High performance text indexing and applications in life sciences

Eric Rivals

High performance text indexing and applications in life sciences

Eric Rivals

December 17, 2018

Address:
Laboratoire d'Informatique de Microélectronique et de Robotique de Montpellier (LIRMM) and Institute of Computational Biology (IBC) CNRS and Université de Montpellier, France
http://www.lirmm.fr/~rivals/

Large corpura of texts or of sequences serve as references and are interrogated through web site or programming interfaces. In life sciences, new sequencing technologies have revolutionised the acquisition of genomic sequences and triggered an exponential accumulation of reference sequences in international, public databases. Several kinds of text queries form the basic operation of programs that analyse genomic sequences. For instance, the webserver of EMBL-EBI receives 27 million queries a day. A typical sequencing experiment yields a hundred million sequencing reads – about 150 nucleotides long – each of which needs to be compared to a reference genome. To analyse such data or to mine public sequence repositories demands very efficient programs and algorithms. Only, the use of complex and specific, indexing data structures allows us to match the needs of Life sciences communities. I will present some indexing data structures that enables high performance computational analyses in genomics, and mention their pracical applications. Beyond text data, such data structures can be adapted to index other types of discrete data like trees or graphs. This will be key for the development of computational pan-genomics.

Funding acknowledgement: "GEM" flagship project from Labex NUMEV; for travelling to the workshop Royal Society (UK), Académie des Sciences, and CNRS.

Abstract of speech for the *UK-France Bilateral International Meeting* on High Performance Computing and Biomathematics, held at Chicheley, UK, on the 20-21 Feb. 2019, which was co-organised by the Royal Society and the French Académie des sciences.