The IS-BioBank Project: A Framework for Biological Data Normalization, Interoperability, and Mining for Cancer Microenvironment Analysis

Michelangelo Ceci[1]
Department of Computer Science "
University of Bari Aldo Moro
Italy
ceci@di.uniba.it

Mauro Coluccia
Dept. of Biomedical Sciences and Human Oncology
University of Bari "Aldo Moro
Italy
Mauro.coluccia@dimo.uniba.it

Fabio Fumarola
Department of Computer Science "
- University of Bari Aldo Moro
Italy
ffumarola@di.uniba.it

Pietro Hiram Guzzi
Dept Surgical and Medical Sciences
Magna Graecia University of Catanzaro
Italy
hguzzi@unicz.it

Federica Mandreoli
Department of Information Engineering
·University of Modena
Italy
mandreoli.federica@unimore.it

Riccardo Martoglia
Department of Information Engineering
· University of Modena
Italy
riccardo.martoglia@unimo.it

Elio Masciari
ICAR-CNR
Rende
Italy
masciar@icar.cnr.it

Massimo Mecella
Dipartimento di Ingegneria Informatica
Automatica e Gestionale A Ruberti
Università La Sapienza
mecella@dis.uniroma1.it

Wilma Penzo
DEIS
University of Bologna
Wilma.penzo@unibo

ABSTRACT

Advances of high throughput technologies have yielded the possibility to investigate human cells of healthy and morbid ones at different levels. Consequently, this has made possible the discovery of new biological and biomedical data and the proliferation of a large number of databases. In this paper, we describe the IS-BioBank (Integrated Semantic Biological Data Bank) proposal. It consists of the realization of a framework for enabling the interoperability among different biological data sources and for ultimately supporting expert users in the complex process of extraction, navigation and visualization of the precious knowledge hidden in such a huge quantity of data. In this framework, a key role has been played by the Connectivity Map, a databank which relates diseases, physiological processes, and the action of drugs. The system will be used in a pilot study on the Multiple Myeloma (MM).

Keywords

Biological Data Normalization, Semantic Similarity, Connectivity Map, Multiple Myeloma.

1. INTRODUCTION

The emergence of high-performance computing systems, and the introduction of high throughput technologies for biological sample investigation are the basis of several projects aiming at establishing new public molecular profile data repositories disease investigation. In particular, the main efforts are devoted to the investigation of cancer data, for instance on clinical cancer and cultured cancer cell lines.

By using such resources, bio-medical researchers may publish and share their data and results and consequently may use the in-lab produced and public data to study a drug candidate, a gene or a disease state to verify hypothesis and generate new cognition.

Some examples of classical bio-technological databases and repositories are as follows: the National Center for Biotechnology

[1] Authors are in alphabetical order.

Information (NCBI) (http://www.ncbi.nlm.nih.gov/). States that host data about genome, proteins, nucleotides, genes, relationships between phenotype and genotype and more than 21 million citations from biomedical literature. Subsequently, other interesting project initiatives have appeared, each of which with the goal of providing useful information with respect to a particular viewpoint of complex biological systems. Gene Ontology [16] (GO focuses on gene product characteristics and gene product annotation data. GO allows users to inquire and to extract knowledge from the built ontology. The KEGG database stores a collection of online databases dealing with genomes and enzymatic pathways. The NCI-60 [14] database offers tools for storing, querying and downloading molecular profile data of 60 diverse human cancer cell lines to screen compounds for anticancer activity.

In addition to the above described ones, we here focus on the Connectivity Map project (hencefort CMap) [11], which is appeared with the challenge of establishing relationships among diseases, physiological processes, and the action of drugs according to the same language. The CMap provides a solution to this problem by; i) describing all biological states (physiological, disease, or induced with a chemical or genetic construct) in terms of genomic signatures, ii) creating a large public database of signatures of drugs and genes, and iii) developing pattern-matching tools to detect similarities among these signatures.

In this paper, we present an ongoing project exploiting the main potentialities of this bio-medical research trend towards a data-centric science by providing a knowledge support to the study of cancer microenvironments. With respect to CMap, the added value of our proposal consists in linking out CMap with the various types of data and partial knowledge stored in different data banks, including those cited above. By performing a comprehensive analysis of databases, data repositories, and ontologies, our ultimate goal is not to replicate existing data, but to design and develop a Web delivery system which:

1. Enables the interoperability among the questionable data sources,

2. Captures the different kinds of relationships that exist among them,

3. Reinforces the cooperation of heterogeneous and distributed databank sources for the query processing target,

4. Supports the users in the complex process of extraction, navigation and visualization of the knowledge hidden in such a huge quantity of data. In particular, to facilitate interoperability (i), we will focus on the normalization problem by creating a semantic layer linking the data sources (ii). On top, innovative algorithms and techniques for querying (iii), mining and visualizing data, models and statistics will enable the extraction of new knowledge (iv).

The main objective of the Web delivery system will be to assist bio-medical researchers in analyzing tumors microenvironments in order to understand them and identify relationships among tumors, the effect of drugs and the patients' biodiversity. Such relationships are of particular interest for drug repositioning (that is, understanding whether drugs typically used for treating some specific tumors can be used for treating other tumors since they report at a normal state the same genes) and for the identification of novel compounds able to overcome resistance or revert it. The system will be used in a pilot study on the Multiple Myeloma (MM), an incurable malignant plasma cell disease

2. OUR APPROACH

Our goal is to introduce an end-to-end system (whose architecture is depicted in Fig. 1) that would provide a technological support to this issue by exploiting the CMap and additional data sources. In this way, bio-medical researchers will be able to study Cancer microenvironments in order to understand their specificities and the effect of drugs considering the patients' biodiversity. In particular, it is possible to understand the relationships of one tumor with other tumors, to understand mechanisms of drug resistance in tumor Cells to drugs in current use, to elucidate the contribution of the tumor environment in conferring drug resistance, to identify candidate compounds for drug repositioning, to identify a set of candidate gene products for extensive study of their role in drug resistance.

To this end, we will pursue the following objectives: – Identification of the data repositories which relate to the CMap;

– Normalization and interoperability of the identified databases and the CMap;

– Use of Data Mining techniques for extracting useful knowledge from data;

– Use of techniques for semantic tagging of the CMap;

– Use of techniques for querying the extended CMap, the identified data repositories and the extracted knowledge as a unique data space;

– Use of Visual Query Languages for querying CMap and extracted knowledge.

Our aim is therefore to give answers to the problems that are mandatory for these objectives. In the following, we survey the main problems we foresee in this context and delineate the research directions toward their solution.

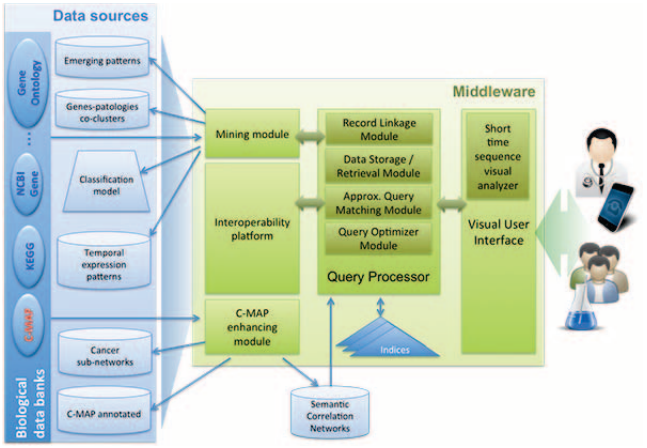

Figure 1: The System Architecture

.

2.1 Biological Data Gathering

Our first goal is to identify the data repositories which can be combined with the Connectivity Map, keeping in mind that the final goal is to allow bio-medical researchers to navigate the stored knowledge as well as to formulate new hypotheses based

on the information stored in the bio-technological data repositories.

As a matter of fact, several works in the literature aim at extending the information stored in the CMap [7,15]. However, they represent adhoc extensions. Our goal, instead, is to provide a systematic approach to extend the CMap and make the information it stores interoperable with data available in other bio-technological database such as NCBI (Gene, Geo, Pubmed), ArrayExpress, Gene Ontology, KEGG, and Drug Bank as reported in Fig. 1

2.2 Data Mining and Semantic Information Extraction

By taking into account the requirements coming from bio-medical researchers we will exploit several mining algorithms in order to extract knowledge from the selected data sources. The goal is to better understand tumors and to identify relationships among them, the effect of drugs and the patients' biodiversity.

These algorithms can significantly be benefited from the identification of semantic relatedness among entities. In particular, the results of the data mining algorithms can be used to enrich/confirm ontologies that can be used to support semantic-based querying. On the other hand, semantically tagged data can be used to identify relationships to be used in the mining processes.

By exploiting co-clustering approaches, it is possible to expose groups of genes whose gene expressions are simultaneously altered by one or more diseases [2]. To this purpose, hierarchical and non-hierarchical co-clustering techniques can be exploited. Moreover, in order to characterize and describe tumors (or classes of tumors) on the basis of the variability of genomic signatures observed in gene products, network based emerging patterns discovery algorithms [3] can be used. Furthermore, we plan to analyze evolution of diseases through short time series analysis techniques [4]. To this end, both visual data mining and temporal patterns extraction algorithms will be defined. Finally, in order to identify the disease's stage on the basis of expression gene values, collective classification algorithms and ensemble-based algorithms can be exploited. While in the case of collective classification [13] it is possible to handle the autocorrelation (according.[2]to which "closer" objects are more related than "furthermost" ones), typically present in data organized in network form (as those considered here, where genes are related to other genes, to diseases, to functional pathways and to miRNAs), in ensemble-based classification [12] different learning models will be combined together (ensemble) in order to define the final model. All the above mentioned mining algorithms will be implemented in the "Mining module" (see Fig. 1).

As regards data correlation analysis, starting from data stored in the CMap, the selected databases and appropriate ontologies, a Semantic Correlation Network will be built. This semantic network will be used to extract sub-networks related to Cancer through the application of network analysis algorithms such as network alignment algorithms [10], clustering [8], and pattern extraction algorithms [6]. The outcomes of this activity will be implemented in the "CMap enhancing module" of the Web delivery system.

2

2.3 Biotechnological data modeling and management

Once all the data repositories have been identified, additional problems raise. Indeed, the biological data to be analyzed are heterogeneous both in their type and format, since they come from several data sources exhibiting different schema. Moreover, another kind of information that is particularly useful for our goal is the knowledge provided by the mining activities. Again, it differs from the bio- logical databanks not only for the format but mainly for the adopted model as it refers to a mining model rather than operational ones. On the other hand, all the above mentioned data sources are inherently connected, thus the availability of normalization and interoperability solutions that allow analysis tools to deal with the information coming from different sources in a unified way is crucial.

In addition, solutions to enrich the CMap with the information gathered from the other biological data sources are necessary to use semantics to search or browse its data. In conclusion, a flexible query model is necessary that allows stake holders to inquire about the knowledge in the data sources in a consistent manner and to get useful results for analysis purposes. Thus, the main challenges for this goal are:

1. Extension of the CMap with semantic information encoded into ontologies;

2. Normalization and interoperability of the set of data sources;

3. Definition of techniques for effectively and efficiently supporting querying.

As regards the first challenge, RDF annotations to the CMap entities with the support of the selected ontologies will be introduced. The output will be stored in a relational database (CMap annotated, see Fig. 1) containing both entities and functional annotations extracted from ontologies, whereas the methods will be implemented in an ad hoc module, called Annotation Module. It will create the first version of CMap annotated, then it will periodically update it by searching for new annotations that can be extracted from publicly available databases.

The second challenge will be dealt with the aim of providing a technological platform to the full interoperability among the selected data banks and the outputs of the mining activities: CMap annotated; the sub-networks related to Cancer; the genes-pathologies co-clusters, the disease (emerging) patterns, temporal expression patterns (extracted from short time sequences) and the disease classification model. As to source participation, the platform will support two alternative options: external sources accessible through Internet querying services and local sources, which the system will have full control and accessibility on.

To this end, the platform will draw inspiration from data spaces [5]. Indeed, a data space follows a data co-existence approach and its main aim is to provide base functionality over all data sources, regardless of how integrated they are, thus shifting the emphasis to a data co-existence. To this end, an ontological language will support the specification of data sources in the form of data schema, mining model or query flow and a mapping language will be used for inter-source mapping specification. As to the latter, the platform will support the gradual specification of schema mapping between sources in a pay-as-you-go fashion [1].

The third challenge is faced through 1) the introduction of a flexible query language and an approximate query matching model that would allow stakeholders to easily query the system and to get useful results, 2) the definition of algorithms and data structures for approximate query answering that would ensure good performances under different system conditions.

The language allows users to specify queries as graphs of biological concepts, biological entities (data instances), predicates on biological entities, and labeled relationships among them. Moreover, it will extend the classical comparison operators with ad-hoc operators to query mined data and models. Query samples that could be specified are "Find all genes that are up-regulated and whose localization is similar to Nucleolus and function is similar to receptor-binding", "Find all the groups of similar genes whose localization is different from nucleolus that are downregulated under the effect of drug X", and many others.

Once a query is issued to the system, the query processor module will approximate the query on the data space by: 1) defining a query plan that selects the involved sources through the interoperability platform, 2) sending to each selected source the appropriate query, collecting and merging the query results through the application of record linkage techniques [9].

In order to support an efficient query processing on the data space, we will study appropriate data structures and algorithms that will support all the different peculiarities of the queried data and of the query language. To this end, two kinds of indices will concurrently be exploited in order to efficiently answer a given query: a) High-level indices; b) Low-level indices (one or more per source). Some of the high-level indices will help in efficiently filtering out unnecessary sources and accessing those having structural and or data properties compatible with the given query only. In this case, ideas from existing successful proposals in different and simpler single-source scenarios will be adapted and exploited (e.g. signatures and bloom filters). Other kinds of high-level indices will instead be helpful when merging answers coming from the queried sources. Some local sources will be equipped with one or more low-level indices, whose kind will depend on the source data. The peculiarities of involved indices and sources will be exploited by algorithms for access plan generation. In case of queries with a very large number of results, the project will study top-k algorithms to maximize the relevance of the results and to minimize the processing timeSubsequent Pages

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be of equal length.

2.4 The Web Delivery System

The Web delivery system will be implemented as a Service Oriented Infrastructure according to which Web-services enabling both access to data and usage of the defined algorithms will be provided. An important feature is the user- friendliness of the whole prototype for potential users. The Web delivery system will then be made accessible by means of a Visual User Interface module that provides biological data experts with a rich user experience during the usage of the tool, both in the querying phase and in the result manipulation phase. The main objective is then the definition of a visual query language specifically targeted to biological data sources analysis and of appropriate visualization techniques.

In order to fully explicate the goal, we intend to obtain let us consider the query and visualization interface of CMap. By using this tool, a query basically consists in providing a signature file and searching for connected objects. The main difficulties for users are in the text-based syntax of the signature file, which almost requires a kind of programming capabilities, as the syntax should be rigorous. In particular, nowadays the way of writing a query is to create an Excel file and to insert specific values into the columns, according to the given sheet format. Conversely, a graphical interface will be developed, in which the user, through drag&drop of the basic elements needed for building a signature (to be taken from a palette available to the user), is able to visually write such a signature and to use it for querying the system. In the same way, currently the results of the queries are viewed in a table format, and then for each of them a click allows for opening the related specification (again an Excel file). Conversely, a graph-based visualization is envisioned, in which results are shown as nodes of a graph, and the edges represent relationships (e.g., due to sharing of some objects in the structure). Different colors, thickness of the edges, etc. convey specific semantics. A more natural interaction modality will also allow for the use of the interface/tool by users equipped with modern devices, such as tablets, during their normal operations in laboratories. Footnotes should be Times New Roman 9-point.

Using the standard Communications of the ACM format for references – that is, a numbered list at the end of the article, ordered alphabetically by first author, and referenced by numbers in brackets [1]. View the examples of citations at the end of this document. Within this template file, use the style named references for the text of your citation.

References should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e. you can give the address to obtain the report within your citation) and may be obtained by any reader. Proprietary information may not be cited. Private communications should be acknowledged, not referenced (e.g., "[Robertson, personal communication]").

3. CASE STUDY ON MULTIPLE MIYELOMA

The web framework will be used in a pilot study on the Multiple Myeloma (MM), an incurable malignant plasma cell disease with an incidence of 5 per 100,000 inhabitants, and for that in NCBI GEO are submitted around 6658 samples. MM locates primarily to the bone marrow (BM) in multiple niches that provide a microenvironment which promotes tumor survival. In this context, the main aim of the system will be to allow bio-medical researches to avoid wasting time and funds for the in-vitro verification of potentially meaningless hypothesis by their testing with in silico techniques. Specifically, thanks to the system, it will be possible to drive the process of hypothesis generation in: 1) understanding the correlation of the MM with other tumors in terms of gene expressions modifications; 2) defining a characterization of the MM in terms of genes; 3) analyzing the evolution of the pathology; 4) automatically identify MM on the basis of gene expressions modifications and additional information stored in other data sources.

This analysis might help the drug repositioning task and the identification of novel compounds able to overcome resistance or revert it in the four drugs in current use (Dexamethasone, Bortezomib, High Dose Melphalan, Lenalidomide).

4. CONCLUSION

In this paper, we presented an ongoing work aiming at the realization of a system for biological and biomedical data normalization and interoperability. It is devoted to knowledge extraction, data querying and knowledge dissemination for supporting biomedical specialist. Currently, it is focused on the analysis of cancer microenvironments as first application case. The system is tailored on the biological data features in order to make it easy to use and provide useful information to the domain experts

.

5. REFERENCES

[1] K. Belhajjame, N. W. Paton, S. M. Embury, A. A. A. Fernandes, and C. Hedeler. Feedback-based annotation, selection and refinement of schema mappings for dataspaces. In Proc. of EDBT, pages 573–584, 2010.

[2] D.Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological net- works and gene expression data. In ISMB, pages 145–154, 2002.

[3] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In KDD, pages 43–52, 1999.

[4] J. Ernst and Z. Bar-Joseph. Stem: a tool for the analysis of short time series gene expression data. BMC Bioinformatics, 2006.

[5] Alon Y. Halevy, Michael J. Franklin, and David Maier. Principles of dataspace systems. In PODS, pages 1–9, 2006.

[6] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kauf- mann, 2000.

[7] Guanghui Hu and Pankaj Agarwal. Human disease-drug network based on genomic expression profiles. PLoS ONE, 4(8):e6536, 08 2009.

[8] A.K.Jain,M.N.Murty,andP.J.Flynn.Dataclustering:areview.A CMComputing Surveys, 31, September 1999.

[9] R. Karmel and D. Gibson. Event-based record linkage in health and aged care services data: a methodological innovation. BMC Health Services Research, 2007.

[10] M Cannataro, PH Guzzi, P Veltri. Protein-Protein Interactin Data: Technologies Databases and Algorithms. ACM Computing Surveys, 44 2010

[11] Justin Lamb. The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7(1):54–60, 2007.

[12] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11:169–198, 1999.

[13] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. AI Magazine, pages 93–106, 2008.

[14] Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer, 6(10):813–823, October 2006.

[15] Marina Sirota, Joel T. Dudley, Jeewon Kim, Annie P. Chiang, Alex A. Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Science Translational Medicine, 3(96):96ra77, 2011Anderson, R.E. Social impacts of computing: Codes of professional ethics. Social Science Computing Review, 2 (Winter 1992), 453-469.

[16] Guzzi P, Mina M, Guerra C, Cannataro M, Semantic Similarity Analysis of Protein Data: Assessmente with biological features and issues. Briefing in Bioinformatics, 44 2012

.

Authors

Michelangelo Ceci received a "laurea"degree with full marks and honors in Informatics from the University of Bari (March 2001). Following graduation, he kept on making research in machine learning and data mining in the Knowledge Acquisition and Machine Learning Laboratory (LACAM), Department of Informatics of the University of Bari. His main research is in Knowledge Discovery from Databases primarily in thedevelopment of Data Mining algorithms for predictive tasks(classification and regression).

Mauro Coluccia MD. He is associate professor at University of Bari where he leads the Experimental Chemioterapy Laboratory. His research interests are the development of drugs for the treatment of cancer. He is author of more than 50 paper published on international peer reviewed journal. He is co-inventor of many industrial patents.

Fabio Fumarola. He is a Ph.D. Student at University of Bari. Main research interests are: the analysis, design and development of a web-service oriented architecture for a knowledge discovery server. He currently focuses on Data Mining and Knowledge

Discovery,Data Stream Mining.

Pietro Hiram Guzzi. PhD. Pietro H. Guzzi is an Assistant Professor of Computer Engineering at the University 'Magna Græcia' of Catanzaro, Italy, since 2008. He received his PhD in Biomedical Engineering in 2008, from Magna Græcia University of Catanzaro. He received his Laurea degree in Computer Engineering in 2004 from the University of Calabria, Rende, Italy. His research interests comprise bioinformatics, the analysis of proteomics data, and the analysis of protein interaction networks. Pietro is an ACM member and serves the scientific community as reviewer for many conferences. He is associate editor of

Information Science journal, and of SIGBioinformatics Record.

Federica Mandreoli. She got the Laurea degree in Computer Science from the University of Bologna (Italy) in 1997. In 1997 she started my Ph.D. experience at DEIS (University of Bologna) and on march 2001 she received the Ph.D. degree in Computer

Science Engineering with a thesis entitled "Temporal Schema Versioning in Object-Oriented Databases". After a two-year position as researcher at DII (University of Modena) in the field of Information Retrieval, at present she is an associate researcher in the same department.

Riccardo Martoglia He got the Laurea degree in Computer Engineering from the University of Modena and Reggio Emilia (Italy) in march 2002. In 2003 he started my Ph.D. experience at DII (University of Modena and Reggio Emilia) in the field of Information Retrieval, Multi-lingual Information Management and Similarity Search and on february 2006 he received the Ph.D. degree in Computer Science Engineering with a thesis entitled "Information Retrieval Techniques for Pattern Matching". At present he is an associate researcher in the same department.

Elio Masciari Elio Masciari received his Laurea degree summa cum laude in Computer Engineering and his Ph.D. in Systems and Computer Engineering from the University of Calabria, Italy. Currently he is a Researcher at the ICAR institute of the Italian National Research Council. His research activity is mainly focused on techniques for the analysis and mining of structured and unstructured data, XML query languages and XML structural similarity.

Massimo Mecella Massimo Mecella holds a Ph.D. in Computing Science and Engineering since 2002, issued by SAPIENZA University of Rome. Currently he is Assistant Professor in the School of Information Engineering, Computer Science and of SAPIENZA, in the Department of Computer Control and Management Science & Engineering ANTONIO RUBERTI (DIS) Caruso. He has been technical manager of the European project WORKPAD (ended in 2009), currently he is technical manager of the European project SM4All, and partecipates / participated with important roles in the European projects GreenerBuildings, SmartVortex, SemanticGOV, EU-PUBLI.com, GENESIS, and in the Italian/national projects eG4M, MAIS, VISPO, DaQuinCIS. He collaborated in organizing the conferences WISE 2003, CoopIS 2001 and the workshop DQCIS 2003 (proceedings chair), he has been workshop chair of BPM 2008 and he regularly is in the program committees of the most important conferences of his area, including the series ICSOC, CoopIS, WISE, WETICE, CEC/EEE, DEECS, ISCRAM

Wilma Penzo. Her scientific activity is set in the area of Information and Knowledge Management, with particular interest in the study of techniques for the effective and efficient retrieval of information in large data collections, which are characterized by aspects such as distribution, heterogeneity in structural representation, ambiguity in the semantics of contents.