Sequence analysis

OMWSA: detection of DNA repeats using moving window spectral analysis

Liping Du[1], Hongxia Zhou[1] and Hong Yan[1,2*]

[1]Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong,

[2]School of Electrical and Information Engineering, University of Sydney, NSW2006, Sydney, Australia.

ABSTRACT

Summary: Repetitive DNA sequences play paramount biological roles, such as gene variation and regulatory functions on gene expressions. Until now, detection of various kinds of DNA repeats accurately is still an open problem. In this paper, we propose a new method and a visualization tool for detecting DNA repeats in a 2D plane of location and frequency by using optimized moving window spectral analysis. The spectrogram can display the general distribution of repetitive sequences while showing the repeat period, length and location without any prior knowledge. Experiment results demonstrate that our method is accurate and robust even under the condition of excessive mutating and interleaving.

Availability: Software is available upon request

Contract: h.yan@cityu.edu.hk

1 INTRODUCTION

DNA repeats play an important role in the evolution of genetic variation (Parkhill *et al*., 2000). Many of the DNA repeats are even part of a structural or functional region. DNA repeats are commonly classified into tandem repeat and interspersed repeat (Brown, 1999). The tandem repeat in a DNA sequence is generated by two or more contiguous, perfect or approximate copies of a nucleotide sequence (repetitive unit). On the other hand, the interspersed repeat is considered as some moderately repetitive sequence distributed in the whole genome. Detecting repeat in DNA sequences has become an important subject in bioinformatics.

In this paper, we focus on the detection of all kinds of tandem repeats. Generally, tandem repeats are not perfect copies of a given repeat unit. The repeat unit can be mutated by mismatch, insertion and deletion so excessively that it is difficult to locate it correctly. A number of detection tools for such tandem repeats have been proposed and several software programs are available, including the Tandem Repeat Finder (TRF), STRING, STAR, MREPS and ATR (Benson, 1999). Among them, TRF is known as an efficient and popular tool for detecting tandem repeats. These tools focus on searching tandem repeats in texts using string matching and statistical algorithms. These methods have several limitations. On the one hand, heuristic methods do not guarantee to find all possible repeats. On the other hand, computational complexity grows exponentially with the repeat unit length in these methods.

Recently, the Fourier spectrogram, which is a text-free digital signal processing based method, has been used for DNA sequence analysis (Anastassiou, 2000, Sussillo *et*. *al*., 2004). This method can also be used for repeat detection. However, it is well-known that the spectrogram obtained using the Fourier transform (FT) contains the windowing or data truncation artifacts and spurious spectral peaks. This problem has been studied extensively in digital signal and image processing, and in modern spectral estimation methods, parametric techniques, such as the autoregressive (AR) model, are used to achieve a high spectral resolution (Stoica *et*. *al*., 2005, Yan 2002). In this paper, we present a new method to graphically display the potential repeats in the location-frequency plane by using optimized moving window spectral analysis (OMWSA). Tandem repeats have the characteristic of a periodical signal which can be revealed by spectral analysis. In our work, we utilize the AR model to analyze the spectrum of the DNA sequences. By combining the moving window spectral analysis, a repetitive sequence can be displayed visually in the spectrogram of a DNA sequence. From the resulting spectrogram, we can not only identify the existence of repeat areas within a DNA sequence, but also the period, length and location for each repeat without any prior knowledge. Thus, according to the information form the spectrogram, we can extract the repeats from the whole genome.

2 METHODS

For a DNA sequence of length N, the numeric sequence is

$$x(n) = WY = W[u_A(n) \quad u_C(n) \quad u_G(n) \quad u_T(n)]^T \tag{1}$$

where the weight vector $W = \mathrm{col}\{w_s\}$, $s = 0, 1, \ldots, S$, $S = 4$ is the number of bases, and $u_A(n)$, $u_C(n)$, $u_G(n)$ and $u_T(n)$ are the binary sequences assigned to each of the four bases, respectively (Anastassiou, 2000). Therefore, the power spectrum of $x(n)$ is

$$X(k) = w_1 U_A(k) + w_2 U_C(k) + w_3 U_G(k) + w_4 U_T(k), \; k = 0,1,2,\cdots,N-1 \tag{2}$$

where k is the frequency index, $X(k)$ and $U(k)$ are the spectra of $x(n)$ and $u(n)$, respectively. In Equation (2), we adaptively adjust the weight vector W to maximize the power spectrum density.

As spectral analysis methods, neither the FT nor the AR model can provide any information about the number of repetitions or the location at which a periodic component occurs. In our method, a moving window is firstly applied to the signal and then the windowed segment of the signal is further analyzed by the optimized spectral analysis method. Thus, we call the method the optimized moving window spectral analysis. For a DNA sequence, the moving window spectral analysis procedure produces the spectrogram at each frequency k and location n in the location-frequency plane. Therefore, the weight vector W can be set dependent on the spectrum at the frequency k and location n. If we show the periodic components as highlighted regions in the spectrogram in the location-frequency plane, then from the coordinates (n, k) of the regions, we can obtain the information of both the periods and locations of DNA repeats.

3 RESULTS

To verify the capability of our algorithm for identifying tandem repeats, we analyze two DNA sequences and compare the results

*To whom correspondence should be addressed.

© Oxford University Press 2005

of our algorithm with those from the FT spectrogram and the Tandem Repeat Finder (TRF).

In Figure 1, we analyze the sequence X64775, which is highly repetitive. We show the spectrograms obtained using our method OMWSA and the FT as two images respectively. Clearly, the spectrogram obtained using OMWSA has less artifacts and is easier to analyze, both visually and numerically. From the spectrogram of OMWSA, we firstly gain an overall visualization of tandem repeats. The vertical axis corresponds to the frequency k, while the horizontal axis shows the relative base pair location. The highlighted areas in the image correspond to spectral peaks and indicate that there are three repeats within X64775, which are listed in Table 1. All the frequencies of these repeats are about 0.33Hz, which means that the repeat period is 3 bases. The first repetitive sequence is between indices 50 and 84. We searched the sequence and found two adjoining repetitive segments. The repeat units are 'ACG' and 'GCG', respectively. The second repetitive sequence is between indices 142 and 190 and the repeat unit is 'GGC'. TRF only detects the repeat between indices 145 and 188 and the repeat unit is 'GGC'.

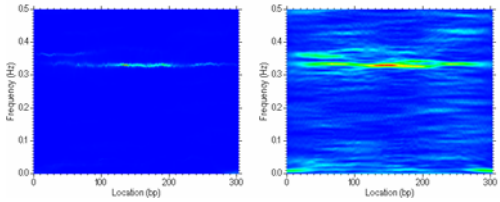

Fig. 1. Spectrograms produced using our OMWSA algorithm (left) and the FT (right) for sequence X64775.

Table 1. Repeats identified by OMWSA and TRF for sequence X64775.

Accession number	Method	Indices	Repeat period	Number of copies
X64775	OMWSA	50 - 60	3	3.6
		61 - 84	3	1.0
		142 - 190	3	2.8
	TRF	145 - 188	3	14.7

In Figure 2, we analyze a section of Homo sapiens complement component 2 (C2) sequence AH005824 between 6800 and 7756 bp. The FT produces a lot of spurious spectral peaks, which make it difficult to identify the genuine frequency components in the signal, both visually and numerically. The AR model does not have this problem, which can be proven theoretically (Stoica *et. al*., 2005, Yan 2002) and is also clearly shown in the spectrogram here. In its spectrogram (image on the left in Figure 2), there is region between 6865 to 7500 bp, which has high spectral intensity values. The frequency of this region is around 0.02 to 0.025 Hz. By searching the sequence, we can find two long interleaved repeats distributed in the region. They are:

CCTCCCTCCCGGACAGGGCGGCTGGCCGGGCGGGGGGCTGACCCCCCCA

and CCTCACTTCTCAGACGGGGCGGCTGCCGGGCGGAGGGGCT.

The repeat periods are 49 and 40, respectively (Table 2). When a sequence contains many copies of a long tandem repeat similar to the above two, TRF may produce several possible repeat units and periods, and it will be rather hard to analyze all these results (Table 2). However, our algorithm does not have such problem. We are able to detect the actual periodic signals more accurately.

We have chosen window length 100 bp for both sequences. The window size should be small enough to achieve a hig spatial resolution but large enough to accommodate several cycles of the longest repeat unit. The optimal prediction order in the AR model can be determined using the Akaike information criterion (AIC) or other criterion functions (Stoica *et. al*., 2005, Yan 2002). In practice, we can vary the window size and the prediction order in order to detect all repeat units in the sequence.

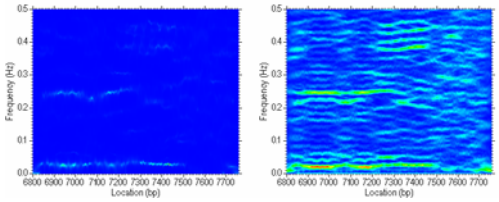

Fig. 2. Spectrograms produced using our OMWSA algorithm (left) and the FT (right) for a section of Homo sapiens complement component 2 (C2) sequence AH005824 between 6800 and 7756 bp.

Table 2. Repeats identified by OMWSA and TRF for a section between 6800 and 7756 in the Homo sapiens sequence AH005824.

Accession Number	Method	Indices	Repeat period	Number of copies
AH005824	OMWSA	6865 - 7043	49	3.6
		7053 - 7089	40	1.0
		7090 - 7227	49	2.8
		7228 - 7381	40	4.0
		7460 - 7500	40	1.0
	TRF	6865 - 7043	49	3.6
		6921 - 7236	176	1.8
		7097 - 7202	49	2.2
		7097 - 7227	49	2.7
		7208 - 7330	40	3.1
		7225 - 7494	39	6.9
		7296 - 7436	76	1.9

In summary, we have presented a new method and a visualization tool for DNA repeat detection. The results demonstrate that the spectrogram obtained by our algorithm can clearly display the general distribution of repetitive sequences. More importantly, the repeats mutating excessively and interleaving each other can be detected with high resolution due to the distinguished ability of parametric spectral analysis in our algorithm. Compared with the traditional FT based spectral analysis, our method produces less artifacts and more reliable results in both visual and numerical analysis. Our OMWSA algorithm also compares with existing software favorably for both short and long repeat detection.

ACKNOWLEDGEMENTS

This work is supported by a grant from the Hong Kong Research Grant Council (project CityU122005).

REFERENCES

Anastassiou, D. (2000) Frequency-domain analysis of biomolecular sequences, *Bioinformatics*, **16**, 1073–1081.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Brown, T. A. (1999) *Genomes*. BIOS Scientific Publishers Ltd., Oxford, UK.

Parkhill, J. *et al*. (2000) Complete DNA sequence of a serogroup a strain of Neisseria meningitidis Z2491. *Nature*, **404**, 502-506.

Stoica, P. *et. al*. (2005) *Spectral Analysis of Signals*, Prentice Hall, New Jersy.

Sussillo, D. *et. al*. (2004) Spectrogram analysis of genomes, *EURASIP J. Applied Signal Processing*, **1**, 29–42.

Yan, H. (ed.) (2002) *Signal Processing for Magnetic Resonance Imaging and Spectroscopy*, Marcel Dekker, New York.