CODAMONO

BioInterchange 2.0

Integrating and Scaling Genomics Data

by

CODAMONO

www.codamono.com
@CODAMONO

+1 647 780 3927

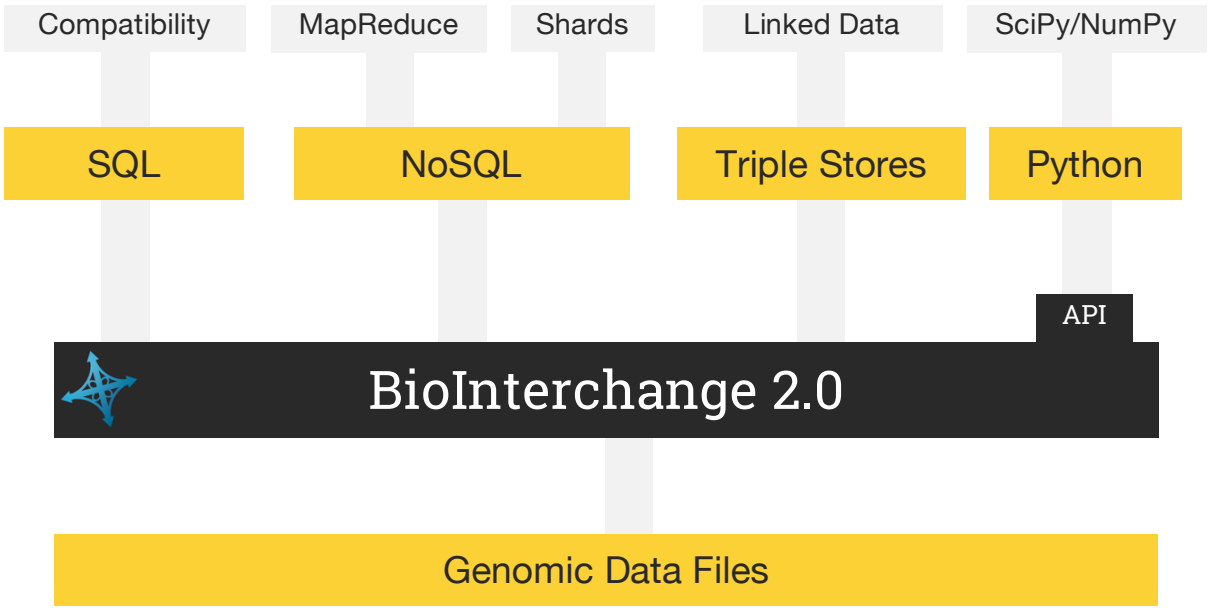
5-121 Marion Street, Toronto, Ontario, M6R 1E6, Canada

Genomics Today: Big Data Challenge

Genomic data — genes, transcription factor binding sites, regulatory elements, proteins — are available in abundance. World-wide data centers provide free access to genomic features and variations via FTP servers. The most commonly shared genomic data files are: GFF3, GVF and VCF.

Genomic data standards have one common denominator: they are incompatible to each other. Another problem are complementary data representation choices across GFF3, GVF and VCF standards, which make it difficult to access data from multiple data providers.

BioInterchange 2. 0 tackles genomics standards by overcoming differences in their specifications, by providing first-class data-access methods, and by enabling the integration of genomic data into modern database management systems.



Unified Data Model

Genomic file standards emerged as a data sharing solution, where large genome centers would offer GFF3, GVF and VCF files for download via FTP. Even today, data sharing via file transfer is the most prevalent method of obtaining genomics data.

File are not a data model though: formatting, encoding and representation of data items differ between GFF3, GVF and VCF files. It takes tremendous efforts to integrate genomics data into a data model that is suitable for data analysis and data mining.

BioInterchange 2.0 solves the data integration problem by providing a single JSON/JSON-LD data model for transparent data access and data processing. This makes it possible to work with genomics data independently of the originating file format. Of course, BioInterchange 2.0 also provides means to convert data in JSON/JSON-LD back into the original genomic file formats.

High-Performance API

Genomics data is inherently big data. Efficient algorithms are required to produce results within acceptable time frames. BioInterchange 2.0 comes with a Python API for first-class data processing. Whether it be for data filtering, data analysis, or data annotation, BioInterchange 2.0's high-performance API has a data throughput that is twice as fast as BioPython or BioRuby.

Data Integration

JSON has become a lingua franca not only in the web- and cloud-development world, but it is also ubiquitously supported by modern database management systems (DBMS). NoSQL DBMS such as MongoDB, RethinkDB, CouchDB, and ArangoDB use JSON at their core. Established SQL DBMS like MySQL and PostgreSQL provide native support for JSON now too. Search servers, for example Elasticsearch, or data warehouse centric systems build on top of Hadoop, such as Apache Hive and Apache Pig, handle JSON as well.

BioInterchange 2.0's discrete use of JSON-LD annotations enables data transformations into data formats of the Resource Description Framework (RDF) where necessary. This makes it possible to integrate genomics data into triple stores as well as other graph databases.

Information Sheet

Requirements

- Mac OS X Yosemite (OS 10.10) or newer; or
- Linux, amd64 architecture

Installation Packages

- CloudBioLinux (Linux)
- Homebrew Science (Mac OS X and Linux)
- DMG file (Mac OS X)
- Debian (jessie) package (Linux)
- Docker images on Docker Hub (Linux)

Supported Data Formats

- Generic Feature Format Version 3 (GFF3)
- Genome Variation Format (GVF)
- Variant Call Format (VCF)
- JavaScript Object Notation (JSON)
- JavaScript Object Notation for Linked Data (JSON-LD)

Software Interfaces

- Command line interface
- Python 3.4 Application Programming Interface (API)

End-User License Agreements

This is an overview only; complete Terms & Conditions are available online.

- Free Tier
    - Trial License: 30 days, non-renewable
    - Peer-Review License: 30 days, renewable
- Monthly Subscriptions
    - Individual License: personal use, non-transferable (United States/Canada only)
    - Team License: up to 5 simultaneous colocated users
    - Cloud, Cluster & HIPAA License: unlimited users, offline license verification