Improved data structures for clustering of transcriptomal and genomic NGS data

Collaboration Partner: Prof. Scott Hazelhurst, School of Electrical and Information Engineering, Wits Bioinformatics, University of the Witwatersrand, Johannesburg, South Africa

UniVr Partner: Dr. Zsuzsanna Lipták, Dept. of Computer Science, Research Group Bioinformatics and Natural Computing, University of Verona, Italy

The need for clustering and assembly of transcript and genomic DNA has been growing with the development of new generation sequencing technologies (NGS), and is becoming ever more pressing as sequencing costs drop. For example, in studying human variation or for personalised medicine, complete exomic or even genomic sequencing will soon be routine. Short read sequence data needs to be post-processed using (a) clustering to find sequences generated by gene products, (b) assembly (either de novo or resequencing), (c) variant detection. All of these tasks require fast approximate string matching algorithms as a basis.

Modified suffix array sa_k

aaa@aacggt@gttaaagt@tcggt@gttat@cgg@acggt@

i	sa	Text from sa[i]	sa_k
15	16	agt@tcggt@gttat@cgg@acggt@	16
16	29	at@cgg@acggt@	29
17	32	cgg@acggt@	6
18	37	cggt@	21
19	6	cggt@gttaaagt@t...gt@	32
20	21	cggt@gttat@cgg@acggt@	37
21	34	g@acggt@	34
22	33	gg@acggt@	33
23	38	ggt@	7
24	7	ggt@gttaaagt@t...t@	22
25	22	ggt@gttat@cgg@acggt@	38
26	39	gt@	8

Suffix-array based approaches have proved very successful in this context; however, as data set sizes and the number of data sets grow, computational demands increase and so does the need to improve performance and quality. Especially for de novo assembly of NGS data, memory usage can be prohibitive. The trade-off between memory usage and running time is thus a crucial parameter to explore.

In KABOOM we developed a filtering method for expression clustering based on modified suffix-arrays [Hazelhurst & Lipták, Bioinformatics, 2011]. We have shown this approach to be very effective, as evidenced by extensive experimentation. We are extending this work in several directions: (a) development of more memory efficient data structures and algorithms, in particular by using compression, (b) extension to assembly and variant calling, (c) parallelisation.

The University of Witwatersrand (Wits) Bioinformatics provides bioinformatics resources and service, and has carried out projects on a large variety of biological data, including Bioinformatics of viral diversity, Evolutionary biology, Experimental algorithms and high-performance computing, Assembly of second generation sequence data, Functional annotation of novel data, Genome-wide Association Studies, Human Diversity Studies, with particular emphasis on genetic data in Africa.