New Aspects in Numerical Representations Involved in DNA Repeats Detection

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

Abstract: - The presence of repeated sequences is a fundamental feature of biological genomes. The detection of tandem 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 arises from the fact that the repeat units can be either exact or imperfect, in tandem or dispersed, and of unspecified length. Many of the methods for detecting repeated sequences are part of the digital signal processing 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. Therefore the numerical representation of genomic signals is very important.
This paper presents results obtained by combining grey level spectrograms with two novel numerical representations to isolate position and length of DNA repeats.

Key-Words: - *Genomic Signal Processing, Sequence Repeats, DNA Representations, Fourier analysis, Spectrograms*

1 Introduction

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. There are tens of billions of bases and sequence records. On average, these databases double in size every 12 months. As a result, computers have become an indispensable tool for biological research because they provide the means for storing large quantities of data and revealing the relationships between them.
A striking genetic difference between species is the size of their genome. Relatively simple organisms may have much larger genome than Homo sapiens for instance. These dramatic differences are due to the presence of repeats. In general, in eukaryotes duplicated genetic material is abundant and can account for up to 60% of the genome. Although some of the mechanisms that generate these repeats are known, from the point of view of evolution, the reasons for such redundancy remain an enigma [1].
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. 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.
Genomic signal processing is now very important in understanding the information contained in the biological genome. 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.
Genomic analysis and bioinformatics is of great topical and offers answers to the problems of the most varied [15][16][17].
This paper presents results obtained by combining grey level spectrograms with two novel numerical representations to isolate position and length of DNA repeats.

2 Interests in Tandem Repeats

Nucleotide and protein sequences contain patterns or motifs that have been preserved through evolution because they are important to the structure or function of the molecule. In proteins, these conserved sequences may be involved in the binding of the protein to its substrate or to another protein, may comprise the active site of an enzyme or may determine the three dimensional structure of the protein. Nucleotide sequences outside of coding

regions in general tend to be less conserved among organisms, except where they are important for function, that is, where they are involved in the regulation of gene expression. Discovery of motifs in protein and nucleotide sequences can lead to determination of function and to elucidation of evolutionary relationships among sequences.
The interest in detecting tandem repeats can be summarized as follows [10]:

- Theoretical interest: related to their role in the structure and evolution of the genome.
- Technical interest: 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.

3 Tandem Repeats

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. Among those, biologists distinguish micro-satellites, mini-satellites, and satellites, according to the length of their repeated unit: between 1 and 6 base-pairs, between 7 and 50 base-pairs, and above 50 base-pairs, respectively. These names are mainly used for repeats located in regions that do not contain genes. In addition to these sub-classes, numerous groups of similar genes that originate from the same ancestor gene are organized in tandem and are termed tandemly repeated genes.
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. These two types of events are named amplification and contraction. Like any other segment of the genome, the repeated copies also change through point mutations: insertion, deletion or substitution of one base [1].
Point mutations give rise to approximate tandem repeats (ATR). The pattern of point mutations along the tandem array of copies informs us on the parent-child relationships between copies. In other words, it gives access to the history of the tandem repeat. The relatively high frequency of these events let these local repeats evolve rapidly. For a given species and at a precise location on the chromosome, a locus, the repeat varies in sequence and/or length in different individuals. Hence, such a locus is said to be polymorphic and each different sequence encountered at this locus is called an allele.
The single-point mutation is a comparatively rare and generally detrimental event. The changing of a single nucleotide only occurs at a rate of 1 in a 100 million nucleotides per generation so it is hard to see how species depending solely on this mechanism would be able to adapt very quickly to changes in their environments. And yet, as the fossil record demonstrates, new species emerge and diversify quite rapidly in geological terms and for species to be able adapt and evolve as quickly as they clearly do, other mechanisms for genetic modification must also be in play.
Point mutations could cause two adjacent copies to diverge so far that their common ancestry is not recognizable anymore from sequence similarity. In this case, it is not a repeat anymore. A major hypothesis is that amplification is favored by the similarity of adjacent patterns, and that when copies have diverged for a long time such former repeat does not undergo amplification anymore. In highly polymorphic loci, like some minisatellites, amplifications and contractions are more probable than point mutations. On the contrary, tandemly repeated genes can accumulate hundreds of mutations and still undergo some amplifications; in this case, amplifications and contractions are less frequent than point mutations.
The repeat sequence might be a simple alternation of two nucleotides or it could be a pattern of dozens or even thousands of nucleotides repeated over and over. These mutations occur at a rate 100,000 times more frequently than single-point mutations do but, unlike the latter, their effect are generally be quite subtle and normally not detrimental the organism. Single-point mutations, on the other hand, are usually either neutral or fatal and only a tiny number of these mutations bestow any benefit at all on the organism.

4 Definitios

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, the repeat is usually called a tandem repeat.
However, perfect tandem repeats are of limited biological interest, since events such as mutations, translocations and reversal events will often render the copies imperfect. 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. Finding ATRs in a sequence is a harder task than finding perfect repeats and has been addressed by several papers during recent years. The scope of ATRs discovered by 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 others, time requirements render the algorithm infeasible for the analysis of whole genomes containing millions of base pairs. Note that in the case of non-integer exponent, a pattern associated with the repeat is not defined uniquely.

5 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. 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:

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

and is obvious that:

$$\sum_j x_{j,k} = 1, \ for \ \ all \ \ k \tag{2}$$

This approach produces a four-dimensional representation yielding an efficient representation for spectral analysis. We used a modified form of this representation in this work to improve TRs detection performances (section 7).
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 [8] 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 [9].
All these representations have advantages for particular analyses but suggest some DNA properties beyond that inherent to them. Starting from these representations we introduced two novel representations to reduce the dimensionality of representation and generate only one numerical sequence for each DNA sequence (section 8).

6 DNA Spectral Analysis

Spectral analysis may be performed by taking the Discrete Fourier Transform (DFT) of each of the indicator sequences [2][3]. Applying DFT definition to all indicator sequences, for alphabet A_4, we obtain another sequences $X_A[k]$, $X_C[k]$, $X_G[k]$, $X_T[k]$:

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

where:

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

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

From (2) and (3) it follows that:

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

The sequences $X_A[k]$, $X_C[k]$, $X_G[k]$, $X_T[k]$ provide the total spectrum of the DNA sequence [5][6][7][8]:

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

In most cases $S[k]$ has a peak at the sample value k=N/3 (Fig. 1), as demonstrated in many papers [11][12]. 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.

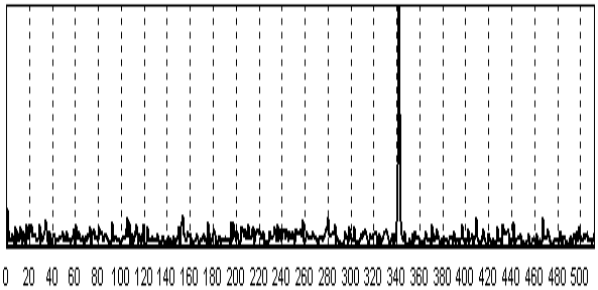

Fig. 1. S[k] showing a strong period-3 property.

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 [11][12]. 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.
The peaks 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 [4] [14] such as:

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

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 $f = 1/p, 2/p, 3/p,\ldots$ 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 Eq. 4 will be equal to zero. To avoid this, a modified product spectrum is defined, as:

$$P[k] = \prod_{\alpha \in \{A,T,G,C\}} (|X_\alpha[k]| + c), k = 0,1,...,N-1 \quad (8)$$

Where c is a small positive constant.

Fig. 2 presents the product spectrum of a genome sequence with tandem repeats using a 512 DFT, which enable to detect the presence of TRs 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 [7]. Now, the candidate peaks can be isolated and the length of TR, $N_i = 1/f_i$ can be estimated.

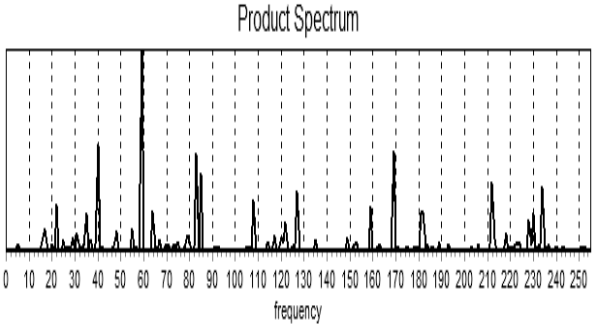

Fig. 2. P[k] showing peaks at different periods

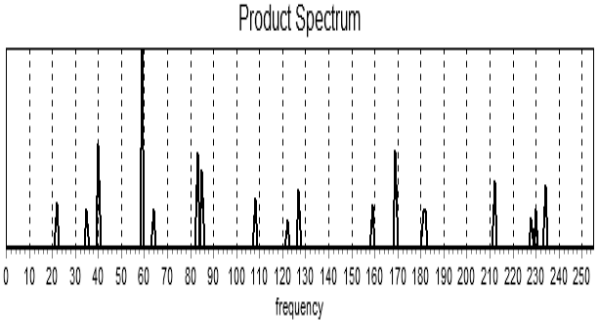

Fig 3. P[k] using a threshold T=3.5

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, $P[k]$

can be used to represent DNA sequence spectra in another way, namely in grey level spectrograms. Fig. 4 shows a spectrogram using DFTs of length 256 of microsatellite M65145 sequence (GenBank). Spectrogram was generated using value T=3.5 for threshold and a global normalization for image. In this way, only significant peaks from *P[k]* will be present and is easier to identify the presence of TRs and the associated length.

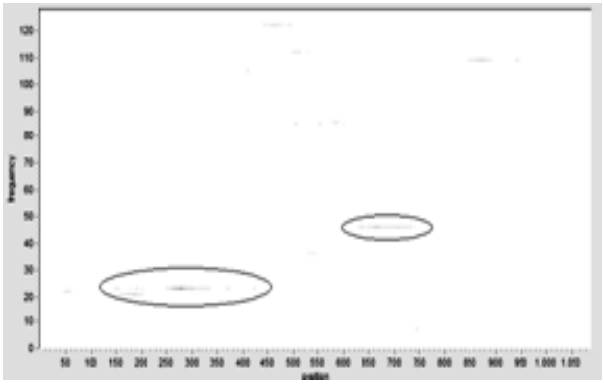

Fig. 4. Product spectrum spectrogram

In this case TRs appear as horizontal lines (more or less continue) at frequencies values $f_1=24$, $f_2=48$. Value $f_1=24$ correspond to a 11mer repeats (256 div 24) while the line at $f_2=48$ suggest that some 5mer TRs are part of 11mer TRs.

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.

Spectrogram offers a global view of product spectrum but is difficult to estimate the location of TRs even if horizontal axis contains nucleotide position. This can be done calculating and representing the values of $P[f_i]$ in a sliding window along the sequence [5] [6] [13].

Fig. 5 presents the product spectrum values *P[f1]* 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 TR) as those where peaks are significant.

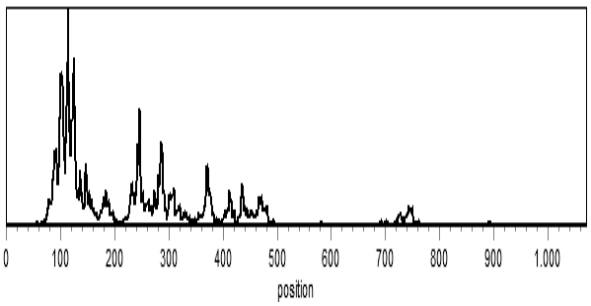

Fig. 5. P[f_1] along DNA sequence

Fig. 6 presents the product spectrum values *P[f2]* of the same sequence. In this case, the peak positions seem to be complementary which suggest that some 5mer TRs are part of 11mer TRs.

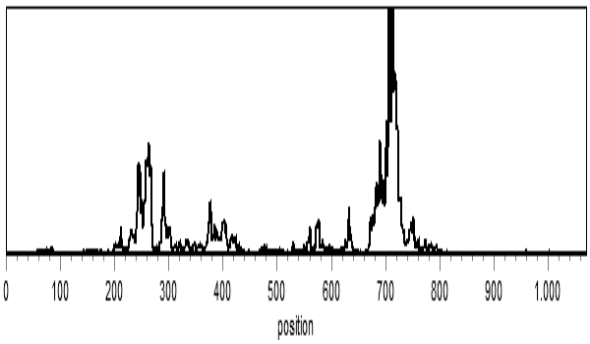

Fig. 6. P[f_2] along DNA sequence

Table 1 list the TRs values and positions from M65145 (GenBank). First eight TRs correspond to the line at $f_1=24$ while last two correspond to the line at $f_2=48$. 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 exact enumeration or even by a heuristic local alignment method.

TABLE 1
11MER REPEATS IN THE MICROSATELLITE M65145

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

This method do not assume any knowledge about the pattern that is being repeated, the size (period) of the pattern, nor the location of the repeats and is insensitive to point mutations. Still, it has disadvantages:

- Repeated sequences are evidenced by horizontal segments more or less obvious;
- Not all repeated sequences are put in evidence;
- Some repeated sequences are reported at a frequency double (half length).

7 Modified Indicator Sequences

Often, TRs pattern 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 same value in sequence:

$$u_{\alpha,k} = \begin{cases} m, & if \ \ a_k = \alpha \\ 0, & otherwise \end{cases} \tag{9}$$

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

$$\begin{aligned} u_A[n] &= 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_C[n] &= 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_G[n] &= 0\ 1\ 0\ 0\ 0\ 0\ 0\ 4\ 4\ 4\ 4 \\ u_T[n] &= 1\ 0\ 0\ 0\ 3\ 3\ 3\ 0\ 0\ 0\ 0 \end{aligned} \tag{10}$$

Second, the expected repeated factors in TR for each nucleotide are included in indicator sequences by limiting the initial repeat factor to expected repeat factor in TR. 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 becomes:

$$\begin{aligned} u_A[n] &= 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_C[n] &= 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_G[n] &= 0\ 1\ 0\ 0\ 0\ 0\ 0\ 3\ 3\ 3\ 3 \\ u_T[n] &= 1\ 0\ 0\ 0\ 2\ 2\ 2\ 0\ 0\ 0\ 0 \end{aligned} \tag{11}$$

Next figures (7…9) shows product spectrum grey-level spectrograms for same microsatellite M65145 sequence (GenBank) using modified indicator sequences with different values for expected nucleotide repeating factors.

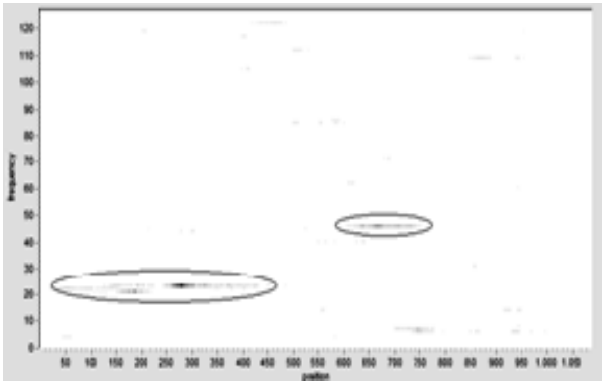

Fig. 7. Product spectrum spectrogram for r_A=1, r_G=1, r_C=1, r_T=2

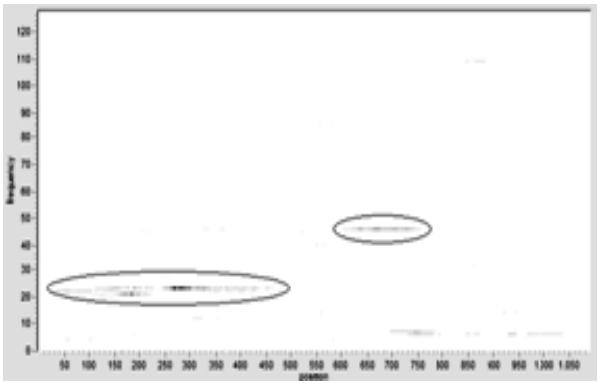

Fig. 8. Product spectrum spectrogram for r_A=1, r_G=2, r_C=1, r_T=2

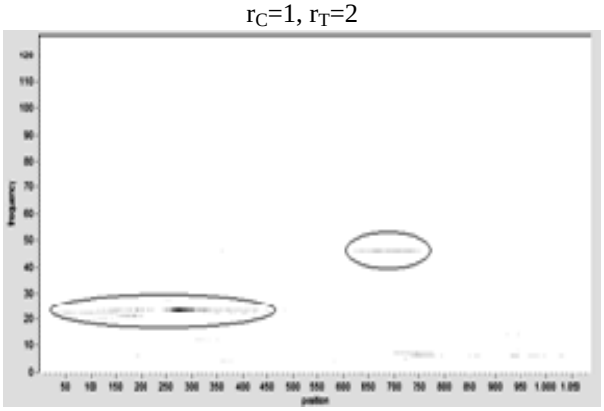

Fig. 9. Product spectrum spectrogram for r_A=1, r_G=2, r_C=1, r_T=3

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 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 TRs can be more easily located.

8 Generating a Single Numerical Sequence

In order to simplify DNA spectral analysis, we propose a sequence representation which takes into account the length of the expected repeats and the number of possible mismatches because of point mutations. Finally, only one set of numerical values is provided for spectral analysis.

For a DNA sequence of length *L* a numerical value is associated in polynomial-like representation:

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

Where V_α is the value of a single nucleotide as follows: A=1, G=2, C=3, T=4. For example,

consider the sequence TCAGA, then the computed value is: 43121. This approach will be used to represent repeated sequences of certain length.

On the other hand point mutations must be taken into consideration. Two sequences that differ by only one nucleotide have a high degree of similarity but will lead, through this approach, two completely different numerical values. As a result is necessary that all similar sequences which differ by the same number of nucleotides (same number of point mutations), lead to the same numerical representation.

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 following input values are needed:

- a DNA sequence of length *N*;
- the length of expected repeated sequence, *L*;
- the maximum number of mismatches (as a result of point mutations) in the repeated sequences, *Mm*.

The algorithm is summarized bellow:

- consider all successive subsequences of length L in the 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)* values.
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.

Here is the algorithm pseudocode description:

```
foreach curr_pos in (0,…, N - L)
   foreach calc_pos in (0,…,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;
```

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):

```
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;
```

An additional parameter (*NRep*) for the number of similar sequences (with same Hamming distance) as a threshold allows reducing the situations in which the consensus values are computed.
Final numerical values from *SeqVal[]* array can be used to compute and represent spectra in grey-level spectrograms.
Next figures (10…18) shows grey level spectrograms obtained for same microsatellite M65145 sequence (GenBank), using DFT of length 128 for spectrum and with *L=11* and different values for *Mm* and *NRep* parameters.

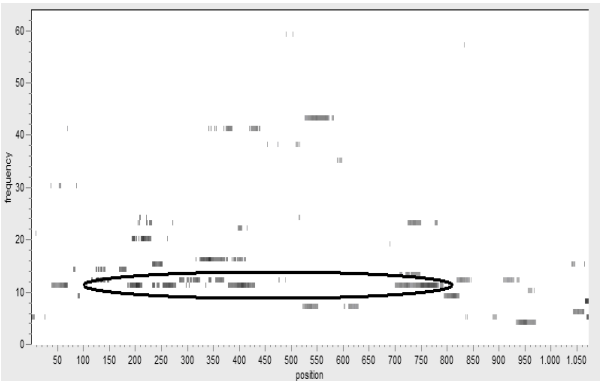

Fig. 10. DNA spectrum for Mm=1, NRep=3

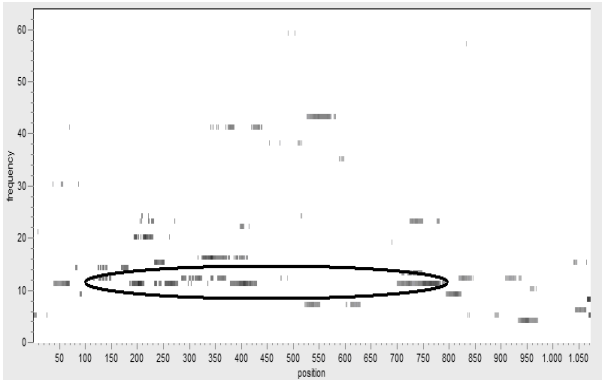

Fig. 11. DNA spectrum for Mm=1, NRep=2

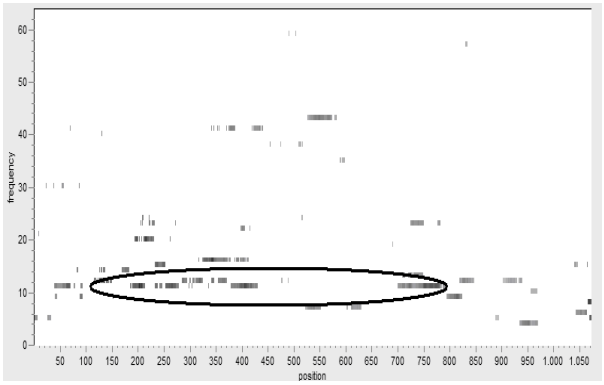

Fig. 12. DNA spectrum for Mm=1, NRep=1

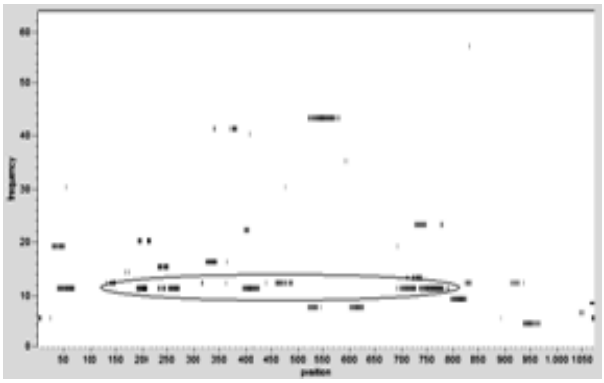

Fig. 13. DNA spectrum for Mm=2, NRep=3

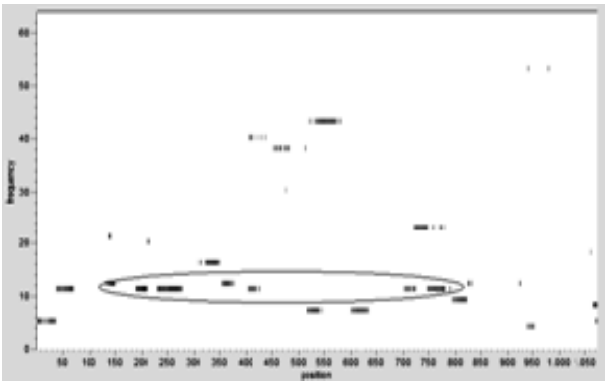

Fig. 14. DNA spectrum for Mm=2, NRep=2

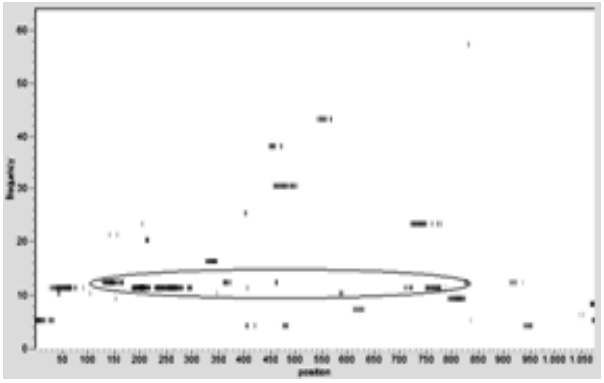

Fig. 15. DNA spectrum for Mm=2, NRep=1

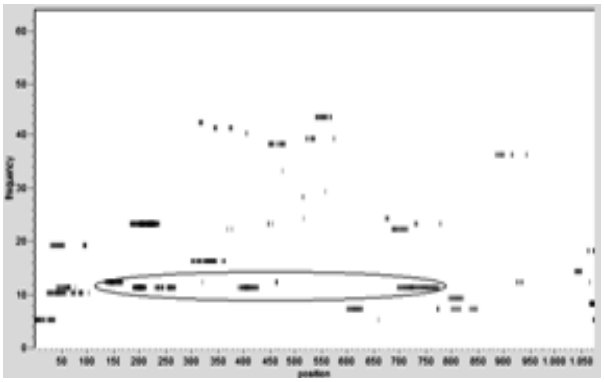

Fig. 16. DNA spectrum for Mm=3, NRep=3

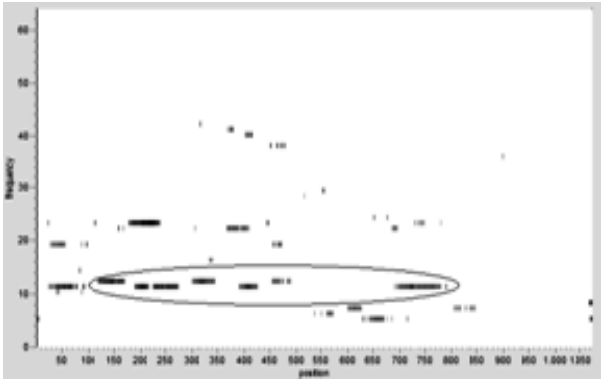

Fig. 17. DNA spectrum for Mm=3, NRep=2

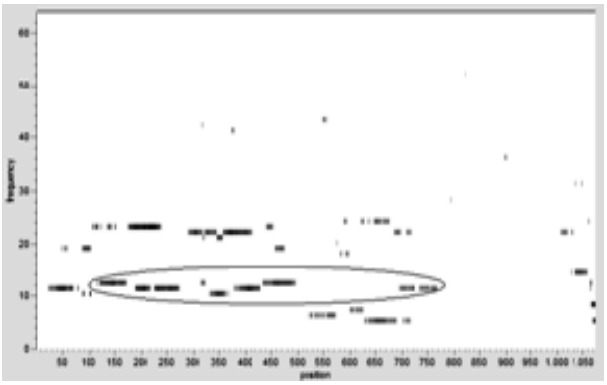

Fig. 18. DNA spectrum for Mm=3, NRep=1

The best results are obtained for *Mm=3* and *NRep=2* (Fig. 17), results that are better than those obtained with product spectrum (Fig. 4) due to the presence of a long line at $f_1=11$ which cover almost all of repeat sequences from Table 1.
Fig. 19 presents the spectrum values $P[f_1]$ of the same sequence using *Mm=3* and *NRep=2*. In this case, is easy to identify the regions containing the repeats (11mer TR) as those where peaks are significant. These peaks cover almost all of repeat sequences from Table 1 and there is no need to represent $P[f_2]$ to locate repeats in region 500-750. In addition, these peaks corresponds to horizontal segments at (and near) value $f_1=11$ which confirms the correctness of the results generated by previous method.

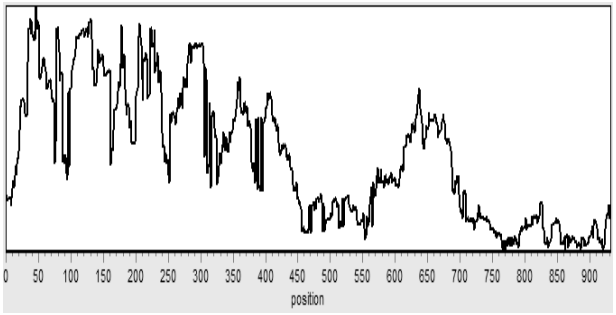

Fig. 19. P[f_1] along DNA sequence

This algorithm has the advantage of simplicity. Also, 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.

9 Conclusion

Pattern discovery is an important area of bioinformatics. The algorithms for pattern discovery use wide range of computer science techniques, ranging from exhaustive search, elaborate pruning techniques, efficient data structures, to machine learning methods and iterative heuristics.
The detection of tandem repeats is important in biology and medicine as it can be used for phylogenic studies and disease diagnosis.
The Fourier product of nucleotide subsequences has shown strong robustness in detecting ATRs, especially those with substitutions and deletions. In this method, the period to be detected in a given DNA sequence is limited by the window length but the method do not assume any knowledge about the pattern that is being repeated, the size (period) of the pattern, nor the location of the repeats.
The modified product spectrum method accuracy can be increased by using a modified form of indicator sequences which include the nucleotide expected repeating factors in target TRs.
A polynomial-like representation of DNA sequences provides a single numerical sequence which can be used directly in spectral analysis and yields improved results. The algorithm is simple, need no addition structures or special memory requirements but needs a priori information about repeat length and the number of mismatches. However, in many situations, biologists know this information before such as this is not a real disadvantage.
In conclusion, these methods can be used as a good screening tools to determine if there are repeated sequences in analyzed sequence and to localize the area they are placed. Then other (exact) methods can be used to determine the pattern of repeated sequences and exact position.

Acknowledgement
This work has been partly supported by the grant project PNCDI-IDEI-334/2007.

References:
[1] A. Krishnan and F. Tang, Exhaustive Whole-Genome Tandem Repeats Search, *Bioinformatics Advance Access*, May 14, 2004.
[2] D. Anastassiou, Genomic signal processing, *IEEE Signal Process. Mag.*, 18 (4) (2001) 8–20.
[3] 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.
[4] 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.
[5] Susillo, A., Kundaje, A., and Anastassiou, D., Spectrogram analysis of genomes, *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 29–42, 2004.
[6] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences, *J. Theor. Biol.*, vol. 206, pp. 323–326, Oct. 2000.
[7] D. Sharma, B. Issac, G.P.S. Raghva, R. Ramaswamy, Spectral Repeat Finder(SRF): identification of repetitive sequences using Fourier transformation, *Bioinformatics*, vol. 20, no. 9, Nov. 2004, pp. 1405-141.

[8] 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.

[9] Borodovsky, M. and et al., GeneMark: A Family of Gene Prediction Programs, *http://opal.biology.gatech.edu/GeneMark/*.

[10] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acid Research*, 27:573–580, 1999.

[11] R. Voss, Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences, *Phys. Rev. Lett.*, vol. 68, 3805-3808, 1992.

[12] 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.

[13] P. P. Vaidyanathan, and B-J. Yoon, The role of signal-processing concepts in genomics and proteonomics, Invited paper, J. Franklin Institute, *Special Issue on Genomics*, 2004, pp. 1-27.

[14] T.T. Tran, 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, Candada, May 2004, vol. 5, pp. 449–452.

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

[16] 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).

[17] 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).