

Systematic Analysis of Structure of Multiple Tandem Repeat Arrays in the Human Genome

Woo-Chan Kim and Dong-Ho Cho

Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

Keywords: Repetitive Element, Tandem Repeat, Multiple Tandem Repeat Array, Systematic Analysis, Human Genome.

Abstract: Repetitive elements constitute the vast majority of the human genome and form many complex but highly-ordered patterns. Tandem repeats whose repeat units are placed next to each other particularly form very highly structured patterns in the human genome when homologous multiple tandem repeats are close together. In this paper, the structure of the multiple tandem repeat array (MTRA) is analyzed based on systematic analysis. The proposed system for analyzing MTRA derives the original tandem repeat units by using the characteristics of homology of MTRA and represents diagram model to show the structure of MTRA easily. The analysis results of the four MTRAs in the human genome are shown and the proposed algorithm is proved to be very efficient for analyzing MTRA by the comparison of three conventional algorithms.

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 (Kazazian, 2004; Prak and Kazazian, 2000). Also, some tandem repeats cause a number of genetic diseases (Sinden, 1999) and they have been used as genetic markers for human identity testing (Christian et al., 2001). 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 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 (Hauth and Joseph, 2002), 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 (Chung et al., 2011; Kim et al., 2012). We, here, focus on the structure of multiple tandem repeats, which is called MTRA (Multiple Tandem Repeat Array).

MTRAs, which consist of multiple homologous tandem repeats dispersed throughout specific sequence regions, are abundant in the genomes of human and mouse (Chung et al., 2011; Kim et al., 2012). Despite of lack of research of MTRA, we expect that MTRA 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 MTRA array since two or more tandem repeats are homologous. By getting a consensus tandem repeat unit, we can find out how much the original MTRA are broken, which can be used as an evidence of the age of the genome.

We analyze four MTRAs 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) (Benson, 1999), SRF (Spectral Repeat Finder) (Sharma et al., 2004), and tandem repeat detection using PT (Period Transform) (Buchner and Janjarasjitt, 2003; Brodzik, 2007). TRF and SRF are the representative program for finding tandem repeat by using string matching algorithm and signal processing algorithm, respectively. A perfect MTRA is constructed by using the derived tandem repeat unit, and it is compared with the original MTRA. Also, the structure of an MTRA is shown in a diagram representation by using the consensus tandem repeat units.

2 SYSTEM MODELING AND ALGORITHM

The modeling of MTRA consists of three stages, which are *TR Extractor*, *TR Analyzer*, and *MTR Analyzer*. Fig. 1 shows the system structure for analysis of MTRA. TR Extractor gets each tandem repeat from a given MTRA. 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.

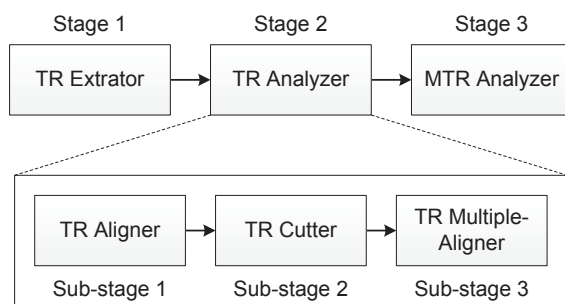


Figure 1: System structure for analysis of MTRA.

2.1 TR Extractor

There are one or more tandem repeats that are homologous each other in an MTRA. We assume that there are N tandem repeats in an MTRA and each tandem repeat has n tandem repeat units whose length is l .

Then, we can express i 'th tandem repeat in an MTRA as $TR_i(n, l)$. TR Extractor divides each tandem repeat from an MTRA, which means that it gets all $TR_i(n, l)$ for the MTRA. However, TR Extractor 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 MTRA. TR Analyzer consists of *TR Aligner*, *TR Cutter*, and *TR Multiple-Aligner*. These three sub-stages are iteratively processed for better performances.

2.2.1 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 (Zhang et al., 2000). 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 for 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.

2.2.2 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.

```

1: procedure TR ALINER(broken_tandem_repeat)
2:   reference_sequence_index ← 0
3:   consensus_unit_index ← -1
4:   while reference_sequence_index ≠ consensus_unit_index do
5:     Calculate identity function ▷ TR Aligner
6:     Get all tandem repeat units ▷ TR Cutter
7:     Get a consensus unit ▷ TR Multiple-Aligner
8:     Substitute the index of the consensus unit to consensus_unit_index
9:   end while
10:  return consensus_unit_index
11: end procedure

```

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

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.

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

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

2.2.3 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 (Wang and Jiang, 1994; Just, 2001). 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 sub-optimal method that utilizes pairwise sequence alignment, which is similar to other suboptimal methods (Humberto and David, 1998; Lipman et al., 1989). 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 MTRA 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. Since the tandem repeats in an MTRA 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 MTRA for the purpose of reducing the computational complexity.

We can describe an MTRA as a diagram by using the derived consensus tandem repeat units. Figure 2 shows an example of a diagram of an MTRA. The MTRA shown in Figure 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 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.

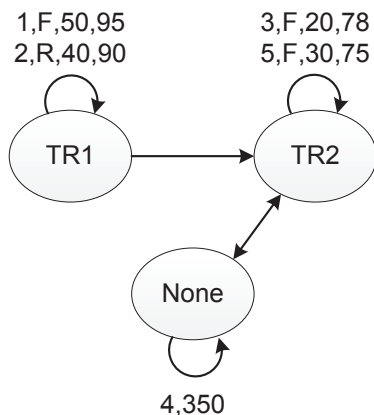


Figure 2: Example of diagram representation of MTRA.

3 EXPERIMENTAL RESULTS

3.1 Analysis of MTRA of the Human Genome

We analyzed four MTRAs of the human genome by using the proposed system modeling. The analyzed MTRAs 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 (Chung et al., 2011; Kim et al., 2012). The dot plot patterns of repetitive elements and repetitive element arrays of individual DNA sequences are shown in Figure 3. According to the protocol of dot plot, a square is the pattern of a tandem repeat and a rectangle shows the relationship between two tandem repeats (Chung et al., 2011; Kim et al., 2012; Edgar and Myers, 2005).

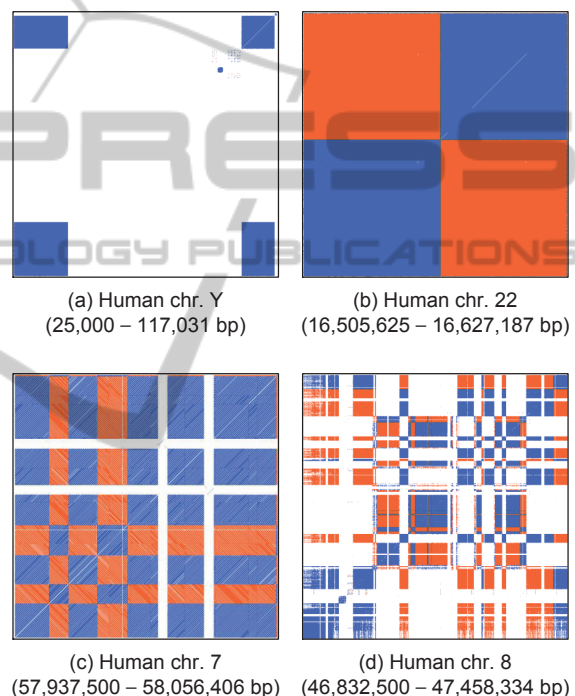


Figure 3: Dot plot pattern of repetitive element arrays of the human genome.

The MTRA of the human chromosome Y in Figure 3 (a) has two tandem repeats and they are homologous directly, whereas the two tandem repeats of the MTRA of the human chromosome 22 are homologous inversely as shown in Figure 3 (b). Also, there are many tandem repeats in the MTRA of the human chromosome 7 but they are all homologous directly or inversely as shown in Figure 3 (c). This means that the tandem repeats all come from a same tandem repeat. In Figure 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 MTRA. Table 1 shows the consensus tandem repeat units of the four target DNA sequences. The average identity is

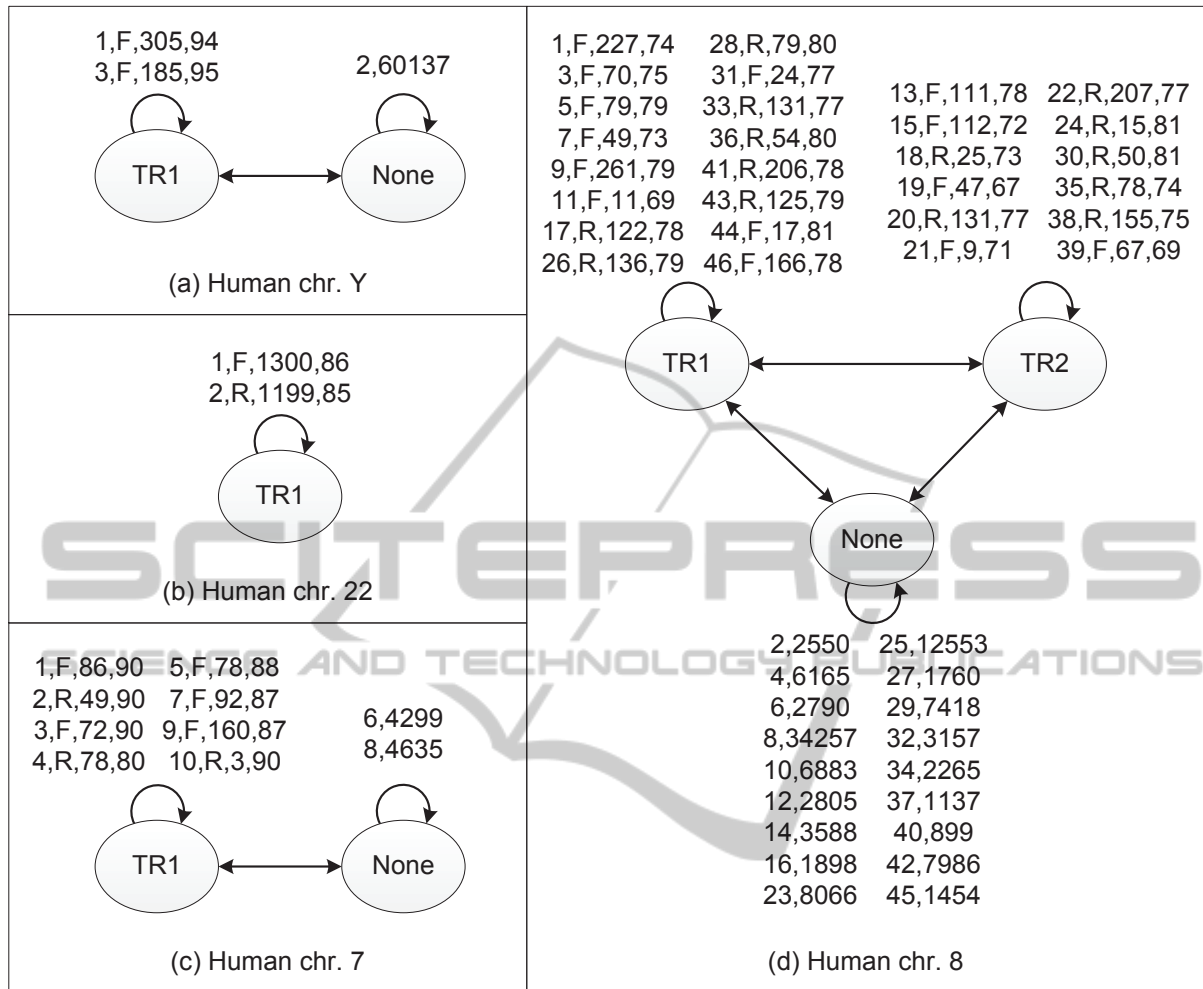


Figure 4: Diagram representation of MTRA of the human genome.

Table 1: Consensus tandem repeat units of each MTRA.

Human Chromosome	Consensus tandem repeat unit	Average identity
Chr. Y	TR1: TAGGTCTCATTGAGGACAGATAGAGAGCAGA CTGTGCAACCTTTAGAGTCTGCATTGGGCC	0.95
Chr. 22	TR1: GCAGCAGTGTTCTGGAATCCTATGTGAGGGACAAACACTCAGAACCCA	0.86
Chr. 7	TR1: TTCAACTCTGTGAGATGAATGCACACATCACAAAGAAGTTTCT CAGAATGCTTCTGTCTAGTTTTTATGTGAAGATATTTCTTTT CCACCATAGGCCTCAAAGTGCTCCAAATGTCCACTTGCAGATT CTACAAAAAGAGTGTTTCAAACACTGCTCAATCAAAAAGAAAGG	0.88
Chr. 8	TR1: CCCACTGAGGCCTATAGTGAAAACTGAATATCCCATGATAAA AACTAGAAAAGAAGCTATCTGTGAAACTGCTTTGTGATGTGTGC ATTCAGCTCACAGAGTTAAACCTTTCTTTTGATTCAGCAGGTT GGAAACACTCTTTTTGTAGAATCTGCAAGGGGATATTTGGAG TR2: CCAAGGAGGCCTCTCCCATCCCAGAAGCCCCAGGGCTGTCCCG GGCGGGCTGTAAAGCCCCAGGCTTTGGAGCAGGGTGCCTGTGTC TCTCGCAGAAGGCCCCCAAGCGAAAACGGGGCCGCAGGGTG GCGTGGGAGGGCCGCAGGGACTCAGGGGGACGTTGAGGCAGGC AGAGGGGAGAAGCGGCGAGACTGCAGGGAATGCTGGGAGCCTC	0.76

Table 2: Comparison of consensus tandem repeat unit of proposed scheme and conventional schemes (Chromosome Y, 22, and 7).

Human Chromosome	Algorithm	Consensus tandem repeat unit	Average identity	Number of fails
Chr. Y	TR Analyzer	TAGGTCTCATTGAGGACAGATAGAGAGCAGA CTGTGCAACCTTTAGAGTCTGCATTGGGCC	0.95	0/2
	TRF	TAGGTCTCATTGAGGACAGATAGAGAGCAGA CTGTGCAACCTTTAGAGTCTGCATTGGGCC	0.95	0/2
	SRF	TAGGTCTCATTGAGGACAGATAGAGAGCAGA CTGTGCAACCTTTAGAGTCTGCATTGGGCC	0.95	0/2
	PTF	TAGGTCTCATTGAGGACAGATAGAGAGCAGA CTGTGCAACCTTTAGAGTCTGCATTGGGCC	0.95	0/2
Chr. 22	TR Analyzer	GCAGCAGTGTCTGGAATCCTATG TGAGGGACAAACACTCAGAACCCA	0.86	0/2
	TRF	GCAGCAGTGTCTGGAATCCTATG TGAGGGACAAACACTCAGAACCCA	0.86	0/2
	SRF	GCAGCAGTGTCTGGAATCCTATG TGAGGGACAAACACTCAGAACCCA	0.86	0/2
	PTF	.	.	2/2
Chr. 7	TR Analyzer	TTCAACTCTGTGAGATGAATGCACACATC ACAAAGAAGTTTCTCAGAATGCTTCTGTC TAGTTTTTATGTGAAGATATTTCTTTT CACCATAGGCCTCAAAGTGCTCCAAATG TCCACTTGCAGATTCTACAAAAAGAGTG TTCAAAACTGCTCAATCAAAGAAAGG	0.88	0/8
	TRF	TTCAACTCTGTGAGATGAATGCACACATC ACAAAGAAGTTTGTGAGAATGCTTCTGT CTAGTTTTTATGTGAAGATATATCTTT TCCACCATAGGCCTCAAAGTGCTCCAAA TGTCCACTGCAGATTCTACAAAAAGAGT GTTTAAAATTGCTCAATCAAAGAAATG	0.88	0/8
	SRF	TTCAACTCTGTGAGATGAATGCACACATC ACAAAGAAGTTTCTCAGAATGCTTCTGTC TAGTTTTTATGTGAAGATATTTCTTTT CCACCATAGGCCTCAAAGCGCTCCAAAT GTCCACTTGCAGATTCTACAAAAAGAGT GTTTAAAACTGCTCAATCAAAGAAAGG	0.79	2/8
	PTF	TTCAACTCTGTGAGGTGAATGCACATATC ATAAAGAAGTTTGTGAGAATGCTTCTGTC TAGTTTTTATGTGAAGATATATCTTTT CCACCATAGGCCCAAAGTGCTCCAAAT GTCCACTGCAGATTCTATAAAAATAGTG TTTTAAAACTGCTCAATTAAAAGTAATG	0.80	6/8

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 MTRA and the brokenness level of the MTRA, which can be graphically shown in Figure 3.

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

sus 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 MTRA. In this subsection, TR Analyzer is compared with the

Table 3: Comparison of consensus tandem repeat unit of proposed scheme and conventional schemes (Chromosome 8).

Human Chromosome	Algorithm	Consensus tandem repeat unit	Average identity	Number of fails
Chr. 8 (TR1)	TR Analyzer	CCCACTGAGGCCTATAGTGAAAACTGAA TATCCCATGATAAAAACTAGAAAAGAAGCT ATCTGTGAAACTGCTTTGTGATGTGTGCA TTCAGCTCACAGAGTTAAACCTTTCTTT TGATTCAGCAGGTTGGAAACACTCTTTT TGTAGAATCTGCAAGGGGATATTTGGAG	0.77	0/16
	TRF	CGCTTTGAGGCCTATGGTGGAAAAGGAAA TATCTTCACATAAAAACTAGACAGAAGCA TTCTCAGAAAATCTTTGTGATGTGTGCA TTCAACTCACAGAGTTGAACCTTCCTTT TGATAGAGCAGTTTTGAAACACTCTTTT TGTAGAATCTGCAAGTGGATATTTGGAG	0.78	0/16
	SRF	CGCATTGAGGCCTATAGTGTA AAAACTGAA TATCCAGTGATAAAAAACAAGAGAGAAGCT ATCTGTGAACCTGCTTAGTGATATGTGGA TTCAGCTCACATAGTTAAACCTTACTTTT GATTCAGCTGTTTGTGGAAACACTCTTTT TGTA AAATCTGCCAATAGACATTTCAAAG	0.73	0/16
	PTF	CCCCCAAAGGCCAAAAGTCAAATCTGAA TATCCCGTAAAAAACTATAAAGAAAAT ATCTGAGAAAATACTTTGTGGTGTAAGA GTCATCTCAGAGAGTTAAACCTTTCTTT TGATAAAAACAATTTGAAAAACTTTTTT GTA AAATCTCTGAAAGGTAATTTTAGAG	0.64	15/16
Chr. 8 (TR2)	TR Analyzer	CCAAGGAGGCCTCTCCCATCCCAGAAGCCCC CAGGGCTGTCCCGGGCGGGCTGTAAAGCCCC AGGCTTTGGAGCAGGGTGCCTGTGTCTCTCG CAGAAGGCCCCCAACAAGCGAAAACGGGGCCG CAGGGTGGCGTGGGAGGGCCGCAGGGACTCA GGGGACGTTGAGGCAGGCAGAGGGGAGAAG CGGCGAGACTGCAGGGAATGCTGGGAGCCTC	0.76	0/12
	TRF	CCAAGGAGGCCTCTCCCATCCCAGAAGCCCC AGGGCTGTCCAGGCAGGCTGTAAAGCCCCA GGCTTTGGAGCAGGGTGCCTGTGTCTCTCGC GGAAGGCCCCACAAGCGAAAACGGGGTGCGA GGGTGGCGTGGGCGGGTCACAGGGACTCAG GGGACATTGAGGCAGGCAGAGGGGAGAAGC AGCAAGACAGCAGGGAATGCTGGGAGCCTC	0.71	3/12
	SRF	CCAGGAGGCCTCTCCCATCCCCGAAGCCCTC AGGGCTGTCCCGACTTGGTGTAAGCCCCA GGCTTTGGAGCAGGGTACTGTGTCTCTGGC GGAAGGCCCTGACAAGCGAAAACGGGGTAGC AGGGTGGCGTGGGCGGGTCATGGGGACTCAG CGGGACGTTGAGGAAGGCCGAGGGGAGAAGC AGCAAGAAAGCAGGGAGTGCTGGGAGCCTC	0.69	2/12
	PTF	TCAAGGAGGCCTCTCCCATTCCAGAAGCCCC CAGGGCTGTTCCTGTTTGATGTA ACTCTTC AGGCTTTGGATTAGGGTACCTGTGTCTCTGG TGGAAGGGCCCCAAAAGCGAGACCCGGGGGC AAGGTGGAAGGTGGCGGGGCAGGGACCCAG GGGAAAGCTGAGACAGGCGGAGGGGAGAAGT GGGAAGACCTCAGGCAATGCTGGGAGCCTT	0.65	11/12

representative conventional schemes that derive the consensus tandem repeat unit, which are TRF (Tandem Repeat Finder) (Benson, 1999), SRF (Spectral Repeat Finder) (Sharma et al., 2004), and tandem repeat detection using PT (Period Transform) (Buchner

and Janjarasjitt, 2003; Brodzik, 2007).

TRF is the representative program of string matching algorithms for finding tandem repeat. It uses pattern recognition criteria that is constructed statistically and it is the most widely used tool for identi-

fication 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 and Table 3 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 and Table 3 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 2 and Table 3. 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.

4 CONCLUSIONS AND FURTHER WORKS

We proposed a system model for analyzing MTRA, which derives the consensus tandem repeat units based on the homology of the multiple tandem repeats and shows the structure of MTRA through a simple diagram representation. The proposed system model was performed on four MTRAs 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 MTRA was performed based on the hypothesis that the homologous tandem repeats of an MTRA are originated from a same tandem repeat and MTRAs are very important to biological phenomenon. This hypothesis is sufficiently plausible considering the high identity of the homologous tandem repeats of an MTRA and their highly structured unique patterns. However, since the hypothesis should be verified biologically, we are going to perform the biological experiments of MTRA with the systematic analysis.

REFERENCES

- Benson, G. (1999). Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Research*, 27(2):573–580.
- Brodzick, A. (2007). Quaternionic periodicity transform: an algebraic solution to the tandem repeat detection problem. *Bioinformatics*, 23(6):694–700.
- Buchner, M. and Janjarasjitt, S. (2003). Detection and visualization of tandem repeats in dna sequences. *IEEE Transactions on Signal Processing*, 51(9):2280–2287.
- Christian, M., Dennis, J., and John, M. (2001). Strbase: a short tandem repeat dna database for the human identity testing community. *Nucleic Acids Research*, 29(1):320–322.
- Chung, B., Lee, K., Shin, K., Kim, W., Kwon, D., You, R., Lee, Y., Cho, K., and Cho, D. (2011). Reminer: a tool for unbiased mining and analysis of repetitive elements and their arrangement structures of large chromosomes. *Genomics*, 98(5):381–389.
- Edgar, R. and Myers, E. (2005). Piler: identification and classification of genomic repeats. *Bioinformatics*, 21(Suppl. 1):i152–i158.
- Hauth, A. and Joseph, D. (2002). Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*, 18(Suppl. 1):S31–S37.
- Humberto, C. and David, L. (1998). The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, 48(5):1073–1082.
- Just, W. (2001). Computational complexity of multiple sequence alignment with sp-score. *Journal of Computational Biology*, 8(6):615–623.
- Kazazian, H. (2004). Mobile elements: drivers of genome evolution. *Science*, 303(5664):1626–1632.
- Kim, W., Lee, K., Shin, K., You, R., Lee, Y., Cho, K., and Cho, D. (2012). Reminer-ii: A tool for rapid identification and configuration of repetitive element arrays from large mammalian chromosomes as a single query. *Genomics*, 100(3):131–140.
- Lipman, D., Altschul, S., and Kececioglu, J. (1989). A tool for multiple sequence alignment. *Proceedings of the*

- National Academy of Sciences of the United States of America*, 86(12):4412–4415.
- Prak, E. and Kazazian, H. (2000). Mobile elements and the human genome. *Nature Reviews Genetics*, 1(2):134–144.
- Sharma, D., Issac, B., Raghava, G., and Ramaswamy, R. (2004). Spectral repeat finder (srf): identification of repetitive sequences using fourier transformation. *Bioinformatics*, 20(9):1405–1412.
- Sinden, R. (1999). Biological implications of the dna structures associated with disease-causing triplet repeats. *American Journal of Human Genetics*, 64(2):346–353.
- Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348.
- Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning dna sequences. *Journal of Computational Biology*, 7(1-2):203–214.

