eMiRo: an ontology-based system for clinical data integration and analysis
Discussion Paper

Pietro Cinaglia, Pierangelo Veltri, and Mario Cannataro

Department of Medical and Surgical Science, University of Catanzaro
{surname}@unicz.it

Abstract. Biological experiments and bioinformatics analysis are producing large datasets of omics data (e.g. genomics, proteomics, interactomics, etc.), stored in diversified sources and databases. Biological data and bioinformatics results can be related to various information, such as clinical data (e.g. cancer stage) or environment data (e.g. place where a patient lives), but this requires novel data integration mechanisms and analysis algorithms. Novel data structures are needed for data integration, while efficient algorithms are necessary for managing integrated data and to analyze results in order to extract knowledge and underline correlation with environmental factors. In this paper an ontology-based system for clinical data integration and analysis is presented. The system, called eMiRo (Electronic MedIcal RecOrd), includes geo-reference features for epidemiological analysis based on geographical and clinical information. Using eMiRo, a physician is able to handle and integrate biological data; moreover, the system supports the retrieval of additional information such as ontology terms and information about genes and diseases. When conducting a study, for each gene relevant to the study, biological function and relations with other genes as well as involvement in biological processes is reported.

1 Introduction

The growth of biological resources and information led to the storage of large data, such as: physiological processes, gene patterns, disease-gene associations, clinical trials; in this context it is necessary the development of bioinformatics tools for data management in order to index heterogeneous resources in common structures and to provide a comprehensive view of the state of knowledge. In recent years, many bioinformatics studies have been focused on gene role in biological processes for preventing and treating diseases. Tools for heterogeneous data integration and association are developed to improve knowledge in experiments related to gene samples from Microarray and Next Generation Sequencing (NGS) technologies. In this scenario it is necessary to develop algorithms able to merge data from multiple sources, with heterogeneous models and formats, to obtain a global information which satisfies the requirements formulated by users (e.g. researchers). The literature presents various solutions to improve and to automatize gene data analysis, for example [1] presents a R-based tool for miRNA data analysis and correlation with clinical ontologies related to a study about modeling and management of biological data. Ontologies are used in computer science for knowledge management; its structure is based on information related to each other by Terms (e.g. concepts, processes, and methods). In biology, Gene Ontology (GO) [2] is used by many algorithms and web systems to retrieve biological information such as: molecular functions, biological processes, and subcellular localizations in drug discovery. Analogously, Disease Ontology (DO) [3] correlates human diseases through the cross-mapping of medical vocabularies (e.g. MeSH, ICD, NCI Thesaurus, SNOMED and OMIM); furthermore, it allows the examination and comparison of genetic variations, phenotypes, proteins, and drugs. In genomics, the data heterogeneity is highly grown in conjunction with the new experimental platforms (e.g. microarray experiments) and consequently data size, type and structure is greatly diversified. Data-integration techniques are crucial to obtain an interdisciplinary view which allows, for example, a biologist to retrieve information of interest from large and heterogeneous dataset [4]. Some applications are able to manage several resources (e.g. physiology, pharmacology, clinical data) using ontologies [5] [6] or correlating concepts with information through integration process based on semantic-analysis [7] [8]. In OAHG [9] a correlation among human protein-coding genes (PCGs), miRNAs, and lncRNAs is established in order to generate a comprehensive functional annotation resources; for this purpose, it uses multi-level ontologies, such as: Gene Ontology, Disease Ontology, and Human Phenotype Ontology (HPO) [10]. The semantics analysis of biomedical data is often necessary to integrate the information from sources that use different schemes as in ontology databases, for example: SNOMED CT (Clinical Terms), ICD-9 and ICD-10 (International Classification of Disease), and Human Phenotype Ontology (HPO) [10]. In recent years the interest of Geographic Information System (GIS) technologies are growing in biomedical studies in order to solve health issues, such as: to evaluate the prevalence of diseases, to plan health interventions for epidemiological studies, to perform a geographical analysis for monitoring of population for prevention purposes. In [11] the biological information are integrated with GIS geographic data for the handling and management of microbial genome data.

In this paper we present a system designed to manage clinical and biological data, and to correlate such data by means of available ontologies. The system, called eMiRo, has been developed to manage clinical and genomic data and integrates them using ontologies and genomic information from external sources. Moreover, eMiRo includes a geographical feature for epidemiological analysis.

2 The eMiRo System

We design a system starting by the following requirements: clinical and biological data management, bioinformatics results from ontologies and semantics data, support for geographical data. The actors identified for the system are listed below:

- Physician: all features are available, it is able to manage data patients (using a model configured by the administrator), performs query to filter the informations, consults the geographical analysis results, and requires data integration features for a specific disease; custom access modes are definable for this profile to create sub-profiles.
- Administrator: refers to the user which is able to handle system and its features, as well as user profiles, component configuration, and data.
- Client: identified with a person or service (e.g. an algorithm) which accessed the data-integration function using the genieR as web-service.

2.1 Architecture

System may be roughly divided into three main components, briefly:

- eMiRo (or 'eMiRo-component', to distinguish this from the entire system called with the same name): it handles the biological data and the interactions with users, as well as data exchanges with other modules (genieR and geoP);
- genieR: it concerns data analysis and data-integration features, also supports retrieving and management of unstructured information from external sources (e.g. ontologies);
- geoP: it is related to geographical features: analysis, and geographic coordinates managing (e.g. it converts raw text into coordinates using the Google Maps API).

Figure 1 shows the interactions among the components; external sources and API calling are also shown.

eMiRo-component is the basic module of the system; it is in charge of the following functions:

- graphical front-end management;
- querying;
- biological data handling;
- database connection and components interaction;
- login, security and user roles management;
- analytics.

User is able to access the system through a Web-Panel which, essentially, consists in a set of Graphical User Interfaces (GUI) that are generated dynamically in according to a predefined model for the information required by user. Geographical data are shown to user using a Map-View (from Google Maps API), and a summary table-view to show only the information of interest for the user (e.g. a physician).

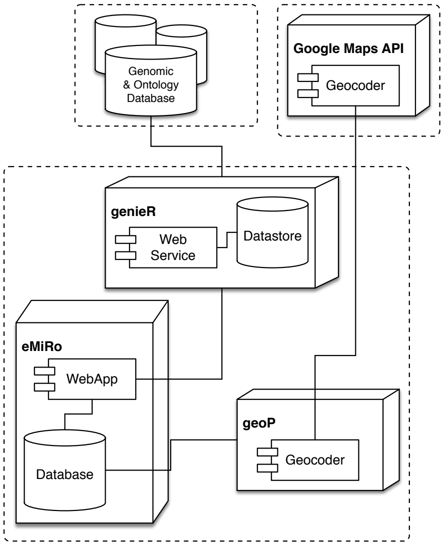


Fig. 1. System deployment

genieR is designed as a web-service to allow external clients (e.g. algorithms or other information systems) to invoke its public functions. This components concerns the data-integration features; specifically it unifies in a custom structure the information given by several external sources. A mapping between Disease Ontology and Gene Ontology is used to integrate annotations in order to obtain the gene-disease associations.

Therefore, it is used for the management and the integration of data from remote genomic and ontology sources. eMiRo supports integration from: Disease Ontology [3], Gene Ontology [2] and DisGenet [12] ("curated gene-disease association" version); others could be added in future updates. The information are structured as objects that structure data using lists and hashmap; this approach allows the management of unstructured data; it is important to emphasize that this component is implemented for deployment on Google App Engine Environment.

genieR main sub-components are:

- GenierServlet: it consists in two modules, dealing respectively to handle requests from clients (even Web-Panel accesses genieR as an external resource) and to structure the output in according to JSON format in order to make faster the data stream and easy the parsing operations.
- Datastore: it periodically retrieves and parse the information from supported external sources and organizes these in a structure compatible with the eMiRo data model.
- DataIntegrator: it is the core module of genieR; it takes care to integrate information in Datastore and to generate an output for the client.

In summary, a physician is able to choose a disease, from a list, to obtain additional information, such as: ontological integration, genes involved for the disease, and meta-data for each gene founded (e.g. description, biological function, biological processes, and relations with other genes). Main commands supported by genieR are 'doget' and 'dolist' ('do': Disease-Object):

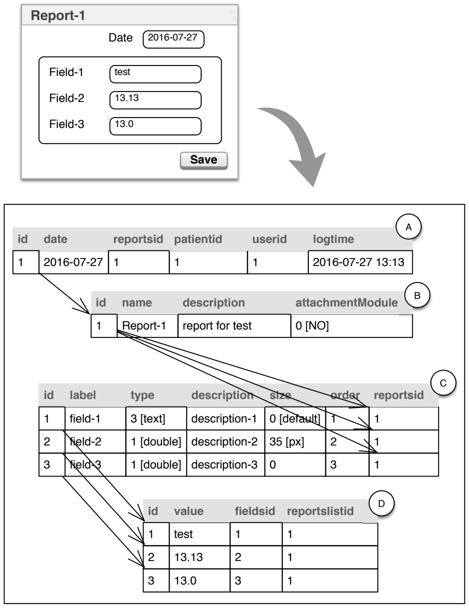
- former returns information obtained from data-integration action for a specific Disease Ontology ID (DOID);
- latter gives the list of diseases that system supports for the integration operation.

geoP is able to convert addresses into geographical coordinates using Google Geocoders API; using this approach a geographical coordinates management through GIS systems is allowed. Furthermore, the system supports analytics features to generate statistical information from geographical and biological data mapping (for statistical analysis are supported a set of parameter, such as: range of values, and threshold level).

2.2 Implementation details

eMiRo system is designed for dynamic creation of reports referred to biological information management; this solution allows users to create new examination and handling data models for existing ones. Data-Type supported by the Database Engine are available during models creation.

A 'model' is composed by a set of information related to a graphic layout and its low-level parameters (e.g. field-name, field-data-type, bounds for graphic elements, and the order of the fields in the resulting GUI), furthermore the persistence of models is granted by a Relational Database Management System (RDBMS). In according to this dynamic solution the Figure 2 shows the data structure for a generic examination (named 'Report-1') which consists of three fields ('date' is a default field whose representation is not required); Figure 2-A and 2-D are referred to a specific examination for a given patient; while, 2-B and 2-C contain information related to the model for GUI generation: fields, data-type, and generic meta-data (e.g. field-order, and field-width in pixel). To implement the system are chosen multiple programming languages related to the purpose for each component; this approach optimizes the deployments and the user experience. genieR and geoP are implemented using Servlet technology to deploy Web-Applications [13] and to expose their features as a service (in according with server policy configuration), for genieR the Google Cloud App Engine SDK is also integrated. eMiRo-component has been developed using the PHP and HTML5 languages, the latter extended using Bootstrap and jQuery frameworks in order to support graphical interface with dynamic elements and a responsive design based on CSS and JavaScript. Oracle MySQL has been chosen as RDBMS for biological and geographic data persistence; in addition, it contains information about the structures and the models used by system during GUI generation.



Date 2016-07-27

Field-1 test

Field-2 13.13

Field-3 13.0

id	date	reportsid	patientid	userid	logtime
1	2016-07-27	1	1	1	2016-07-27 13:13

id	name	description	attachmentModule
1	Report-1	report for test	0 [NO]

id	label	type	description	size	order	reportsid
1	field-1	3 [text]	description-1	0 [default]	1	1
2	field-2	1 [double]	description-2	35 [px]	2	1
3	field-3	1 [double]	description-3	0	3	1

id	value	fieldsid	reportslistid
1	test	1	1
2	13.13	2	1
3	13.0	3	1

Fig. 2. Data model (low level representation)

2.3 Data-Integration: approach

During the test, the eMiRo data-integration component (named genieR) has been used to retrieve the Open Biological Ontologies datasets from the server of Gene-Ontology, Disease-Ontology, and Disgenet. OBO Foundry initiative supports the evolution of ontologies for biomedical data integration promoting ontologies designed to be interoperable and logically well-formed in order to incorporate accurate representations of biological information [14]. A dataset in OBO format contains several tags (e.g. id, name, description, synonyms, cross-references) organized in Terms linked to each other; thus, a dataset may be represented using, for example, a graph data structure. genieR pipeline is based on two phases: preprocessing and analysis. During Preprocessing the data are integrated and organized in a single structure: (i) the ontological datasets are downloaded on cloud-memory (the data can be refreshed when the sources update their datasets); (ii) subsequently, each dataset is parsed to extract information of interest, and it is collapsed within a single structure with others. Using this approach, genieR is able to create a new large ontology (managed by the Data-store component) that contains the heterogeneous information extracted from each ontological dataset, and the novel relations found among these. Preprocessing reduces the data in the cloud-memory to improve the overall performance during the analysis. For the Analysis phase, genieR implements a component named 'DataIntegrator': when a client requests information about a disease the DataIntegrator performs an analysis of its integrated data (that represents its knowledge), and returns a results in JSON Format. Summing, genieR is able to provide for a specific disease the relevant genes and for each the biological function, and its involvement in biological processes, and the relations with other genes.

3 Security

In eMiRo, the user requests and the access policy are handled by web-application and DBMS Engine [15]; first grants the user access based on his role, second checks the read and write operations by the components, as well as the direct requests to database. The communication between the components (internal and external) is performed through HTTPS protocol which allows server authentication, privacy protection and maintaining data integrity, as well as checks data exchanged between the parts and allows SSL/TLS encryption. Furthermore, account credentials are encrypted using the Secure Hash Algorithm (SHA) before being stored within the database.

4 Conclusion

This paper presents the architecture and the security insights of a cloud-based information system, named eMiRo, for biological data mapping; georeferencing of patients is also supported. A light-beta version is used to handling biological data by infectious diseases group of Magna Graecia University.

References

1. F. Cristiano and et al., "An r-based tool for mirna data analysis and correlation with clinical ontologies," *Proceeding BCB '14 Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2014.
2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.
3. DO Official WebSite, *Disease Ontology*, 2016.
4. D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner, "Data integration in the era of omics: current and future challenges," *BMC Syst Biol*, vol. 8 Suppl 2, p. I1, 2014.
5. S. M. Wimalaratne, P. Grenon, R. Hoehndorf, G. V. Gkoutos, and B. de Bono, "An infrastructure for ontology-based information systems in biomedicine: RICORDO case study," *Bioinformatics*, vol. 28, pp. 448–450, Feb 2012.
6. B. De Bono, R. Hoehndorf, S. Wimalaratne, G. Gkoutos, and P. Grenon, "The RICORDO approach to semantic interoperability for biomedical data and models: strategy, standards and solutions," *BMC Res Notes*, vol. 4, p. 313, 2011.

7. K. M. Livingston, M. Bada, W. A. Baumgartner, and L. E. Hunter, “KaBOB: ontology-based semantic integration of biomedical databases,” *BMC Bioinformatics*, vol. 16, p. 126, 2015.
8. C. Pang, A. Sollie, A. Sijtsma, D. Hendriksen, B. Charbon, M. de Haan, T. de Boer, F. Kelpin, J. Jetten, J. K. van der Velde, N. Smidt, R. Sijmons, H. Hillege, and M. A. Swertz, “SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data,” *Database (Oxford)*, vol. 2015, 2015.
9. L. Cheng, J. Sun, W. Xu, L. Dong, Y. Hu, and M. Zhou, “Oahg: an integrated resource for annotating human genes with multi-level ontologies,” *Sci Rep*, vol. 6, p. 34820, Oct 2016.
10. S. Kohler and at al., “The human phenotype ontology project: linking molecular biology and disease through phenotype data,” *Nucl. Acids Res.*, 2014.
11. S. Jelokhani-Niaraki, M. Tahmoorespur, Z. Minuchehr, and M. R. Nassiri, “An Ontology-Based GIS for Genomic Data Management of Rumen Microbes,” *Genomics Inform*, vol. 13, pp. 7–14, Mar 2015.
12. J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, “DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes,” *Database (Oxford)*, vol. 2015, 2015.
13. Oracle, *Java Servlet Technology Overview*, 2016.
14. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration,” *Nat. Biotechnol.*, vol. 25, pp. 1251–1255, Nov 2007.
15. G. Miklau and D. Suciu, “A formal analysis of information disclosure in data exchange,” *J. Comput. Syst. Sci*, 2007.