A Time and Space Efficient Heuristic Algorithm for Alignment of Sparse Biomarkers over Whole Genomes

정우근
Chung Woo-Keun
부산대학교 컴퓨터공학과
wkchung@pusan.ac.kr

ABSTRACT

It is not an easy task to align DNA sequences of whole genomes to compare useful biomarkers, as the whole genome usually has more than 100 mega bases for eukaryotes. Thus, we do not apply the straightforward Smith-Waterman $O(N^2)$ time and space algorithms. In this paper we propose another alignment problem that consists of only two kinds of symbols, biomarkers and other non-biomarker regions. Thus, the whole genome can be regarded as a binary string of '1'(biomarkers) and '0'(others). We propose an alignment algorithm for the simplified binary string to reveal the linear structural similarity of biomarkers over two different genomes. The time complexity of this algorithm is $O(M^2)$, where M denotes the number of biomarkers appearing in the genomes. We studied the structural similarities of all HERV(Human Endogenous RetroViruses) elements of four typical primates, Human, Chimpanzee, Orangutan and Rhesus monkey to show the usefulness of our algorithm. Our system successfully revealed the similar distribution of retro elements for the four primates.

KEYWORDS Sequence, Alignments, Biomarkers, Whole Genomes, Retro Elements.

1 Motivation

Sequence alignment, a classical problem in sequence analysis, is applied to diverse problems for biological information processing. Modified heuristic algorithms have been developed to overcome the computing time and space requirements due to the huge size of genomes for genome-wide applications. The pairwise alignment of genomes with weighted patterns has many applications, including original genome alignment and biomarker alignment. This problem is so different from the original genome alignment since the patterns can be sparse and the scoring of patterns are diverse. Therefore, more appropriate time-efficient and space-efficient algorithms can be developed. If we can extract some biological information by aligning a few markers or a set of short segment without considering whole sequence, it would be very efficient and desirable. For example GR-aligner only compares a series of short subsequences containing rearrangement events in genome scale[8].

If we have to compare a few specified biomarkers on a very long whole genome without considering other subsequences, we do not use a heavy and traditional optimal alignment, such as Smith-Waterman base. We need to develop another special alignment algorithm to deal with the long strings of only two types of symbols for meaningful biomarkers and other meaningless symbol strings. However, we do not delete the meaningless strings, since the in-between distance of two adjacent biomarkers is crucial to investigate biological meaning.

In order to show the excellence of our method, we will compare the distribution structure of Human Endogenous Retro Virus(HERV), since these could be some clues in understanding the regulations of gene functions[2, 10]. HERVs were a sequence fragment which is originated from virus in the evolution process. HERVs were first identified almost 30 years ago and are classified into the numerous families. So especially HERVs play an important key in evolutionary analysis[2]. In this paper, we want to compare the whole configuration of HERVs over several chromosomes of four typical primates, where the length of the whole chromosome is greater than 250 mega bases. But the current alignment tools can not be applied to compare HERVs due to the more than 300 mega long sequence. That is why we need to to develop another tool for aligning a few HERVs scattered over the 300 mega long genomes.

2 Related Work

Pairwise alignment algorithms of biological sequences have been developed since the Needleman/Wunsch's dynamic programming algorithm. Local alignment methods were developed following Smith and Waterman's dynamic programming method to find biological conserved patterns. FASTP / FASTN algorithm, which is a more practical heuristic algorithm to search huge sequence. It was subsequently improved to FASTA. The BLAST algorithm, which has been most commonly used by drastically improving computing time to almost linear time, searches seed patterns using query sequence W-mer automata and extends them.

There are two main issues in edit-distance based alignment problem. One is how to define the scoring matrix(weight matrix in alignment, which should be studied in the biological side. Another computational problem is how to align the very long sequences such as human genomes of more than 100 mega bases. So developing time/space efficient algorithm(or software) is the most important problem in bioinformatics and computational biology. Since the plain alignment needs $O(N^2)$ time and space complexity, it can not be used for the very long sequences. So lot of algorithms have been developed to reduce the computing memory required from quadratic space to linear space for sequence alignment[7].And Crochemore et al.

developed a sub-quadratic sequence alignment algorithm for unrestricted cost matrices by dividing the dynamic programming algorithm into variable sized blocks as induced Lempel-Ziv parsing.[3, 8] Recently, a new alignment algorithm for a run-length-encoded string with M runs and an uncompressed string of length n was developed with the complexity $O(M \cdot n)$[16]. But this algorithm can not be applied to biological string without modification, since it is hard to transform a biological string into a run-length codes.

Recently according to the super-fast sequencing machine, it is available to get the whole genome. So we have to devise another alignment algorithm to deal with the whole genome. There are a few alignment tools devised only for the whole genome string. AVID algorithm[6], for pairwise alignment of huge sequences, determines matches using suffix trees and selects anchor. The EMAGEN algorithm[13], which aligns closely related multiple whole genomes, uses combinations of suffix arrays, graph theoretical formulation and ClustalW. The SGA algorithm, which aligns two genomes at high speed, is a grammar-based method based on the Yang-Kieffer algorithm[3]. And GRAT[14] have been developed to map a huge set of short sequences on genome sequences at high speed. We recently found an optimal alignment(mapping) algorithm for scaffolding and validation of bacterial genome assemblies without considering the whole sequences[17]

3 Fast Alignment for Biomarkers

In this section, we specify an alternative alignment algorithm for binary string with a constrained match/mismatch weight matrix. Many alignment algorithms had been announced[7]. Generally, the common alignment(global and local alignment) can be easily optimized by the simple dynamic programming approach.

Alignment algorithms can be classified into two classes; optimal alignments and heuristic alignments. The basic Smith-Waterman $O(N^2)$ time and space algorithm is inefficient for large strings such as whole genome scale. For example if the input string is greater than multi-mega bases, then the straightforward implementation for optimal alignment can not complete within a reasonable time. Many heuristic alignment algorithms over come these problems, BLAST is a famous well-known heuristic. There have been a few specialized alignment algorithms for restricted input cases[11, 13, 3]. Here, we are very interested in run-length code alignment[16]. Their algorithm computes optimal alignment in $O(M \cdot n)$ for an uncompressed string of length n and a compressed string with M runs. This is a great improvement compared with plain $O(MN)$ algorithm.

In this paper we are concerned with aligning BBS (Biological Binary Sequence). Each BBS string consists of only two symbols $\{0, M_i\}$. For an alignment of two BBS strings, $S_a, S_b \in BBS$, the matching score is defined as $match(M_i \ , M_j) = +K_{i,j} > 0$ and $match(0,0) = +k > 0$, where $K > k$. The penalty score for a mismatching is defined as $match(0, M_i) = -p \cdot |M_i| < 0$, $match(0, gap) = -q < 0$, $match(M_i, gap) = -q \cdot |M_i| < 0$, where $p, q > 0$. The BBS alignment problem is to determine the optimal global alignment configuration with the a $match()$ function satisfying the above constraints. In this paper, the gap insertion penalty is $-q$ regardless of marker M_i or 0.

Definition 1 *Let $S \in BBS$. Then $S = B_1 \cdot M_1 \cdot B_2 \cdot M_2 \cdot B_3 \cdot M_3 \ldots \cdot B_n \cdot M_n \cdot B_{n+1}$, where B_i denotes a 'blank' (corresponding to non-biomarkers) substring and M_i denotes a 'biomarker' in the i-th order. $|B_i|$ and $|M_i|$ represents the size(string) of B_i and M_j, respectively.* □

Thus, a whole genome can be simplified into a shorter BBS. In the following, we assume that K is sufficiently greater than k to make the problem simple; that is, $K > n \cdot k$ for any integer n. Thus the alignment score can be reduced in aligning the following meta-symbols $match(B_i, B_j)$,$match(M_i, M_j)$. Since $match(0,0) = k$ and $match(0, gap) = -p$, so we derive the following.

$$match(B_i, B_j) = k \cdot \min\{|B_i|, |B_j|\} - p \cdot ||B_i| - |B_j||$$

$match(M_i, M_j)$ should be determined in the biological context. For example if M_i is HERV, then the alignment score of two HERV sequences would be used. Let us consider all mismatch scores. Since B_i is totally different to M_j, $match(B_i, M_j)$ should contain at least $\min\{|B_i|, |M_j|\}$ mismatches.

The optimal alignment of $match(B_i, M_j)$ is dependent on p and q.This can be simply calculated. If we align B_i to M_j with the overlapping interval of length L, then the $match(B_i, M_j) = L \cdot p - q \cdot \{(|B_i| - L) + (|M_j| - L)\}$. So any arbitrary p and q, we can find the optimal L maximizing $L \cdot p - q \cdot \{(|B_i| - L) + (|M_j| - L)\}$. But in this paper we simply assumed that $q(gappenalty)$ is quite greater than $p(mismatchpenalty)$, so the smaller one of B_i and M_j) should be aligned in the middle of the larger one to make the best alignment. In a similar way, we can easily get $match(B_i, gap) = |B_i| \cdot match(0, gap)$ and $match(M_i, gap) = |M_i| \cdot match(0, gap)$ by definition.

The genome string alignment problem is reduced to an alignment of simplified BBS string consisting of B_i and M_j with another matching matrix for B_i, M_j and gap symbols. Thus, we do not describe the plain global alignment procedure based on a dynamic programming approach.

Now we consider the time complexity of $S_a, S_b \in BBS$ alignment. Let n be the number of M_i contained in a BBS. Then the size of the simplified BBS string is equal to $O(2 \cdot n)$, so the time and space complexity of BBS alignment is $O(4n^2)$. This is much less than $O(|S_a| \cdot |S_b|)$.

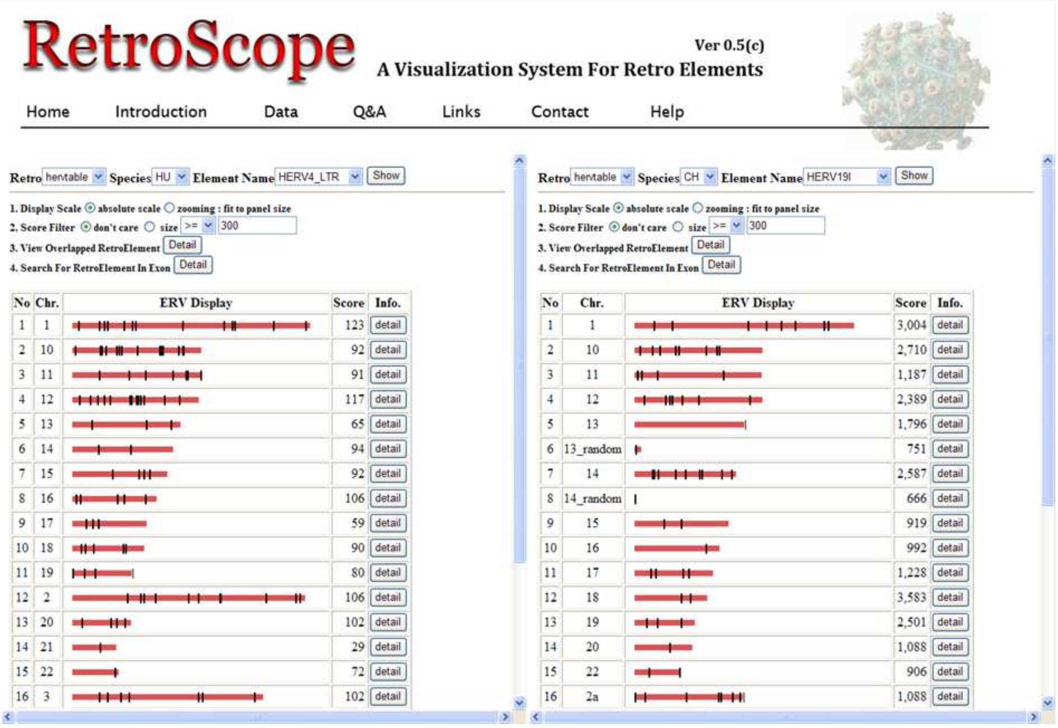
4 Experiments

4.1 Data Preparation

This section shows that our alignment works successfully in comparing the linear structure of HERV elements on four primates genomes including Human, Chimpanzee, Orangutan and Rhesus monkey. It was recently reported that the Endogenous RetroVirus (ERV) elements have been shown to contribute the promoter sequence to interfere with the transcription of adjacent genes[10]. Thus, it is interesting the physical proximity of ERV and genes on a whole genome scale. In addition, other researchers have studied how to cluster HERVs to find the mutual relationships to determine the phylogenetic taxonomies using a median self-organizing map[20].

One of main reasons of studying the identification and characterization of HERV is suggest the most-likely evolutionary process of species. Especially if we are interested in investigating the evolutionary process of primates, it is crucial to study the distribution of HERV variants over multiple primate genomes. We implemented RETROSCOPE[1] in previous work. It is a web-based visualization and DB system to search for HERV and ALU elements over whole genome scale. You can easily identify the physical location of each HERV elements over four primate genomes. In addition, you can investigate and visualize the information of each HERV, including the number and location.

We obtained the complete set $\{H_i\}$ of HERV sequences from the Genetic Information Research Institute (GIRI), http://www.girinst.org/repbase/index.html to prepare test data. Then we searched all homologous subsequences of H_i over whole genomes released by the UCSC genome center. Next, we collected a part of the homologous sequences whose scores were greater than a threshold(BLAT score 100). In this process we applied BLAT to determine the homologous sequences for a query HERV sequence. The collected HERV statistics appear in the following table.The Alu family is a family of repetitive elements in the Human genome[5]. Alu sequences are about 300 base pairs long and are therefore classified as short interspersed elements (SINEs) amongst the class of repetitive DNA elements. As a kind of retro elements, Alu is becoming important in the study of transposon mechanism[5]. The following table shows the number of HERVs, Alu and SVA(SINE-VNTR-Alu) in each species[1].



Retro hervtable Species HU Element Name HERV4_LTR

1. Display Scale ● absolute scale ○ zooming : fit to panel size
2. Score Filter ● don't care ○ size >= 300
3. View Overlapped RetroElement Detail
4. Search For RetroElement In Exon Detail

No	Chr.	ERV Display	Score	Info.
1	1		123	detail
2	10		92	detail
3	11		91	detail
4	12		117	detail
5	13		65	detail
6	14		94	detail
7	15		92	detail
8	16		106	detail
9	17		59	detail
10	18		90	detail
11	19		80	detail
12	2		106	detail
13	20		102	detail
14	21		29	detail
15	22		72	detail
16	3		102	detail

Retro hervtable Species CH Element Name HERV19I

1. Display Scale ● absolute scale ○ zooming : fit to panel size
2. Score Filter ● don't care ○ size >= 300
3. View Overlapped RetroElement Detail
4. Search For RetroElement In Exon Detail

No	Chr.	ERV Display	Score	Info.
1	1		3,004	detail
2	10		2,710	detail
3	11		1,187	detail
4	12		2,389	detail
5	13		1,796	detail
6	13_random		751	detail
7	14		2,587	detail
8	14_random		666	detail
9	15		919	detail
10	16		992	detail
11	17		1,228	detail
12	18		3,583	detail
13	19		2,501	detail
14	20		1,088	detail
15	22		906	detail
16	2a		1,088	detail

그림 1. RetroScope: A Visualization System for Retro Elements over Whole Genome. The user can select up to five HERV elements to see their relative positions. Each small black dot-bar represents the HERV. In this Figure, we show the location of HERV4_LTR for the Human and Chimpanzee[1]. RETROSCOPE is available at http://neobio.cs.pusan.ac.kr/∽retroscope/

	HU.	CH.	OR.	RH.
HERV	51	50	51	51
Alu	52	52	53	51
SVA	6	6	6	6

4.2 Linear Similarity of HERV Distributions on Four Primates

We believe the comparison of the linear configuration of HERV distribution gives good clues to understand phylogeny. We want to identify the most “similar” configuration of the HERV layout on a one-dimensional linear genome scale. We need to align HERVs are distributed on the whole genome using the proposed algorithm. First, we need to get the corresponding binary sequence. The whole DNA sequence of a chromosome is transformed into a binary string of $\{0, H_i\}$. H_i denotes each HERV element, whose size is denoted $|H_i|$. This is defined as the number of DNA bases of H_i. With the exception of all H_i, each base is transformed into ’0’ and each non-HERV DNA region between a pair of adjacent HERVs is treated as a symbol with its length in our algorithm.

Now we need to explain how to construct the scoring matrix for these binary $\{0, H_i\}$ strings. Since it is important to get pairs of matched HERVs, the match score of H_i is sufficiently greater than that of any number of consecutive '0' matches. Here, we give a scoring matrix $M[\ ][\ ]$ for our meta-symbol string as follows.

M[ ][ ]	H_j	'0'	gap
H_i	$+K_{i,j}$	$-p$	$-q$
'0'	$-p$	$+k$	$-q$
gap	$-q$	$-q$	

Note that $K_{i,j} \gg k > 0$, which means $K_{i,j} \gg N \cdot k$ for the size N of the non-HERV DNA region. We simply find that $K_{i,j} = \max\{|C_1|, |C_2|\}$ in comparing two chromosome C_1 and C_2. Thus, in our HERV alignment the score due to matched HERVs dominates the alignment score. The mismatch and gap penalty constants p and q are arbitrarily determined depending on the application objective in our binary alignment. We set $K_{i,j} \geq 300,000,000$ in the experiment, since the size of largest chromosome of primates is less than 300 mega bases. p, q should be sufficiently small not to cause the total alignment score to become negative. We empirically determine $p = 0.1$, $q = 0.001$ in this experiment.

Using this $M[\ ][\ ]$, we tried to find the most *"similar"* HERV configuration between Human and – HU., CH. , OR., RH. ˝. 4.2 shows that the most similar chromosome to Human chromosome No.3 is chromosome No.10 of the Chimpanzee with respect to HERV4_LTR measured by the scoring matrix $M[\ ][\ ]$. Human chromosome No.6 is similar to chromosome No.12 of Orangutan with respect to HERV4_LTR element. In this experiment, we considered the HERV elements with a score of more than BLAT 300 points.

First, Fig.2 shows the best alignment between all chromosomes in the four primates with respect to HERV=*HERV4_LTR*. Our algorithm clearly showed that Hu.1 is very similar to CH.1.

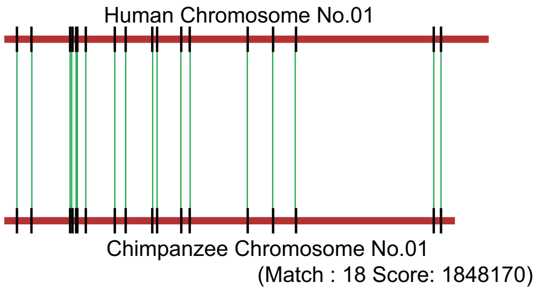


그림 2. The alignment of Human chromosome No.1 versus CH. No.1, the alignment with the highest score among all pairs of chromosomes.

Next, the four different alignment results of HERV4_LTR of Human chromosome No.3 are shown. In this alignment we allocate a lower penalty for gap insertion than for the mismatch penalty. The order of good alignment is from CH.10(the best alignment) > RH.3 > OR.12 > CH.6(the worst alignment). First, we give a brief description of the test chromosomes holding HERV4_LTR in following Table2.

We show four different(the best four) alignments based on HERV4_LTR in Human Ch.3 in Fig.3. As was shown in Fig.3, our HERV alignments mostly prefer the number of paired HERVs, discouraging unmatched HERV remaining and gap insertion. Though the alignment (b) with RH.3 seems to be better than

표 1. The Result of Fast Binary-String Alignment between Human and other three primates – CH. , OR., RH. ″. with respect to the HERV4_LTR element

Human Ch.	Species	Ch. No.	Matched HERV	Score
1	CH.	1	19	1,848,170
3	CH.	10	10	891,376
4	RH.	3	10	907,213
5	CH.	5	11	997,452
6	OR.	12	10	925,694
8	OR.	14	5	408,014
9	OR.	17	4	358,725
10	CH.	1	18	1,583,280
11	RH.	13	8	768,809
12	RH.	3	14	1,273,280
13	CH.	13	4	367,219
14	OR.	14	3	267,596
16	OR.	6	6	563,425
17	CH.	9	4	375,165
18	OR.	6	7	646,843
19	RH.	20	5	446,640
20	RH.	10	5	470,855
21	CH.	22	2	194,468
22	CH.	8	2	199,204

표 2. Data description for test chromosomes including HERV4_LTR

Species	Ch.No.	Size(base)	Copies
HU	3	199,501,827	9
RH	3	196,418,989	14
CH	10	135,001,995	9
CH	6	173,908,612	12
OR	12	136,387,465	10

(a) due to the parallel alignment lines(green), it does not have a better score, since one and three HERVs remained unmatched on the right and left sides, respectively.

We tried to find the best match with respect to HERV4_LTR of all chromosomes in Human, Chimpanzee and Orangutan. Fig.4 shows the most 'similar' pairs of chromosomal structures with respect to HERV4_LTR, denoted by the green solid edges, where we do not illustrate 'weak' pairs whose match scores are less than a threshold value.

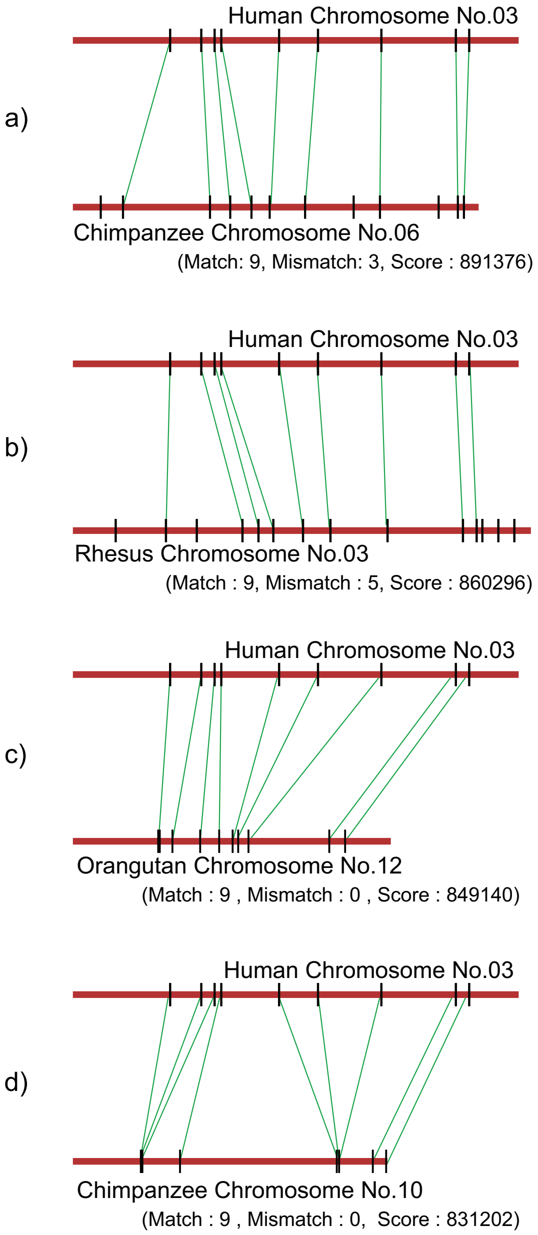
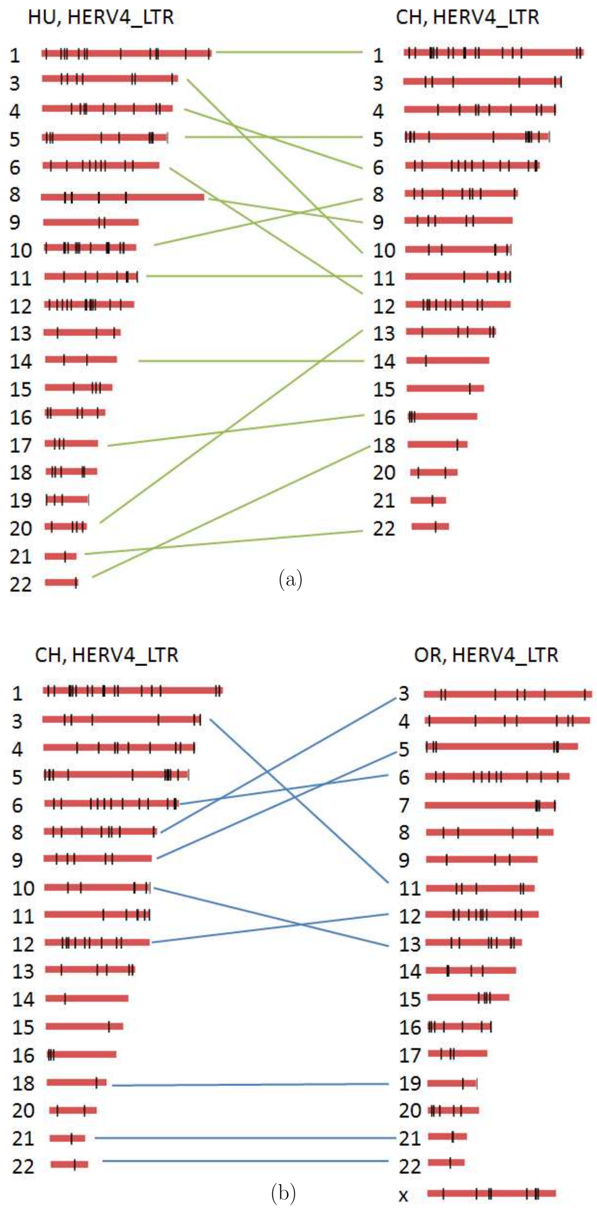

a)
Human Chromosome No.03
Chimpanzee Chromosome No.06
(Match: 9, Mismatch: 3, Score : 891376)
b)
Human Chromosome No.03
Rhesus Chromosome No.03
(Match : 9, Mismatch : 5, Score : 860296)
c)
Human Chromosome No.03
Orangutan Chromosome No.12
(Match : 9 , Mismatch : 0 , Score : 849140)
d)
Human Chromosome No.03
Chimpanzee Chromosome No.10
(Match : 9 , Mismatch : 0, Score : 831202)

그림 3. Four best alignment HERV4_LTR in Human Chromosome No.3 to all other chromosomes of CH., OR. and RG. (a) A comparative view of best match between HU.3 and CH.10. The green edge denotes the matched pair of HERV elements on both sides. (b)(c) and (d) the second, third and fourth best alignment for HERV4_LTR in Human Chromosome No.3. The mismatch value denotes the number of HERV copies unselected in our alignment. In this experiment, we set $p = 0.001$ for mismatch penalty and $q = 0.01$ for gap penalty.

5 Conclusion and Future Work

Sequence alignment is a very important and fundamental problem in many analysis methods for genomic studies. However, the alignment of a whole genome is problematic compared with the simple and short



(a)

(b)

그림 4. Pairs of the matched chromosomes of (a) Human and Chimpanzee, (b) Chimpanzee and Orangutan with respect of HERV4_LTR distribution. We considered all HERVs with BLAT homologous score greater than 300.

alignment, since its complexity in time and space is quite high. It is nearly impossible to find the globally and locally optimal alignment over more than mega-base size for whole genomes. Thus, until now, many of heuristic or sub-optimal alignment algorithms have been introduced. In addition, a few specialized alignments algorithm were studied in constrained input cases, such as run-length encoded string or string of regular expression.

In this paper, we propose a simple and fast alignment algorithm that is specially devised for very long binary strings. The main contribution of our paper is as follows.

- We developed $O(P \cdot Q)$ time and space alignment for a arbitrary long biological binary strings, where P and Q are the number of '1'(biomarker) symbols appearing in two strings. These '1' symbols can be regarded as biologically meaningful biomarkers, such as HERV, Alu or SNP elements.
- One requirement of our alignment is that the weight matrix should be constrained. The match value for the '1' symbol should greater than that of an arbitrary length of '0' strings.
- We tried to determine similar chromosomal linear structures with respect to HERV distribution on whole genomes of four primates. An experiment showed that our algorithm is appropriate to identify the similar configuration of HERVs in terms of relative position and in-between distance(e.g., the length of '0' strings between two consecutive '1' symbols.
- By controlling the gap penalty constants p, q, we can obtain more flexible alignment results for diverse applications. Thus, if we prefer to conserve the inter-distance between a pair of adjacent biomarkers, then we should increase the gap penalty.

Though our binary alignment problem can be solved by previous methods[16], our complexity is relatively low, since our method only depends on the number of '1' symbols, not on the chromosomal size. The sparser the biomarker is, the better the algorithm's performance. The length of biomarkers can be very diverse and the normalization step could help to represent the biological meaning of match score more succinctly. Much biological research is based on genome-wide work and aims to discover additional biomarkers. We could determine more applications for the algorithm and will modify/improve it to suit specific requirements.

Currently, we are developing a more accurate alignment procedure that considers the score of homologous, $K_{i,j}$, between H_i and H_j. In this paper we only set a constant $c = 300,000,000$ to create the simple and fast code. We do not have a clear idea how to set the gap penalty and mismatch penalty. These should be carefully determined on the biological meaning of the final alignment result. We will devise a general alignment algorithm for binary strings without a strong $K \gg k$ constraint in the match/mismatch scoring matrix.

참고 문헌

1. Retroscope: A web-based visualization system for retro element, http://neobio.cs.pusan.ac.kr/∽retroscope/.
2. Identification and characterization of novel human endogenous retroviral sequences prefentially expressed in undifferentiated embryonal carcinoma cells. *Nucl. Acids Res.*, 19(7):1513–1520, 1991.
3. Sga: A grammar-based alignment algorithm. *Computer Methods and Programs in Biomedicine*, 86(1):17 – 20, 2007.
4. S. Aftab, L. Semenec, J. Chu, and N. Chen. Identification and characterization of novel human tissue-specific rfx transcription factors. *BMC Evolutionary Biology*, 8(1):226, 2008.
5. J. A. Bailey, G. Liu, and E. E. Eichler. An alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genetics*, 73(4):823–834, October 2003.
6. N. Bray, I. Dubchak, and L. Pachter. AVID: A Global Alignment Program. *Genome Research*, 13(1):97–102, 2003.
7. K. M. Chao, R. C. Hardison, and W. Miller. Recent developments in linear-space alignment methods: A survey. *Journal of Computational Biology*, 1(4), 1994.
8. T.-C. Chu, T. Liu, D. T. Lee, G. C. Lee, and A. C.-C. Shih. GR-Aligner: an algorithm for aligning pairwise genomic sequences containing rearrangement events. *Bioinformatics*, 25(17):2188–2193, 2009.

9. C. J. Cohen, W. M. Lock, and D. L. Mager. Endogenous retroviral ltrs as promoters for human genes: A critical assessment. *Gene*, In Press, Corrected Proof:–, 2009.

10. A. B. Conley, J. Piriyapongsa, and I. K. Jordan. Retroviral promoters in the human genome. *Bioinformatics*, 24(14):1563–1567, 2008.

11. M. Crochemore, G. M. Landau, and M. Ziv-Ukelson. A sub-quadratic sequence alignment algorithm for unrestricted cost matrices. In *Proc. SODA '02*, pages 679–688, Philadelphia, PA, USA, 2002. Society for Industrial and Applied Mathematics.

12. A. Delcher, S. Kasif, R. Fleischmann, J. Peterson, O. White, and S. Salzberg. Alignment of whole genomes. *Nucl. Acids Res.*, 27(11):2369–2376, 1999.

13. J. S. Deogun, J. Yang, and F. Ma. Emagen: an efficient approach to multiple whole genome alignment. In *Proc. APBC '04*, pages 113–122, 2004.

14. E. Kindlund, M. T. Tammi, E. Arner, D. Nilsson, and B. Andersson. Grat–genome-scale rapid alignment tool. *Computer Methods and Programs in Biomedicine*, 86(1):87 – 92, 2007.

15. S. Kurtz. Approximate string searching under weighted edit distance. In *Third South American Workshop on String Processing*. Carleton Univ. Press, 1996.

16. J. J. Liu, G. S. Huang, Y. L. Wang, and R. C. T. Lee. Edit distance for a run-length-encoded string and an uncompressed string. *Inf. Process. Lett.*, 105(1):12–16, 2008.

17. N. Nagarajan, T. D. Read, and M. Pop. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, 24(10):1229–1235, 2008.

18. G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.

19. M. Oja, J. Peltonen, J. Blomberg, and S. Kaski. Methods for estimating human endogenous retrovirus activities from est databases. *BMC Bioinformatics*, 8(Suppl.2), 2007.

20. M. Oja, P. Somervuo, S. Kaski, and T. Kohonen. Clustering of human endogenous retrovirus sequences with median self-organizing map. In *Proc. of WSOM 2003*, 2003.