A VISUAL TOOL FOR DNA REPEATS LOCALIZATION

Petre G. POP
Technical University, Cluj-Napoca, Comm. Dept.,
E-mail:Petre.Pop@com.utcluj.ro

Abstract: 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. This paper presents results obtained by combining grey level spectrograms with a novel numerical representation to isolate position and length of tandem repeats (TRs) in DNA sequences.

***Key words**: Genomic Signal Processing, Fourier analysis, Spectrograms, Sequence Repeats.*

I. INTRODUCTION

The presence of repeated sequences is a fundamental feature of genomes. In general, in eukaryotes, organisms whose cells bear a kernel, 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].

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: between 1 and 6 base-pairs, between 7 and 50 base-pairs, and above 50 base-pairs, respectively. In addition to these sub-classes, numerous groups of similar genes that originate from the same ancestor gene are organized in tandem. They 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. Like any other segment of the genome, 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].

The pattern of point mutations along the tandem array of copies gives access to the history of the tandem repeat. Tandem repeats can also be used for disease diagnosis. 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.

II. ASSIGNMENT OF NUMERICAL VALUES

Biomolecular sequences, like DNA and proteins, 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. There are many representations proposed and adapted to the type of analysis.

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:

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

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 [3]. This representation was useful for nucleotide quantization to amino acids and in autocorrelation analysis. To preserve DNA’s reverse complementary properties, discrete numerical sequence symmetric about y-axis, inspired from pulse amplitude modulation, can be used, in which A=-1.5, G=-0.5, C=0.5, T=1.5 [6]. For statistical approaches using Markov models, a four Galois field assignment can be used in which A=0, C=1, T=2, G=3.

All these representations have advantages for particular analyses but suggest some DNA properties beyond that inherent to them.

Manuscript received February 28, 2008; revised June 6, 2008

III. DNA SPECTRAL ANALYSIS

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

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

In most cases *S[k]* has a peak at the sample value k=N/3 (Fig. 1), often called a *period-3* property of the DNA sequences.

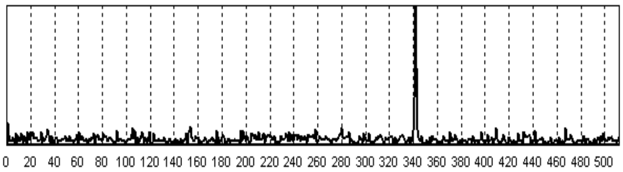

Figure 1. S[k] showing a strong period-3 property.

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 indicator of gene location [3][4][7].

Fourier analysis can be used to detect hidden periodicity and is robust in the presence of substitutions, insertions, and deletions. The algorithm is based on a Fourier product spectrum [5] defined as:

$$P[k]=\prod_{\alpha\in\{A,T,G,C\}}(\left|X_\alpha[k]\right|+c),\ k=0,1,\ldots,N-1 \quad (3)$$

Where c is a small positive constant and $X_\alpha[k]$ 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). 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.

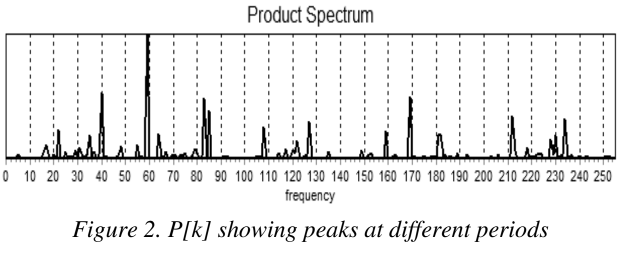

Figure 2. P[k] showing peaks at different periods

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. Now, the candidate peaks can be isolated and associated information (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, P[k] can be used to represent DNA sequence spectra in another way, namely in grey level spectrograms. Fig. 3 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. In this case TRs appear as horizontal lines (more or less continue) at frequencies values f_1=24, f_2=48 (vertical axis). 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 precise 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 [4] [5].

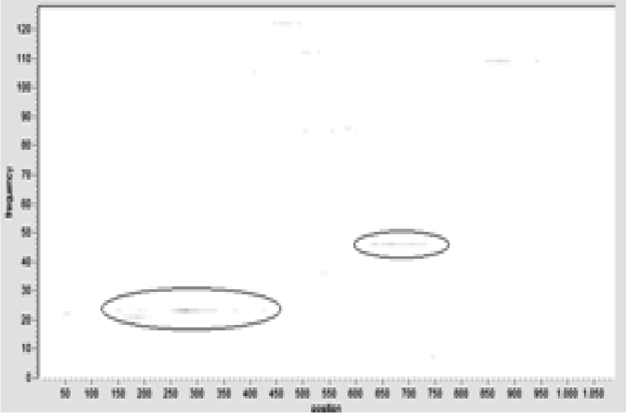

Figure 3. Product spectrum spectrogram

Table I list the TRs values and positions from M65145 (GenBank).

As one can see, first eight TRs correspond to the line at f1=24 while last two correspond to the line at f2=48. Since the length of the repeat (1/fi) 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 I
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

IV. ALGORITHM FOR GENERATING NUMERICAL VALUES

In order to simplify spectral analysis, we propose a novel 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:

$$V = \sum_{k=0}^{L-1} V_{\alpha_k} 10^{k} , \quad \alpha \in \{ A , G , C , T \} \qquad (4)$$

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 TCCGA, then the computed value is: 43321.

In passing from DNA sequence to numerical values, we need to use Hamming distance and consensus value. 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 in the repeated sequences, *Mm*.

The algorithm is summarized bellow:

- determine all the positions in original sequence for which the Hamming distance is less or equal the prefixed mismatches number;
- determine the consensus sequence for all subsequences of length L, starting at these positions;
- compute the numerical value for consensus sequence and assign this value to all these 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 to exclude first *L* subsequences starting from current position (for which is no sense to evaluate the distance).

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 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;
```

Another improvement is obtained through the use of an additional parameter (*NRep*) for the number of similar sequences (with same Hamming distance) as a threshold which allows reducing the situations in which the consensus values are computed.

Next figures shows grey level spectrograms obtained for same microsatellite M65145 sequence (GenBank), using 128 DFT for spectrum and with *L=11* and different values for *Mm* and *NRep* parameters.

The best results are obtained for *Mm=3* and *NRep=2* (Fig. 8), results that are better than those obtained with product spectrum (Fig. 3) due to the presence of a long line at $f_1=11$ which cover almost all of repeat sequences from Table I.

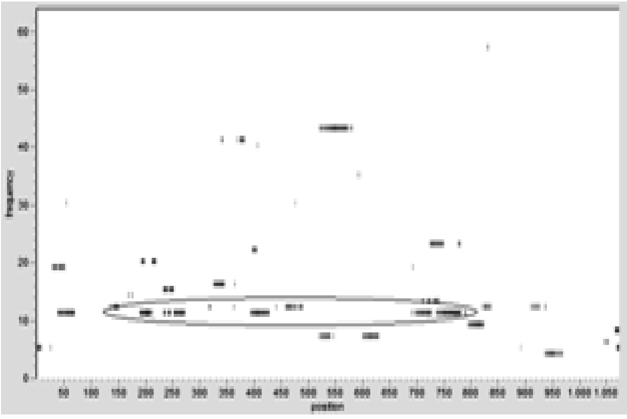

Figure 4. DNA spectrum for Mm=2, NRep=3

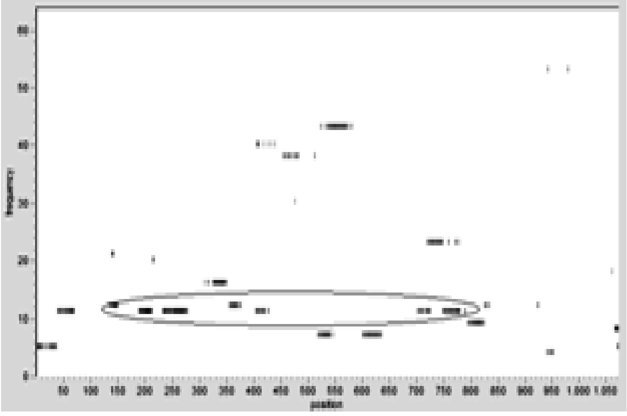

Figure 5. DNA spectrum for Mm=2, NRep=2

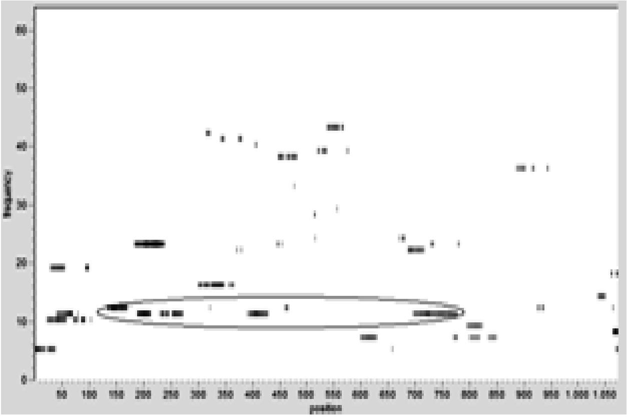

Figure 6. DNA spectrum for Mm=3, NRep=3

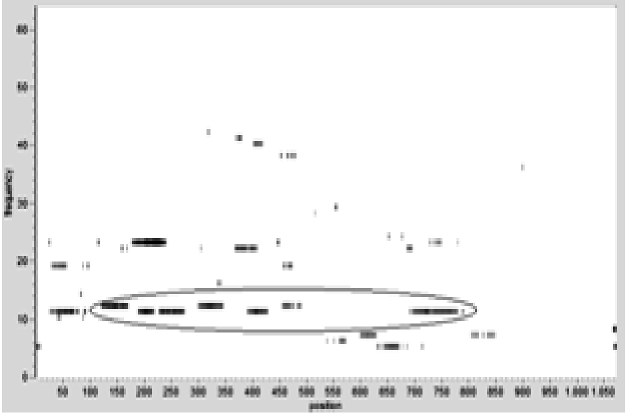

Figure 7. DNA spectrum for Mm=3, NRep=2

Fig. 8 presents the spectrum values *P[f_1]* of the same microsatellite M65145 sequence (GenBank) using parameters values *Mm=3* and *NRep=2*. In this case, is easier to identify the regions containing the repeats (11mer TR) as those where peaks are significant. These peaks cover cover almost all of repeat sequences from Table I 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.

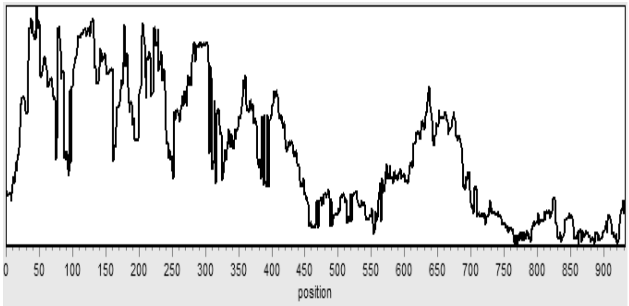

Figure 8. P[f1] along DNA sequence

This algorithm is simple to implement. Also, no additional structures or special memory requirements are needed. The main limitation is related to a priori information about the repeat length and the maximum number of mismatches. However, in many situations, biologists know this information before such as this is not a real disadvantage.

V. CONCLUSIONS

This work presents a novel nucleotide sequence representation used to provide a single numerical sequence for repeats localization with spectral analysis using grey level spectrograms. Results obtained are better than those obtained with Fourier product method. 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.

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] Y. Wexler, Z. Yakhini, Y.Kashi, D.Geiger, "Finding Approximate Tandem Repeats in Genomic Sequences", *RECOMB'04,* March 27–31, 2004, San Diego, California.

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

[4] 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.

[5] 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.

[6] 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.

[7] Susillo, A., Kundaje, A., and Anastassiou, D., "Spectrogram analysis of genomes,", *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 29–42, 2004.