University of Pennsylvania
ScholarlyCommons

Database Research Group (CIS) | Department of Computer & Information Science

January 2007

BioGuideSRS: Querying Multiple Sources with a user-centric perspective

Sarah Cohen-Boulakia
University of Pennsylvania, sarahcb@seas.upenn.edu

Olivier Biton
University of Pennsylvania, biton@seas.upenn.edu

Susan B. Davidson
University of Pennsylvania, susan@cis.upenn.edu

Christine Froidevaux
chris@lri.fr

Follow this and additional works at: http://repository.upenn.edu/db_research

Cohen-Boulakia, Sarah; Biton, Olivier; Davidson, Susan B.; and Froidevaux, Christine, "BioGuideSRS: Querying Multiple Sources with a user-centric perspective" (2007). *Database Research Group (CIS)*. Paper 8.
http://repository.upenn.edu/db_research/8

© The Authors, 2007. Pre-print version. Published in *Bioinformatics*, Advance Online on March 7, 2007.
Publisher URL: doi:10.1093/bioinformatics/btm088

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/db_research/8
For more information, please contact repository@pobox.upenn.edu.

BioGuideSRS: Querying Multiple Sources with a user-centric perspective

Abstract

Summary: Biologists are frequently faced with the problem of integrating information from multiple heterogeneous sources with their own experimental data. Given the large number of public sources, it is difficult to choose which sources to integrate without assistance. When doing this manually, biologists differ in their preferences concerning the sources to be queried as well as the strategies, i.e. the querying process they follow for navigating through the sources. In response to these findings, we have developed BioGuide to assist scientists search for relevant data within external sources while taking their preferences and strategies into account.

In this paper, we present BioGuideSRS, a user-friendly system which automatically retrieves instances of data by using BioGuide on top of the SRS system. BioGuideSRS is an Applet that can be run from its web page on any system with Java 5.0.

Availability: http://www.bioguide-project.net

Keywords

bioinformatics, biological databases, paths between sources

Comments

© The Authors, 2007. Pre-print version. Published in *Bioinformatics*, Advance Online on March 7, 2007.
Publisher URL: doi:10.1093/bioinformatics/btm088

This journal article is available at ScholarlyCommons: http://repository.upenn.edu/db_research/8

BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–3

BioGuideSRS: Querying Multiple Sources with a user-centric perspective

Sarah Cohen-Boulakia [a,b*], Olivier Biton [b], Susan Davidson[b], Christine Froidevaux[a]

[a]Laboratoire de Recherche en Informatique, CNRS UMR 8023, Université Paris-Sud XI, 91405 Orsay, France; [b]Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut St, PA-19104, Philadelphia, USA

ABSTRACT

Summary: Biologists are frequently faced with the problem of integrating information from multiple heterogeneous sources with their own experimental data. Given the large number of public sources, it is difficult to choose which sources to integrate without assistance. When doing this manually, biologists differ in their *preferences* concerning the sources to be queried as well as the *strategies*, *i.e.* the querying process they follow for navigating through the sources. In response to these findings, we have developed BioGuide to assist scientists search for relevant data within external sources while taking their preferences and strategies into account. In this paper, we present BioGuideSRS, a user-friendly system which automatically retrieves instances of data by using BioGuide on top of the SRS system. BioGuideSRS is an Applet that can be run from its web page on any system with Java 5.0.

Availability: http://www.bioguide-project.net

Contact: sarahcb@seas.upenn.edu

1 INTRODUCTION

To enable scientific discovery, biological data coming from multiple heterogeneous sources must be combined. When doing this manually, scientists exhibit *preferences* concerning the sources and the cross-references to use; e.g. they trust a source that is highly curated more than one that is not. Scientists also follow different *strategies* or querying processes when they navigate through the sources, depending on the kind of answer they are interested in.

Over the past ten years, there has been an exponential increase in the number of public biological sources (Galperin (2007)), and manually choosing which sources to use has become an overwhelming task. BioGuide (Cohen-Boulakia *et al.* (2005)) was therefore designed to assist scientists with data searching, taking into account their preferences and query strategies. BioGuide generates a set of *paths* to be followed between sources, *i.e.* a ranked list of sequences of sources and links that can be used to answer a given query. In this paper, we introduce BioGuideSRS which places BioGuide on top of the popular Sequence Retrieval System (Etzold *et al*. (1996)), to automatically provide instances of data.

2 MAIN FUNCTIONALITIES

BioGuideSRS's graphical user interface is shown in part (A) of Fig. 1. The BioGuideSRS framework consists of a high-level, semantic view of the scientific domain called the *Entity graph* as well as a model of the data sources available called the *Source-Entity graph*. The Entity graph consists of biological entities (e.g. Gene, Disease) and relationships between them (e.g. causes). This graph is then mapped to the *Source-Entity graph*, which consists of linked data sources (e.g. EntrezGene, OMIM) which provide information about the entities of interest as well as the implementation of relationships (e.g. EntrezGene provides information about Genes and has a cross-reference (CrossRef) to OMIM implementing Causes).

Browsing. By clicking on an entity or relationship in the Entity Graph, users can determine which sources implement the entity or are involved in the relationship, as well as the cross-references they share. For example, in Fig. 1 (A), the user has selected the "causes" relationship (left hand side) and can visualize the network formed by cross-references between sources providing information about genes and diseases (right hand side).

Basic Querying. Users pose queries over the Entity graph by double-clicking on entities and possibly the relationships between them, and by specifying keywords to be searched for each given entity. They are then given a set of ranked paths in the Source-Entity graph representing alternative ways of implementing their query (Fig. 1 (B)). For example, the following question, Q1: *What information may I get about narcolepsy and the genes related to this disease?*, can be expressed by selecting two entities (GENE and DISEASE, in orange in Fig. 1 (A)) and by specifying "narcolepsy" as a keyword to be searched for the DISEASE entity (the entity name then turns into yellow, as Disease in Fig. 1 (A)).

Filtering and ranking. Since numerous alternative combinations of sources-entities and links can be returned, BioGuide provides advanced functionalities to filter and rank the paths. Default settings are provided based on the most frequent choices of our users.

First, BioGuide provides **Strategy criteria**, which are alternative approaches for characterizing source-entity paths corresponding to selected entities. The strategy criteria can be selected through the graphical user interface (see top of Fig. 1 (A)), and their combination forms the *query strategy*. Users can specify whether or not they want to (i) follow an order on the entities (Ordered entities); (ii) explore other, unspecified, entities (Only given entities); and (iii) visit a source more than once (Source once for all). Selecting one or several criteria ensures that only paths which meet the criteria are

*to whom correspondence should be addressed

© Oxford University Press 2006.

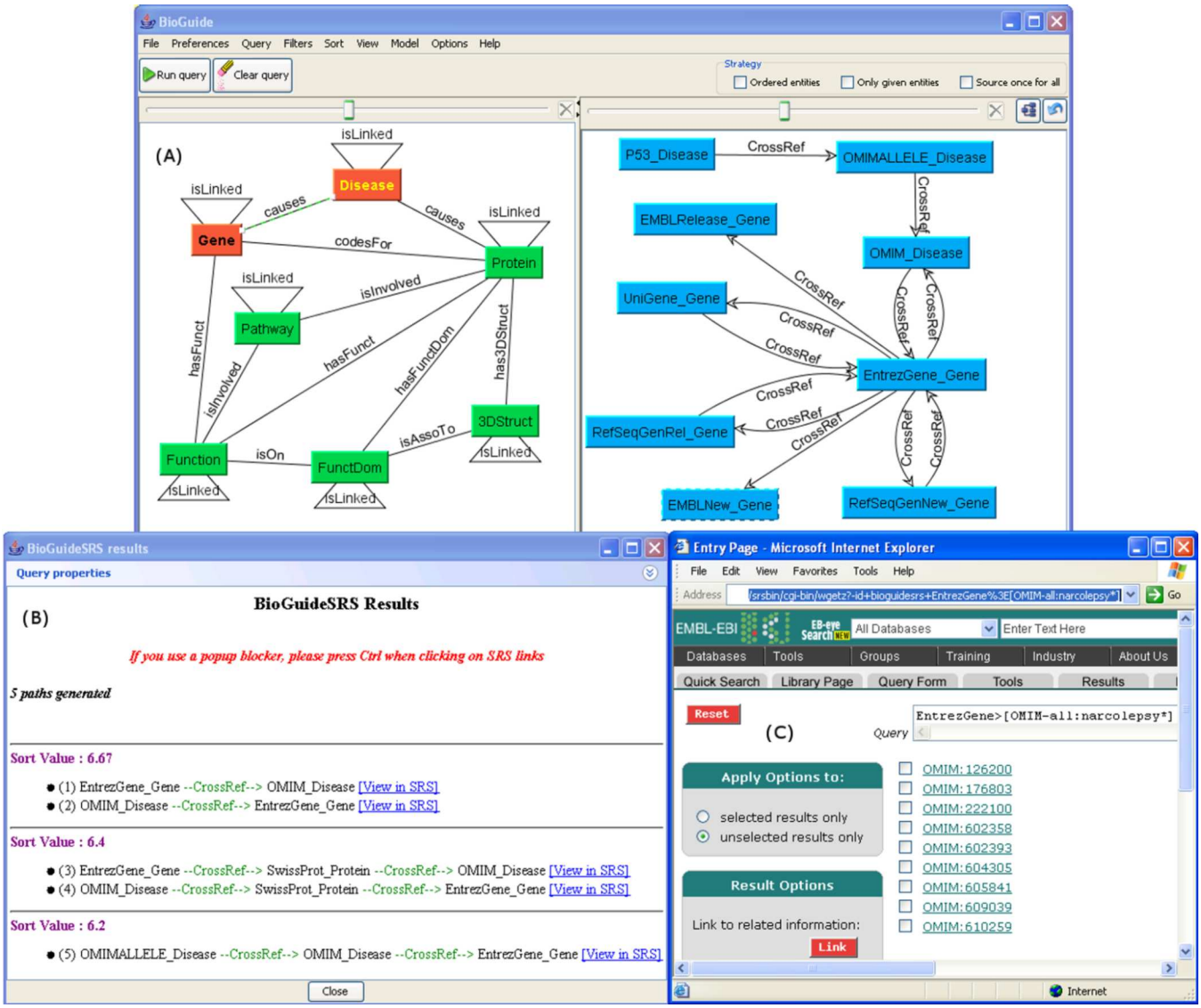

Fig. 1. (A) BioGuideSRS main graphical interface; (B) List of paths obtained; (C) Data obtained from SRS for the first path.

returned as a result. These criteria have been identified during the study of user requirement we conducted on biological data source browsing (Cohen-Boulakia *et al*. (2005)).

Secondly, BioGuide considers the user **preferences** about the sources to be used. Preference values are used by BioGuideSRS to rank as well as filter the answers according to the wishes of the user (Filters and Sort menus). Example of preference-filters include "no more than 3 cross-references must be followed per path", and "only reliable sources should be consulted". BioGuideSRS also helps scientists quantify the confidence they have in the sources by providing additional interfaces to adapt the preference values to their needs (Preferences menu, Fig. 1 (A)).

Adapting BioGuideSRS. BioGuideSRS can be customized by each user: modifying preference values, adding new kinds of preferences, adding/removing/modifying links (relationships and cross-references) and nodes (entities and sources) of the Entity and Source-Entity graphs (Model menu). The resulting configuration can then be saved to an XML file (File menu) for future use, and exchanged between users (see BioGuideSRS user manual for more information).

3 BENEFIT OF USING ALTERNATIVE PATHS

We present here results obtained for the query Q1. Assume that the user exploits the strategy provided by default (i.e. considering every ordering between entities, allowing intermediate entities and each source to be visited several times), and specifies the following preference-filters: No more than 3 cross-references must be followed per path; only very reliable sources should be consulted (reliability level higher than 7); and only complete sources should be considered for the gene entity (completeness level higher than 5). As a consequence, five alternative paths are found by

BioGuideSRS (Fig. 1, (B)). Each path describes which source can be queried to provide a given entity and which cross-reference can be followed. As an example, (1) indicates that gene and disease information can be found in EntrezGene (EntrezGene_Gene) and OMIM (OMIM_Disease), respectively. A cross-reference linking these two sources can be followed.

By selecting each of these paths, the user obtains the corresponding instances of data (e.g., instances corresponding to path (1) are shown in Fig. 1 (C)). Note that the user did not have to specify the sources to be queried nor indicate the links to be followed; paths were automatically generated. Obtaining instances of data from SRS for each path was also performed automatically by BioGuideSRS. Thus querying is automated from the beginning to the very end.

BioGuideSRS is a *multi-strategy* approach, in which complementary information is obtained and the scientist is guided in the analysis of the results. Continuing with our example, the entry giving precise knowledge about the general form of narcolepsy is found by path (3), which links genes to diseases by passing through the proteins of SwissProt, but not by path (1). BioGuide thus finds a rich set of information about the disease. On the other hand, path (5) provides a single entry, the HCRT gene, which is well-known to be responsible for narcolepsy; the HCRT gene is also found by paths (2) and (3). Knowing that this entry is given by several reliable paths increases the confidence the user has in the results.

Complete examples of use are provided on the BioGuideSRS site.

4 CONCLUSION

BioGuideSRS is a path-based system (Cohen-Boulakia *et al*. (2006)) in the same spirit as Biozon (Birkland and Yona (2006)) and BioNavigation (Lacroix *et al*. (2004)). It is the first system to provide a multi-strategy approach, allowing various querying capabilities of path systems to be expressed, and implements on-the-fly queries using SRS rather than a warehouse.

There are several advantages of BioGuideSRS: First, queries phrased in terms of *biological entities* are posed through the BioGuide user-friendly interface. Second, *Preferences* on the kind of sources to be accessed can be easily specified. Third, *Links* between the sources are systematically followed according to the strategy of the user, thus *alternative* and *complementary* ways of finding data are explored. Important information that may have been missed by a user following a single path can be found using multiple paths. Finally, an intermediate level between queries and data is offered: Each path yields a given set of data, thus the user always knows the *origin* of the data obtained (the sources and links followed), and paths can be explored one after the other following their ranked order (corresponding to their order of *preference*).

BioGuideSRS is available for use from its web site which has been accessed by more than 1300 visitors since January 2006, and several visitors have returned more than twice per month. Current BioGuideSRS users include members of the Children's Hospital of Philadelphia. The default configuration of BioGuideSRS – including the design of the graphs, their mapping, the choice of preferences – has been done in close collaboration with its users. Currently, adding a new SRS source to the system is easily done using the configuration file and the user interface; the user then has to map at least one entity with the source. In the future, BioGuideSRS may benefit from text-mining tools to automate the mapping between the graphs (mapping between entities/relationships and sources/cross-references).

ACKNOWLEDGEMENT

BioGuideSRS would not have come into being without the participation of many scientists, who are acknowledged on the web site. This research is supported by the National Science Foundation under Grants No. 0415810 and 0513778[1].

REFERENCES

Birkland. A., and Yona, G. (2006) Biozon: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*, in press.

Cohen-Boulakia, S., Davidson, S., Froidevaux, C., Lacroix, Z., and Vidal, M-E (2006) Path-based systems to guide scientists in the maze of biological data sources. *Journal of Bioinformatics and Computational Biology (JBCB)*, **4(5)**, 1069-95.

Cohen-Boulakia, S., Davidson, D., Froidevaux, C. (2005) A User-centric Framework for Accessing Biological Sources and Tools. *Proc. Data Integration for the Life Sciences (DILS)*, Springer-Verlag, Lecture Notes in Bioinformatics, **3615**, 3-18.

Etzold, T., Ulyanov, A. and Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*, **266**, 114-128.

Galperin, Y. (2007) The Molecular Biology Database Collection: 2007 update. (2007) *Nucleic Acids Research*, **35**, D3-D4.

Lacroix, Z., Raschid, L., and Vidal, M-E (2004) Efficient Techniques to Explore and Rank Paths in Life Science Data Sources. *Proc. Data Integration for the Life Sciences (DILS)*, Springer-Verlag, Lecture Notes in Bioinformatics, **2994**, 187-202.

[1] Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.