Representation of Semantic Networks of Biomedical Terms

André Vechina[1], Joel P. Arrais[2], and José Luís Oliveira[1]

[1] Department of Electronics Telecommunications and Informatics (DETI),
Institute of Electronics and Telematics Engineering of Aveiro (IEETA),
University of Aveiro, 3830-193 Aveiro, Portugal
{andrevechina,jlo}@ua.pt

[2] Department of Informatics Engineering (DEI),
Centre for Informatics and Systems of the University of Coimbra (CISUC),
University of Coimbra, Portugal
jpa@dei.uc.pt

Abstract. The recent technological developments in computing science and information systems are promoting innovative methods and tools of capital relevance in present biomedical research. By taking advantage of those new methods, it was possible to expand the set of well-known associations between different biomedical terms. However, biomedical knowledge is spread over multiple independent databases. The integration of these data into a single repository allows the representation of all biomedical associations as a network and the use of network theory to extract new findings.

This paper presents the BioMedNet system, a web-based application that can be used to visualize and study the relations between biomedical concepts. The system provides a variety of tools that allows professionals to extract relevant biomolecular evidences among diseases, genes and other biomedical entities.

Keywords: System Biology, Biological networks, PPI, Visualization

1 Introduction

The evolution in computing sciences and informatics helped to create a complete revolution in several research fields, such as genomics, molecular biology and biomedicine [1, 2]. Moreover, recent discoveries on biological researches , led to an overall growth of the semantic networks that represent these biomedical entities [3].

Nowadays, many public databases already store and provide access to large-scale networks, typically generated through several methods [4]: manual curation of small subsets of high-quality experimental results, computational technology of high-throughput screening of biomedical interactions, automatic data mining of scientific literature, and computational prediction from data integration. Although the manual curation produces the most reliable interactions, it would not be possible to retrieve such an extended set of resultswithout the use of computational methods. Despite the lower confidence rate, there has been large developments in some new techniques which attempt to maximize the efficiency of computational processes and minimize the false positive rates [5, 6].

While computational methods prove their importance to generate data, the human action defines the quality of the entire process. Using a graphical representation of a graph, it is possible to create a very simple, but still powerful, abstraction of the network. The human ability to interpret and analyse a network, as well as all the professional experience with biological information, should not be despised. It is also important, when analysing the computational prediction results, that biomedical professionals are capable of understanding the reasons behind a specific prediction.

In this paper, we present BioMedNet, a web-based application that, thought the representation of a semantic network of biomedical terms as a graph, allowing domain experts to extract relevant biomolecular evidences among multiple biomedical terms.This solution integrates biological data from several sources and provides a variety of different computational tools and methods to visualize and help to understand the relationships among diseases, genes and other biomedical entities. BioMedNet is freely available at `http://bioinformatics.ua.pt/BioMedNet/`.

2 Related Work

Graphs have the ability to model and represent complex biological systems [4, 7]. However, information about relevant biomedical entities and their associations is scattered along many data repositories.

STRING[3], for instance, is a database of known and predicted protein interactions that provides a large set of associations, including proteins related with the human species (*Homo sapiens*). Ensembl Protein[4] is the notation used to identify all those proteins. The OMIM[5] database is an on-line catalogue of human disorders and diseases, human genes and associations between them. This repository uses its own notation for both diseases and genes. More information about gene ontologies is also provided by Gene Ontology[6]. Through data provided by the Gene Ontology Annotation database[7] it is possible to obtain a series of relations between gene ontologies and human genes. KEGG[8] is another resource which provides an extended set of information about biological systems, including a set of known relations between human genes and pathways.

These are some of the data sources that store and provide information about biomedical terms and their relations. All these sources were used to collect the

[3] STRING, `http://string-db.org/`

[4] Ensembl Protein, `http://www.ensembl.org/`

[5] Online Mendelian Inheritance in Man, `http://www.omim.org/`

[6] Gene Ontology, `http://www.geneontology.org/`

[7] Gene Ontology Annotation, `http://www.ebi.ac.uk/GOA/`

[8] Kyoto Encyclopedia of Genes and Genomes, `http://www.kegg.jp/`

entire set of biomedical information used in the BioMedNet system. Several techniques can be applied to share and access biomedical information. Depending on the source provider in this work two approaches are used: download and parse *flat-files* with large sets of data or dynamically retrieve information from Web Services.

3 Methods and Materials

3.1 Software as a Service

The services provided through the Internet have grown in popularity and acceptance in IT related areas. These services, also known as Cloud Computing Services, grant the access to a rich set of resources that can be split in three major types [8, 9]: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

The SaaS layer represents the top of the stack, since its existence depends on the underlying layers. Within this layer, it is possible to find a wide set of different web-based applications. These applications can frequently be accessed through their web portal interfaces, using Web Services [10] and Web 2.0 technologies [11].

There are several advantages in deploying an application as SaaS [12]. From a provider perspective it is easier to develop, update and maintain an application that is deployed in a uniform and controlled environment. This fact simplifies the development and allow to avoid a large set of problems coming from the diversity of infrastructures and systems. The final user will also be able to enjoy a richer user experience from an application which encapsulates both infrastructure and platform. Due to the ubiquity of services and data, the user can access the application using a simple web-browser from anywhere, at any time. In order to take advantage of those benefits, the BioMedNet system was developed using a SaaS paradigm.

3.2 Graphical Representation

As the interaction with the end user was one the main concerns of the system, it was essential to develop an interface that maximizes the visualization and analysis of data by the user. From the human perspective, one of the easiest ways to access a network is through the graphical representation of graphs.

Cytoscape Web[9] is one of the available options to incorporate graph representation and interaction in a web page. It is a visual component capable of rendering graphs inside HTML pages. This tool has a JavaScript API which is able to control either data displayed, visual styles and behaviour of the component.

The interaction with a Cytoscape Web component is quite simple and straightforward. With the exclusive assistance of a mouse the user can navigate within

[9] Cytoscape Web, `http://cytoscapeweb.cytoscape.org/`

the graph area, zoom in and out and even select elements (vertices or edges). It is possible to define different algorithms which toggle the position of vertices and edges into several layouts: circular, radial, force directed, etc. We can also customize the visual options of the graph by changing vertices shape and caption or edges thickness and colour. It is also possible to create a powerful interaction assigning mouse click events to different graph elements.

Lopes et al.[13] conducted several performance tests showing that Cytoscape Web can efficiently handle graphs with up to 2000 elements (800 vertices and 1200 edges).

3.3 Logical Representation

Although the graphical representation of a graph is the better way to interact with users, it is not suitable for data exchanges between applications or computational components. For that purpose it is more appropriate to have a logical representation of a graph.

GraphML[10] is a XML document format that is able to represent graphs, its elements and topology. This format has the ability to represent small and simple graphs, but also complex mathematical concepts that were not covered in our system. One of the greatest advantages of using GraphML is that it allows to directly import and export data into and from Cytoscape Web.

3.4 Graph Modelling and Analysis

Once a network is modelled into a graph it is possible to apply the mathematical findings of graph theory to study and analyse the network. JUNG[11] is a software framework that fits this problem. It provides a good set of features and algorithms that allow to process the data in a graph topology. Since it is an open-source solution, it is also easy to create custom plugins and algorithms to solve different problems.

4 BioMedNet System

The BioMedNet system was entirely developed using Java programming language and some related technologies. This way, it is possible to assure independence from any operating system. Benefiting from Java multi-platform capacities, any manager can easily configure and deploy the application, on any production environment.

4.1 Components

The BioMedNet system was created based on three major components (Fig. 1) named: BMNetwork, BioMedNet and BioMedNetPortal.

[10] GraphML, `http://graphml.graphdrawing.org/`

[11] Java Universal Network/Graph Framework, `http://jung.sourceforge.net/`

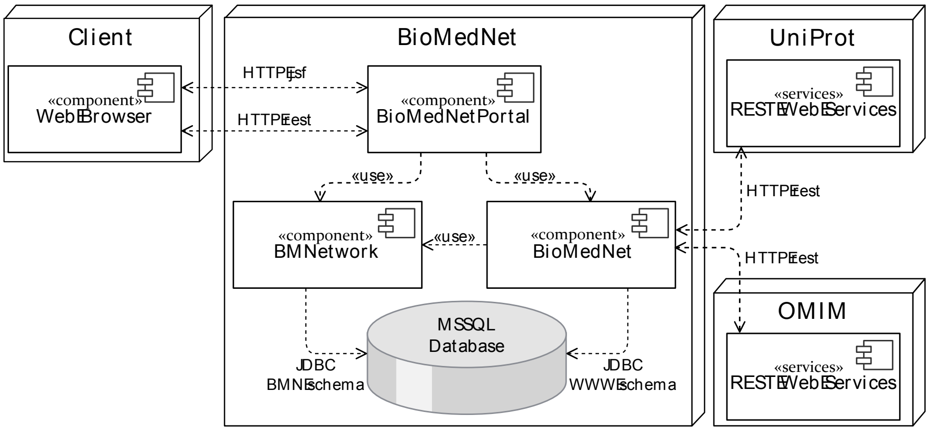


Fig. 1. BioMedNet Deployment Diagram

The BMNetwork component deals with all storage processes, allowing to store, access, edit and process any graph. This framework was developed in a generic way to handle different types of nodes and edges, which makes it possible to use it in several different scenarios. Although the component manages the data manipulations, it requires a database solution to store all the information. Different databases may need to implement different connection modules. When used properly, the framework also allows to process the stored graph using JUNG native features and some of the plugins developed especially to this project.

The BioMedNet component represents the core of the application. It extensively uses the BMNetwork component (to create and access a biomedical graph) as well as the SQL database. The two main functionalities of this component are:

Application Script - It implements a specific procedure necessary to initiate the BioMedNet system. This preprocessing is responsible for collecting and storing all the biological data required by the system.

Library - It also incorporates the data access layer (DAL) and business logic layer (BLL) used by the Web application.

The BioMedNetPortal component handles the Web portal elements: pages or RESTful services. It implements a REST interface and keeps track of all Java Managed Beans necessary for the proper operation of all Web pages.

4.2 Collecting Data

The biological data used by BioMedNet system needs to be collected before deploying the Web application. In order to gather all the data some external flat files and Web Services are used (Fig. 2).

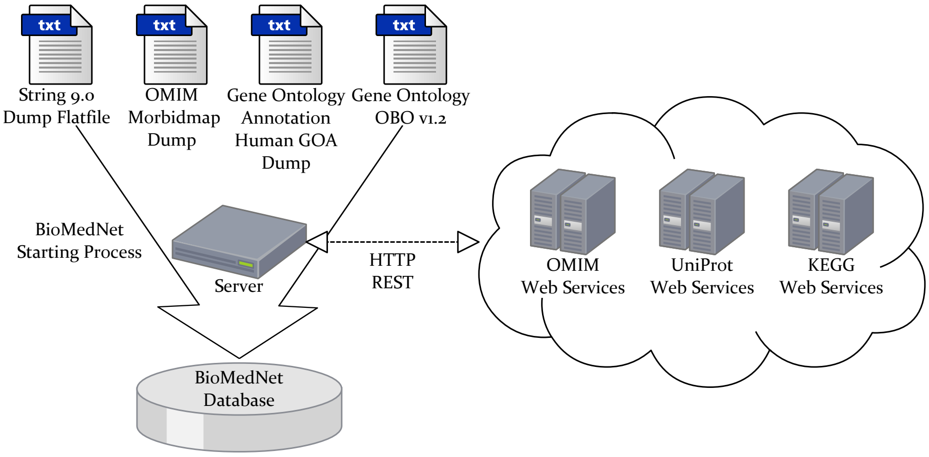


Fig. 2. Data Collection Process

The created graph has four types of vertices, identified by four different notations: UniProt (proteins or genes), MIM (disorders or diseases), GO (ontologies, biological processes, cellular components or molecular functions) and KEGG (pathways). It also includes four types of relations among these elements: Protein - Protein, Disease - Protein, Ontology - Protein, and Pathway - Protein.

5 Results

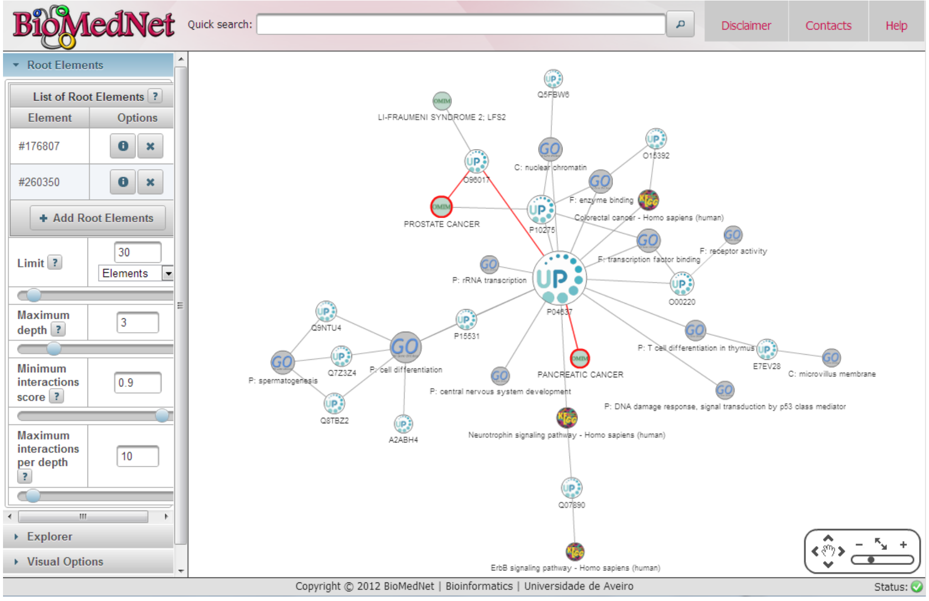
BioMedNet can be accessed through its Web-based interface or its RESTful Web Services. The Web interface is composed of two essential elements: a side bar control menu and the work area where we can visualize a specific subgraph (Fig. 3).

The main entry point is the identification of the *root elements* which are the biological concepts under study. The user can define root elements by directly selecting or searching for them. The *search* features was implemented using UniProt and OMIM Web Services.

For each set of root elements it is represented a subgraph that is a subsection of the full logical graph. Several visualization options can be controlled in order to filter the graph by: maximum number of vertices or edges, maximum subgraph depth, minimum edge confidence score and maximum number of edges per depth.

The user can also apply several graph processing features, such as, search for relations between two elements or rank elements according to different types of measures: degree centrality, betweenness centrality and closeness centrality. Furthermore, it is possible to apply several different visual styles that can help in the process of analysing each sub-network.

In the following example (Fig. 3), we are trying to study some possible relations between pancreatic cancer and prostate cancer. The two diseases were



Root Elements

List of Root Elements ?

Element	Options
#176807	i ✖
#260350	i ✖

+ Add Root Elements

Limit ? 30 Elements

Maximum depth ? 3

Minimum interactions score ? 0.9

Maximum interactions per depth ? 10

Copyright © 2012 BioMedNet | Bioinformatics | Universidade de Aveiro

Status: ✓

BioMedNet
Quick search:
Disclaimer
Contacts
Help
Explorer
Visual Options
Q5FBW6
LI-FRAUMENI SYNDROME 2; LFS2
C: nuclear chromatin
O15392
O96017
F: enzyme binding
PROSTATE CANCER
P10275
Colorectal cancer - Homo sapiens (human)
F: receptor activity
F: transcription factor binding
P: rRNA transcription
O00220
P04637
Q9NTU4
P15531
P: T cell differentiation in thymus
E7EV28
C: microvillus membrane
Q7Z3Z4
P: cell differentiation
P: spermatogenesis
PANCREATIC CANCER
P: central nervous system development
P: DNA damage response, signal transduction by p53 class mediator
Q8TBZ2
A2ABH4
Neurotrophin signaling pathway - Homo sapiens (human)
Q07890
ErbB signaling pathway - Homo sapiens (human)

Fig. 3. BioMedNet Web Interface - Relations between Prostate and Pancreatic Cancers

previously selected as *root elements*, so they are represented with a thick red border. Since we are seeking for relations among these two entities, the result is shown through red edges.

From the obtained result, it is possible to understand that genes P04637 and O96017 play an important role in the relations between the two diseases.

The size of the nodes represent the measure that each node scored in a betweenness centrality ranking. It is graphically comprehensible that the gene P04637 is the most relevant in the paths between all the displayed entities. If we had applied a different ranking method, different conclusions would have been obtained from the newly calculated scores.

6 Conclusion

Life sciences has benefited from the possibility of integrating disperse data in order to create new scientific knowledge. The presented application integrates in a graph representation, a network of biomedical concepts.Though the visualization of the network representation and application of some graph theory features, researchers can improve the way they study and understand relations between biological elements.

There are many qualities in the BioMedNet system that make it a practical tool for different biological researches including: the ability to filter the logical graph by several central points, to seek the relations between different elements, and to classify the entities by several metrics.

References

1. Kenneth H Wolfe and Wen-Hsiung Li. Molecular evolution meets the genomics revolution. *Nature genetics*, 33:255–265, March 2003.
2. Hans V Westerhoff and Bernhard O Palsson. The evolution of molecular biology into systems biology. *Nature biotechnology*, 22(10):1249–1252, October 2004.
3. Joel P Arrais, João Fernandes, João Pereira, and José Luís Oliveira. GeneBrowser 2: an application to explore and identify common biological traits in a set of genes. *BMC bioinformatics*, 11:389, January 2010.
4. Xuebing Wu and Shao Li. Cancer gene prediction using a network approach. *Cancer Systems Biology*, pages 191–212, 2010.
5. Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M. Humphrey. Using literature-based discovery to identify disease candidate genes. *International journal of medical informatics*, 74(2-4):289–298, March 2005.
6. Joel P Arrais and José Luís Oliveira. Using biomedical networks to prioritize gene-disease associations. *Open Access Bioinformatics*, pages 123–130, 2011.
7. Alexander Martin, Maria E Ochagavia, Laya C Rabasa, Jamilet Miranda, Jorge Fernandez-de Cossio, and Ricardo Bringas. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC bioinformatics*, 11(1):91, January 2010.
8. Rajkumar Buyya, James Broberg, and Andrzej M. Goscinski. *Cloud Computing: Principles and Paradigms*. Wiley Series on Parallel and Distributed Computing. John Wiley & Sons, 2011.
9. Eriksson J. Melicio Monteiro, Luis A. Bastiao Silva, and Carlos Costa. CloudMed: Promoting telemedicine processes over the cloud. *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on. IEEE*, pages 1–6, 2012.
10. Subbu Allamaraju. *RESTful Web Services Cookbook: Solutions for Improving Scalability and Simplicity*. Oreilly Series. O'Reilly Media, Inc., 2010.
11. Miltiadis D. Lytras, Ernesto Damiani, and Patricia Ordóñez de Pablos. *Web 2.0: The Business Model*. Springer, 2008.
12. Nick Antonopoulos and Lee Gillam. *Cloud Computing: Principles, Systems and Applications*. Computer Communications and Networks. Springer, 2010.
13. Christian T. Lopes, Max Franz, Farzana Kazi, Sylva L. Donaldson, Quaid Morris, and Gary D. Bader. Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18):2347–2358, September 2010.