Distributed Management and Analysis of Omics Data

Mario Cannataro* and Pietro Hiram Guzzi

Department of Medical and Surgical Sciences, Bioinformatics Laboratory,
University Magna Græcia of Catanzaro, Italy
{cannataro,hguzzi}@unicz.it

Abstract. The omics term refers to different biology disciplines such as, for instance, genomics, proteomics, or interactomics. The suffix -ome is used to indicate the objects of study of such disciplines, such as the genome, proteome, or interactome, and usually refers to a totality of some sort. This paper introduces omics data and the main computational techniques for their storage, preprocessing and analysis. The increasing availability of omics data due to the advent of high throughput technologies poses novel issues on data management and analysis that can be faced by parallel and distributed storage systems and algorithms. After a survey of main omics databases, preprocessing techniques and analysis approaches, the paper describes some recent bioinformatics tools in genomics, proteomics and interactomics that use a distributed approach.

Keywords: Omics Data, Genomics, Proteomics, Interactomics, Distributed Computing.

1 Introduction

The omics term refers to different biology disciplines such as, for instance, genomics, proteomics, or interactomics. The suffix -ome is used to indicate the objects of study of such disciplines, for instance the genome, proteome, or interactome, and usually refers to a totality of some sort. Main omics disciplines are thus genomics, proteomics, and interactomics, that respectively study the genome, proteome and interactome. The term omics data is used here to refer to experimental data regarding the genome, proteome or interactome of an organism.

The development of novel technologies for the investigation of the omics disciplines had caused the increased availability of omics data. Consequently the need of both support and spaces for data storing as well as procedures and structures for data exchanging arises. The resulting scenario is thus characterized by the introduction of a set of methodologies and tools enabling the management of data stored in geographically distributed databases using distributed tools often implemented as services.

* Corresponding author.

M. Alexander et al. (Eds.): Euro-Par 2011 Workshops, Part II, LNCS 7156, pp. 43–52, 2012.

Distribution of data may improve data availability allowing scalability in terms of data and users, parallel data manipulation from different users allows to improve the overall knowledge stored in distributed databases and of course enhances performance.

Main requirements of distributed management of omics data are:

- the introduction of a common shared data model able to capture both raw data of the experiment and related metadata;
- the definition of an uniform and widely accepted access and manipulation strategy for such large datasets;
- the design of algorithms that are aware of data distribution and thus may improve their performance;
- the design of ad-hoc infrastructures for efficient data transfer.

For instance the distributed processing of protein interaction data involves the following activities: (i) *Sharing and dissemination of PPI data among different databases*; (ii) *Collection of data stored in heterogeneous databases*; and (iii) *Parallel and distributed analysis of data.*

The first activity requires the development of both standards and tools to manage the process of data curation and exchange between interaction databases. Currently there is an ongoing project, namely the International Molecular Exchange Consortium (IMEx)[1], that aims to standardize the exchange of interactomics data. The second activity requires to solve the classical bioinformatic problem of linking identical data identified with different primary keys. Finally the rationale for the third activity is due to the algorithmic nature of problems regarding graphs. A big class of algorithms that mine interaction data can be re-conducted to classical problems of graph and subgraph isomorphism that are computationally hard. So the need for high-performance computational platforms as well as parallel algorithms arises.

The rest of the paper is structured as follows. Section 2 discusses the management issues of omics data and presents some omics databases. Section 3 recalls main techniques for analysing omics data, while Section 4 describes some parallele and distributed bioinformatics tools for the analysis of omics data. Finally. conclusions and future work are reported in Section 5.

2 Management of Omics Data

2.1 Genomics Databases

These databases store information about the primary sequence of proteins. Each sequence is generally annotated by several information, e.g. the name of the scientist that discovered the sequence or about the post translational modifications. User can query these databases by using a protein identifier or a fragment of sequence in order to retrieve the most similar proteins.

[1] `http://imex.sourceforge.net`

The EMBL Nucleotide Sequence Database[2] [5], maintained at the European Bioinformatics Institute (EBI) collects nucleotide sequences and annotations from public available sources. The database, involved in an international collaboration, is synchronized with DDBJ (DNA Data Bank of Japan) and GenBank (USA) (see next Sections). Core data are the protein and nucleotide sequences. The Annotations section of this database describes the following items: (i) Function(s) of the protein; (ii) Post-translational modification(s); (iii) Domains and sites; (iv) Disease(s) associated with deficiencies; and (v) Secondary structure.

The GenBank database[3] [4] stores information about nucleotide sequences maintained by the National Center of Biotechnology Information (NCBI). GenBank entries are structured as flat files (like the EMBL database) and share the same structure with EMBL and DDBJ. All the entries are grouped following both taxonomic and biochemical criteria. GenBank is accessible through a web interface. Through the ENTREZ system, the entries of GenBank are integrated with many datasources, enabling the search of information about proteins and its structures, as well as literature about the functions of genes.

Finally, the UniProt [11] consortium is structured on three main knowledge bases: (i) UniProt (also referred to as UniProt Knowledge base), that is the main archive storing information about protein sequences and annotations extracted from Swiss-Prot, TrEMBL and PSD-PIR; (ii) UniParc (Uniprot archive) that contains information about proteins extracted from the main publicly available archives; and (iii) UniRef (Uniprot reference), a set of databases that organize entries of UniProt by their similarity sequence, e.g. the UniRef90 groupes in a single record entries of UniProt that present at least the 90% of sequence similarity.

2.2 Proteomics Databases

The Global Proteome Machine Database[4] [12] was constructed to utilize the information obtained by the different servers included into the Global Proteome Machine project (GPM), to validate peptide MS/MS spectra and protein coverage. GPM is a system for analyzing, storing, and validating proteomics information derived from tandem mass spectrometry. The system is based on a relational database, on different servers for data analysis, and on a user-friendly interface to retrieve and analyze data. This database has been integrated into GPM server pages. The gpmDB data model is based on a modification of the Hupo-PSI Minumun Information About Proteomic Experiment (MIAPE) [16] scheme. System is available both through a web interface and as a stand alone application allowing users to compare their experimental results with the other ones that have been previously observed by other scientists.

PeptideAtlas[5] [13] is a database that aims to annotate the human genome with protein-level information. It contains data coming from identified peptides

[2] `http://www.ebi.ac.uk/embl`
[3] `http://www.ncbi.nlm.nih.gov/genbank`
[4] `http://www.thegpm.org/GPMDB/index.html`
[5] `http://www.peptideatlas.org/overview.php`

analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) and thus mapped onto the genome. PeptideAtlas is not a simple repository for mass spectrometry experiments, but uses spectra as primary information source to annotate the genome, combining different information. Consequently the population of this database involves two main phases: (i) a proteomic phase in which samples are analyzed through LC-MS/MS, and resulting spectra are mined to identify the contained peptides, (ii) an in silico phase in which peptides are processed by applying a bioinformatic pipeline and each peptide is used to annotate a genome. Resulting derived data, both genomics and proteomics, are stored in the PeptideAtlas database.

2.3 Interactomics Databases

The accumulation of protein interaction data caused the introduction of several databases [6]. Here we distinct on **databases of experimental determined interactions**, that include all the databases storing interactions extracted from both literature and high-throughput experiments, and databases of **predicted interactions** that store data obtained by in silico prediction. Another important class that we report is constituted by **integrated databases** or meta-databases, i.e. databases that aim to integrate data stored in other publicly available datasets. Currently, there exist many databases that differ on biological and information science criteria: the covered organism, the kind of interactions, the kind of interface, the query language, the file format and the visualization of results.

Although the existence of many databases the resulting amount of data presents three main problems [10]: (i) the low overlap among databases, (ii) the resulting lack of completeness with respect to the real interactome, and (iii) the absence of integration. Consequently, in order to perform an exhaustive data collection, (e.g. for an experiment), researchers should manually query different data sources. This problem is faced with the introduction of databases based on the integration of existing ones. Nevertheless, in the interactomics field, the integration of existing databases is a complex problem not yet completely solved.

In such a scenario many different laboratories are producing data by using different experimental techniques. Then, data can be modeled as a graph and stored in repositories by using different technical solutions. Finally, data stored in such databases can be mined to derive novel interactions or to extract functional modules, i.e subgraphs of the PPI network that have a biological meaning.

The distributed processing of protein interaction data consequently involves the following activities: (i) *Sharing and dissemination of PPI data among different databases*; (ii) *Collection of data stored in heterogeneous databases*; and (iii) *Parallel and distributed analysis of data.*

The first activity requires the development of both standards and tools to manage the process of data curation and exchange between interaction databases. Currently there is an ongoing project, namely the International Molecular Exchange Consortium (IMEx)[6], devoted to build an enabling framework for data

[6] `http://imex.sourceforge.net`

exchange. It is based on an existing standard for protein interaction data, the HUPO PSI-MI format. Databases that participate in this consortium accept the deposition of interaction data from authors, helping the researchers to annotate the dataset through a set of ad hoc developed tools.

The second activity requires to solve the classical bioinformatic problem of linking identical data identified with different primary keys. The cPath[7] tool [9] is an open source software for collecting and storing pathways coming from different data sources. From a technological point of view this software is an open source database integrated in a web application capable of collecting data from different data sources and exporting these data through a web service interface.

The third activity is related to the possibility of processing omics data in a parallel way. Issues include the development of parallel bioinformatics algorithms and the development of collaborative analysis platforms (collaboratories) where remote users can analise data in a collaborative way.

3 Omics Data Analysis

3.1 Microarray Data Analysis

The typical dimension of microrarray dataset is growing for two main reasons: the dimension of files generated when using a single chip and the number of the arrays involved in a single experiment are increasing. Let us consider, for instance, two common Affymetrix microarray files (also known as CEL files): the older Human 133 Chip CEL file that has a dimension of 5MB and contains 20000 different genes and the newer Human Gene 1.0 st that has a typical dimension of 10 MB and contains 33000 genes. Moreover a single array of the Exon family (e.g. Human Exon or Mouse Exon) can have up to 100 MB of size. Moreover the recent trend in genomics is to perform microarray experiments considering a large number of samples (e.g. coming from patients and controls) [1].

From this scenario, the need for the introduction of tools and technologies to process such huge volume of data in an efficient way arises. A possible way to develop the efficient preprocessing of microarray data is represented by the parallelization of existing algorithms on multicore architectures. In such a scenario the whole computation is distributed onto different processors, that perform computations on smaller sets of data and results are finally integrated. Such scenario requires the design of new algorithms for summarisation and normalisation that take advantage of the underlying parallel architectures. Nevertheless a first step in this direction can be represented by the replication on different nodes of existing preprocessing softwares that runs on smaller datasets.

Despite its relevance, the parallel processing of microarray data is a relatively new field. An important work is represented by affyPara [15], that is a Bioconductor package for parallel preprocessing of Affymetrix microarray data. It is freely available from the Bioconductor project. Similarly the μ-CS project presents a framework for the analysis of microarray data based on a distributed

[7] `http://cbio.mskcc.org/software/cpath`

architecture made of different web-services internally parallel for the annotation and preprocessing of data. Compared to affyPara, such an approach presents three main differences: (i) the possibility to realize more summarisation scheme such as Plier, (ii) the easily extension to newer SNP arrays, (iii) it does not require the installation of Bioconductor platform.

3.2 Mass Spectrometry Data Analysis

Mass Spectrometry-based proteomics is becoming a powerful, widely used technique in order to identify molecular targets in different pathological conditions. Classical bioinformatics tasks, such as protein sequence alignment, protein structure prediction, peptide identification, etc., are more and more combined with data mining and machine learning algorithms to obtain powerful computational platforms. Mass spectrometry produces huge volumes of data, said spectra, that may be affected by errors and noise due to sample preparation and instrument approximation. As a results preprocessing and data mining algorithms require huge amount of computational resources. The collection, storage, and analysis of huge mass spectra can leverage the computational power of Grids, that offer efficient data transfer primitives, effective management of large data stores (e.g. replica management), and high computing power.

3.3 Protein-to-Protein Interaction Data Analysis

Once that an interaction network is modeled by using graphs, the study of biological properties can be done using graph-based algorithms [6], and associating graph properties to biological properties of the modeled PPI. Algorithms for the analysis of local properties of graphs may be used to analyze local properties of PPIs networks, e.g. dense distribution of nodes in a small graph region may be associated to proteins (nodes) and interactions (edges) relevant to represent biological functions.

The rationale for the distributed analysis of PPI data is due to the algorithmic nature of problems regarding graphs. A big class of algorithms that mine interaction data may be faced using classical algorithms for solving the graph and subgraph isomorphism problems that are computationally hard. So the need for high-performance computational platforms arises. Currently, different softwares that mine protein interaction networks are available through web interfaces. For instance NetworkBlast[8] and Graemlin[9], that allow the comparison of multiple interaction networks are both available through a web-interface. Alignment algorithms usually employ different heuristics to face with the subgraph isomorphism problem. Although this, they are usually time consuming and the dimension of input data is still growing, so the development of high performance architectures will be an important challenges in the future.

[8] `http://www.cs.tau.ac.il/~bnet/networkblast.htm`
[9] `http://graemlin.stanford.edu`

4 Tools for Distributed Management of Omics Data

4.1 Micro-CS

μ-CS (Microarray Cel file Summarizer) [14], is a distributed tool for the automatic normalization, summarization and annotation of Affymetrix binary data. μ-CS is based on a client-server architecture. The μ-CS client is provided both as a plug-in of the TIGR M4 (TM4) platform and as a Java standalone tool and enables users to read, preprocess and analyse binary microarray data, avoiding the manual invocation of external tools (e.g. the Affymetrix Power Tools), the manual loading of preprocessing libraries, and the management of intermediate files. The μ-CS server automatically updates the references to the summarization and annotation libraries that are provided to the μ-CS client before the preprocessing. The μ-CS server is based on the web services technology. Thus μ-CS users can directly manage binary data without worrying about locating and invoking the proper preprocessing tools and chip-specific libraries. Moreover, users of the μ-CS plugin for TM4 can manage and mine Affymetrix binary files without using external tools, such as APT (Affymetrix Power Tools) and related libraries.

4.2 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 collection, storage and analysis of huge mass spectra produced in different laboratories can leverage the services of Computational Grids, that offer efficient data transfer primitives, effective management of large data stores, and large computing power. MS-Analyzer [7] is a software platform that uses ontologies and workflows to combine spectra preprocessing tools, efficient spectra management techniques, and off-the-shelf data mining tools to analyze proteomics data on the Grid. Domain ontologies are used model bioinformatics knowledge about: (i) biological databases; (ii) experimental data sets; (iii) bioinformatics software tools; and (iv) bioinformatics processes. MS-Analyzer adopts the Service Oriented Architecture and provides both specialized spectra management services and public available off-the-shelf data mining and visualization software tools. Composition and execution of such services is carried out through an ontology-based workflow editor and scheduler, and services are discovered with the help of the ontologies. Finally, spectra are managed by a specialized database.

4.3 IMPRECO

Starting from protein interaction data, a number of algorithm for the individuation of biologically meaningful modules has been introduced such as algorithms for prediction of protein complexes. Protein complexes are a set of mutually interacting proteins that play a common biological role. The individuation of

protein complexes in protein interaction networks is often made by searching small dense subgraphs. The performance of a prediction algorithm is therefore influenced by: (i) the kind and the initial configuration of the used algorithm, and (ii) the validity of the initial protein to protein interactions (i.e., reliability of edges in the graph representing of the input interaction network). IMPRECO (IMproving PREdiction of COmplexes) is a tool that combines the results of different predictors using an integration algorithm which is able to gather (partial) results from different predictors and eventually produce novel predictions [8]. IMPRECO is based on a distributed architecture that implements the IMPRECO integration algorithm and demonstrates its ability to predict protein complexes. The proposed meta-predictor first invokes different available predictors wrapped as services in a parallel way, then integrates their results using graph analysis, and finally evaluates the predicted results by comparing them against external databases storing experimentally determined protein complexes.

4.4 OntoPIN

PPI databases are often publicly available on the Internet offering to the user the possibility to retrieve data of interest through simple querying interfaces. Users, in fact, can conduct a search through the insertion of: (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. Often it is impossible to formulate even simple queries involving biological concepts, such as all the interactions that are related to glucose synthesis.

The OntoPIN project [2], conversely, demonstrates the effectiveness of the use of ontologies for annotating interaction starting from the annotation of nodes and the subsequent use for querying interaction data. The OntoPIN project is based on three main modules:

- A framework able to extend existing PPI databases with annotations extracted from ontologies: at the bottom of the proposed software platform there is an annotation module able to extend an existing PPI database with annotation extracted from the Gene Ontology Annotation Database [3] (GOA). For each protein three kind of annotations are currently provided: biological process, cellular compartment and molecular function.
- A system to annotate interactions starting from the annotations of interacting proteins: usually annotated databases contain annotations only for single proteins, nor for interactions. For instance, if the protein A is annotated with terms T_1, T_2, and T_3, and the protein B is annotated with terms T_1, T_2, T_4, and T_5, then the annotation of the interaction (A, B) is the common set: $\{T_1, T_2\}$.
- A system for querying such database using semantic similarity in addition to key-based search. The realized query interface supports the following querying parameters: (i) protein identifier, (ii) molecular function annotation, (iii)

cellular process annotation, (iv) cellular compartment. The user can insert a list of parameters that will be joined in a conjunctive way, i.e. the system will retrieve interactions whose participants are annotated with all the selected terms.

5 Conclusion and Future Work

Nowadays the efficient management and analysis of omics data has a big impact in molecular biology and is a key technology in genomics as well as in molecular medicine and clinical applications. The storage and analysis of omics data is becoming the bottleneck in this process, so 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. The huge dimension of experimental data is a 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. The paper introduced main omics data types and described the use of distributed management and analysis techniques along the whole pipeline of analysis, from data storage, to data analysis and knowledge extraction.

References

1. Guzzi, P.H., Cannataro, M.: Challenges in microarray data management and analysis. In: Proceedings of the 24th IEEE International Symposium on Computer-Based Medical Systems, Bristol, United Kingdom, June 27-30 (2011)
2. Cannataro, M., Guzzi, P.H., Veltri, P.: Using ontologies for querying and analysing protein-protein interaction data. Procedia CS 1(1), 997–1004 (2010)
3. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., Apweiler, R.: The GOA database in 2009–an integrated Gene Ontology Annotation resource. Nucleic Acids Research 37, D396–D403 (2009)
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: GenBank. Nucleic Acids Research 36(Database issue) (2008)
5. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31(1), 365–370 (2003)
6. Cannataro, M., Guzzi, P.H., Veltri, P.: Protein-to-protein interactions: Technologies, databases, and algorithms. ACM Comput. Surv. 43 (2010)
7. Cannataro, M., Guzzi, P.H., Mazza, T., Tradigo, G., Veltri, P.: Using ontologies for preprocessing and mining spectra data on the grid. Future Generation Comp. Syst. 23(1), 55–60 (2007)
8. Cannataro, M., Guzzi, P.H., Veltri, P.: Impreco: Distributed prediction of protein complexes. Future Generation Comp. Syst. 26(3), 434–440 (2010)
9. Cerami, E., Bader, G., Gross, B.E., Sander, C.: Cpath: open source software for collecting, storing, and querying biological pathways. BMC Bioinformatics 7(497), 1–9 (2006)

10. Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E.E., Futschik, M.E.: UniHI: an entry gate to the human protein interactome. Nucl. Acids Res. 35(suppl. 1), D590–D594 (2007)
11. The UniProt Consortium: The universal protein resource (UniProt) in 2010. Nucleic Acids Research 38(suppl. 1), D142–D148 (2010)
12. Craig, R., Cortens, J.P., Beavis, R.C.: Open source system for analyzing, validating, and storing protein identification data. Journal of Proteome Research 3(6), 1234–1242 (2004)
13. Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N., Aebersold, R.: The peptideatlas project. Nucleic Acids Research 34(suppl. 1), D655–D658
14. Guzzi, P.H., Cannataro, M.: mu-cs: An extension of the tm4 platform to manage affymetrix binary data. BMC Bioinformatics 11, 315 (2010)
15. Schmidberger, M., Vicedo, E., Mansmann, U.: Affypara: a bioconductor package for parallelized preprocessing algorithms of affymetrix microarray data
16. Taylor, C.F., Hermjakob, H., Julian, R.K., Garavelli, J.S., Aebersold, R., Apweiler, R.: The work of the human proteome organisation's proteomics standards initiative (HUPO PSI). OMICS 10(2), 145–151 (2006)