Methods xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying

Marco Masseroli *, Abdulrahman Kaitoua, Pietro Pinoli, Stefano Ceri

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

ARTICLE INFO

Article history:
Received 1 May 2016
Accepted 12 September 2016
Available online xxxx

Keywords:
Genomic data management
Data modeling
Data interoperability
Metadata management
Query languages
Operations for genomics

ABSTRACT

While a huge amount of (epi)genomic data of multiple types is becoming available by using Next Generation Sequencing (NGS) technologies, the most important emerging problem is the so-called *tertiary analysis*, concerned with sense making, e.g., discovering how different (epi)genomic regions and their products interact and cooperate with each other. We propose a paradigm shift in tertiary analysis, based on the use of the Genomic Data Model (GDM), a simple data model which links genomic feature data to their associated experimental, biological and clinical metadata. GDM encompasses all the data formats which have been produced for feature extraction from (epi)genomic datasets. We specifically describe the mapping to GDM of SAM (Sequence Alignment/Map), VCF (Variant Call Format), NARROWPEAK (for called peaks produced by NGS ChIP-seq or DNase-seq methods), and BED (Browser Extensible Data) formats, but GDM supports as well all the formats describing experimental datasets (e.g., including copy number variations, DNA somatic mutations, or gene expressions) and annotations (e.g., regarding transcription start sites, genes, enhancers or CpG islands). We downloaded and integrated samples of all the above-mentioned data types and formats from multiple sources. The GDM is able to homogeneously describe semantically heterogeneous data and makes the ground for providing data interoperability, e.g., achieved through the GenoMetric Query Language (GMQL), a high-level, declarative query language for genomic big data. The combined use of the data model and the query language allows comprehensive processing of multiple heterogeneous data, and supports the development of domain-specific data-driven computations and bio-molecular knowledge discovery.

1. Introduction

Extraordinary advances in genomics are made possible by Next Generation Sequencing (NGS), a family of technologies that is progressively reducing the time and cost of reading an individual genome[1]; therefore, huge amounts of sequencing data of many genomes, in multiple biological and clinical conditions, are continuously collected and made publicly available, often organized by worldwide consortia such as ENCODE [1], Roadmap Epigenomics [2], TCGA [3] and the 1000 Genomes Project [4].

So far, the bioinformatics research community has been mostly challenged by *primary analysis* (production of sequences in the form of short DNA or RNA segments, or "reads") and *secondary analysis* (alignment of reads to a reference genome and search for specific features of genome regions, such as variants and peaks of binding or expression intensities) [5]. The most important emerging problem is the so-called *tertiary analysis* [6], concerned with sense making, e.g., discovering how different (epi)genomic regions and their products interact and cooperate with each other. Tertiary analysis requires integrating heterogeneous DNA features, such as variations (e.g., a mutation in a given DNA position), or peaks of binding or expression (i.e., genomic regions with higher read density), or structural properties of the DNA (e.g., break points, where the DNA is damaged, or junctions, where the DNA creates loops). Such data are collected within numerous and heterogeneous files (both in formats and semantics); they are usually distributed within different repositories, and lack an attribute-based organization for systematically expressing features as high-level attributes. Furthermore, they lack a systematic description of their metadata, i.e., of their biological and clinical properties, which are greatly heterogeneous. Answers to crucial biomedical questions may be hidden within already existing open and public collections of heterogeneous data, but the methods and tools which are made

* Corresponding author.
E-mail addresses: marco.masseroli@polimi.it (M. Masseroli), abdulrahman.kaitoua@polimi.it (A. Kaitoua), pietro.pinoli@polimi.it (P. Pinoli), stefano.ceri@polimi.it (S. Ceri).

[1] Recently below the barrier of 1000 US$ for a human genome (https://www.genome.gov/sequencingcosts/).

http://dx.doi.org/10.1016/j.ymeth.2016.09.002
1046-2023/

available for knowledge extraction are still rather poor and specialized.

We propose a paradigm shift in tertiary genomic data management, based on the introduction of a simple data model which links genomic features to their associated metadata. This model is able to homogeneously describe semantically heterogeneous data and makes the ground for providing data interoperability, which can be achieved through a high-level, declarative query language for genomic big data. The combination of the data model and query language provides the right concepts for information extraction from genomic data repositories, and allows the development of domain-specific data-driven computations required by tertiary data analysis and bio-molecular knowledge discovery.

2. Genomic Data Model

The Genomic Data Model (GDM) that we propose is based on the notions of *datasets* and *samples*, and on two abstractions: one for *genomic regions*, which represent portions of the DNA and their features, and one for their *metadata*. Datasets are collections of samples, and each sample consists of two parts: the *region data*, which describe the characteristics and DNA location of genomic features (e.g., called through the processing of raw NGS data after their alignment to a reference genome), and the *metadata*, which describe general properties of the sample.

2.1. Motivation

Genomic region/feature data are very valuable for molecular investigation and precision medicine; they describe a broad variety of molecular aspects, which are individually measured, and provide single views on biomolecular phenomena. Their integrated evaluation would provide a systemic view on how they interact and cooperate towards the triggering and regulation of biological functions. Yet, they are available in a variety of formats which hamper their integration and comprehensive assessment.

GDM provides a schema to genomic feature data of DNA regions; thus, it makes such heterogeneous data self-describing, as advocated by Jim Gray [7], and interoperable. This is obtained by simple mapping of the data from data files in their original format into the GDM format when they are used, without including them into a database, so as to preserve the possibility for biologists to work with their usual file-based tools. The provided data schema has a fixed part, which guarantees the comparability of regions produced by different kinds of processing, and a variable part reflecting the "feature calling process" that produced the regions and describing the region features determined through various processing types. DNA regions are sequences of nucleotides[2], usually represented by strings of letters[3]; GDM identifies them through their genomic coordinates and associates them with a list of one or more features (e.g., produced by NGS data secondary analysis).

Metadata are paramount to characterize the high heterogeneity of genomic feature data and guide their correct processing; however, they are collected in a broad variety of data structures and formats that constitute barriers to their use and comparison. To cope with the lack of agreed standards for metadata, GDM models metadata simply as free arbitrary semi-structured attribute-value pairs, where attributes may have multiple values (e.g., the *Disease* attribute can have both "*Cancer*" and "*Diabetes*" values). We expect metadata to include at least the considered organism, tissue, cell line, experimental condition (e.g., antibody target – in the case of NGS ChIP-seq experiments, treatment, etc.), experiment type, data processing performed, feature calling and analysis method used for the production of the related data; in the case of clinical studies, individual's descriptions including phenotypes.

2.2. Definitions

A *genomic region r* is a well-defined portion of the genome identified by the quadruple of values <*chr*, *left*, *right*, *strand*>, called *region coordinates*, where *chr* represents the DNA chromosome where the region is located, *left* and *right* are the positions of the two ends of the region along the DNA coordinates[4]; *strand* indicates the DNA strand on which the region is read, as well as the direction of DNA reading[5] (encoded as either '+' or '−'), and can be missing (encoded as '∗') when the region is not assigned to a specific strand, e.g., in the case of DNA binding regions identified through NGS ChIP-seq experiments.

A *sample s* is formally modeled as a triple <*id*; *R*; *M*> where:

- *id* is the sample *identifier* of type *long*
- *R* is the set of *regions* of the sample, built as pairs <*c*; *f*> of *coordinates c* and *features f*. Coordinates are composed of four fixed attributes *chr*, *left*, *right*, *strand* which are respectively typed *string*, *long*, *long*, *char*. Features are made of typed attributes; we assume attribute names of features to be different, and their types to be any of *Boolean*, *char*, *string*, *int*, *long*, *double* (GDM types are available in several programming languages, including Java and Scala, and frameworks for cloud computing, such as Apache Pig[6], Apache Flink[7] and Apache Spark[8]). The *region schema* of *s* is the list of attribute names used for the identifier, the coordinates and the features.
- *M* is the set of *metadata* of the sample, built as *attribute-value* pairs <*a*; *v*>, where we assume the type of each value *v* to be *string* (numerical values can then be casted to a numerical type, such as *int*, *long*, or *double*, when used). The same attribute name *a* can appear in multiple pairs of the same sample (in which case we say that *a* is multi-valued).

A *dataset* is a collection of samples with the same region schema and with features having the same types; sample identifiers are unique within each dataset. Each dataset can be thought as grouping related data samples, in case produced within the same project (either at a genomic research center or within an international consortium) by using the same or equivalent technology and tools, but with different experimental conditions, described by metadata.

2.3. Implementation example

According to GDM, each dataset can be stored using two data structures (e.g., two tables), one for regions and one for metadata. An example of two tables to represent a particular experiment, called *ChIP-seq*, is shown in Fig. 1, where two small samples are

[2] Nucleotides are the individual molecular components of the DNA macromolecule, and are of four different types (Adenine, Cytosine, Guanine, and Thymine).

[3] DNA can be abstracted as a string of billions of four different letters (A, C, G, T), each representing a nucleotide molecule, subdivided in chromosomes (23 in humans), which are disconnected intervals of the string.

[4] Species are associated with their *reference genome*. DNA samples are aligned to these references, hence referred to the same system of coordinates; for humans, several reference genomes were progressively defined, the latest is *hg20* (also known as *GRCh38* or *hg38*). According to the University of California at Santa Cruz (UCSC) notation, we use *0-based*, *half-open inter-base coordinates*, i.e., the considered genomic sequence is [*left*; *right*). In this coordinate system, left and right ends can be identical (e.g., when they represent a *splicing junction*), or consecutive (e.g., when the region represents a *single nucleotide polymorphism*).

[5] DNA is made of two strands rolled-up together in anti-parallel directions, i.e., they are read in opposite directions by the biomolecular machinery of the cell.

[6] https://pig.apache.org/.

[7] http://flink.apache.org/.

[8] http://spark.apache.org/.