Ranking-aware integration and explorative search of distributed bio-data

Marco Masseroli✉, Matteo Picozzi, Giorgio Ghisalberti
Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Motivation and Objectives

High-throughput production of both biomolecular data and their annotations is providing a rapidly increasing amount of very valuable information that can potentially help finding also long-searched answers to fundamental biomedical questions. Yet, such data deluge makes difficult to extract the information most reliable and most related to the increasingly complex biomedical questions to be answered, which can simultaneously regard many heterogeneous aspects of single or multiple organisms, biological tissues, cells or biomolecular entities. To address such complex questions, many bio-data about several heterogeneous topics, which are available but dispersed in different data sources, must be searched, extracted, integrated and comprehensively queried.

Different approaches have been proposed to combine individual search services available on the Web in order to support such heterogeneous searches (Hull et al., 2006; Nekrutenko, 2010). Yet, they rarely rely on a general model of the services to be integrated and none considers, in the integration process, the often available partial rankings of the data to be integrated. Lately, Search Computing (Ceri et al., 2010) has been proposed as a new software framework to build answers to complex search queries by interacting with a collection of cooperating search services and using ranking and joining of results as the dominant factors for service composition. By leveraging the peculiar features of search services, it offers query approaches, execution plans, plan optimization techniques, query configuration tools, and exploratory user interfaces.

Here, we report and discuss our work aimed at supporting the explorative search of heterogeneous distributed bio-data and the automatic integration and global ranking of their individual search results, also taking into account the partial rankings of individual searches. In so doing, we make a step towards the computational support required for complex biomedical question answering and biomedical knowledge discovery.

Methods

According to the Service Mart modeling approach of Search Computing (Ceri et al., 2010), we selected an initial set of typical biomolecular topics (i.e. Protein, Gene, Gene Expression and Biological Function) and modeled the Service Marts (i.e. the generalized and normalized conceptual description) of the bioinformatics services that provide data regarding such topics. We did so by identifying their main and common attributes and normalizing their names. We also defined the semantic Connection Patterns, i.e. the pair-wise coupling, between Service Marts of services that provide data about different topics. This was done by identifying pairs of normalized attributes of the connected Service Marts and defining their comparison predicates, as conjunctive Boolean expressions, that allow joining their values semantically. In so doing, we defined the Semantic Resource Framework (SRF) depicted in Figure 1, which constitutes the reference used by Search Computing to enable the exploration of the services registered in the framework and integrate the data that they provide (Ceri et al., 2010).

Then, using available Search Computing tools, we registered in the Search Computing framework five bioinformatics search services that provide data about the topics and semantic associations described in the biomolecular

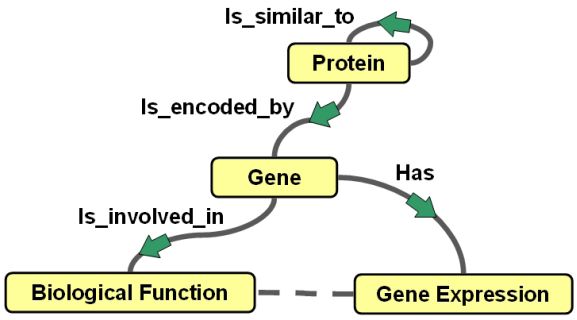

Figure 1: Biomolecular Semantic Resource Framework defined through modeling of data provided by bioinformatics search services and created through service registration. Boxes represent topics of the data provided by the search services registered in the Search Computing framework; lines represent the defined semantic connections created, at registration time, between the registered services.

SRF in Figure 1, i.e. the NCBI Blast (Johnson et al., 2008) and WU-BLAST (Lopez et al., 2003) protein sequence alignment search services, the Array Express gene expression search service (Parkinson et al., 2005), and two access services to the protein coding genes and their biological function annotations (e.g. Gene Ontology annotations) in our Genomic and Proteomic Data Warehouse (GPDW, http://www.bioinformatics.dei.polimi.it/GPKB/). Thus, through service registration, the biomolecular SRF in Figure 1, previously described at conceptual level, is created. To do so, for each service, first we created a wrapper, i.e. an adapter that matches the service attributes to their normalized version defined in a modeled Service Mart, and associated the wrapper with such a Service Mart. Then we defined one or more Access Patterns and Service Interfaces for the service. The latter ones map an access pattern to the end point of a concrete service data source, whereas the former ones are specific signatures of a Service Mart, with the characterization of each attribute as input (I) or output (O), depending on the role that the attribute plays in the service call; furthermore an output attribute can be characterized as ranked (R), if the service produces its results in an order that depends on the value of that attribute. An example Access Pattern for the GPDW Gene to Biological Function Feature (BFF) service is:
(GPDW _ Gene2BFF-Name _ byGeneID(GeneID^{I}, GeneIDName^{I}, BFFName^{I}, BFFID^{O}, BFFIDName^{O}, BFFName^{O}, BFFDefinition^{O}))

Specific Connection Patterns between individual registered services are then automatically derived from the Connection Patterns defined at conceptual level between the modeled Service Marts that have been associated with the registered services.

Results and Discussion

Leveraging the Search Computing framework and biomolecular SRF, which we constructed as previously reported in (Masseroli et al., 2011) and briefly described in the Methods section, we created the Bio Search Computing (Bio-SeCo) application. In particular, in the work here reported, we modeled and registered in Bio-SeCo two additional services and created a Web interface that offers public access to Bio-SeCo at http://www.bioinformatics.dei.polimi.it/bio-seco/seco/. It enables explorative search, automatic integration and global ranking of bio-data individually provided by the services registered in the framework. In this way and thanks to the additional services integrated, Bio-SeCo supports explorative answering of even more complex biomedical questions and biomedical knowledge discovery.

As an example, let us consider the following complex question: Which are the genes (if they exist) that encode proteins, in different organisms, with high sequence similarity to a protein X and have some biomedical features in common (e.g. up/down significantly co-expressed in the biological tissue or condition Y and involved in the biological function Z)? Using Bio-SeCo, a user can first input the UniProt ID of a protein X and run a sequence alignment search, by using the NCBI Blast or WU-BLAST service, to look for amino acid sequences similar to the protein X in a user selected protein database (e.g. UniProtKB Swiss-Prot). Then, he/she can select the most similar proteins found (or some of them, e.g. only those of some selected organisms) and automatically retrieve the coding gene of each of them by using the GPDW protein coding gene query service. Next, the user can search for biomedical features shared among the retrieved genes. For instance, by using the Array Express and GPDW gene biological function annotation services, he/she can explore if some of such genes are significantly co-expressed in the same biological tissue or condition Y and are known to be involved in the biological function Z. For example, the user can set the human *Paired box protein Pax-6 isoform a* protein (UniProt ID P26367) as input protein X, *tumor* as pathological biological condition Y, and *regulation of apoptotic process* as biological function Z. By doing so, unpredictably, on July 20th 2012, Bio-SeCo discovered the human PAX7 and PAX2, mouse Pax8 and human PAX8 genes, ranked by their global score of 0.90661, 0.90407, 0.90354 and 0.90289, respectively (with 1.0 as best score). This global score is computed by Bio-SeCo according to a score function defined as a combination of partial scores of intermediate ranked results, i.e. of the ranked sequence alignment expectation and gene expression p-value. To compute the global score, we adopted the Fagin method (Fagin et al., 2004), which resulted to be very fast and less computationally demanding than a recently proposed and very promising approach for ranking composition (Cohen-Boulakia et al., 2011). The 4 genes found

encode, respectively, the human Paired box protein Pax-7, human Paired box protein Pax-2, mouse Paired box protein Pax-8 and human Paired box protein Pax-8 (which have 1.35413 E^{-76}, 1.72295 E^{-70}, 3.22281 E^{-69} and 1.16475 E^{-67} expectation of sequence similarity to the input human Paired box protein Pax-6 isoform a protein) and all 4 genes are significantly co-expressed in tumor with a 1.0 E^{-11} p-value.

As the described methods and results demonstrate, Bio-SeCo provides a public extremely useful automated support for exploratory searches at the base of Life Science data driven knowledge discovery. It enables the user to explore the very large and very heterogeneous bio-data available, allowing he/she to easily make different attempts, inspect obtained partial results and move forward and backward in the construction of the global query that would eventually find the most relevant results, in case after several unsuccessful attempts.

Acknowledgements

This research is part of the Search Computing (SeCo) project (2008-2013) funded by the European Research Council (ERC), IDEAS Advanced Grant.

References

1. Ceri S, Abid A, et al. (2010) Search Computing: an approach for managing complex search queries. IEEE Internet Comput 14(6):14-22.
2. Cohen-Boulakia S, Denise A, Hamel S (2011) Using medians to generate consensus rankings for biological data. In: Cushing JB, French J, Bowers S (Eds.) Scientific and Statistical Database Management. LNCS, Vol. 6809. Springer, Heidelberg, D, pp. 73-90.
3. Fagin R, Kumar R, et al. (2004) Comparing and aggregating rankings with ties. Proceedings ACM Symposium on Principles of Database Systems (PODS '04). pp. 47-58.
4. Hull D, Wolstencroft K, et al. (2006) Taverna: A tool for building and running workflows of services. Nucleic Acids Res 34(Web Server issue):729-732.
5. Johnson M, Zaretskaya I, et al. (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(Web Server issue):W5-W9.
6. Lopez R, Silventoinen V, et al. (2003) WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res 31(13):3795-3798.
7. Masseroli M, Ghisalberti G, Ceri S (2011) Bio-Search Computing: Integration and global ranking of bioinformatics search results. J Integr Bioinform 8(2):166, p. 1-9.
8. Nekrutenko A (2010) Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the Life Sciences. Genome Biol 11(8):R86.
9. Parkinson H, Sarkans U, et al. (2005) ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33(Database issue):D553-D555