Toward a Semantic Framework for the Querying, Mining and Visualization of Cancer Microenvironment Data

Michelangelo Ceci[4], Fabio Fumarola[4], Pietro Hiram Guzzi[3], Federica Mandreoli[6], Riccardo Martoglia[6], Elio Masciari[1], Massimo Mecella[2], and Wilma Penzo[5]

[1] ICAR-CNR, Italy
[2] La Sapienza University, Italy
[3] Magna Graecia University, Italy
[4] University of Bari
[5] DEIS - University of Bologna, Italy
[6] DII - University of Modena and Reggio Emilia
ceci@di.uniba.it, fabiofumarola@gmail.com, hguzzi@unicz.it, {federica.mandreoli,martoglia.riccardo}@unimo.it, masciari@icar.cnr.it, mecella@dis.uniroma1.it, wilma.penzo@unibo.it

Abstract. Over the last decade, the advances in the high-throughput omic technologies have given the possibility to profile tumor cells at different levels, fostering the discovery of new biological data and the proliferation of a large number of bio-technological databases. In this paper we describe 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. The system will be used in a pilot study on the Multiple Myeloma (MM).

1 Introduction

The emergence of affordable high-performance computers, and the high-throughput omic technologies are the basis of several projects aiming at building new public molecular profile databases and data repositories on clinical cancer and cultured cancer cell lines. Using such new public databases, biomedical researchers can i) publish their data and results making them available to the scientific community, and ii) use the in-lab produced and public data to study a drug candidate, a gene or a disease state in a biological system in order to verify hypothesis and generate new knowledge. Major examples of public "bio-technological databases and repositories" are: the National Center for Biotechnology Information (NCBI) located in United States, the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). They host data about genome (sequences, maps, chromosomes, assemblies, and annotations), proteins, nucleotides, genes (reference sequences, maps, pathways,

C. Böhm et al. (Eds.): ITBAM 2012, LNCS 7451, pp. 109–123, 2012.

variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources), relationships between phenotype and genotype and more than 21 million citations from biomedical literature. Subsequently, other interesting project initiatives have come out, each of which with the goal of providing useful information with respect to a particular viewpoint of complex biological systems. Gene Ontology (GO) project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data. GO allows users to query and to extract knowledge from the built ontology. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database stores a collection of online databases dealing with genomes and enzymatic pathways. The KEGG pathway data bank records networks of molecular interactions in the cells, and variants of them specific to particular organisms. The DrugBank, the KEGG DRUG and the Chemical Entities of Biological Interest (ChEBI) databases offer different kinds of bioinformatics and cheminformatics resources that combine detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The U.S. National Cancer Institute (NCI) with the NCI-60 database offers tools for storing, querying and, downloading molecular profile data of 60 diverse human cancer cell lines used since 1990 to screen compounds for anticancer activity. In addition to the above described ones, a particularly interesting project is the Connectivity Map (CMap), which is born with the challenge of establishing relationships among diseases, physiological processes, and the action of drugs (small molecules). The CMap provides a solution to this problem by:

- describing all biological states (physiological, disease, or induced with a chemical or genetic construct) in terms of genomic signatures;
- creating a large public database of signatures of drugs and genes;
- developing pattern-matching tools to detect similarities among these signatures.

Prevalently, in the literature, CMap is queried by researchers using signatures obtained from comparative gene expression analysis (e.g. disease compared with normal state, treated drugs versus not treated ones) to identify drug response profile that either correlate or anti-correlate with it. When a signature is derived from clinical samples representing a disease, the discovered connections represent a list of sample-drugs which either mimics or reverts the disease signature, while a signature obtained from drug-treated cells can be used to retrieve a list of chemical compounds with similar effect. CMap plays a central role because it relates, from a genomic point of view, diseases, genes' functions and drugs' actions according to the same language.

The added value of the framework we propose in this paper is obtained by 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 aim is not to replicate existing data, but to design and develop a Web delivery system which:

1. enables the interoperability among the queryable data sources;
2. captures the different kinds of relationships that exist among them;
3. reinforces the cooperation of heterogeneous and distributed data bank 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 (1), we will focus on the normalization problem by creating a semantic layer linking the data sources (2). On top, innovative algorithms and techniques for querying (3), mining and visualizing data, models and statistics will enable the extraction of new knowledge (4). This would support 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 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 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 particular all the functionalities of our project will be exploited to understand the relationships of MM with other tumors, to understand mechanisms of drug resistance in MM Cells (MMCs) to 4 drugs (Dexamethasone, Bortezomib, High Dose Melphalan, Lenalidomide) in current use, to elucidate the contribution of the tumor environment in conferring drug resistance in MMCs, to identify novel compounds able to overcome resistance or revert it in MMCs to drugs in current use, to identify a set of candidate gene products for extensive study of their role in drug resistance in MMCs.

2 Background

In this section, we will discuss the preliminary concepts mandatory to our system implementation that emerged after a deep analysis performed with the support of biological data experts. In particular for each topic we will discuss in detail the state of the art and the limitations of current proposals that lead us to the system architecture described in Section 3.

2.1 The Connectivity Map

Biological data features made traditional approaches to data analysis inadequate for their efficient analysis and understanding thus a great effort has been devoted to research on these fields. A main problem is to integrate data coming from different data sources and obtained with different acquisition technologies. In this perspective, the Connectivity Map (CMap) plays an important role[34]. Briefly, it is a freely available search engine that may be used to retrieve information about diseases, drugs and gene expression levels. It stores raw data

about the gene expression published in different papers, and information about the impact of different drugs on the expression level. The main utility of CMap, as evidenced in recent papers, is the possibility to generate in silico hypothesis, by analyzing gene expression data. The usage of the CMap as a starting point for our activities is motivated by the increasing interest of the scientific community in the prediction of novel associations among diseases, physiological processes, and the action of small therapeutic molecules [3]. Several papers in the literature proposed different solutions in order to extend the information stored in the Connectivity Map with other bio-technological databases. Examples are [25],[35],[18],[39] and [44] where the CMap is extended by the definition of new genomic signatures for diseases and chemical compounds extracted from GEO Dataset and Series, new distance measures for chemical compounds and pathologies based on protein network interaction and PUBMED abstracts. However, all the above contributions represent ad-hoc extensions of the CMap. At the moment, there is no systematic approach which allows us to extend and integrate the information stored in the CMap with data available in other bio-tecnological database such as NCBI (Gene, Geo, Pubmed), ArrayExpress, Gene Ontology, KEGG, and Drug Bank. Finally, when normalizing data a crucial activity is data de-duplication. A proposed approach consists in the adoption of a hierarchical clustering method [29], equipped with a suitable record matching scheme, that leverages accurate field-wise similarity metrics to match corresponding record tokens. To overcome efficiency and effectiveness problems, the adoption of a hash-based index has been proposed[6].

2.2 Interoperability

The dataspace principles have been recently introduced in the literature[20] as a data management abstraction alternative to data integration where, unlike fully integrating heterogeneous data sources, the coexistence of data, which is autonomously modeled and loosely connected through relationships for sharing purposes, is supported. This new paradigm better fits the data scenario envisioned by the project, where a full control of data sources is not always available. Currently, a huge amount of biological data can be naturally represented by graphs. Several works have been proposed on the graph query problem, which has been extensively studied in the context of a graph database consisting of a set of relatively small graphs, while little attention has been paid to the context of a single large graph[47], which is the context common to most biological networks. The major challenge in this scenario is to reduce the number of pairwise subgraph isomorphism checkings, since subgraph isomorphism is known to be a NP-complete problem. A number of graph indexing techniques, where different structural patterns are examined to help prune the candidate search space, have been proposed to address this challenge (e.g. [24]). As to the problem of supporting approximate graph matching, which is a decisive feature to support queries on heterogeneous data, only few works have been proposed[48]. Much work has been done w.r.t. the problem of querying both databases and mining datasets. As to the relational data model, a data mining language based on the principles

of closure (the results of a query can be further manipulated) and cross-over between data and rules[27] has been introduced. Extensions to the object-oriented data model[12] have also been reported in the literature. However, approximate querying where graph-based data and mining datasets co-exist is a research issue that has never been addressed. Also, given the different syntactic and semantic representations and the massive scale of the datasets, checking whether multiple data instances are actually the same entity is a very challenging problem. The proposed record linkage techniques are either id-based, when ids are available, or apply syntactic approximations on the data. Most of these approaches are integration-oriented and are not suitable to the data coexistence scenario envisioned by the project. Very few works present on-the-fly techniques[28] that will be considered as the starting point to propose the hybrid approach needed to dynamically support the different connection alternatives for data sources.

2.3 Data Mining

As regards the mining activities on biological data a key task is the discovery of groups of genes whose gene expressions are simultaneously altered by one or more pathologies. This analysis would provide useful information for drug repositioning[2], 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 (e.g. it is a recent finding the fact that chemical compounds typically used in treatments for tumors have positive effects on patients suffering from the Alzheimer's disease[7]). A possible improvement is the exploitation of co-clustering discovery approaches, which are the most suited tools to identify clusters of objects of different nature (in this case genes-patologies). Indeed, the application of co-clustering techniques in the biological context is not new[22]. However, most of the existing approaches focus on the algorithms that, if applied to large datasets, present the problem of a high number of extracted co-clusters. Moreover, most of the existing approaches suffer from the impossibility of extracting overlapping co-clusters (in our case a gene can be involved in several regulation networks)[8]. Finally, existing co-clustering algorithms do not consider possible relationships that involve first class objects considered in the analysis (i.e. genes and pathologies) or relationships between these first class objects and other objects (possibly of a different type, such as functional pathways and mRNA). In order to overcome these limitations recent studies in Collective Classification[42] has been presented. They allow to take into account possible autocorrelation in the data. Another interesting research line related to our project is the study of evolution of pathologies through short time series analysis techniques. Unlike traditional time series analysis, whose main problem is related to the length of time series, short time series are characterized by very few temporal points. This is a characteristic of our extended CMAP. According to [14], relevant tasks for short time series are: classification, clustering and anomaly detection. The use of interactive visual interfaces for cycles identification to perform classification of sequences, to perform comparisons among sequences as well as to perform pattern matching and temporal pattern

search have been addressed in literature [43], but very few of them can be used for short time series, a proposal on this field is in [46]. Recent studies focus on the definition of tools for the identification of the pathology stage on the basis of expression gene values. To this end, approaches of collective classification will be particularly studied[45] due to their peculiarity to handle the autocorrelation aspects typically present in data organized in network/graph form. Indeed, it has been proved in the literature that the co-occurrence of autocorrelation with high-density neighborhood of data could bias the selection of features in the task of relational classification. Evolutionary algorithms (EA)[15] are heuristics that mimic the processes of natural evolution in order to solve global search problems. They differ from more traditional optimization techniques in that they involve a search starting with a "population" of solutions (i.e. a string of bits), not with a single point. Recombination, crossover and mutation operators are used to generate new solutions that are biased towards different regions of the search space. Genetic programming (GP)[32] is an extension of genetic algorithms (GAs) that iteratively evolves a population of (typically) trees of variable size, by applying variation operators. The use of hybrid techniques, i.e. EAs and data mining ensembles, together with efficient implementations and with new models of distributed computations enables these kinds of algorithm to cope with hard classification problems. Bagging and boosting, introduced in [41] and [16] are well known ensemble techniques that repeatedly run a learning algorithm on different distributions over the training data.

2.4 Semantic Tagging and Ontologies

Semantic relationships among biological entities are actually a growing research field. These relationships are usually derived from biological knowledge encoded into biological ontologies. There exist many active projects that aim to organize such knowledge into structured vocabularies of concepts and taxonomies of concepts themselves. For instance Gene Ontology (geneontology.org/)[23] stores concepts about the localization, processes and function of genes and gene products, while Protein Ontology (proteinontology.org.au/)[37] contains concepts related to proteins and Disease Ontology (diseaseontology.sourceforge.net) relationships among diseases and genes. In computational biology, ontologies are often used to annotate biological concepts, i.e. to associate to a biological concept such as gene BRCA1 its description using only concept from a biological field[11]. The use of ontologies to support modeling and querying of biological data has been explored in [19]. In recent years, many approaches for the evaluation of the similarity of two or more concepts belonging to the same ontology have been developed. Thus, starting from two entities that are annotated with terms belonging to the same ontology, it is possible to define a semantic similarity by the similarity of the concepts used for annotating them. In this way, it is possible to define all pairwise similarities of entities belonging to the same domain and annotated with the same ontology (even if this approach may be extended in the case of different ontologies) [19]. The whole set of entities and their similarities may be efficiently represented into a single comprehensive model by using graph

theory [17]. Finally, there exist also different computational approaches developed both for the knowledge extraction of biological networks (e.g. clusterings) [1],[5] and for concepts belonging to the same ontology (e.g. semantic similarity measures)[19].

2.5 Visual Query Languages

A Visual Query System (VQS) is a system that uses a visual representation for both the domain of interest and the related requests on it. Two forms of query creation (SQL and of Query By Example (QBE)) have been recently compared through experiments [26]. The authors found that the time requested for query formulation applying a QBE-based approach was shorter than the time requested using a SQL approach. Interestingly, there were no remarkable differences regarding the accurateness of the queries between the two approaches. An example of diagram-based Visual Query Language is QBD*[4]. A framework that allows both intensional and extensional queries, developed following the paradigm Query By Browsing (QBB), is described in [40]. QBI[36] is a pure iconic Visual Query Language, which provides tools for an intensional browsing of databases. Finally, as stated by [31], Visual Analytics is "the science of analytical reasoning facilitated by interactive visual interfaces", which means to help the decision making process by turning the information overload into an opportunity. Human perception plays an important role in such a tool because a visualization which does not consider cognitive principles may lead to misunderstandings and wrong interpretations of the data.

3 Our Approach

As previously discussed, a fundamental challenge that arises throughout biomedicine is the need to establish relations among diseases, physiological processes, and the action of small-molecule therapeutics. 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, the 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.

As stated before, such relationships can give better insights about tumors as well as can be used for drug repositioning and for the identification of novel compounds able to overcome resistance or revert it in to drugs in current use. Our aim is to offer the following functionalities within a user friendly Web delivery system:

- Identification of the data repositories and databases which relate to the CMap;
- Normalization and interoperability of the identified databases and the CMap;
- Extraction of useful knowledge from the data by means of data Mining techniques;

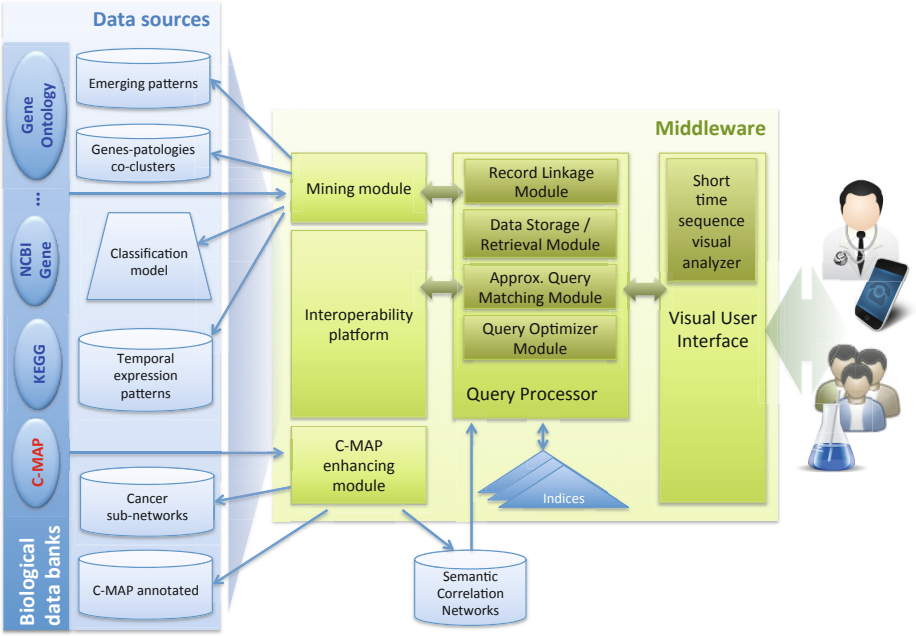

Fig. 1. The system architecture

– Semantic tagging of CMap;
– Querying of the extended CMap, the identified data repositories and the extracted knowledge as a unique dataspace;
– Querying CMap and extracted knowledge by means of a Visual Query Language.

To this end, we aim to give innovative answers to the problems that are mandatory for these objectives, by applying several methodologies and techniques.

3.1 Bio-technological 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 and databases. In this perspective, the central role of the Connectivity Map in this project is motivated by the increasing interest of the scientific community in the prediction of novel association among diseases, physiological processes, and the action of small therapeutic molecules. As a matter of fact, as mentioned above several works in the literature aim at extending the information stored in the CMap. However, they represent ad-hoc 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-tecnological database such as NCBI (Gene, Geo, Pubmed), ArrayExpress, Gene Ontology, KEGG, and Drug Bank as reported in Fig. 1. The interoperability issues arising from accessing such a wide set of data sources are discussed in Section 3.3.

3.2 Data Mining and Semantic Information Extraction

By taking into account the requirements coming from bio-medical researchers we exploit several mining algorithms in order to extract knowledge on Cancer microenvironments from the selected datasources. The goal is to better understand tumors and to identify relationships among tumors, the effect of drugs and the patients' biodiversity. As stated before, such relationships can be used for drug repositioning and for the identification of novel compounds able to overcome resistance or revert it in drugs in current use. Due to both the novelty and the nature of the these problems, ad hoc data mining algorithms are used. Indeed, a significant benefit is the identification of semantic relatedness among domain-specific entities. In particular, the results of the data mining algorithms are 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 that can be considered in the mining processes. Accordingly, we introduce the notion of semantic relatedness that relates CMap entities. By exploiting co-clustering approaches, we discover groups of genes whose gene expressions are simultaneously altered by one or more pathologies [9]. To this purpose, hierarchical and non-hierarchical co-clustering techniques are exploited. Moreover, in order to characterize and describe pathologies (or classes of pathologies) on the basis of the variability of genomic signatures observed in gene products, network based emerging patterns discovery algorithms [10] are used. Furthermore, we analyze evolution of pathologies through short time series analysis techniques [13]. To this end, both visual data mining and temporal patterns extraction algorithms are defined. Finally, in order to identify the pathology stage on the basis of expression gene values, collective classification algorithms and ensemble-based algorithms are exploited. While in the case of collective classification [42] it is possible to handle the autocorrelation (according to which "closer" objects are more related than "furthermost" ones), typically present in data organized in network/graph form, in ensemble-based classification [38] 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" of our system (see Fig. 1). As regards data correlation analysis, starting from data stored in the CMap, the selected databases and the ontologies, a Semantic Correlation Network will be built. This semantic graph will be used to extract sub-networks related to Cancer through the application of network analysis algorithms such as network alignment algorithms [33], clustering [29], and pattern extraction algorithms [21]. The outcomes of this activity isimplemented in the "CMap enhancing module" of the Web delivery system.

3.3 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. Once again, it differs from the biological data 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 would allow analysis tools to deal with 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. Finally, a flexible query model is necessary that would allow stakeholders to easily query the knowledge in the data sources in a uniform way and to get useful results for analysis purposes. Therefore, the main challenges related to this goal are:

1. Extension of the Connectivity Map with semantic information encoded into ontologies;
2. Normalization and interoperability of the set of data sources;
3. Definition of techniques effectively and efficiently supporting the querying of the data sources.

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 (*C-Map 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 *C-Map 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 various tasks: C-Map 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. The interoperability platform will then support the co-existence of all these sources, each of which can participate both as internal source or as external one, through two languages, a *Data Delivery Definition Language* (DDDL) for source specification and a *Mapping Language* (ML) for inter-source relationship specification. Specifically, the interoperability platform will include low-level repository functionalities, including: a) (for all data sources) storage and retrieval of data source associated descriptions, including DDDL and ML; b) (only for complete access sources) storage /retrieval / update of the data itself; c) (only for external sources) storage and retrieval of wrapping patterns.

The third challenge is faced through 1) the creation of a flexible query language (equipped with 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 both data and mining 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 down-regulated 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 dataspace 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 [30].

3.4 Definition of the Web Delivery System for the Easy-Access to the Bio-technological Data and the Derived Knowledge

All the project results will be made available to the research community through the Web delivery system as depicted in Fig. 1. The 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 explain the goal we intend to obtain, consider the following example: the query and visualization interface of CMAP – http://www.broadinstitute.org/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, etc. 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, in our system a graphical interface is 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, etc.). 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, etc. and therefore do not impose the use of a desktop.

4 Case Study: Supporting Bio-Medical Researchers in the Study of the Multiple Myeloma through the Web Framework

The web framework will be used in a pilot study on the Multiple Myeloma (MM), an incurable malignant plasma cell disease. MM is particularly interesting since it allows researchers to concentrate on microenvironments which promote tumor survival. Main challenges related to this goal can be tackled in our system, such as:

1. Better understanding the Multiple Myeloma (MM) by expoiting semantic record linkage techniques. This is performed by identifying other blood tumors that are semantically related to the MM not only on the basis of information stored in the CMap, but also on the basis of information stored in other datasources such as PubMed (literature) and Gene Ontology;
2. Provide the bio-medical researchers with a tool for querying the dataspace identified for the MM pilot study;
3. Support the work of bio-medical researchers with Data Mining techniques.

To this purpose, the aim is 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. In other words, on the basis of results obtained through the application of Data Mining techniques, it is possible to drive the process of hypothesis generation in:

- understanding the correlation of the MM with other tumors in terms of gene expressions modifications;
- defining a characterization of the MM in terms of genes;
- analyzing the short time evolution of the pathology;
- generating classification models to automatically identify MM on the basis of gene expressions modifications and additional information stored in other datasources. This analysis might help the drug repositioning task and the identification of novel compounds able to overcome resistance or revert it in drugs in current use;
- provide the bio-medical researchers with semantic network analysis techniques. Indeed, since drug resistance in the MM is possibly related to drug resistance in other (blood) tumors, we use semantic network analysis techniques in order to identify the semantic distance between MM and each other (blood) tumor (again, to support drug repositioning).

5 Conclusion

In this paper we presented a system for biological data normalization and interoperability devoted to information extraction, data querying and knowledge dissemination for supporting biomedical specialists in the analysis of cancer microenvironments. The system is modular and is tailored on the biological data features in order to make it easy to use and provide useful information to the domain experts. The system will be used for assisting bio-medical researcher in the analysis of information related to Multiple Myeloma.

References

1. Aittokallio, T., Schwikowski, B.: Graph-based methods for analysing networks in cell biology. Brief Bioinform. 7(3), 243–255 (2006)
2. Ashburn, T.T., Thor, K.B.: Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery 3, 673–683 (2004)
3. Boutros, P.C.: Fun with microarrays part iii: Integration and the end of microarrays as we know them. Hypothesis 6(1) (2008)
4. Catarci, T., Santucci, G.: Query by diagram: A graphical environment for querying databases. In: Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, May 24-27, p. 515. ACM Press (1994)
5. Ciriello, G., Guerra, C.: A review on models and algorithms for motif discovery in protein interaction networks. Briefings in Functional Genomics and Proteomics 7(2), 147–156 (2008)
6. Costa, G., Manco, G., Ortale, R.: An incremental clustering scheme for data deduplication. Data Min. Knowl. Discov. 20(1), 152–187 (2010)
7. Cramer, P.E., Cirrito, J.R., Wesson, D.W., Lee, C.Y.D., Karlo, J.C., Zinn, A.E., Casali, B.T., Restivo, J.L., Goebel, W.D., James, M.J., Brunden, K.R., Wilson, D.A., Landreth, G.E.: Apoe-directed therapeutics rapidly clear ß-amyloid and reverse deficits in ad mouse models. Science 335(6075), 1503–1506 (2012)
8. Deodhar, M., Gupta, G., Ghosh, J., Cho, H., Dhillon, I.S.: A scalable framework for discovering coherent co-clusters in noisy data. In: Pohoreckyj Danyluk, A., Bottou, L., Littman, M.L. (eds.) ICML. ACM International Conference Proceeding Series, vol. 382, p. 31. ACM (2009)
9. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. In: ISMB, pp. 145–154 (2002)
10. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: KDD, pp. 43–52 (1999)
11. Plessis, L.D., Kunca, N., Dessimoz, C.: The what, where, how and why of gene ontology a primer for bioinformaticians. Briefings in Bioinformatics (2011)
12. Elfeky, M.G., Saad, A.A., Fouad, S.A.: ODMQL: Object Data Mining Query Language. In: Dittrich, K.R., Oliva, M., Rodriguez, M.E. (eds.) ECOOP-WS 2000. LNCS, vol. 1944, pp. 128–140. Springer, Heidelberg (2001)
13. Ernst, J., Bar-Joseph, Z.: Stem: a tool for the analysis of short time series gene expression data. BMC Bioinformatics (2006)
14. Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics 21(suppl. 1), i159–i168

15. Fogel, D.B.: Evolutionary computation - toward a new philosophy of machine intelligence, 3rd edn. Wiley-VCH (2006)
16. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: ICML, pp. 148–156 (1996)
17. Golumbic, M.C.: Algorithmic Graph Theory and Perfect Graphs. Academic Press, New York (1980)
18. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: Predict: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. 7 (2011)
19. Guzzi, P.H., Mina, M., Guerra, C., Cannataro, M.: Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in Bioinformatics (2011)
20. Halevy, A.Y., Franklin, M.J., Maier, D.: Principles of dataspace systems. In: PODS, pp. 1–9 (2006)
21. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2000)
22. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics 18(suppl. 1), S145–S154 (2002)
23. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., Tonellato, P., Jaiswal, P., Seigfried, T., White, R.: The gene ontology (go) database and informatics resource. Nucleic Acids Res. 32, 258–261 (2004)
24. He, H., Singh, A.K.: Closure-tree: An index structure for graph queries. In: ICDE, pp. 38–49 (2006)
25. Hu, G., Agarwal, P.: Human disease-drug network based on genomic expression profiles. PLoS One 4(8), e6536 (2009)
26. Hvoreckya, J., Drlikb, M., Munk, M.: The effect of visual query languages on the improvement of information retrieval skills. Procedia - Social and Behavioral Sciences 2(2), 717–723 (2010)
27. Imielinski, T., Virmani, A.: Msql: A query language for database mining. Data Min. Knowl. Discov. 3(4), 373–408 (1999)
28. Ioannou, E., Nejdl, W., Niederée, C., Velegrakis, Y.: On-the-fly entity-aware query processing in the presence of linkage. PVLDB 3(1), 429–438 (2010)
29. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31 (September 1999)
30. Karmel, R., Gibson, D.: Event-based record linkage in health and aged care services data: a methodological innovation. BMC Health Services Research (2007)
31. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual Analytics: Scope and Challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
32. Koza, J.R.: Genetic Programming On the Programming of Computers by Means of Natural Selection. MIT Press (1992)

33. Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., Przulj, N.: Topological network alignment uncovers biological function and phylogeny. J. of the Royal Society (2010)
34. Lamb, J.: The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer 7(1), 54–60 (2007)
35. Li, J., Zhu, X., Chen, J.Y.: Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. PLoS Comput. Biol. 5(7), e1000450 (2009)
36. Massari, A., Pavani, S., Saladini, L., Chrysanthis, P.K.: Qbi: Query by icons. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, p. 477. ACM Press (1995)
37. Natale, D., Arighi, C., Barker, W., Blake, J., Chang, T.-C., Hu, Z., Liu, H., Smith, B., Wu, C.: Framework for a protein ontology. BMC Bioinformatics 8 (2007)
38. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198 (1999)
39. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. Journal of Computational Biology a Journal of Computational Molecular Cell Biology 18(2), 133–145 (2011)
40. Polyviou, S., Evripidou, P., Samaras, G.: Query by browsing: A visual query language based on the relational model and the desktop user interface paradigm. In: The 3rd Hellenic Symposium on Data Management (2004)
41. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning 37(3), 297–336 (1999)
42. Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective classification in network data. AI Magazine 29(3), 93–106 (2008)
43. Shah, M., Corbeil, J.: A general framework for analyzing data from two short time-series microarray experiments. IEEE/ACM Trans. Comput. Biol. Bioinformatics 8(1), 14–26 (2011)
44. Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A., Sage, J., Butte, A.J.: Discovery and preclinical validation of drug indications using compendia of public gene expression data. Science Translational Medicine 3(96), 96–77 (2011)
45. Stojanova, D., Ceci, M., Appice, A., Džeroski, S.: Network Regression with Predictive Clustering Trees. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 333–348. Springer, Heidelberg (2011)
46. Wang, X., Wu, M., Li, Z., Chan, C.: Short time-series microarray analysis: Methods and challenges. BMC Systems Biology 2 (2008)
47. Zhao, P., Han, J.: On graph query optimization in large networks. Proc. VLDB Endow. 3(1-2), 340–351 (2010)
48. Zhu, L., Ng, W.K., Cheng, J.: Structure and attribute index for approximate graph matching in large graphs. Inf. Syst. 36(6), 958–972 (2011)