An Approach for Discovering and Exploring Semantic Relationships between Genes

Nicoletta Dessì, Matteo Pani, Barbara Pes, and Diego Reforgiato Recupero
Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica,
Via Ospedale 72, 09124 Cagliari, Italy
dessi@unica.it, matteopani@yandex.com, {pes,diego.reforgiato}@unica.it

Abstract. This paper presents an approach for extracting, integrating and mining the annotations from a large corpus of gene summaries. It includes: i) a method for extracting annotations from several ontologies, mapping them into concepts and evaluating the semantic relatedness of genes, ii) the definition of a NoSQL graph database that leverages a loosely structured and multifaceted organization of data for storing concepts and their relationships, and iii) a mechanism to support the customized exploration of stored information. A prototype with a user-friendly interface fully enables users to visualize all concepts of their interest and to take advantage of their visualization for formulating biomedical hypotheses and discovering new knowledge.

Keywords: Gene Relatedness, Graph Database, Neo4j, Biomedical text annotation, Data integration, Knowledge discovery.

1 Introduction

Ontologies are paramount to annotate biomedical texts with unambiguous topics about biomolecular entities such as genes and their protein products. These annotations consist of specialized terms related by semantic relations and have a great potential in supporting the interpretation of biological events and the formulation of biomedical hypotheses. Often the use of different ontologies for annotating a single text can help overcoming the incompleteness of annotations extracted using a single ontology and understanding complex biological processes and multi-faceted biomedical questions.

Several tools have been created to store, search and use multiple ontologies at the same time. Among them, we cite the National Institute of Health the supports UMLS (Unified Medical Language System) services [1], the EBI (European Bioinformatics Institute) Ontology Lookup Service [2], the NCBO (National Center for Biomedical Ontology) BioPortal [3]. Despite their specificities, these portals share an overall common objective: providing to the biomedical community a unique service for exploiting semantic resources. Usually, this service comes both with a programmatic and a friendly Web user interface: users insert the text to be annotated, select the ontologies of their interest and download the annotations in their own computers without the need of installing any additional software.

The integration of concepts resulting from the annotation of a corpus of documents is considered the next step of text-mining research community [4] and presents tremendous opportunities for accelerating investigations already under way and taking advantage of additional knowledge revealed by data. Unfortunately, annotations are stored in a plethora of downloaded files and manually integrating their content is hard and error prone. Thus, a computational environment to support this integration requires a) to structure annotated concepts in a database and b) to build a tool that supports user queries over this database. To the best of our knowledge, both a) and b) are not yet completely solved challenges.

To overcome these limits, our idea is to leverage NoSQL [5] databases that exhibit a loosely structured and multifaceted organization of data. In particular, graph databases support data representation in a graph structure where nodes denote biomedical concepts and edges between nodes represent their relationships. Studying the topological structure originated by a graph database from annotated concepts can contribute to explain or investigate causes of biomedical events. As such, the paper proposes an approach for extracting, mining and visualizing the annotations obtained from a large corpus and their relationships.

Specifically, we considered annotating the corpus of gene summaries supplied by NCBI[1] (National Center for Biotechnology Information), one of the most important and largest collections of documents about genes that is freely available on the Internet. Compiled by expert curators, summaries condense into a short text the description of functions and processes in which genes are involved. To annotate these summaries, we exploited biomedical ontologies compliant with the UMLS standard. The resulting annotations can serve several purposes: they not only provide a visual representation of gene relatedness but, at the same time, standardized UMLS concepts[2] can help infer new gene functionality.

Designed to fully help users in exploring semantic annotations, the approach we propose in this paper includes:
(i) a method for extracting annotations from several ontologies and evaluating the semantic relatedness of genes;
(ii) a graph NoSQL database that stores concepts and their relationships,
(iii) a mechanism and a demo that supports the customized exploration of concepts and their relationships that can be freely downloaded[3].

A user-friendly graphical interface supports the easy composition of multi-topic queries and their semantic expansion upon the integrated annotations; such an interface fully enables users to visualize all concepts of their interest that match, syntactically or semantically, the performed query, to further explore concepts and take advantage of this visualization for biomedical hypotheses formulation and knowledge discovery.

More in detail, this paper is organized as follows. Section 2 presents how the annotation process is carried out within the proposed approach. Section 3 describes the procedural component that sets up and populates the graph database. The

[1] https://www.ncbi.nlm.nih.gov/
[2] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/
[3] https://github.com/mattux/qgenes

customized exploration of concepts and their relationships is presented in section 4. Section 5 presents the related work and section 6 outlines conclusions.

2 Annotating content from Web resources

Biomedical texts deal with many specific domains that are not covered by a single ontology. Moreover, the description of biomedical concepts in natural language requires tackling lexical issues such as recognition of synonyms and redundant parts of speech, disambiguation of terms etc. Finally, annotations need to be filtered out because they are not always pertinent to the domain of interest.

To face the above concerns, we took advantage of the NCBO annotator available in Bioportal[4] that supports the selection of several ontologies and performs annotations on a highly efficient syntactic concept recognition (i.e. using concept names and synonyms). Direct annotations are created from raw text according to a dictionary that uses terms from a set of ontologies. Different components expand the first set of annotations using ontology semantics. Most important, when the selection includes ontologies compliant with UMLS, the user can filter annotations by specifying a list of UMLS semantic types. According to this list, NCBO Annotator maps each annotated term to one or more UMLS CUIs (Concept Unique Identifier). This avoids the above mentioned lexical problems because a single CUI identifies a specific concept which can be described by different terms in several ontologies.

In the approach we propose, the annotation process is carried out by specialized procedural components that perform the following tasks:

Task 1 – Obtaining gene summaries
This task obtains gene summaries of protein-coding genes (11840 summaries in total) using the Entrez Programming Utilities provided by the Bio.Entrez library[5] and available from Biopython[6].

Task 2 – Annotating summaries
This procedural component programmatically accesses NCBO Annotator using REST (REpresentational State Transfer) APIs and performs the annotations of summaries by setting up NCBO Annotator with a list of ontologies and semantic types.

This task results in a file where each record pertains to a specific summary and contains the name of the gene the summary is about and its annotations, namely its bag-of-concepts (BOC). Each summary received an average of 18 annotations.

Task 3 - Evaluating the relatedness of gene pairs.
The correlation between a pair of genes is usually measured by their semantic similarity, i.e. their distance evaluated on the graph of the ontology. This measure is fully dependent on the ontology structure and treats the terms at the same topological

[4] https://www.bioontology.org/annotator-service
[5] http://www.ncbi.nlm.nih.gov/books/NBK25501/
[6] http://biopython.org/DIST/docs/api/Bio.Entrez-module.html

level equally. Usually ontologies do not include relations other than *is-a* and *part-of*. However, gene similarity is multifaceted in nature as genes interact in many and heterogeneous ways.

These motivations have suggested recent research [6,7] to consider a more suitable measure called the relatedness, that accounts for many relationships including *causes, treats, affects, symptom-of* etc. All similar concepts are related, but not all related concepts are similar. For example, *bronchitis* and *cough* are related (*bronchitis* can cause *cough*), but they really are not similar because *bronchitis is-not-a chough* and vice versa. The relatedness of two entities is usually asserted by evaluating the overlap of the specialized terms the entities share within a specialized corpus [8].

Given an input gene A, we evaluate its relatedness (R) with another gene B by calculating their functional similarity fraction (i.e., the overlapping of their BOCs) according to the *Jaccard* similarity index:

$$R = (\mid A \cap B \mid) / (\mid A \cup B \mid)$$

This index represents the size of the intersection of the genes' BOCs divided by the size of their union. A value of 0 indicates no overlap; a value of 1 indicates perfect agreement. This task results in a file where each record contains the name of a gene, the list of its related genes and the measure of their relatedness.

3 Setting up and populating the database

Neo4j[7] is a highly scalable, native graph database built to leverage not only data but also its relationships. Within Neo4j the user naturally stores, manages, analyzes, and uses his data within the context of connections, like the circles and lines drawn on whiteboards. That said, Neo4j organizes data in a graph structure where: (i) nodes model entities labeled with one or more labels, (ii) nodes have properties expressed by key-value couples, (iii) oriented and named links represent relationships between nodes, and (iv) links have properties.

Accordingly, we model relations between genes in a graph where nodes represent genes and links express their relationships. Specifically, a node has the following properties: *GID* = the Entrez identifier of the gene, *Name* = the gene name, *BOC* = the list of concepts associated to the gene, *Summary* = the gene summary, *Type* = the gene type (i.e. protein-coding). Links are featured by the following properties: *Weight* = the measure of the relatedness between the linked gene pair, *SharedCUIs* = the list of concepts shared by the linked gene pair.

We hosted the graph database in a specialized database server in order to completely separate the database structure from the logic applications that leverage it. Fig. 1 shows an excerpt from the graph database.

[7] https://neo4j.com/

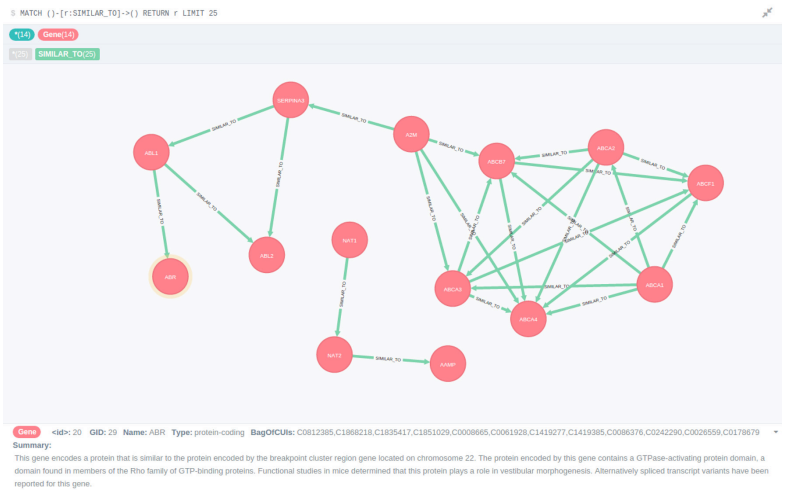

Fig. 1. An excerpt of gene database in Neo4j.

4 A prototype for exploring gene relatedness

As shown in Fig. 2, the general architecture of our approach consists of a client that supports user interaction and a Web Service that sends the user query to the Neo4j server and returns the answer to the client. The query is coded in a JSON message and sent to the Web Service that parses the message, converts the query and sent it to the Neo4j server. To speed queries up, genes are identified by their ID property. The Neo4j server executes the query and sends the answer (again a JSON message) to the Web Service that, in turns, sends it to the client.

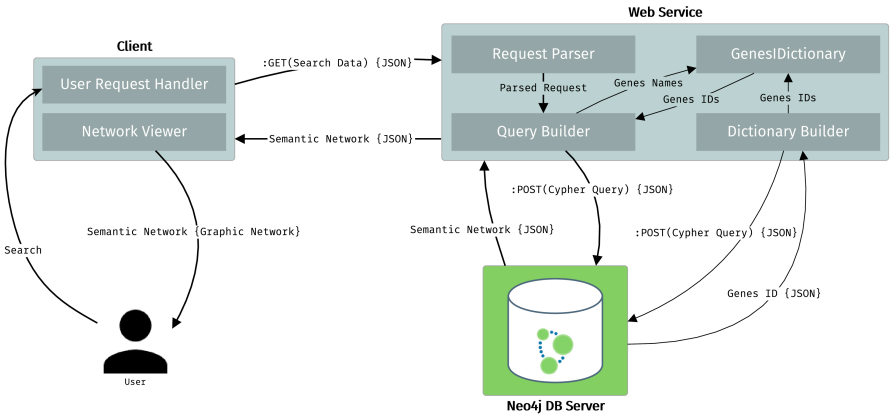

Fig. 2. The architecture of the implemented prototype.

Within our approach, this general architecture serves as a reference for different application contexts. We have developed a prototype that provides search and visualization facilities to interactively explore gene relatedness. It can be freely downloaded[8]. Neo4j Community Edition (licensed under the free GNU General Public License (GPL)) provided the database support. Neo4j has quickly risen to become the most popular graph database in the world as rated by DB-Engines and by hundreds of thousands of Neo4j community members.

For implementing the prototype, we used Java as programming language and the following open source technologies:

- Glassfish[9], an open-source application server for the Java EE platform.
- Jersey RESTful Web Services framework[10], an open source, production quality, framework for developing RESTful Web Services in Java that supports JAX-RS.
- JSONPath[11], a Java library for managing JSON documents.
- Vis.js[12], a dynamic, browser based visualization library to handle large amounts of dynamic data, and to enable manipulation of and interaction with the data.
- The Graphical User Interface was developed using JQuery UI[13].

5 Related Work

As in our knowledge, few work has been done to explore the potential of Neo4j in supporting biomedical applications. Similar to our approach, [9] proposes a framework for mining associations between concepts in a large dictionary using different data sources. The framework benefits from the visualization capabilities of Neo4j for analyzing concepts and their associations.

Many tools exist for visually exploring semantic networks [10]. Some of them are applicable to a wide range of problems, others are specialized for specific applications such as protein-protein interactions, pathways analysis, gene networks. Among broadly applicable tools, Cytoscape [11] is currently a gold standard for large scale network visualization. It can support directed, undirected and weighted graphs and provides customizable visual styles that allow the user to change the properties of nodes or edges. Moreover, it incorporates statistical analysis as well as network filtering capabilities. A broad variety of additional features are made available as plug-ins (apps) mainly developed by high-experienced users.

Already established tools are also BioLayout Express3D [12], ProViz [13] and VisANT [14] that provide support for handling very large graphs. If integration of heterogeneous data is a major challenge, Ondex [15] offers possible solutions. When dealing with highly interconnected data that involve multiple biological relationships, tools featuring multi-edge networks, such as Medusa [16] are a suitable choice.

---

[8] https://github.com/mattux/qgenes
[9] https://glassfish.java.net/
[10] https://jersey.java.net/
[11] https://www.npmjs.com/package/JSONPath
[12] http://visjs.org/
[13] https://jqueryui.com/

It has been observed [17] that an adequate support for exploiting the semantic web to its full potential requires structured knowledge representations, mappings between resources, data sharing and use of semantic web technology standards.

In this vein, a number of recently developed tools profit from Web 2.0 technology. In particular, NaviCell [18] relies on Google Maps for supporting user-friendly exploration of large-scale maps at different scales. Similarly, CellPublisher [19] and PathVisio [20] exploit the geographical metaphor for navigating within the maps. Payao [21] and WikiPathways [22] are web-based platforms for collaborative annotation and curation of biological networks.

The importance of interaction is stressed by COWB (COllaborative Workspaces in Biomedicine) [23], a cloud-based framework which supports collaborative knowledge management in the context of biomedical communities. Supported by a NoSQL database, public and private workspaces provide an accessible representation of the biomedical entities that are graphically visualized and structured using highly interactive graphical interfaces.

Finally, SSAIIB (Smart Spaces for Adaptive Information Integration in Bioinformatics) [24] is a reference framework for designing "Smart Spaces", i.e. software environments that adaptively support specific user's activities such as discovering, aggregating and delivering contents from web resources according to user's goals, tasks and concerns. A case study is presented that shows the functionality of a smart space for annotating biomedical text.

6 Conclusions

This paper has proposed an approach for capturing, mining and visualizing associations between concepts from a text corpus using domain ontologies as reference knowledge. Although created for exploring gene relatedness from gene summaries, the method is quite modular and can be adapted for annotating and exploring biomedical literature. A demo of our approach can be freely downloaded.

The proposed approach can be improved in several aspects. An interesting aspect could be investigating the extension of the type of associations between concepts in order to include additional links between nodes, such as virtual links introduced by the researcher to test some hypothesis. Another research aspect could be the comparison of clusters originated by semantic relatedness with clusters detected by traditional data mining methods [25].

References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. PMID: 14681409 (2004).
2. Coté, R., Reisinger, F., Martens, L., Barsnes, H., et al.: The Ontology Lookup Service: bigger and better. Nucleic Acids Res. PMID: 20460452(2010).
3. Whetzel, P.L., and NCBO Team: NCBO Technology: Powering semantically aware applications. J Biomed Semantics, 4(Suppl 1): S8 (2013).

4. Li, C., Liakata, M., Rebholz-Schuhmann D.: Biological network extraction from scientific literature: state of the art and challenges. Briefings in Bioinformatics 15 (5), 856-877 (2013).
5. Stonebraker, M.: SQL databases v. NoSQL databases. Commun. ACM 53(4): 10-11 (2010).
6. Rybinski, M., Aldana-Montes, J.F.: tESA: a distributional measure for calculating semantic relatedness, Journal of Biomedical Semantics 7:67. (2016).
7. Ben Aouicha, M., Hadj Taieb, M.A., Ben Hamadou, A.: SISR: System for integrating semantic relatedness and similarity measures. Soft Comput (2016).
8. Dessì, N., Dessì, S., Pascariello, E., Pes, B.: Exploring the relatedness of gene sets. In Proceedings of 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, CIBB 2014, LNCS, Vol. 8623, pp. 44-56, Springer (2015).
9. Sadoddin, R., Driollet, O.: Mining and Visualizing Associations of Concepts on a Large-Scale Unstructured Data. BigDataService 2016: 216-224.
10. Pavlopoulos, G.A., Wegener A., Schneider R: A survey of visualization tools for biological network analysis. BioData Mining 2008, 1:12.
11. Shannon, P., Markiel A., Ozier, O., Baliga, N.S, et al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498-2504 (2003).
12. Goldovsky, L., Cases, I., Enright, A.J., Ouzounis, C.A.: BioLayout (Java): versatile network visualisation of structural and functional relationships. Appl Bioinform, 4(1):71-74, (2005).
13. Iragne, F., Nikolski, M., Mathieu, B., Auber, D., Sherman, D.: ProViz: protein interaction visualization and exploration. Bioinformatics, 21(2):272-274 (2005).
14. Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., Delisi, C.: VisANT: data-integrating visual framework for biological networks and modules. Nucleic Acids Res. 2005, 33, W352–W357.
15. Kohler, J., Baumbach, J., Taubert, J., Specht, M., et al.: Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 2006, 22(11):1383-1390.
16. Pavlopoulos, G.A., Hooper, S.D., Sifrim, A., Schneider, R., Aerts, J.: Medusa: A tool for exploring and clustering biological networks. BMC Research Notes 2011, 4:384.
17. Kuperstein, I., Cohen, D.P.A., Pook, S., Viara, E., et al.: NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. BMC Systems Biology 2013, 7:100.
18. Machado, C.M., Rebholz-Schuhmann, D., Freitas, A.T., Couto F.M.: The semantic web in translational medicine: current applications and future directions, Briefings in Bioinformatics 16 (1), 89-103 (2013).
19. Flórez, L.A., Lammers, C.R., Michna, R., Stülke, J: Cell Publisher: a web platform for the intuitive visualization and sharing of metabolic, signalling and regulatory pathways. Bioinformatics 2010, 26:2997–2999.
20. Van Iersel, M.P., Kelder, T., Pico, A.R., Hanspers, K., et al.: Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 2008, 9:399.
21. Matsuoka, Y., Ghosh, S., Kikuchi, N., Kitano, H.: Payao: a community platform for SBML pathway model curation. Bioinformatics 2010, 26:1381–1383.
22. Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., et al: WikiPathways: pathway editing for the people. PLoS Biol 2008, 6:e184.
23. Dessì, N., Milia, G., Pascariello, E., Pes, B.: COWB: A cloud-based framework supporting collaborative knowledge management within biomedical communities, Future Generation Computer Systems 54 (2016) 399–408
24. Dessì, N., Pes, B.: Smart Spaces for Adaptive Information Integration in Bioinformatics, Future Generation Computer Systems 68 (2017) 407–415
25. Reforgiato, D., Gutierrez, R., Shasha, D.: GraphClust: A Method for Clustering Database of Graphs, Journal of Information & Knowledge Management, Volume 07, Issue 04 (2008)