Indexation et requêtage de grandes collections de données de séquençage

Indexing and query large collections of sequencing data

Soutenu par le GDR BIM

cnrs GDR Groupement de recherche BIM Bioinformatique moléculaire

15h-15h30

Camille Marchet (CRIStAL, CNRS, Lille)

Data-structures for querying large k-mer (collections of) sets

High-throughput sequencing datasets are usually deposited in public repositories, e.g. the European Nucleotide Archive, to ensure reproducibility. As the amount of data has reached petabyte scale, repositories do not allow to perform online sequence searches; yet such a feature would be highly useful to investigators. Towards this goal, in the last few years several computational approaches have been introduced to index and query large collections of datasets. In this seminar I propose an overview of methods for representing and indexing sets of k-mer efficiently. Then we will review how these techniques were adapted to index collections of thousands of datasets (and more) for membership queries. I will propose application examples for these techniques with a focus on RNA and splicing.

15h30-16h

Mikhail Karasikov (Biomedical Informatics Group, Department of Computer Science, ETH Zurich, Zurich, Switzerland)

Scalable Indexing of Sequencing Data in Annotated Genome Graphs

Technological advances have led to an exponential growth in the amount of high-throughput sequencing data available to the scientific community. However, most of this data, commonly stored in repositories such as the NCBI Sequence Read Archive, which currently contains over 50 petabases of sequences, is only available in raw format. Transforming it into a searchable representation that is easily accessible to life science researchers for large scale analysis and search remains an unsolved problem.
In my talk, I will review the state-of-the-art approaches for indexing large cohorts of sequencing data. Then, I will describe MetaGraph, a method that efficiently indexes petabase-scale cohorts of sequencing experiments in annotated de Bruijn graphs and that supports k-mer search and sequence-to-graph alignment. Internally, MetaGraph represents input data as collections of k-mer sets encoded in succinct data structures, offering practically relevant trade-offs between the index size and the query performance. This flexibility allows running MetaGraph at different scales and on different hardware, from laptops to research compute clusters and distributed cloud environments.
I will pay special attention to the methods and data structures used in MetaGraph for representing graph annotations, including non-binary attributes with Counting de Bruijn graphs, e.g., for representing gene expression and genome coordinates. Finally, I will conclude with real-

world applications, such as indexing a portion of all publicly available whole-genome sequencing samples from the Sequence Read Archive, currently including over 90% of all Microbe, Fungi, Plant, and Human, as well as indexing all reference genome sequences (RefSeq), the RNA-Seq Genotype-Tissue Expression dataset (GTEx), and significantly more diverse metagenomic data, such as the entire catalog of 286,997 reference genome sequences from the human gut microbiome (UHGG), all 242,619 publicly available human gut microbiome short read sequencing samples, and a set of 4,220 public transit surface microbiomes (MetaSUB).

16h-16h30

Thérèse Commes (IRMB, University of Montpellier, INSERM U1183)

TranSipedia: a novel framework for large scale RNAseq data analysis with applications in cancer from research to diagnosis

Driven by myriads of projects, public RNAseq databases are exploding, to date, over 850,000 RNA-seq are deposited on SRA for humans. This huge body of publicly available RNAseq libraries is a precious resource to identify specific transcriptional events. The challenges lie in the complexity of RNA biological content and the exponential increase in data volume. We want to make RNA-seq data easily accessible providing a better capture of the whole transcriptome complexity, in the context of biological and human health applications. New computational methods that perform indexing of k-mers across huge datasets constitute interesting solutions to interrogate "Omics data" at a large scale from dataset collections. Here, we developed TranSipedia, a new framework, based on kmer approach, constructed with several modules: 1/ The RNA-seq indexing step constructed with Reindeer (REad Index for abuNDancE quERy; Marchet et al., 2020), a novel computational method that serves as an efficient platform to request all transcribed information, 2/a completed module to generate kmers as signature of transcripts (Kmerator; Riquier et al, 2021), 3/ a supporting web site to facilitate the queries easily shared by biologists (TranSipedia, https://transipedia.montp.inserm.fr).
Reindeer performs indexing of k-mers and records their counts across a large collection of datasets. Interestingly it associated k-mers to their counts instead of only recording the presence/absence of k-mers as frequently done in previous works. Moreover, Reindeer provides an ultra-fast performance in the query process while indexing several thousands of RNAseq. One of the great advantages of indexing raw data is also that it integrates reference-free and annotation-free approaches. For applications where gene expression level is required, the kmer count must be sufficiently sensitive and representative to be applicable. The quantitative accuracy with k-mers counts from Reindeer indexed datasets was compared to classical quantitative methods like Kallisto. Secondly, the design kmer module uses Kmerator, a tool develop to construct specific kmers already avalailable on github. Thirtly, the website is already available to facilitate index queries by the biologists with kmer on fasta file format. The Transipedia platform now deserves several thousands of datasets from public and private collections mainly from acute myeloid leukemia (AML) for cancer applications. We indexed whole CCLE cohort representing 1019 RNAseq samples for a total of 10To and f indexes from Leucegene, BEAT-AML, SRA and TCGA collections represented with more than 1000 RNAseq samples for AML application. Concerning biological applications, we already requested in selected public dataset gene co-expression, biomarker tissue specificity, as well as tumor specific signatures comparing normal/tumor, for simple and useful medical usage. In perspectives, based on data structures, such as k-mer features, diagnosis applications are in development. Moreover, Machine learning models could be used to search for signatures and explore better diagnostic and prognosis models.

16h30-17h Coffee Break

17h-17h30

Karel Břinda (INRIA/IRISA Rennes)

The Tree of Life enables efficient and robust compression and search of microbes

Comprehensive collections of genomes have been instrumental in the study of life using sequencing; however, their storage, transmission, and analysis have become challenging due to the exponential growth of sequence data. This raises the question of designing efficient computational solutions for storing and indexing large data sets, such as the recently created comprehensive data set of 661k bacterial genomes, uniformly assembled from the European Nucleotide Archive.

Here, we present MOF (Microbes on a Flash Drive), a method for large-scale lossless compression and search of microbial collections, using the Tree of Life as a biological prior on their redundancy structure. MOF uses state-of-the-art tools and databases from population genomics and metagenomics to infer the geometrical structure of a given collection and identify the likely position of individual microbes on the Tree of Life, and then uses this information for guiding data compression using standard approaches.

We show MOF losslessly compresses 425 thousand microbial de Bruijn graphs from 16 TB (McCortex) to 52.2 GB (xz) (319×), 661 thousand bacterial assemblies from 805 GB (gzip) to 29 GB (xz) (27×), and their associated k-mer index from 937 GB (COBS) to 110 GB (xz) (8.5×). When combined together as building blocks, this enables performing BLAST-like alignments to the 661k data set on a standard desktop computer within several hours, as we demonstrate with the examples of resistance genes, plasmids, and nanopore reads as queries. The optimization of data structures using the Tree of Life has broad applications across computational biology and provides a fundamental design principle for future genomics infrastructure.

17h30-18h

Open table & discussions