SlideSort: A fast and exact tool for finding all similar pairs from next-generation sequencing data

Kana Shimizu and Koji Tsuda

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)

shimizu-kana@aist.go.jp, koji.tsuda@aist.go.jp, slidesort@m.aist.gojp

We have designed and implemented fast algorithm that can efficiently analyze next-generation sequencing (NGS) data. We address the problem of finding all similar pairs from a set of short reads, which appears in various NGS data analyses, such as de novo assembly, read clustering and frequent sequence pattern search. Given distance threshold *d*, SlideSort exactly finds all similar pairs whose edit distance is at most *d* from a set of short reads. By using efficient pattern growth algorithm, SlideSort finds chains of common k-mers to narrow down the search. Comparing to using single k-mer, SlideSort drastically reduces the number of edit distance calculations. Experimental results show that SlideSort is much faster than state-of-the art methods with comparable memory size on the dataset downloaded from NCBI SRA. The tool equips useful function of calculating minimum spanning trees, which can be directly applicable to read clustering. Executable binary files and C++ libraries are available at http://www.cbrc.jp/~shimizu/slidesort/ for Linux and Windows.