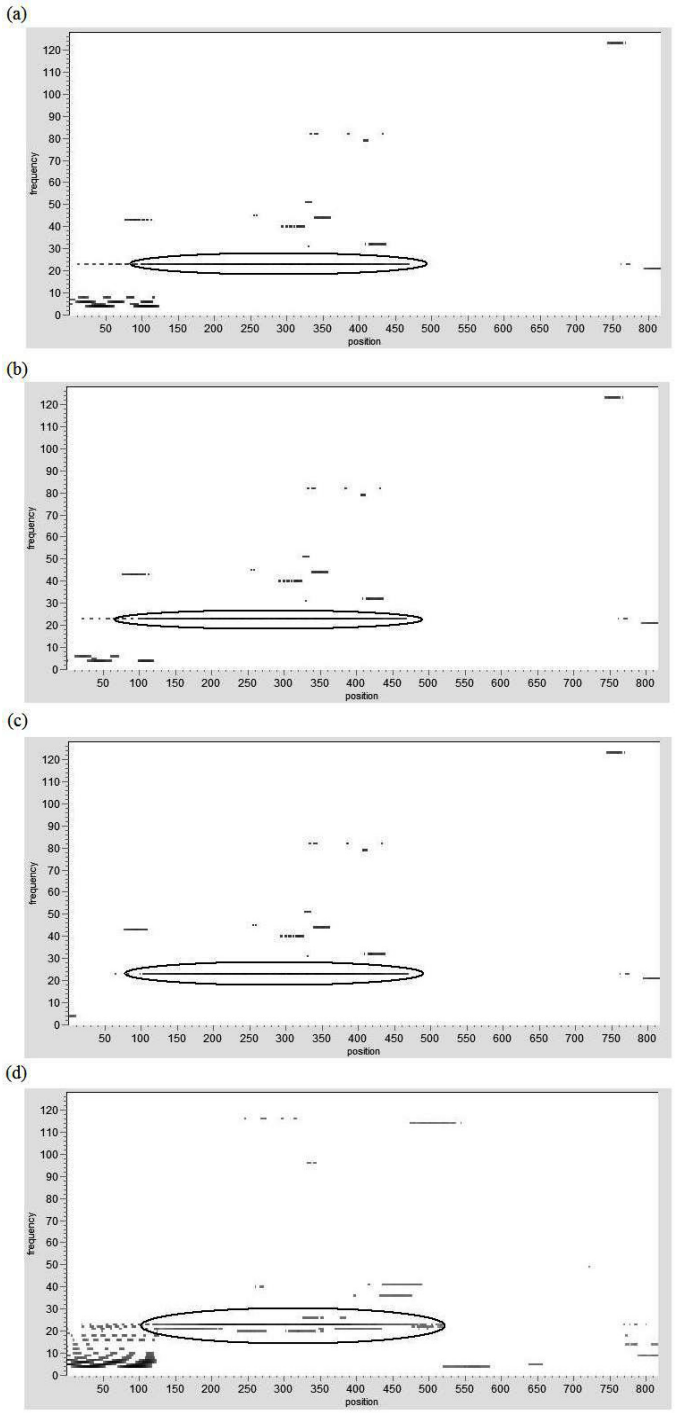

Figure 6. Spectrograms for M65145 using quartic mapping (256 DFT). a) a=1, g=2, c=3, t=4; b) a=0, g=1, c=2, t=3; c) using PAM inspired values; d) with pseudopotential values.

– Repeats of the upper zone are not highlighted by any variant;

– PAM inspired values variant gave the best results.

Figures 7 shows grey-level spectrograms for satellite AC017075 sequence (GenBank) using different values for coefficients implied in (10). Repeats length (≈171bp) is shown by the first horizontal line at a frequency f≈12 (171 ≈ 2048 div 12). Repetition number (16) should be given by the number of equidistant lines starting from f≈12.

As can be seen:

– All off the approaches used in this research allows length determination of the repeated sequences but the repetition number is not clearly evidenced by any alternative;

– Again, PAM inspired values variant gave the best results (Figure 7.c).

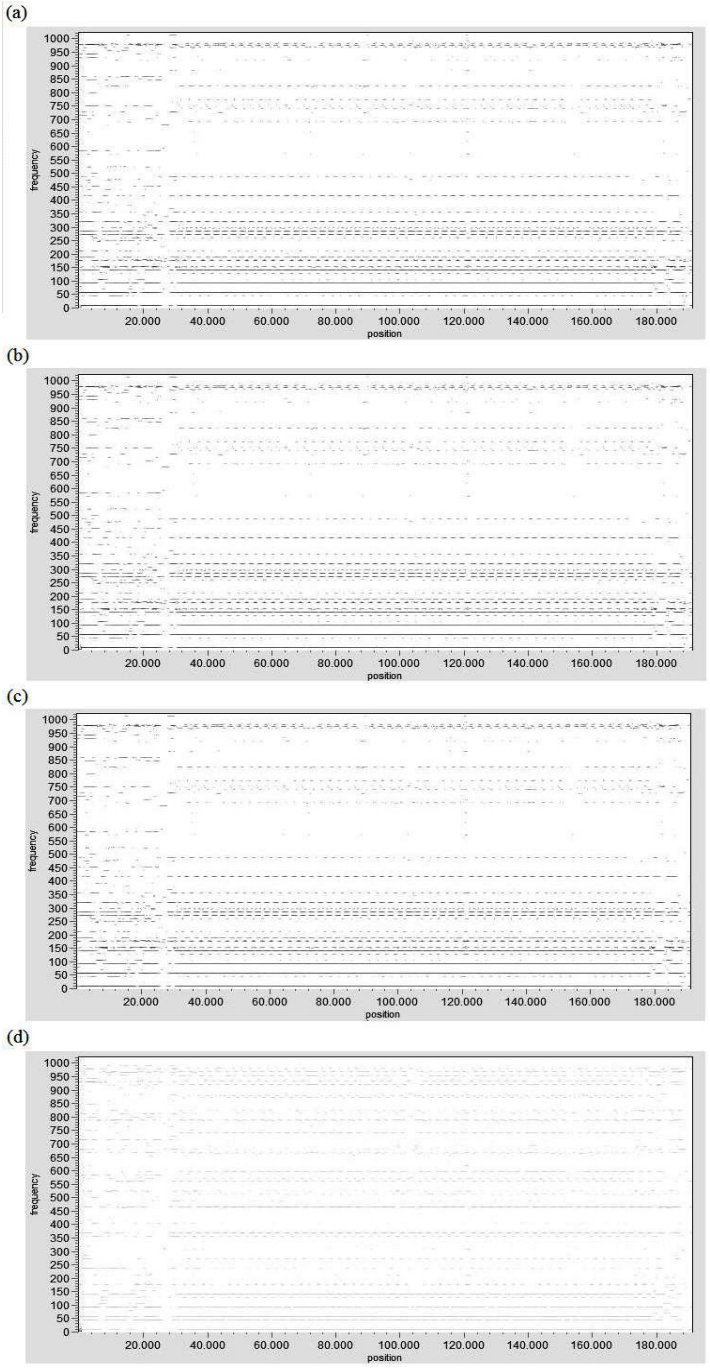

Figure 7. Spectrograms for AC017075 using quartic mapping (2048 DFT). a) a=0, g=1, c=2, t=3; b) a=1, g=2, c=3, t=4; c) using PAM inspired values; d) with pseudopotential values.

DNA Spectral Analysis Using Polynomial Representation

In this case, the new sequence, x[n] is calculated using (12) and algorithm (14) with different values for expected repeats length (L) and number of mismatches (M_m). Finally, x[n] is used to compute power spectrum which is represented using grey-level spectrograms.

Several experiments were conducted with different values for the parameters L and M_m. Below are the best results for certain values of parameter L.

Figure 8 shows grey-level spectrograms for microsatellite M65145 sequence (GenBank) using different values for L (repeat length) and those M_m values with best results (number of mismatches).

In this case:

- All M_m values allow a good highlighting of repeats in the lower zone and upper zone;
- Values M_m=2 and M_m=3 allow the direct evidence 11mers repeats in all zones.

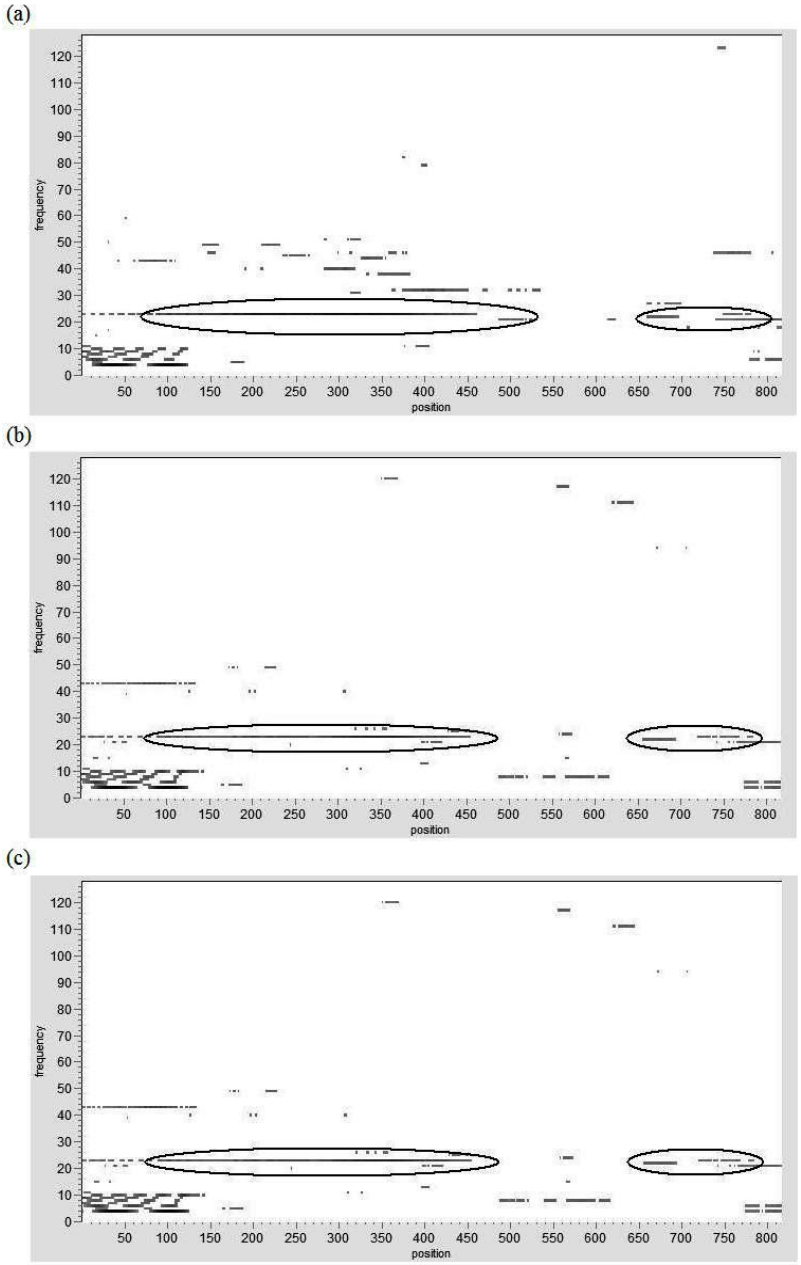

Figure 8. Spectrograms for M65145 using polynomial mapping (256 DFT). a) L=11, M_m=2; b) L=11, M_m=3; c) L=11, M_m=4.

Next Figure (9) shows grey-level spectrograms for satellite AC017075 sequence (GenBank) using different values for the same parameters L (repeat length) and M_m values with the best results (number of mismatches).

M_m values affect the quality of the results. The best results were obtained for values of 30-40% of L value.

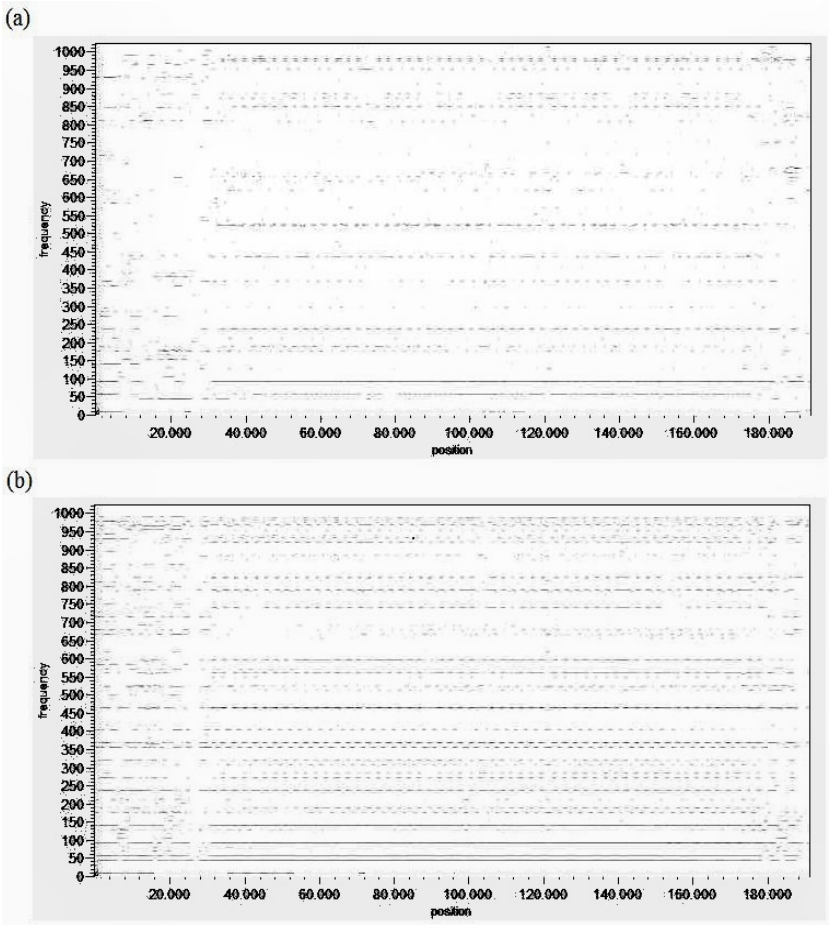

Figure 9. Spectrogram for AC017075 using polynomial mapping (2048 DFT).
a) L=9, M_m=4; b) L=19, M_m=6.

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

- All figures allow the isolation of the area of a high-order repeat alpha satellite (27000 bp-176000 bp) and areas with monomeric alpha satellite in the front domain and back domain of genomic sequence;
- The length of the repeat is best evidenced when L is increased (L=19);
- Repetition number appears clearly only in Figure 9.b (L=19);
- It is sufficient to use divisors of repeat length (171: 9, 19) for L values; this allows a significant reduction in the number of searches;

4. Conclusions

Fourier analysis and grey level spectrograms provide a robust detection method for DNA repeats. Repeats are easily recognizable by regular horizontal lines, which give information about repeat length and number of repeats. However, DNA numerical representation affects the results of the analysis. We've used several numerical representations and we've evaluated the qualitative differences that have appeared in the spectrograms. Tests were conducted on short sequences with microsatellites and on long sequences containing alpha satellite DNA repeats.

In case of short sequences with short repeats (like M65145):

- Using plain indicator sequences with sum spectrum does not lead to good results;
- Using plain indicator sequences with product spectrum allows significantly better results than sum spectrum;
- Adding a supplementary information (repeating factor for nucleotides within target repeats) to the indicator sequences leads to slightly better results for sum spectrum and better results for product spectrum;
- Quartic mapping allows obtaining very good results for repeats in the lower area but does not emphasize repeats in the high area. The best results are obtained using PAM inspired values, which respect the nucleotides complementary properties.
- Using a polynomial representation (with additional information related to the repeats' length) allows obtaining the best results, with evidence of all repetitions for an adequate value of M_m (number of mismatches).

In case of long sequences with DNA alpha satellites (AC017075):

- Using plain indicator sequences and sum spectrum allows a good estimate of the repeat's length and number of repetitions while product spectrum does not reveal the number of repetitions but allows a better localization of areas which contain repetitions;
- Quartic mapping allows obtaining good results with all variants. Again, PAM inspired values give the best results;
- Using a polynomial representation (with additional information related to the repeats' length) allows obtaining the best results, with evidence of all repetitions for an adequate value of M_m. In both cases, for polynomial representation, M_m values affect computational effort and the quality of the results. If the values used for M_m are too small or too large, this may damage results. The values of this parameter are chosen using biological criteria.

Inclusion of additional information (such as nucleotide repeating factor within target repeat) or using a numerical representation which respects DNA's chemical properties allows the obtaining of better results. If the length of the searched repeats is known, polynomial representations lead to the best results regardless of repeat length. In addition, if the length of repeated sequence admits divisors, computational effort can be reduced substantially but a preprocessing stage is necessary to obtain the associated numerical sequence.

Acknowledgements

This work was partly supported by CNCSIS–UEFISCSU, project number PN2-PARTENERIATE-42127/2008 and partly by the project "Doctoral studies in engineering sciences for developing the knowledge based society-SIDOC" contract no. POSDRU/88/1.5/S/60078, project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

1. KRISHNAN, A., TANG, F, **Exhaustive Whole-Genome Tandem Repeats Search**, Bioinformatics Advance Access, vol. 20(16), 2004, pp. 2702-2710.
2. RIVALS, E., **A Survey on Algorithmic Aspects of Tandem Repeats Evolution**, Intl, J. Foundations of Computer Science, vol. 15, 2004, pp. 225- 257.
3. WEXLER, Y., Z. YAKHINI, Y. KASHI, D. GEIGER, **Finding Approximate Tandem Repeats in Genomic Sequences**, RECOMB'04, March 27–31, 2004, San Diego, California, USA.
4. RUDD, M. K., H. F. WILLARD, **Analysis of the Centromeric Regions of the Human Genome Assembly**, TRENDS in Genetics, 2004, vol. 20(11), pp. 529-533.
5. COWARD, E., **Equivalence of Two Fourier Methods for Biological Sequences**, Journal of Mathematical Biology, vol. 36, 1997, pp. 64-70.

6. AFREIXO, V., P. J. S. G. FEREIRA, D. SANTOS, **Fourier Analysis of Symbolic Data: A Brief Review**, Digital Signal Processing, vol. 14, 2004, pp. 523-530.

7. ANASTASSIOU, D., **Genomic Signal Processing**, IEEE Signal Processing Magazine, vol. 18(4), pp. 8-20.

8. CHAKRAVARTHY, K. et al., **Autoregressive Modelling and Feature Analysis of DNA Sequences**, EURASIP Journal on Applied Signal Processing, vol. 1, 2004, pp. 13-28.

9. POP, G. P., E. LUPU, **DNA Repeats Detection using BW Spectrograms**, IEEE-TTTC Intl. Conf. on Automation, Quality and Testing, Robotics, AQTR 2008, May 22-25, 2008, Romania, Tome III, pp. 408-412.

10. POP, G.P., **Spectral Representations of Alpha Satellite DNA**, WSEAS Trans. Information Science and Applications 2009, vol. 5(6), pp. 819-828.

11. PAAR, V, 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(1), Nov. 3, 2008, p. 466.

12. ACHUTHSANKAR, S. N, P. S. SIVARAMA, **A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopotential (EIIP)**, Bioinformation, vol. 1(6), 2006, pp. 197-202.

13. HAMMING, R. W., **Error Detecting and Error Correcting Codes**, Bell System Technical Journal, vol. 29(2), 1950, pp. 147-160.

14. SHARMA, D., B. ISSAC, G. P. S. RAGHVA, R. RAMASWAMY, **Spectral Repeat Finder (SRF): Identification of Repetitive Sequences using Fourier Transformation**, Bioinformatics, vol. 20(9), 2004, pp. 1405-1411.

15. DODIN, G, P. VANDERGHEYNST, P. LEVOIR, C. CORDIER, L. MARCOURT, **Fourier and Wavelet Transform Analysis, A Tool for Visualizing Regular Patterns in DNA Sequences**, Journal of Theoretical Biology, vol. 206, 2000, pp. 323-326.

16. SUSILLO, A., A. KUNDAJE, D. ANASTASSIOU, **Spectrogram Analysis of Genomes**, EURASIP Journal on Applied Signal Processing, vol. 1, 2004, pp. 29-42.

17. TIWARI, S., S. RAMACHANDRAN, A. BHATTACHARYA., S. BHATTACHARYA, R. RAMASWAMY, **Prediction of Probable Genes by Fourier Analysis of Genomic Sequences**, Computer Applications in the Bioscience, vol. 13(3), 1997, pp. 263-270.

18. VOSS, R., **Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences**, Physical Review Letters, vol. 68, 1992, pp. 3805-3808.

19. HERZEL, H., O. WEISS, E. N. TRIFONOV, **10-11 bp Periodicities in Complete Genomes Reflect Protein Structure and Protein Folding**, Bioinformatics, vol. 15, 1999, pp. 187-193.

20. TRAN, T. T., V. A. EMANUELE II, G. T. ZHOU, **Techniques for Detecting Approximate Tandem Repeats in DNA**, Proceedings of the International Conference for Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, May 17-21, 2004, vol. 5, pp. 449-452.

21. EMANUELE II, V. A., T. T. TRAN, G. T. ZHOU, **A Fourier Product Method for Detecting Approximate Tandem Repeats in DNA**, IEEE Workshop on Statistical Signal Processing, Bordeaux, 2005, July 17-20, pp. 1390-1395.

22. VAIDYANATHAN, P. P., B.-J. YOON, **The Role of Signal-Processing Concepts in Genomics and Proteonomics**, J. Franklin Institute (Special Issue on Genomics), 2004, pp. 1-27.