High Performance Management and Analysis of Omics Data

Mario Cannataro

Department of Medical and Surgical Sciences,
University "Magna Græcia" of Catanzaro,
88100 Catanzaro, Italy
emails: cannataro@unicz.it

1. Introduction

Main omics disciplines, such as genomics, proteomics, and interactomics, respectively refer to the study of the genome, proteome and interactome of an organism. Such disciplines are gaining an increasing interest in the scientific community due to the availability of novel, high throughput platforms for the investigation of the cell machinery, such as mass spectrometry, microarray, next generation sequencing, that are producing an overwhelming amount of experimental omics data.

On the other hand, 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. Moreover, both raw experimental data and derived information extracted by raw data are more and more stored in various databases spread all over the Internet.

Thus 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.

Biological databases and bioinformatics tools are key tools for organizing and exploring omics data, however 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. Thus, high-performance computing may play an important role in all steps of the life sciences research pipeline, from raw data management and processing, to data integration and analysis, till data exploration and visualization [1].

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. 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.

The importance and increasing use of high performance computing for life sciences and bioinformatics is demonstrated by recent publications and workshops, such as the Euro-Par High Performance Bioinformatics and Biomedicine workshop, the ICCS Biomedical Bioinformatics Challenges to Computer Science. The rest of the paper describes some parallel and distributed bioinformatics tools for the preprocessing and analysis of omics data.

2. micro-CS (Microarray Cel file Summarizer)

micro-CS (Microarray Cel file Summarizer) [2], is a distributed tool for the automatization of the microarray analysis pipeline. It supports the automatic normalization, summarization and annotation of Affymetrix binary data that require the use of specialized bioinformatics software (e.g. the Affymetrix Power Tools – APT) coupled with chip-specific libraries for summarization and annotation. micro-CS is based on a client-server architecture. The micro-CS client is provided both as a plug-in of the TIGR M4 (TM4) platform and as a Java standalone tool. It includes (via code wrapping) the binary code of the Affymetrix Power Tools and enables users to read, preprocess and analyse binary microarray data avoiding the manual loading of the Affymetrix libraries and the management of intermediate files. The micro-CS server that is implemented as a web service automatically updates the references to the summarization and annotation libraries that are made available to the micro-CS client. Thus users can preprocess microarray data without worrying about locating and invoking the proper preprocessing tools and chip-specific libraries.

3. MS-Analyzer

The analysis of mass spectrometry proteomics data requires the combination of large storage systems, effective preprocessing techniques, and data mining and visualization 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 [3] 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 ontologies are used model: (i) biological databases; (ii) experimental data sets; (iii) and bioinformatics software tools. MS-Analyzer uses the Service Oriented Architecture and provides both specialized spectra management services and public available data mining and visualization 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. MS-Analyzer has been used both as a standalone proteomics pipeline in oncology [4], and as a component of a comprehensive data analysis platform. In fact, MS-Analyser has been connected in a grid-enabled pipeline with the BioDCV machine learning platform for unbiased predictive analysis. We exploited the middleware and computing resources of the EGEE Biomed VO grid infrastructure thus providing a complete grid environment for proteomics data analysis [5].

4. IMPRECO

Protein complexes are a set of mutually interacting proteins that play a common biological role. The prediction of protein complexes in protein interaction networks is usually performed by searching small dense subgraphs. The performance of a prediction algorithm is influenced by: the initial configuration of the used clustering algorithm and the reliability of the protein-protein interactions. IMPRECO (IMproving PREdiction of COmplexes) is a tool that invokes in parallel different complexes predictors and combines their results using an integration algorithm that eventually produces novel predictions not present in the results of each predictor [6]. IMPRECO uses a distributed architecture that comprises the elementary predictors and the IMPRECO integration algorithm. The IMPRECO meta-predictor invokes in parallel different predictors wrapped as services, then integrates their results using graph analysis, and then evaluates the predicted results against external databases storing experimentally determined protein complexes.

5. OntoPIN

Protein-protein interaction (PPI) databases offer to the user the possibility to retrieve data of interest through simple querying interfaces [7]. Query inputs include: (i) one or more protein identifiers, (ii) a protein sequence, or (iii) the name of an organism. Results may consist of, respectively, a list of proteins that interact directly with the seed protein or that are at distance k from the seed protein, or the list of all the interactions of an organism. A main drawback is related to the fact that it is impossible to formulate even simple queries involving biological concepts, such as all the interactions that are related to a biological function. The OntoPIN tool [8] uses ontologies for annotating proteins interactions and for querying the resulting annotated interaction data. The OntoPIN project is based on:

- A framework able to extend existing PPI databases with annotations extracted from Gene Ontology, in particular annotations are extracted from the Gene Ontology Annotation Database (GOA) [9]. For each protein three kind of annotations are provided: biological process, cellular compartment, and molecular function.
- A system for querying the annotated PPI database using semantic similarity in addition to key-based search. Semantic queries may include: (i) protein identifier, (ii) molecular function annotation, (iii) cellular process annotation, (iv) cellular compartment.

References

1. Cannataro M. (Editor), Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, Medical Information Science Reference, IGI Global Press, Hershey, USA, May 2009. http://www.igi-global.com/reference/details.asp?ID=34292, ISBN: 978-1-60566-374-6
2. Pietro H Guzzi and Mario Cannataro, μ-CS: An extension of the TM4 platform to manage Affymetrix binary data, BMC Bioinformatics 2010, Volume 11:315, doi:10.1186/1471-2105-11-315, Published: 10 June 2010.
3. M. Cannataro, P. H. Guzzi, T. Mazza, P. Veltri, Using Ontologies for Preprocessing and Mining Spectra Data on the Grid, Future Generation Computer Systems – Special Issue on Data Mining in Grid Computing Environments, Vol. 23, n. 1, pp. 55-60, January 2007. (Available online 27 June 2006. http://dx.doi.org/10.1016/j.future.2006.04.011). ISSN=0167-739X.
4. Barbara Quaresima, Telma Crugliano, Marco Gaspari, Maria Concetta Faniello, Paola Cosimo, Rosa Valanzano, Maurizio Genuardi, Mario Cannataro, Pierangelo Veltri, Francesco Baudi, Patrizia Doldo, Giovanni Cuda, Salvatore Venuta, Francesco Costanzo, A proteomics approach to identify changes in protein profiles in serum of familial adenomatous polyposis patients, Cancer Letters 272(1):40-52. 2008. http://dx.doi.org/10.1016/j.canlet.2008.06.021
5. M. Cannataro, A. Barla, R. Flor, G. Jurman, S. Merler, S. Paoli, G. Tradigo, P. Veltri, C. Furlanello, A grid environment for high-throughput proteomics, IEEE Transaction on NanoBiosciences, 6(2):117-123. 2007. http://dx.doi.org/10.1109/TNB.2007.897495
6. M. Cannataro, Pietro H. Guzzi and P. Veltri, IMPRECO: Distributed prediction of protein complexes, Future Generation Computer Systems 26(3):434-440. 2010. http://dx.doi.org/10.1016/j.future.2009.08.001
7. Mario Cannataro, Pietro Hiram Guzzi, Data Management of Protein Interaction Networks, Wiley-IEEE Computer Society Press, Wiley Book Series on Bioinformatics, USA, (December 2011), ISBN-10: 0470770406, ISBN-13: 978-0470770405.
8. Mario Cannataro, Pietro Hiram Guzzi, Pierangelo Veltri: Using ontologies for querying and analysing protein-protein interaction data. Procedia CS 1(1): 997-1004 (2010). International Conference on Computational Science (ICCS 2010), Amsterdam, The Netherland, May 31- June 2, 2010. http://dx.doi.org/10.1016/j.procs.2010.04.110
9. Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucl. Acids Res., 32 (suppl 1):D262–266, January 2004.