Read indexing

Nicolas Philippe, Mikael Salson, Thierry Lecroq, Martine Léonard, Thérèse Commes, Eric Rivals

Read indexing

Nicolas Philippe, Mikael Salson, Thierry Lecroq, Martine Leonard, Therese Commes, Eric Rivals
Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS, équipe MAB, Université de Montpellier II, Montpellier, France,
LITIS, Univ. Rouen, Mont Saint Aignen, France
CRBM, UMR 5237 CNRS, Montpellier, France
http://www.lirmm.fr/~rivals

The question of read indexing remains broadly unexplored. However, the increase in sequence throughput urges for new algorithmic solutions to query large read collections efficiently. We propose a solution, named Gk arrays, to index large collections of reads, an algorithm to build the structure, and procedures to query it. Once constructed, the index structure is kept in main memory and is repeatedly accessed to answer various types of queries. We compare our data structure to other possible solutions to investigate its scalability and computational efficiency. Gk arrays are implemented in a general purpose library, which may prove useful for assembly purposes, for evaluating the expression level in RNA-seq, and others high throughput sequencing applications.

References

1. Querying large read collections in main memory: a versatile data structure. N. Philippe, M. Salson, T. Lecroq, M. Leonard, T. Commes and E. Rivals. BMC Bioinformatics, Vol. 12, p. 42, doi:10.1186/1471-2105-12-242, 2011.

Relevant Web sites

2. http://crac.gforge.inria.fr/gkarrays/
3. http://www.atgc-montpellier.fr/ngs/