SLIDESORT : Developing an exact method to find similar pairs with small edit-distance

Kana Shimizu, Koji Tsuda

Computational Biology Research Center, National Institute of Advanced Science and Technology, JAPAN

Contact: shimizu-kana@aist.go.jp

ABSTRACT: In this study, we propose an exact method that finds all similar pairs from a string pool in terms of edit distance without any duplication on small memory in O(N). Using an efficient pattern growth algorithm, our method discovers chains of common k-mers to narrow down the search. Compared to existing methods based on single k-mer, our method is more effective in reducing the number of edit distance calculations. In large short read datasets, our method was 20-3000 times faster than the state-of-the-art method, BWA.

Motivation

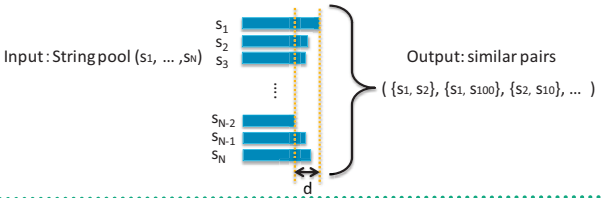
Due to the dramatic improvement of DNA sequencing, it required to evaluate sequence similarities among a huge amount of fragment sequences such as short reads. Finding similar pairs from a string pool is an important first step for many biological sequence analyses.
In this work, we would like to address the problem of enumerating all neighbor pair in a large string pool in terms of edit distance. This problem is conventionally called "all pairs similarity search", that will help ...

- ▶ Shot reads analyses
  - preprocess of de nove assenmbly
  - Clustering reads. (e.g. online tree construction, greedy clustering etc ...)
- ▶ genome analyses
  - Repeat analyses
  - Sequence pattern analyses etc...

Problem setting

Given n strings of similar length (maximal length−minimal length$\leqq d$), $s_1,...,s_N$, the task is to find all pairs whose edit distance is at most d,

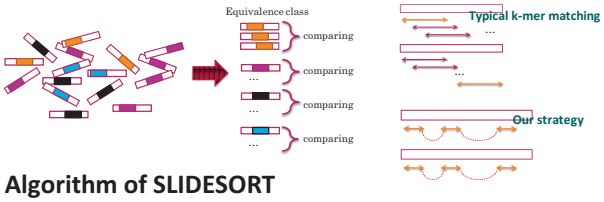
$$E = \{(i,j) \mid EditDist(s_i, s_j) \leq d, i < j\}$$



Method

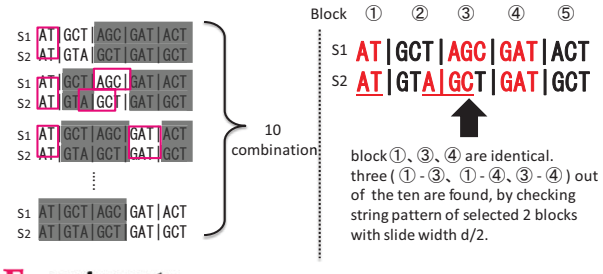
Strategy of SLIDESORT

Like other method based on k-mer matching, SLIDESORT finds common substrings and verify the matches. The key idea of SLIDESORT is to find longer common substrings to narrow down search space. We developed an efficient pattern growth algorithm inspired by multiple sorting algorithm (*Uno 2008 PAKDD*).



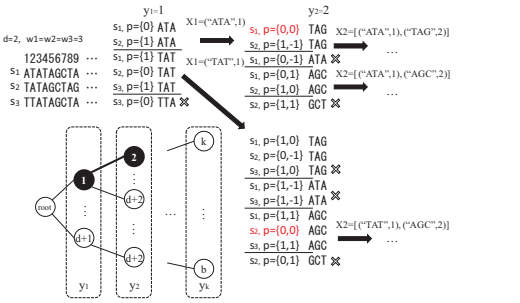
Theory of SLIDESORT

How SLIDESORT can find longer common substrings? Given two sequences that are divided into $k+d$ blocks, If edit distance of the two sequences is at most d, there exist at least k blocks that are the same string patterns between the two within slide width d/2. Based on this theory, SLIDESORT regards concatenation of k blocks as a common substring.

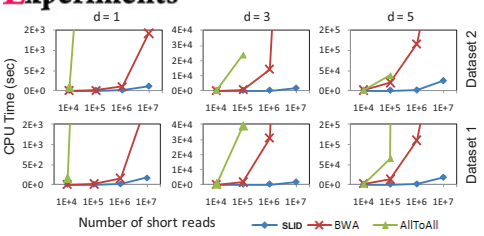


Algorithm of SLIDESORT

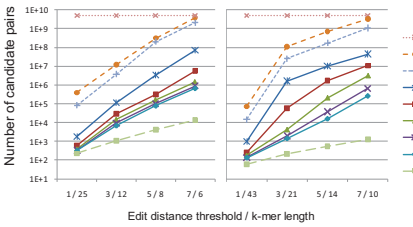
A pattern corresponds to a sequence of substrings. The space of all patterns is organized as a tree and systematically traversed. SLIDESORT uses radix sort to find equivalent strings in pattern growth.



Experiments



Evaluation on two types of short read datasets. The graphs mainly compares performance difference in size of the dataset.



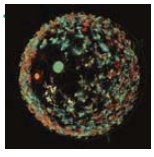
Comparison of number of candidate pairs. The evaluations were done on 100,000 short reads. The proposed method was examined with k=1,...,5. 'Neighbor pairs' represent the actual number of neighbor pairs in data. 'k-mer/nonredundant' and 'k-mer/redundant' represent two variants of the single seed method.

Dataset from NCBI SRA
ERR1081 : Read length 51
SRR020262 : Read length 87
Computation environment
Implemented by C++
Linux PC with Intel Xeon X5570

SLIDESORT is available from www.cbrc.jp/~shimizu/slidesort/

Application

One of an application of SLIDE-SORT is constructing a tree. The right graph is a minimum spanning forest of 112,995 short reads that was generated by processing a pair of sequences (an edge) one-by-one. Most of the case, computation time of each process is smaller than interval time to find next pair.



d	SLIDE-S	BWA Index Time	BWA Search Time	AllToAll
1	0.53	2.34	3.25	8743.46
3	2.71	2.37	562.63	23796.1
5	20.64	2.19	19697.67	38179.5
7	131.88	2.51	10161.15	50922.8

This work is supported by Grant-in-Aid for Young Scientists (22700319) of Japan Society for the Promotion of Science.