Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 53(67), Fascicola 2, 2008

Detecting DNA Tandem Repeats With a Modified Fourier Product Method and Spectrograms

Petre G. Pop[1], Eugen Lupu[1]

Abstract – The presence of repeated sequences is a fundamental feature of 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. This paper presents results obtained by combining the modified product spectrum and grey level spectrograms with a numerical representation to isolate position and length of tandem repeats (TRs) in DNA sequences.

Keywords: tandem repeats, genomic signal processing.

I. INTRODUCTION

A striking genetic difference between species is the size of their genome. These dramatic differences are due to the presence of repeats. 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.

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 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, 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. 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.

Point mutations give rise to approximate tandem repeats (ATR). The pattern of point mutations along the tandem array of copies 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.

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. NUMERICAL SEQUENCES

DNA 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.

In passing from symbolic to numeric data, the set of symbols is first mapped 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 $(u_{\alpha,\,k})$, k=0,..,N-1 such that:

$$u_{\alpha,k} = \begin{cases} 1, & if\ \ a_k = \alpha \\ 0, & otherwise \end{cases} \qquad (1)$$

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.

[1] Technical University of Cluj-Napoca, Comm.Dept.
Baritiu str., 26-28, 400027, Cluj-Napoca, e-mail petre.pop@com.utcluj.ro

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

A representation which preserve DNA's reverse complementary properties 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.

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

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

First, the indicator sequences are modified to include the repeating factor:

$$u_{\alpha,k} = \begin{cases} m, & if \ \ a_k = \alpha \\ 0, & otherwise \end{cases} \qquad (2)$$

Where *m* is the number of consecutive positions with same value in sequence.

Consider the next nucleotide sequence: TGACTTTGGGG. The initial indicator sequences are:

$$\begin{aligned} u_A[n] &= 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_C[n] &= 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_G[n] &= 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1 \\ u_T[n] &= 1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 0\ 0 \end{aligned} \qquad (3)$$

The modified indicator sequences which include the repeating factors are:

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

Second, the expected repeated factors in TR for each nucleotide are included in indicator sequences by limiting de 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:

$$\begin{aligned} u_A[n] &= 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_C[n] &= 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ u_G[n] &= 0\ 1\ 0\ 0\ 0\ 0\ 0\ 3\ 3\ 3\ 3 \\ u_T[n] &= 1\ 0\ 0\ 0\ 2\ 2\ 2\ 0\ 0\ 0\ 0 \end{aligned} \qquad (5)$$

III. DETECTING TANDEM REPEATS

Spectral analysis techniques provide widely interpretable results but require a mapping of symbols to numbers. 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 [4] [5] and consist of next steps.

First, convert the DNA sequence into four nucleotide subsequences $u_A[n]$, $u_T[n]$, $u_G[n]$, $u_C[n]$.

Second, take Fourier transform of the mean removed processes:

$$X_\alpha[k] = \sum_{n=0}^{N-1} (u_\alpha[n] - m_\alpha) e^{-j\frac{2\pi}{N}kn}, k = 0,1,\ldots,N-1 \qquad (6)$$

where:

$$m_\alpha = \frac{1}{N}\sum_{n=0}^{N-1} u_\alpha[n], \quad \alpha \in \{A,T,G,C\} \qquad (7)$$

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

In the next step, form the Fourier product spectrum:

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

It results that if any of the $u_\alpha[n]$ is periodic with period p, then X[k] will also be periodic with period p. Thus, P[k] should have peaks at frequencies f =1/p, 2/p, 3/p,… reflecting any periodicities in the indicator sequences $u_\alpha[n]$. The period p can thus be inferred from the peak location but the period is limited by the window length (N).

Multiplication as a nonlinear operation is used to enhance peaks in a product spectrum. Since multiplication is a nonlinear operation, it is expected that peaks are enhanced while the "noise floor" is suppressed in a product spectrum. 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. 8 will be equal to zero.

To avoid this, a modified product spectrum is defined, as:

$$P[k] = \prod_{\alpha \in \{A,T,G,C\}} (|X_\alpha[k]| + c), k = 0,1,\ldots,N-1 \qquad (9)$$

where c is a small positive constant.

But not all peaks are significant. The significance of any spectral line should be assessed with respect to the spectral average. A threshold T can be used to find peak candidates such that:

$$\frac{P[k]}{P_m} > T \qquad (10)$$

where P_m is the frame spectral product average.
Peaks with T greater than 2 typically begin to be significant, but from a number of studies, a threshold of T=3.5 is more useful to locate repeats [6].
Fig. 1 presents the product spectrum of the same sequence using threshold T=3.5 to eliminate weak peaks. Now, the candidate peaks can be isolated and associated information (the length of TR, $N_i = 1/f_i$) can be estimated. But 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.

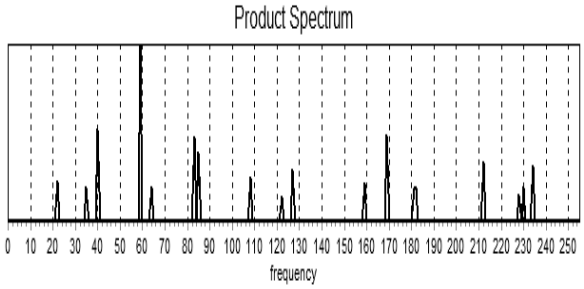

Fig. 1. P[k] using a threshold T=3.5

Instead, P[k] can be used to represent DNA sequence spectra in another way, namely in grey level spectrograms. Spectrograms of DNA sequences simultaneously provide local frequency information for all four bases. Fig. 2 shows a spectrogram using DFTs of length 256 of microsatellite M65145 sequence (GenBank).
The horizontal axis shows the relative nucleotide locations starting from nucleotide 1 while the vertical axis corresponds to the frequencies k from 1 to 128, due to spectra conjugate symmetry. 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 at frequencies values f_1=24, f_2=48. Value f_1=24 correspond to a 11mer repeats (256 div 24).

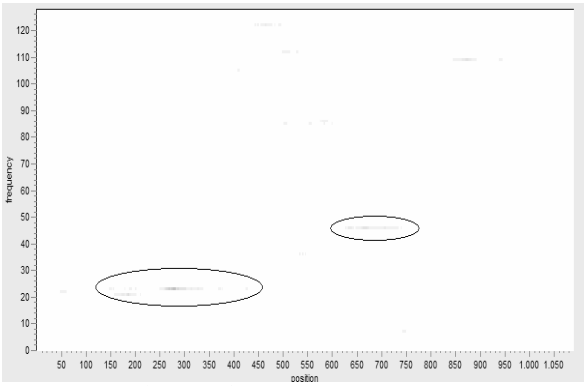

Fig. 2. Product spectrum spectrogram

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 [3] [6].
Fig. 3 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 TR) as those where peaks are significant.
Fig. 4 presents the product spectrum values $P[f_2]$ 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.
Table I list the TRs values and positions from M65145 (GenBank). As one can see, first 8 TRs correspond to the peaks from Fig. 3 while last 2 TRs appears in Fig. 4. 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.

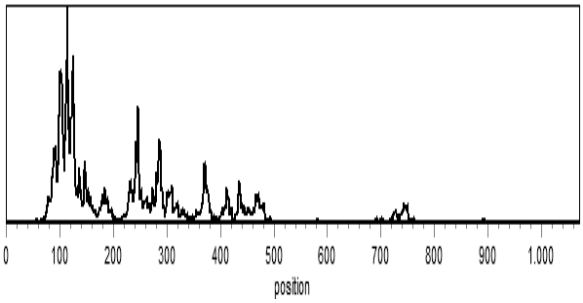

Fig. 3. P[f1] along DNA sequence

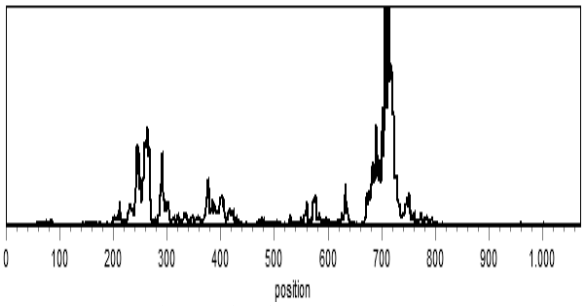

Fig. 4. P[f2] along DNA sequence

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

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

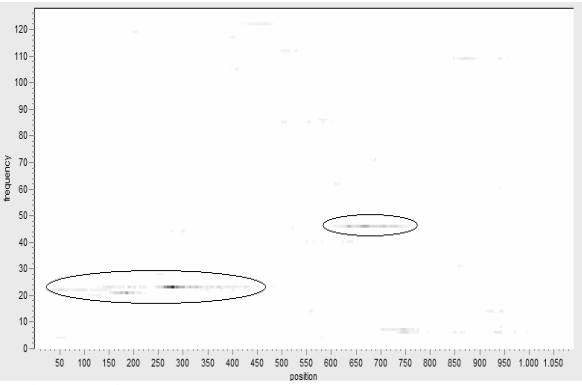

Fig. 5. Product spectrum spectrogram for r_A=1, r_G=1, r_C=1, r_T=2

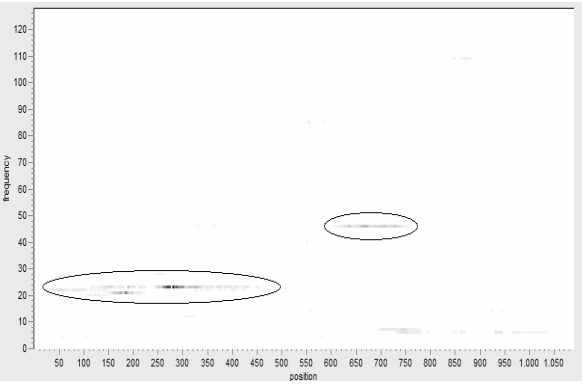

Fig. 6. Product spectrum spectrogram for r_A=1, r_G=2, r_C=1, r_T=2

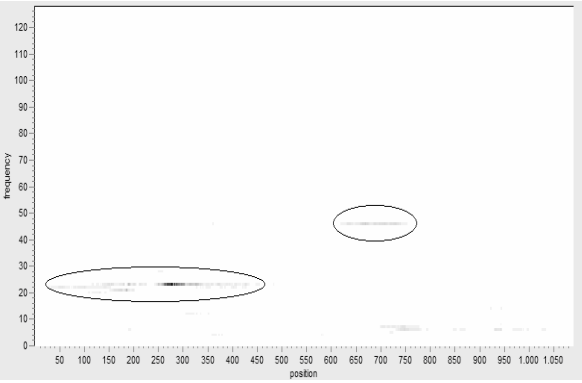

Fig. 7. Product spectrum spectrogram for r_A=1, r_G=2, r_C=1, r_T=3

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 I. 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.

IV. CONCLUDING REMARKS

The Fourier product of nucleotide subsequences has shown strong robustness in detecting TRs, especially those with substitutions and deletions. 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 and spectrograms allows isolating TRs (position and length) on a whole sequence not on a frame by frame basis. The method accuracy can be increased by using a modified form of indicator sequences which include the nucleotide expected repeating factors in target TRs. This allows to be used as a good screening tool in finding tandem repeats.

REFERENCES

[1] 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.

[2] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences", Bioinformatics, vol. 16, no. 12, pp. 1073-1082, Dec. 2000.

[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] T.T. Tran, V.A. Emanuele II, G.T. Zhou, "Techniques for detecting approximate tandem repeats in DNA," in Proceedings of the International Conference for Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Candada, May 2004, vol. 5, pp. 449–452.

[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] 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.