Letter for SoICT draft Improvement

Dear Editor,

We are very happy to submit our SoICT 2017 draft for the reviewing process of INFORMATICA journal. We tried to improve throughly the first draft presented in SoCIT 2017 conference. One of the substantial improvement of the first draft is that we have reconstructed the DNA sequence search model presented to speed up the searching time for a relative short query sequences again the whole genome. We accomplished around 10 times speed up in human genome testing experiment. In previous experiments, we used the artificially modified DNA segment for the searching query. For the new draft, we have used a biologically meaningful query sequence LTR(Long Terminal Repeat). Long terminal repeats (LTRs) are identical sequences of DNA that repeat hundreds or thousands of times found at either end of retrotransposons or proviral DNA formed by reverse transcription of retroviral RNA. So locating LTRs over whole genome scale is a crucial step to reveal the evolutionary process(such as constructing the phylogenetic tree).

We are asked to locate one LTR sequence over the human chromosome No.1 from evolution biology group since they did not successfully locate LTRs with other tools. In this draft we newly present how we successfully done this job effectively and efficiently with new figures. Also we have changed the following features in new draft.

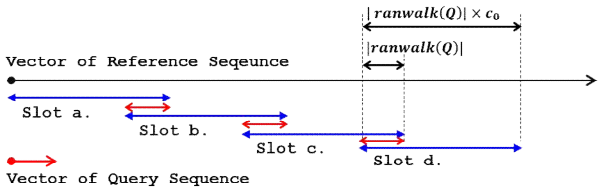
- the paper title
- Newly added references for alignment-free methodology [Ref.2]
- Added a description and formular for the improved query searching algorithm [Chapter 3.5 Reference Sequence Slot]
- Explained LTRs used in the experiment [Chapter 4.1 Dataset Preparation]
- Added and changed more experimental results with visualized figures [Fig.13-19]
- Added the new comparison shots using our visualization framework [Fig.20]

To make easy and insightful understanding, we provide the changing one by one in the following pages. We know the amount of updated contents are more than 30% of the previous SoICT draft, which are required to promote our work to the INFORMATICA reviewing process. Please let me know if we have to do more work to improve this version. We all authors wish to publish our work in the prestigious journal INFORMATICS.

Best Regards,

In the following, the blue statements are the newly added contents.

Three main improvement is description of new model, experiment dataset and visualization result of query search. Very detailed modifications are attached to the file(*docu_compare(1in1).pdf, docu_compare(2in1).pdf*) as a change tracking history using the pdf comparison tool.

In chapter 3. New method using 3D Random Plot (Add new model description)	
SoICT2017	**Informatica**
3.1 3D Random Plot Representation for DNA Sequence 3.2 Vector Allocation for Walk Plot 3.3 Vector Extraction from Random Walk Plot 3.4 Computing Similarity and Search on Walk plot	3.1 3D Random Plot Representation for DNA Sequence 3.2 Vector Allocation for Walk Plot 3.3 Vector Extraction from Random Walk Plot 3.4 Computing Similarity and Search on Walk plot 3.5 Reference Sequence Slot If the length of the query is long enough, the sequence information is compressed at an appropriate rate during visualization in a limited space. Therefore, it is possible to perform in the on-memory state by applying the same compression ratio when searching in the reference sequence. However, sequences with short lengths, such as the LTR sequence, are only kilo-bytes in size and remain uncompressed in the visualization process. In this case, vector information becomes large, and query search becomes impossible in on-memory state. In order to compensate for this, when the length of the reference sequence differs by more than 200 times, the reference sequence is divided into an appropriate number of slots to perform the search. By reducing the search range by multiple of the query length at a certain point in time, the method described above can be applied even in a case where a search is required at a low compression ratio in a large size sequence. $$\|Slot(Q,R)\| = \frac{\|ranwalk(R)\| - c_0 \bullet \|ranwalk(Q)\|}{\|ranwalk(Q)\| \bullet (c_0 - 1)}$$ The expression \ref{slot_numm} is the number of slots created when a query and reference sequence are given. Q and R are Query and Reference sequence respectively, and $len(ranwalk(X))$ represents the length of the whole vector information when X sequence is expressed as a random walk plot. c_0 is a control constant, which is the size of the space in which a vector should be searched when a certain size query vector is given. In this paper, c_0 is set to around 200.0. Since the query may exist at the point where the slot is divided, the boundaries of each slot are overlapped by the length of the query vector. The figure \ref{slotting} shows that the vector of the reference sequence is divided into slots.  Figure 9: The result of slot division in reference sequence vector based on the vector length of the query sequence.

In chapter 4. Experiments (Add Biological Dataset, Search Result)	
SoICT2017	**Informatica**
4.1 Data Set Table 2. Specification of artificial data of 1M, 2M size extracted from Human chromosome 1 and comparison result Table 3. Specification of artificial data of 4M, 10M size 4.2 Experiment:Similarity Comparison 4.3 Experiment:Sequence Search Over whole genome sequence	4.1 Dataset Preparation Table 2. Specification of artificial data of 1M, 2M size extracted from Human chromosome 1 and comparison result Table 3. Specification of artificial data of 4M, 10M size Table 4. Specification of biological data for reference 4.2 Experiment:Comparison Between Modification ratio and Similarity based proposed Model 4.3 Experiment:Artificial Sequence Search over whole genome sequence 4.4 Experiment:Biological Sequence Search over whole genome sequence The table \ref {search_bio_result} shows the results of searching a biological query sequence in a whole genome sequence. The search for the LTR sequence (Q-F-1) extracted from the flatfish chromosome 1 resulted in a similarity of 85.7\% within 90 B.P. of the actual query position within about 0.4 seconds of search time. On the other hand, the HER-V sequence (Q-H-1) extracted from human chromosome 1 took relatively longer time, longer than 40 seconds because the length of the query sequence was short and the length of the reference sequence was long. The difference between the actual position and the search result is about 2000 B.P., which is relatively accurate considering that the length of the reference sequence is more than 200M. The figures \ref{find_ltr_f1, find_ltr_f2, find_ltr_f3, find_ltr_f5} visualize the flatfish chromosome 1,2,3,5 sequences, respectively. The red one is a visualization of the whole genome of a flatfish, and the area marked in blue is where each query was searched. The figures \ref{find_ltr_f3, find_ltr_f5} show that the marked positions are almost identical to the origin, reflecting that the Q-F-3 and Q-F-5 queries are actually located within 0.5 \% of the flatfish whole genome sequence. On the other hand, the figures \ref{find_ltr_f1, find_ltr_f2} reflect that the marked positions are relatively far away from the origin, that the positions of the Q-F-1 and Q-F-2 queries are actually located within 7\% and 10\% of the flatfish whole genome sequence. The figure \ref{find_ltr_h1} visualizes the human chromosome 1 sequence and marks the result of searching the Q-H-1 query. It is well reflected that the Q-H-1 query is actually located in the early 63 \% (about 155 MB.P.) of the human sequence. The figure \ref{find_ltr_h1_E} is the result of original query sequence(Q-H-1) and enlarged subsequence of the reference sequence(R-H-1) at searched position. The similarity of the searched sequence in the reference(green plot) was 78\%, and it can be confirmed that the query is very similar to the query when matched with the query sequence. Table 7. The result of sequence search for biological query sequence in flatfish and human chromosome

In chapter 4. Experiments (Add and change Search Result Figure)	
SoICT2017	**Informatica**

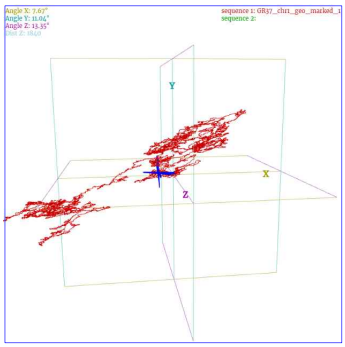


Figure 12: The visualization of search result for query sequence of 1M size in human chromosome 1. Red random walk plot represents human chromosome 1, and blue point represents the location where the query was searched.

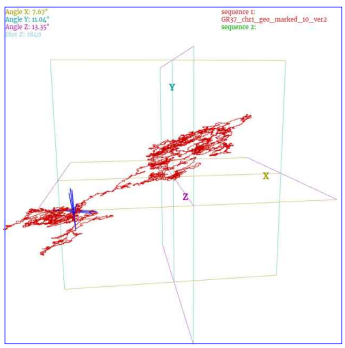


Figure 13: The visualization of search result for query sequence of 10M size in human chromosome 1.

Newly added or changed figures to show the final results.

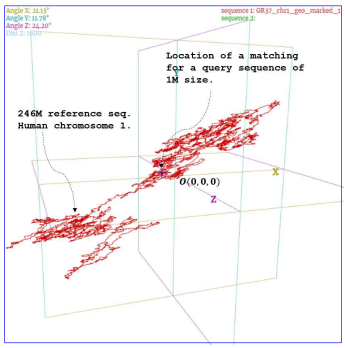


Figure 13: Searching result of query sequence(A1-0) in reference sequence(human chromosome 1). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

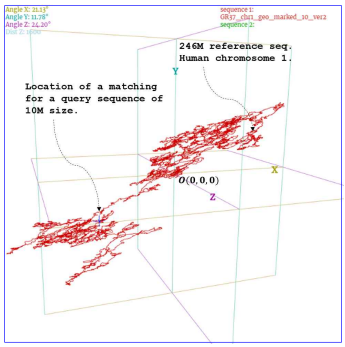


Figure 14: Searching result of query sequence(A10-0) in reference sequence(human chromosome 1). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

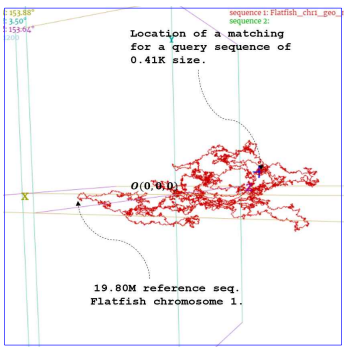


Figure 15: Searching result of query sequence(QF-1) in reference sequence(R-F-1). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

In chapter 4. Experiments (Add and change Search Result Figure)	
SoICT2017	**Informatica**
	Newly added figures to show the final results.

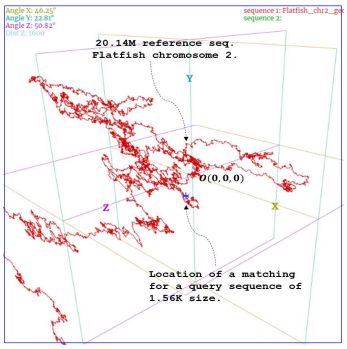


Figure 16: Searching result of query sequence(QF-2) in reference sequence(R-F-2). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

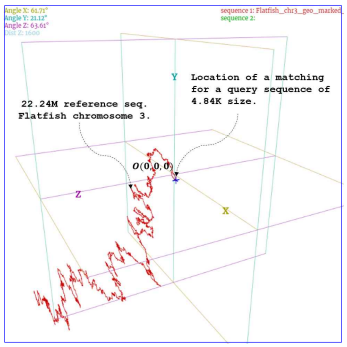


Figure 17: Searching result of query sequence(QF-3) in reference sequence(R-F-3). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

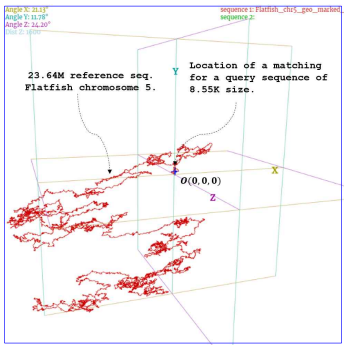


Figure 18: Searching result of query sequence(QF-5) in reference sequence(R-F-5). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

In chapter 4. Experiments (Add and change Search Result Figure)	
SoICT2017	**Informatica**
	Newly added figures to show the final results.

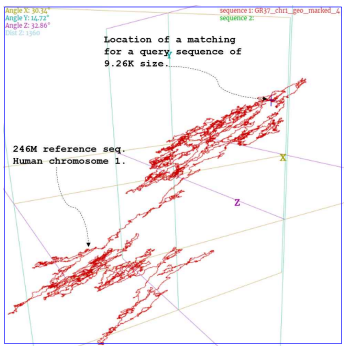


Figure 19: Searching result of query sequence(QH-1) in reference sequence(R-H-1). Red plot represents reference sequence and blue cross point represents the position of searched query sequence.

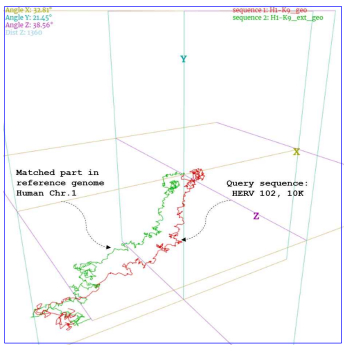


Figure 20: Matching result between the query sequence(Q-H-1) and the extended subsequence of reference sequence(R-H-1), which was depicted as a blue cross in figure 19.