Systematic Analysis of Homologous Tandem Repeat Family in the Human Genome

Woo-Chan Kim(✉) and Dong-Ho Cho

Department of Electrical Engineering, Korea Advanced Institute
of Science and Technology, Daejeon, South Korea
wckim@comis.kaist.ac.kr, dhcho@kaist.ac.kr
http://umls.kaist.ac.kr

Abstract. The vast majority of the human genome consists of repetitive elements that form many complex but highly-ordered patterns. In particular, tandem repeats, whose repeat units are placed adjacent to each other, form highly structured patterns in the human genome when homologous tandem repeats are close together. Herein, the structure of the homologous tandem repeat family (HTRF) is assessed using systematic analysis. In the proposed system for analyzing HTRF, the original tandem repeat units are derived using the characteristics of homology of HTRF, and represented in a diagram in order to show the structure of HTRF easily. The analysis results of the four HTRFs in the human genome are shown here and the proposed algorithm may be seen to be very efficient for analyzing HTRF via the comparison of three conventional algorithms.

Keywords: Repetitive element · Tandem repeat · Homologous tandem repeat array · Systematic analysis · Human genome

1 Introduction

There are many repeated DNA sequences in the genome of most organisms, which is called *repetitive element*. The two major classes of repetitive elements are interspersed repeats and tandem repeats. Interspersed repeats are usually present as single copies and distributed widely throughout the genome, whereas tandem repeats are DNA sequences of which repeat units are placed next to each other. Although the functions of many repetitive elements have not yet been known, their impact and importance on genomes is evident. Mobile repeat elements have been a critical factor in gene evolution [1,2]. Also, some tandem repeats cause a number of genetic diseases [3] and they have been used as genetic markers for human identity testing [4]. Therefore, analyzing repetitive elements is very important and we study tandem repeats especially in this paper.

Tandem repeats are classified into three types, which are satellite, minisatellite, and microsatellite. Satellites form arrays of 1,000 to 10 million repeat units particularly in the heterochromatin of chromosomes. Minisatellite form arrays

M. Fernndez-Chimeno et al. (Eds.): BIOSTEC 2013, CCIS 452, pp. 162–175, 2014.
DOI: 10.1007/978-3-662-44485-6_12

of several hundred repeat units of 7 to 100 bp in length. They are present everywhere with an increasing concentration toward the telomeres. Microsatellites are composed of units of one to six nucleotides, repeated up to a length of 100 bp or more.

Although tandem repeats have been characterized by some features, which are the position in the genome, sequence, size, number of copies, and presence or absence of coding regions, there are much more complex tandem repeats in the human genome. In [5], the authors researched complex pattern structures of variable length tandem repeat (VLTR) and multi-periodic tandem repeat (MPTR). Also, our previous studies to find and visualize all repetitive elements in the genomes showed that the structure of the unknown as well as known repetitive elements is very complex but highly organized [6,7]. We, here, focus on the structure of multiple tandem repeats, which is called HTRF (Homologous Tandem Repeat Family).

HTRFs, which consist of multiple homologous tandem repeats dispersed throughout specific sequence regions, are abundant in the genomes of human and mouse [6,7]. Despite of lack of research of HTRF, we expect that HTRF plays an important role involving biological functions from its abundance and unique structure. Also, we can easily find a consensus tandem repeat unit of an HTRF array since two or more tandem repeats are homologous. By getting a consensus tandem repeat unit, we can find out how much the original HTRF are broken, which can be used as an evidence of the age of the genome.

We analyze four HTRFs from the human genome, which are chromosome 7 (57,937,500 – 58,056,406 bp), chromosome 8 (46,832,500 – 47,458,334 bp), chromosome 22 (16,505,625 – 16,627,187 bp), and chromosome Y (25,000 – 117,031 bp). The method for getting a consensus tandem repeat unit that are proposed in this paper is compared with the three conventional programs or algorithms, which are TRF (Tandem Repeat Finder) [8], SRF (Spectral Repeat Finder) [9], and tandem repeat detection using PT (Period Transform) [10,11]. TRF and SRF are the representative program for finding tandem repeat by using string matching algorithm and signal processing algorithm, respectively. A perfect HTRF is constructed by using the derived tandem repeat unit, and it is compared with the original HTRF. Also, the structure of an HTRF is shown in a diagram representation by using the consensus tandem repeat units.

2 System Modeling and Algorithm

The modeling of HTRF consists of three stages, which are *TR Extractor*, *TR Analyzer*, and *MTR Analyzer*. Figure 1 shows the system structure for analysis of HTRF. TR Extrator gets each tandem repeat from a given HTRF. The individual extracted tandem repeat is analyzed by TR Analyzer. The analysis results of TR Analyzer are the original tandem repeats as well as the properties of the individual tandem repeats such as repeat unit, number of repeat unit, and homology. MTR Analyzer, then, analyzes the relationships among the individual tandem repeats by using the results of TR Analyzer.

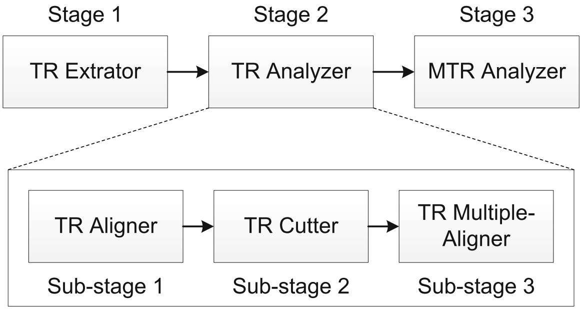


Fig. 1. System structure for analysis of HTRF.

2.1 TR Extractor

There are one or more tandem repeats that are homologous each other in an HTRF. We assume that there are N tandem repeats in an HTRF and each tandem repeat has n tandem repeat units whose length is l. Then, we can express i'th tandem repeat in an HTRF as $TR_i(n, l)$. TR Extractor divides each tandem repeat from an HTRF, which means that it gets all $TR_i(n, l)$ for the HTRF. However, TR Extrator defines tandem repeat when the number and the length of tandem repeat units is greater than δ_n and δ_l, respectively.

2.2 TR Analyzer

TR Analyzer derives a consensus tandem repeat from each tandem repeat in an HTRF. TR Analyzer consists of *TR Aligner*, *TR Cutter*, and *TR Multiple-Aligner*. These three sub-stages are iteratively processed for better performances.

TR Aligner. Two DNA sequence blocks which are a reference sequence and a target sequence are aligned, and the identity of the two sequence blocks are recorded in TR Aligner. If we assume that the length of the sequence block is B and a sequence block that begins from i'th nucleotide base of a tandem repeat is $S(i)$, the reference sequence of the firstly performed TR Aligner is generally $S(1)$. Also, the target sequence is moved from $S(1)$ to $S(x - B + 1)$ where x is the length of the tandem repeat.

The alignment of the two DNA sequences is conducted by dynamic algorithm or greedy algorithm [12]. The identity of the two sequences as a result of the alignment is recorded to $I(i, j)$ where i is the index of the reference sequence and j is the index of the target sequence. Since the reference sequence is fixed in TR Aligner, $I(i, j)$ is a function of variable j. Then, $I(i, j)$ of a perfect tandem repeat becomes 1 when $j = i + l \times k$ and $1 \leq i, j \leq x$ where k is an integer since same sequence blocks are arranged periodically in a perfect tandem repeat. The identity may have its peak point even if the tandem repeat is broken because it still has the attribute of the repetition of the tandem repeat. By using this characteristic of the identity of the tandem repeat, we can get the index of each tandem repeat unit.

TR Cutter. The identity function of a broken tandem repeat is generally fluctuated because its original perfect tandem repeat is randomly broken by biological phenomena such as insertion, deletion, and substitution. Thus, TR Cutter performs two processes to derive the index of each tandem repeat unit. First, TR Cutter makes the identity function be smoothed by averaging it locally. That is, a smoothed version of identity function $M(i,j)$ is defined as follows.

Algorithm 1. Recursive process of three sub-stages of TR Analyzer: TR Aligner, TR Cutter, and TR Multiple-Aligner.

```
procedure TR ALINER(broken_tandem_repeat)
    reference_sequence_index ← 0
    consensus_unit_index ← −1
    while reference_sequence_index ≠ consensus_unit_index do
        Calculate identity function                              ▷ TR Aligner
        Get all tandem repeat units                               ▷ TR Cutter
        Get a consensus unit                             ▷ TR Multiple-Aligner
        Substitute the index of the consensus unit to consensus_unit_index
    end while
    return consensus_unit_index
end procedure
```

$$M(i,j) = \frac{\sum_{k=j-\lfloor W/2 \rfloor}^{j+\lceil W/2 \rceil -1} I(i,k)}{W} \qquad (1)$$

where W is the smoothing window size.

The smoothing process removes the fluctuation of the identity function so that only the start points of real tandem repeat units have their local peak value of identity. The smaller the window size of the smoothing process is, the more peak points that are not the start of tandem repeat unit exist. However, too large window size of the smoothing process may remove the local peak point of a real tandem repeat unit. Therefore, the proper window size is required to leave only the local peak points of the real tandem repeat units in the smoothing process of the identity function.

Then, TR Cutter gets the start index of all the tandem repeat units by differentiating the function $M(i,j)$. The differentiated function $M'(i,j)$ of $M(i,j)$ is as follows.

$$M'(i,j) = M(i,j+1) - M(i,j). \qquad (2)$$

After calculating $M'(i,j)$, TR Cutter can find all the tandem repeat units by recording the points when $M'(i,j)$ is 0.

TR Multiple-Aligner. The tandem repeat units that are gotten by TR Cutter are aligned by TR Multiple-Aligner. By using the multiple sequence alignment of the tandem repeat units, TR Multiple-Aligner can get the consensus tandem repeat unit. A direct method of the multiple sequence alignment is the dynamic programming technique to identify the globally optimal alignment solution [13,14]. However, computational complexity of the direct method is basically too high, which takes $O(l^n)$ time where l and n are the average length and the number of tandem repeat units, respectively. Thus, we here use a suboptimal method that utilizes pairwise sequence alignment, which is similar to other suboptimal methods [15,16]. In our method, all pairwise sequence alignments between each pair of tandem repeat units are performed, and the tandem repeat unit that has the highest average alignment score with other tandem repeat units is chosen as the consensus tandem repeat unit. The proposed suboptimal method for finding the consensus tandem repeat unit takes $O((nl)^2)$ time, which reduces many computations compared with the dynamic programming technique particularly when l and n are large.

After the consensus tandem repeat unit is chosen, the three sub-stages of TR Analyzer, which are TR Aligner, TR Cutter, and TR Multiple-Aligner, are re-performed to get a more accurate consensus tandem repeat unit. The recursive process of TR Analyzer is conducted until the reference sequence is not changed in TR Aligner. Algorithm 1 shows the pseudo code of TR Analyzer.

2.3 MTR Analyzer

After TR Extractor divides all tandem repeats from the target HTRF and TR Analyzer derives the consensus tandem repeats of the individual tandem repeats, MTR Analyzer derives a consensus tandem repeat unit among all the tandem repeats. Sine the tandem repeats in an HTRF are highly homologous and are expected to be an identical tandem repeat originally, the consensus tandem repeat unit that is gotten from the multiple tandem repeats increases the reliability of the originality. Also, there are many reverse-complement directional homologous tandem repeats as well as forward directional homologous tandem repeats. Thus, we should not only consider the forward direction but also the reverse-complement direction of homology.

The derivation of the consensus tandem repeat unit is performed by multiple sequence alignment. Thus, we also apply the sub-optimal method of multiple sequence alignment that is used in TR Multiple-Aligner to the derivation of the consensus tandem repeat unit of HTRF for the purpose of reducing the computational complexity.

We can describe an HTRF as a diagram by using the derived consensus tandem repeat units. Figure 2 shows an example of a diagram of an HTRF. The HTRF shown in Fig. 2 has two different tandem repeat units and each tandem repeat appears twice. The tandem repeats made by the first tandem repeat unit are shown twice in forward and reverse directions, and then the tandem repeats made by the second tandem repeat unit are shown twice in only forward direction. Also, a region that is not a tandem repeat exists between the two

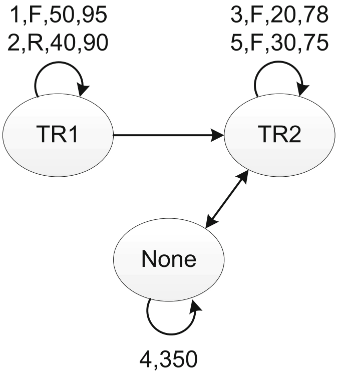


Fig. 2. Example of diagram representation of HTRF.

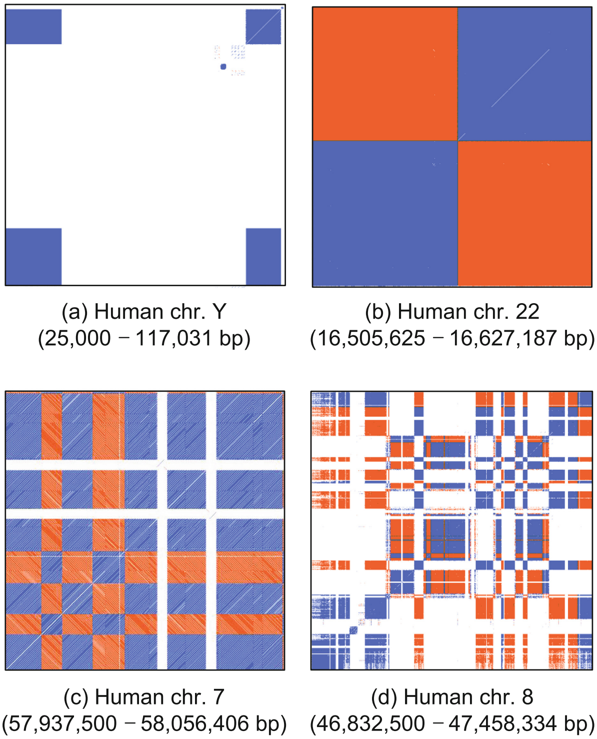


Fig. 3. Dot plot pattern of repetitive element arrays of the human genome.

tandem repeats made by the second tandem repeat unit. The four elements that are written above tandem repeat unit in the diagram are order, direction (F is forward direction and R is reverse direction), number of repetitions, and identity in percentage. Also, the two elements below *None* vertex is order and number of nucleotide bases.

Table 1. Consensus tandem repeat units of each MTRA.

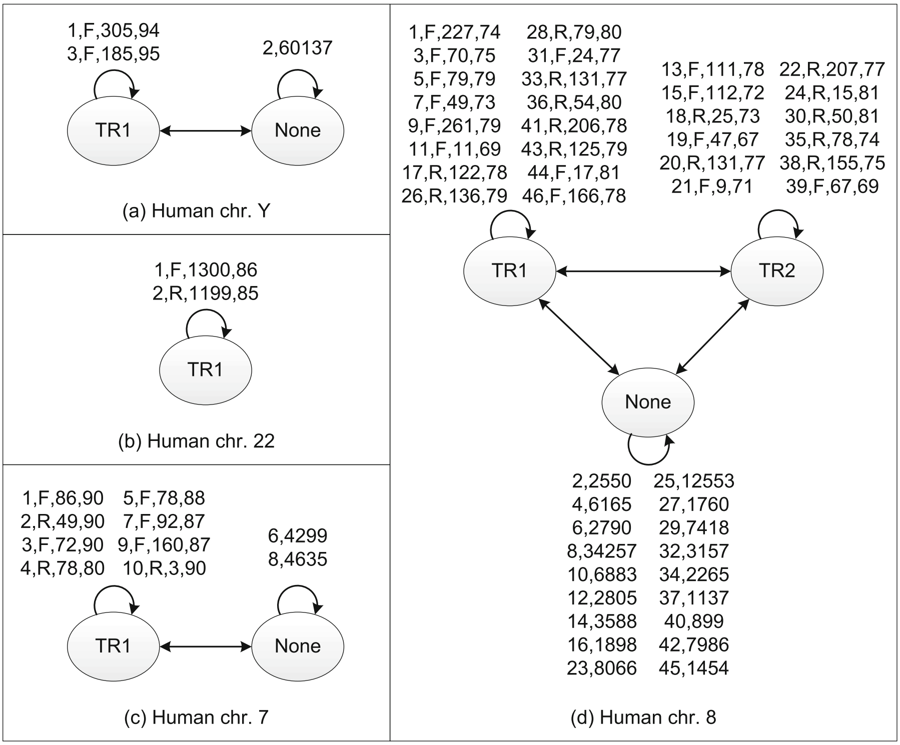
Human chromosome	Consensus tandem repeat unit	Average identity
Chr. Y	TR1: *TAGGTCTCATTGAGGACAGATAGAGAGCAGA*	0.95
	CTGTGCAACCTTTAGAGTCTGCATTGGGCC	
Chr. 22	TR1: *GCAGCAGTGTTCTGGAATCCTATG*	0.86
	TGAGGGACAAACACTCAGAACCCA	
Chr. 7	TR1: *TTCAACTCTGTGAGATGAATGCACACATC*	0.88
	ACAAAGAAGTTTCTCAGAATGCTTCTGTC	
	TAGTTTTTATGTGAAGATATTTCCTTTTC	
	CACCATAGGCCTCAAAGTGCTCCAAATG	
	TCCACTTGCAGATTCTACAAAAAGAGTG	
	TTTCAAAACTGCTCAATCAAAAGAAAGG	
Chr. 8	TR1: *CCCACTGAGGCCTATAGTGAAAAACTGAA*	0.76
	TATCCCATGATAAAAACTAGAAAGAAGCT	
	ATCTGTGAAACTGCTTTGTGATGTGTGCA	
	TTCAGCTCACAGAGTTAAACCTTTCTTT	
	TGATTCAGCAGGTTGGAAACACTCTTTT	
	TGTAGAATCTGCAAGGGGATATTTGGAG	
	TR2: *CCAAGGAGGCCTCTCCCATCCCAGAAGCCCC*	
	CAGGGCTGTCCCGGGCGGGCTGTAAAGCCCC	
	AGGCTTTGGAGCAGGGTGCCTGTGTCTCTCG	
	CAGAAGGCCCCCACAAGCGAAAACGGGGCCG	
	CAGGGTGGCGTGGGAGGGCCGCAGGGACTCA	
	GGGGGACGTTGAGGCAGGCAGAGGGGAGAAG	
	CGGCGAGACTGCAGGGAATGCTGGGAGCCTC	

3 Experimental Results

3.1 Analysis of HTRF of the Human Genome

We analyzed four HTRFs of the human genome by using the proposed system modeling. The analyzed HTRFs are chromosome 7 (57,937,500 – 58,056,406 bp), chromosome 8 (46,832,500 – 47,458,334 bp), chromosome 22 (16,505,625 – 16,627, 187 bp), and chromosome Y (25,000 – 117,031 bp). The human genome were obtained from the NCBI (National Center for Biotechnology Information) databases.

We first analyzed the repetitive elements and repetitive element arrays of the DNA sequences by using our analysis program, REMiner and REMiner Viewer [6,7]. The dot plot patterns of repetitive elements and repetitive element arrays of individual DNA sequences are shown in Fig. 3. According to the protocol of



(a) Human chr. Y

(b) Human chr. 22

(c) Human chr. 7

(d) Human chr. 8

Fig. 4. Diagram representation of HTRF of the human genome.

dot plot, a square is the pattern of a tandem repeat and a rectangle shows the relationship between two tandem repeats [6,7,17].

The HTRF of the human chromosome Y in Fig. 3 (a) has two tandem repeats and they are homologous directly, whereas the two tandem repeats of the HTRF of the human chromosome 22 are homologous inversely as shown in Fig. 3 (b). Also, there are many tandem repeats in the HTRF of the human chromosome 7 but they are all homologous directly or inversely as shown in Fig. 3 (c). This means that the tandem repeats all come from a same tandem repeat. In Fig. 3 (d), there are much more tandem repeats and they come from two original tandem repeats.

The consensus tandem repeat units are derived as results of our proposed analysis tool for HTRF. Table 1 shows the consensus tandem repeat units of the four target DNA sequences. The average identity is the mean values of the identity between the perfect tandem repeat that is made by the consensus tandem repeat unit and individual broken tandem repeat. The average identity shows the homology of each tandem repeat in the HTRF and the brokenness level of the HTRF, which can be graphically shown in Fig. 3.

Table 2. Comparison of consensus tandem repeat unit of proposed scheme and conventional schemes (chromosome Y and 22).

Human chromosome	Algorithm	Consensus tandem repeat unit	Average identity	Number of fails
Chr. Y	TR analyzer	*TAGGTCTCATTGAGGACAGAT*	0.95	0/2
		AGAGAGCAGACTGTGCAACC		
		TTTAGAGTCTGCATTGGGCC		
	TRF	*TAGGTCTCATTGAGGACAGAT*	0.95	0/2
		AGAGAGCAGACTGTGCAACC		
		TTTAGAGTCTGCATTGGGCC		
	SRF	*TAGGTCTCATTGAGGACAGAT*	0.95	0/2
		AGAGAGCAGACTGTGCAACC		
		TTTAGAGTCTGCATTGGGCC		
	PTF	*TAGGTCTCATTGAGGACAGAT*	0.95	0/2
		AGAGAGCAGACTGTGCAACC		
		TTTAGAGTCTGCATTGGGCC		
Chr. 22	TR analyzer	*GCAGCAGTGTTCTGGAATCCTATG*	0.86	0/2
		TGAGGGACAAACACTCAGAACCCA		
	TRF	*GCAGCAGTGTTCTGGAATCCTATG*	0.86	0/2
		TGAGGGACAAACACTCAGAACCCA		
	SRF	*GCAGCAGTGTTCTGGAATCCTATG*	0.86	0/2
		TGAGGGACAAACACTCAGAACCCA		
	PTF	·	·	2/2

Based on the consensus tandem repeat units, we describe each HTRF as a diagram in Fig. 4. By using the new representation of HTRF, we can easily see the overall structure of HTRF and the relationships among individual tandem repeats in HTRF. Furthermore, the original perfect tandem repeat array of the HTRF can be restored and the brokenness level of the HTRF can be calculated by using the consensus tandem repeat units.

3.2 Proposed Algorithm vs. Conventional Algorithm

There are many conventional algorithms that find tandem repeats although they did not consider multiple homologous tandem repeats simultaneously. Most of them can also derive the consensus tandem repeat unit of a tandem repeat. Thus, the conventional algorithms can be used to derive the consensus tandem repeat unit of a tandem repeat that is the function of TR Analyzer in our proposed system for the analysis of HTRF. In this subsection, TR Analyzer is compared with the representative conventional schemes that derive the consensus tandem repeat unit, which are TRF (Tandem Repeat Finder) [8], SRF (Spectral Repeat Finder) [9], and tandem repeat detection using PT (Period Transform) [10,11].

TRF is the representative program of string matching algorithms for finding tandem repeat. It uses pattern recognition criteria that is constructed statistically

Table 3. Comparison of consensus tandem repeat unit of proposed scheme and conventional schemes (chromosome 7).

Human chromosome	Algorithm	Consensus tandem repeat unit	Average identity	Number of fails
Chr. 7	TR analyzer	*TTCAACTCTGTGAGATGAATGC*	0.88	0/8
		ACACATCACAAAGAAGTTTCTC		
		AGAATGCTTCTGTCTAGTTTTT		
		ATGTGAAGATATTTCCTTTTC		
		CACCATAGGCCTCAAAGTGCT		
		CCAAATGTCCACTTGCAGATT		
		CTACAAAAAGAGTGTTTCAAA		
		ACTGCTCAATCAAAAGAAAGG		
	TRF	*TTCAACTCTGTGAGATGAATGC*	0.88	0/8
		ACACATCACAAAGAAGTTTGT		
		CAGAATGCTTCTGTCTAGTTT		
		TTATGTGAAGATATATTCTTT		
		TCCACCATAGGCCTCAAAGTG		
		CTCCAAATGTCCACTGCAGAT		
		TCTACAAAAAGAGTGTTTGAA		
		ATTGCTCAATCAAAAGAAATG		
	SRF	*TTCAACTCTGTGAGATGAATGC*	0.79	2/8
		ACACATCACAAAGAAGTTTCTC		
		AGAATGCTTCTGTCTAGTTTT		
		TATGTGAAGATATTTCCTTTT		
		CCACCATAGGCCTCAAAGCGC		
		TCCAAATGTCCACTTGCAGAT		
		TCTACAAAAAGAGTGTTTAAA		
		ACTGCTCAATCAAAAGAAAGG		
	PTF	*TTCAACTCTGTGAGGTGAATGC*	0.80	6/8
		ACATATCATAAAGAAGTTTGTC		
		AGAATGCTTCTGTCTAGTTTT		
		TATGTGAAGATATATCCTTTT		
		CCACCATAGGCCCCAAAGTGC		
		TCCAAATGTCCACTGCAGATT		
		CTATAAAAATAGTGTTTTAAA		
		ACTGCTCAATTAAAAGTAATG		

Table 4. Comparison of consensus tandem repeat unit of proposed scheme and conventional schemes (chromosome 8 - TR1).

Human chromosome		Consensus tandem repeat unit	Average identity	Number of fails
Chr. 8 (TR1)	TR analyzer	*CCCACTGAGGCCTATAGTGAAA*	0.77	90/16
		AACTGAATATCCCATGATAAAA		
		ACTAGAAAGAAGCTATCTGTGA		
		AACTGCTTTGTGATGTGTGCA		
		TTCAGCTCACAGAGTTAAACC		
		TTTCTTTTGATTCAGCAGGTT		
		GGAAACACTCTTTTTGTAGAA		
		TCTGCAAGGGGATATTTGGAG		
	TRF	*CGCTTTGAGGCCTATGGTGGAA*	0.78	0/16
		AAGGAAATATCTTCACATAAAA		
		ACTAGACAGAAGCATTCTCAGA		
		AACTTCTTTGTGATGTGTGCA		
		TTCAACTCACAGAGTTGAACC		
		TTCCTTTTGATAGAGCAGTTT		
		TGAAACACTCTTTTTGTAGAA		
		TCTGCAAGTGGATATTTGGAG		
	SRF	*CGCATTGAGGCCTATAGTGTAA*	0.73	0/16
		AACTGAATATCCAGTGATAAAA		
		ACAAGAGAGAAGCTATCTGTGA		
		ACCTGCTTAGTGATATGTGGAT		
		TCAGCTCACATAGTTAAACCTT		
		ACTTTTGATTCAGCTGTTTGTG		
		GAAACACTCTTTTTGTAAAAT		
		CTGCCAATAGACATTTCAAAG		
	PTF	*CCCCCAAAGGCCAAAAGTCAAA*	0.64	15/16
		ATCTGAATATCCCGTGAAAAAA		
		ACTATAAAGAAAATATCTGAGA		
		AAATACTTTGTGGTGTAAAGA		
		GTCATCTCAGAGAGTTAAAAC		
		TTTCTTTTGATAAAACAATTT		
		GAAAAAACTTTTTTGTAAAAT		
		CTCTGAAAGGTAATTTTAGAG		

Table 5. Comparison of consensus tandem repeat unit of proposed scheme and conventional schemes (chromosome 8 - TR2).

Human chromosome	Algorithm	Consensus tandem repeat unit	Average identity	Number of fails
Chr. 8 (TR2)	TR analyzer	*CCAAGGAGGCCTCTCCCATCCC*	0.76	0/12
		AGAAGCCCCCAGGGCTGTCCCG		
		GGCGGGCTGTAAAGCCCCAGGC		
		TTTGGAGCAGGGTGCCTGTGTC		
		TCTCGCAGAAGGCCCCCACAAG		
		CGAAAACGGGGCCGCAGGGTGG		
		CGTGGGAGGGCCGCAGGGACTC		
		AGGGGGACGTTGAGGCAGGCA		
		GAGGGGAGAAGCGGCGAGACT		
		GCAGGGAATGCTGGGAGCCTC		
	TRF	*CCAAGGAGGCCTCTCCCATCCCAG*	0.71	3/12
		AAGCCCCAGGGCTGTCCCAGGCAG		
		GCTGTAAAGCCCCAGGCTTTGGAG		
		CAGGGTGCCTGTGTCTCTCGCGGA		
		AGGCCCCACAAGCGAAAACGGGGT		
		CGCAGGGTGGCGTGGGCGGGTCAC		
		AGGGACTCAGGGGACATTGAGGCA		
		GGCAGAGGGGAGAAGCAGCAAGA		
		CAGCAGGGAATGCTGGGAGCCTC		
	SRF	*CCAGGAGGCCTCTCCCATCCCCGA*	0.69	2/12
		AGCCCTCAGGGCTGTCCCGGACTT		
		GGTGTAAAGCCCCAGGCTTTGGAG		
		CAGGGTGACTGTGTCTCTGGCGGA		
		AGGCCCTGACAAGCGAAAACGGGG		
		TAGCAGGGTGGCGTGGGCGGGTCA		
		TGGGGACTCAGCGGGACGTTGAGG		
		AAGGCCGAGGGGAGAAGCAGCAAG		
		AAAGCAGGGAGTGCTGGGAGCCTC		
	PTF	*TCAAGGAGGCCTCTCCCATTCCAG*	0.65	11/12
		AAGCCCCCAGGGCTGTTCCTGTTT		
		GATTGTAACTCTTCAGGCTTTGGA		
		TTAGGGTACCTGTGTCTCTGGTGG		
		AAGGGCCCCAAAAGCGAGACCCGG		
		GGGCAAGGTGGAAGGTGGCGGGGG		
		CAGGGACCCAGGGGAAAGCTGAGA		
		CAGGCGGAGGGGAGAAGTGGGAAG		
		ACCTCAGGCAATGCTGGGAGCCTT		

and it is the most widely used tool for identification of tandem repeat for its high accuracy. SRF is the representative program of signal processing algorithms to identify tandem repeat. It finds repetitions by converting the target DNA sequence from time domain to frequency domain using Fourier transform. The tandem repeat detection using PT, which is called *PTF* in this paper, is one of the algorithms for detecting tandem repeat based on signal processing. However, it does not use Fourier transforms but uses period transform to find repetitions.

We performed experiments to derive the consensus tandem repeat unit of each tandem repeat of the human genome by using the conventional schemes. Table 2 through Table 5 compare the results of our proposed scheme with those of three conventional schemes. Among the conventional schemes, TRF finds the most exact consensus tandem repeat unit in all the given tandem repeats, whereas PTF shows poor performance to detect consensus tandem repeat units. The *number of fails* in Table 2 through Table 5 means the number of the cases that a consensus tandem repeat is not detected because the given DNA sequence is not determined to be a tandem repeat. Thus, PTF is only usable to find the consensus tandem repeat units of the tandem repeats of human chromosome Y. This is because PTF does not consider the mutations of insertion and deletion of nucleotide bases. Although the performance of TRF is similar to TR Analyzer, TRF is inadequate to find the consensus tandem repeat unit of tandem repeats that are lengthy and highly broken like *TR2* of human chromosome 8 as shown in Table 5. Therefore, TR Analyzer is the most appropriate tool to derive the consensus tandem repeat unit to date, though there are many other tools that can be substituted (Tables 3 and 4).

4 Conclusions and Further Works

We proposed a system model for analyzing HTRF, which derives the consensus tandem repeat units based on the homology of the multiple tandem repeats and shows the structure of HTRF though a simple diagram representation. The proposed system model was performed on four HTRFs of the human genome, which are chromosome 7 (57,937,500 – 58,056,406 bp), chromosome 8 (46,832,500 – 47,458,334 bp), chromosome 22 (16,505,625 – 16,627,187 bp), and chromosome Y (25,000 – 117,031 bp). The algorithm for deriving a consensus tandem repeat unit of a tandem repeat in the proposed system model can be substituted by a conventional scheme that finds tandem repeat. However, in view of deriving an exact consensus tandem repeat unit, the experimental results showed that the proposed algorithm is the most appropriate for deriving a consensus tandem repeat unit to date.

The analysis of HTRF was performed based on the hypothesis that the homologous tandem repeats of an HTRF are originated from a same tandem repeat and HTRFs are very important to biological phenomenon. This hypothesis is sufficiently plausible considering the high identity of the homologous tandem repeats of an HTRF and their highly structured unique patterns. However, since the hypothesis should be verified biologically, we are going to perform the biological experiments of HTRF with the systematic analysis.

References

1. Kazazian, H.H.: Mobile elements: drivers of genome evolution. Sci. **303**, 1626–2632 (2004)
2. Prak, E.T., Kazazian, H.H.: Mobile elements and the human genome. Nat. Rev. Genet. **1**, 134–144 (2000)
3. Sinden, R.R.: Biological implications of the DNA structures associated with disease-causing triplet repeats. Am. J. Hum. Genet. **64**, 346–353 (1999)
4. Christian, M., Dennis, J., John, M.: STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. **29**, 320–322 (2001)
5. Hauth, A.M., Joseph, D.A.: Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinform. **18**, S31–S37 (2002)
6. Chung, B.I., Lee, K.H., Shin, K.S., Kim, W.C., Kwon, D.N., You, R.N., Lee, Y.K., Cho, K., Cho, D.H.: REMiner: a tool for unbiased mining and analysis of repetitive elements and their arrangement structures of large chromosomes. Genomics **98**, 381–389 (2011)
7. Kim, W.C., Lee, K.H., Shin, K.S., You, R.N., Lee, Y.K., Cho, K., Cho, D.H.: REMiner-II: A tool for rapid identification and configuration of repetitive element arrays from large mammalian chromosomes as a single query. Genomics **100**, 131–140 (2012)
8. Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**, 573–580 (1999)
9. Sharma, D., Issac, B., Raghava, G.P., Ramaswamy, R.: Spectral repeat finder (SRF): identification of repetitive sequences using fourier transformation. Bioinform. **20**, 1405–1412 (2004)
10. Buchner, M., Janjarasjitt, S.: Detection and visualization of tandem repeats in DNA sequences. IEEE Trans. Signal Process. **51**, 2280–2287 (2003)
11. Brodzik, A.K.: Quaternionic periodicity transform: an algebraic solution to the tandem repeat detection problem. Bioinform. **23**, 694–700 (2007)
12. Zhang, Z., Schwartz, S., Wagner, L., Miller, W.: A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7**, 203–214 (2000)
13. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. J. Comput. Biol. **1**, 337–348 (1994)
14. Just, W.: Computational complexity of multiple sequence alignment with SP-score. J. Comput. Biol. **8**, 615–623 (2001)
15. Humberto, C., David, L.: The multiple sequence alignment problem in biology. SIAM J. Appl. Math. **48**, 1073–1082 (1998)
16. Lipman, D.J., Altschul, S.F., Kececioglu, J.D.: A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. U.S.A. **86**, 4412–4415 (1989)
17. Edgar, R.C., Myers, E.W.: PILER: identification and classification of genomic repeats. Bioinform. **21**, i152–i158 (2005)