The role of parallelism, web services and ontologies in bioinformatics and omics data management and analysis

Mario Cannataro✉, **Pietro Hiram Guzzi**

University of Catanzaro, Italy

Received 14 July 2013; **Accepted** 12 September 2013; *Published* 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

The increasing availability of omics data poses new challenges to bioinformatics applications regarding the efficient storage and integration of experimental data, their efficient and high-throughput preprocessing and analysis, the building of reproducible "in silico" experiments, the integration of analysis results with pre-existing knowledge repositories stored into ontologies, like Gene Ontology, or into specialised databases, as those available in pharmacogenomics. This paper presents an overview of how parallelism, service orientation, and knowledge management techniques can be used to face those challenges presenting some recent bioinformatics tools and projects that employ such technologies in different stages of the bioinformatics analysis's pipeline.

Motivation and Objectives

The increasing availability of *omics* data poses new challenges to bioinformatics applications that need to face an overwhelming availability of raw data. Main challenges regard: (i) the efficient storage, retrieval and integration of experimental data; (ii) their efficient and high-throughput preprocessing and analysis; the building of reproducible "in silico" experiments; (iii) the integration of analysis results with pre-existing knowledge repositories stored into ontologies or into specialized databases.

This paper presents an overview of how parallelism, service orientation, and knowledge management techniques can be used to face those challenges presenting some recent bioinformatics tools and projects that employ such technologies in different stages of the bioinformatics analysis's pipeline.

Methods

Main *omics* disciplines are gaining an increasing interest in the scientific community due to the availability of high throughput platforms and computational methods which are producing an overwhelming amount of *omics* data.

The increased availability of *omics* data poses new challenges both for the efficient storage and integration of the data and for their efficient preprocessing and analysis.

Hence, managing *omics* data requires both support and spaces for data storing as well as procedures and structures for data preprocessing, analysis, and sharing. The resulting scenario comprises a set of methodologies and bioinformatics tools, often implemented as web services, for the management and analysis of data stored in geographically distributed biological databases.

As the storage, preprocessing and analysis of raw experimental data is becoming the main bottleneck of the analysis pipeline, due to the increasing size of experimental data, **high-performance computing is playing an important role in all steps of the life sciences research pipeline**, from raw data management and processing, to data integration and analysis, up to data exploration and visualization.

Web services and workflows are used to face the complexity of the bioinformatics pipeline that comprises several steps. Finally, **ontologies** and knowledge management techniques **are used to connect pre-existing knowledge** in biology and medicine **to the omics experimental data and analysis results**.

We present an overview of how parallelism, service orientation, and knowledge management techniques can be used to face those challenges presenting some recent bioinformatics tools and projects that employ such technologies in different stages of the bioinformatics analysis's pipeline, with special focus on the analysis of *omics* data. Moreover, we briefly introduce some recent emerging architectures (multicore systems, GPUs) and programming models (MapReduce, Cloud Computing) that will have a key role to face the overwhelming volumes of data generated by *omics* platforms.

Results and Discussion

The role of parallelism, web services and ontologies in bioinformatics

In these last years, both well-known high performance computing techniques such as Parallel and Grid Computing, as well as emerging computational models such as Graphics Processing and Cloud Computing, are more and more used in bioinformatics and life sciences (Cannataro, 2009).

The huge dimension of experimental data is the first reason to implement large distributed data repositories, while high performance computing is necessary both to face the complexity of bioinformatics algorithms and to allow the efficient analysis of huge data. In such a scenario, novel parallel architectures (e.g. multicore systems, GPU, FPGA, hybrid CPU/FPGA, CELL processors) coupled with emerging programming models (e.g. Service Oriented Architecture, MapReduce) may overcome the limits posed by conventional computers to the mining and exploration of large amounts of data.

On the other hand, the modeling of complex bioinformatics applications as collections of web services composed through workflows, is an emerging approach to face the high complexity of bioinformatics applications and enabling the repeatability of "in silico" experiments, and thus the reproducibility of the same experiment by different research groups.

Workflows are used to combine such web services forming reusable bioinformatics applications that may be deployed on several distributed or parallel architectures, such as Grids or clusters. Moreover, using parameter sweep technology, a single workflow may be instantiated in various forms to test in parallel different algorithms on some of the steps of the bioinformatics pipeline.

Knowledge management techniques and especially ontologies are more and more used to model pre-existing knowledge in medicine and biology. For instance Gene Ontology[1] (GO) is used to annotate experimental data or results data with external information.

Ontologies are not only useful to annotate data, but also to support the composition of bioinformatics workflows. By modeling the application domain of a bioinformatics application and the analysis techniques used to analyse data, ontologies may be used to guide the development of bioinformatics workflows, suggesting tools needed to implement specific steps of preprocessing or analysis or alerting the user when some constraints are going to be violated, e.g. when the user tries to apply a wrong preprocessing tool or a wrong sequence of tools.

Parallel preprocessing of gene expression microarray data.

The dimension of microarray datasets is becoming very large since the dimension of files encoding a single chip and the number of the arrays involved in a single experiment, are increasing. The system developed in (Guzzi and Cannataro, 2010a) uses a master/slave approach, where the master node computes partitions of the input dataset (i.e., it sets a list of probesets intervals) and calls in parallel several slaves each one wrapping and executing the apt-probeset-summarize program, that is applied to the proper partition of data. Such system showed a nearly linear speedup up to 20 slaves.

Web Services-based preprocessing of gene expression microarray data

micro-CS (Microarray Cel file Summarizer) (Guzzi and Cannataro, 2010b) is a distributed tool for the automation of the microarray analysis pipeline that supports the automatic normalization, summarization and annotation of Affymetrix binary data, providing a web service that collects on behalf of the user the right and most updated libraries.

Workflow-based preprocessing and analysis of mass spectrometry-based proteomics data

The analysis of mass spectrometry proteomics data requires the combination of large storage systems, effective preprocessing techniques, and data mining and visualisation tools. The management and analysis of huge mass spectra produced in different laboratories can exploit the services of computational grids that offer efficient data transfer primitives, effective management of large data stores, and large computing power.

MS-Analyzer (Cannataro, 2007) is a software platform that uses ontologies and workflows to combine specialized spectra preprocessing algorithms and well known data mining tools, to analyze mass spectrometry proteomics data on the Grid. Data mining and mass spectrometry

[1] http://www.geneontology.org/

ontologies are used to model: (i) biological databases; (ii) experimental data sets; (iii) and bioinformatics software tools.

MS-Analyzer uses the Service Oriented Architecture and provides both specialised spectra management services and publicly available data mining and visualisation tools. Composition and execution of such services is performed through an ontology-based workflow editor and scheduler, and services are classified with the help of the ontologies.

Ontology-based annotation and querying of protein interaction data

Protein-protein interaction (PPI) databases store interactions among proteins and offer to the user the possibility to retrieve data of interest through simple querying interfaces. Thus, even simple queries like "retrieve all the proteins related to glucose synthesis" are usually hard to express.

OntoPIN[2] (Cannataro *et al.*, 2010) is a software platform that uses ontologies for automatically annotating proteins interactions and for querying the resulting annotated interaction data. OntoPIN includes a framework able to extend existing PPI databases with annotations extracted from GO and an interface for querying the annotated PPI database using semantic similarity in addition to key-based search.

Semantic similarity-based visualisation of protein interaction networks

The use of such annotations for the analysis of protein data is a novel research area. Semantic similarity measures evaluate the similarity of two or more terms belonging to the same ontology, thus they may be used to evaluate the similarity of two genes or proteins measuring the similarity among the terms extracted from the same ontology and used to annotate them (Guzzi *et al.*, 2012).

Recently, we used semantic similarity measures among proteins to develop a novel visualization method for protein interaction networks implemented into CytoSevis[3], a plugin of Cytoscape[4]. CytoSevis visualizes protein interaction networks in a semantic similarity space (Guzzi and Cannataro, 2012). CytoSevis exploits semantic similarity analysis and provides a graphical user interface that enables the visualization of networks in such a semantic space.

Emerging architectures and programming models

High performance computing is more and more used in biology, medicine and bioinformatics, to face the increasing amount of available experimental data. In the following we briefly introduce some emerging architectures and programming models that will have a key role to face the overwhelming volumes of data generated by *omics* platforms.

Multi-core and many-core systems

A multi-core processor is a single computing element containing two or more independent CPUs, called "cores", which read and execute program instructions in parallel. A multi-core processor usually comprises two, four, six or eight independent processor cores on the same silicon chip and connected through an on-chip bus.

Multi-core processors execute threads concurrently and often use less power than coupling multiple single-core processors. On the other hand, when increasing the number of cores the on-chip bus becomes a bottleneck, since all the data travel through the same bus, limiting the scalability of multi-core processors. Many-core processors put more cores in a thermal container than the corresponding multi-core processors.

General Purpose Graphics Processing Units

The Graphics Processing Unit (GPU) is a specialised electronic device initially used to accelerate the building of images to be sent to a display. The term General Purpose GPUs (GPGPU) indicates GPUs that are used for general-purpose computation. GPGPUs are mainly used for embarrassingly parallel computations and they are well suited to applications that exhibit large data-parallelism. One of the main manufacturers of GPUs and GPGPUs is NVIDIA.

MapReduce and Apache Hadoop

MapReduce is a recent programming model well suited for programming embarrassingly parallel applications that need to process large volumes of data. MapReduce is also the name of the Google programming model (Dean and Ghemawat, 2008). The MapReduce model is inspired by the **map** and **reduce** functions used in functional programming. A very popular free

[2] http://www.ontopin.org/
[3] http://sites.google.com/site/cytosevis/
[4] http://www.cytoscape.org/

implementation of MapReduce is Apache Hadoop[5].

Cloud Computing

Cloud Computing allows to access computers, services and eventually infrastructures as a utility, through the Internet. Developers can build novel Internet services without the need to buy large and costly hardware to deploy them as well as the human expenses to operate them (Ahmed *et al.*, 2012). Cloud Computing encompasses technology, economics and business model aspects so finding a complete definition is an issue.

According to Ahmed *et al.* (Ahmed *et al.*, 2012): "Cloud computing is a way of leveraging the Internet to consume software or other information technology services on demand". Using Cloud computing both resources and costs are shared. Also Ahmed *et al.* conclude that Cloud computing is more a business model than a computing paradigm.

Cloud-based Bioinformatics

Clouds are more and more used to host and deploy bioinformatics applications. Amazon EC2 has made available two main bioinformatics datasets in its publicly available repository[6]: the Annotated Human Genome Data provided by ENSEMBL, and UniGene provided by the National Center for Biotechnology Information. Dudley and Butte (Dudley and Butte, 2010) point out that clouds not only can offer elastic computational power to bioinformatics applications, but the availability of instances of bioinformatics applications that are stored and shared in the cloud can make computational analyses more reproducible. Schatz *et al.* (Schatz *et al.*, 2010) report a list of bioinformatics resources made available through the cloud[7]. Recently, several cloud-based platforms for bioinformatics and biomedical applications have been deployed.

Conclusion

Parallelism, web services and ontology technologies are key tools for modern emerging bioinformatics applications. The paper discussed the role of such technologies in many case studies regarding especially the management, preprocessing and analysis of *omics* data, with special focus on genomics, proteomics and interactomics data. The discussion is completed through the presentation of several bioinformatics tools exploiting those technologies.

Future work will regard a more comprehensive assessment of such technologies through the definition of quantitative and qualitative application requirements that can be fulfilled by adopting those technologies.

References

Ahmed M, Chowdhury ASMR, *et al.* (2012) An Advanced Survey on Cloud Computing and State-of-the-art Research Issues. *International Journal of Computer Science* **9**, 201-207.

Cannataro M (2009) *Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare*, Medical Information Science Reference, IGI Global Press, Hershey, USA, May 2009.

Cannataro M, Guzzi PH, *et al.* (2007) Using Ontologies for Preprocessing and Mining Spectra Data on the Grid. *Future Generation Computer Systems* **23**(1), 55-60. doi:10.1016/j.future.2006.04.011

Cannataro M, Guzzi PH, Veltri P (2010) Using ontologies for querying and analysing protein-protein interaction data. *Procedia CS* **1**(1), 997-1004. doi:10.1016/j.procs.2010.04.110

Dudley JT, Butte AJ (2010) In silico research in the era of cloud computing. *Nature Biotechnology* **28**, 1181-1185. doi:10.1038/nbt1110-1181

Guzzi PH, Cannataro M (2010a) Parallel Pre-processing of Affymetrix Microarray Data. Euro-Par Workshops 2010, *Springer Lecture Notes in Computer Sciences* LNCS 6586, 2011, 225-232. doi:10.1007/978-3-642-21878-1_28

Guzzi PH, Cannataro M (2010b) μ-CS: An extension of the TM4 platform to manage Affymetrix binary data. *BMC Bioinformatics* **11**, 315. doi:10.1186/1471-2105-11-315.

Guzzi P, Cannataro M (2012) Cyto-Sevis: semantic similarity-based visualisation of protein interaction networks. *EMBnet. journal* **18**(A), 32-33.

Guzzi PH, Mina M, Guerra G, Cannataro M (2012) Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in Bioinformatics* **13**(5), 569-585. doi:10.1093/bib/bbr066

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race. Nature Biotechnology 28, 691–693. doi:10.1038/nbt0710-691

[5] http://hadoop.apache.org/

[6] http://aws.amazon.com/publicdatasets

[7] http://www.nature.com/nbt/journal/v28/n7/fig_tab/nbt0710-691_T1.html