Querying and analyzing biological data with BioGraph

Antonio Messina[1], Antonino Fiannaca[1], Laura La Paglia[1], Massimo La Rosa[1], Alfonso Urso[1]

[1] CNR - ICAR, National Research Council of Italy, Palermo, Italy

Corresponding Author:

Antonio Messina[1]

via Ugo La Malfa 153, 90146 Palermo, Italy

Email address: antonio.messina@icar.cnr.it

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3309v1 | CC BY 4.0 Open Access | rec: 2 Oct 2017, publ: 2 Oct 2017

Querying and analyzing biological data with BioGraph

Antonio Messina[1], Antonino Fiannaca[1], Laura La Paglia[1], Massimo La Rosa[1] and Alfonso Urso[1]

[1] CNR - ICAR, National Research Council of Italy, Palermo, Italy

Introduction

Nowadays, a huge amount of biomedical data of different biological entities is provided by many online databases and services, each with its own data model, user interface and query language. However, typical bioinformatics scenarios require the use of more than one resource. Therefore, the availability of a single bioinformatics platform that integrates many biological resources and services is a fundamental issue.

Some attempts to go beyond the drawbacks of a handcrafted combination of different resources have already been made. HumanMine [1], Java BioWareHouse [2], and Bio4J [3] are some examples of such tools. Regardless of focus and biological data sources considered by them (and also by others), they all lack one or more high-level functional features, such as web interface availability, dynamic data visualization, custom queries support, analytics, and expandability.

Here, we present BioGraph, a web application that allows to query, visualize and analyze biological data belonging to several online available resources. BioGraph is built on top of our previously developed graph database called BioGraphDB [4], which collects and integrates heterogeneous biological data and makes them available through a common structure and Gremlin [5] as unique query language. BioGraph makes use of state-of-the-art technologies and provides some pre-compiled bioinformatics scenarios, as well as the possibility to perform custom queries and obtaining an interactive and dynamic visualization of results.

Methods

BioGraph is characterized by a highly modular and scalable architecture to assure responsiveness and performances. The application's stack is made up of three levels. From the bottom to the top:

- *Graph Data level*, overlooked by Apache Tinkerpop (http://tinkerpop.apache.com) with his Gremlin Server, which provides a way to remotely execute Gremlin queries against graph instances hosted within it, such as BioGraphDB, built as a Neo4j (https://neo4j.com) instance;
- *Microservices level*, composed of a set of microservices used to deal with the management, transformation and production of queries results emitted by the graph engine, to implement the word auto-completion features, and to compute p-values using the right-tailed Fisher exact test.
- *Web Application level*, built with Bootstrap (https://getbootstrap.com), a popular framework to create responsive web front-ends. Dynamic content manipulation, event handling, effects, and asynchronous data transfers are managed by jQuery (http://jquery.com). Graphs visualizations and interactions are handled by Cytoscape.js (http://js.cytoscape.org), a powerful JavaScript library which allows easy display and manipulation of rich interactive graphs, via all the common gestures, such as panning, box selection, pic-to-zoom, et cetera.

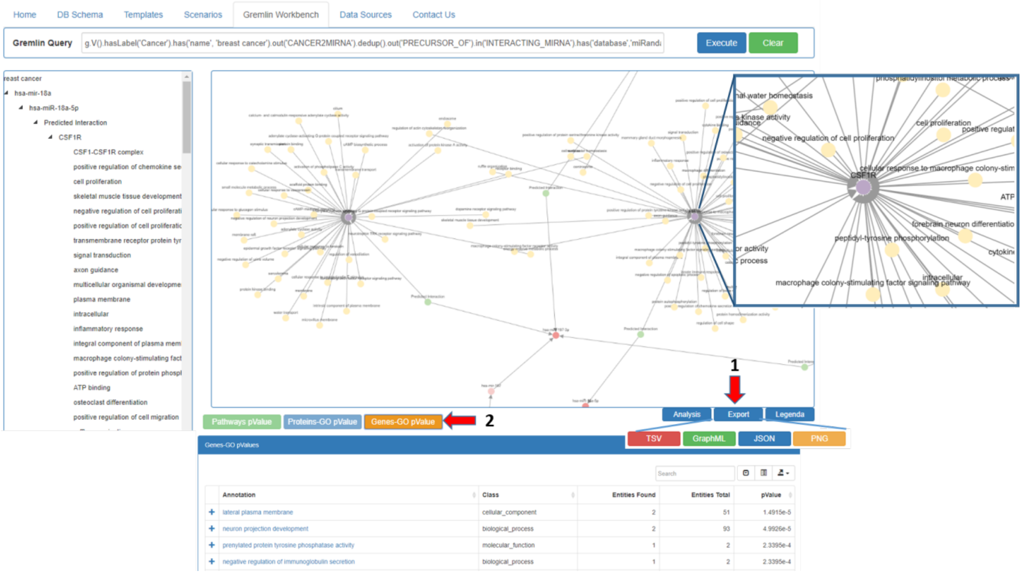
The web user interface is organized in tabs, and the *Gremlin Workbench* is the place where the most of user's activities are performed (typing and executing queries, interactive browsing of results in the tree and graph views, displaying data details). Also, simple and complex predefined queries are available in the *Templates* and *Scenarios* tabs.

Results

The proposed system has been used to deal with a common bioinformatics scenario, that is the functional analysis of deregulated microRNA (miRNA) in breast cancer. Using traditional services and databases, that case study needs at least four different resources. First of all, in fact, given the pathology, a set of deregulated miRNAs should be selected, for example from miRCancer. Then, using miRBase and online databases collecting miRNA-target interactions such as miRanda, a set of mRNA related to deregulated miRNAs can be identified. Finally, using Gene Ontology (GO) it is possible to obtain statistical meaningful functional annotations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3309v1 | CC BY 4.0 Open Access | rec: 2 Oct 2017, publ: 2 Oct 2017

BioGraph, on the other hand, makes available a single platform with an integrated knowledge base and a single query language to face the scenario. It, in fact, allows avoiding all those annoying steps by composing a Gremlin query that defines the starting point and the sequence of available resources to consult, eventually specifying custom filters and query parameters. For the proposed case study, BioGraph provides a result such as the one depicted in Figure 1. There, on the left part there is an interactive tree showing nodes and edges traversed by executing the Gremlin query. In the central part of Figure 1, there are the final and intermediated query results represented as an interactive graph. The graph's leaves represent the desired functional annotations, given the set of deregulated miRNAs related to a specific cancer. Each element of the graph, if clicked, provides a set of detailed information that are visualized in the Details Section, on the low part of Figure 1.



Gremlin Query: g.V().hasLabel('Cancer').has('name', 'breast cancer').out('CANCER2MIRNA').dedup().out('PRECURSOR_OF').in('INTERACTING_MIRNA').has('database','miRand

reast cancer
- hsa-mir-18a
  - hsa-miR-18a-5p
    - Predicted Interaction
      - CSF1R
        - CSF1-CSF1R complex
        - positive regulation of chemokine se
        - cell proliferation
        - skeletal muscle tissue development
        - negative regulation of cell proliferati
        - positive regulation of cell proliferatic
        - transmembrane receptor protein tyr
        - signal transduction
        - axon guidance
        - multicellular organismal developme
        - plasma membrane
        - intracellular
        - inflammatory response
        - integral component of plasma mem
        - macrophage colony-stimulating fact
        - positive regulation of protein phospl
        - ATP binding
        - osteoclast differentiation
        - positive regulation of cell migration

	Annotation	Class	Entities Found	Entities Total	pValue
+	lateral plasma membrane	cellular_component	2	51	1.4915e-5
+	neuron projection development	biological_process	2	93	4.9926e-5
+	prenylated protein tyrosine phosphatase activity	molecular_function	1	2	2.3395e-4
+	negative regulation of immunoglobulin secretion	biological_process	1	2	2.3395e-4

Figure 1: Typical visualization of the results obtained by a query in BioGraph

Results can be exported in textual and graphical formats, including JSON that can be imported into Cytoscape framework. Furthermore, it is possible to select the "Gene-GO pValue" function to compute the p-value of the returned functional annotations.

Conclusions

BioGraph is a web application that allows to query, visualize and analyze biological data that are originally collected in different online databases. BioGraph is built on top a graph database that integrates several bioinformatics resources in order to provide a single platform useful to deal with bioinformatics scenarios that involve more than one resource. BioGraph can be easily expanded with new biological resources, increasing this way the kind and number of supported scenarios, and updated with the latest version of the already integrated data. BioGraph is available at http://biograph.pa.icar.cnr.it.

References

1. Smith, R.N., et al. "InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data.", Bioinformatics 28(23), 3163–5 (2012)
2. Vera, R., et al. "JBioWH: an open-source Java framework for bioinformatics data integration.", Database 2013, 051–051 (2013)
3. Pareja-Tobes, P., et al. "Bio4j: a high-performance cloud-enabled graph-based data platform". Technical report, Era7 bioinformatics (2015)
4. Fiannaca, A., et al. "BioGraphDB: a New GraphDB Collecting Heterogeneous Data for Bioinformatics Analysis". BIOTECHNO 2016: The Eighth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, pp. 28–34 (2016)
5. Rodriguez, M.A. "The Gremlin graph traversal machine and language". Proceedings of the 15th Symposium on Database Programming Languages, pp. 1–10. ACM Press, New York, USA (2015)