Fast search of locally repetitive elements based on auto-correlation property in genome

Kyung-Seop Shin*, Byung-Chang Chung†, Woo-Chan Kim‡ and Dong-Ho Cho§
Korea Advanced Institute of Science and Technology (KAIST)
Email: ksshin@comis.kaist.ac.kr*, bcchung@comis.kaist.ac.kr†, wckim@comis.kaist.ac.kr‡, dhcho@kaist.ac.kr§

***Abstract*—Since the beginning of a human genome project, 12 years have passed. There are many studies regarding the meaning of human genome sequences and the effort to identify the whole genome of other species. Although genes significantly affect phenotype, the importance of other factors have increased. In this paper, we propose an autocorrelation based method to arrange the repetitive elements which demonstrate a major part of the genomic sequences. The search for the entire genome based on a simple mathematical analysis will be given. The performance of our proposed self autocorrelation based method will be compared with that of conventional scheme for the human chromosome sequence. Fast scanning of the genome sequence based on our proposed scheme can give a clue to analyze the complex function of the genomic sequences.**

***Index Terms*—autocorrelation, repetitive element, genomic sequence**

I. Introduction

Since the first human genome was reported to be decoded in 2001, the genome sequences of several humans as well as a range of other species have been published. Public availability of these genome databases allows for in-depth investigations into the biology of the genome, primarily in regard to their roles in phenotypic determination and disease processes.

There are repetitive elements in majority of the species' genome sequences. It is controversial that these repetitive elements are meaningful in identifying the characteristics of a number of species. From many recent biological researches, the repetitive elements are determined to show the origin of various diseases such as Huntington's disease, Friedreich's ataxia, et al. Besides the repetitive elements' roles, it is necessary to analyze these repetitive elements from the whole genome [1], [2]. The fast search of the repetitive elements can help the understanding of the genome sequences' structure.

The repetitive elements are categorized in two groups by the gap between the repetitive elements. First, tandem repeat is the repetitive element in which each repeat is located adjacently. These tandem repeats can also be divided into three subgroups corresponding to the repeat unit's length. Microsatellite repeat is a tandem repeat with a unit length of 1-4 base pairs, minisatellite is composed of tandem repeats with 6-64 base pairs, and satellite repeats is constructed by a long tandem repeat with a unit length of 5-171 base pair. Second, interspersed repeat is a repetitive element in which each repeat of a family exists far away. It is also divided into several subgroups: short interspersed nuclear elements, long interspersed nuclear elements, transposons, long terminal repeats and so on.

In recent years, there have been several researches regarding the examinations of repetitive elements. Tandem repeat finder(TRF) [3] searched tandem repeats using the computational algorithms employing two criteria: sum of head test and apparent size test. After the TRF study, there are some attempts to find the tandem repeats. ATRHunter[4] found tandem repeats-like repeats using a two phase search which is composed of screening phase and verification phase. TandemSWAN [5] determined fuzzy tandem repeats using the stochastic approach called mask probability and MotifF model. Spectral repeat finder(SRF) [6] indentified the repetitive elements including both tandem repeats and interspersed repeats via discrete fourier transform approach.

The analysis using various mathematical methods was done for the genomic sequences. A frequency-domain analysis for biomolecular sequences used discrete complex value to analyze DNA sequences [7]. Using autocorrelation function, Herzel et al. found 10-11 base pair periodicities in the genome [8], while in another study, they employed the entire 4×4 dimensional covariance matrix of DNA sequences [9]. It was proved that a computational algorithm to find the periodic DNA sequences' pitch with 10-11 period reflects the characteristics of periodic nucleotide sequences [10]. A critical review for the genomic analysis using a correlation structure was summarized in [11].

Diverse populations of the repetitive elements are abundantly present in the genomes of most eukaryotes, with constituting 45% or more portions in the human genome. However, less than 3 % of the sequence constitutes the entire set of ~25,000 to 30,000 typical protein-coding sequences (genes). Although individual repetitive elements' participation in biological processes are well-studied, the investigation of the properties of complex repetitive element arrangements, as a functional genome unit, has not drawn much attention from the field. The comprehensive profile of a genome-wide distribution of repetitive element arrangement architecture in the human genome and other species have not been established yet. In this study, we will identify the genomic repetitive element arrangement using a self autocorrelation function for whole genome or individual chromosome.

The remainder of this paper is organized as follows. Section II will present the operation and mathematical analysis of our proposed algorithm. Self autocorrelation function will be explained with a new mapping function from the DNA sequence to the numerical value based on the correlation between two nucleotide bases. In Section III, we will evaluate

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

our proposed scheme using the results for human chromosome Y. Conventional spectral repeat finder and proposed self autocorrelation based approach will be discussed. Finally, conclusions are made in Section IV.

II. Auto-correlation Based Method

A. Methods

It is necessary to indicate a target sequence as a numerical value. Converting a character sequence to a numerical sequence has been done in various researches. Some of them indicate the DNA sequences as a sequence of complex numbers, for example, convert A to $1+j$, G to $1-j$, C to $-1+j$, and T to $-1-j$ where $j^2=-1$. When the probability of each base-pair appears to be the same, this converted sequence has mean of 0. Using this mapping structure, the DNA can be analyzed sequences mathematically. But, we use the functional mapping rather than this individual mapping of the nucleic acids.

The numerical mapping of the DNA sequences used in our paper will be presented as follows. Instead of mapping each nucleotide base as a numerical value, the correlation of two DNA sequences is selected to measure a sequence mapping. The autocorrelation function is used to detect a periodic signal at initial step. When discrete real-numbered signal X_n has $E\{X_n\}=\mu$, then the value of autocorrelation $r(k)$ is defined as the expected multiplication of corresponding signal X_n and delayed sequence X_{n+k} as follows.

$$r(k) \triangleq E\{(X_n-\mu)(X_{n+k}-\mu)\} = E\{X_n X_{n+k}\} - \mu^2 \quad (1)$$

If this function is maximized with τ over $\tau \in (0, N]$ where N is the length of X_n, then τ can be a minimal possible period. Whether it is a real periodic or asymptotically periodic is not clear yet.

Let's consider original DNA sequence D_n, where $D_i \in \{A,G,C,T\},\ i=1,2,...,N$. A mapping function $I(a,b)$ expresses the correlation of two nucleotides a and b, as a real number, which is defined as

$$I(a,b) = \begin{cases} 3 & ,a=b \\ -1 & ,a \neq b \end{cases} \quad (2)$$

where the function $I(a,b)$ has mean 0 when the sequences are identically and independently distributed(i.i.d.). From the definition of function I(a,b), a highly correlated sequence tends to have higher values than using an individually characterized mapping. Then, when the candidate period of a sequence is given as t, the modified autocorrelation function called self autocorrelation(SAC) function $s(t)$ can be described as follows.

$$s(t) \;=\; \sum_{n=1}^{N-t+1} I(D_n, D_{n+t}) \quad (3)$$

Again, if the sequences are independent of each other, the expected value of this self autocorrelation function becomes 0. Since $I(a,b)$ is an increasing function for the same value of a and b, we can see that at a minimum possible period, $s(t)$ has a maximum value. The predicted period $\hat{T}$ is the minimum value of t which maximizes $s(t)$.

$$\hat{T} = arg\max_t \ s(t) \quad (4)$$

This predicted value will be used to find locally repeated elements.

B. Design

The extraction algorithm for locally repetitive elements is designed based on the self autocorrelation method presented in the methodology subsection. The locally repetitive elements include tandem repeats and interspersed repeats which appear in a small range due to the insertion or deletion phenomenon between them. This small gap of interspersed repeat is denoted by γ. The locally repetitive element extraction algorithm is implemented in the following four steps as shown in Fig. 1.

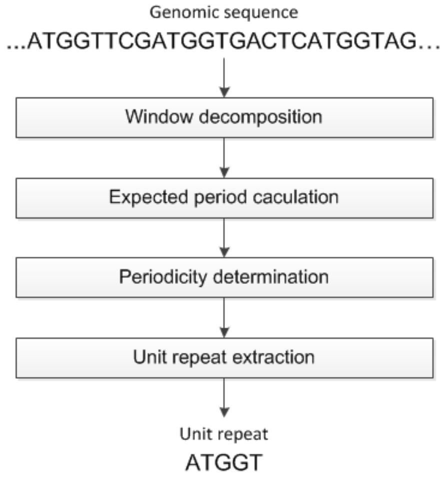
Genomic sequence
...ATGGTTCGATGGTGACTCATGGTAG...
Window decomposition
Expected period caculation
Periodicity determination
Unit repeat extraction
Unit repeat
ATGGT

Figure 1. Flow chart of self autocorrelation based search algorithm

In the first step, raw DNA sequences are decomposed to several windows with length W, in which the finding algorithm operates. As a unit of an operation, a window is employed to manage a huge amount of a genome. The window is a partition of a whole DNA sequence, which is the target for the repetitive elements' extraction. When the size of a genome is large, the computational complexity will increase dramatically. Self autocorrelation function in the sequence with fixed length will decrease this unexpected boost of computational complexity. While this block operation increases the speed of computation, a repetitive element which locates outside of the block would not be seen. Although we cannot search all repetitive elements which are scattered in the whole genome, we can extract locally repetitive element efficiently.

At the second step, we should compute an expected period $\hat{T}$ in a local window. For the window sequence, we apply self autocorrelation function $s(t)$. This function has a higher value when a periodic signal exists. For instance, if a DNA sequence Y_n is a trinucleotide tandem repeat, then the sequence Y_{n+3} delayed by 3 from Y_n has the same sequence as Y_n. In this case, $s(t)$ will be maximized when $t = 3$. It is not possible that the multiples of the period will be another maxima. In the previous example, when t is selected as 6, then, $s(6) < s(3)$ because the summation's end point in $s(t)$ decreases.

In the next step, we need to identify if a repetitive sequence candidate is an actual repetitive sequence that we want to find. Some sequences can not be repeated purely, that is, the maximum value of $s(t)$ can be small, when the repeat is not periodic. Thus, a self autocorrelation threshold η determines whether the predicted period $\hat{T}$ is obtained from periodic DNA sequences where $s(\hat{T}) > \eta$. In addition, a filtering is necessary so that $\hat{T}$ satisfies $T_L < \hat{T} < T_H$ in order to classify the type of repetitive elements among the repeat elements.

Finally, the repeat unit and the characteristics of the locally repetitive elements are identified. From the predicted period $\hat{T}$, we construct a unit window R_n where $n = 1, 2, ..., \hat{T}$. The window sequence is a part of an original sequence, so $R_{k,n} = D_{n'}$, where $n' = k, ..., k + \hat{T} - 1$. The correlation between the window and the original sequence will be determined to find repetitive element candidates. Then, the repeat sequence unit $\hat{R}_n$ is the same as $R_{\hat{k},n}$ where

$$\hat{k} = arg\max_{k} \sum_{j=0}^{N-\hat{T}} \sum_{n=1}^{\hat{T}} I(R_{k,n\ mod\ \hat{T}}, D_{j+n}) \tag{5}$$

III. Results and Discussions

Spectral Repeat Finder(SRF) identifies the repetitive elements such as tandem and interspersed repeats based on Fourier Transform. The SRF finds repetitive elements using a mathematical analysis just as our proposed algorithm does. The comparison between this Discrete Fourier Transform(DFT) approach and our autocorrelation based approach will be presented. The SRF finds repetitive elements using a commonly defined DFT $S(f)$ as follows.

$$S(f) = \sum_{\alpha} \frac{1}{N^2} \left| \sum_{j=1}^{N} U_\alpha(j) e^{2\pi i f j} \right|^2 \tag{6}$$

Here, $U_\alpha(j)$ is 1 when DNA sequence $D_j = \alpha$, otherwise 0. f is a frequency that is measured by a reciprocal of the period or repeat unit's length. This spectral analysis can achieve clear result when each DNA sequence D_n is highly repeated sequences with low mutations. When the other sequence is mixed, the value of $S(f)$ will decrease. That's why an arbitrary window size makes it difficult to find the repetitive element as shown in Fig. 2. Since we don't know the length of this unknown repeat, it is dangerous to approximate the length of this repeat.

The self autocorrelation(SAC) method can give a solution to this unpredicted phenomenon. In a uncorrelated region, the effect of random sequences is removed since it has a mean of 0. As we just focus on the comparatively maximum value of a selected window size, we don't need to consider the other region that is not periodic. The length of the repeat in a periodic region can be identified clearly using the self autocorrelation function. The proposed scheme achieved much improved result in view of acquiring a period from the same sequence as shown in Fig. 3.

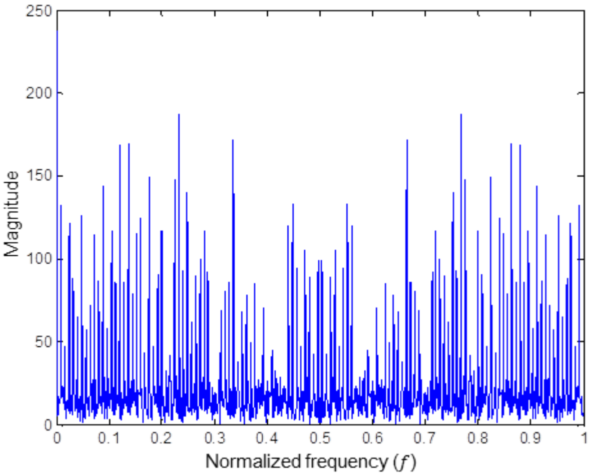

Figure 2. DFT result for Human Y chromosome(16,440 Kbase, W=1000)

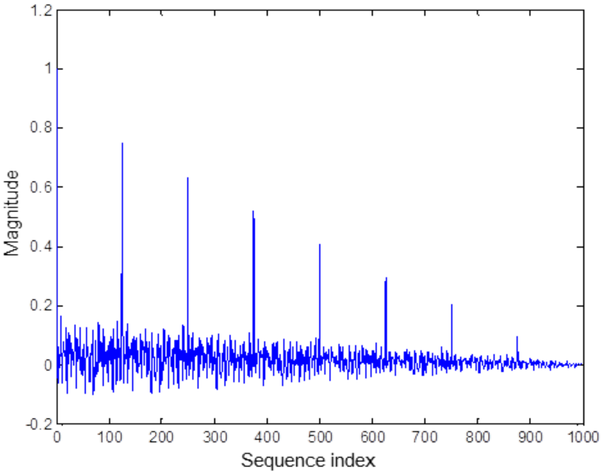

Figure 3. Self autocorrelation result of Human Y chromosome(16,440 Kbase, W=1000)

Now, the results of the SRF and the SAC for the same sequence are presented in Table. I. Our proposed scheme finds the largest pattern that contains all of these findings. As shown in Table. I, the smaller pattern is involved in a larger pattern. Using our algorithm, when we set the window to a small value to the corresponding block, we can find these small repeat patterns as well. It was shown that our proposed algorithm identifies more than 100 nucleotide bases while the conventional scheme doesn't.

From the NCBI database's human chromosome Y sequences, we extracted locally repetitive element data as shown in Fig. 4. When the window size increases, the number of searched patterns decreases at the same threshold value. As the window sizes gets bigger, the probability that locally repetitive genomic sequences do not appear gets larger due to randomness effect. That is to say, the locally repetitive

Table I
Searched Repeat Pattern of Human Y chromosome(16,440 Kbase)

	Searched Repeat Pattern	Period
SRF	TT	2
	TGTTC	5
	TCCAGA	6
	TGCAGACTT	9
	TATGTCCAGAGTTT	14
	TCCAGAGTTTCTTCCTTC	18
	TGTCCAGAGTTTCTTCCTTCT	21
	GTTGTGTCCAGAGTTTCTTCCTTCT	25
	TTGTGTCCAGAGTTTCTTCCTTCTG	25
SAC	TTTGTTCCTTCAGATGTGTCCAGATTTTCT TTCTTCTTGCAGTTTCATGGTCTTGCTCAC TTCAAGAATGAAGCTCCAGACCTTTACGGT GAGTTTTACAGCACTTAAATGTGTTATATC CAGAG	125

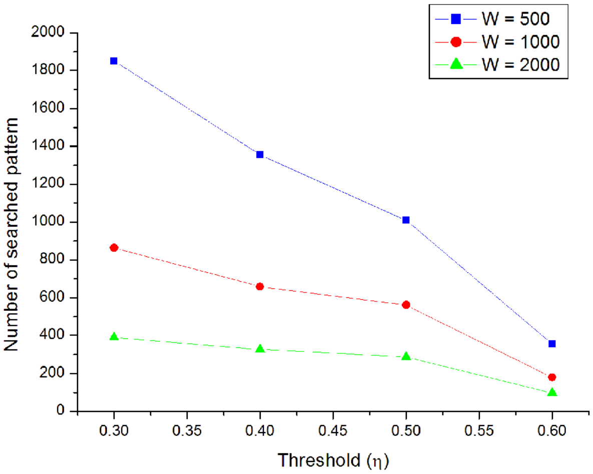

Figure 4. Number of locally repetitive elements vs threshold in Human Y chromosome

elements are located in a small region. We can easily imagine that the number of searched patterns increases with a smaller threshold η. Based on the result of our research, we can expand the coverage of our research by investigating the genome of various species.

The SAC finds more expansive pattern compared to the conventional repeat finder. The results from the SRF have redundancies which are duplicated, while the SAC finds the largest set of locally repeated pattern. Furthermore, by adapting various parameters here, we can acquire selective repeat patterns in the windows. When we extend our schemes to use a cross-correlation function, an interspersed repeat pattern can be derived. We expect that this fast genomic calculation can be applied to the next-generation genome-wide search of repetitive elements.

IV. Conclusions

We have proposed the self autocorrelation(SAC) method to find locally repetitive elements in genomic sequences. In DNA sequences, the locally repetitive element is defined as the repeat elements that are interspersed in a small range such as the unit length of a repeat. An autocorrelation based mapping function between the DNA sequences and the numerical value has been proposed to find the locally repetitive elements in the genomic sequences. The numerical result of our proposed scheme has shown a more clear and clever arrangement of repetitive elements. Our proposed method will give fast scanning of the whole genome sequence.

References

[1] W.-C. Kim, K.-H. Lee, K.-S. Shin, R.-N. You, Y.-K. Lee, K. Cho, and D.-H. Cho, “Reminer-ii: A tool for rapid identification and configuration of repetitive element arrays from large mammalian chromosomes as a single query,” *Genomics*, vol. 100, no. 3, pp. 131–140, Sep 2012.

[2] K.-H. Lee, S. Chiu, Y.-K. Lee, D. G. Greenhalgh, and K. Cho, “Age-dependent and tissue-specific structural changes in the c57bl/6j mouse genome,” *Experimental and Molecular Pathology*, vol. 93, no. 1, pp. 167–172, Dec. 2012.

[3] G. Benson, “Tandem repeats finder: a program to analyze dna sequences,” *Nucleic Acids Research*, vol. 27, no. 2, pp. 573–580, Nov 1999.

[4] Y. Wexler, Z. Yakhini, Y. Kashi, and D. Geiger, “Finding approximate tandem repeats in genomic sequences,” *Journal of Computational Biology*, vol. 12, no. 7, pp. 928–942, Oct 2005.

[5] V. Boeva, M. Regnier, D. Papatsenko, and V. Makeev, “Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression,” *Bioinformatics*, vol. 22, no. 6, pp. 676–684, Jan 2006.

[6] D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy, “Spectral repeat finder (srf): identification of repetitive sequences using fourier transformation,” *Bioinformatics*, vol. 20, no. 9, pp. 1405–1412, Feb 2004.

[7] D. Anastassiou, “Frequency-domain analysis of biomolecular sequences,” *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, Jul 2000.

[8] H. Herzel, O. Weiss, and E. N. Trifonov, “10-11 bp periodicities in complete genomes reflect protein structure and dna folding.” *Bioinformatics*, vol. 15, no. 3, pp. 187–193, Mar 1999.

[9] H. Herzel, E. Trifonov, O. Weiss, and I. Gro?, “Interpreting correlations in biosequences,” *Physica A: Statistical Mechanics and its Applications*, vol. 249, no. 1-4, pp. 449 – 459, Jan 1998.

[10] E. N. Trifonov and J. L. Sussman, “The pitch of chromatin dna is reflected in its nucleotide sequence,” *Proceedings of the National Academy of Sciences*, vol. 77, no. 7, pp. 3816–3820, Jul 1980.

[11] W. Li, “The study of correlation structures of dna sequences: a critical review,” *Computers and Chemistry*, vol. 21, no. 4, pp. 257–271, Mar 1997.