ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Large-scale biological meta-database management

Edvard Pedersen *, Lars Ailo Bongo

Department of Computer Science and Center for Bioinformatics, University of Tromsø, Tromsø, Norway

HIGHLIGHTS

- We present the GeStore system for biological meta-database management.
- We evaluate the performance characteristics of GeStore.
- We integrate GeStore with a workflow manager, and evaluate the benefits.
- Using the GeStore approach to meta-data management yields significant benefits.

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form
18 December 2015
Accepted 17 February 2016
Available online xxxx

Keywords:
Bioinformatics
Big data management
Hadoop
Data-intensive computing
Metagenomics

ABSTRACT

Up-to-date meta-databases are vital for the analysis of biological data. However, the current exponential increase in biological data leads to exponentially increasing meta-database sizes. Large-scale meta-database management is therefore an important challenge for production platforms providing services for biological data analysis. In particular, there is often a need either to run an analysis with a particular version of a meta-database, or to rerun an analysis with an updated meta-database. We present our GeStore approach for biological meta-database management. It provides efficient storage and runtime generation of specific meta-database versions, and efficient incremental updates for biological data analysis tools. The approach is transparent to the tools, and we provide a framework that makes it easy to integrate GeStore with biological data analysis frameworks. We present the GeStore system, an evaluation of the performance characteristics of the system, and an evaluation of the benefits for a biological data analysis workflow.

1. Introduction

Recent advances in scientific instruments, such as next-generation sequencing machines, have the potential of producing data that provides views of biological processes at different resolutions and conditions, opening a new era in molecular biology and molecular medicine [1]. Many of the data analysis techniques developed for analyzing such biological data integrate data from many experiments with metadata from multiple knowledge bases. The information in the meta-databases [2] is essential for understanding the biological content of the experiment data. For example, the results of DNA sequencing may not become truly useful before the UniProtKB [3] meta-database is used to map sequence bases to genes, the gene expression results are compared to results from other experiments, and the differences in expression values have been mapped to biological functions using the GO [4] meta-database.

The low cost of next-generation sequencing machines and other biotechnological instruments has caused an exponential growth of biological data [6]. Analysis of all this data produces many results, which are added to meta-databases such as UniProtKB. Such meta-databases are frequently updated and therefore growing rapidly (Fig. 1). For example, the June 2015 release of UniProtKB/TrEMBL contains 48.744.721 entries and is 137 GB in size. Compared to the previous May 2015 release, the number of entries increased by 5%, and 45% of the entries were updated. Each update may provide novel insights when reanalyzing old experiment data [7]. Updating experiment data with new meta-data is especially important for servers that provide search analysis services based on integrated data analysis [8].

For many analyses, it is also important that a specific meta-database version is used. For example, it is common to compare analysis results against gold standard results that are calculated using a specific meta-database version.

There are four main requirements for an infrastructure system that maintains large-scale biological meta-databases. First, multiple versions of the meta-database must be maintained to ensure repeatability of the analysis. Such repeatability is a cornerstone

* Corresponding author.
E-mail addresses: edvard.pedersen@uit.no (E. Pedersen), larsab@cs.uit.no (L.A. Bongo).

http://dx.doi.org/10.1016/j.future.2016.02.010
0167-739X/

Please cite this article in press as: E. Pedersen, L.A. Bongo, Large-scale biological meta-database management, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.02.010

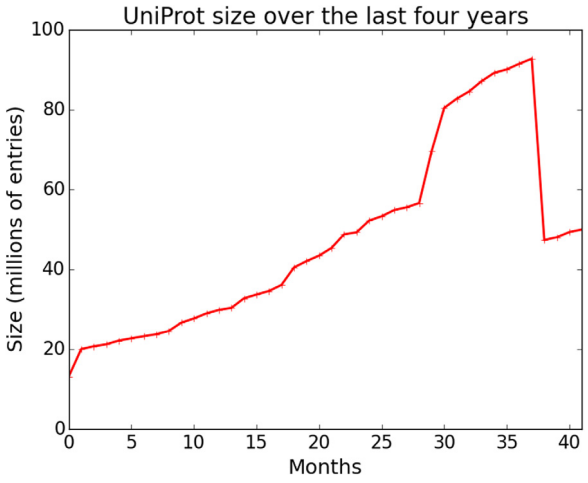

Fig. 1. Number of entries in UniProtKB from July 2011 to June 2015. The dip in early 2015 is due to removal of redundant proteomes [5].

in the scientific process, but has often been hard to achieve in bioinformatics [9]. Second, the system should enable efficient methods for integrating biological compendium with new or updated meta-data, since the computational cost of the integration can be orders of magnitude larger than the cost of producing the data [10]. Third, the system must be transparent to data analysis tools since it is not practical to implement and maintain modified versions of the many analysis tools used in biological data analysis [10]. Fourth, the system must easily integrate with biological data analysis frameworks to ensure adaptation in production systems.

Current popular biological data analysis frameworks such as Galaxy [9], Taverna [11], and Bioconductor [12] do not satisfy the first two requirements, since the user manually maintains and specifies meta-data versions. In addition, meta-database updates typically require re-executing the analysis for each meta-data update. Such full updates increase the computational cost, often to the point where reanalysis is not done.

Incremental update systems [13] for large-scale data [14–19] solve the first two requirements. These systems maintain several versions of the experiment data compendia and meta-databases, and greatly reduce the cost of reanalysis by using incremental updates that limits the computation to new and updated data. However, they do not provide a transparent approach for adding incremental updates to existing biological analysis workflows. Instead, they require either porting applications to a specific framework (such as Dryad [20], MapReduce [21], or Spark [22]) or implementing ad hoc scripts for input generation and output merging.

Data warehouse approaches for biological data, such as Turcu et al. [23], may provide incremental updates for specific tools, but do not easily allow adding new tools, nor integrating with biological data analysis frameworks.

We use the GeStore system [24] for large-scale biological meta-database management. It satisfies all four requirements listed above. GeStore provides an efficient transparent file based approach for incremental updates. It uses HBase to implement distributed data structures with efficient compression for multiple versions of large meta-databases. It uses Hadoop MapReduce for scalable parallel generation of specific database versions and increments. The transparent approach enables easy integration of GeStore with data processing frameworks, and does not require any changes to data analysis tools. Our contributions are threefold:

1. We describe the design and implementation of a system for large-scale biological meta-database management.

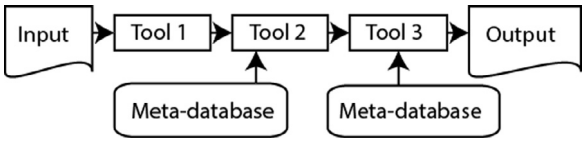

Fig. 2. A biological data analysis workflow.

2. We demonstrate how the approach can be integrated with biological data analysis frameworks with minimal changes to the framework code, and no changes to data analysis tools.
3. We present experimental evaluation of the performance, overhead and resource usage of the approach using a biological analysis workflows and real large-scale meta-databases.

We find that large-scale biological meta-databases can be efficiently maintained using data-intensive computing systems, and that our approach can easily be integrated with biological data analysis frameworks.

2. Background

We provide a background describing biological data analysis implementation, configuration, and execution. Further examples can be found in [25,26].

2.1. Data analysis workflows

A computer system for analyzing biological data typically consists of four main components: input data, meta-data, a set of tools in a workflow, and finally output data for interactive analysis (Fig. 2). Biotechnology instruments such as short-read sequencing machines produce the input data. The data can also be downloaded from public repositories such as GEO [27] and ENA [28]. There are hundreds of meta-databases with human or machine curated meta-data extracted from the published literature and analysis of experimental data [2]. The datasets and databases range in size from megabytes to petabytes.

A series of tools process the data in a pipeline where the output of one tool is the input to the next tool. The data transformations include file conversion, data cleaning, normalization, and data integration. There are many libraries [9,12,29] with hundreds of tools, ranging from small, user-created scripts to large, complex applications. A specific biological data analysis project often requires a deep workflow that combines many tools [30].

2.2. Workflow managers

The analyst specifies, configures, and executes the workflow using a workflow manager. The workflow manager provides a way of specifying the tools and their parameters, management of data and meta-data, and execution of the tools. In addition, a workflow manager may enable data analysis reproducibility by maintaining provenance data such as the version and parameters of the executed tools. It may also maintain the content of input data files, meta-databases, output files, and possibly intermediate data.

A workflow manager may comprise of a set of scripts run in a specific platform, or a system that maps high-level workflow configuration to executable jobs for many platforms. There are also managers that provide a GUI for workflow configuration, and a backend that handles data management and tool execution.

2.3. Hardware platforms

The workflow manager typically executes the workflow on a fat server, high performance computing clusters, or a data-intensive

Please cite this article in press as: E. Pedersen, L.A. Bongo, Large-scale biological meta-database management, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.02.010